Can Machine Learning Help Teachers Predict and Improve Student Outcomes? Evidence from a Randomized Controlled Trial*

Daniel Sabey Jake Meyer

October 19, 2026

Preliminary draft

Abstract

Can more precise information about students improve teaching and achievement? We randomly provide elementary-school teachers with machine-learning forecasts of students' end-of-year test scores across six charter-school campuses serving 3,068 students in grades 1–6. Linking repeated teacher surveys to administrative monitoring records and test scores, we measure responses in beliefs, student priorities, and instructional practices. Access to forecasts reduces teachers' measured prediction error and increases their reported focus on students predicted to struggle. However, teachers report less independent information gathering and intervention planning, and shift instruction from review toward new material. We estimate an average achievement decline of 0.099 standard deviations, with larger losses among students predicted to score below proficiency. These findings distinguish identifying students likely to struggle from providing effective support: predictive tools should be paired with effective interventions and not evaluated on accuracy alone.

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 Please do not cite or circulate.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

1 Introduction

Academic achievement gaps emerge early and persist through schooling. Skills measured at school entry predict later achievement [Duncan et al., 2007], and income-related achievement gaps already are substantial at kindergarten entry and do not appreciably narrow as children progress through school [Reardon, 2011]. Improving achievement among struggling students requires teachers to identify learning difficulties early and decide how to address them. This is especially hard when teachers get new sets of students every year, and when the students are young and have not yet taken many assessments. Teachers must assemble information from a variety of sources, such as assessments, gradebooks, prior teachers, and other administrative systems while learning about students through classroom interactions. By the time a teacher develops a clear understanding of a student's needs, opportunities for early support may have passed. More precise information could accelerate this process by making existing records more digestible, thereby helping teachers identify students who are likely to struggle before those difficulties become apparent in the classroom.

Whether more accurate information provided early in the school year improves achievement depends on how teachers respond to it. Earlier identification could help teachers target additional instruction and prepare interventions. But forecasts also can change how much information teachers gather themselves, how they allocate their time, their optimism about their students, and what they teach. A prediction of a student's academic outcomes does not establish how to help that student succeed. Teachers might interpret a low forecast as evidence that additional effort is unlikely to be productive or lower their expectations for specific students. More precise information about likely outcomes therefore need not translate into more effective instructional support without additional training and guidance in how to effectively use this information.

In this paper we study teachers' responses to more precise information in a randomized experiment providing elementary-school teachers with machine-learning forecasts of their students' end-of-year achievement. The experiment covered 3,068 students in grades 1–6 across six campuses of a U.S. charter school network during the 2025–2026 academic year. We randomly assigned 18 of the network's 36 campus-by-grade groups to receive prediction reports, with the remaining groups serving as controls. Teachers in treated groups received an initial report near the beginning of the school year and four subsequent updates. Each report displayed students' predicted scores, the proficiency levels tied to those scores, and a list of students whose predictions were nearest to a proficiency boundary. The experiment allows us to estimate how access to these forecasts changes teachers' beliefs, their reported practices, their monitoring of individual students, and student achievement. We

find that teachers' forecasts become more accurate, they focus more on struggling students, but student achievement declines. The evidence is consistent with substitution away from productive investigation and planning and a shift in classroom focus toward lower-ability students.

The prediction models combine penalized linear models and tree-based methods, trained on earlier cohorts within the network. They use prior assessment scores, curriculum placement, gradebook performance, attendance, and current-year assessment results as these become available. Teachers in the treated and control groups could access the underlying records; the intervention provided a more digestable summary of that information. We forecast scores on each student's three end-of-year assessments: two in reading and one in mathematics. The assessments vary by grade and include the network's reading benchmark, nationally normed tests, and state tests.

A distinctive feature of the study is the detailed information we observe about teachers' responses. Repeated surveys we conducted elicit teachers' predictions of individual students' test scores, how teachers report allocating instructional and planning time, and how often they consult existing student records. Teachers also identify the five students they are most focused on helping, allowing us to connect their stated priorities to the forecasts those students received. We link these surveys to the network's Learning Progress Charts (LPCs), which teachers submit weekly to administrators to flag students struggling in reading, mathematics, spelling, or behavior. We collect snapshots of these charts at three points during the year. Together, these sources let us examine changes in beliefs, students on whom the teachers focus, reported instructional practices, and administrative monitoring. This combination provides a more detailed account of the response to prediction reports than achievement outcomes alone can supply.

We first establish that the model predicts achievement more accurately than teachers and that access to its forecasts improves the accuracy of teachers' predictions. We measure prediction error as the absolute difference between a predicted score and the student's realized end-of-year score, expressed in test-score standard deviations and percentile error. In the baseline comparison of forecasts, the model's average error is 0.2 standard deviations and about 7.9 percentile points lower than the teacher's prediction error. In subsequent survey waves, access to the reports reduces teachers' average prediction error by 0.051 standard deviations and 2.12 percentile points.

Teachers also change their documentation of struggling students and their reported priorities. By the end of the year, treated teachers record 2.72 additional cumulative LPC flags per student, an increase of 44 percent relative to the control mean. In the surveys, treatment shifts teachers' reported focus toward students predicted below proficiency: the treatment

effect on being named a focus student is 11 percentage points larger for these students than for students predicted at or above proficiency. These measures indicate that the reports influence both teachers' stated priorities and the students they document as struggling in reports to administrators.

These changes coincide with reductions in other reported activities. Treated teachers report consulting assessment records less frequently and spending less time planning student interventions. They also report shifting whole-class instruction away from reviewing previously taught material and toward introducing new material. The share devoted to review falls by 8.4 percentage points, while the share devoted to new material rises by 9.1 percentage points. Thus, increased identification and monitoring of struggling students accompany changes in how teachers prepare and deliver instruction.

Teachers working under time constraints may use the forecasts to reduce the time spent assembling and reviewing student records. However, predicting which students will struggle does not reveal their specific learning difficulties or how to address them. Reviewing records may therefore support both prediction and instructional planning. The reductions in reported record consultation and intervention planning are consistent with forecasts substituting for this work, although the experiment does not separately identify this mechanism's contribution to the achievement decline.

Student achievement declines following the intervention. Treatment reduces our pooled standardized achievement index by 0.099 standard deviations. The estimated effect is negative for each of the assessments, and the largest decline, 0.226 standard deviations, occurs among students predicted to score below proficiency—the group teachers more often identify as a priority in the surveys.

Taken together, these results show that more precise information alone is not enough to help students, and indeed it can be harmful. The intervention changes teachers' beliefs and practices, but the behavioral responses do not translate into improved learning. Reduced intervention planning and changes in whole-class instruction provide possible explanations for the achievement losses. The findings suggest that supplying forecasts alone can be insufficient and potentially counterproductive, motivating the question of how to pair predictive information with support for teachers to use the information more effectively.

We contribute to the literature on algorithmically assisted decision-making by being the first to examine how teachers respond to information given via algorithm and how that impacts students' outcomes. Predictive models have the potential to improve decisions in settings such as pretrial release and medical testing [Kleinberg et al., 2018, Mullainathan and Obermeyer, 2022], and Bergman et al. [2021] demonstrate gains from algorithmic course placement in higher education. Yet deployment does not uniformly improve outcomes: eval-

uations of judicial risk assessments and human–algorithm collaboration in radiology show that results depend on how recipients use the information [Stevenson and Doleac, 2024, Imai et al., 2023, Agarwal et al., 2023]. Recommendations, discretionary overrides, and responses to erroneous scores further shape the consequences of prediction tools [Albright, 2024, Angelova et al., 2026, De-Arteaga et al., 2020].[1]

This dependence on recipients' actions connects algorithmic assistance to the classic principal–agent problem: a principal can supply information, but the agent retains discretion over how to act on it. The multitask framework of Holmström and Milgrom [1991] emphasizes that the allocation of effort across activities matters, not just its total amount. Our finding that forecasts improve teachers' prediction accuracy but reduce student achievement is consistent with the conclusion of Barwahwala et al. [2024] that model accuracy does not necessarily translate into improved outcomes. Education is a particularly interesting setting for this question because teachers have substantial discretion over how to act on information, and their actions can directly affect the outcomes being predicted. In this setting where teachers have substantial ability to effect the outcome, we find that, without additional supports to help teachers better use the information, information alone can be insufficient and even counterproductive for students.

We also contribute to the teacher expectation and bias literature by impacting the accuracy of teacher's predictions for their whole class, not for specific students. Prior evidence suggests that learning about students may be an important part of teachers' work: Hill and Jones [2018] find achievement gains when students have the same teacher more than once, consistent with benefits from increased information about a student, although repeated matches also change relationships and other classroom conditions. Our study tests the information pathway directly.

Teachers' judgments are informative but imperfect [Südkamp et al., 2012, Kaufmann, 2020, Djaker et al., 2024]. Their expectations also can influence student outcomes. Rosenthal and Jacobson [1968] experimentally manipulate teachers' expectations, while Papageorge et al. [2020] use differences between two teachers' expectations for the same student to estimate how expectations affect educational attainment. Research on stereotypes and grading biases provides further evidence that teachers' beliefs and assessments can shape students' achievement, educational choices, and social integration [Alan et al., 2018, Carlana, 2019, Terrier, 2020, Alan et al., 2023].[2]

[1]Related work studies predictions of high-risk opioid prescriptions and shooting victimization [Hastings et al., 2020, Heller et al., 2024], human discretion and racial disparities in child welfare [Cheng et al., 2022], and workers' deviations from algorithmic instructions [Sun et al., 2022]. Within education, administrative data predict dropout and college success [Adelman et al., 2018, Bird et al., 2025]; Escueta et al. [2020] review the broader experimental evidence on educational technology.

[2]Related studies document differences in teacher assessments by ethnicity, English-learner status, and

Together, these findings motivate providing information that could help teachers learn about their students earlier. Rather than targeting a particular demographic group or revealing teachers' implicit stereotypes, as in Alesina et al. [2024], we supply forecasts for every student in treated elementary-school classrooms. We test a classroom-wide approach to improving teachers' information, rather than an intervention designed specifically to correct demographic bias.[3] Classroom-wide information can change what teachers do: giving teachers maps of their students' friendship networks reduces social isolation and antisocial behavior [Alan et al., 2025]. For achievement, Brown et al. [2026] add granular information on students' topic mastery, and then student-specific teaching plans, on top of leveled mathematics workbooks in Pakistan; all three arms raise scores by similar amounts, so neither the information nor the guidance adds detectably to the materials. Our design isolates the information itself: treated teachers receive forecasts with no accompanying materials or instructional guidance. Providing forecasts without accompanying instructional support allows us to evaluate a potentially low-cost intervention before considering more resource-intensive approaches that combine information with training or teaching materials.

Our detailed data also expand on the education literature by measuring changes in teachers' beliefs, the students they prioritize, and their reported instructional practices in response to new information. Repeated surveys elicit student-specific forecasts, the names of students teachers prioritize, their consultation of existing records, and their reported allocation of planning and instructional time. Learning Progress Charts provide a separate administrative record of students flagged as struggling in weekly reports to school leaders, observed through three snapshots during the year. These measures connect changes in teachers' assessments to the choices emphasized in research on whom teachers target [Neal and Schanzenbach, 2010], the level at which they teach [Duflo et al., 2011], and the effectiveness of different instructional practices [Zakharov and Carnoy, 2021]. They allow us to distinguish incorporating information and documenting concern from changes in reported instructional support. Teachers make more accurate forecasts and report greater focus on students predicted to struggle but also report less intervention planning and a shift away from review; achievement subsequently declines.[4] Linking these measures provides a fuller account of teachers'

teacher–student demographic match [Burgess and Greaves, 2013, Gershenson et al., 2016, Zhu, 2024], and associations with students' socio-emotional experiences [Gnas et al., 2024].

[3]Other approaches expand assessment through universal screening [Card and Giuliano, 2016] or provide information and support to parents [Bergman et al., 2018].

[4]The role of confidence in sustaining motivation in Bénabou and Tirole [2002] suggests that forecasts might change teachers' perceptions of what their effort can accomplish. Bénabou and Tirole [2003] offer a related framework in which an informed principal's actions convey information that affects an agent's motivation. We do not directly measure perceived returns to effort or separately identify teacher discouragement or student responses to teacher signaling.

responses than achievement outcomes or a single behavioral measure alone. The experiment identifies effects on each of these outcomes, without separately identifying how much each behavioral response contributes to the achievement decline.

The findings highlight a distinction between identifying students who may struggle and improving their achievement. Teachers incorporate the forecasts into their assessments and change their reported priorities and practices, yet student achievement declines. For schools considering predictive tools, the question is therefore not accuracy alone, but whether the information can be paired with effective interventions. Our results motivate evaluating forecasts together with an educational intervention that helps teachers use the information productively, rather than assuming that identifying students at risk is sufficient to improve their learning.

2 Setting and Experimental Design

We ran an experiment 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 are falling behind and where additional attention might help. Prior to our intervention, teachers 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. The network places substantial emphasis on assessed academic performance and maintains assessment records beginning in kindergarten. Teachers therefore work in a setting with an established testing infrastructure, rather than one in which the intervention introduces student assessment for the first time. For returning students, these records can provide a history of achievement even in the early elementary grades.

Most instruction takes place in larger homeroom classes. Students also attend mathematics and reading breakout work groups organized by achievement level. These groups are generally led by teacher aides and emphasize student work rather than a separate block of teacher-led instruction. Administrators determine group placements, and students can be reassigned based on performance. We describe the assessment and classroom arrangements relevant to the experiment without naming the network or its state, consistent with 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 (Table 2). 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.

Table 2: End-of-year assessments by grade

	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 we generated use these same cutoffs to identify predicted proficiency bands and students near a boundary. We standardize each score by subtracting the control-group mean and dividing by the control-group standard deviation for the same assessment and grade. We then average each student's standardized scores to construct the pooled achievement index. Section 4 describes the pooled achievement index in more detail.

