Personalized Teacher Attention and Student Achievement*

Minahil Asim † Ronak Jain ‡ Vatsal Khandelwal §

We study how personalized teacher attention affects student achievement by randomizing whether students receive messages on behalf of their teacher containing either information on past performance, information with teacher-set performance expectations, additional peer pairing for encouragement, or no message. We find that messages containing information or expectations improve math achievement by 0.18–0.21σ, with students interpreting both as signals of teacher attention. Effects are largest among low-performing students and those randomized to receive more ambitious expectations. Peer pairing only improves outcomes when students are similar. Our findings show that signaling teacher attention is a cost-effective input in the education production function.

Keywords: Teacher Attention, Expectations, Information, Peer Effects, Performance

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

*The draft was previously circulated as: *Great Expectations? Leveraging Teachers to Improve Student Performance.* 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 identify which inputs in the education production function can effectively raise student achievement. While considerable advances are being made in understanding the role of material inputs in the classroom (Evans and Popova, 2016; Glewwe and Muralidharan, 2016), the role of *relational* inputs —that is, personal interactions between students and teachers— remains less well understood. Such interactions can signal *personalized teacher attention* and affect whether a student feels noticed, monitored, or supported. In this paper, we study whether signaling teacher attention through personalized communication can improve student achievement.

Identifying the causal effects of personalized teacher attention is challenging due to various reasons. Teachers typically endogenously decide which students to pay particular attention to, making it difficult to separate the impact of teacher attention from student characteristics or unobserved classroom dynamics. Further, teachers routinely give performance feedback to students via standardized report cards, so isolating the marginal effect of additional attention is non-trivial. The effects of signaling personalized teacher attention are also theoretically ambiguous. Personalized messages conveying performance information or expectations may motivate students by signaling teacher care and raising aspirations. At the same time, such messages may discourage effort if students interpret increased attention as close monitoring or pressure. Additionally, the effect of these messages may also depend on peer interactions. Peer encouragement can be reinforcing in supportive environments but could also trigger demotivating social comparisons.

In this paper, we study how signaling teacher attention through personalized communication affects student achievement using a clustered randomized controlled trial that varies whether students receive personalized messages on behalf of their teacher and the content of those messages. We randomize students across 288 classrooms in Pakistan to receive either: (i) personalized information about past math performance, (ii) the same information bundled with a teacher-set performance expectation randomly framed as either attainable or ambitious, (iii) additional peer pairing for mutual encouragement, or (iv) no message. This design allows us to causally identify the role of signaling personalized teacher attention, test whether communicating teacher-set performance expectations adds further value in signaling teacher attention, and examine how the ambitiousness of expectations and peer interactions affect students' responses to teacher communication.

We find that both personalized communication of performance information and teacher-set expectations increase math achievement by 0.18–0.21σ, with no statistically significant difference between them. Effects are largest among lower-performing students, consistent with personalized teacher attention being particularly motivating for students at the bottom of the distribution. While communicating personalized expectations does not outperform communicating past performance information on average, students randomized to receive more ambitious expectations experience larger gains, and those with larger gaps between baseline performance

and expectations show persistent improvements 12–18 months later. In contrast, peer pairing has no average effect and only improves outcomes when students are matched with friends or peers similar in achievement and expectations; mismatched peers have lower performance, suggesting demotivating effects from negative social comparison. Together, these findings suggest that signals of teacher attention can be an effective input in the education production function.

We conducted our experiment in partnership with a large private school chain in Pakistan, catering to middle- and upper-middle-income families. To construct realistic and meaningful teacher-authored personalized messages, we reminded all teachers of each student's most recent math test score[1] and asked them to set two student-specific performance benchmarks. For each student, teachers completed both statements: (a) "I expect the student to work hard and improve to achieve at least X (out of 100%) in upcoming exams and tests" (*High Expectations*) and (b) "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" (*Very High Expectations*).[2] Eliciting these benchmarks for all students prior to randomization allows us to isolate the causal effect of *communicating* personalized expectations—holding expectation setting fixed—and to test whether more ambitious benchmarks produce stronger responses as signals of personalized teacher attention.

We then randomly divided the classrooms into four groups. In the *Information Arm*, students received a private message—sent on behalf of their teacher—reporting their past math performance. In the *Expectations Arm*, students received the same performance information bundled with a personalized teacher-set performance benchmark, framed either as the *High Expectations* or *Very High Expectations* statement. In the *Peer Arm*, students received the expectations message and were additionally paired with a randomly chosen classmate for mutual encouragement. Students in the *Control* group received no message. Messages were delivered privately through the school's online platform, and teachers were blinded to student treatment status to prevent differential instruction or encouragement across students.

This design delivers three sources of identifying variation. First, randomizing the delivery of personal messages from the teacher identifies the effect of signaling personalized teacher attention. Second, comparing expectations to information isolates whether teacher-set goals add value beyond past performance information. Third, random variation in goal ambitiousness and peer pairing allows us to assess when ambitious goals and peer interactions strengthen or weaken students' responses to personalized teacher communication.

Our main outcomes are student scores in standardized, high-stakes Mathematics and English exams, administered 2–4 weeks and six months after the intervention. To study longer-term effects, we use administrative test score data 12 and 18 months after the start of the intervention, and we conduct a follow-up student survey to understand how the messages are interpreted.

[1] This mitigates concerns that teachers may not know students' performance (Djaker et al., 2024) or that reported benchmarks reflect stereotypes rather than ability. We find no systematic differences in elicited benchmarks by student gender, wealth, or age (Figure A.4.1).

[2] Teachers were informed that these statements may be communicated to students, but did not know for which students or classes.

The experiment yields three main findings. First, signaling teacher attention through personalized communication substantially increases math achievement. Students who receive teacher-set expectations score 0.21σ higher than the control group ($p<0.01$), and students who receive performance information alone score 0.18σ higher. Importantly, the effects of communicating personalized expectations and past performance information are statistically indistinguishable, suggesting that it is the act of personalized teacher communication—rather than its precise content—that drives much of the impact. This is also consistent with students interpreting both messages as personalized attention and encouragement from the teacher in our follow-up survey. Effects are concentrated among lower-performing students, consistent with teacher attention being particularly meaningful at the bottom of the distribution.

Second, while expectations do not outperform information on average, the content of expectations still matters. Students randomly assigned to receive the more ambitious (*Very High*) expectation experience larger gains, and treatment effects increase with the gap between the communicated benchmark and the student's baseline performance: a 10 percentage point increase in this gap raises achievement by about 2 percentage points.

Third, peer matching does not improve outcomes on average and reduces achievement relative to communicated expectations alone. However, peer effects are sharply heterogeneous: encouragement helps when students are paired with friends or peers who are similar in baseline achievement, but it harms performance when paired with higher-performing peers or peers with higher teacher-set performance expectations. Survey responses indicate that unfavorable social comparisons play an important role, highlighting peer interactions as a constraint on scaling peer-based encouragement.

As expected, in the longer term (12 and 18 months after the start of the intervention), when we no longer sustain personalized messages, we do not detect any positive average treatment effects. However, students whose teacher-set expectations substantially exceeded baseline performance initially continue to perform better, suggesting that ambitious goals can generate persistent gains for a subset of students even when average effects fade. Despite detecting modest negative spillovers to English in the short term, we also do not detect any negative spillovers in the longer run.

Our paper makes three contributions. First, we provide causal evidence that signaling teacher attention through personalized communication can substantially improve student achievement. In doing so, we contribute to the literature on information provision to parents and feedback to students (Andrabi et al., 2017; Barrera-Osorio et al., 2020; Bobba and Frisancho, 2022; Friedlander, 2020) by demonstrating that personalized performance messages can be helpful even when they provide little new information, with students interpreting them as signals of teacher attention and care. Importantly, by experimentally separating teacher-set expectations from performance information, we show that performance information alone can be as effective as communicated expectations. This is informative because performance information and high expectations are often bundled together in successful schooling models and reforms (Angrist et al., 2013; Fryer Jr, 2014) and our design allows us to isolate their impacts.

Second, we evaluate the causal effect of teacher-set expectations as a form of personalized communication. Expectations are typically endogenous and selectively conveyed (Friedrich et al., 2015; Jussim and Harber, 2005; Papageorge et al., 2020), making their causal effects difficult to identify. Our design holds expectation formation fixed by eliciting student-specific expectations from all teachers prior to randomization and then varying whether these expectations are communicated. We show that communicating expectations improves achievement, and that more ambitious expectations generate larger gains, alleviating concerns that ambitious goals necessarily discourage effort. In doing so, we also complement the large literature on goal-setting interventions (Damgaard and Nielsen, 2018; Dobronyi et al., 2019; Morisano et al., 2010; Oreopoulos and Petronijevic, 2019; Schippers et al., 2015) by showing that ambitious teacher-set goals can be particularly effective at improving student achievement.

