Better Predictions, Worse Outcomes: An Experimental Evaluation of Giving Teachers Machine-Learning Forecasts of Student Achievement*

Daniel Sabey Jake Meyer

August 31, 2026

Preliminary draft

Abstract

Teachers allocate attention and effort across students based on their own noisy assessments of who needs help. We test whether giving teachers algorithmically generated predictions of their students' end-of-year test scores improves student achievement by better focusing efforts. In a randomized experiment across the six campuses of a U.S. charter school network, each term treated teachers received reports containing machine-learning forecasts of each student's end-of-year score on the assessments the schools already administer. The information treatment landed, and it landed where it was aimed: treated teachers' own predictions of student performance improved in accuracy, they flagged struggling students more intensively on internal monitoring charts, and when asked which students they were focusing most on helping they named the students the model had flagged eleven percentage points more often than control teachers did. Treated teachers reduced their own data consultation, cut time spent planning interventions, and shifted instruction away from review. And we then observe that student achievement *fell*, and it fell most for the students who were flagged as being below proficent by the model.

*DRAFT — preliminary and incomplete. Please do not cite or circulate. We thank our partner charter school network for their collaboration.

1 Introduction

Improving student outcomes is a perennial challenge for educators. One difficulty felt by some educators is that by the time the symptoms of academic struggle are apparent, the root causes are already ingrained. One way to address such a difficulty could be to provide information about students very early in the academic year. In particular, machine learning may offer potentially more accurate information about student trajectories, which could be used by teachers effectively to direct their efforts in the classroom. At the same time, this information approach could have counterproductive effects if it somehow crowds out teacher effort or deflates optimistic attitudes.

The potential for teachers to use predictive aids represents one of many contexts where human decisionmakers may or may not benefit from using predictive models. While administrative data and modern computational abilities combine to offer better information than ever, implementing these tools has not been straightforward. Even in industries where these tools are integrated, there remain many concerns and limitations [Kleinberg et al., 2018, Mullainathan and Obermeyer, 2022, Stevenson and Doleac, 2022, Albright, 2024, Angelova et al., 2024, De-Arteaga et al., 2020]. The effects of such tools in elementary education are even less known. Teachers face a unique decision environment with complex and dynamic information and choices. In this setting, it is unclear what the contribution of a prediction model would be, or how teachers would respond to this information.

In partnership with a small charter school network in the United States, we study the effects of providing teachers with machine learning predictions about students' end-of-year performance in a randomized control trial. The experiment was conducted in partnership with six elementary schools from a US charter school network during the 2025-2026 academic year. At the start of each term, teachers in treated school-grades received a report with predictions of each of their student's end-of-year scores on standardized tests. All teachers were surveyed each term to measure beliefs about student performance as well as efforts in and out of the classroom to help students. This design allows us to document the relative accuracy of teacher and machine learning predictions throughout the year, as well as to measure the effect of the information on teachers beliefs and efforts, and in turn on student outcomes.[1]

We find that machine learning predictions improve the accuracy of teacher beliefs, but that students in classrooms where teachers received the predictions score 0.099 standard deviations lower on standardized exams. The predictions are more accurate than teachers,

[1]The experiment was pre-registered with the Center for Open Science: https://osf.io/75f3a/ and approved by Cornell's IRB: .

especially at the beginning of the year. On the same students, the model's mean absolute error is 13.9 percentile points (0.51 standard deviations) against 21.8 percentile points (0.71 standard deviations) for teachers. Treated teachers' own forecasts move toward the model: over the rest of the year their prediction errors fall by 2.1 percentile points (0.05 standard deviations), with the gain concentrated among lower-performing students on the standardized-error metric. What the reports changed is the size of the error rather than its direction. Control teachers are mildly optimistic on average (+0.07 standard deviations), and treated teachers slightly less so, especially for the level of the forecasts for the weakest students.

Teacher efforts also shift after receiving the reports. By the end of the year, treated teachers had recorded 2.72 more flags per student on the network's low-performance monitoring charts, 44 percent above the control mean and concentrated among students the model predicted would fall below proficiency. Asked each term to name the five students they were focusing most on helping, treated teachers named model-flagged students 12 percentage points more often than control teachers. At the same time, treated teachers consulted the network's data systems less, with queries to academic and assessment records falling by 25 and 23 log points, spent 33 log points less time planning student interventions, and shifted whole-class instruction away from reviewing previously taught material (a decrease of 8.4 percentage points of class time) and toward introducing new material (an increase of 9.1 percentage points).

Ultimately, students whose teachers received the predictions have worse scores on standardized tests. On an index pooling all five exams, treatment reduces scores by 0.099 standard deviations ($p = 0.03$ with cluster-robust standard errors, $p = 0.09$ to 0.11 under randomization inference and a wild cluster bootstrap). The coefficient is negative for each of the five exams, for every predicted-proficiency group, and for all sixteen demographic subgroups we examine. The largest loss, 0.226 standard deviations, falls on the students the model flagged as likely to be below proficient: the same students teachers monitored more and said they were focusing on.

These findings contribute to a body of other studies investigating the beliefs of teachers and their effect on student outcomes. The starting point for this literature is that a teacher's belief about a student is not just a forecast but an input into what the student goes on to do. Rosenthal and Jacobson [1968] told teachers that a randomly selected group of students were about to have a growth spurt, and those students improved. More recent work with better identification points the same direction. Papageorge et al. [2020] find that teachers are on average optimistic about their students, and that this optimism pays off: higher expectations lead to better outcomes. Hill and Jones [2018] find similar effects on test scores in grades

four through eight, the setting closest to ours, and Terrier [2020] and Gnas et al. [2024] find that students whose teachers overestimate them go on to do better on test scores and non-cognitive measures, respectively. A second set of papers documents that teacher beliefs are wrong in systematic ways. Teachers overestimate low achievers in India and Bangladesh [Djaker et al., 2024], under-assess Black pupils in England [Burgess and Greaves, 2013], hold lower expectations for English learners with the same test scores [Zhu, 2024], and pass over low-income students who universal screening later identifies as gifted [Card and Giuliano, 2016]. Gershenson et al. [2016] and Carlana [2019] show that these gaps depend on the match between teacher and student and that they matter for outcomes. Meta-analyses put the correlation between teacher judgments and measured achievement around 0.6 [Südkamp et al., 2012, Kaufmann, 2020], so there is room for a prediction model to add information. The interventions closest to ours correct beliefs directly. Alesina et al. [2024] show teachers their own grading bias against immigrant students and find they correct it, and Bergman et al. [2018] finds that parents who receive real-time information about their child's performance revise their optimistic beliefs and their children do better. Teacher beliefs are inaccurate, so better information should help; but optimism has returns, so correcting it could hurt. Finally, Escueta et al. [2020] provides a recent summary of educational interventions that use technology to try and improve student outcomes. We investigate a novel potential solution by providing teachers with the best possible information about their students.

This paper also contributes to the growing literature on prediction tools and their use by human decisionmakers. Machine learning models can beat experts at prediction in consequential settings, including pretrial release [Kleinberg et al., 2018], medical testing [Mullainathan and Obermeyer, 2022], opioid prescribing [Hastings et al., 2020], and gun violence [Heller et al., 2024]. In education, administrative data can predict dropout [Adelman et al., 2018] and college success [Bird et al., 2025]. Whether these gains survive handing the prediction to a person is a separate question. Stevenson and Doleac [2022] find that a validated risk score placed in judges' hands produced little lasting change, and Albright [2024] and Angelova et al. [2024] find that when judges do respond, it is as much about who bears the blame for a visible mistake as about the information in the score. Imai et al. [2023] find limited effects of a pretrial risk assessment on judges' decisions or case outcomes. In child welfare, De-Arteaga et al. [2020] and Cheng et al. [2022] show that screeners partly compensate for errors in the algorithm rather than deferring to it. Agarwal et al. [2023] find that radiologists given machine learning predictions do worse than the machine learning predicitons alone, because they do not weight the model's signal efficiently. In other settings, Sun et al. [2022] find that warehouse managers routinely override algorithmic prescriptions, and Barwahwala et al. [2024] find that an accurate model for catching bogus firms was of

limited use once handed to tax inspectors. The closest paper to ours in education is Bergman et al. [2021], who use an algorithm to place college students into remedial courses and find that it improves on the status quo placement rules. A common thread in this work is that the decisionmaker, not the model, determines what the prediction is worth. We contribute by studying a novel setting, finding that while predictive models can improve on human assessment, the provision of updated information can be a hindrance rather than a help.

2 Setting and Experimental Design

The experiment took place during the 2025–26 academic year at six campuses of a U.S. charter school network. Teachers in grades 1–6 regularly have to decide which students were falling behind and where additional attention might help. They had substantial information on each student, but that information was spread across test records, gradebooks, curriculum placements, attendance files, and benchmark assessments. The intervention combined those records into predicted end-of-year scores that teachers could review during the school year.

2.1 The school setting and the teacher's task

The six campuses use a common curriculum, a shared student information system, and the same assessment calendar. Teachers place students into leveled reading and mathematics groups and update those placements as students progress. We do not identify the network or its state under our agreement with the partner.

Table 1 describes the students in the experimental grades. Across the network, 71% are White and 35% are Hispanic/Latino. These federal categories are recorded separately, so the shares need not sum to one. Thirty percent receive free or reduced-price lunch, 23% are classified as English learners, and 14% receive special-education services. The campuses serve different populations: free-lunch shares range from 12% to 56%, while English-learner shares range from 6% to 44%.

Table 1: Student Characteristics by Campus

Campus	1	2	3	4	5	6	All
Students	142	502	920	373	548	583	3,068
Share in treated cells	0.29	0.68	0.52	0.50	0.67	0.35	0.53
Below proficient — predicted at baseline							
Share of students	0.62	0.23	0.35	0.43	0.65	0.59	0.45
Share of assessments	0.41	0.11	0.18	0.23	0.43	0.32	0.25
Below proficient — realized at end of year							
Share of students	0.58	0.28	0.32	0.47	0.70	0.63	0.47
Share of assessments	0.39	0.14	0.18	0.31	0.50	0.41	0.30
White	0.79	0.64	0.61	0.97	0.73	0.74	0.71
Hispanic/Latino	0.00	0.14	0.27	0.14	0.66	0.60	0.35
Black	0.01	0.04	0.05	0.02	0.16	0.13	0.08
Asian	0.05	0.34	0.35	0.02	0.09	0.12	0.21
Other race	0.15	0.04	0.02	0.03	0.05	0.05	0.04
Female	0.52	0.54	0.55	0.49	0.58	0.55	0.55
Free/reduced lunch	0.25	0.12	0.16	0.20	0.51	0.56	0.30
English learner	0.09	0.19	0.11	0.06	0.44	0.38	0.23
Special education	0.15	0.12	0.12	0.18	0.11	0.16	0.14

Proficiency is measured against each assessment's official cutoff, over the three different assessments. The two panels differ in what is counted, not in the cutoff. "Share of students" is the share below proficient on *at least one* of their assessments; "share of assessments" pools assessments and asks what fraction fall below. The student measure is roughly double the assessment measure because one miss is enough to count. The predicted panel uses the baseline forecast and is a prediction, not an outcome; the realized panel uses end-of-year scores.

Each student takes three end-of-year assessments: two in reading and one in mathematics. All grades take a curriculum-based reading benchmark. Students in grades 3–6 also take the state accountability tests in English language arts and mathematics; students in grades 1–2 instead take nationally normed reading and mathematics tests.

	Grades 1–2	Grades 3–6
Reading 1	Benchmark reading	Benchmark reading
Reading 2	National reading	State ELA
Mathematics	National math	State math

Each assessment has official proficiency cutoffs. Across the network, students score below proficiency on 30% of their end-of-year assessments, on average; the campus rates range from

14% to 50%. The reports used these same cutoffs to identify predicted proficiency bands and students near a boundary. Because score scales differ across assessments and grades, our main achievement outcome standardizes and pools the three scores each student receives. Section 4 defines that index, and Appendix C reports the results separately by assessment.

Teachers also maintain weekly "Learning Progress Charts" (LPCs). On these internal spreadsheets, they flag students who are identified as being below proficent in mathematics, reading, spelling, or behavior. Network administrators had been encouraging more teachers to complete the charts and to keep them current. Updating an LPC therefore requires a teacher to check and record a student's progress. This makes the charts a useful high-frequency measure of which students teachers are monitoring most closely.

2.2 From existing records to student forecasts

Teachers in both treatment and control cells could see the records used to build the forecasts. These included prior-year assessment scores, current curriculum placement, gradebook entries, attendance, program classifications, and current-year benchmark scores as they became available. Teachers accessed those records through the student information system and master academic spreadsheets. Treatment did not change access to any underlying record.

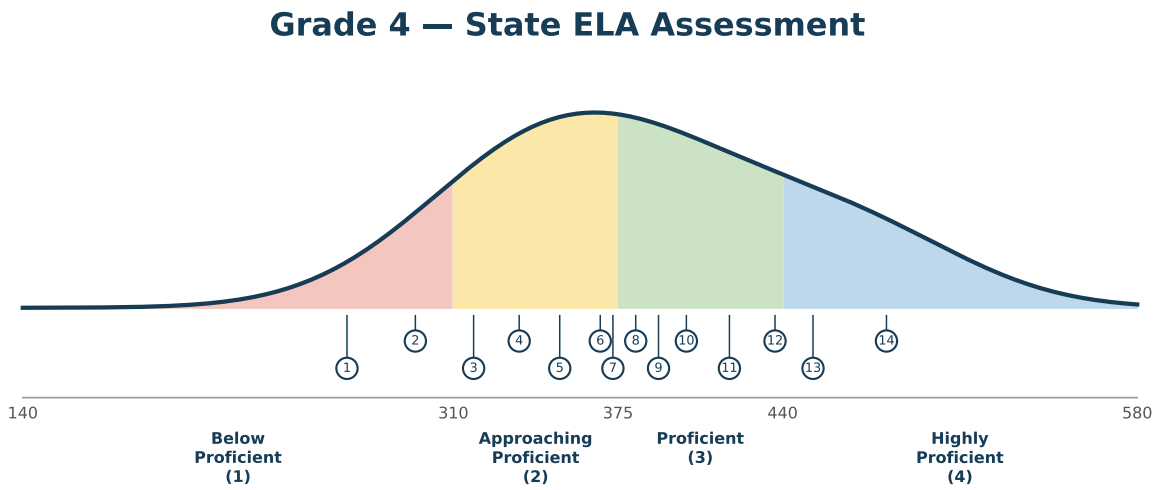
What differed was the summary. Treated teachers received a report with three pieces of information for every student in their homeroom:

1. a predicted end-of-year score on each assessment taken by the student's grade and where it lands on the district's distribution;

2. the proficiency band containing that predicted score; and

3. a near-cutoff flag for predictions within one-quarter of a standard deviation of a proficiency-band boundary, where the standard deviation is calculated from the grade-wide predicted-score distribution for that assessment.