Teachers also maintain weekly "Learning Progress Charts" (LPCs). 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 monitored 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 include 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.

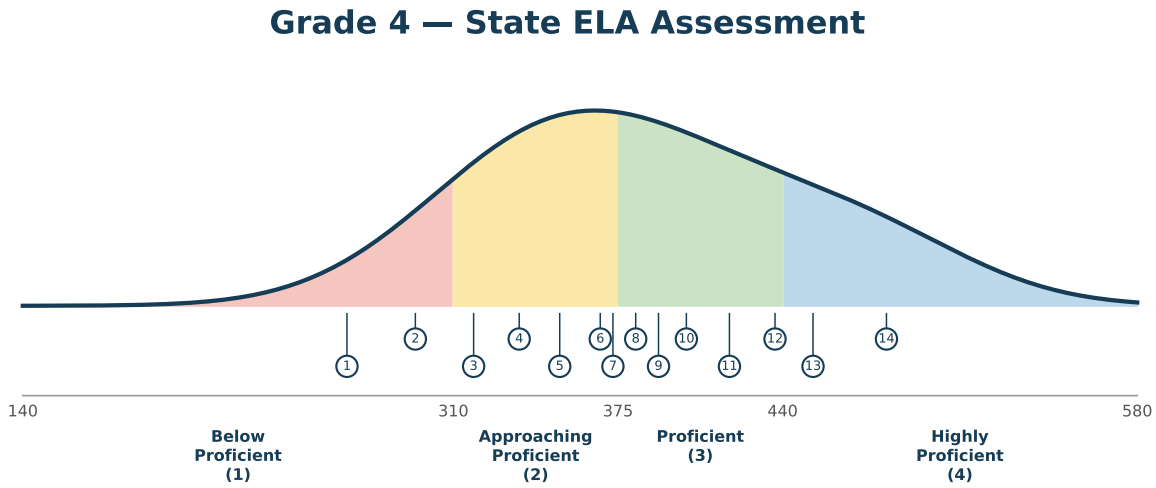
Our information intervention is designed to help make this information easier to digest. Treated teachers received a report with three pieces of information for every student in their homeroom:

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 therefore could 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 train assessment-specific forecasting models on earlier cohorts in the same school network, using realized end-of-year scores as outcomes. Predictors include prior scores, curriculum placement and progress, gradebook performance and completion, attendance, student characteristics, and grade. For each report update, we construct historical predictors corresponding to that point in the school year, incorporating additional records as they become available.

We estimate separate models for five assessments and accommodate differences in available student histories through three versions: returning students with linked gradebooks, returning students without gradebook predictors, and students without prior-year histories. Each version combines penalized linear models and tree-based learners in a stacked ensemble. Appendix A describes the training samples, feature construction, fitting procedure, and historical evaluation diagnostics, including their limitations.

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. The design minimizes the possibility of spillovers, since there is much less cross-grade communication than within grade. Our regressions include campus and grade fixed effects, and we cluster standard errors at the campus-by-grade level. With only 36 clusters, we also report robust inference estimates in Section 5.5.

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.

Presenting this information to teachers in advance minimizes the possibility that treated

teachers simply failed to understand what they were sent, and it means control teachers also knew that the reports existed and what information they contained, although they did not receive forecasts for their own students. Any resulting behavioral response by control teachers would likely narrow the contrast between the treatment and control groups 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.

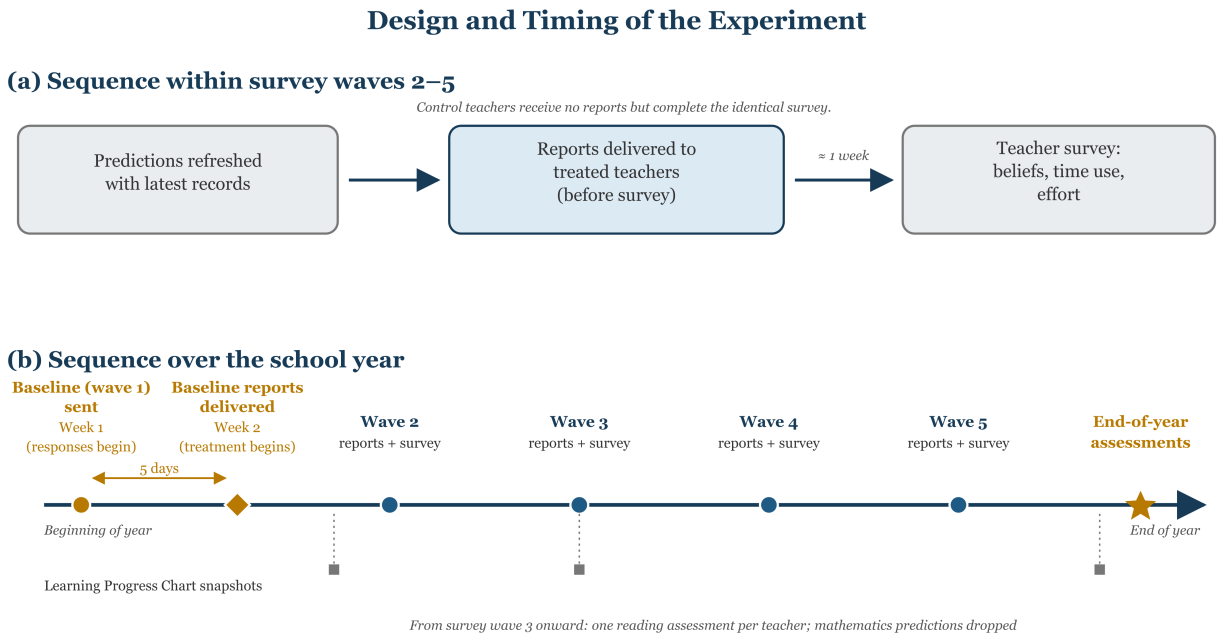


Figure 2: Design and timing of the experiment.

The reports continued to cover every relevant assessment, but we shortened the teacher survey after discussions with the schools to reduce the burden on teachers. We number survey rounds as waves 1–5: wave 1 is the initial survey, and wave 2 follows school term 1, at the start of school term 2. In waves 1 and 2, teachers predicted up to three assessment scores for each student. From wave 3 onward, after school term 2, they predicted scores for one reading

assessment and no longer provided mathematics predictions. This change limits comparisons of mathematics prediction accuracy over time and changes the mix of assessments observed across survey waves. Our prediction-accuracy regressions include assessment-by-wave fixed effects to account for this changing mix.

3 Data

We combine student records, machine forecasts, teacher surveys, and homeroom curriculum trackers.

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 5.5 shows that presence in each score file is balanced between treatment and control groups.

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.

Homeroom curriculum trackers. We also use the network's homeroom curriculum trackers to measure recorded lesson coverage. These trackers list the planned curriculum alongside dated entries for each teacher and subject. We count distinct numbered lessons or chapters recorded during the school year, counting repeated entries once. This measure covers science and history in grades 1–6 and mathematics in grades 1, 3, 5, and 6, where numbered entries permit lesson counts.

Teacher surveys. Teachers were surveyed at baseline (wave 1) and after each school term (waves 2–5). There are 231 teacher-wave 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. Post-treatment time-use questions refer to the previous week; baseline questions ask about typical practice.

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 order in which students appeared in the prediction questions was randomized. In waves 2–5, the survey followed report delivery by roughly one week; baseline survey responses straddled the first report delivery. 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 waves 2–5 and 69% including wave 1, with an average of 1.5 waves per student.

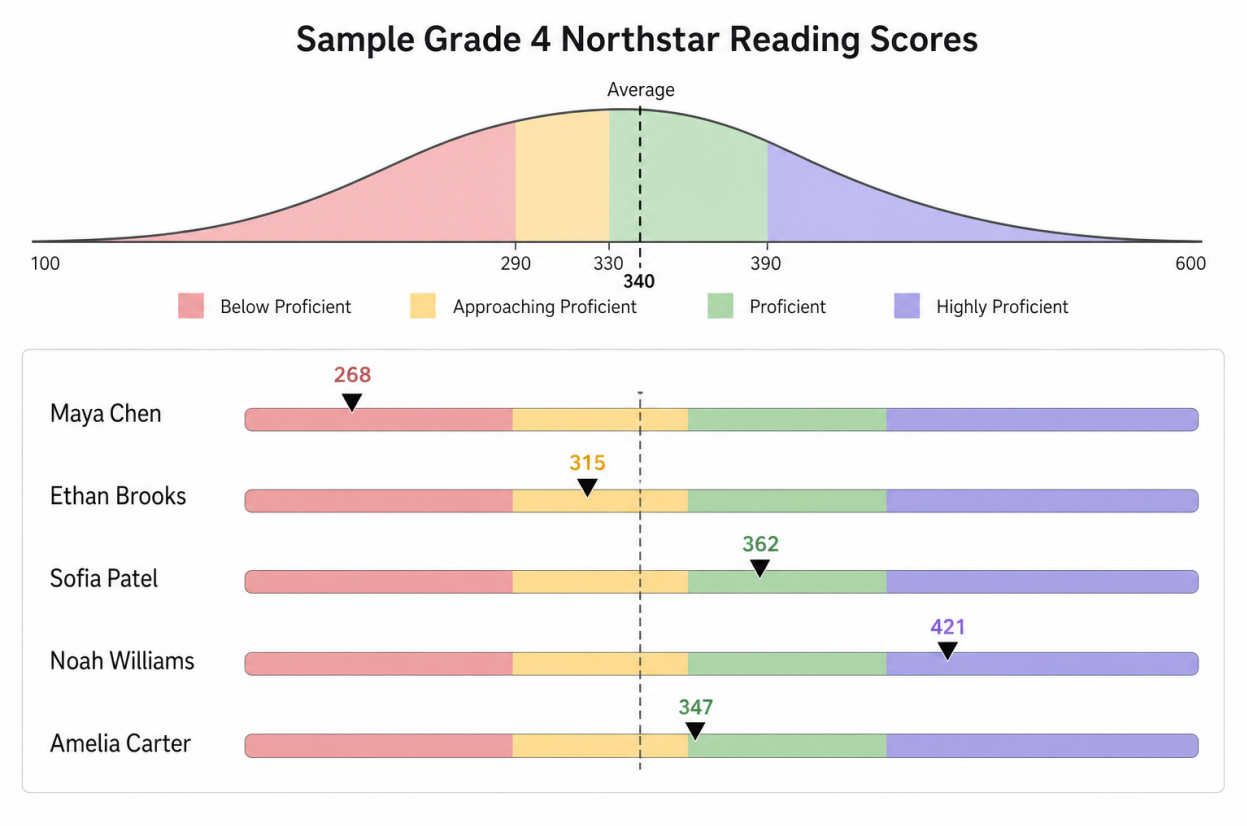


Figure 3: Anonymized reconstruction of the survey interface eliciting teacher predictions of student scores.

The elicitation is an unbalanced panel. Teachers were asked about the full assessment battery in survey waves 1 and 2 and about a single reading assessment in waves 3–5, following the survey shortening described in Section 2.4.

Table 3 compares baseline characteristics across the treatment and control groups before restricting the sample to students with observed end-of-year scores. None of the reported individual differences in the baseline predicted score index or demographic indicators is statistically significant at the 5% level, and the campus- and grade-adjusted joint F-test does not reject balance ($p = .67$). With only 36 randomized campus-grade groups, some chance imbalance is unsurprising: treated students have a higher baseline predicted-achievement index, but the adjusted difference is not statistically significant ($p = .20$). This initial advantage favors the treatment group and therefore does not provide a simple explanation for the negative achievement effect.

Table 3: Baseline Balance

			Raw			Campus and grade FE			
	Control	Treatment	Diff.	SE	p	Diff.	SE	p	N
Baseline predicted score index (z)	-0.054	0.045	0.098	0.130	0.448	0.074	0.057	0.199	3,068
Free lunch	0.246	0.225	-0.020	0.054	0.706	-0.000	0.012	0.995	3,068
English learner	0.156	0.158	0.001	0.051	0.981	-0.008	0.028	0.763	3,068
Joint F-test p-value				0.875			0.669		

Differences are treated minus control, shown both raw and adjusted for campus and grade fixed effects. 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. Because randomization was stratified by grade, the campus- and grade-adjusted columns are the design-consistent comparison; neither the raw nor the adjusted joint test rejects balance.

4 Empirical Framework

Treatment is assigned once to a campus-by-grade group. Teachers report student-specific predictions and time use in repeated survey waves. Final student achievement is measured once per year, using assessments administered at the end of the school year. Each outcome family requires a separate estimating equation. Table 4 reports the level of observation for each outcome family.

Table 4: Level of observation by outcome family

Outcome family	Level of observation	Equation	N
Teacher prediction accuracy	teacher × student × assessment × wave	(1)	5,982
Teacher effort and time use	teacher × wave	(2)	≈165
Focus-student selection	teacher × wave × roster slot	(3)	4,547
Student achievement, proficiency	student	(4)	3,009
Monitoring-chart flags	student	(4)	3,068

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. The focus-student N refers to the predicted-proficiency interaction in Table 8; samples differ across its panels.

4.1 Teacher Prediction Accuracy

Each wave 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:

$$A_{ijmt} = \alpha + \beta T_{cg} + \phi_{mt} + \delta_c + \eta_g + \varepsilon_{ijmt}, \tag{1}$$

Here, A_{ijmt} is teacher j's absolute prediction error for student i on assessment m in wave t. The indicator T_{cg} equals one for students in treated campus c and grade g groups and zero for controls, and β measures the adjusted treatment–control difference. The terms ϕ_{mt}, δ_c, and η_g are assessment-by-wave, campus, and grade fixed effects, respectively; α is the intercept, and ε_{ijmt} is the error term.