Third, we contribute to the large literature on peer effects in education (e.g.: Bifulco et al. (2011); Bursztyn et al. (2019); Calvó-Armengol et al. (2009); Jackson et al. (2023); Lavy et al. (2012); Shan and Zölitz (2025); Wu et al. (2023)) by showing that the effectiveness of peer encouragement depends on peer characteristics. Exploiting random peer assignment, we find that peer encouragement helps only when peers are friends or similar in achievement and teacher-set performance benchmarks, but can harm outcomes otherwise. This highlights social comparisons as a constraint that limits the effectiveness of peer-based interventions.

From a policy perspective, signaling personalized teacher attention is low-cost, non-invasive, and highly scalable, meeting the generalizability criteria proposed in List (2022). Delivering the intervention leverages existing school infrastructure and costs less than ten cents per 0.1σ gain in test scores, placing it among the most cost-effective learning interventions documented (Beteille and Evans, 2019; Glewwe and Muralidharan, 2016).[3] Importantly, our partner school chain resembles many higher-income school systems in that class sizes are relatively small (average of 20 students) and students receive regular performance feedback. The fact that personalized messages still generate sizable gains in this context suggests that effects may be even larger in low- and middle-income settings where larger class sizes and higher student–teacher ratios make signals of personalized teacher attention even more valuable.

We proceed as follows. Section 2 describes the setting; Sections 3 and 4 outline the design and empirical strategy; Section 5 presents the results; Section 6 discusses mechanisms and cost-effectiveness; and Section 7 concludes.

2 Context

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). We partnered with a large private school chain 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.

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

We conducted our study in grades 3 to 8 across 288 classrooms in 15 geographically dispersed schools (Appendix Figure A.1).[4] 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. Nearly 90% of the schools report that parents have medium or high levels of literacy. The average class size is around 20.

2.1 Data Sources

2.1.1 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 (2) December 2022, i.e., after 12 and 18 months after the start of our intervention.

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.

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. Then, we conducted a follow-up student survey six months after the end of the intervention (June 2022) to understand how students interpreted various components of 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.[5]

2.1.3 Personalized Messages from Teachers

We collected data from all teachers to construct personalized messages for students. To do this, we reminded teachers of each student's most recent math test score and asked them to set two

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

[5] Additionally, we conducted two rounds of online surveys and independent tests with students during the intervention period and two rounds of surveys with teachers before and after the intervention. Response rates were lower than expected, which limits statistical power to detect treatment effects using these measures. We therefore do not focus on these results in the main text, but report them in an online appendix for completeness.

student-specific performance benchmarks by filling 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 statements for all students, 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.

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.1.3). About a third of teachers report concerns about class-

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.

room disruption, attendance, or students not completing their homework (Table A.1.2). 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 performance 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 attention to students through performance-related expectations but do not prioritize students at the bottom end of the score distribution when 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. In fact, 53% re-

port that performance-related expectations should be conveyed to those who would be most successful in encouraging others. This motivated the inclusion of the peer pairing arm in our study design.

3 Experiment Design

Details of the design were pre-registered with the AEA RCT Registry under AEARCTR-0007846. 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 personalized information about past math performance is bundled with a teacher-set performance expectation conveyed individually to a student), one-third to the Peer Arm (where in addition to conveying student-specific high teacher expectations individually to a student, students were randomly matched with another classmate and asked to encourage each other),[6] and one-third to a Comparison Group. Half of the Comparison Group classrooms were randomized to receive a personalized reminder about their last test score (Information Arm), and half were randomly selected to receive no messages (Control Group). Further, half the students in the Expectations and Peer Arms were randomly chosen to receive the "High" teacher expectation statement, and half received the "Very High" teacher expectation statement with the corresponding statements outlined earlier (Section 2.1.3).

Importantly, all teachers were blind to the treatment status of students to ensure they did not selectively change their efforts towards any students. The randomization was stratified along grade,[7] 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. Using historical test score data from our setting, we also conducted power calculations indicating that the experiment can detect minimum effects of 0.13 standard deviations between treatment and control arms.[8]

3.1 Timeline

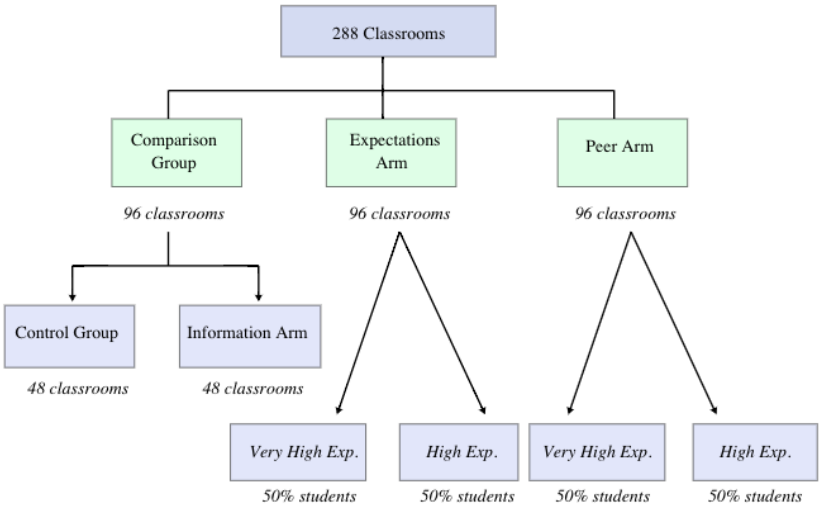
The timeline is as follows. Informed parental consent and student assent were obtained between March and May 2021. Personalized teacher messages were elicited and delivered by mid-June. We collected administrative test score data on student performance in June/early July. We sent two reminders to students –one at the start of the summer holidays and another at the beginning of the new academic year in August. A final round of personalized messages with updated design graphics and scores was sent in November 2021 before the school conducted its end-of-term exams in December 2021. A follow-up survey with school administrators and students was conducted between March to May 2022. Long-term student test score

[6] 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.

[7] 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).

[8] Power calculations and deviations from the pre-analysis plan are detailed in Supplementary Appendix Section H.

Figure 1: Randomization Design



data were collected in June and December 2022.

3.2 Format and Delivery of Personalized Messages

We delivered personalized messages via emails using the virtual learning infrastructure (Google Classrooms)[9] 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.

Figure 2 shows the designed graphics sent out to each group. Students in the Information Arm received a graphic with a simple image of a boy or a girl with their most recent Math score. The graphic used to deliver teacher expectations positions each student as a superhero who can work towards achieving the teacher-set performance 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.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.

[9]This platform was regularly used by teachers to engage digitally with students.

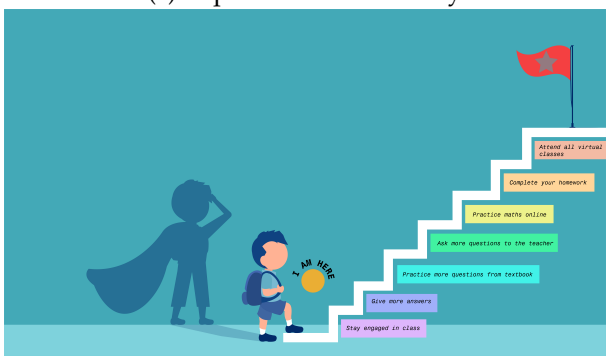
For the second round of intervention, we re-designed the graphics (Appendix Figure A.2.2). The performance information displayed was updated using each student's most recent math score, but the teacher-set expectations were not updated. Keeping expectations fixed preserves the experimental variation from the initial elicitation and avoids endogeneity that would arise if teachers revised expectations in response to interim performance or treatment-induced improvements. These graphics were emailed before the end-of-term exams.