Figure 1 shows one page of the report. Each page covers one assessment. The top panel plots the grade-wide distribution of predicted scores, shaded by proficiency band, and marks the teacher's students on that distribution. The lower tables list predicted scores and bands and identify students near a cutoff. A teacher could therefore see both the class as a whole and the students whose predicted scores were close to an official boundary.



4th Graders

Name	Pred. Score	Band
Avery Bright	268	Below Prof. (1)
Jordan Castillo	295	Below Prof. (1)
Riley Dunn	318	Approaching (2)
Sam Escobar	336	Approaching (2)
Casey Fontaine	352	Approaching (2)
Quinn Garber	368	Approaching (2)
Alex Hooper	373	Approaching (2)
Jamie Iqbal	382	Proficient (3)
Morgan Jeffers	391	Proficient (3)
Taylor Kim	402	Proficient (3)
Rowan Lester	419	Proficient (3)
Devin Moss	437	Proficient (3)
Skyler Novak	452	Highly Prof. (4)
Emerson Ortiz	481	Highly Prof. (4)

Students Near Cutoff

Name	Proximity	Band
Riley Dunn	+8	Approaching (2)
Quinn Garber	-7	Approaching (2)
Alex Hooper	-2	Approaching (2)
Jamie Iqbal	+7	Proficient (3)
Devin Moss	-3	Proficient (3)
Skyler Novak	+12	Highly Prof. (4)

Figure 1: Example page of the treatment report. *Notes:* The figure uses fictional students, scores, and a fictional score distribution; no student data appear.

The reports did not reveal a new administrative record. They supplied a new statistical summary of records teachers could already inspect. Control teachers received no reports and were not shown the model's predictions. Both groups completed the same surveys and maintained the same LPCs.

2.3 Forecast construction

We trained the forecasting models on earlier cohorts in the same network, using realized end-of-year scores as outcomes. The predictors were the records available to teachers: prior scores and proficiency bands, curriculum placement, gradebook performance and completion, attendance, program classifications, and grade level. We rebuilt the features for each wave and removed every field generated after the prediction date. Thus each forecast used only information available at the time it was produced.

We estimated separate models for the five assessments and refreshed the forecasts as new records accumulated. The final prediction combined penalized linear models with tree-based learners in a stacked ensemble. We maintained different versions for returning students with complete gradebooks, returning students with incomplete gradebooks, and students new to the network. Kindergarten students had too little prior information to support a comparable forecast, so kindergarten classrooms were excluded before randomization. Appendix A describes the features, model fitting, and validation in detail.

2.4 Randomization and implementation

We randomized treatment at the campus-by-grade level. The six campuses and six eligible grades produced 36 assignment cells; 18 received reports and 18 served as controls. Within each grade, exactly three campuses were assigned to treatment. Assignment remained fixed for the school year. The experimental roster contains 3,068 students. There were 1,611 treated students and 1,457 control students.

The network preferred assignment within campuses because every principal wanted some teachers to receive reports. Grade teams were also the main unit of joint planning, making the campus-by-grade cell a natural level for assigning access. And such a design does seem to minimize the possibilty of spillovers, since there is much less cross-grade communication than within grade. Our regressions include campus and grade fixed effects and cluster standard errors at the campus-by-grade level. With only 36 clusters, we also report robust inference estimates in Section 5.4.

Before the school year began we visited with teachers from each of the six campuses and presented the study at a faculty meeting. Teachers were told that reports containing predicted end-of-year scores would be sent to about half of the grade teams in the network, and were walked through an example report: what the predicted score meant, how to read the distribution, how the proficiency bands were drawn, and how to read the list of students near a band boundary. Every teacher therefore saw the instrument and understood the study, whether or not their own grade team would receive reports.

This has two consequences worth stating. It rules out the possibility that treated teachers simply failed to understand what they were sent, and it means control teachers were not blind: they knew reports existed and knew what information those reports contained, even though they never received one for their own students. Any resulting behavioral response by control teachers would likely narrow the contrast between arms rather than widen it, so it works against finding an effect of either sign.

Reports were produced five times: once near the start of the year and four times as the year progressed. In the four verified post-treatment cycles, the models first incorporated the records available at that date. Treated teachers then received updated reports, and roughly one week later all teachers completed the prediction survey. The lag gave treated teachers time to read the report while keeping the survey close to the information release. Control teachers completed the same survey on the same schedule.

The start-of-year cycle, which we call the baseline wave, straddled the first report delivery. We sent the survey invitation in the first week of the school year and sent the first reports to treated teachers five days later.

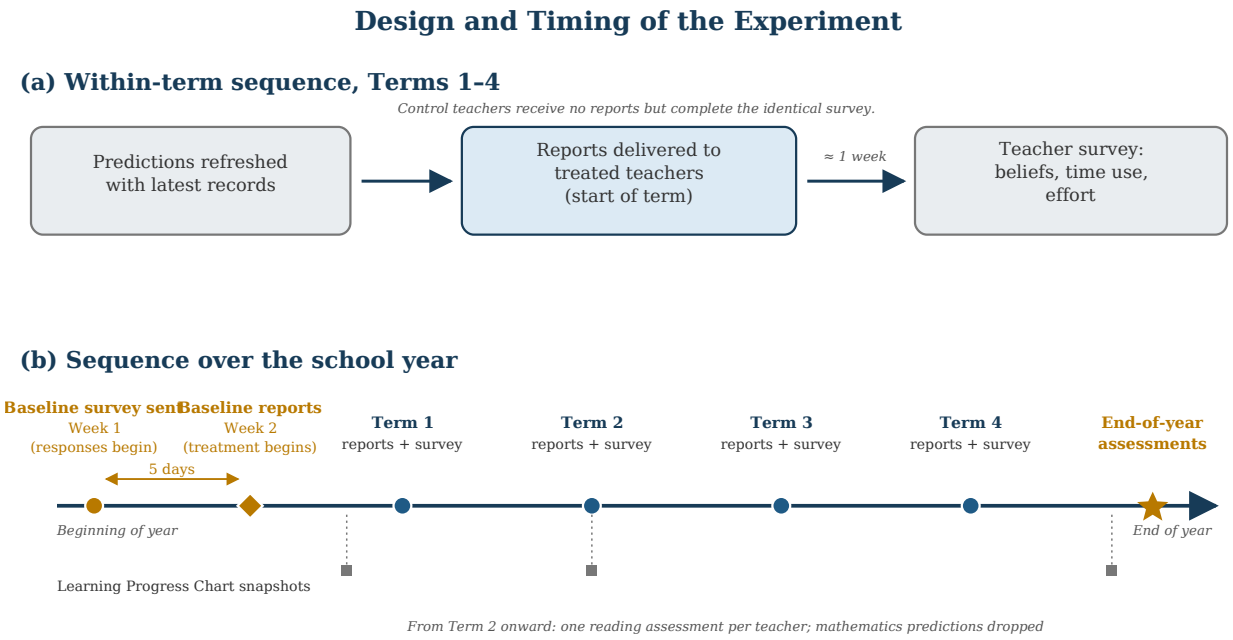


Figure 2: Design and timing of the experiment.

The reports continued to cover every relevant assessment, but the survey used to measure teacher beliefs became shorter. In terms 0 and 1, teachers forecast up to three scores for every student. From term 2 onward, the survey asked each teacher about one reading assessment and no longer asked for mathematics forecasts. Section 3 explains how the analysis handles

the resulting unbalanced panel.

3 Data

We assemble five data sources, linked by a network-wide student identifier.

Actual test scores. End-of-year outcomes come from the network's master academic records: end-of-year reading benchmark scores for 2,979 students (97% of the analysis sample), state ELA and math scores for 1,722 students in grades 3–6 (84%), and national test reading and math scores for 791 students in grades 1–2 (77%). Section 5.4 shows that presence in each score file is balanced across arms.

Machine-predicted scores. We use the baseline machine forecasts as the baseline ability measure throughout. These algorithmic forecasts were fixed before reports were delivered and are distinct from the teacher predictions collected in the baseline survey. They are therefore usable as pre-treatment moderators, whereas later machine forecasts incorporate current-year inputs (gradebooks and benchmark scores) that treatment itself may have affected.

Learning Progress Charts. The schools maintain a Learning Progress Chart (LPC) that flags students who are struggling in different areas. We digitize each campus's LPC at three points in the year, October 2025 (T_1), January 2026 (T_2), and May 2026 (T_3), and count, for each student, cumulative flags in math, reading, spelling, and behavior. Students absent from the chart are coded as zero flags.

Teacher surveys. Teachers were surveyed at baseline and after each term. There are 231 teacher-term observations across 85 homerooms in the experimental grades. The surveys measure time allocation (share of class time in whole-class instruction, independent work, and peer work; share of whole-class time reviewing old material versus introducing new material), one-on-one time allocation across student ability levels, counts of consultations of the network's data systems, parent and staff contacts, and minutes spent planning student interventions.

Response was voluntary and teachers were told response would not impact their evaluations or be shared with their administrators, as requested by the IRB. Of the 517 teacher-wave invitations issued to homeroom teachers in the experimental grades, 38% were answered, with per-wave rates of 36%, 31%, 28%, 40% and 57%. Participation is broader than any single wave suggests: 70% of the 107 teachers answered at least one of the five waves. Treated

teachers responded more often than control teachers (75% versus 65% ever responding). Because the belief results are estimated on responders, these should be read as effects on the teachers who responded rather than on the full roster.

Teacher predictions of student scores. In each survey wave, teachers were asked to predict each of their students' end-of-year scores on the relevant assessments. The survey question mirrors the report format deliberately: teachers see the same roster of their own students and are asked for a point score on the same scale the reports use, so treated teachers' answers and the machine forecasts are directly comparable, and control teachers answer the same question without ever having seen a forecast. Figure 3 shows an anonymized reconstruction of the question as teachers saw it. Teachers moved a marker along the assessment's score scale; the displayed numerical score updated as the marker moved and changed color whenever it crossed an official proficiency cutoff. The dashed vertical line marked the grade-wide average. The reconstruction uses fictional student names, scores, and an assessment title to protect participants' identities. In total, 6,189 predictions for students in the experimental grades can be matched to a realized score and a treatment cell.

Coverage at the student level is wider than the teacher response rates imply, because a teacher who answers predicts for a whole homeroom. Of the 3,068 students in the analysis sample, 65% were predicted by their teacher in at least one of the four terms and 69% including the baseline wave, with an average of 1.5 waves per student.

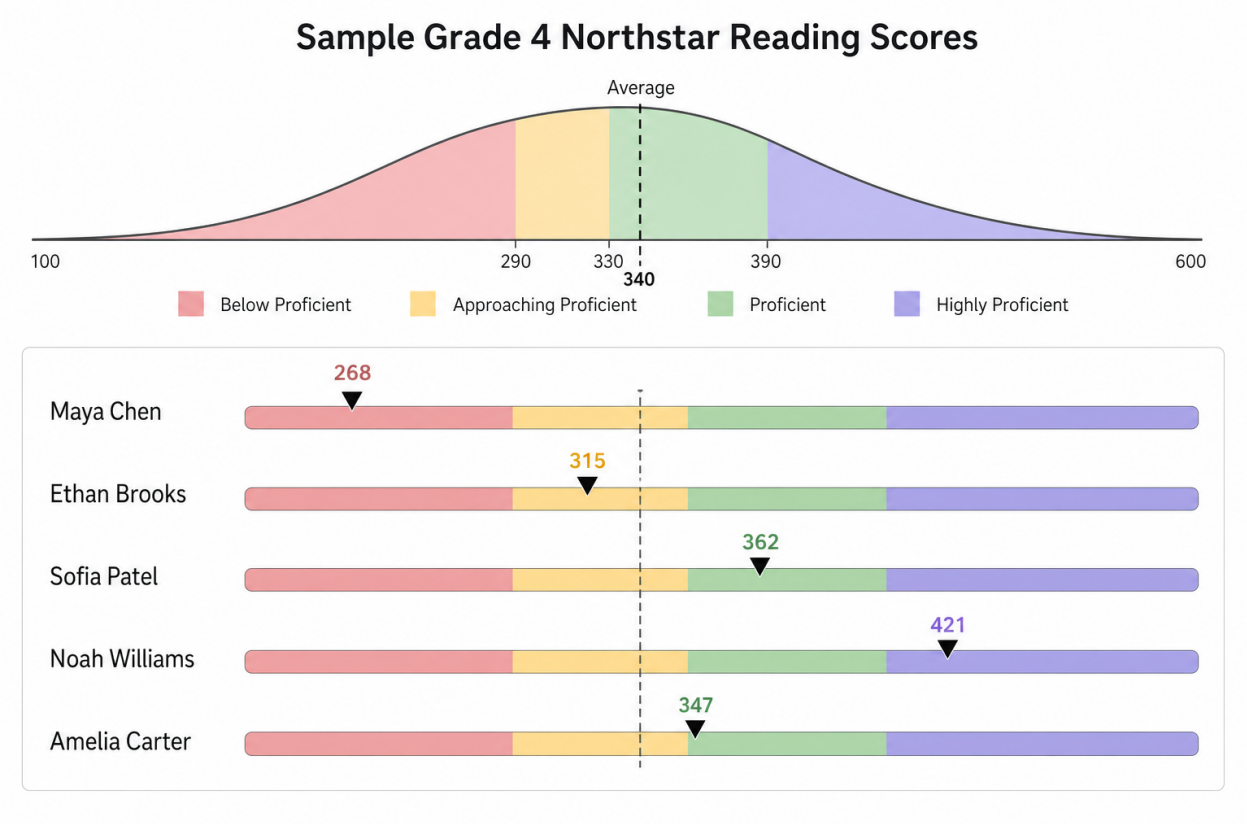


Figure 3: Anonymized reconstruction of the survey interface eliciting teacher predictions of student scores.