From wave 3 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. Assessment-by-wave fixed effects account for differences in the assessments elicited in each wave.

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. The headline sample uses waves 2–5: baseline-wave responses straddled the first report delivery, so that wave is neither cleanly pre- nor cleanly post-treatment.

4.2 Teacher Effort and Time Use

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

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

Here, E_{jt} is teacher j's reported effort or time-use outcome in wave t, and T_{cg} indicates treatment assignment for campus c and grade g. The coefficient β measures the adjusted treatment–control difference. E_j^0 is the corresponding wave-1 outcome, averaged within the teacher's homeroom when multiple responses are available. When this value is missing, we replace it with the mean among observed wave-1 responses and set the separate indicator M_j to one; otherwise, M_j equals zero [following McKenzie, 2012]. The coefficients γ and π adjust for these two variables. The terms δ_c, η_g, and τ_t are campus, grade, and wave fixed effects; α is the intercept, and ε_{jt} is the error term.

Shares enter in percentage points and are compositional within a block. Consultation and contact outcomes enter as counts, and intervention planning enters in minutes; their

corresponding baseline controls use the same units. Standard errors are clustered by campus-by-grade assignment group, the level at which treatment was randomized.

4.3 Focus-Student Selection

From wave 2 the survey asked teachers which five students they were focusing most on helping. The unit is the roster slot the teacher was offered:

$$\begin{aligned} F_{ijt} \;=\; \alpha \;+\; \beta_1 T_{cg} \;+\; \beta_2 B_i \;+\; \beta_3 (T_{cg} \times B_i) \\ +\; \delta_c \;+\; \eta_g \;+\; \tau_t \;+\; \varepsilon_{ijt}, \end{aligned} \tag{3}$$

where F_{ijt} is one if teacher j named student i in wave 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. The terms δ_c, η_g, and τ_t are campus, grade, and wave fixed effects. The overall-selection model omits the moderator and interaction terms.

4.4 Student Achievement

The primary specification is a covariate-adjusted ITT regression. Random assignment identifies the treatment effect without conditioning on baseline forecasts. We include the forecasts to improve precision: they summarize prior-year performance and student characteristics through prediction models that allow nonlinear relationships and interactions among these predictors. For student i in campus c and grade g:

$$Y_{icg} \;=\; \alpha \;+\; \beta\, T_{cg} \;+\; \gamma\, Y^0_{icg} \;+\; \pi\, M_{icg} \;+\; \delta_c \;+\; \eta_g \;+\; \varepsilon_{icg}, \tag{4}$$

Here, 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.

We use the same framework for proficiency and LPC outcomes. For proficiency, the main outcome is the share of observed assessments on which a student meets the proficiency threshold. The baseline covariate is the share of assessments the baseline forecasts predict

will meet the threshold. Assessment-specific proficiency outcomes are binary indicators, estimated using a linear probability model. For LPC outcomes, we omit lagged flag counts because every available snapshot follows initial report delivery, retaining the campus and grade fixed effects in equation (4).

4.5 Heterogeneity

Heterogeneous effects augment equation (4) 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 $H_0 : \beta_1 = \cdots = \beta_K$. We examine heterogeneity by predicted proficiency level, proximity to a cutoff, and student characteristics: free and reduced lunch, English learner status, special education, sex, and race or ethnicity. These moderators are measured before treatment.

Because our primary achievement outcome pools scores across assessments, we also use a single student-level proficiency classification for the subgroup analysis. We classify each student using the lowest proficiency band among their available baseline assessment forecasts. For example, a student predicted to be proficient on both reading assessments but below proficient in mathematics is classified as below proficient. 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. Only 38 students are in the highest band on every assessment; we therefore merge the top two levels and estimate three: below, approaching, and proficient or above.

Our measure of proximity to a cutoff reproduces 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. 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.

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.

4.6 Inference

Standard errors are clustered at the campus-grade cell throughout. With 36 cells, cluster-robust asymptotics can overstate precision, so Section 5.5.3 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 pre-analysis plan provides a common estimating framework for the study's outcomes. We adapt this framework to each outcome's level of observation, measurement, and timing. The illustrative 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. Appendix B compares all headline results with and without grade fixed effects, holding the sample and other regressors fixed within each outcome.

5 Results

We present results in the order of the causal chain the intervention presumes: teacher beliefs first, teacher behavior second, and student achievement third.

5.1 Effect on Teacher Beliefs

Among responding teachers, report access brings forecasts closer to the algorithm and reduces measured prediction error. We first establish the scope for improvement: Table 5 compares teacher and algorithm accuracy before report delivery and in control classrooms later in the year, while Figure 4 shows the algorithm's initial advantage across the forecast distribution. Figure 5 shows how forecast accuracy evolves over the school year for treated and control teachers. Table 6 estimates the reduction in teacher prediction errors caused by report access, and Figure 6 examines how these effects vary across the baseline forecast distribution. Finally, Figure 7 shows that treated teachers' forecasts align more closely with the algorithm, providing evidence of belief updating that does not depend on realized achievement.

Teachers meet to discuss returning students and can consult their prior test scores and previous teachers' assessments. Panel A of Table 5 compares baseline teacher and algorithm forecasts for these students. As the table shows, the algorithm provides more accurate

information. Mean absolute percentile error is 13.9 points for the algorithm and 21.8 points for teachers; the corresponding standardized errors are 0.51 and 0.71.

That gap narrows sharply once the year is underway. Panel B reports average prediction errors in control classrooms, pooling 2,326 student–assessment–wave predictions for 710 students across waves 2–5. Mean absolute percentile error is 17.4 points for teachers and 14.9 points for the algorithm; the corresponding standardized errors are 0.604 and 0.552. The subgroup rows distinguish observations in the bottom and top halves of the baseline algorithm-forecast distribution for the corresponding assessment. Teachers start the year at a real disadvantage relative to the model and close much of it by learning about their students.

Table 5: Who forecasts student achievement more accurately

	Percentile error		Error in SD		Observations	
	Teacher	Algorithm	Teacher	Algorithm	Students	Predictions
Panel A. Wave 1, before any report was delivered						
Students with a teacher forecast	21.8	13.9	0.714	0.512	369	1,000
All students	—	14.3	—	0.507	2,569	7,233
Panel B. Waves 2–5, control classrooms						
Students with a teacher forecast	17.4	14.9	0.604	0.552	710	2,326
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 reports all available baseline algorithm forecasts matched to realized scores, including students without a baseline teacher forecast. This second row uses a broader student sample and has no corresponding teacher comparison. Panel B pools control-classroom predictions across survey waves 2–5 and requires a teacher forecast, an initial-wave algorithm forecast for the same assessment, and a realized score. All three rows use this initial-forecast-eligible sample. The subgroup rows split observations at the median of the *baseline machine forecast* for the same assessment, the same moderator used in Table 6. Prediction counts add across the two halves; unique student counts can overlap because a student may belong to different halves for different assessments.

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 from treatment and control groups, 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.

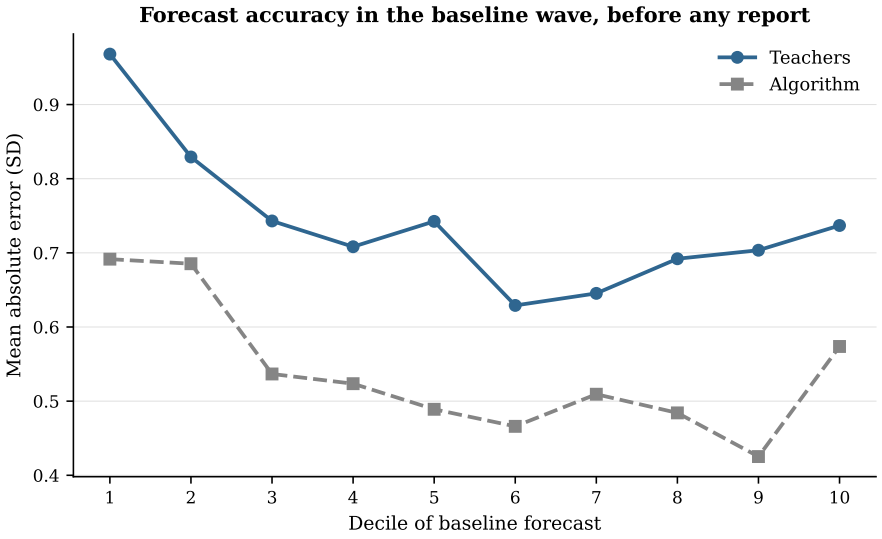


Figure 4: Forecast accuracy in the baseline wave, before any report was delivered. *Notes:* Mean absolute error in standard deviations, over the 2,036 baseline-wave predictions covering 755 students for whom a teacher forecast, a machine forecast and a realized score all exist. Teachers from treatment and control groups are pooled, 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 that prediction errors decline from 0.71 to 0.62 SD for control teachers and from 0.72 to 0.58 SD for treated teachers between waves 1 and 5. The algorithm remains more accurate throughout, while treated teachers have lower errors than control teachers in three of the four subsequent waves.

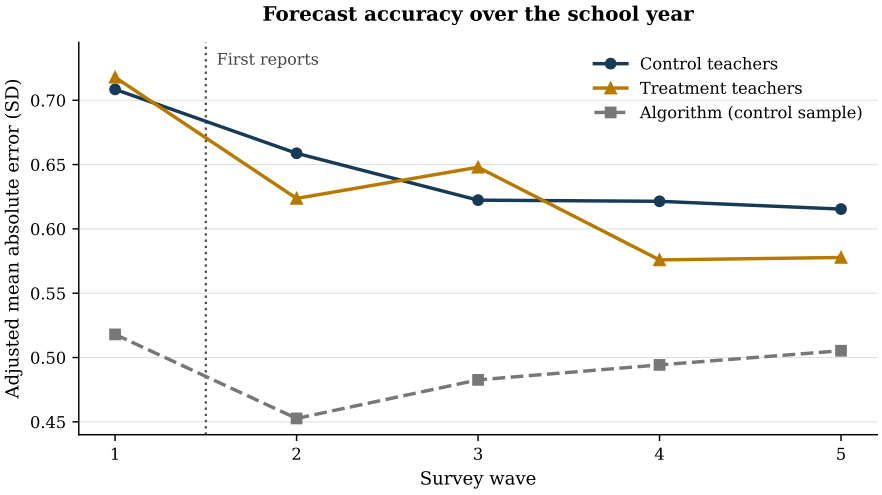


Figure 5: Forecast accuracy over the school year. *Notes:* Mean absolute error (SD), adjusted for campus and grade fixed effects.

Table 6 compares treated and control teachers over waves 2–5. 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 ($p = .035$).

We define subgroups using each student's pre-treatment algorithm forecast for the corresponding assessment. Because treatment may change realized scores, students in the same realized-score group may have differed before treatment. Requiring an assessment-specific baseline forecast leaves 5,101 of the 5,982 post-treatment predictions, which we use for both the pooled and subgroup estimates.

Student order in the prediction questions was randomized, but the treatment and control forecast samples differ in baseline predicted achievement (Appendix Table E3, panel A). Adding flexible controls for baseline predicted performance leaves both estimated treatment effects on forecast accuracy nearly unchanged (panel B).

Teacher prediction error falls by 2.11 percentile points in the bottom half of the baseline forecast distribution and 1.31 points in the top half. The corresponding standardized reductions are 0.083 and 0.001 SD. These point estimates suggest larger reductions in the bottom half, but we cannot reject equal treatment effects across halves ($p = .58$ for percentile error; $p = .11$ for standardized error). Figure 6 plots treatment effects on absolute prediction error by baseline forecast decile.

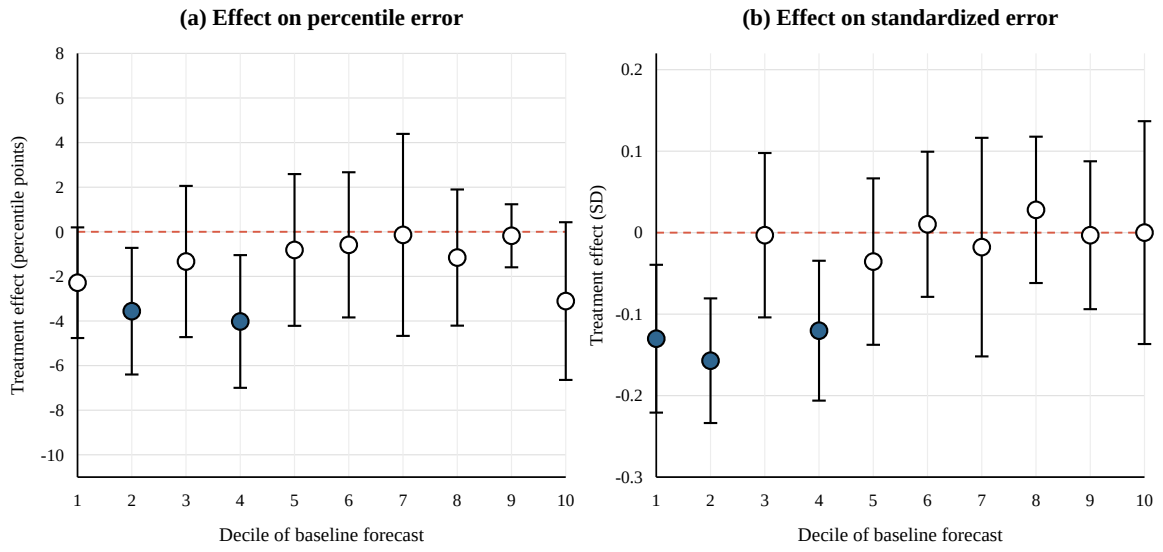


Figure 6: Teacher prediction accuracy across the forecast distribution. *Notes:* Completed responses, waves 2–5. Students are grouped by the decile of the baseline machine forecast for the same assessment, fixed before any report was delivered. A negative value means treated teachers were more accurate. Both panels include assessment × wave, campus, and grade fixed effects. Bars are 95% confidence intervals clustered at campus-grade.