Figure 2: Treatment Delivery Design Variations - Round 1

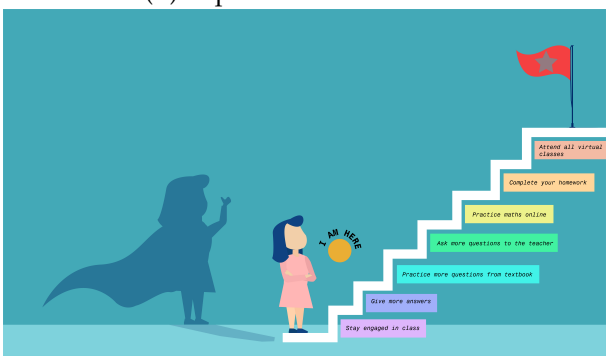
(a) Information Arm - Boy

(b) Information Arm - Girl

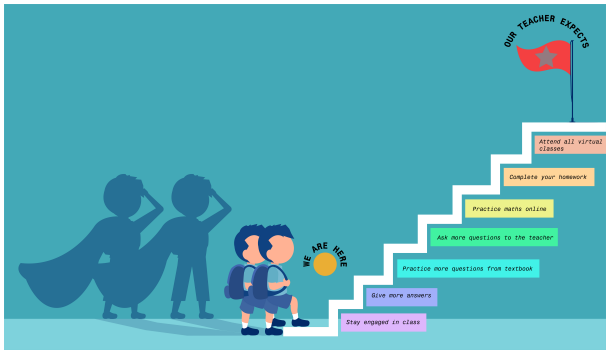
(c) Expectations Arm - Boy



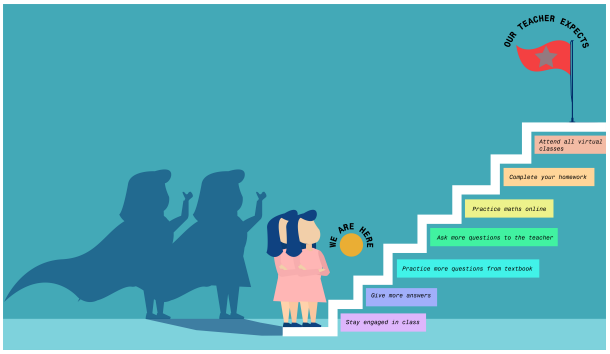
(d) Expectations Arm - Girl



(e) Peer Arm - Boy



(f) Peer Arm - Girl



3.3 Descriptives for Teacher-set Performance Expectations

We now briefly describe teacher-set performance benchmarks 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.4.3 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 expectations, we minimized the risk of gender or wealth-related stereotypes driving their expectations. Figure A.4.4 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.4.4 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.

Finally, among students assigned to receive expectations in the Expectations and Peer arm, 78% have teacher expectations that are equal to or exceed their baseline performance. For the remaining students for whom teacher expectations were below their past performance, the infographic displayed only the teacher's expectation and omitted the prior score.[10] We do not exclude these students to avoid any selection effects and interpret our treatment effects as conservative ITT estimates. Importantly, the share of students with expectations below baseline performance does not differ across treatment arms.

4 Empirical Strategy

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.[11] We use the pooled sample combining data from the midline and endline waves for the main results to maximize power.[12]

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

[10] This was done to avoid risks of potential demotivation, as required by our IRB.

[11] 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. Therefore, 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.

[12] 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.

Standard errors are clustered at the class level (unit of treatment). We include fixed effects ϕ_s for each stratum s and round fixed effects λ_t for midline and endline waves. We present results on standardized test scores and raw test scores, controlling for baseline student performance in a value-added specification in the latter case.

4.1 Balance

4.1.1 Balance in 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.3.4, B.1, and B.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.3.5. 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).

4.1.2 Balance in Teacher-set Performance 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.4.3). 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.3.4, Figure A.4.3), 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

5.0.1 Effect on Math Performance

Table 2 presents the treatment effects from our (pre-registered) 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.[13] 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.[14] This suggests that receiving a personalized message on behalf of the teacher that contains a reminder about the student's past performance can also increase student performance and be just as effective as a message with teacher expectations. We find that this is driven by students interpreting the messages as a signal of teacher attention and care, which we discuss further in the next section.

The results from the pooled specification are also consistent with the treatment effects estimated separately for the midline and endline waves (Tables C.1 and C.2). Further, as shown in Table C.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 6 months if signals of teacher attention are sustained through reminders. Together, these findings suggest that students are highly responsive to personalized teacher attention.

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.[15]

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 G.1).

[13] 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.

[14] 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.

[15] 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.

5.0.2 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 C.4 and C.5). This could suggest that, over time, subject-specific expectations may induce students to reallocate effort toward subjects in which they receive greater attention from the teacher.

5.1 Heterogeneous Effects

5.1.1 Magnitude of Expectations

First, we exploit the exogenous variation in the type of teacher-set performance expectation 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 expectation benchmarks 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 performance expectation benchmark 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 impact of the Expectations Arm. This implies that receiving an ambitious benchmark 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.

5.1.2 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.[16] As shown in Table A.5.7, 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-set performance benchmarks, classroom motivation, parental wealth, and number of friends in the classroom. We construct the index by first generating the squared differences in terms

[16] 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.

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 Pan-

els 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 0.38σ and 0.28σ, 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 0.31σ higher than being matched with someone with a higher baseline score.

When compared with the Control Group, Appendix Table D.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.

5.1.3 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. 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 (Figure A.5.6). 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.5.6 where we employ the method proposed by Abadie et al. (2018). We predict 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.5.6 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, personalized teacher attention through information and performance benchmarks is particularly motivating for students at the bottom of the distribution.

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.

5.1.4 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 Expectations and Peer Arms. Figure A.5.5 shows the results.

We then categorize individuals into four groups based on their predicted performance under treatment, ranging from lowest to highest. Comparing baseline characteristics across these groups reveals systematic differences in students' baseline scores and in their matched peers' relative achievement, underscoring the importance of our previous findings. The procedure is discussed in the Supplementary Appendix Section C.1.

5.2 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.[17] We do not find any significant average treatment effects in Math for the Expectation, Peer, or Information Arms, as shown in Table A.7.8. Notably, these test scores capture a time period of six months and a year, respectively, without receiving any communication from the teacher. This shows that reminders are critical for sustaining the impact of teacher attention and care in the long run.[18] Additionally, we do not detect any effects on English test scores as shown in Table A.7.9.

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.7.10 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

We use school administrative data, surveys with head teachers, a follow-up survey with students six months after the end of the intervention, and findings from the heterogeneity analysis to explore potential mechanisms underlying the treatment effects.

6.1 Why does information about past performance improve outcomes?

Given that both the Information and Expectations Arms include information about past performance, and their effects are statistically indistinguishable, we interpret that the impact arises primarily due to personalized teacher communication itself—likely a salient signal of teacher attention—rather than from the specific content. We therefore first focus on understanding

[17] 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.

[18] 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.

how students interpreted the information about past performance.

How did students interpret the personalized message? To understand how this information was interpreted, we draw on evidence from our follow-up student survey. Responses suggest that students interpret personalized messages delivered on behalf of their teacher as a signal that their teacher pays attention to them. Nearly 35% of students in our follow-up survey who viewed the Information Arm infographic interpreted that the teacher intended 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.8.9a). When asked about how they would feel in response to the image, 51% of students said that they would feel happy or motivated to receive this message from the teacher (Figure A.8.9b). This suggests that the belief that it is a targeted, personalized message from the teacher is effective in motivating students.

Did the infographic provide new information? We also examine 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.[19] 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). We also find no heterogeneity in treatment effects by schools that send report cards more frequently. Furthermore, 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 solely by pure information effects.

Did the infographic provide information in an accessible format? 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 E.1). It is also unlikely that the treatment effects observed in this arm can be attributed solely to the format of the information delivered. This is because expressing scores as a percentage (out of 100%) used in our intervention aligns with common practices in schools: 69% of the schools in our sample give out scores in percentages, and 88% of schools use raw scores.

At the same time, the ease of interpreting information could be an additional factor contributing to the effects of this arm. Seventy-five percent 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 E.2 shows significantly lower effects of information in schools with high parental literacy compared to schools with lower parental literacy. Supporting this, over 50% of students reported that even though they remember their scores, they still find the image helpful as a reminder.

[19] Report cards are standardized and indicate scores from all subjects separately. Importantly, the report cards do not combine any other teacher evaluations, like participation or homework.

6.2 Do teacher-set performance expectations provide additional benefits?

Although the Expectations Arm does not outperform the Information Arm on average, we find evidence that the ambitiousness of communicated goals matters. In particular, more ambitious expectations and larger gaps relative to baseline scores predict larger gains, suggesting an additional motivational role for ambitious teacher expectations beyond the attention signal common to both treatment arms.

Does the magnitude of the expectation matter? We find that those randomized to receive the "Very High" or especially ambitious 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 is consistent with students working toward the expectation set for them by their teachers when this was communicated to them. In particular, Figure A.6.7 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.

How did students interpret the message? Consistent with the above findings, 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 signal about how smart they are, or inferring that they are lagging or being monitored (Figure A.8.9a). Seventy-six percent reported that they would feel motivated or happy if they were sent the image, as opposed to feeling stressed or disappointed (Figure A.8.9b). At the same time, 92% reported feeling motivated by teacher expectations. Our follow-up survey also reveals that the majority of students find the potential for improvement most helpful in the infographic, rather than the information about their previous score or their expectation considered separately (Figure A.8.8).

Did the provision of tips in the infographic also affect the observed treatment effects? We find that the generic tips on how to improve (delivered in the infographic in this treatment arm) are unlikely to drive the treatment effects. In particular, they 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.).

Did setting expectations lead to changes in teacher behavior? We note that the observed treatment effects cannot be driven by changes in teacher behavior, as teachers were blind to student treatment status by design, and expectations were elicited from all teachers in both the control and treated classrooms. Supporting this, over half of the students in our endline survey confirmed that their teacher did not spend extra time with them after class, with no

statistically significant differences between treatment and control groups.