The elicitation is an unbalanced panel, and the structure matters for how we estimate. In the first two terms,teachers were asked about the full battery. After discussions with the schools from the third term onward the survey was shortened. We asked teachers to predict scores for a single reading assessment.

Table 2 reports baseline balance. Differences between arms are individually insignificant for the baseline predicted score index, prior-year test scores, and demographic indicators, and a joint F-test does not reject balance ($p = .88$). Sample sizes differ across rows because coverage does: prior-year state scores exist only for the upper grades, since grade-3 students had not yet taken a state test.

Table 2: Baseline Balance

	Control	Treatment	Diff.	SE	p	N
Baseline predicted score index (z)	-0.054	0.045	+0.098	0.130	0.448	3,068
Prior-year benchmark reading score	364.137	359.141	-4.996	61.432	0.935	2,594
Prior-year state ELA score	389.816	380.916	-8.900	22.206	0.689	1,261
Prior-year state math score	363.708	356.085	-7.623	18.018	0.672	1,252
Free lunch	0.246	0.225	-0.020	0.054	0.706	3,068
English learner	0.156	0.158	+0.001	0.051	0.981	3,068
Joint F-test p-value						0.875

Differences are raw, treated minus control. Standard errors are clustered at campus-grade. The joint test is an F-test of the characteristics observed for every student – the baseline index, free lunch and English learner – in a regression of treatment on those characteristics, so that it spans all 36 randomization cells; the prior-year state assessments are given only in the upper grades. Because randomization was stratified by grade, we also ran the joint test with campus and grade fixed effects, which likewise does not reject ($p = 0.67$).

4 Empirical Framework

Treatment is assigned once, to a campus-grade cell, but the outcomes it acts on are recorded at four different levels of aggregation. A report is delivered to a teacher; the teacher's beliefs are elicited about individual students, several times a year; the teacher's use of time is reported once a term; and the achievement we ultimately care about is measured once, per student, at the end of the year. Each of these needs its own estimating equation, and the unit of observation differs in each. Table 3 fixes what an observation is in each case before the equations follow.

Table 3: Unit of observation by outcome family

Outcome family	Unit of observation	Equation	N
Student achievement, proficiency	student	(1)	3,009
Monitoring-chart flags	student	(1)	3,068
Teacher effort and time use	teacher × term	(2)	≈165
Teacher prediction accuracy	teacher × student × assessment × term	(3)	5,982
Focus-student selection	teacher × term × roster slot	(4)	4,450

Every specification clusters standard errors at the campus-grade cell, the level at which treatment was assigned. There are 36 such cells: six campuses by six grades. Kindergarten is excluded throughout because it never received reports.

4.1 Student achievement

The primary specification is a covariate-adjusted ITT regression. For student i in campus c and grade g:

$$Y_{icg} = \alpha + \beta T_{cg} + \gamma Y^0_{icg} + \pi M_{icg} + \delta_c + \eta_g + \varepsilon_{icg}, \tag{1}$$

where T_{cg} is the campus-grade treatment indicator, Y^0_{icg} is the baseline predicted score for the same outcome, M_{icg} indicates a student with no baseline forecast, and δ_c and η_g are campus and grade fixed effects.

Test scores are standardized within assessment $\times$ grade on control-group moments, so β is in control-group standard deviations. The primary achievement outcome is the *normalized score*: the simple average of a student's standardized scores across the assessments they take. Assessment-by-assessment estimates appear alongside the pooled row in Table 9.

The same equation estimates the proficiency outcomes, as a linear probability model, with Y^0 replaced by the share of assessments the baseline forecast placed the student above the proficiency bar. It also estimates the monitoring-chart flags, with no baseline control: those charts are cumulative and the first snapshot already follows the first report delivery, so conditioning on it absorbs part of the effect being measured.

4.2 Teacher effort and time use

Effort outcomes are reported by the teacher once per term, so the unit is the teacher-term. For teacher j in campus c and grade g, in term t:

$$E_{jt} = \alpha + \beta T_{cg} + \gamma E^0_j + \pi M_j + \delta_c + \eta_g + \tau_t + \varepsilon_{jt}, \tag{2}$$

adding term fixed effects τ_t and controlling for the teacher's own baseline value of the same outcome with a missing indicator [following McKenzie, 2012]. Shares enter in percentage points and are compositional within a block; counts enter as $\log(1+x)$, so those coefficients read as approximate proportional changes. Clustering remains at campus-grade, which is coarser than the teacher: teachers are nested within cells, and it is the cell that was randomized.

4.3 Teacher prediction accuracy

Each term we ask teachers to predict end-of-year scores for the students in their class, so a single teacher contributes one observation per student per assessment per wave:

$$A_{ijmt} \;=\; \alpha \;+\; \beta\, T_{cg} \;+\; \phi_{mt} \;+\; \delta_c \;+\; \eta_g \;+\; \varepsilon_{ijmt}, \tag{3}$$

where A_{ijmt} is teacher j's absolute error about student i on assessment m in term t, and ϕ_{mt} are *interacted* assessment-by-term effects. The interaction is not cosmetic. From term 2 the survey stopped asking about mathematics and assigned each teacher a single reading assessment by grade band, so the mix of questions changes across waves; additive assessment and term effects would compare teachers answering different questions.

Because the state rescaled its ELA metric in 2025–26, raw point errors are not comparable across all assessments, so we use two scale-free measures, each computed within assessment $\times$ grade pooling campuses. The *percentile error* is the absolute difference between the percentile rank of the teacher's prediction and the percentile rank of the student's realized score. The *standardized error* is the absolute difference after converting both to z-scores using each distribution's own moments. The percentile metric weights ranking mistakes evenly across the distribution; the z-score metric preserves distances and so weights tail errors more heavily. Both measure relative-position accuracy rather than calibration of levels. We retain only surveys marked complete, and the headline sample is Terms 1–4: baseline-wave completion straddled the first report delivery, so that wave is neither cleanly pre- nor cleanly post-treatment.

4.4 Focus-student selection

From term 1 the survey asked teachers which five students they were focusing most on helping. The unit is the roster slot the teacher was offered:

$$F_{ijt} \;=\; \alpha \;+\; \beta_1 T_{cg} \;+\; \beta_2 B_i \;+\; \beta_3 \left(T_{cg} \times B_i\right) \;+\; \eta_g \;+\; \tau_t \;+\; \varepsilon_{ijt}, \tag{4}$$

where F_{ijt} is one if teacher j named student i in term t and B_i is a pre-treatment moderator, either predicted below proficiency or presence on the report's near-cutoff list. Because teachers name a roughly fixed number of students, β_1 is close to mechanical and β_3 is the object of interest: it measures whether treatment changed *which* students were named, not how many.

4.5 Heterogeneity

Heterogeneous effects augment equation (1) by interacting T_{cg} with a full set of subgroup indicators and omitting the constant, so that each subgroup's treatment effect is read directly rather than as a deviation from an omitted category:

$$Y_{icg} \;=\; \sum_{k=1}^{K} \beta_k \left(T_{cg} \times G_i^k\right) \;+\; \sum_{k=2}^{K} \theta_k G_i^k \;+\; \gamma Y_{icg}^0 \;+\; \pi M_{icg} \;+\; \delta_c \;+\; \eta_g \;+\; \varepsilon_{icg}, \qquad (5)$$

with the reported equality test being $H_0 : \beta_1 = \cdots = \beta_K$. Moderators are of three kinds, all measured before treatment.

First, *predicted proficiency level*. We assign each student to the official band their baseline forecast places them in, taking the lowest band across the student's assessments. Harmonizing bands takes some care. The benchmark reading and state tests define four bands, but the national tests define only three — below average, average, above average, with no "approaching" category — so we map bands by label rather than by position. Label mapping also guards against a quirk in the state ELA files, two of which carry a spurious leading band that a positional rule would shift every category by one. Because a student takes the lowest level across the assessments they sit, and the national tests' compressed predicted scores rarely reach a top band, only 38 students are top-band on everything; we therefore merge the top two levels and estimate three: below, approaching, and proficient or above.

Second, *student characteristics*: free and reduced lunch, English learner status, special education, sex, and race or ethnicity.

Third, *proximity to a cutoff*, an indicator reproducing the rule the report generator itself used. A student is flagged if any baseline predicted score lies within 0.25 standard deviations of a proficiency-band boundary, with the standard deviation taken *within the teacher's own class* rather than across the grade. The distinction matters: a classroom is far more homogeneous than a grade, and a grade-pooled bandwidth flags roughly twice as many students as the reports actually listed. We verified the classroom-relative rule against the lists printed on the baseline reports, which it reproduces for 92% of listed students. Because the printed list exists only for treated teachers, we apply the rule to both arms so that control students carry a counterfactual flag.

All moderators are built from baseline forecasts only. Ninety students entering the network without a prior-year record were instead scored by a separate model built for students with no history; its forecasts sit on a visibly different scale and correlate poorly with the main model's where both exist, so we treat these students' moderators as missing. They still contribute to the ITT estimates.

A note on how to read every subgroup result in the paper. The pre-analysis plan designates the breakdowns by student characteristic, proficiency band and assessment as *exploratory*; the confirmatory outcomes are the pooled ones. We therefore report these splits descriptively, without multiple-testing adjustment, and we do not rest any claim on an individual subgroup contrast. A reader should treat a single starred cell among a dozen as what it is.

4.6 Inference

Standard errors are clustered at the campus-grade cell throughout. With 36 cells, cluster-robust asymptotics can overstate precision, so Section 5.4.2 re-evaluates the headline coefficient under randomization inference that respects the stratified design and under a wild cluster bootstrap.

4.7 Relation to the pre-analysis plan

The experimental design, outcome families, treatment indicator and level of clustering follow the pre-analysis plan. Three implementation choices differ from the plan. First, its estimating equation included campus fixed effects but inadvertently omitted grade fixed effects, even though treatment was randomized within grade. We include grade fixed effects to reflect those randomization strata. Second, the plan specified lagged scores, demographics and special-education status as controls. Our preferred specification uses the baseline machine forecast and a missing-forecast indicator. The forecast is a pre-treatment predicted score constructed from lagged achievement and other baseline records, including all of the variables indicated in the pre-analysis plan. It captures their joint predictive content but is not equivalent to entering each registered covariate separately. Appendix B therefore reports both the registered covariates and the preferred forecast control, with and without grade fixed effects.

Third, the plan defined teacher prediction error in raw exam points. The state rescaled its ELA assessment during the study, and the set of assessments asked about on the teacher survey changed across terms. Raw errors are therefore not comparable across the full prediction panel. We instead use the percentile and standardized errors defined above and include assessment-by-term fixed effects.

5 Results

We present results in the order of the causal chain the intervention presumes: information first, behavior second, achievement third.

5.1 Effect on Teacher Beliefs

We had access to all of the school's administrative records, as do teachers. But on top of the administrative data, teachers have a lot of soft data. They see students' work ethic, they know which students are struggling with which concepts, and they have a sense of students' moods or mindsets. So there is a first order question on if the machine learning algorithm or teachers themselves have more accurate predictions.

Teachers meet together to discuss their incoming students, so if a student is not new, teachers know test scores and prior teacher's perceptions of students. Panel A of Table 4 gives us insight into if teachers or the algorithm is more accurate for these returning students. The algorithm is much better. It misses the realized within-grade percentile by 14.3 points against the teacher's 22.0, and by 0.54 SD against 0.72.

That gap narrows sharply once the year is underway. Panel B repeats the comparison over terms 1-4 in control classrooms, and by then the algorithm's edge is small: 15.8 percentile points against the teacher's 17.1, closer on 53% of predictions, and no difference at all on the standardized metric (0.61 SD each). Teachers start the year at a real disadvantage relative to the model and mostly close it by learning their students.

Table 4: Who forecasts student achievement more accurately

	Percentile error		Error in SD		Observations	
	Teacher	Algorithm	Teacher	Algorithm	Students	Predictions
Panel A. Baseline wave, before any report was delivered (returning students)						
Same students	21.8	13.9	0.714	0.512	369	1,000
Algorithm, all its forecasts	—	14.3	—	0.507	2,569	7,233
Panel B. Terms 1–4, control classrooms						
All students	17.3	15.5	0.605	0.601	738	2,415
Bottom 50%	17.3	14.4	0.640	0.594	450	1,352
Top 50%	17.5	15.6	0.552	0.493	368	974

Mean absolute error of each forecast against the realized score. Percentile ranks and standard deviations are taken within grade for each assessment. Panel A uses the baseline survey wave, before any report was delivered, and is restricted to returning students, the only ones the model can forecast at that point. Its first row holds the sample fixed across the two forecasters; the second shows the algorithm on every baseline forecast it produced. Panel B uses control classrooms only, over the students for whom both forecasts exist. It is split at the median of the *baseline machine forecast* for the same assessment, the same moderator used in Table 5, so the two tables cut the sample the same way; this requires an assessment-specific forecast that predates the year, which is why Panel B's subgroup rows do not sum to the row above them.

The algorithm's baseline-wave advantage holds everywhere in the distribution. Figure 4 plots mean absolute error in standard deviations by decile of the baseline machine forecast, pooling teachers across arms, which is unambiguous in the baseline wave because no report had been delivered. The algorithm is more accurate in all ten deciles. Both forecasters do worst at the bottom, where teachers miss by 0.97 SD against the algorithm's 0.69, and the gap between them narrows but never closes as the forecast rises.

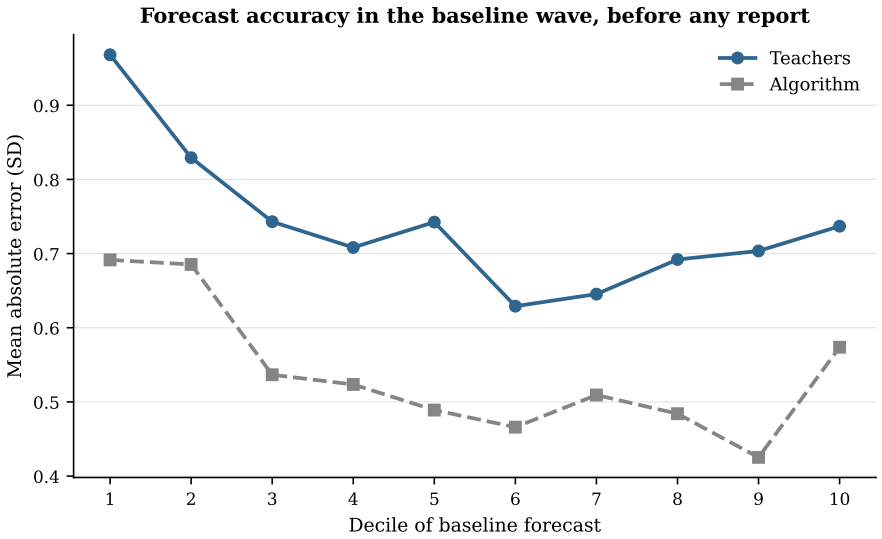


Figure 4: Forecast accuracy in the baseline wave, before any report was delivered. *Notes:* Mean absolute error in standard deviations, over the 2,036 baseline-wave predictions covering 755 students for whom a teacher forecast, a machine forecast and a realized score all exist. Teachers are pooled across arms, since no report had yet been delivered. Deciles are of the baseline machine forecast within assessment and grade, the same moderator used throughout the paper.