Table 6: Teacher prediction accuracy, waves 2–5

	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	-0.80		0.578	
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	-0.082		0.107	

Unit of observation is a teacher-student-wave prediction. The 5,101 predictions cover 1,684 students and 69 teachers; a student appears once per assessment per wave 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 × wave fixed effects, so they compare teachers answering the same question in the same wave, plus campus and grade fixed effects. *Difference* is the bottom-half effect minus the top-half effect, estimated in one model that interacts every regressor, fixed effects included, with the bottom-50% indicator. Standard errors clustered at campus-grade. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Treated teachers' forecasts track the algorithm more closely across much of the baseline forecast distribution. In control classrooms, teacher forecasts place students 1.74 percentile points above their realized achievement on average, with greater overprediction in the bottom half of the baseline forecast distribution (Appendix Table F1). Descriptively, treated teachers report lower mean forecasts across the bottom seven deciles and higher forecasts at the upper end (Figure 7). This suggests that report access may change forecasts differently across the distribution, even though the pooled effects on signed error and forecast level are not statistically distinguishable from zero.

Because treatment also affected achievement, lower forecast errors could reflect scores moving toward teachers' forecasts rather than changes in the forecasts themselves. Appendix F examines this possibility. Adjusting treated students' scores to undo estimated achievement effects within baseline forecast deciles leaves the accuracy effect nearly unchanged. We also test whether treatment moved scores toward teachers' forecasts and find no evidence of that pattern. Finally, teacher forecasts move closer to the fixed pre-treatment machine forecasts, a comparison unaffected by changes in realized scores. The persistence of the accuracy gain after these adjustments, together with the small estimated change in average forecast levels, is consistent with teachers revising individual students' forecasts upward and downward in directions that reduce errors without substantially changing the overall average forecast.

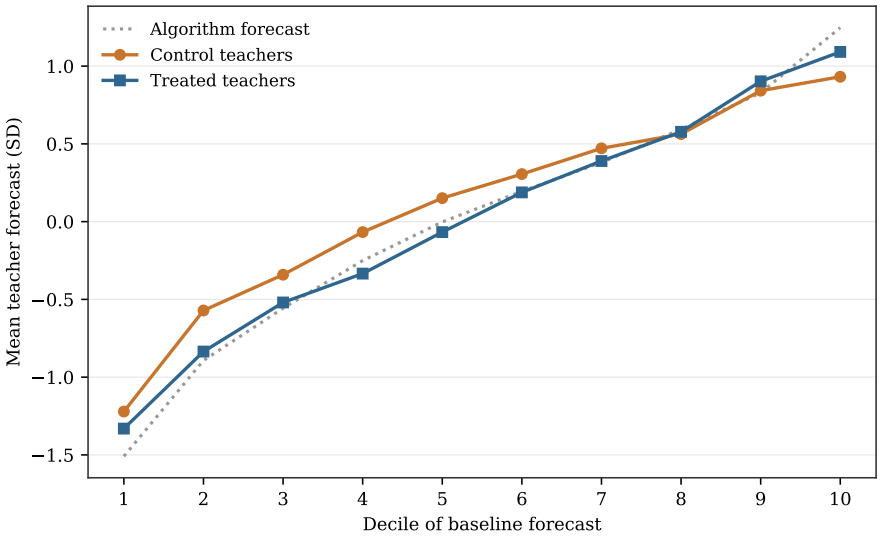


Figure 7: Level of the teacher's forecast, by baseline decile. *Notes:* Completed responses, waves 2–5, for students with a pre-treatment baseline forecast for their own assessment. Each point is a mean within one decile of that forecast. The dotted line is the algorithm's own forecast. Unlike signed error, the level of the forecast does not involve the realized score, which the treatment itself moved.

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 examine teacher effort in three ways. First, we examine whether treatment caused teachers to be more active in monitoring students, as measured by the Learning Progress Charts (LPCs). Second, we estimate changes in self-reported teacher effort, 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

We measure teachers' documentation of student difficulties using cumulative flags on the Learning Progress 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. We do not detect a statistically significant change in behavior flags, which measure teachers' documentation of behavioral concerns rather than student behavior directly.

Table 7 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 these estimates.

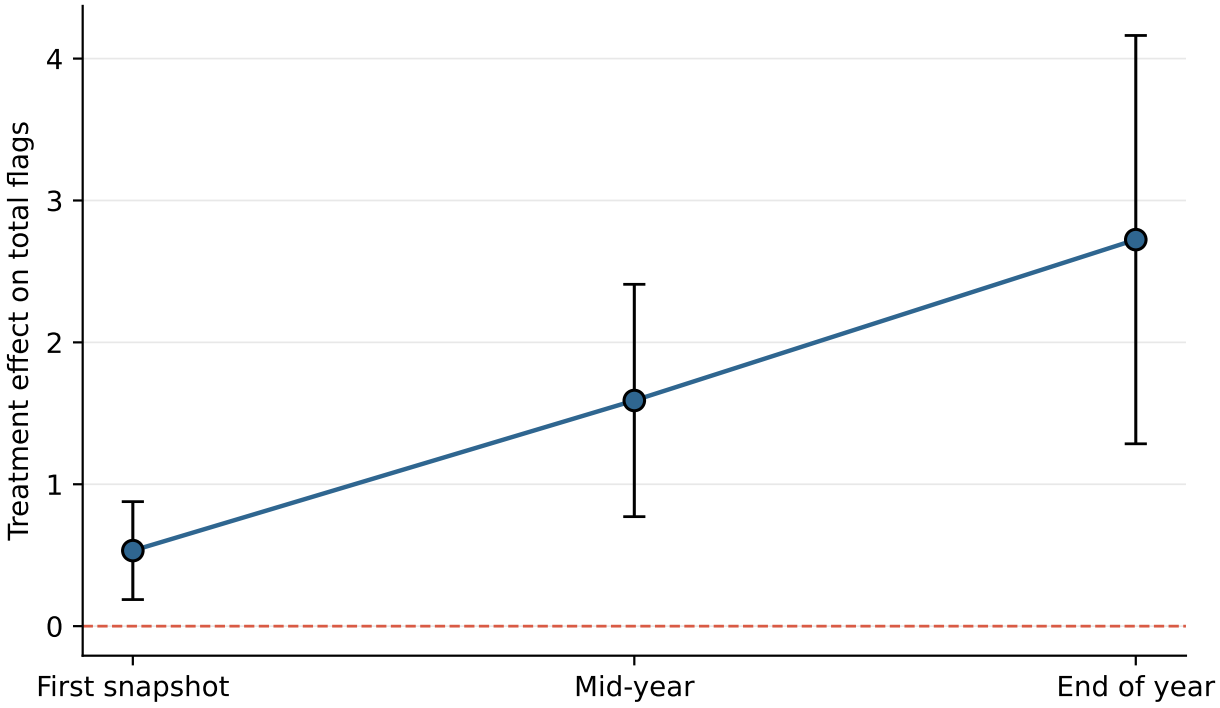


Figure 8: Treatment effects on teacher monitoring, by snapshot. *Notes:* 95% confidence intervals; SEs clustered at campus-grade.

Table 7: 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. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

The reports increased teachers' documentation of struggling students. To examine whether report access also changed teachers' priorities, we asked them to identify the five students they were most focused on helping.

5.2.2 Teacher Reports of Focus Students

The survey asked teachers, "Which five students are you focusing most on helping right now?" Teachers selected students from their own roster, providing a direct measure of their stated priorities for support.

Panel B of Table 8 reports a treatment effect on focus-student selection that is 11.0

percentage points larger for students predicted below proficiency than for students predicted at or above proficiency (SE 5.3 percentage points; $p = .036$). The adjusted treatment effects are +7.7 percentage points for the below-proficiency group and −3.3 points for the other group. Raw focus-selection rates are 38.4% in treated classrooms and 29.3% in control classrooms for students predicted below proficiency, compared with 8.5% and 10.8% for other students. Since teachers name a roughly fixed number of students, this is a composition effect rather than an increase in attention overall: treatment reallocates reported focus toward students predicted below proficiency.

Panel C of Table 8 compares the effect of report access on focus-student selection for students on and off the near-cutoff list. The estimated difference is −1.7 percentage points (SE 3.8 percentage points; $p = 0.65$), so we do not detect a differential treatment effect.

Table 8: Teacher Reports of Focus Students

	Estimate	SE	p
Panel A. Any focus selection			
Treatment	−0.0061	(0.0093)	0.509
Control mean	0.161		
Observations	5,473		
Panel B. By predicted proficiency			
Treatment	−0.0330**	(0.0152)	0.030
Below proficiency	0.2094***	(0.0272)	0.000
Treatment × below proficiency	0.1101**	(0.0526)	0.036
Control mean, below / not below	0.293 / 0.108		
Observations	4,547		
Panel C. By the near-cutoff list			
Treatment	−0.0012	(0.0215)	0.956
On the near-cutoff list	0.0521***	(0.0192)	0.007
Treatment × on the list	−0.0172	(0.0379)	0.649
Control mean, on / not on list	0.171 / 0.125		
Observations	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-wave-student slot, waves 2–5. All models include campus, grade, and wave fixed effects, with standard errors clustered at campus-grade. Control means are the raw share of control-classroom slots named as a focus student, overall and within each moderator group. Panels B and C are smaller because each moderator is defined only for students with a baseline forecast. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Treated teachers make more accurate forecasts, record more LPC flags, and are more likely to name students predicted below proficiency as focus students. We next examine whether report access changes teachers' reported time use.

5.2.3 Teacher Time Use

Time-use outcomes are measured at the teacher-wave level. Each regression controls for the corresponding baseline outcome and its missing indicator, plus campus, grade, and wave fixed effects [following McKenzie, 2012]. Figure 9 shows time-share effects in percentage points and other outcomes as percentages of their control means. Appendix Table D1 reports the estimates in their original units and the control means.

Report access reduces assessment-record consultations by 1.29 and intervention-planning time by 7.56 minutes per week. The estimated reduction in master academic record consultations is 1.64, but is not statistically distinguishable from zero ($p = .106$). Within whole-class instruction, the share of time devoted to new material rises by 9.1 percentage points, while the share devoted to reviewing old material falls by 8.4 points.

Homeroom trackers show no statistically detectable change in recorded lesson coverage: the estimated effect is -0.79 lessons (SE 2.43), compared with a control mean of 70.84 lessons per teacher-subject. The increase in the reported share of whole-class time devoted to new material, together with the absence of a detectable change in recorded lesson coverage, is consistent with teachers teaching the same material more slowly.

The reports also reduced the time teachers spent planning interventions by 7.56 minutes per week. If teachers adjusted whole-class instruction to support struggling students, they may have perceived less need to plan individualized interventions for those students. The reduction in assessment-record consultations is also consistent with this interpretation.

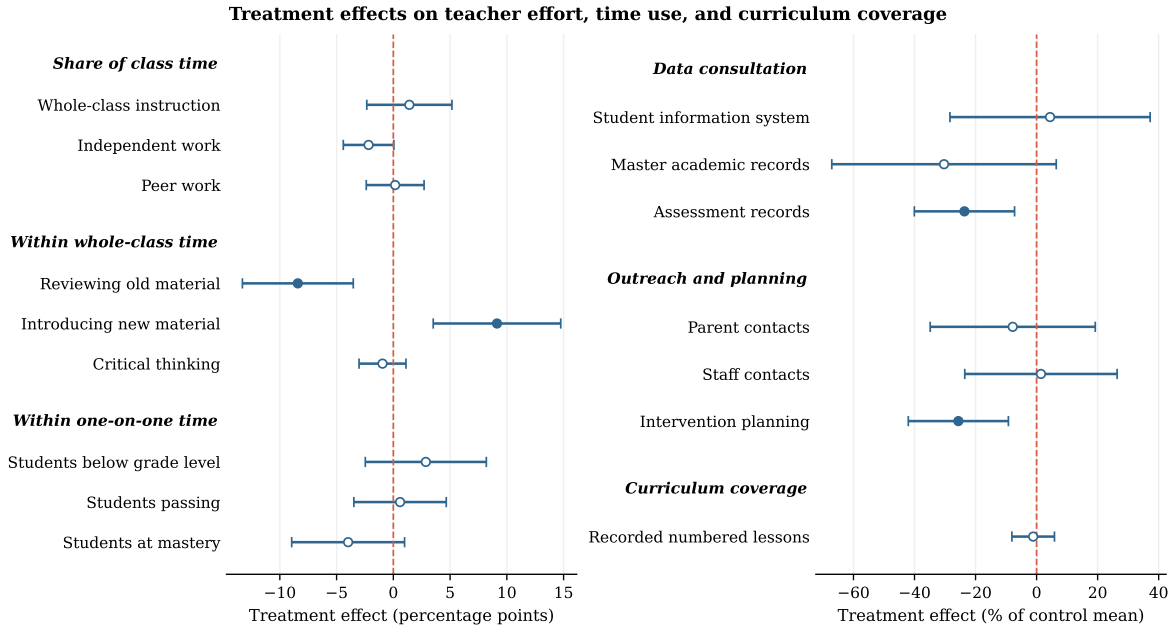


Figure 9: Treatment effects on teacher effort, time use, and curriculum coverage. *Notes:* The left panel shows time-share effects in percentage points. In the right panel, estimates, standard errors, and confidence-interval endpoints are divided by the corresponding control mean in Appendix Table D1 and multiplied by 100, treating these means as fixed scaling factors. Bars are 95% confidence intervals with SEs clustered at campus-grade; filled markers indicate $p < .05$. See Appendix Table D1 for model and sample details.