6.3 Why does peer pairing have heterogeneous effects?

While the Information and 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 find that morale effects due to interpersonal comparisons between matched students are likely an important factor driving heterogeneity in treatment effects. Importantly, the effects do not arise due to unfavorable classroom norms that discourage effort. Our baseline survey evidence suggested 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.

However, while the performance and expectations of the peer were not revealed to either student, we find that around one-third of students reported that they would try to find out what their peer scored and the teacher's expectations for them. As shown in Figure A.8.10, the majority of the students reported that they would feel disappointed or stressed 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. This pattern is consistent with the heterogeneity in treatment effects we observe when the matched peer differs in characteristics such as baseline scores and teacher expectations.

6.4 Cost-benefit Analysis

Our findings offer encouraging evidence of the potential scalability of communicating personalized teacher messages 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.9.11) 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 messages 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).

6.5 Generalizability

We find encouraging evidence on all four dimensions of the SANS framework (Selection, Attrition, Naturalness, and Scaling) when assessing scalability and generalizability (List, 2022).

First, selection is unlikely to pose a concern in our setting, as classrooms were randomly chosen from an existing large private school network. As 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 Pakistan. If anything, we expect the effects of personalized teacher communication to be at least as large in public and low-cost private schools, where larger class sizes and higher student-teacher ratios may limit the amount of individualized attention teachers can provide (Qureshi and Razzaque, 2021).

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: performance expectations were elicited from regular class teachers and delivered using existing school channels, closely mirroring real-world conditions. We find little evidence that the detected effects arise 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.

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. Importantly, our setting also resembles many higher-income contexts in that students receive performance information regularly through report cards. The fact that personalized performance reminders still generate sizable gains suggests that the intervention operates less through correcting information constraints and more through signaling personalized teacher attention. In this sense, the results highlight the role of relational inputs in schooling—how signals of attention through teacher-initiated personalized messages can motivate effort even when performance information is already conveyed through report cards. 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 studies how signaling teacher attention through personalized communication affects student achievement. Using a clustered randomized experiment, we vary whether students receive personalized messages from teachers and the content of those messages: past-performance information, student-specific performance goals (framed as attainable or ambitious), and peer encouragement through randomized pairing. We find that signals of personalized teacher attention substantially increase math achievement. Communicated teacher-set expectations raise scores by 0.21σ and past-performance information raises scores by 0.18σ, with effects statistically indistinguishable on average. Gains are concentrated among lower-

performing students, and more ambitious goals generate larger improvements. Additional peer pairing does not improve outcomes on average: it helps when peers are similar or friends, but has a negative effect when pairings create unfavorable interpersonal comparisons.

Together, these results suggest that teacher-initiated personalized communication is a powerful and scalable input in the education production function. The fact that performance reminders are as effective as communicated high performance expectations—despite widespread report-card provision—indicates that the impact primarily operates through the signal of personalized teacher attention rather than through new information. Additionally, variation within the expectations arm shows that the ambitiousness of communicated goals matters, highlighting the importance of providing tailored yet ambitious goals. At the same time, results from the Peer Arm highlight social comparisons as an important constraint: peer pairings can motivate, but can also reduce effort when such comparisons are unfavorable.

From a policy perspective, signaling personalized teacher attention through performance reminders or ambitious expectations is low-cost, non-invasive, and easy to embed in routine school operations. A 0.1σ increase in test scores costs approximately \$0.08 per student in the Expectations Arm and \$0.09 in the Information Arm, placing these interventions among the most cost-effective strategies for improving learning outcomes. This is particularly relevant in resource-constrained settings that face persistent learning shortfalls (World Bank, 2017).

Two directions for future work are promising. First, understanding the complementarities between personalized teacher communication and parental engagement may help to reinforce and sustain student motivation over time. Second, future research can test how personalized teacher messages across multiple subjects affect effort allocation across subjects, course choices, and long-term educational aspirations.

A Tables and Figures

A.1 Context

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 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.

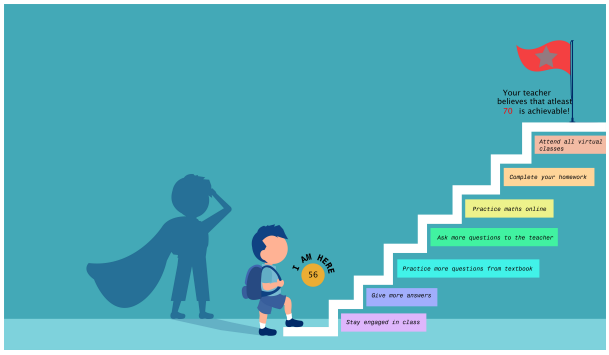
Table A.1.3: Summary Statistics of Teachers

	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

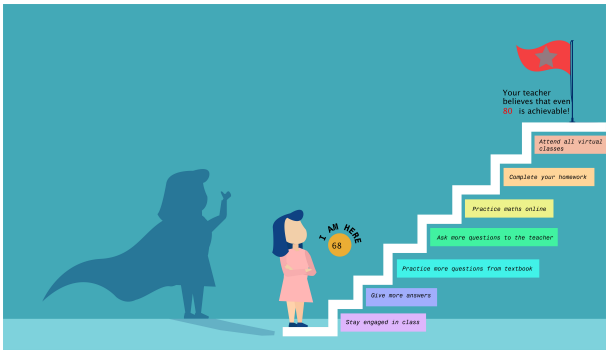
Note: The statistics are from the baseline teacher survey. We asked teachers to rank from 1-3 who they thought would benefit the most from the communication of teacher expectations, e.g., 52% of teachers ranked the top of the achievement distribution as 1.

A.2 Treatment Infographics

Figure A.2.1: Treatment Delivery Illustrations - Round 1



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

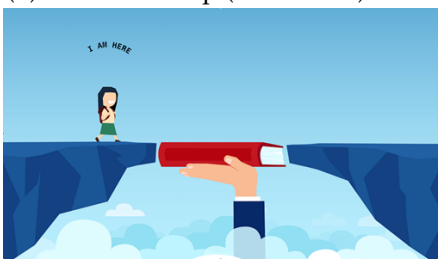


(b) Illustration for Student-Specific "Very High" Teacher Expectation – Girl

Figure A.2.2: Treatment Delivery Variations- Round 2

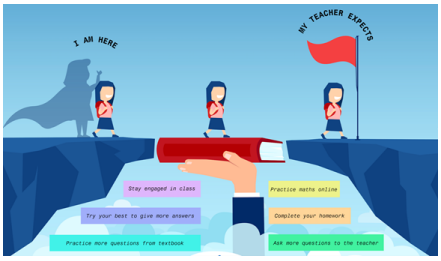
(a) Control Group (with Score) - Boy

(b) Control Group (with Score) - Girl



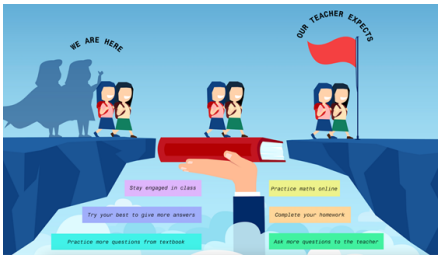
(c) Individual Arm - Boy

(d) Individual Arm - Girl



(e) Peer Arm - Boy

(f) Peer Arm - Girl



A.3 Balance Tables

Table A.3.4: 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).

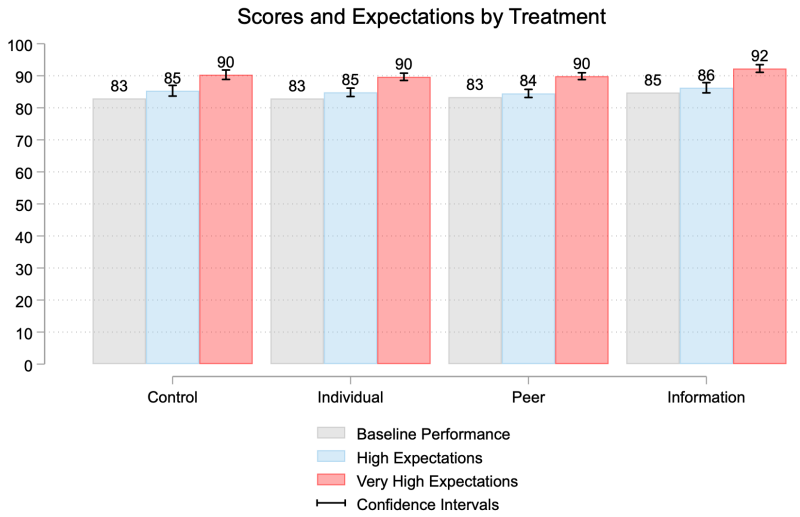
Table A.3.5: Balance Table of Class Characteristics

	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.