Figure 5 shows how the accuracy evolves throughout the school year. Teachers improve steadily, from 0.72 SD in the baseline wave to 0.59 by term 4. The algorithm does not: it sits near 0.50 all year and is no better at term 4 than at term 1, despite being re-estimated each term on more of the current year's data. A large advantage in August is mostly gone by spring. The reports were at their most informative exactly when they were first delivered, and progressively less so as teachers learned their own students.

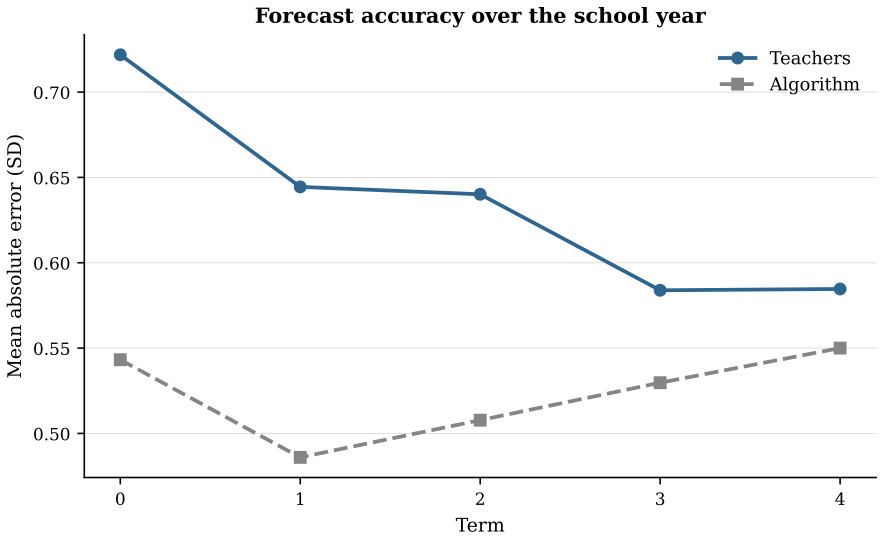


Figure 5: Forecast accuracy over the school year. *Notes:* Mean absolute error in standard deviations; teachers pooled across arms. Within each term both forecasters are evaluated on the same student-by-assessment cells. Algorithm forecasts for terms 1–4 are recovered from the class rosters printed on the reports.

Table 5 compares treated and control teachers over terms 1–4. Treated teachers' absolute percentile error is 2.12 points lower ($p = .007$) against a control mean of 17.5, and their standardized error is 0.051 SD lower ($p = .035$).

Students are split by the decile of the baseline machine forecast for their own assessment. That moderator is fixed before any report was delivered. Splitting instead on the student's realized decile would condition on the outcome: a student lands in a low realized bin partly because of what the treatment did to them, so the arms would not be composed alike within a bin. The cost is sample. The forecast must be assessment-specific, since a student may be forecast strong in reading and weak in mathematics, and it must genuinely predate the year, so students whose only forecast was generated in a later term drop out. Together these leave 5,101 of the 5,982 post-treatment predictions.

The gain is larger at the bottom of the distribution. Percentile error falls 2.11 points among the bottom 50% against 1.31 points among the top 50%; in standard deviations the contrast is -0.083 against -0.001. The difference between the two groups is not itself significant on either metric ($p = .38$ and $p = .08$), so the concentration is a pattern in the point estimates rather than something the data establish. The more informative comparison may be the top half on its own: an effect of -0.001 SD on 2,545 predictions is a precisely estimated zero, not an underpowered one. Figure 6 shows the same thing decile by decile.

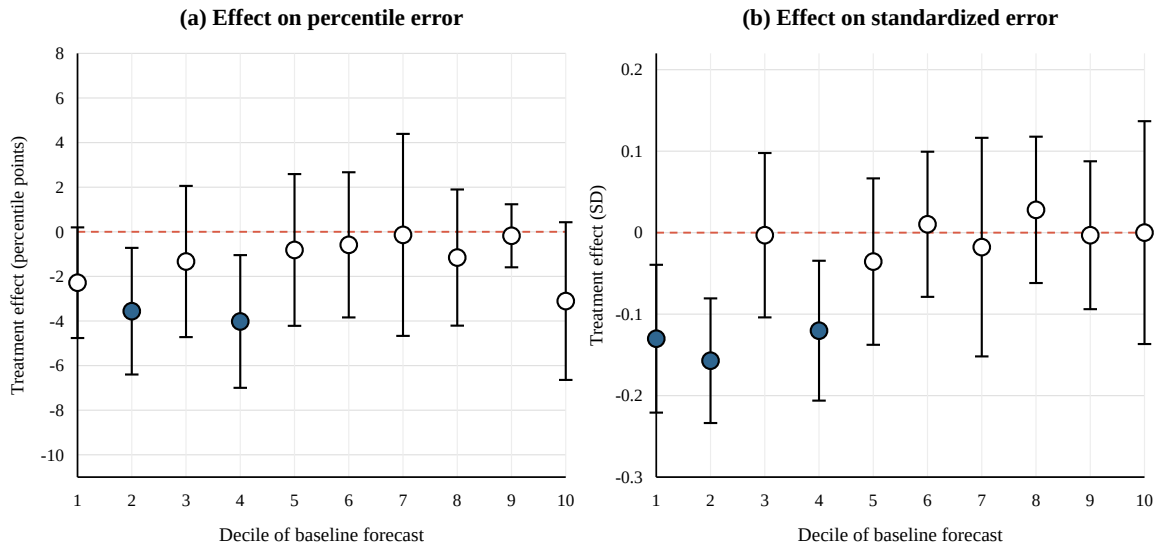


Figure 6: Teacher prediction accuracy across the forecast distribution. *Notes:* Completed responses, terms 1–4. Students are grouped by the decile of the baseline machine forecast for the same assessment, fixed before any report was delivered. A negative value means treated teachers were more accurate. Both panels include assessment × term, campus, and grade fixed effects. Bars are 95% confidence intervals clustered at campus-grade.

Table 5: Teacher prediction accuracy, terms 1–4

	ITT	Control mean	p	N
Panel A. Percentile error (0–100)				
All students	-2.12^{***}	17.59	0.007	5,101
	(0.78)			
Bottom 50% of students	-2.11^{*}	17.69	0.065	2,556
	(1.15)			
Top 50% of students	-1.31	17.46	0.153	2,545
	(0.91)			
Difference	-1.21		0.379	
Panel B. Standardized error (SD)				
All students	-0.051^{**}	0.610	0.035	5,101
	(0.024)			
Bottom 50% of students	-0.083^{***}	0.652	0.007	2,556
	(0.031)			
Top 50% of students	-0.001	0.552	0.990	2,545
	(0.040)			
Difference	-0.079^{*}		0.081	

Unit of observation is a teacher-student-term prediction. The 5,101 predictions cover 1,684 students and 69 teachers; a student appears once per assessment per term their teacher was surveyed. Negative coefficients mean treated teachers were more accurate. Bottom 50% is deciles 1–5 of the *baseline machine forecast* for the same assessment, within assessment and grade. That moderator is fixed before any report was delivered, so the split does not condition on the outcome; it requires an assessment-specific forecast that predates the year, which drops 881 of the 5,982 post-treatment predictions. All models include interacted assessment × term fixed effects, so they compare teachers answering the same question in the same wave, plus campus and grade fixed effects. *Difference* is the treatment interaction with the bottom-50% indicator. Standard errors clustered at campus-grade. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A natural worry is that treated teachers bought this accuracy by being talked into lower expectations. Figure 7 suggests they did lower them, modestly and only at the bottom. Control teachers forecast above the algorithm across the bottom six deciles, by +0.29, +0.32 and +0.22 SD in deciles 1 through 3; they are more optimistic than the machine about weaker students. Treated teachers sit almost on the algorithm's line, at +0.18, +0.06 and +0.04. Above the seventh decile the three series are all but indistinguishable.

We are cautious about how much weight this bears. The treated-control gap in forecast level is −0.094 SD in the bottom half and does not reach conventional significance ($p = .35$), the decile-by-decile effects are individually insignificant, and Appendix D finds no detectable movement in signed error either. What the figure shows is therefore a direction, not an

established effect.

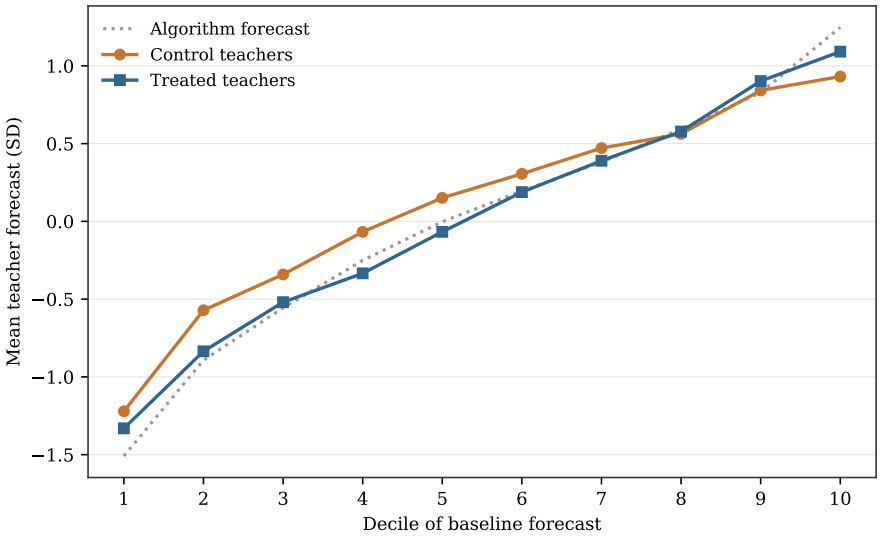


Figure 7: Level of the teacher's forecast, by baseline decile. *Notes:* Completed responses, terms 1–4, for students with a pre-treatment baseline forecast for their own assessment. Each point is a mean within one decile of that forecast. The dotted line is the algorithm's own forecast. Unlike signed error, the level of the forecast does not involve the realized score, which the treatment itself moved.

Taken together, Figure 7 shows the reports pulling teachers' forecasts toward the model. The convergence is most visible in the bottom six deciles, where control teachers forecast above the model and treated teachers track its predictions much more closely. A regression of the teacher forecast on the model forecast quantifies the pattern: the slope rises from 0.572 in control classrooms to 0.885 in treated classrooms, a difference of 0.314 ($p = .038$), or about 55%. This comparison complements the pre-registered accuracy result: it does not use realized scores, so it cannot be produced mechanically by treatment changing achievement. The reports therefore both shifted stated beliefs in the direction of the information they provided and reduced teachers' ex post forecast error. We next ask whether these changes in beliefs translated into changes in teacher behavior.

5.2 Effect on Teacher Efforts

We try to examine teacher efforts in three ways. First, we see if treatment caused teachers to be more active in monitoring students, as measured by the Learning Progress Charts (LPCs). Second, we look at self-reported teacher efforts, including time spent on various activities and one-on-one support. Third, we analyze whether the teachers reported focusing on students who were flagged as struggling by the reports.

5.2.1 Flagging of Struggling Students

The first measure is the measure of teachers monitoring struggling students more on the LPC charts. Treated teachers logged significantly more flags on the Learning Progress Charts: 2.72 additional cumulative flags per student by the end of the year, driven by math and reading, with no effect on spelling or behavior flags.

Table 6 reports the full set of LPC results. Because the first reports were delivered before the first snapshot, even that early chart is post-treatment, and tracing the effect across the year's three snapshots reveals a clean dose-response pattern: treated students carry +0.53 more cumulative flags at the first snapshot, +1.59 by mid-year, and +2.72 by year's end. That is a 44% increase over the control mean of 6.25. Figure 8 plots the pattern.

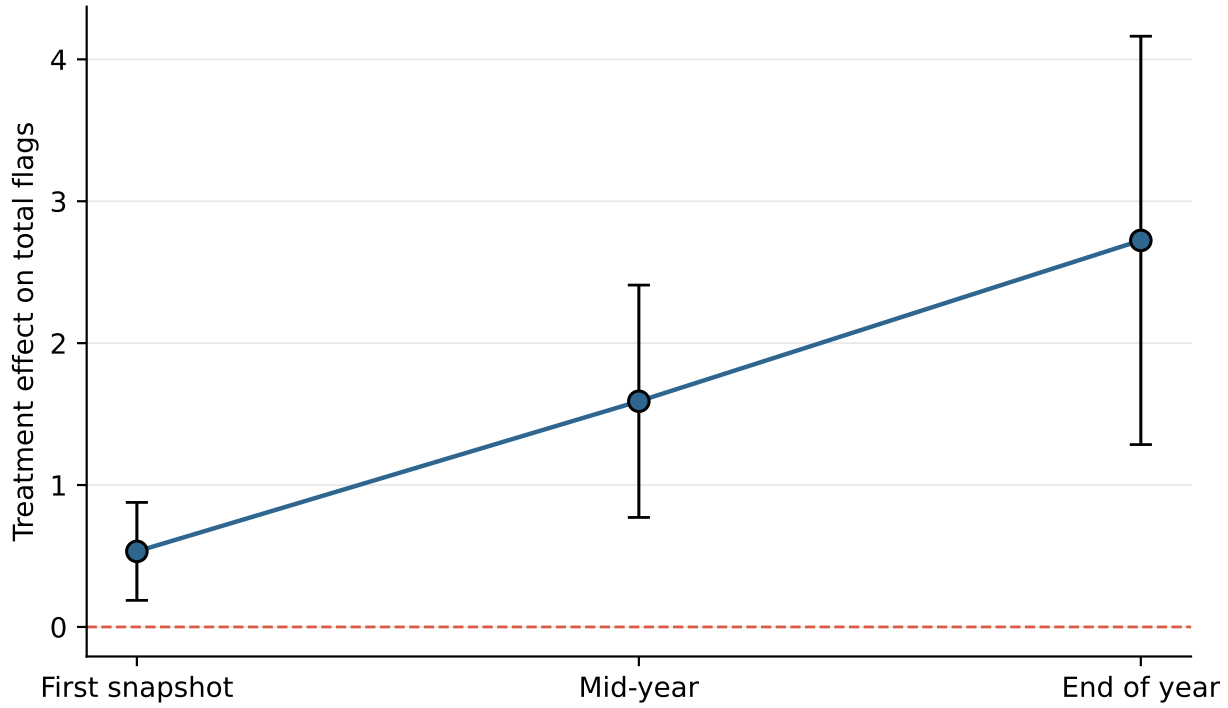


Figure 8: Treatment effects on teacher monitoring, by snapshot. *Notes:* 95% confidence intervals; SEs clustered at campus-grade.