5.3 Effect on Student Achievement

We next estimate the effects of report access on student achievement. Table 9 reports treatment effects on student achievement. Panel A reports results for normalized test scores. Each assessment score is standardized using the control-group mean and standard deviation within grade. For the pooled outcome, we average each student's available standardized assessment scores. Panel B reports results for test 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. (See table 2 for assessment labels and grade coverage.) 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.

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 available standardized scores for each student. 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 score with a missing indicator and include campus and grade fixed effects. Standard errors clustered at campus-grade in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

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. Panel B reports linear probability estimates.

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.

5.3.1 Heterogeneity

The focus-student responses indicate that teachers shifted their reported attention toward students predicted below proficiency. One hypothesis is that this shift improved achievement among students predicted below proficiency at the expense of other students. Table 10 provides evidence on this hypothesis using two subgroup comparisons. 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.[5]

Panel A compares students on the near-cutoff lists with other students. The estimated effects are similar across the two groups, and we do not reject equality for either normalized scores ($p = .822$) or proficiency ($p = .979$).

Panel B groups students by their lowest baseline predicted proficiency band across assessments. On the normalized score every band is negative. Surprisingly, the largest loss falls on students predicted to be *below* proficiency. Students predicted proficient or above also lose ground. For students predicted to be approaching proficiency, the estimated effect is −0.063 standard deviations (SE 0.059). On proficiency the same ordering appears, but none of the subgroup estimates is statistically significant at the 5% level.

Treated teachers were more likely than control teachers to name students predicted below proficiency as focus students and flagged these students more often on monitoring charts. Yet the estimated achievement effect for this group was negative.

[5]Heterogeneity by student characteristics is reported in Appendix Table C2.

Table 10: Heterogeneity in the achievement effect

	Normalized score			Proficiency		
Subgroup	ITT	Control mean	N	ITT	Control mean	N
Panel A. On the near-cutoff list						
On the list	−0.088	-0.108	1,525	−0.018	0.659	1,525
	(0.066)			(0.028)		
Not on the list	−0.105*	0.066	1,484	−0.019	0.731	1,484
	(0.054)			(0.020)		
Equality (p)	0.822			0.979		
Panel B. Baseline predicted proficiency band						
Below	−0.226**	-0.912	497	−0.070	0.285	497
	(0.096)			(0.044)		
Approaching	−0.063	-0.234	638	−0.005	0.641	638
	(0.059)			(0.038)		
Proficient or above	−0.148***	0.556	1,434	−0.026	0.917	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. Control means are raw means among control students in the subgroup. Panel A uses the near-cutoff rule from the report generator, applied to both the treatment and control groups; 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. 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 examine 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 group, 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 examines how treatment effects vary across the achievement distribution using two approaches: separate regressions within baseline forecast deciles in panel (a), and quantile regressions in panel (b). Both yield broadly negative estimates across the distribution, with the largest estimated loss in each panel occurring in the lower deciles.

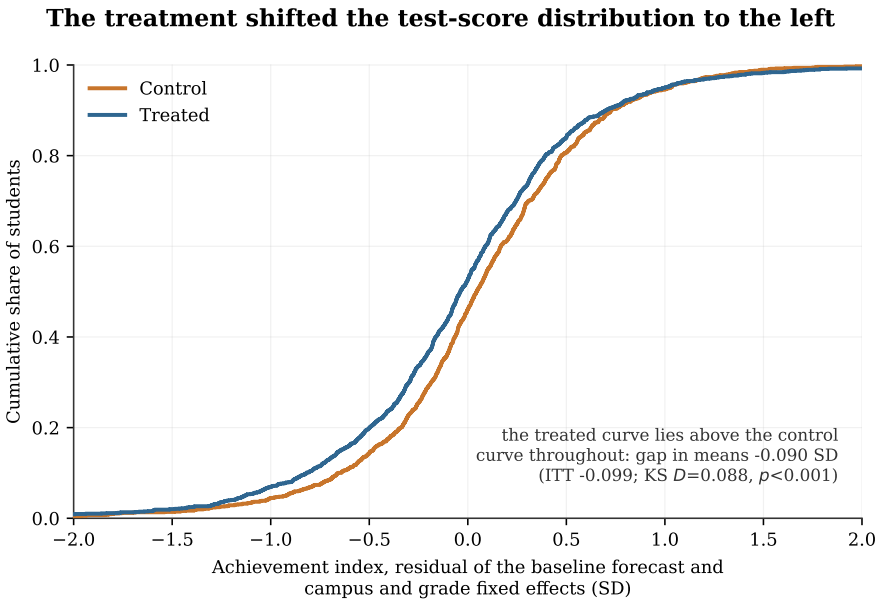


Figure 10: The achievement distribution by treatment status. *Notes:* Empirical cumulative distributions of the normalized achievement index, after residualising on the baseline forecast, its missing indicator, and campus and grade fixed effects. Units are control-group standard deviations. The gap in 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.

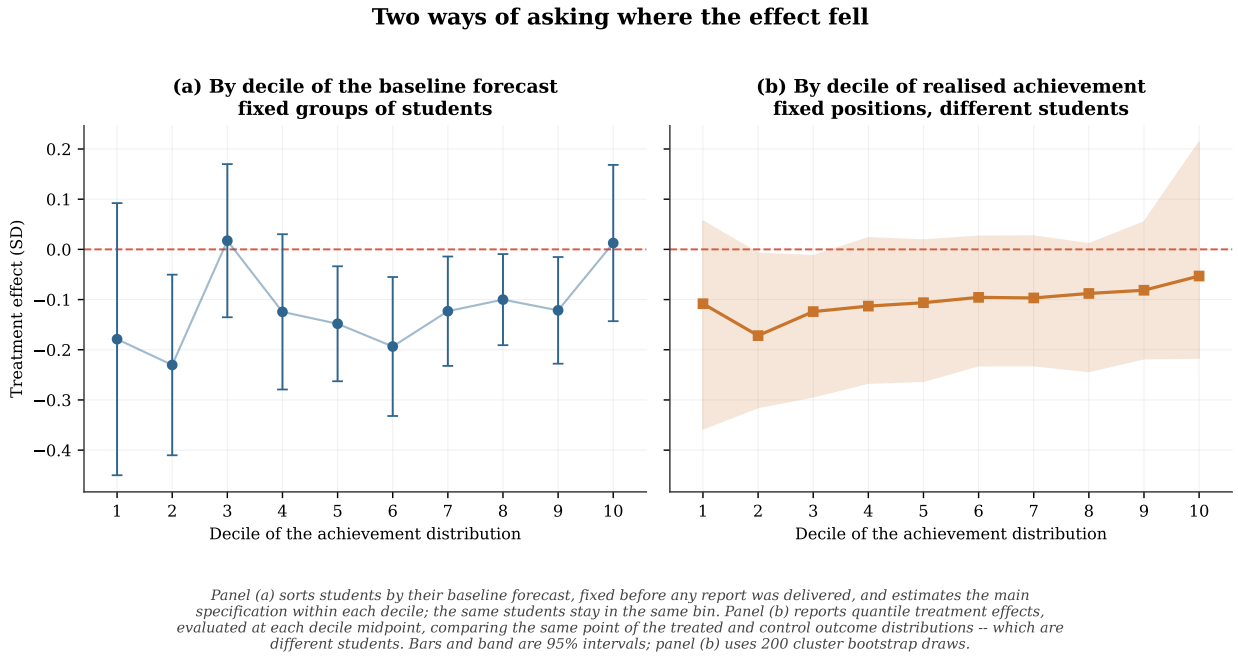


Figure 11: Where the effect fell: fixed groups against fixed positions. *Notes:* Both panels are indexed by decile of the achievement distribution. Panel (a) estimates equation (4) within each decile of the baseline machine forecast, restricted to the 2,652 students who have one; bars are 95% confidence 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.

5.4 Reported Use of the Forecasts

Twenty-four of the 54 treated teachers answered the end-of-year question about report use: 13 reported using the forecasts, and 11 reported not using them. Table 11 compares pooled achievement among students of these teachers with the full control group. Panel A groups teachers by reported use, while Panel B groups them by reported usefulness.

Table 11: Pooled achievement by reported use of the forecasts

Treated teachers who...	Adjusted difference vs. control	p	Teachers	Students
Panel A. Reported use				
used the reports	−0.123	0.028	13	375
	(0.056)			
did not use them	−0.058	0.438	11	323
	(0.075)			
Panel B. Reported usefulness				
moderately or very useful	−0.130	0.009	9	274
	(0.050)			
slightly useful	−0.084	0.340	10	285
	(0.088)			
not useful	0.133	0.335	4	106
	(0.138)			

The outcome is the pooled normalized achievement index. Each row estimates equation (4) on the full control group (1,428 students, control mean −0.019) plus the students of the indicated treated homeroom teachers. Models control for the baseline predicted score and its missing indicator and include campus and grade fixed effects. The rows share a control group. Use and usefulness are self-reported after treatment and are not randomly assigned. These are descriptive comparisons rather than causal effects of use. Panel A includes 24 teachers who answered the use question; Panel B includes 23 teachers who answered the usefulness question. Missing answers are excluded separately for each panel. Students are linked to their homeroom teacher. Standard errors clustered at campus-grade are in parentheses.

Students whose teachers reported using the forecasts have a pooled achievement index 0.123 units lower than controls (SE 0.056, $p = .028$). The corresponding difference for reported nonusers is −0.058 (SE 0.075). The negative achievement estimate is therefore also present among reported users. Because report use and survey response are not randomly assigned, these comparisons do not identify the causal effect of use or establish that effects differ between users and nonusers.

We discuss possible explanations for these achievement effects in Section 6 and turn next to robustness checks.

5.5 Robustness

5.5.1 Attrition and Matching

Differential attrition from the administrative test-score sample could bias the estimated treatment effect. Table 12 checks whether treatment affects which students have observed end-of-year scores: Panel A compares score availability, and Panel B compares baseline

forecasts among tested students. These checks provide no strong evidence of differential attrition.

Table 12: Test-score availability and baseline-forecast comparisons

Panel A. End-of-year score availability

Assessment	Grades	Eligible N	Control (%)	Treatment (%)	Adjusted difference (pp)	p-value
Benchmark reading	1–6	3,068	97.9	97.9	0.16 (0.43)	0.717
State ELA	3–6	2,039	92.5	94.2	1.46 (0.91)	0.111
State math	3–6	2,039	92.5	93.5	0.42 (0.97)	0.664
National reading	1–2	1,029	78.2	81.3	−0.47 (3.57)	0.895
National math	1–2	1,029	78.8	82.2	−0.73 (3.79)	0.848

Panel B. Baseline forecast balance among tested students

Assessment	Adjusted baseline forecast difference	p-value	Tested with forecast (N)
Benchmark reading	0.037 (0.068)	0.584	2,647
State ELA	−0.023 (0.050)	0.650	1,713
State math	0.032 (0.034)	0.348	1,705
National reading	0.137 (0.108)	0.204	691
National math	0.015 (0.226)	0.948	696

Panel A reports raw score-availability rates in the eligible grades and adjusted treatment–control differences, in percentage points, from linear probability models with campus and grade fixed effects. Panel B compares assessment-specific standardized baseline forecasts between treatment and control among students with an observed score and a nonmissing forecast, adjusting for campus and grade fixed effects. Differences are treatment minus control, in within-grade forecast-standardization units. Standard errors in parentheses are clustered at campus-grade. p-values use a normal reference, as in the main tables. Benchmark reading uses 36 clusters, state assessments 24, and national assessments 12.

5.5.2 Survey Nonresponse

Survey outcomes are observed only for responding teachers, and treated teachers responded more often than control teachers. Differential response could therefore affect comparisons of reported practices and forecast accuracy. Appendix E reports Manski worst-case bounds [Manski, 1989] and Lee trimming bounds [Lee, 2009] for unadjusted survey comparisons. The worst-case bounds are wide and include zero. Lee bounds for forecast accuracy also include zero; after allowing for sampling uncertainty, most time-use bounds include zero, although

intervention planning remains negative. We therefore interpret the survey estimates as suggestive evidence of behavioral responses to the reports, subject to selective nonresponse.

5.5.3 Inference with 36 Clusters

Treatment is assigned to 36 campus-grade cells, and cluster-robust asymptotics can overstate precision at that scale. To address this, we consider alternative inference methods. To start, the teacher could be viewed as the fundamental unit of outcome correlation because students share instructional inputs. Teacher-level clustering provides a useful comparison under the assumption that remaining dependence is confined within teachers. School-grade clustering additionally allows dependence across teachers within the same treatment-assignment group and is our preferred method throughout the paper. We also report two methods that deal with small numbers of clusters more directly: wild cluster bootstrap p-values [Cameron et al., 2008] and randomization-inference p-values [Young, 2019]. Table 13 reports these four approaches, and Appendix Table C1 reports assessment-specific results. Implementation details appear in the table notes.

The main results are broadly supported across inference procedures. The pooled achievement effect of -0.099 is statistically significant at the 10% level under teacher clustering ($p = .009$), school-grade clustering ($p = .039$), and the wild cluster bootstrap ($p = .086$); the randomization-inference p-value of .105 lies just above that threshold. Most reported main effects remain significant at the 10% level under at least three of the four approaches, and effects on monitoring, reviewing old material, introducing new material, and weekly intervention planning remain significant at the 5% level under all four. The exceptions are standardized forecast error, whose bootstrap and randomization-inference p-values are .108 and .110, and master-record consultation, which is not statistically significant under any approach.