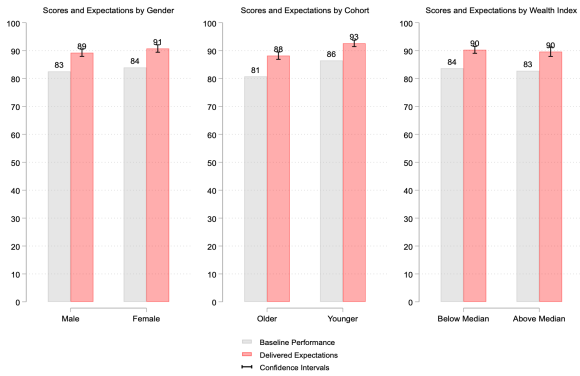
A.4 Teacher Expectations

Figure A.4.3: Raw scores, High and Very High Teacher Expectations

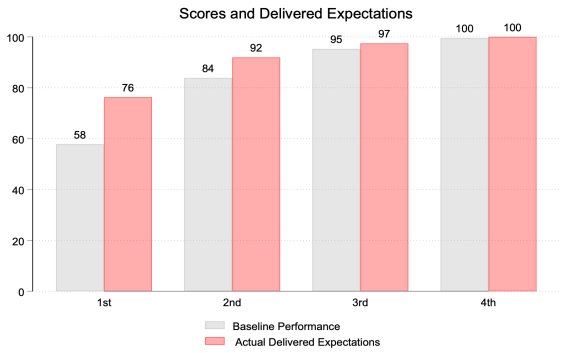


Note: The figure shows students' baseline math scores and the elicited 'High' or 'Very High' teacher expectations across treatment and control arms.

Figure A.4.4: Teacher expectations across student demographics and baseline score quartile.



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

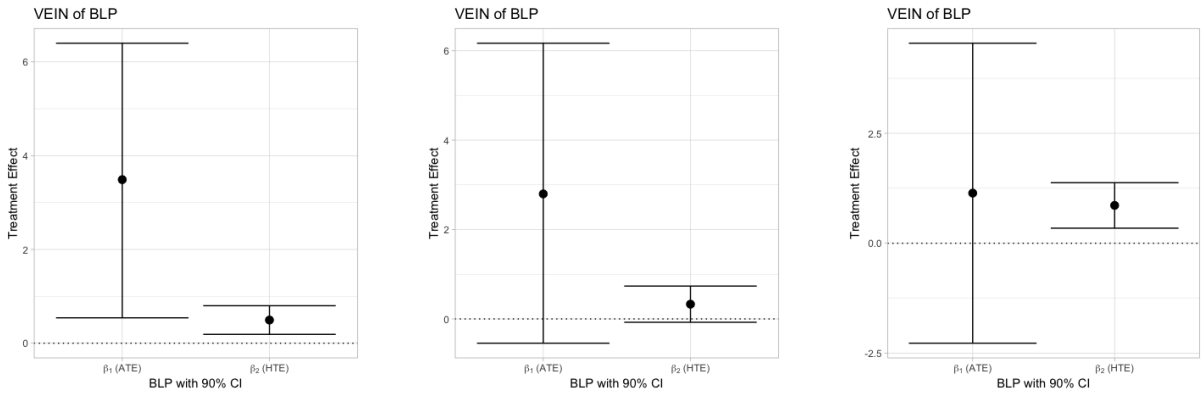


(b) Teacher Expectations by Treatment Arms

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

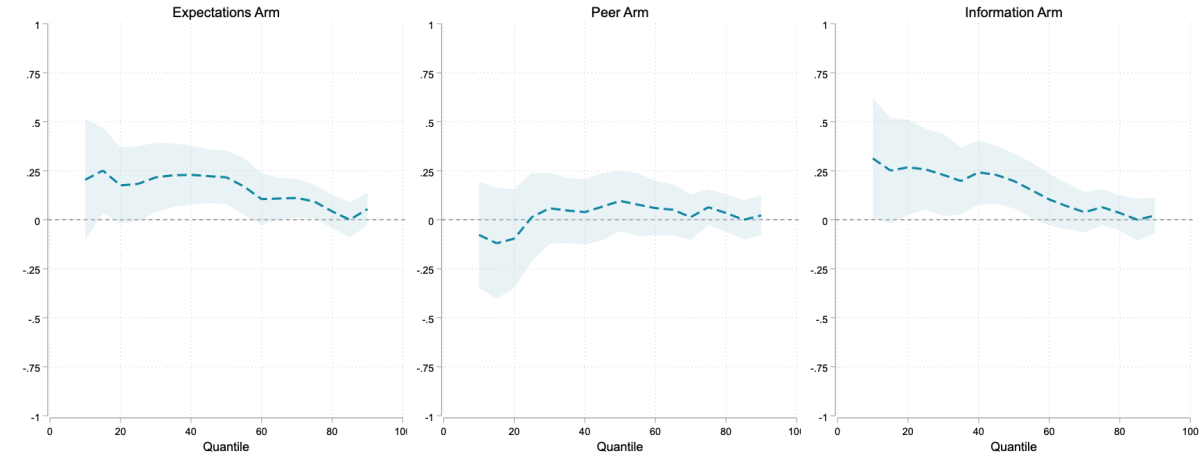
Figure A.5.5: Heterogeneous Treatment Effects



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

Note: We apply the method of Chernozhukov et al. (2018) to test for heterogeneity in treatment effects. The sample is repeatedly split into two equal parts. In the first part, machine learning methods (Lasso, SVM, and Random Forest) are used to model test scores as a function of baseline characteristics separately for treated and control students. These models are then used in the second half to predict potential outcomes under treatment and control, yielding a predicted individual treatment effect $S(Z_i)$. Test scores are then regressed on the treatment indicator, its interaction with $S(Z_i)$ (capturing heterogeneous effects via β_2), and additional controls including strata and round fixed effects. Standard errors are clustered at the class level. Results shown correspond to the median coefficients corresponding to the best-performing learner across splits.

Figure A.5.6: 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.5.6: Heterogeneous Treatment Effects by Predicted Performance: Leave One Out Estimator

VARIABLES	(1) Group 1	(2) Group 2	(3) Group 3	(4) 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.

Table A.5.7: Treatment Effects by Peer Friendship Status

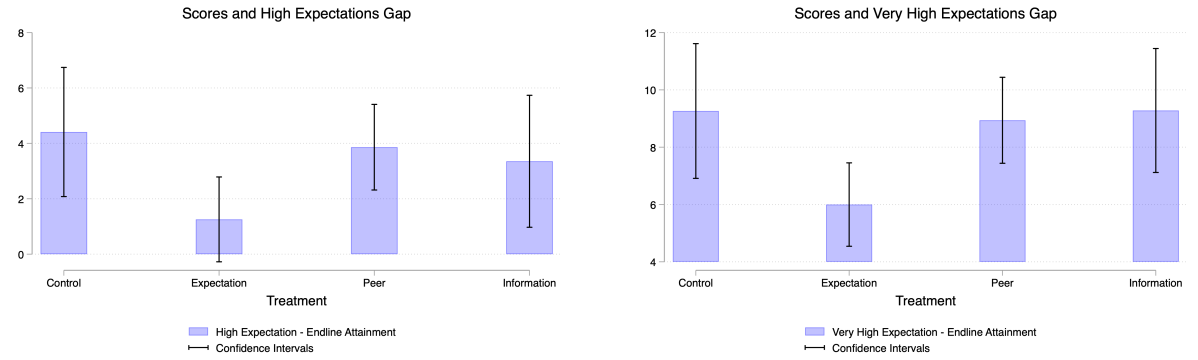
	(1) Standardized	(2) Raw Scores
Friend	0.245*	3.451*
	(0.135)	(1.772)
Baseline Score		0.452***
		(0.076)
Constant	-0.183	40.452***
	(0.127)	(6.484)
Observations	595	589

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.6 Gap Between Teacher Expectations and Endline Performance

Figure A.6.7: Gap Between Teacher Expectations and Endline Achievement by Treatment Arm



Note: The left panel presents the gap between the "High Expectations" and endline performance by treatment arm. The right panel presents the gap between "Very High Expectations" statement and endline performance by treatment arm. Error bars represent 95% confidence intervals.

A.7 Long Run Results

Table A.7.8: 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.7.9: 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.