Table 6: Treatment effects on monitoring-chart flags

Outcome	ITT	Control mean	p	N
Panel A. End of year, by category				
Total flags	$+2.724^{***}$	6.25	0.000	3,068
	(0.734)			
Math flags	$+1.741^{***}$	3.21	0.000	3,068
	(0.438)			
Reading flags	$+1.068^{***}$	2.54	0.003	3,068
	(0.362)			
Spelling flags	-0.107^{*}	0.29	0.093	3,068
	(0.064)			
Behavior flags	+0.021	0.20	0.580	3,068
	(0.039)			
Panel B. Total flags, by snapshot				
First snapshot	$+0.532^{***}$	1.40	0.002	3,068
	(0.176)			
Mid-year	$+1.590^{***}$	3.45	0.000	3,068
	(0.418)			
End of year	$+2.724^{***}$	6.25	0.000	3,068
	(0.734)			

Flags are failure indicators on the network internal monitoring charts; the charts are cumulative, so counts are weakly increasing within a student. No snapshot is used as a control: the first snapshot already follows the first report delivery, so conditioning on it absorbs part of the effect. The extensive margin, appearing on a chart at all, is omitted because it is mechanically weakly increasing and was flat throughout. All models include campus and grade fixed effects, with standard errors clustered at campus-grade. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

We read these results as a first stage in behavior: the reports changed who teachers tracked and monitered more closely. And to confirm this we then ask teachers to list the students they are most concerned about.

5.2.2 Which Students Teachers Say They Are Focusing On

The LPCs show teachers flagging struggling students more. They do not neccesarily show if a teachers focus on those specific students more. So, we asked directly. Which five students are you focusing most on helping right now? Teachers named them from their own roster. This is the most direct measure of attention allocation we have..

Treated teachers did redirect attention, and they redirected it exactly where the reports

pointed. Among students the model predicted would fall below proficiency, the probability of being named a focus student is 37.3% in treated classrooms against 29.6% in control classrooms. Among everyone else it falls, from 11.5% to 8.5%. Since teachers name a roughly fixed number of students, this is a composition effect rather than an increase in attention overall: the treatment reallocated focus toward flagged students and away from the rest.

The near-cutoff list does not drive this. Students on the report's near-cutoff list are more likely to be named in both arms, but treatment does not add to that (-0.045, $p = .28$, column 3). Teachers responded to the level the report assigned a student, not to how close the student was to a boundary.

Table 7: Which students teachers say they are focusing on

	Estimate	SE	p
Panel A. Any focus selection			
Treatment	-0.0074	(0.0086)	0.385
Panel B. By predicted proficiency			
Treatment	-0.0229	(0.0180)	0.203
Below proficiency	$+0.1839^{***}$	(0.0245)	0.000
Treatment × below proficiency	$+0.1185^{**}$	(0.0522)	0.023
Panel C. By the near-cutoff list			
Treatment	$+0.0018$	(0.0185)	0.923
On the near-cutoff list	$+0.0479^{***}$	(0.0177)	0.007
Treatment × on the list	-0.0137	(0.0360)	0.704
Observations	5,473	4,547	4,779

Linear probability models, one per panel. The outcome is an indicator for the student being named when the teacher was asked which five students they were focusing most on helping. Unit of observation is the teacher-term-student slot, terms 1–4. All models include grade and term fixed effects, with standard errors clustered at campus-grade. The observation counts read left to right as Panels A, B and C; Panels B and C are smaller because each moderator is defined only for students with a baseline forecast. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

So it seems that treated teachers held more accurate beliefs, flagged struggling students more often on the LPCs, and named those same students as the ones they were focusing on. To get a better understanding of how teachers changed their behavior, we next look at self-reported measures of teacher time use.

5.2.3 Teacher Time Use

Time use outcomes are measured at the teacher-term level and estimated following McKenzie [2012]: each regression controls for the teacher's own baseline value of the outcome and a missing indicator, plus school, grade, and term fixed effects. Shares are in percentage points and counts enter as $\log(1+x)$, so those coefficients read as approximate proportional changes.

Relative to control teachers, treated teachers report reducing the number of times they consulted the student information system and other records (specifically, consultations of master academic records fell by 25 log points and assessment records fell by 23 log points), and spending less time planning interventions (a decrease of 33 log points), while increasing the share of class time spent on new material (which rose by 9.1 percentage points as review time fell by 8.4 points).

The fact that they spent more time covering new material could mean one of two things. It could mean they taught more material, or it could mean that they taught the same material but more slowly. Given the focus on struggling students it seems more likely that they taught the same material but more slowly.

And perhaps most importantly, the reports reduced the time planning interventions (by those 33 log points). This is consistent with the idea that they were changing their teaching focus onto the struggling students. And thus, spent less unique time planning interventions for specific students because the focus of the whole-class instruction was more focused on the struggling students already.

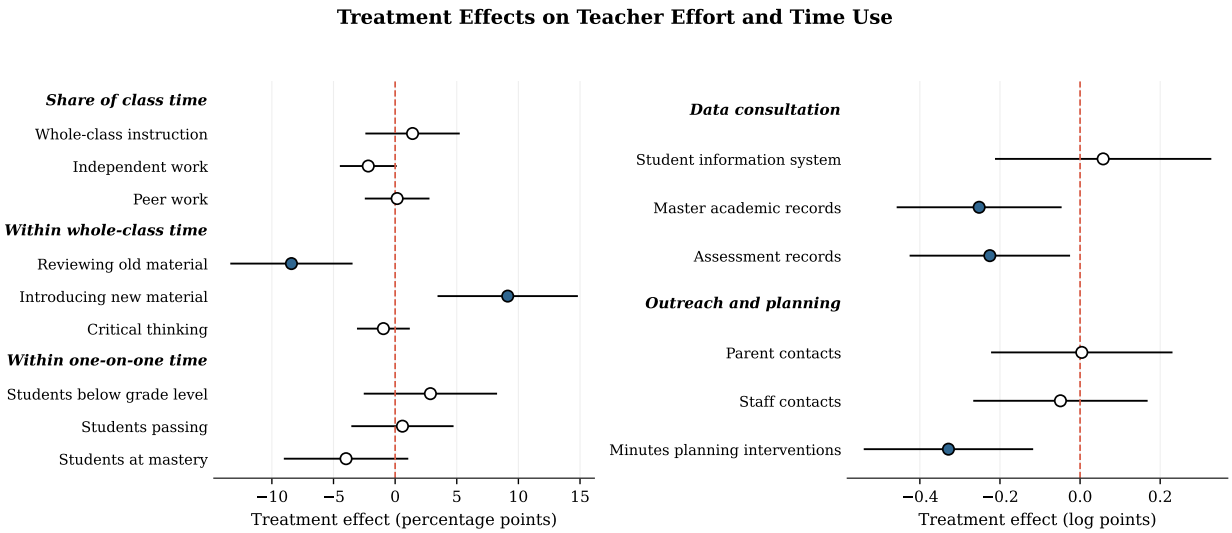


Figure 9: Treatment effects on teacher effort and time use. *Notes:* Filled markers denote $p < 0.05$; bars are 95% confidence intervals with SEs clustered at campus-grade.

Table 8: Treatment effects on teacher effort and time use

Outcome	ITT (SE)	Control mean	p	N
Panel A. Share of class time (percentage points)				
Whole-class instruction	+1.413 (1.914)	70.12	0.460	168
Independent work	−2.180* (1.137)	19.59	0.055	168
Peer work	+0.163 (1.299)	10.29	0.900	168
Panel B. Within whole-class time (percentage points)				
Reviewing old material	−8.414*** (2.492)	30.81	0.001	166
Introducing new material	+9.133*** (2.866)	51.52	0.001	166
Critical thinking	−0.953 (1.054)	17.67	0.366	166
Panel C. Within one-on-one time (percentage points)				
Students below grade level	+2.862 (2.720)	47.56	0.293	162
Students passing	+0.593 (2.077)	33.00	0.775	162
Students at mastery	−3.982 (2.537)	19.44	0.116	162
Panel D. Data consultation (log points)				
Student information system	+0.058 (0.137)	2.62	0.673	160
Master academic records	−0.252** (0.104)	1.63	0.015	161
Assessment records	−0.225** (0.101)	1.67	0.026	160
Panel E. Outreach and planning (log points)				
Parent contacts	+0.004 (0.114)	1.64	0.970	161
Staff contacts	−0.049 (0.110)	2.00	0.656	161
Minutes planning interventions	−0.329*** (0.107)	3.24	0.002	161

Observations are teacher-terms. Each model controls for the teacher own baseline value of the outcome with a missing indicator, and includes school, grade and term fixed effects; standard errors are clustered at school-grade. Panels A through C are percentage shares and are compositional within a panel. Panels D and E enter as $\log(1 + x)$, so coefficients read as approximate proportional changes. Control means are taken over post-treatment teacher-terms. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

So we see that the reports changed teacher beliefs, and that those changes in beliefs led to changes in teacher behavior. The next question is whether those changes in behavior translated into changes in student achievement.

5.3 Effect on Student Achievement

Table 9 reports the paper's central result. Panel A uses the normalized score: each student's realized score on each assessment, standardized within assessment and grade on control-group moments, then averaged over the assessments that student sits. Panel B uses proficiency.

Every model controls for the baseline predicted score, includes campus and grade fixed effects, and clusters standard errors at the campus-grade cell that treatment was assigned to.

Treatment reduced the normalized score by 0.099 standard deviations. The effect is negative on all five assessments. Two are individually significant: national reading and state math. The other three are negative and imprecise, which is what 36 randomization cells should be expected to deliver. The consistency of sign across two subjects, three assessment systems, and six grades is more informative than any single coefficient.

A normalized score and a proficiency rate can move in opposite directions. If the reports pulled attention toward students sitting just under a cutoff, the share clearing that cutoff could rise even as mean achievement falls — an unattractive trade in welfare terms, but the trade most accountability systems reward. Panel B estimates it directly, by linear probability model.

Proficiency comes from each assessment's own reported performance level. The first row of Panel B is the share of a student's assessments where they were proficient; the rows beneath it are student level per-assessment probabilities.There is no threshold gain. Treatment moved the share of assessments passed by -0.019 against a control mean of 0.696.

Table 9: Effect of the reports on student achievement

Outcome	ITT	Control mean	p	N
Panel A. Normalized score				
Normalized score (pooled)	-0.099^{**}	-0.019	0.032	3,009
	(0.046)			
Benchmark reading	-0.053	-0.000	0.309	3,003
	(0.052)			
State ELA	-0.001	-0.000	0.988	1,910
	(0.053)			
State math	-0.108^{**}	0.000	0.010	1,902
	(0.042)			
National reading	-0.267^{***}	-0.000	0.000	826
	(0.050)			
National math	-0.082	-0.000	0.550	834
	(0.136)			
Panel B. Proficiency				
Proficient (share of assessments)	-0.019	0.696	0.251	3,009
	(0.016)			
Benchmark reading	0.009	0.811	0.632	3,003
	(0.019)			
State ELA	-0.030	0.551	0.132	1,910
	(0.020)			
State math	-0.021	0.600	0.190	1,902
	(0.016)			
National reading	-0.051^{**}	0.728	0.041	821
	(0.025)			
National math	-0.059	0.829	0.114	829
	(0.037)			

Each row is a separate regression. Panel A outcomes are standardized within assessment and grade on control-group moments; the pooled row averages over the assessments a student sits. Panel B outcomes are indicators for reaching proficiency, estimated by linear probability model; the first row is the share of the assessments a student sits that they pass. All models control for the baseline predicted baseline with a missing indicator and include campus and grade fixed effects. Standard errors clustered at campus-grade in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

5.3.1 Heterogeneity

All of the evidence we are seeing seems ot suggest that teachers reallocated effort toward the students the reports flagged as at risk. It is natural to wonder if that helped those students

but harmed the other students. That story suggests that losses should fall on students the model expected to do well, and the flagged students should gain. Table 10 looks for that pattern in two ways. First, we split students by whether they were on the near-cutoff list printed on the reports. And second, we split students by the baseline predicted proficiency band.[2]

First, we look at the near cutoff lists. Panel A shows nothing resembling bubble-teaching. Listed and unlisted students have statistically indistinguishable effects on the normalized score and on proficiency.

Panel B splits on the level the baseline forecast placed a student in, taking the lowest band across the assessments they take. On the normalized score every band is negative. Suprisingly, the largest loss falls on students predicted to be *below* proficiency. Students predicted proficient or above also lose ground; the middle band seems relatively unaffected. On proficiency the same ordering appears but nothing is individually significant.

This is suprising given the evidence that teachers said they were focusing on struggling students, and named them at a rate eleven percentage points higher than control teachers did. They flagged them more often on the monitoring charts. And yet they had worse scores.

[2]Heterogeneity by student characteristics is reported in Appendix Table 17.