Table 13: The main results under four approaches to inference

Result	Table	Coef.	p-value			
			Teacher clustering	School-grade clustering	Wild cluster bootstrap	Randomization inference
Beliefs						
Percentile error	6	−2.121	0.026	0.011	0.081	0.051
Standardized error	6	−0.051	0.095	0.044	0.108	0.110
Monitoring						
Total flags, end of year	7	2.724	0.000	0.001	0.011	0.008
Focal students						
Treatment × predicted below proficient	8	0.110	0.019	0.045	0.064	0.079
Teacher effort						
Reviewing old material	D1	−8.414	0.001	0.002	0.008	0.019
Introducing new material	D1	9.133	0.002	0.003	0.015	0.030
Master academic records	D1	−1.645	0.112	0.116	0.141	0.132
Assessment records	D1	−1.290	0.038	0.008	0.044	0.065
Minutes planning interventions	D1	−7.558	0.021	0.005	0.024	0.040
Achievement						
Pooled normalized score	9	−0.099	0.009	0.039	0.086	0.105

Point estimates and samples are held fixed across columns. Teacher and School-grade report CR1 cluster-robust p-values with a t reference using $G - 1$ degrees of freedom at the indicated clustering level; the main tables use a normal reference. Teacher clustering allows dependence within teachers; school-grade clustering also allows dependence across teachers within an assignment group. WCB is a wild cluster bootstrap with the null imposed and Rademacher weights at school-grade (1,999 replications). RI redraws the grade-stratified assignment, treating three of six schools within each grade, over 2,000 draws and recomputes the school-grade-clustered t-statistic in each draw. RI tests the sharp null of no treatment effect on any unit. For the focus-student interaction, WCB imposes a zero interaction coefficient while retaining the treatment main effect; RI rebuilds both treatment and its interaction and tests the sharp global null, rather than a null of zero differential effect alone. Teacher-effort coefficients use percentage points, consultation counts, or weekly planning minutes, matching Appendix Table D1.

6 Discussion

The reports improve teachers' measured forecast accuracy and increase their reported focus on students predicted to struggle. Teachers also report less record consultation and intervention planning, and devote a larger share of whole-class instruction to introducing new material rather than reviewing previously taught material. These responses accompany a decline in pooled achievement, with the largest negative point estimate among students predicted to score below proficiency. Two mechanisms could help explain this pattern: changes in student priorities and instructional pacing, and using the reports as a substititute for

productive diliberation about students. These mechanisms are not mutually exclusive.

6.1 Changes in Student Priorities and Instructional Pacing

Teachers may have adjusted instruction to accommodate students whose forecasts indicated that they would struggle. Figure 7 shows descriptively lower forecasts among treated teachers in the bottom seven baseline forecast deciles and higher forecasts at the upper end. The estimated forecast-level change in the bottom half is negative but imprecise (-0.094, $p = .35$) and is very similar in scale to the actual decline in scores we see. The treatment effect on focus-student selection is approximately 11 percentage points larger for students predicted below proficiency than for other students, and reports increase cumulative monitoring flags by 2.72 per student. These responses are consistent with teachers directing more attention toward students with lower predicted achievement.

The instructional-time estimates show an 8.4 percentage-point reduction in reviewing previously taught material and a 9.1 percentage-point increase in introducing new material. Recorded lesson coverage changes by -0.79 lessons (SE 2.43, $p = .75$), relative to a control mean of 70.84. More time devoted to new material without a detectable increase in recorded coverage is consistent with spending more time teaching each lesson, or in other words a slower pace of teaching. This could reduce time available for review or individualized support.

6.2 The Reports Replaced Productive Deliberation

A second mechanism is that forecasts reduce teachers' perceived need to review student records. Teachers may reasonably view a forecast as a substitute for assembling information to predict a student's final score. However, reviewing the underlying records can also help identify specific skill deficits and inform instructional plans. A forecast identifies students likely to struggle without necessarily explaining which skills need attention or how to address them. If prediction and planning draw on the same review process, replacing that process with a forecast could reduce the information available for choosing instruction.

The reported time-use changes are consistent with this mechanism. Treated teachers consult assessment records 1.29 fewer times per week and spend 7.56 fewer minutes per week planning interventions. Consultation of master academic records also declines, although that estimate is less precise. These changes could reduce individualized support even as teachers report greater attention to struggling students.

7 Conclusion

We study how elementary-school teachers respond to machine-learning forecasts of student achievement in a randomized experiment across six campuses of a charter-school network. Report access improves teachers' measured forecast accuracy and increases their reported focus on students predicted to struggle. Teachers also report less record consultation and intervention planning, and shift whole-class instructional time from review toward new material. Yet pooled achievement declines by 0.099 units of the normalized achievement index, with the largest negative point estimate among students predicted to score below proficiency. The findings are consistent with forecasts changing instructional priorities and substituting for activities that support teaching, although the experiment does not separately identify these mechanisms.

More accurate predictions and greater reported attention to struggling students are therefore insufficient to establish an educational tool's value. Evaluations should measure how information changes professional behavior and the outcomes of the people those decisions affect. Our results motivate testing whether forecasts paired with guidance on specific learning difficulties and instructional responses produce different outcomes. Such guidance may help translate predictions into effective support, but its effects remain to be evaluated.

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.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

Sule Alan, Seda Ertac, and Ipek Mumcu. Gender Stereotypes in the Classroom and Effects on Achievement. *The Review of Economics and Statistics*, 100(5):876–890, 2018. doi: 10.1162/rest_a_00756.

Sule Alan, Enes Duysak, Elif Kubilay, and Ipek Mumcu. Social Exclusion and Ethnic Segregation in Schools: The Role of Teachers' Ethnic Prejudice. *The Review of Economics and Statistics*, 105(5):1039–1054, 2023. doi: 10.1162/rest_a_01111.

Sule Alan, Michela Carlana, and Marinella Leone. Inclusive Teaching: Spotting Social Isolation in the Classroom. *American Economic Journal: Economic Policy*, 2025. Forthcoming.

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. *The Review of Economic Studies*, 93(4):2250–2283, 2026. URL https://www.nber.org/papers/w31747.

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.

Roland Bénabou and Jean Tirole. Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics*, 117(3):871–915, 2002. doi: 10.1162/003355302760193913.

Roland Bénabou and Jean Tirole. Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, 70(3):489–520, 2003. doi: 10.1111/1467-937X.00253.

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.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Christina Brown, Tahir Andrabi, Asim Ijaz Khwaja, and Isabel MacDonald. Back to Basics: Experimental Evidence on Targeted Instruction. Working paper, 2026.

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.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

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.

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.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774, 2011. doi: 10.1257/aer.101.5.1739.

Greg J. Duncan, Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Crista Japel. School Readiness and Later Achievement. *Developmental Psychology*, 43(6):1428–1446, 2007. doi: 10.1037/0012-1649.43.6.1428.

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.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Bengt Holmström and Paul Milgrom. Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization*, 7(special issue):24–52, 1991. doi: 10.1093/jleo/7.special_issue.24.

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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

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 S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009. doi: 10.1111/j.1467-937X.2009.00536.x.

Charles F. Manski. Anatomy of the selection problem. *Journal of Human Resources*, 24(3):343–360, 1989. doi: 10.2307/145818.

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.

Derek Neal and Diane Whitmore Schanzenbach. Left Behind by Design: Proficiency Counts and Test-Based Accountability. *The Review of Economics and Statistics*, 92(2):263–283, 2010. doi: 10.1162/rest.2010.12318.

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.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648, 2018.

Sean F. Reardon. The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations. In Greg J. Duncan and Richard J. Murnane, editors, *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, pages 91–116. Russell Sage Foundation, New York, 2011.

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. *American Economic Journal: Economic Policy*, 16(4):382–414, 2024. doi: 10.1257/pol.20220620.

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.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Alwyn Young. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *The Quarterly Journal of Economics*, 134 (2):557–598, 2019. doi: 10.1093/qje/qjy029.

Andrey Zakharov and Martin Carnoy. Does Teaching to the Test Improve Student Learning? *International Journal of Educational Development*, 84:102422, 2021. doi: 10.1016/j.ijedudev.2021.102422.

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 forecast timing. We train assessment-specific models on historical student records from the same network, using realized end-of-year scores as outcomes. The returning-student input files cover outcome years 2021–2025 for Acadience and RISE and 2022–2025 for SAT. The input files for the version without prior-year histories cover 2020–2025 for Acadience and 2021–2025 for RISE and SAT. Training observations are student-year records; a student can appear in more than one historical year.

For each report update, we reconstruct historical gradebook features using the corresponding point in the school year. At baseline, these features summarize the preceding year. For subsequent updates, they include records from completed school terms in the outcome year, alongside prior-year summaries where available. The training code excludes outcome-score fields from the predictors and removes fields identified as future-term information. The gradebook-free version also removes annual and snapshot summary fields. Thus the intended predictor set corresponds to the information available at the report date.

Predictors and feature preparation. The predictors include prior assessment scores and proficiency categories; reading and mathematics placement and curriculum progress; gradebook performance, score variation, and assignment-completion measures; attendance; student demographic and program indicators; and grade. Later versions also use available current-year assessment and gradebook information. Student identifiers and names, the configured subject-teacher identifier fields, and outcome columns are excluded from the model inputs.

To summarize curriculum placement, the preparation code ranks curriculum levels by their mean historical outcome score for the corresponding assessment. A progress measure adds the lesson number divided by 180 to this rank. Missing ranks are filled with a median or a grade-specific historical average, depending on the student-history version. Selected demographic and program fields are filled with zero, and missing campus or teacher labels with an unknown category during preparation. Remaining numeric missing values are handled within the learner pipelines: the penalized regressions and random forest use median imputation. The later saved training notebook allows LightGBM and CatBoost to retain numeric missing values, while the earlier scripts apply median imputation to these inputs as well. Categorical predictors are one-hot encoded, and the penalized regressions use standardized inputs.

Assessment and student-history versions. We estimate models separately for Acadience reading, RISE reading/language arts, RISE mathematics, SAT reading, and SAT

mathematics. We pool grades within each assessment and include grade as a categorical predictor. Returning-student models cover grades 1–6 for Acadience, 3–6 for RISE, and 1–2 for SAT. The version serving new entrants also allows historical kindergarten observations for Acadience and SAT; the experimental sample remains grades 1–6. Observations with missing target scores are excluded from the corresponding model.

Each assessment has three versions (Table A1). The first is trained on returning students whose records link to the gradebook features. The second is trained on all available returning students, with gradebook summaries, prior-year gradebook summaries, and gradebook changes removed from its predictor set. The third serves students new to the network and removes prior-year and change predictors from the broader historical input files. Its training sample includes records that also appear in the returning-student files; it is not restricted to historical new entrants.

Table A1: Forecast model versions

Version	Students served	Training sample	Predictors
Returning, with gradebook	Returning students with usable linked gradebook features	Returning students with linked gradebooks	Prior scores and proficiency bands, group placement and curriculum progress, gradebook performance and completion, attendance, program indicators, grade; current-year records as they arrive
Returning, without gradebook	Returning students without usable linked gradebook features	Returning students, gradebook summaries removed	Same as above, without gradebook summaries
New to the network	Students new to the network	Historical records without requiring prior-year histories	Placement, student characteristics, grade; available current-year assessments and gradebook features; no prior-year history

Each version is estimated separately for the five assessments (benchmark reading, state reading/language arts, state mathematics, nationally normed reading, nationally normed mathematics), pooling grades within an assessment with grade as a categorical predictor. All versions use the same stacked ensemble, described below, and are refreshed as new records accumulate during the year.

Learners, tuning, and stacking. The ensemble combines LASSO [Tibshirani, 1996], ridge regression [Hoerl and Kennard, 1970], a random forest [Breiman, 2001], LightGBM

[Ke et al., 2017], and CatBoost [Prokhorenkova et al., 2018], using a ridge meta-learner [Wolpert, 1992]. For each assessment and student-history version, we tune each base learner with 20 Optuna trials [Akiba et al., 2019]. The objective is mean root mean squared error (RMSE) under three-fold cross-validation on the eligible pooled data. The search varies the penalty for the linear models, the number and depth of trees for the random forest, the number of boosting iterations and learning rate for LightGBM, and those parameters plus tree depth for CatBoost.

After tuning, the code randomly splits the eligible student-year records into an 80% fitting sample and a 20% evaluation sample, using a fixed seed of 42. Five-fold cross-validation within the fitting sample generates base-learner predictions for fitting the ridge meta-learner. The base learners are then fitted on the full 80% sample. The saved ensemble is this fitted model; the code does not subsequently refit it on the evaluation observations.

Evaluation and interpretation. Table A2 reports saved evaluation RMSEs from training runs logged as report updates 1 and 4, corresponding to survey waves 2 and 5. RMSE is measured in each assessment's score points, so its magnitude is not directly comparable across assessments. These are historical training diagnostics, distinct from the teacher–algorithm comparisons using experimental students. Changes across the two saved runs need not reflect only the addition of new predictors, since the eligible records and data preparation can also differ.

The evaluation observations are excluded from the final ensemble fit, but not from every preceding step. Hyperparameter tuning uses the eligible data before the split, and curriculum-level rankings and selected imputations are also constructed before that split. In particular, the curriculum rankings use historical outcome scores. Moreover, the split is by student-year record, without grouping by student or withholding an entire outcome year. The diagnostics therefore do not constitute a fully independent evaluation of the complete forecasting procedure or a test on a wholly new cohort.

Table A2: Saved historical forecast evaluation diagnostics

	Returning students		No prior-year
Assessment	Linked gradebooks	Gradebooks excluded	history
Report update 1 (survey wave 2)			
Acadience reading	62.92	62.03	74.62
RISE ELA	42.86	42.63	54.96
RISE mathematics	19.68	19.40	27.32
SAT reading	23.48	25.06	29.83
SAT mathematics	24.71	24.99	30.07
Report update 4 (survey wave 5)			
Acadience reading	52.42	54.26	60.70
RISE ELA	41.21	41.49	49.84
RISE mathematics	19.36	19.61	24.66
SAT reading	21.26	23.40	25.36
SAT mathematics	22.65	24.81	29.09

Entries are root mean squared errors in assessment score points, transcribed from saved training-notebook outputs for runs logged as updates 1 and 4. Each assessment and student-history version has its own random 20% evaluation sample. The no-prior-year-history version serves new entrants but is trained on broader historical records with history predictors removed. Tuning and outcome-based feature preparation precede the split, and records are not grouped by student or year; these values are diagnostics rather than fully independent validation estimates.