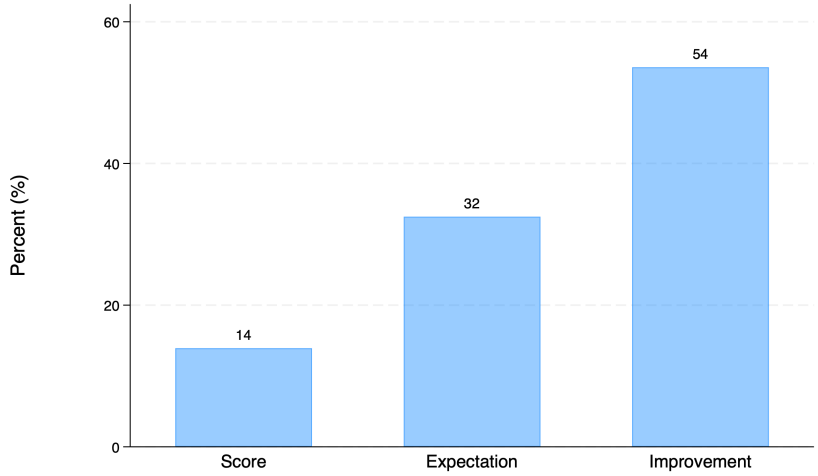
Table A.7.10: Treatment Effects on Long Run Math Test Scores by the Gap between Expectations and Baseline Score

	(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.

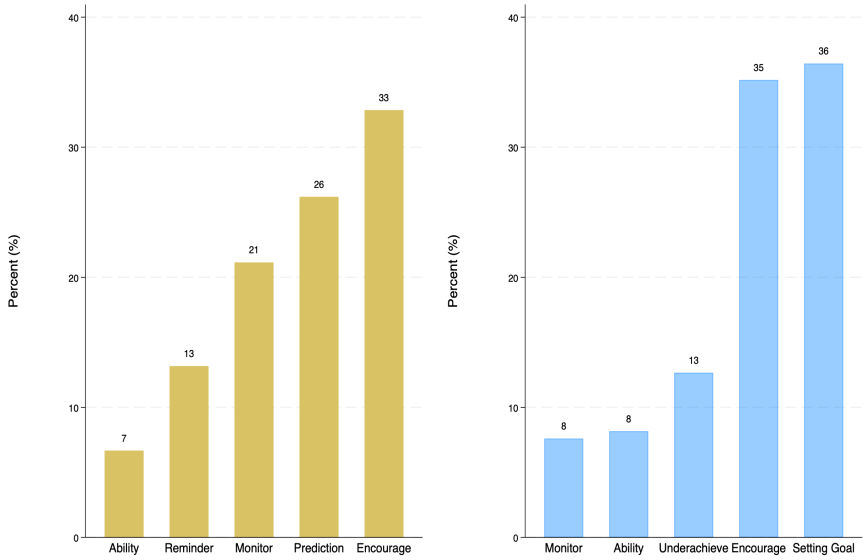
A.8 Student Interpretations: Follow-up Survey Results

Figure A.8.8: What Students Notice in the Expectations Arm Image

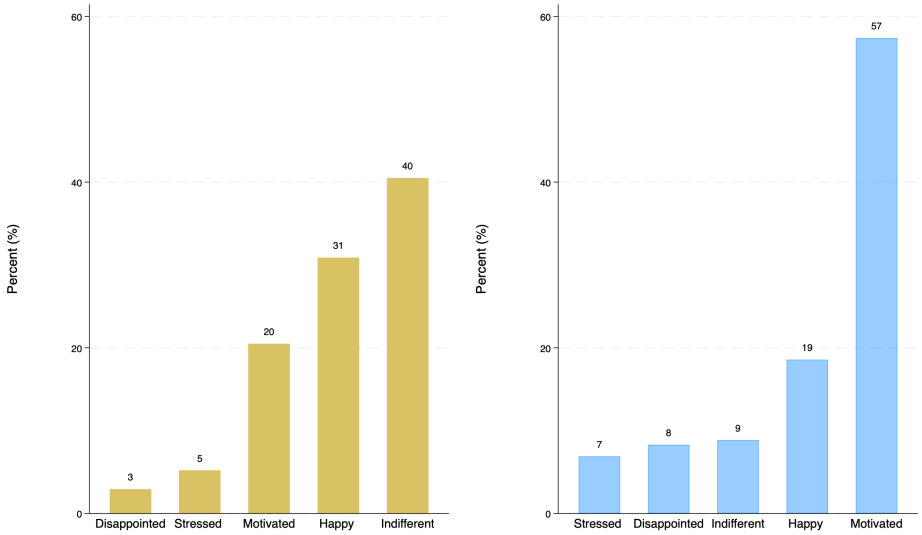


Note: The student follow-up survey sample size was 997 students. The figure illustrates survey responses to the question: 'What do you notice most or find most helpful in this picture?' Respondents had three options: 'Information about your current performance' (labeled as 'Score'), 'How much I can improve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

Figure A.8.9: Inferences and Feelings about Images - Expectations v/s Information Arm



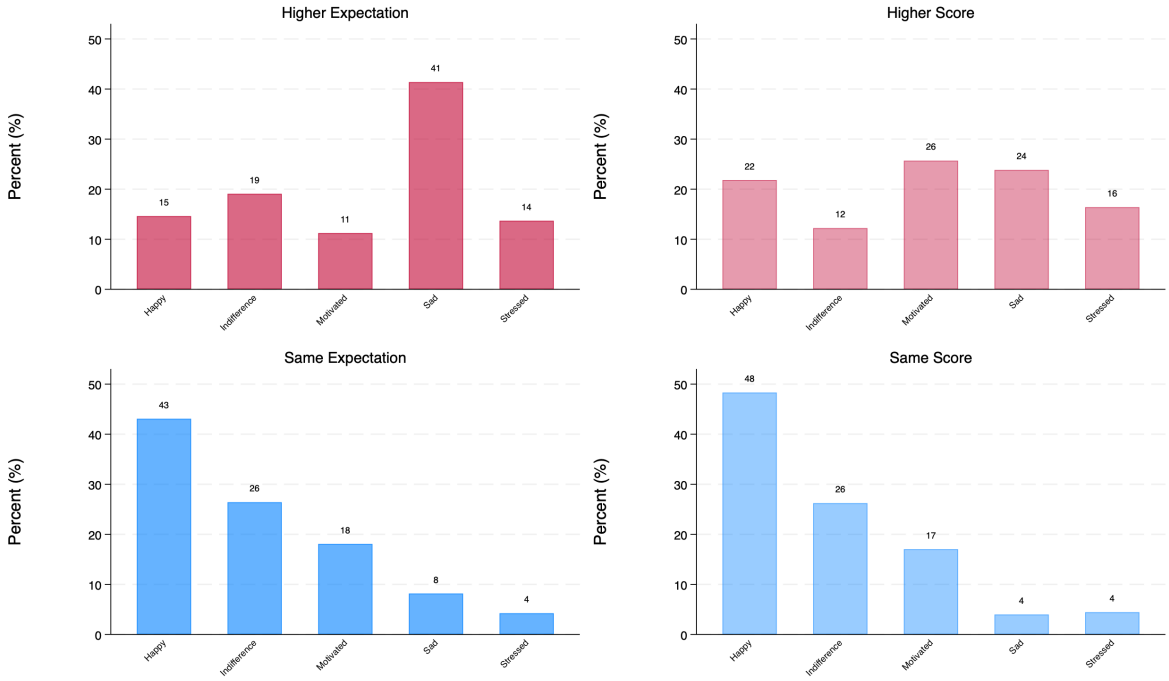
(a) Student Inferences from the Information and Expectations Arm Images



(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.8.10: Feelings about Peer Matching Scenarios



Note: This figure presents students' reactions when asked how they would feel if they were matched with a peer with higher teacher expectations (top left), higher baseline score (top right), same teacher expectations (bottom left), or same baseline score (bottom right).

A.9 Cost-effectiveness

Table A.9.11: 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).

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.

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.

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.

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.

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.

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.

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

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.

Evans, D. K. and Popova, A. (2016), 'What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews', *The World Bank Research Observer* **31**(2), 242–270.

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.

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.

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., 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.

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* .

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.

Shan, X. and Zölitz, U. (2025), 'Peers affect personality development', *Review of Economics and Statistics* pp. 1–45.

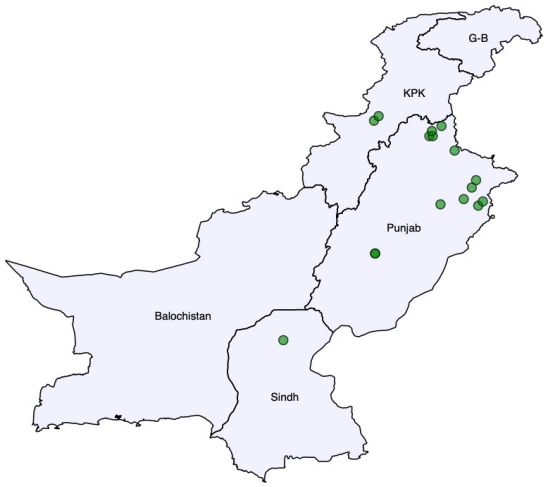
World Bank (2017), *World development report 2018: Learning to realize education's promise*, The World Bank.

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.

A Supplementary (Online) Appendix

A Context

Figure A.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.