Table 10: Heterogeneity in the achievement effect

	Normalized score		Proficiency	
Subgroup	ITT	N	ITT	N
Panel A. On the near-cutoff list				
On the list	−0.088	1,525	−0.018	1,525
	(0.066)		(0.028)	
Not on the list	−0.105*	1,484	−0.019	1,484
	(0.054)		(0.020)	
Equality (p)	0.822		0.979	
memo: no baseline prediction		357		357
Panel B. Baseline predicted proficiency band				
Below	−0.226**	497	−0.070	497
	(0.096)		(0.044)	
Approaching	−0.063	638	−0.005	638
	(0.059)		(0.038)	
Proficient or above	−0.148***	1,434	−0.026	1,434
	(0.045)		(0.016)	
Equality (p)	0.099		0.472	

Treatment effects estimated within each subgroup from a single regression interacting treatment with a full set of subgroup indicators, so no category is omitted and each coefficient reads directly. The equality row tests that the subgroup effects are the same. Panel A uses the near-cutoff rule from the report generator, applied to both arms; the two rows partition the estimation sample. A student with no baseline prediction could not appear on the printed list – the reports carry a separate students-without-predictions block – and is counted as not listed; the memo line gives that count. Panel B uses the lowest band across the assessments a student sits, and is defined only for students with a baseline prediction, so it does not partition the full sample. Standard errors clustered at campus-grade. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Another way to look at the same question is to ask what happened to the shape of the achievement distribution rather than to particular subgroups. Figure 10 plots the cumulative distribution of the achievement index in each arm, after partialling out the baseline forecast and the campus and grade effects. The treated curve lies above the control curve at every point (Kolmogorov–Smirnov $D = 0.088$, $p < .001$): the whole distribution shifts left rather than one tail being pulled down.

Figure 11 sets the two heterogeneity questions side by side on a common horizontal scale. Panel (a) estimates the main specification within each decile of the baseline forecast, so students are sorted by a moderator fixed before any report was delivered and stay in the same bin. Panel (b) reports quantile treatment effects at each decile midpoint, comparing the same

point of the treated and control outcome distributions. Both are broadly negative across the distribution. The two panels answer different questions — fixed groups of students against fixed positions occupied by different students — and coincide only under rank invariance, so we read them together rather than against each other.

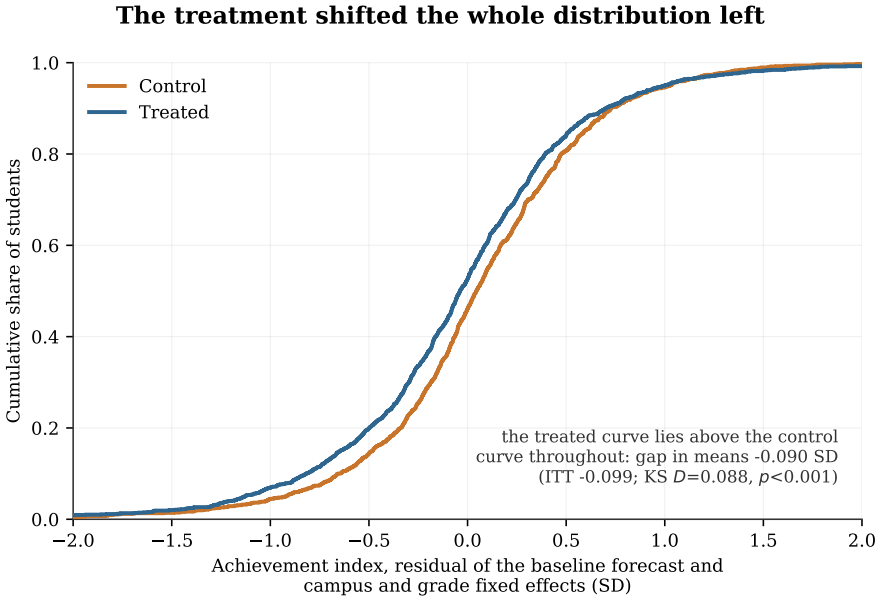


Figure 10: The achievement distribution by arm. *Notes:* Empirical cumulative distributions of the normalized achievement index, after residualising on the baseline forecast, its missing indicator, and campus and grade fixed effects. Units are control-group standard deviations. The gap in means between the curves is −0.090 SD, close to the regression ITT of −0.099; the two differ slightly because treatment is not exactly orthogonal to the controls.

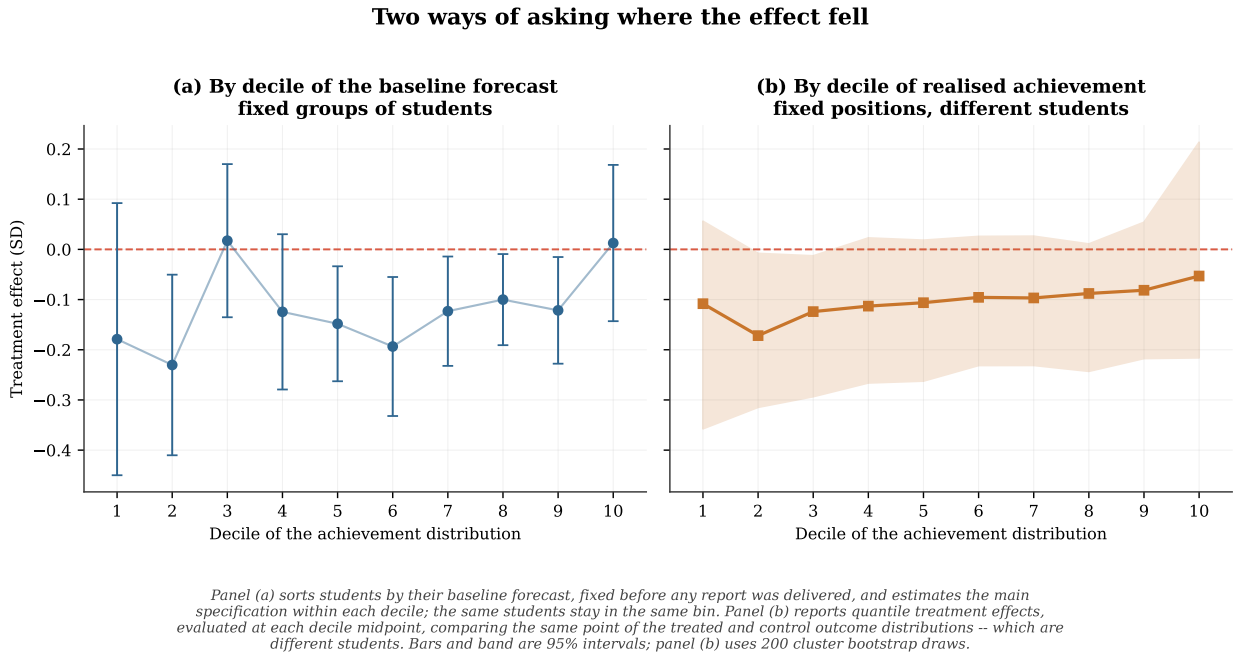


Figure 11: Where the effect fell: fixed groups against fixed positions. *Notes:* Both panels are indexed by decile of the achievement distribution. Panel (a) estimates equation (1) within each decile of the baseline machine forecast, restricted to the 2,652 students who have one; bars are 95% confidence intervals clustered at campus-grade. Panel (b) plots quantile treatment effects with the same controls, with a 95% band from 200 bootstrap draws resampling the 36 campus-grade cells, evaluated at each decile's midpoint. A quantile treatment effect compares two marginal distributions at the same point and is not the effect on a given student.

These results is a bit flumexing. But we leave our interpretation of the achievement result to the discussion section, and turn next to robustness checks.

5.4 Robustness

5.4.1 Attrition and matching

Because outcomes come from administrative score files, differential presence in those files could mimic treatment effects. It does not. The share of eligible students with an end-of-year score is statistically indistinguishable across arms for the benchmark test (97.1% treated vs. 97.0% control), the state tests (84.1% vs. 84.8%), and the national test (77.0% vs. 75.3%). Interactions of treatment with baseline score in predicting file presence are all null (no selective attrition), and within matched samples baseline predicted scores are balanced across arms (all $p > .15$).

5.4.2 Inference with 36 clusters

Treatment is assigned to 36 campus-grade cells, and the national tests are identified from only the 12 grade 1–2 cells. Cluster-robust asymptotics can overstate precision at that scale. We report cluster-robust p-values in the main tables because they are the convention and because the alternatives require judgment calls we would rather make visible than bury; Appendix Table 14 reports two design-based alternatives alongside them. The first is randomization inference, re-drawing placebo assignments that respect the stratified design — within each grade, exactly three of six campuses treated — over 2,000 draws. The second is a wild cluster bootstrap with the null imposed and Rademacher weights at the campus-grade level, over 1,999 replications.

Both temper the asymptotic p-values, and they agree with each other. The pooled effect of -0.099 carries $p = .11$ under randomization inference and $p = .09$ under the bootstrap, against .03 asymptotic. State math moves from $p = .010$ to .03 and .05. National reading, the largest estimate in magnitude, is the most affected: $p < .001$ asymptotic becomes .11 and .13, reflecting how few cells sit behind it. A reader who prefers design-based inference should read the achievement result as suggestive rather than conventionally significant. We think that is the honest reading, and it does not change the paper's argument, which rests on the joint pattern across beliefs, monitoring, time use, and scores rather than on any single p-value.

The same question applies to every other main result, so Table 11 reports all of them the same three ways. The pattern is consistent: the bootstrap and randomization inference move p-values in the same direction and by similar amounts, and nothing reverses sign or changes magnitude, because these procedures rebuild the sampling distribution rather than the estimate. What changes is which results clear conventional thresholds.

Three groups emerge. The monitoring effect is the most robust: +2.72 flags, with $p = .001$ asymptotic and $p = .011$ and .008 under the bootstrap and randomization inference, so treated teachers demonstrably logged more problem weeks. The teacher-effort results mostly hold — reviewing old material, introducing new material and time spent planning interventions all stay under .05 on all three, and the shift toward flagged focus students holds at .045 and .055; consultation of assessment records is the exception, moving from .033 to .103 and .120. The achievement and belief results are the ones that do not survive: the pooled score goes from .039 to .086 and .105, and both belief metrics land between .05 and .11.

We read this as the paper's central caveat rather than a footnote to it. The behavioural results — teachers monitored more, shifted what they taught, and spent less time on records and intervention planning — are estimated on outcomes that vary within cluster and survive

design-based inference. The achievement effect, which is what a reader most wants to know, is the one that 36 clusters can least support. The argument rests on the joint pattern, and we would not want a reader to take the score effect as established on its own.

Table 11: The main results under three inference procedures

			p-value		
Result	Table	Coef.	CRVE	WCB	RI
Beliefs					
Percentile error	5	−2.121	0.011	0.081	0.051
Standardized error	5	−0.051	0.044	0.108	0.110
Monitoring					
Total flags, end of year	6	+2.724	0.001	0.011	0.008
Focus students					
Treatment × predicted below proficient	7	+0.119	0.030	0.045	0.055
Teacher efforts					
Reviewing old material	8	−8.414	0.002	0.008	0.019
Introducing new material	8	+9.133	0.003	0.015	0.030
Master academic records	8	−0.252	0.021	0.035	0.069
Assessment records	8	−0.225	0.033	0.102	0.120
Minutes planning interventions	8	−0.329	0.004	0.026	0.043
Achievement					
Pooled normalized score	9	−0.099	0.039	0.086	0.105

Every row tests the same coefficient on the same sample; the columns differ only in how the sampling distribution is formed. CRVE is the cluster-robust p-value reported throughout the paper, with the CR1 correction and $G - 1$ degrees of freedom. WCB is a wild cluster bootstrap with the null imposed and Rademacher weights (1,999 replications), resampled at campus-grade. RI is randomization inference, replaying the 3-of-6 grade-stratified lottery 2,000 times on unchanged outcomes. It permutes the *studentized* statistic, recomputing the cluster-robust standard error within each draw, following Chung and Romano [2013]: permuting the raw coefficient is exact only under the sharp null of no effect for any student, and becomes conservative for the average effect when effects are heterogeneous, because treatment then inflates the variance of the pool being permuted. That inflation is material here — the placebo coefficients for the pooled score have a standard deviation 1.5 times the analytic one, and permuting the raw coefficient would report $p = .17$ rather than .11. For the focus-student row the coefficient is an interaction: the bootstrap imposes the null on the interaction term, and randomization inference rebuilds both the treatment and the interaction column from each permuted assignment. Subgroup regressions are not reported here.

The experiment therefore detects the behavioral response more strongly than the achievement consequence. That is what its ex-ante power implies: behavior is measured at the teacher level with a tight pre-post design, while score effects of 0.1 SD sit near the edge of

what 36 cells can distinguish from zero. What lends the achievement result credibility despite marginal formal significance is its consistency — all five assessment-specific estimates are negative, across two subjects, three assessment systems, and six grades.

5.4.3 Take-up: who actually used the reports

Assignment guarantees delivery, not use. Delivery records show a 92% open rate, but opening an email is a weak proxy for consulting a forecast, and the resulting first stage of 0.957 leaves the Wald estimate indistinguishable from the ITT. Our best measure of actual use is an end-of-year survey administered to treated teachers only, which asked directly how the reports were used. Twenty-six of the fifty-four treated teachers responded, and eleven of them, 41% of respondents, report never having used the reports at all. Most of the rest consulted them about once per term.

Instrumenting self-reported use with assignment gives a first stage of 0.578 ($F = 35$) and a local average treatment effect of -0.181 (SE 0.116), roughly twice the ITT, as one expects when two-fifths of the treated group are non-compliers. Splitting the treated group is more informative than rescaling it. Table 12 compares each group of treated students against the full control group, ordering treated teachers first by whether they used the reports and then by how useful they found them. The estimates line up monotonically: the more a teacher engaged with the forecasts, the worse her students did.

Table 12: Achievement effects by reported take-up

Treated teachers who…	ITT vs. control	Teachers	Students
Panel A. Reported use			
used the reports	−0.135	15	362
	(0.085)		
did not use them	−0.041	11	250
	(0.063)		
Panel B. Reported usefulness			
moderately or very useful	−0.111	9	244
	(0.086)		
slightly useful	−0.075	10	246
	(0.119)		
not useful	+0.132	4	69
	(0.148)		

Each row estimates equation (1) on the full control group plus the indicated subset of treated students, so the rows share a comparison group and are not a partition of the treated arm. Take-up is self-reported on the end-of-year survey and is not randomly assigned, so these contrasts describe which teachers engaged rather than identifying an effect of engagement. Standard errors clustered at campus-grade in parentheses. Three teachers who used the reports did not answer the usefulness item, so the panels do not sum identically.