Applying the saved models. For each update, the prediction code loads the corresponding saved assessment and student-history model. Returning students are routed between the linked-gradebook and gradebook-free versions according to the availability of a designated mathematics gradebook summary. Students classified as new entrants use the no-prior-year-history version. The code restricts predictions to the assessment's eligible grades and arranges inputs to match the feature names stored with the fitted model. Absent inputs are supplied as missing values for the model's preprocessing steps. The resulting point predictions feed the teacher reports.

B Grade Fixed Effects

Treatment was randomized within grade — three of six campuses in each grade — but the pre-analysis plan's illustrative equation included campus fixed effects and inadvertently omitted grade fixed effects. We apply its common estimating framework with the baseline controls, moderators, and timing effects appropriate to each outcome. Table B1 compares these models with and without grade fixed effects for the same headline results as Table 13.

Table B1: Main results with and without grade fixed effects

Result	With grade fixed effects	Without grade fixed effects
Beliefs		
Percentile error	-2.121^{***}	-1.967^{**}
	(0.780)	(0.915)
Standardized error	-0.051^{**}	-0.041
	(0.024)	(0.030)
Monitoring		
Total flags, end of year	2.724^{***}	2.643^{***}
	(0.734)	(0.867)
Focal students		
Treatment × predicted below proficient	0.110^{**}	0.107^{**}
	(0.053)	(0.051)
Teacher effort		
Reviewing old material	-8.414^{***}	-7.013^{*}
	(2.492)	(3.647)
Introducing new material	9.133^{***}	7.965^{**}
	(2.866)	(3.728)
Master academic records	-1.645	-1.963^{*}
	(1.017)	(1.099)
Assessment records	-1.290^{***}	-1.568^{**}
	(0.457)	(0.613)
Minutes planning interventions	-7.558^{***}	-8.523^{***}
	(2.466)	(2.511)
Achievement		
Pooled normalized score	-0.099^{**}	-0.053
	(0.046)	(0.118)

The rows match the headline results in Table 13. Each pair uses the same observations, campus fixed effects, and outcome-specific baseline controls, moderators, and timing effects; only additive grade fixed effects differ. Standard errors in parentheses are CR1 clustered at campus-grade. Belief errors use percentile points or standard deviations; monitoring uses flag counts; the focal-student row is the treatment × predicted-below-proficient interaction in probability units. Teacher-effort outcomes use percentage points, weekly consultation counts, or weekly planning minutes. Achievement is in control-group standard deviations. Samples are 5,101 predictions for beliefs, 3,068 students for monitoring, 4,547 student-wave observations for focal students, 160–166 teacher-wave observations for effort, and 3,009 students for achievement. Stars use normal-reference p-values, as in the main results tables: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

All ten estimates retain the same sign when grade fixed effects are omitted, but most become less precise. For pooled achievement, the estimate changes from −0.099 (SE 0.046) to −0.053 (SE 0.118). Our preferred models retain grade fixed effects to reflect the randomization strata.

C Additional Inference and Heterogeneity

Table 9 in the main text reports achievement effects for each assessment. Table C1 supplements those estimates with randomization-inference and wild-cluster-bootstrap p-values. These checks are particularly relevant for the nationally normed tests, which cover grades 1–2 and only 12 randomization cells. The state tests cover grades 3–6 (24 cells), while benchmark reading spans all experimental grades. Assessment-specific results are exploratory and are not adjusted for multiple testing; the pooled achievement index is the confirmatory outcome.

Table C1: Design-based inference by assessment

		p-value			
Outcome	ITT	CRVE	WCB	RI	N
Normalized score (pooled)	−0.099	0.039	0.086	0.105	3,009
Benchmark reading	−0.053	0.316	0.441	0.432	3,003
State ELA	−0.001	0.988	0.992	0.991	1,910
State math	−0.108	0.016	0.049	0.031	1,902
National reading	−0.267	$< .001$	0.127	0.108	826
National math	−0.082	0.559	0.770	0.818	834

Same procedures as Table 13; the pooled row repeats its achievement row. CRVE uses a t reference with $G - 1$ degrees of freedom, so it is slightly larger than the normal-reference p-values in Table 9. WCB is a wild cluster bootstrap with the null imposed and Rademacher weights (1,999 replications); RI replays the 3-of-6 grade-stratified lottery 2,000 times, permuting the studentized statistic. 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 C2 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.

Table C2: Heterogeneous Treatment Effects on the Pooled Achievement Index

	Subgroup effect				
	Coef.	SE	Control mean	N	Diff. p
Panel A. By predicted proficiency level (baseline forecast)					
Below	−0.226	0.096	−0.912	497	
Approaching	−0.063	0.059	−0.234	638	
Proficient or above	−0.148	0.045	0.556	1,434	
Equality of the three effects					0.099
Panel B. By student characteristic					
Free/reduced lunch	−0.180	0.069	−0.219	901	0.111
not free/reduced lunch	−0.065	0.049	0.073	2,108	
English learner	−0.146	0.083	−0.292	677	0.533
not English learner	−0.091	0.049	0.068	2,332	
Special education	−0.016	0.083	−0.634	406	0.246
not special education	−0.106	0.045	0.077	2,603	
Female	−0.112	0.049	−0.007	1,435	0.739
not female	−0.126	0.051	0.054	1,367	
White	−0.147	0.049	−0.045	2,000	0.029
not White	−0.042	0.051	0.192	802	
Hispanic/Latino	−0.178	0.070	−0.277	978	0.120
not Hispanic/Latino	−0.079	0.041	0.179	1,824	
Black	−0.124	0.083	−0.095	217	0.931
not Black	−0.118	0.045	0.034	2,585	
Asian	−0.011	0.051	0.365	580	0.020
not Asian	−0.144	0.050	−0.064	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. Control means are raw means of the index among control students in the subgroup. 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 C3 and C4 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 estimates 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 C3: 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 C2 carries the precision. "—" means the subgroup is empty or too small to estimate for that assessment.

Table C4: 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 C2 carries the precision. "—" means the subgroup is empty or too small to estimate for that assessment.

Proficiency by Assessment

Table C5 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 C5: 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Belief Accuracy with Teacher Fixed Effects

Table C6 re-estimates the accuracy comparison with teacher fixed effects. Because no teacher changes treatment status, a teacher fixed effect absorbs the treatment indicator, so Panel A is a difference-in-differences identified only from the 15 teachers observed on both sides of the delivery date (5 treated and 10 control). None of the six estimates is distinguishable from zero. Panel B adds controls cumulatively to the pre-treatment wave alone. The estimate is small and insignificant until campus and grade fixed effects are both included, at which point it rests on 14 clusters.

Table C6: Belief accuracy with teacher fixed effects

	(1) Pooled	(2) Teacher FE	(3) Teacher FE + test×wave
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 wave-2 to wave-5 response, or a baseline-wave response submitted on or after the day the reports were sent. Because no teacher changes treatment status, 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. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

D Teacher Effort, Time Use, and Curriculum Coverage

Table D1: Treatment effects on teacher effort, time use, and curriculum coverage

Outcome	ITT (SE)	Control mean	p	N
Panel A. Share of class time (percentage points)				
Whole-class instruction	1.41 (1.91)	70.12	0.460	168
Independent work	−2.18* (1.14)	19.59	0.055	168
Peer work	0.16 (1.30)	10.29	0.900	168
Panel B. Within whole-class time (percentage points)				
Reviewing old material	−8.41*** (2.49)	30.81	0.001	166
Introducing new material	9.13*** (2.87)	51.52	0.001	166
Critical thinking	−0.95 (1.05)	17.67	0.366	166
Panel C. Within one-on-one time (percentage points)				
Students below grade level	2.86 (2.72)	47.56	0.293	162
Students passing	0.59 (2.08)	33.00	0.775	162
Students at mastery	−3.98 (2.54)	19.44	0.116	162
Panel D. Data consultation (counts)				
Student information system	0.80 (3.02)	18.07	0.792	160
Master academic records	−1.64 (1.02)	5.42	0.106	161
Assessment records	−1.29*** (0.46)	5.46	0.005	160
Panel E. Outreach (counts) and planning (minutes)				
Parent contacts	−0.41 (0.73)	5.31	0.571	161
Staff contacts	0.11 (0.95)	7.42	0.911	161
Minutes planning interventions	−7.56*** (2.47)	29.48	0.002	161
Panel F. Recorded curriculum coverage (lessons)				
Recorded numbered lessons	−0.79 (2.43)	70.84	0.747	265

Panels A–E use teacher-wave observations. Outcomes and their corresponding baseline controls enter in original units. Each survey model includes the baseline-missing indicator and campus, grade, and wave fixed effects. Standard errors in parentheses are clustered at campus-grade; p-values for Panels A–E use the normal reference distribution. Panels A–C are percentage points and are compositional within each panel. Panel D and contact outcomes in Panel E are weekly counts; intervention planning is minutes per week. Control means use post-treatment teacher-waves. Panel F counts distinct numbered lessons or chapters recorded during the school year; repeated entries count once. Observations are teacher-subject records with at least ten numbered-entry dates, covering science and history in grades 1–6 and math in grades 1, 3, 5, and 6. Formula-imported columns are excluded. The coverage model includes campus and grade-by-subject fixed effects, without a baseline coverage control; its p-value uses a t_{35} reference. The Panel F control mean is the school-year mean per teacher-subject. Counts are provisional reconstructions of tracker entries. These outcome-specific estimates are exploratory and are not adjusted for multiple testing. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

E Sensitivity to Survey Nonresponse

Survey outcomes are observed only for teachers who respond, and treated teachers responded more often than control teachers. Random assignment protects the comparison of *invited* teachers, but not the comparison of responders. This appendix reports how far the survey results could move if nonresponse were related to treatment. The exercise is a sensitivity analysis, not a correction: it does not replace the adjusted estimates in the main text. It applies only to the survey outcomes, not to the administrative achievement and monitoring results.

Frames. The bounds use the same units as the regressions in the main text, so response is measured at the level at which outcomes go missing. For teacher effort, the frame is the 417 classroom-wave invitations in waves 2–5; the response rate is 35% for control and 44–45% for treated classrooms, varying slightly by outcome. For prediction accuracy, the frame is the 14,801 teacher–student–assessment–wave opportunities with an observed end-of-year score and a baseline machine forecast, of which 5,101 have a usable prediction (33% control, 36% treated). These rates are lower than the 70% of teachers who answered at least one survey because a teacher who answered once is missing in every other wave.

Bounds. *Manski* worst-case bounds [Manski, 1989] fill each missing outcome with the smallest or largest value it could take, separately in each group, and require no assumption about who responds. *Lee* trimming bounds [Lee, 2009] assume that treatment moves teachers into responding but never out of it (monotone selection). They remove the excess responders from the higher-response group, from either the top or the bottom of its outcome distribution, and report the resulting range for the effect on teachers who would respond under either assignment. Outcomes without a verified upper limit, such as counts of record consultations and standardized errors, have unbounded Manski intervals, so no finite value is shown.

Reading the tables. Three features of Tables E1 and E2 limit how the results should be read. First, the bounds are computed on unadjusted means and therefore do not bracket the adjusted regression coefficients; the first column is shown for reference only. Second, Lee bounds describe always-responders, whereas Manski bounds describe the full eligible frame, so their endpoints need not nest. Third, for accuracy we hold fixed the reference distributions used to rank and standardize forecasts; we do not re-rank the full population.

Results. Worst-case bounds are wide, as they must be when fewer than half of invited observations respond. Every Manski interval that can be computed includes zero, and sev-

eral are unbounded. The Lee point bounds are more informative. For effort, the estimates for independent work, reviewing old material, introducing new material, assessment-record consultations and intervention planning keep their signs. Their approximate 95% envelopes, which add sampling uncertainty, include zero except for intervention planning, whose envelope is $[-23.61, -0.97]$ minutes per week. For prediction accuracy, both the percentile and standardized measures have Lee bounds that include zero. Selective response could therefore account for the measured improvement in accuracy, although these results do not show that it did. Allowing the direction of selection to differ by wave and assessment gives the same pattern. A Lee bound that excludes zero should not be read as a statistically significant estimate that is robust to nonresponse.