B Additional Balance Tests

Table B.1: Balance Table of Student Characteristics (Midline Student Scores Sample)

	Mean				P-values		
	(1)	(2)	(3)	(4)			
	Control	Exp	Peer	Info	(1)-(2)	(1)-(3)	(1)-(4)
Baseline Math Score	82.93	82.92	83.33	84.77	0.64	0.89	0.34
Female	0.44	0.48	0.34	0.38	0.01**	0.01**	0.45
High Parental Income	0.15	0.14	0.09	0.13	0.77	0.18	0.74
Adults peer Room	0.57	0.55	0.55	0.59	0.55	0.60	0.14
High Parental Literacy	0.10	0.11	0.10	0.06	0.77	0.76	0.78
Number of Friends in the Classroom	4.16	4.19	3.90	4.06	0.25	0.05*	0.93
Weekly Hours Studying Math	4.04	3.74	3.88	3.45	0.82	0.79	0.18
Weekly Hours doing Math Homework	3.27	2.68	3.25	2.61	0.08*	0.10	0.18
Teacher Takes Interest in Studies	0.96	0.95	0.95	0.96	0.99	0.86	0.94
How often do you discuss math with your teacher?	1.70	1.70	1.73	1.73	0.75	0.71	0.88
How often do you discuss math with your parent?	1.81	1.80	1.69	1.96	0.59	0.04**	0.02**
How often do you discuss math with your peers?	0.94	1.02	1.01	0.85	0.37	0.37	0.02**
Intrinsic Motivation Index	0.85	0.84	0.84	0.86	0.37	0.93	0.11
Extrinsic Motivation Index	0.84	0.83	0.80	0.79	0.31	0.19	0.19
Observations:	273	532	503	229			

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 B.2: Balance Table of Student Characteristics (Endline Student Scores Sample)

	Mean				P-values		
	(1) Control	(2) Exp	(3) Peer	(4) Info	(1)-(2)	(1)-(3)	(1)-(4)
Baseline Math Score	82.60	83.13	83.00	85.53	0.75	0.86	0.16
Female	0.42	0.48	0.32	0.39	0.01***	0.01**	0.81
High Parental Income	0.15	0.13	0.07	0.10	0.60	0.08*	0.87
Adults peer Room	0.57	0.55	0.56	0.58	0.28	0.84	0.47
High Parental Literacy	0.10	0.11	0.08	0.06	0.59	0.81	0.94
Number of Friends in the Classroom	4.26	4.09	3.94	3.99	0.69	0.19	0.83
Weekly Hours Studying Math	4.12	3.89	3.94	3.47	0.99	0.91	0.17
Weekly Hours doing Math Homework	3.30	2.78	3.21	2.46	0.29	0.22	0.05*
Teacher Takes Interest in Studies	0.96	0.95	0.94	0.95	0.92	0.58	0.73
How often do you discuss math with your teacher?	1.73	1.69	1.66	1.76	0.89	0.40	0.55
How often do you discuss math with your parent?	1.84	1.84	1.72	1.99	0.59	0.04**	0.02**
How often do you discuss math with your peers?	0.90	1.01	1.02	0.83	0.34	0.35	0.03**
Intrinsic Motivation Index	0.86	0.85	0.85	0.87	0.43	0.59	0.21
Extrinsic Motivation Index	0.84	0.82	0.79	0.78	0.36	0.13	0.16
Observations:	234	434	411	202			

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Endline 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).

C Midline and Endline Results (Separately)

Table C.1: Treatment Effects on Midline Math Test Scores

	(1) Standardized	(2) Raw Scores
Expectations	0.207**	3.751**
	(0.098)	(1.900)
Peer	0.080	1.757
	(0.101)	(1.834)
Information	0.201*	3.798**
	(0.116)	(1.864)
Baseline Score		0.481***
		(0.039)
Observations	1492	1422
Comparisons (p-values)		
Exp vs Peer	0.154	0.159
Exp vs Info	0.955	0.975
Info vs Peer	0.247	0.136

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 C.2: Treatment Effects on Endline Math Test Scores

	(1) Standardized	(2) Raw Scores
Expectations	0.214**	3.280*
	(0.105)	(1.801)
Peer	0.054	0.344
	(0.114)	(1.936)
Information	0.158	2.219
	(0.115)	(2.028)
Observations	1281	1281
Comparisons (p-values)		
Exp vs Peer	0.053	0.036
Exp vs Info	0.530	0.500
Info vs Peer	0.273	0.263

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 C.3: Treatment Effects on Math Test Scores Over Time

	(1) Standardised	(2) Raw
Information	0.482**	3.437
	(0.234)	(2.091)
Information x Endline	-0.356	-0.737
	(0.310)	(2.957)
Individual	0.446*	3.358*
	(0.229)	(1.918)
Expectations x Endline	-0.222	0.225
	(0.309)	(2.712)
Peer	0.298	1.150
	(0.241)	(1.967)
Peer x Endline	-0.248	-0.225
	(0.301)	(2.727)
Endline	0.400	2.178
	(0.275)	(2.379)
Constant	-1.517***	75.416***
	(0.389)	(2.524)
Observations	2773	2773
Info Effect (End-Mid)= Exp Effect (End-Mid)	0.505	0.662
Peer Effect (End-Mid)= Exp Effect (End-Mid)	0.894	0.812
Peer Effect (End-Mid)= Info Effect (End-Mid)	0.566	0.816

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) 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. Endline is a binary variable equal to 1 for the scores collected during the endline round and 0 for the midline round. The t-tests reported below the table, labelled "End-Mid", check if the change in the effect of the treatment arms is differential across arms.

Table C.4: Treatment Effects on Midline English Test Scores

	(1) Standardized	(2) Raw Scores
Expectations	0.044	2.190*
	(0.139)	(1.300)
Peer	-0.162	0.357
	(0.190)	(1.477)
Information	0.123	0.892
	(0.164)	(1.510)
Baseline Score		0.245***
		(0.025)
Observations	1189	1159
Comparisons (p-values)		
Exp vs Peer	0.248	0.126
Exp vs Info	0.614	0.335
Info vs Peer	0.160	0.691

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 C.5: Treatment Effects on Endline English Test Scores

	(1) Standardized	(2) Raw Scores
Expectations	-0.446**	-2.307
	(0.207)	(1.750)
Peer	-0.618***	-2.813
	(0.221)	(1.720)
Information	-0.184	-0.481
	(0.215)	(1.789)
Observations	1224	1224
Comparisons (p-values)		
Exp vs Peer	0.395	0.701
Exp vs Info	0.177	0.175
Info vs Peer	0.030	0.069

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.

C.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.5.5 for the expectations, information, and peer arms, respectively. As shown in Figure A.5.5, 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).

D Heterogeneity by Matched Peer Characteristics

Table D.1: Treatment Effects by Peer Achievement (Pooled)

	(1) Standardized	(2) Raw Scores
Expectations	0.201**	3.295**
	(0.086)	(1.374)
Information	0.161*	2.850**
	(0.088)	(1.436)
Peer score is higher	-0.029	-0.612
	(0.121)	(1.988)
Peer score is lower	0.019	0.285
	(0.097)	(1.668)
Peer score is same	0.219**	3.604**
	(0.108)	(1.678)
Constant	-2.213***	41.819***
	(0.223)	(3.862)
Observations	2355	2355
Comparisons (p-values)		
High vs Same	0.049	0.043
Low vs Same	0.074	0.071
High vs Low	0.652	0.616

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. Regressions control for individuals own score and include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.2: Treatment Effects by Peer Achievement at Midline

	(1) Standardized	(2) Raw Scores
Expectations	0.222**	3.849**
	(0.110)	(1.909)
Information	0.205*	3.997**
	(0.112)	(1.890)
Peer score is higher	-0.027	-0.279
	(0.153)	(2.636)
Peer score is lower	0.037	1.086
	(0.120)	(2.162)
Peer score is same	0.326***	5.886***
	(0.125)	(2.074)
Constant	-2.623***	35.365***
	(0.275)	(4.957)
Observations	1251	1251
Comparisons (p-values)		
High vs Same	0.010	0.006
Low vs Same	0.029	0.025
High vs Low	0.634	0.540

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. Regressions control for individuals own score and include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.3: Treatment Effects by Peer Expectations Gap at Midline

	(1) Standardized	(2) Standardized	(3) Raw Scores	(4) Raw Scores
Absolute difference between the pair's teacher expectations	-0.011**		-0.114	
	(0.005)		(0.092)	
Own Expectation	0.040***	0.036***	0.381***	0.330***
	(0.007)	(0.008)	(0.111)	(0.115)
Peer exp is same		0.397**		4.422
		(0.160)		(2.841)
Peer exp is lower		0.324*		4.459
		(0.168)		(3.189)
Observations	305	305	303	303

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) and (2) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.4: Treatment Effects by Peer Scores Gap at Midline