We report this as suggestive. Use is self-selected rather than assigned, so these contrasts compare teachers who differ in unobserved ways; the measure is self-reported after outcomes were known; only 48% of treated teachers responded, and response is unlikely to be independent of use. What can be said is that the negative effect does not appear to be driven by teachers who ignored the intervention. If anything it is concentrated among those who engaged with it, which is difficult to reconcile with an account in which the reports simply failed to reach anyone.

6 Discussion

The experiment establishes an unusually complete chain. The forecasts were more accurate than teachers' own, most of all at the start of the year and most of all for students at the bottom of the distribution. Treated teachers read them, their stated beliefs moved toward the truth, again concentrated in the bottom of the distribution. They acted on them, and not only in the cheap, legible way of adding flags to a monitoring chart, but they said that they were focusing most on helping the students the model had flagged. And their reported

use of time changed: less consultation of the underlying records, a third less time planning interventions, and a large shift of whole-class instruction away from review.

Then achievement fell, and fell most for the flagged students.

What makes this hard to explain away is that every link before the last one worked. The usual consolation when an information intervention disappoints is that the information never reached the point of decision. Here it plainly did. We set out two explanations for the last link. They are not mutually exclusive. Both are consistent with everything we observe, and we suspect both operated.

6.1 Lower expectations and a slower class

The first explanation runs through pacing. Treated teachers lowered their expectations for most students and directed attention toward the lowest performers, and instruction slowed to accommodate them.

Both halves of that are visible. On expectations, Figure 7 shows control teachers forecasting above the algorithm across the bottom six deciles — more optimistic than the machine about weaker students, by +0.29, +0.32 and +0.22 SD in deciles 1 through 3 — while treated teachers sit almost exactly on the algorithm's line. The treated-control gap in forecast level is −0.094 SD in the bottom half. It does not reach conventional significance on its own ($p = .35$), and the difference between halves is marginal ($p = .10$), so we read this as a direction rather than an established effect. What moved is calibration, downward, for the students who had the furthest to fall.

On attention, treated teachers named model-flagged students as their focus twelve percentage points more often, and logged 2.72 more cumulative flags per student by year's end, 44% above the control mean. On pacing, whole-class instruction moved sharply away from reviewing previously taught material (−8.4 percentage points) and toward introducing new content (+9.1).

If a class is pitched to its weakest students, everyone moves more slowly through material. The strong are held back, and the weak receive class-wide accommodation in place of the targeted support they had been getting: one-on-one time with below-grade students does not rise enough to compensate, while intervention planning falls 33 log points. This account fits the breadth of the loss — a slower class is a slower class for everyone — and it fits the one departure from breadth, that the flagged students fare worst (−0.226 SD against −0.148 for students predicted proficient or above) despite receiving more of the teacher's stated attention. Attention that arrives as a lowered ceiling is not obviously worth having.

Two things this explanation is *not*. It is not a reallocation from strong students to weak

ones: the strong did not absorb the cost, they lost too. And it does not operate through teachers writing students off in any way we can detect in their stated beliefs. If teachers were doing that, their signed prediction errors should have turned pessimistic about flagged students. They did not: treated and control teachers show identical signed errors overall ($p = .99$) and identically across the forecast distribution. What we cannot rule out is a channel running through classroom behaviour or student perception without moving the teacher's stated belief, and we have no student-side data to test it.

6.2 The reports replaced productive deliberation

The second explanation runs through what the forecast displaced rather than what it caused. Control teachers, lacking a bottom-line forecast, had to construct one. They consulted records, worked through each student's situation, and — because diagnosis and treatment are joint products of the same deliberation — emerged with intervention plans and a disposition to shore up weak material. Treated teachers received the diagnosis pre-computed and economized on the process that would have produced it.

On this reading the deliberation was not overhead but input. Working through a student's records does not merely produce a number; it produces an understanding of *why* that student is likely to struggle — which skills are weak, which lessons did not land — and that understanding is what makes an intervention worth delivering. The algorithm supplied the number without the understanding. The treatment operated less like giving a doctor a better diagnostic test and more like giving them the diagnosis while skipping the examination: the prescribed treatment suffers even though the diagnosis is more accurate.

The time-use evidence is what recommends this account. Treated teachers cut queries to academic records by 25 log points and to assessment records by 23, and cut time planning student interventions by 33 log points — the largest effort response in the survey, and the one most directly tied to converting a diagnosis into an action. These are not the marks of teachers who used the forecast as one input among several. They are the marks of teachers who substituted it for the work.

Like the first explanation, this one fits the breadth of the loss: if what declined is classroom-level practice rather than student-specific targeting, everyone bears some cost. And it accommodates the flagged students faring worst, since they received more of the teacher's stated attention and less of the individualized planning that attention would otherwise have consisted of.

6.3 What we can and cannot distinguish

The two explanations are close to observationally equivalent in our data. Both run through the displacement of individualized planning; both predict a broad loss with the flagged students worst off; both are consistent with the same survey responses. They differ in what replaced the planning — a slower whole-class pace in the first, nothing in particular in the second — and separating them would take instructional pacing data, how far through the curriculum each class actually got, which we do not have.

What these accounts share is more useful than what separates them. In both, the information did its job and the failure occurred downstream, in the machinery that converts knowing which students are struggling into doing something that helps them. That machinery — the deliberation, the planning, the pacing — is where the value of a prediction is realized or lost, and it is precisely the part that a forecasting tool does not touch.

7 Conclusion

We randomized whether elementary school teachers received accurate, individualized machine-learning forecasts of their students' end-of-year test scores. The forecasts moved teachers' beliefs toward the model, made them more accurate, and focused their monitoring on the students the model identified as struggling. They also reduced student achievement by about 0.10 standard deviations ($p = 0.03$ with cluster-robust standard errors, 0.09 to 0.11 under randomization inference and a wild cluster bootstrap). The loss is broad but it is largest among the very students the model flagged as struggling. Every link in the causal chain worked except the last one: the forecasts were accurate, teachers believed them, and teachers acted on them. What changed alongside was the machinery through which attention becomes instruction. Handed the conclusion, teachers did less of the record-checking, intervention planning and review through which they would otherwise have reached it; and they seemed to lowered their expectations for the students furthest behind, pitching the class to match. We cannot say which of these did the damage, and it may be both.

The result cautions against a common implicit assumption in the deployment of decision-support tools: that supplying a high-quality answer weakly improves on letting the professional work it out, since they could always ignore the tool. When the process of working it out has direct productive value, automation of judgment can subtract more than it adds. In our setting the prediction was the cheapest part of what teachers do with student data — and the only part the algorithm could replace.

References

Melissa Adelman, Francisco Haimovich, Andres Ham, and Emmanuel Vazquez. Predicting school dropout with administrative data: new evidence from Guatemala and Honduras. *Education Economics*, 26(4):356–372, July 2018. doi: 10.1080/09645292.2018.1433127. URL https://doi.org/10.1080/09645292.2018.1433127.

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. Technical Report NBER Working Paper 31422, National Bureau of Economic Research, July 2023. URL https://www.nber.org/papers/w31422.

Alex Albright. The Hidden Effects of Algorithmic Recommendations, November 2024. URL https://researchdatabase.minneapolisfed.org/concern/publications/g732d927w.

Alberto Alesina, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. Revealing Stereotypes: Evidence from Immigrants in Schools. *American Economic Review*, 114(7):1916–1948, July 2024. doi: 10.1257/aer.20191184. URL https://www.aeaweb.org/articles?id=10.1257/aer.20191184.

Victoria Angelova, Will Dobbie, and Crystal S Yang. Algorithmic Recommendations and Human Discretion. *Working Paper*, 2024.

Taha Barwahwala, Aprajit Mahajan, Shekhar Mittal, and Ofir Reich. Is Model Accuracy Enough? A Field Evaluation Of A Machine Learning Model To Catch Bogus Firms. Technical Report NBER Working Paper 32705, National Bureau of Economic Research, July 2024. URL https://www.nber.org/papers/w32705.

Peter Bergman, Chana Edmond-Verley, and Nicole Notario-Risk. Parent skills and information asymmetries: Experimental evidence from home visits and text messages in middle and high schools. *Economics of Education Review*, 66:92–103, October 2018. doi: 10.1016/j.econedurev.2018.06.008. URL https://www.sciencedirect.com/science/article/pii/S027277571630629X.

Peter Bergman, Elizabeth Kopko, and Julio Rodriguez. A Seven-College Experiment Using Algorithms to Track Students: Impacts and Implications for Equity and Fairness. Technical Report w28948, National Bureau of Economic Research, Cambridge, MA, June 2021. URL http://www.nber.org/papers/w28948.pdf.

Kelli A. Bird, Benjamin L. Castleman, and Yifeng Song. Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, 44(2):379–402, 2025. doi: 10.1002/pam.22569. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.22569.

Simon Burgess and Ellen Greaves. Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. *Journal of Labor Economics*, 31(3):535–576, July 2013. doi: 10.1086/669340. URL https://www.journals.uchicago.edu/doi/full/10.1086/669340.

David Card and Laura Giuliano. Universal screening increases the representation of low-income and minority students in gifted education. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13678–13683, November 2016. doi: 10.1073/pnas.1605043113. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5137751/.

Michela Carlana. Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics*, 134(3):1163–1224, August 2019. doi: 10.1093/qje/qjz008. URL https://doi.org/10.1093/qje/qjz008.

Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–22, New York, NY, USA, April 2022. Association for Computing Machinery. doi: 10.1145/3491102.3501831. URL https://doi.org/10.1145/3491102.3501831.

EunYi Chung and Joseph P. Romano. Exact and Asymptotically Robust Permutation Tests. *The Annals of Statistics*, 41(2):484–507, 2013. doi: 10.1214/13-AOS1090.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, April 2020. doi: 10.1145/3313831.3376638.

Sharnic Djaker, Alejandro J. Ganimian, and Shwetlena Sabarwal. 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, October 2024. doi: 10.1016/j.econedurev.2024.102575. URL https://www.sciencedirect.com/science/article/pii/S0272775724000694.

Maya Escueta, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan. Upgrading Education with Technology: Insights from Experimental Research. *Journal of Economic Literature*, 58(4):897–996, December 2020. doi: 10.1257/jel.20191507. URL https://www.aeaweb.org/articles?id=10.1257/jel.20191507&ArticleSearch%5Bwithin%5D%5Barticletitle%5D=1&ArticleSearch%5Bwithin%5D%5Barticleabstract%5D=1&ArticleSearch%5Bwithin%5D%5Bauthorlast%5D=1&ArticleSearch%5Bq%5D=&JelClass%5Bvalue%5D=I2&journal=2&from=j.

Seth Gershenson, Stephen B. Holt, and Nicholas W. Papageorge. Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, 52:209–224, June 2016. doi: 10.1016/j.econedurev.2016.03.002. URL https://www.sciencedirect.com/science/article/pii/S0272775715300959.

Jessica Gnas, Julian Urban, Markus Daniel Feuchter, and Franzis Preckel. Socio-emotional experiences of primary school students: Relations to teachers' underestimation, overestimation, or accurate judgment of their cognitive ability. *Social Psychology of Education*, 27(5):2417–2454, October 2024. doi: 10.1007/s11218-024-09915-1. URL https://doi.org/10.1007/s11218-024-09915-1.

Justine S. Hastings, Mark Howison, and Sarah E. Inman. Predicting high-risk opioid prescriptions before they are given. *Proceedings of the National Academy of Sciences*, 117(4):1917–1923, January 2020. doi: 10.1073/pnas.1905355117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1905355117.

Sara B Heller, Benjamin Jakubowski, Zubin Jelveh, and Max Kapustin. Machine Learning Can Predict Shooting Victimization Well Enough to Help Prevent It, 2024. Working paper.

Andrew J. Hill and Daniel B. Jones. A teacher who knows me: The academic benefits of repeat student-teacher matches. *Economics of Education Review*, 64:1–12, June 2018. doi: 10.1016/j.econedurev.2018.03.004. URL https://www.sciencedirect.com/science/article/pii/S0272775717306635.

Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: application to pretrial public safety assessment*. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167–189, May 2023. doi: 10.1093/jrsssa/qnad010. URL https://academic.oup.com/jrsssa/article/186/2/167/7024674.

Esther Kaufmann. How accurately do teachers' judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology*, 63:101902, October 2020. doi: 10.1016/j.cedpsych.2020.101902. URL https://www.sciencedirect.com/science/article/pii/S0361476X20300679.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293, February 2018. doi: 10.1093/qje/qjx032. URL https://doi.org/10.1093/qje/qjx032.

David McKenzie. Beyond Baseline and Follow-Up: The Case for More T in Experiments. *Journal of Development Economics*, 99(2):210–221, 2012. doi: 10.1016/j.jdeveco.2012.01.002.

Sendhil Mullainathan and Ziad Obermeyer. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *The Quarterly Journal of Economics*, 137(2):679–727, April 2022. doi: 10.1093/qje/qjab046. URL https://academic.oup.com/qje/article/137/2/679/6449024.

Nicholas W. Papageorge, Seth Gershenson, and Kyung Min Kang. Teacher Expectations Matter. *The Review of Economics and Statistics*, 102(2):234–251, May 2020. doi: 10.1162/rest_a_00838. URL https://doi.org/10.1162/rest_a_00838.

Robert Rosenthal and Lenore Jacobson. Pygmalion in the classroom. *The Urban Review*, 3(1):16–20, September 1968. doi: 10.1007/BF02322211. URL https://doi.org/10.1007/BF02322211.

Megan T. Stevenson and Jennifer L. Doleac. Algorithmic Risk Assessment in the Hands of Humans, September 2022. URL https://papers.ssrn.com/abstract=3489440.

Jiankun Sun, Dennis J. Zhang, Haoyuan Hu, and Jan A. Van Mieghem. Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations. *Management Science*, 68(2):846–865, February 2022. doi: 10.1287/mnsc.2021.3990. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2021.3990.

Anna Südkamp, Johanna Kaiser, and Jens Möller. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3):743–762, August 2012. doi: 10.1037/a0027627. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027627.

Camille Terrier. Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77:101981, August 2020. doi: 10.1016/j.econedurev.2020.101981. URL https://www.sciencedirect.com/science/article/pii/S0272775718307714.

Maria Zhu. Differences in Teachers' Assessments of Students by English Learner Status. *AEA Papers and Proceedings*, 114:523–527, May 2024. doi: 10.1257/pandp.20241019. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20241019.