Table E1: Sensitivity of teacher effort and time use to survey nonresponse

	Paper	Respondent	Lee bounds		Manski	Response
Outcome	estimate	difference	Point	Approx. 95%	bounds	(%, C / T)
Panel A. Share of class time (percentage points)						
Whole-class instruction	1.41	2.25	[−0.65, 7.36]	[−7.16, 13.81]	[−56.96, 62.81]	35 / 45
Independent work	−2.18	−3.04	[−6.34, −1.07]	[−10.77, 2.95]	[−64.15, 55.62]	35 / 45
Peer work	0.16	0.79	[−2.12, 2.73]	[−5.67, 6.86]	[−63.34, 56.43]	35 / 45
Panel B. Within whole-class time (percentage points)						
Reviewing old material	−8.41	−7.03	[−12.23, −3.54]	[−20.81, 4.50]	[−65.12, 55.58]	35 / 44
Introducing new material	9.13	8.64	[4.03, 14.05]	[−4.57, 23.45]	[−56.40, 64.30]	35 / 44
Critical thinking	−0.95	−1.61	[−4.15, 0.20]	[−8.05, 4.18]	[−63.86, 56.84]	35 / 44
Panel C. Within one-on-one time (percentage points)						
Students below grade level	2.86	1.80	[−4.52, 9.22]	[−17.30, 19.99]	[−60.73, 61.85]	35 / 43
Students passing	0.59	−0.71	[−5.26, 3.39]	[−12.06, 12.13]	[−62.90, 59.68]	35 / 43
Students at mastery	−3.98	−1.09	[−8.14, 1.87]	[−16.85, 10.43]	[−64.12, 58.46]	35 / 43
Panel D. Data consultation (counts)						
Student information system	0.80	−4.57	[−9.43, −1.98]	[−19.43, 10.36]	$(-\infty, \infty)$	34 / 43
Master academic records	−1.64	−1.62	[−2.75, −0.79]	[−4.98, 1.13]	$(-\infty, \infty)$	34 / 43
Assessment records	−1.29	−2.04	[−2.99, −1.34]	[−4.56, 0.64]	$(-\infty, \infty)$	34 / 43
Panel E. Outreach (counts) and planning (minutes)						
Parent contacts	−0.41	−0.83	[−2.00, −0.11]	[−4.20, 2.61]	$(-\infty, \infty)$	34 / 43
Staff contacts	0.11	−0.58	[−2.74, 0.51]	[−5.60, 3.33]	$(-\infty, \infty)$	34 / 43
Minutes planning interventions	−7.56	−12.05	[−16.74, −8.90]	[−23.61, −0.97]	$(-\infty, \infty)$	34 / 43

Paper estimate is the adjusted regression coefficient reported in the main text. Respondent difference, Lee bounds and Manski bounds concern unadjusted, observation-weighted mean contrasts, so the bounds are not intervals around the paper estimate and need not contain it. Lee bounds trim the higher-response group under individual-level monotone selection and describe the always-observed population; Manski bounds fill every missing outcome with its worst case and describe the full eligible frame. Approximate 95% envelopes extend each Lee endpoint by 1.96 times its standard error from 1,999 resamples of the 36 campus-grade cells within the treatment and control groups; they are exploratory, not exact confidence sets. Response is the share of eligible observations with a usable outcome, control / treated. The frame is the 417 classroom-wave invitations in waves 2–5; item response varies by outcome. Manski bounds for consultation, contact and planning outcomes are unbounded because these counts have no verified upper limit. Time shares are bounded outcome by outcome and are not jointly attainable.

Table E2: Sensitivity of teacher prediction accuracy to survey nonresponse

Outcome	Paper estimate	Respondent difference	Lee bounds		Manski bounds	Response (%, C / T)
			Point	Approx. 95%		
Panel A. Percentile error						
All students	−2.12	−0.70	[−3.41, 0.55]	[−9.58, 5.90]	[−66.68, 64.53]	33 / 36
Bottom half of baseline forecasts	–	−1.30	[−3.31, 2.63]	[−9.08, 8.72]	[−61.75, 63.56]	39 / 35
Top half of baseline forecasts	–	−0.06	[−7.39, 5.36]	[−15.09, 12.60]	[−72.20, 65.64]	26 / 36
Panel B. Standardized error (SD)						
All students	−0.051	−0.002	[−0.093, 0.042]	[−0.302, 0.223]	$(-\infty, \infty)$	33 / 36
Bottom half of baseline forecasts	–	−0.033	[−0.103, 0.099]	[−0.299, 0.312]	$(-\infty, \infty)$	39 / 35
Top half of baseline forecasts	–	0.049	[−0.182, 0.231]	[−0.452, 0.454]	$(-\infty, \infty)$	26 / 36

Paper estimate is the adjusted regression coefficient reported in the main text. Respondent difference, Lee bounds and Manski bounds concern unadjusted, observation-weighted mean contrasts, so the bounds are not intervals around the paper estimate and need not contain it. Lee bounds trim the higher-response group under individual-level monotone selection and describe the always-observed population; Manski bounds fill every missing outcome with its worst case and describe the full eligible frame. Approximate 95% envelopes extend each Lee endpoint by 1.96 times its standard error from 1,999 resamples of the 36 campus-grade cells within the treatment and control groups; they are exploratory, not exact confidence sets. Response is the share of eligible observations with a usable outcome, control / treated. The frame is 14,801 teacher-student-assessment-wave opportunities, of which 5,101 have a usable prediction. Errors are held on the scale of the main analysis: reference distributions for ranking and standardizing are those of the observed forecasts. Subgroups use fixed medians of baseline forecasts, so they do not reproduce the paperś decile-based subgroup coefficients (shown as –). Manski bounds for standardized error are unbounded because the measure has no verified finite support.

Forecast-Sample Composition and Accuracy

Table E3 reports baseline balance in the exact forecast-accuracy sample and compares the accuracy estimates with and without flexible controls for baseline predicted performance. Treatment observations have higher baseline predicted achievement, but the estimated accuracy effects are nearly unchanged after adjustment. These checks address observed composition within the prediction sample; they do not eliminate selection on unobserved characteristics.

Table E3: Forecast-sample balance and accuracy sensitivity

Panel A. Baseline balance in the forecast-accuracy sample

Characteristic	Control mean	Treatment mean	Adjusted difference (standard error)	p	N
Baseline predicted-achievement index	-0.183	0.138	0.155 (0.057)	0.006	5,101
Free lunch (%)	33.6	25.4	4.28 (2.32)	0.065	5,101
English learner (%)	23.7	16.1	-6.10 (3.75)	0.103	5,101

Panel B. Accuracy estimates with added baseline forecast controls

	Main specification		Flexible baseline forecast controls	
Outcome	Estimate (SE)	p	Estimate (SE)	p
Percentile error (points)	-2.121 (0.780)	0.007	-2.133 (0.764)	0.005
Standardized error (SD)	-0.0508 (0.0241)	0.035	-0.0504 (0.0239)	0.035

Panel A reports treatment-minus-control differences adjusted for campus, grade, and assessment-by-wave fixed effects. Group means are unadjusted, so their difference need not equal the adjusted gap. Free-lunch and English-learner differences are percentage points. The baseline index averages standardized assessment forecasts and is not itself standardized to have an SD of one. Observations are prediction-weighted: a student appears once per observed assessment and wave.

Both Panel B columns use the same 5,101 predictions from 1,684 students and 69 teachers in 30 campus-grade groups. Negative estimates mean smaller absolute forecast errors. Both retain campus, grade, and assessment-by-wave fixed effects. The second column adds a four-degree-of-freedom cubic spline of the baseline predicted-achievement index and four-degree-of-freedom cubic splines of the assessment-specific baseline forecast, with separate coefficients by assessment. These are pre-treatment forecasts; realized scores are not added as controls. Standard errors are CR1 clustered at campus-grade and p-values use a normal reference, as in the main tables. These checks address observed composition within the prediction sample and do not eliminate selection on unobserved characteristics.

F Is the Accuracy Gain Driven by Changes in Achievement?

Table 6 measures accuracy against realized scores, and the treatment also affected those scores. Treated forecasts could therefore end up closer to the outcome because the outcome moved rather than because the forecasts improved. Separating the two would require each treated student's score without treatment, which we do not observe, so we do not split the gain into shares. We instead ask what an achievement channel would have to look like to produce the gain, and whether the data show it. All estimates use the sample and specification of Table 6.

Forecast changes across the distribution. Figure 7 shows descriptively lower mean forecasts among treated teachers across the bottom seven baseline deciles and higher forecasts at the upper end. These differences need not produce a shift in the pooled average. Table F1 reports no statistically detectable overall treatment effect on either signed error or forecast level. Signed error, where positive means the forecast exceeded the realized score, changes by -0.02 percentile points overall ($p = .99$) and -1.86 in the bottom 50% ($p = .46$), against a control mean of $+1.74$. The forecast level changes by $+0.036$ SD overall ($p = .59$) and by an imprecise -0.094 SD for low-forecast students ($p = .35$).

A shift in average achievement cannot produce the gain. If treatment moved every score by the same amount, it would shrink the errors on one side of zero and enlarge those on the other. Because signed errors are centered near zero, the two nearly offset, and the net change in mean absolute error is small. Panel A of Table F2 removes assumed achievement effects from treated students' scores and re-estimates the accuracy effect. Removing a uniform decline of 0.10 SD, the size of the pooled achievement effect in the main text, changes the estimate from -0.051 to -0.059 SD. In this forecast sample, treated students in fact scored slightly higher ($+0.058$ SD, $p = .37$); removing that shift leaves -0.042.[6]

Only convergence toward forecasts could produce it. Let F be the forecast, $Y(0)$ and $Y(1)$ the untreated and treated scores, and $\Delta = Y(1) - Y(0)$. For squared error,

$$E\big[(F - Y(1))^2\big] - E\big[(F - Y(0))^2\big] = E[\Delta^2] - 2\,E\big[(F - Y(0))\,\Delta\big].$$

[6]The standardized error normalizes forecasts and scores separately within assessment and grade. Using one control-group score distribution for both, on the assessments whose scale did not change during the year, the effect on absolute error is -0.070 SD ($p = .01$; 3,038 predictions).

The first term is positive: achievement effects unrelated to the forecast error add noise and make forecasts look worse. Achievement effects can raise measured accuracy only through the second term, by moving scores toward forecasts, that is, by raising the scores of students whose forecasts were too high and lowering those whose forecasts were too low. Such convergence would itself be substantive. It is the mechanism by which expectations shape outcomes in the literature beginning with Rosenthal and Jacobson [1968] and sharpened by Papageorge et al. [2020].

Panel B tests for convergence by interacting treatment with two pre-treatment proxies for the untreated signed error, in a regression of the standardized score. The first is the control-group mean signed error in the student's baseline algorithm-forecast decile. The second is the teacher's own forecast before any report was delivered, minus the algorithm forecast; in control classrooms it predicts the later signed error. Convergence predicts a positive interaction. Both estimates are negative: -0.141 SD per SD of the decile proxy ($p < .001$) and -0.110 for the teacher-specific proxy ($p = .41$). Where the data are informative, treatment moved scores away from forecasts, which works against the measured gain. Consistent with this, removing decile-specific achievement effects leaves the accuracy effect at -0.052. Even when each decile's effect is placed at the edge of its 95% confidence interval in the direction of convergence, the estimate is -0.029, more than half its original size, though no longer statistically significant ($p = .26$).

A benchmark that treatment could not move. Panel C replaces the realized score with the student's pre-treatment algorithm forecast, which treatment could not affect. The distance between teacher and algorithm forecasts falls by 0.124 baseline SD ($p = .03$). This does not by itself show better prediction of the truth: the algorithm can be wrong, treated teachers received its forecasts, and the estimate is weaker under the wild cluster bootstrap ($p = .18$). It does show that treated forecasts moved relative to a fixed reference, not only relative to outcomes that treatment changed.

What we conclude. The analyses do not support a uniform lowering of forecasts, a shift in average achievement, or achievement effects that track the part of forecast error predictable before treatment as explanations for the accuracy gain. They cannot rule out convergence along teachers' private information about individual students. If such convergence occurred, it would mean treated teachers' expectations shaped achievement, not that the accuracy gain is an artifact of measurement.

Table F1: 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	−3.25		0.256	
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	−0.111		0.434	

Appendix companion to Table 6, on the same sample of 5,101 predictions and the same baseline forecast decile split; Table 6 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. Neither overall treatment effect is statistically distinguishable from zero. All models include interacted assessment × wave fixed effects plus campus and grade fixed effects, with standard errors clustered at campus-grade. *Difference* is the bottom-half effect minus the top-half effect, estimated in one model that interacts every regressor, fixed effects included, with the bottom-50% indicator. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table F2: Is the accuracy gain driven by changes in achievement?

	Estimate	SE	Control mean	p	N
Panel A. Effect on absolute standardized error after removing an assumed achievement effect					
No adjustment (Table 6)	-0.051^{**}	(0.024)	0.610	0.035	5,101
Uniform effect of -0.10 SD	-0.059^{**}	(0.024)	0.610	0.013	5,101
Uniform effect of 0.058 SD (estimated here)	-0.042^{*}	(0.025)	0.610	0.092	5,101
Decile-specific estimated effects	-0.052^{**}	(0.026)	0.610	0.042	5,101
Decile-specific, at 95% bound toward forecasts	-0.029	(0.025)	0.610	0.260	5,101
Panel B. Achievement effect × pre-treatment predicted forecast error (per SD)					
Predicted error from baseline decile	-0.141^{***}	(0.035)		$< .001$	5,101
Teacher's pre-report forecast minus algorithm	-0.110	(0.132)		0.405	2,415
Panel C. Benchmark fixed before treatment (baseline algorithm-forecast SD)					
Distance to baseline algorithm forecast	-0.124^{**}	(0.057)	0.970	0.030	5,101

Same sample of 5,101 predictions, fixed effects (assessment × wave, campus, grade) and campus-grade clustering as Table 6. Panel A subtracts an assumed treatment effect on achievement from treated students' standardized scores, holding forecasts, normalization and sample fixed, and re-estimates the effect on absolute standardized error. Decile-specific effects are estimated separately within each baseline algorithm-forecast decile; the last row moves each to the edge of its 95% confidence interval in the direction that brings scores toward forecasts. Panel B interacts treatment with a standardized pre-treatment proxy for the untreated signed error (forecast minus score) in a regression of the standardized score; a positive coefficient would mean treatment moved scores toward forecasts. The first proxy is the control-group mean signed error in the student's baseline decile, excluding the student's own campus-grade cell. The second is the teacher's forecast before any report was delivered, minus the algorithm forecast, both standardized within assessment and grade; it is available for 2,415 predictions in 21 cells, and in control classrooms a one-SD higher value predicts a 0.082 SD (SE 0.032) higher signed error. Panel C is the absolute distance between the teacher forecast and the same student's pre-treatment algorithm forecast, scaled by the SD of baseline algorithm forecasts within assessment and grade; realized scores do not enter it. Control means are unchanged across Panel A because only treated scores are adjusted; Panel B reports interactions, for which a control mean does not apply. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.