	(1) Standardized	(2) Standardized	(3) Raw Scores	(4) Raw Scores
Absolute difference between the pair's baseline scores	-0.010*		-0.184*	
	(0.005)		(0.093)	
Own baseline performance	0.027***	0.029***	0.453***	0.486***
	(0.006)	(0.007)	(0.102)	(0.107)
Peer score is same		0.399**		7.080**
		(0.174)		(2.907)
Peer score is lower		0.035		1.018
		(0.188)		(3.069)
Observations	303	303	303	303

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) and (2) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.5: Treatment Effects by Peer Baseline Characteristics at Midline

	(1) Standardized	(2) Raw Scores
Expectations	0.204**	3.797**
	(0.099)	(1.901)
Information	0.204*	3.914**
	(0.118)	(1.891)
Peer	-1.206***	-12.252**
	(0.379)	(4.955)
Peer x Homophily	1.688***	18.489***
	(0.437)	(5.658)
Observations	1309	1251

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. All regressions control for the individual's own characteristic that is being compared to the peer. Regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.6: Treatment Effects by Peer Achievement at Endline

	(1) Standardized	(2) Raw Scores
Expectations	0.176	2.654
	(0.108)	(1.774)
Information	0.118	1.646
	(0.118)	(2.037)
Peer score is higher	-0.028	-0.905
	(0.151)	(2.545)
Peer score is lower	0.004	-0.476
	(0.129)	(2.191)
Peer score is same	0.108	1.202
	(0.143)	(2.328)
Constant	-1.602***	51.077***
	(0.251)	(4.444)
Observations	1104	1104
Comparisons (p-values)		
High vs Same	0.411	0.465
Low vs Same	0.476	0.506
High vs Low	0.772	0.826

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. Regressions control for individuals own score and include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.7: Treatment Effects by Peer Expectations Gap at Endline

	(1) Standardized	(2) Standardized	(3) Raw Scores	(4) Raw Scores
Absolute difference between the pair's teacher expectations	-0.002		-0.001	
	(0.005)		(0.086)	
Own Expectation	0.031***	0.025***	0.457***	0.389***
	(0.006)	(0.006)	(0.120)	(0.111)
Peer exp is same		0.343**		4.569*
		(0.134)		(2.374)
Peer exp is lower		0.244*		3.297
		(0.127)		(2.543)
Observations	286	286	286	286

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) and (2) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.8: Treatment Effects by Peer Scores Gap at Endline

	(1) Standardized	(2) Standardized	(3) Raw Scores	(4) Raw Scores
Absolute difference between the pair's baseline scores	0.000		-0.013	
	(0.004)		(0.072)	
Own baseline performance	0.021***	0.019***	0.374***	0.356***
	(0.006)	(0.006)	(0.103)	(0.107)
Peer score is same		0.225		3.458
		(0.191)		(3.415)
Peer score is lower		0.038		0.532
		(0.137)		(2.511)
Observations	286	286	286	286

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) and (2) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

Table D.9: Treatment Effects by Peer Baseline Characteristics at Endline

	(1) Standardized	(2) Raw Scores
Expectations	0.211**	2.624
	(0.105)	(1.774)
Information	0.156	1.603
	(0.113)	(2.038)
Peer	-0.525*	-5.125
	(0.315)	(4.525)
Peer x Homophily	0.730*	6.476
	(0.378)	(5.129)
Observations	1158	1104

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. All regressions control for the individual's own characteristic that is being compared to the peer. Regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

E Mechanisms

Table E.1: Heterogeneous 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.

Table E.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.

F Long Run Results

Table F.1: Treatment Effects on Long Run Math Test Scores by the Type of Expectation Delivered

	(1) Standardized	(2) Raw Scores
Expectations (Very High)	0.164	0.919
	(0.124)	(1.848)
Expectations (High)	0.151	2.419
	(0.133)	(2.095)
Peer (Very High Expectation)	0.011	-0.277
	(0.123)	(1.764)
Peer (High Expectation)	0.199	2.681
	(0.122)	(1.655)
Information	0.114	-0.214
	(0.136)	(1.824)
Baseline Score		0.351***
		(0.038)
Observations	1648	1601
Comparisons (p-values)		
Exp Very High vs Info	0.672	0.497
Exp Very High vs High	0.901	0.330

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.

G Robustness

Table G.1: Treatment Effects using PDS Lasso

	(1) Standardized	(2) Raw Scores
Information Arm	0.149*	2.604*
	(0.084)	(1.444)
Expectations Arm	0.244***	4.043***
	(0.082)	(1.319)
Peer Arm	0.071	1.118
	(0.077)	(1.292)
Constant	-2.671***	36.177***
	(0.310)	(4.249)
Observations	2687	2687
Comparisons (p-values)		
Exp vs Peer	0.003	0.002
Info vs Peer	0.224	0.198
Exp vs Info	0.164	0.211

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Standard errors clustered on the classroom level. The estimations pool midline and endline scores. Regressions include strata and round fixed effects. Missing values of any baseline characteristics are imputed to be equal to the mean value of the characteristic in the class.

Table G.2: Heterogenous Treatment Effects by Homophily using PDS Lasso

	(1) Standardized	(2) Raw Scores
Information	0.168**	2.909**
	(0.083)	(1.432)
Expectations	0.259***	4.297***
	(0.081)	(1.311)
Peer	-0.346*	-7.382*
	(0.198)	(3.849)
Peer x Homophily	0.562**	11.555***
	(0.225)	(4.395)
Constant	-2.585***	37.770***
	(0.346)	(4.711)
Observations	2329	2329

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Standard errors clustered on the classroom level. The estimations pool midline and endline scores. Regressions include strata and round fixed effects. Missing values of any baseline characteristics are imputed to be equal to the mean value of the characteristic in the class.

Table G.3: Heterogenous Treatment Effects by Homophily using PDS Lasso (Midline)

	(1) Standardized	(2) Raw Scores
Information	0.158	3.134*
	(0.110)	(1.898)
Expectations	0.248**	4.314**
	(0.103)	(1.805)
Peer	-0.499**	-9.261**
	(0.244)	(4.508)
Peer x Homophily	0.783***	14.832***
	(0.279)	(5.191)
Constant	-3.252***	29.638***
	(0.417)	(5.965)
Observations	1274	1274

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Standard errors clustered on the classroom level. The estimations only include midline results. Regressions include strata and round fixed effects. Missing values of any baseline characteristics are imputed to be equal to the mean value of the characteristic in the class.

Table G.4: Heterogenous Treatment Effects by Homophily using PDS Lasso (Endline)

	(1) Standardized	(2) Raw Scores
Information	0.144	2.003
	(0.117)	(2.057)
Expectations	0.238**	3.682**
	(0.111)	(1.821)
Peer	-0.027	-2.359
	(0.228)	(4.284)
Peer x Homophily	0.094	3.482
	(0.257)	(4.842)
Constant	-2.047***	44.747***
	(0.277)	(4.229)
Observations	1120	1120

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Standard errors clustered on the classroom level. The estimations only include endline results. Regressions include strata and round fixed effects. Missing values of any baseline characteristics are imputed to be equal to the mean value of the characteristic in the class.

H Power Calculations and Deviations from Pre-Analysis Plan

The analysis follows the pre-registered plan in its key elements, including the randomization design, primary outcomes, and main heterogeneity analyses. Below we summarize power calculations, as specified in our pre-analysis plan, and discuss key deviations.

Power Calculations.

Power calculations were conducted using standardized math scores from historical administrative school data. We estimated an intra-cluster correlation of 0.14 and the correlation between math scores over time (using the most recent term's mid-term and end-term scores) of 0.7987. With significance level $\alpha = 0.05$, power of 0.80, and a conservative estimate of average class size of 5 students, the experiment was powered to detect minimum effects of 0.13 standard deviations between each treatment arm and the control group, and between any two treatment arms. Our main treatment effects of 0.18–0.21 standard deviations exceed the pre-specified minimum detectable effect, confirming that the study was adequately powered to detect effects of the magnitude observed.

Key Deviations.

The key deviations from our pre-analysis plan are as follows.

First, in the pre-analysis plan, classrooms randomized to receive a reminder about their last test score were classified as part of the comparison group, as this information was not anticipated to be new. In the paper, we relabel this group as the "Information Arm" and treat it as a distinct treatment arm. This reclassification reflects the observed treatment effects: students who received performance reminders showed significant improvements in math achievement (0.18σ), comparable to those who received teacher-set expectations. Treating the Information Arm as a separate treatment condition allows for cleaner comparison of performance information alone versus performance information bundled with expectations.

Second, the pre-analysis plan specified survey-based outcomes including student motivation, effort, and non-cognitive skills. It also included research-team-designed math assessments for cross-validation of school test scores and a parental survey measuring engagement, investment, and beliefs. Survey response rates for students and parents were significantly lower than anticipated, so we focus our main analysis on high-stakes school-administered test scores, which provide sufficient statistical power and allow us to answer our main question of interest. We do not detect effects on most survey outcomes.