A Forecast Construction

Training data and features. We train the models on earlier student cohorts in the same network, using each cohort's realized end-of-year score as the outcome. The features include prior-year scores and proficiency bands; current reading and mathematics placement; gradebook means, standard deviations, completion rates, and within-year changes; prior-year versions of the same gradebook measures; attendance; program classifications; and grade. We convert curriculum levels and lessons into ordinal ranks using maps estimated from historical outcomes. Before producing a forecast for wave t, we remove every field derived from a later wave.

Model fitting. The stacked ensemble contains LASSO and ridge regressions, a random forest, LightGBM, and CatBoost. Each learner is estimated in a pipeline that median-imputes numeric variables, mode-imputes and one-hot-encodes categorical variables, and standardizes inputs for the penalized regressions. We tune each learner with 20 trials of Bayesian optimization, minimizing three-fold cross-validated root mean squared error. A ridge meta-learner combines the five out-of-fold predictions using five-fold cross-validation. We then refit the selected ensemble and evaluate it on a held-out 20% sample.

We estimate a separate ensemble for each assessment: state ELA, state mathematics, national reading, national mathematics, and benchmark reading. We pool grades within an assessment and include grade as a feature. This gives the models more training observations at the cost of some grade-specific flexibility.

Student histories. For each assessment, we maintain three model versions. The richest version covers returning students whose gradebooks merge successfully. A second version covers returning students without merged gradebooks and imputes those features. A third covers students new to the network, whose predictions rely on placement, demographics, and any current-year records. We do not model kindergarten because entering students have little predictive history.

B Alternative Achievement Specifications

The pre-analysis plan specified the student-outcome regression as

$$Y_{icg} = \alpha + \beta T_{cg} + \mathbf{X}'_{icg}\gamma + \delta_c + \varepsilon_{icg}, \tag{6}$$

where $\mathbf{X}_{icg}$ contains lagged test scores, demographics and special-education status, and δ_c denotes campus fixed effects. The plan also states that treatment was randomized within grade, but grade fixed effects were inadvertently omitted from equation (6). Our preferred equation (1) adds grade fixed effects and replaces $\mathbf{X}_{icg}$ with the baseline forecast index and a missing-forecast indicator.

Table 13 compares these choices on the same 3,009 students.

Table 13: Alternative specifications for the pooled achievement effect

Specification	ITT	SE	p
No covariates; campus and grade fixed effects	−0.045	0.047	0.344
Baseline forecast index; campus and grade fixed effects (preferred)	−0.099	0.046	0.032
PAP covariates; campus and grade fixed effects	−0.082	0.047	0.077
PAP covariates; campus fixed effects only (PAP equation as written)	−0.020	0.105	0.852
Baseline forecast index; campus fixed effects only	−0.053	0.118	0.651
Baseline forecast index and PAP covariates; campus and grade fixed effects	−0.089	0.045	0.049

The outcome is the pooled achievement index in control-group standard deviations. Every row uses the same 3,009 students and clusters standard errors at campus-grade. "PAP covariates" are lagged test scores, demographics and special-education status. Specifications using the baseline forecast also include a missing-forecast indicator.

Rows 2, 3 and 6 change the covariate set while retaining the grade indicators; the estimated effect moves modestly across them. Rows 4 and 5 show that omitting grade fixed effects matters more because it ignores the randomization strata and substantially reduces precision. We therefore retain equation (1) as the preferred specification while reporting the equation specified in the plan in row 4.

C Results by Individual Assessment

The main text reports per-assessment ITT effects alongside the pooled normalized score (Table 9). This appendix collects the supporting material: the design-based and multiplicity-adjusted inference behind those estimates (Table 14) and heterogeneity broken out by assessment. Recall that grades 1–2 take the national reading and math tests while grades 3–6 take the state ELA and math tests, so those columns are identified off roughly half the sample each (12 and 24 randomization cells respectively); the benchmark reading test spans

all grades.

Every point estimate is negative. The two significant ones — national reading and state math — come from different grade bands, different subjects, and different testing systems, which argues against a test-specific artifact. In the heterogeneity tables that follow, a dash marks a proficiency level that is simply not populated for a given assessment: the national tests define only three bands and their predicted scores are compressed enough that in practice they place students in only two levels. Table 14 reports the design-based p-values for these estimates. The by-assessment split is exploratory, and the pooled normalized score is the confirmatory outcome; no individual assessment should be read as a result on its own.

Table 14: Design-based inference by assessment

Outcome	ITT	p (CRVE)	p (RI)	p (WCB)
Normalized score (pooled)	−0.099	0.032	0.113	0.093
Benchmark reading	−0.053	0.309	0.431	0.441
State ELA	−0.001	0.988	0.991	0.989
State math	−0.108	0.010	0.034	0.052
National reading	−0.267	0.000	0.111	0.130
National math	−0.082	0.550	0.818	0.767

CRVE is the cluster-robust p-value reported in the main text. RI is randomization inference re-drawing placebo assignments that respect the stratified design (2,000 draws). WCB is a wild cluster bootstrap with the null imposed and Rademacher weights (1,999 replications). The by-assessment split is exploratory under the pre-analysis plan and is reported without multiple-testing adjustment; the pooled normalized score is the confirmatory outcome.

Heterogeneity by student characteristics

Table 17 reports treatment effects within subgroups defined by free and reduced lunch, English learner status, special education, sex, and race/ethnicity, alongside the predicted-band split carried in the main text. All sixteen demographic estimates are negative. Two contrasts reach nominal significance — White students, and Asian students showing essentially no effect — but eight characteristics are tested here, this breakdown is exploratory, and the race contrasts drop the one campus with unusable demographic records; two nominal rejections out of eight is close to what chance alone produces. The groups a targeting technology is most often advocated for, namely English learners, special-education students, and low-income students, show effects at least as negative as everyone else's.

D Decomposing the Accuracy Gain

Table 5 shows that treated teachers made smaller absolute errors. Absolute error cannot say where that came from. An accuracy gain can arise from a teacher lowering her forecasts or from her ordering students better, and the two have different implications: a literature beginning with Rosenthal and Jacobson [1968] and sharpened by Papageorge et al. [2020] holds that teacher expectations are an input to achievement rather than only a forecast of it, so a treatment that lowers expectations could depress outcomes directly. Table 15 separates the two on the same sample and the same baseline decile split used in the main text.

We find no evidence for the lowered-expectations channel. Signed error, where positive means the teacher forecast exceeded the realized score, is unmoved: -0.02 percentile points overall ($p = .99$) and -1.86 in the bottom 50% ($p = .46$), against a control mean of only $+1.74$ points, so control teachers were barely over-optimistic to begin with. The level of the teacher's own forecast is likewise essentially unchanged overall ($+0.036$ SD, $p = .59$) and imprecisely negative for low-forecast students (-0.094 SD, $p = .35$).

Decomposing mean squared forecast error into bias, scale and rank components tells the same story from the other direction. Treated error falls from 0.627 to 0.609. Reduced bias accounts for about a quarter of that, and a higher correlation between forecast and outcome for the rest, slightly offset by greater dispersion in treated forecasts. The bias reduction lands near zero rather than overshooting into pessimism: control teachers sit at $+0.070$ SD, treated teachers at -0.001. Whatever produced the achievement decline, it does not appear to be teachers being talked into lower expectations.

Table 15: Decomposing the accuracy gain: signed error and forecast level

	ITT	Control mean	p	N
Panel A. Signed percentile error (0–100)				
All students	−0.02	1.74	0.994	5,101
	(3.10)			
Bottom 50% of students	−1.86	3.33	0.460	2,556
	(2.52)			
Top 50% of students	+1.38	-0.48	0.724	2,545
	(3.92)			
Difference	−3.26		0.233	
Panel B. Level of the teacher forecast (SD)				
All students	+0.036	-0.005	0.587	5,101
	(0.066)			
Bottom 50% of students	−0.094	-0.441	0.349	2,556
	(0.101)			
Top 50% of students	+0.017	0.600	0.884	2,545
	(0.115)			
Difference	−0.163*		0.097	

Appendix companion to Table 5, on the same sample of 5,101 predictions and the same baseline forecast decile split; Table 5 reports absolute error and is not repeated here. Panel A reports *signed* error, where positive means the teacher forecast exceeded the realized score; Panel B reports the level of the teacher forecast itself. Together they separate an accuracy gain achieved by lowering forecasts from one achieved by ordering students better: a treated-control gap in absolute error with no gap in either signed error or forecast level points to sharper ranking rather than a uniform downward shift. All models include interacted assessment × term fixed effects plus campus and grade fixed effects, with standard errors clustered at campus-grade. *Difference* is the treatment interaction with the bottom-50% indicator. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 16: Belief accuracy with teacher fixed effects

	(1) Pooled	(2) Teacher FE	(3) Teacher FE + test×term
Panel A. Difference-in-differences, treatment × post			
Percentile error	+0.569	−0.378	−0.754
	(1.592)	(2.348)	(1.731)
Standardized error	−0.065	−0.066	−0.083
	(0.083)	(0.110)	(0.086)
Observations	8,274	8,274	8,274
Panel B. Pre-treatment wave alone, percentile error			
no controls	+0.182	(1.638)	$p = 0.911$
+ test FE	+0.438	(1.311)	$p = 0.738$
+ test + campus FE	+3.832	(5.194)	$p = 0.461$
+ test + campus + grade FE	+14.342***	(3.292)	$p = 0.000$

Panel A is a difference-in-differences. A post observation is a term-1 to term-4 response, or a baseline-wave response submitted on or after the day the reports were sent. Because no teacher changes arm, a teacher fixed effect absorbs the treatment indicator outright, so the coefficient is identified only from teachers observed on both sides of that date: 5 treated and 10 control. Panel B adds controls cumulatively to the pre-treatment wave alone (886 predictions, 14 clusters). Standard errors clustered at campus-grade. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 17: Heterogeneous Treatment Effects on the Pooled Achievement Index

	Subgroup effect			
	Coef.	SE	N	Diff. p
Panel A. By predicted proficiency level (baseline forecast)				
Below	−0.226	0.096	497	
Approaching	−0.063	0.059	638	
Proficient or above	−0.148	0.045	1,434	
Equality of the three effects				0.099
Panel B. By student characteristic				
Free/reduced lunch	−0.180	0.069	901	0.111
not free/reduced lunch	−0.065	0.049	2,108	
English learner	−0.146	0.083	677	0.533
not English learner	−0.091	0.049	2,332	
Special education	−0.016	0.083	406	0.246
not special education	−0.106	0.045	2,603	
Female	−0.112	0.049	1,435	0.739
not female	−0.126	0.051	1,367	
White	−0.147	0.049	2,000	0.029
not White	−0.042	0.051	802	
Hispanic/Latino	−0.178	0.070	978	0.120
not Hispanic/Latino	−0.079	0.041	1,824	
Black	−0.124	0.083	217	0.931
not Black	−0.118	0.045	2,585	
Asian	−0.011	0.051	580	0.020
not Asian	−0.144	0.050	2,222	

Each row is the treatment effect *within* the stated subgroup; this table is descriptive and no significance claim rests on it. Proficiency level is the lowest band the student's baseline forecasts place them in across all three assessments they sit; students without a baseline forecast are excluded from Panel A. "Diff. p" tests equality of the two subgroup effects. Standard errors clustered at campus-grade. Sample sizes differ down Panel B because race and sex are unusable at one campus, which is dropped from those rows only.

Heterogeneous effects by assessment

Tables 18 and 19 decompose the heterogeneity analysis of Section 5.3.1 by individual assessment. Cell sizes shrink considerably — a proficiency level within a single grade-band-specific test can rest on a few hundred students in twelve randomization cells — so individual esti-

mates are noisy and we draw no conclusions from any single cell. The consistency of sign is what carries over: the large majority of subgroup-by-test estimates are negative, and the significant equality tests do not replicate across assessments.

Table 18: Treatment effects by predicted proficiency level, by assessment

	Benchmark Reading	State ELA	State Math	National Reading	National Math
Below	−0.236	−0.077	−0.186	−0.190	−0.098
Approaching	+0.000	−0.012	−0.155	−0.290	+0.003
Proficient or above	−0.108	−0.047	−0.083	−0.251	−0.150

Treatment effect within each subgroup, estimated separately for each assessment. Descriptive; no significance claim rests on this table, so standard errors and stars are omitted and the pooled-index version in Table 17 carries the precision. "—" means the subgroup is empty or too small to estimate for that assessment.

Table 19: Treatment effects by student characteristic, by assessment

	Benchmark Reading	State ELA	State Math	National Reading	National Math
Free/reduced lunch	−0.105	−0.170	−0.226	−0.081	+0.101
English learner	−0.141	−0.109	−0.218	−0.065	+0.283
Special education	+0.060	+0.002	−0.002	−0.032	+0.101
Female	−0.086	+0.009	−0.082	−0.321	−0.105
White	−0.087	−0.040	−0.103	−0.399	−0.195
Hispanic/Latino	−0.124	+0.007	−0.149	−0.441	−0.186
Black	−0.119	+0.021	−0.094	−0.188	−0.122
Asian	−0.011	+0.162	−0.044	−0.201	+0.041

Treatment effect within each subgroup, estimated separately for each assessment. Descriptive; no significance claim rests on this table, so standard errors and stars are omitted and the pooled-index version in Table 17 carries the precision. "—" means the subgroup is empty or too small to estimate for that assessment.

Proficiency by assessment

Table 20 reports the binary proficiency results of Section 5.3 assessment by assessment. Control-group proficiency rates vary widely across the five — from 0.56 on state ELA to 0.83 on the benchmark reading test and national math — which matters for interpretation, since a linear probability model has least room to detect movement where the rate is already near one. The two assessments with the most room, state ELA and state math, show the effects closest to conventional significance.

Table 20: Treatment effects on proficiency, by assessment

Assessment	ITT	Control mean	p	N
Benchmark Reading	+0.0093	0.811	0.632	3,003
	(0.0194)			
State ELA	-0.0302	0.551	0.132	1,910
	(0.0201)			
State Math	-0.0212	0.600	0.190	1,902
	(0.0162)			
National Reading	-0.0513**	0.728	0.041	821
	(0.0251)			
National Math	-0.0590	0.829	0.114	829
	(0.0373)			

Linear probability models. Standard errors clustered at campus-grade in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.