similar rates as others. Other student characteristics varied substantially by district, with District E representing mostly white students, while Districts A and D included mostly students of color; Districts B and C were more racially diverse overall and in our study. District C, the participating CMO schools, tended to have a higher proportion of white students and a lower proportion of students receiving free or reduced price lunch than their surrounding districts.

In each participating classroom, teachers chose six students to collect work from every day during our visit. Teachers photographed this work, replacing students' names with research-team generated identifiers. All students in participating classrooms were eligible to be chosen but were first required to return a signed parental consent form. Approximately 56% of students in our classes returned a consent form.[3] Among those who returned parental consent, we asked teachers to identify students who had mostly regular attendance, and then choose two students each from three proficiency levels: two students who were far below or below grade level, two students who were approaching or at grade level, and two students who were above grade level based on benchmark assessments or prior standardized test scores. Once teachers chose their six students, known as focal students, they photographed their work throughout the school year. This allowed the research team to connect the same student's work across time, connect individuals' work to their final grades in the course and to other measures of student experience gleaned from student surveys.

All students who returned a parental consent took part in student surveys if they were in 3rd grade or above, even if they were not one of the six students chosen to also have their work photographed.

TABLE A.1 | **DESCRIBING PARTICIPATING DISTRICTS AND CLASSROOMS**

		District A	District B	District C	District D	District E
District description:		Urban district	Urban district	CMO with three schools in separate urban districts	Urban district	Rural district
District size (range, to protect anonymity)		10,000 – 50,000	50,000 – 100,000		50,000 – 100,000	1,000 – 2,000
Classrooms						
Total classrooms		100	114	58	110	74
Percent of district's students in a study classroom		5%	2%	---	3%	64%
Grade-Level	K-2	17%	18%	29%	5%	22%
	3-5	28%	20%	28%	9%	38%
	6-8	28%	23%	25%	53%	16%
	9-12	27%	39%	16%	33%	24%
Subject	CTE	2%	4%	0%	3%	4%
	ELA	33%	39%	41%	37%	39%
	Math	47%	39%	29%	27%	41%
	Science	8%	11%	7%	22%	9%
	Social Studies	10%	7%	10%	11%	7%
	Multi-Subject[a]	0%	0%	12%	0%	0%
Course Type[b]	Honors/AP/Dual Enroll	9%	27%	4%	42%	14%
	Remedial/ Intervention	21%	12%	18%	7%	12%

		District A	District B	District C	District D	District E
Teachers[b]						
Total teachers		56	59	33	55	52
Gender	Female	77%	72%	72%	85%	69%
	Male	16%	24%	22%	15%	29%
	Preferred not to say	7%	3%	6%	0%	2%
Race/Ethnicity	Black	7%	5%	0%	15%	2%
	Latinx	2%	7%	0%	33%	0%
	White	79%	78%	84%	39%	96%
	Other or multi-racial	2%	2%	6%	11%	0%
	Preferred not to say	11%	9%	9%	2%	2%
Teaching Experience	0-4 Years	20%	50%	22%	46%	48%
	5-9 Years	27%	22%	31%	19%	17%
	10-19 Years	52%	19%	41%	28%	21%
	20+ Years	2%	9%	6%	7%	13%

a: Some participating elementary classrooms in District C were part of a multi-disciplinary elective, so our team used math, ELA, science, or social studies tools as appropriate, depending on the day.

b: Self-reported by teachers. All teachers weighted equally though some are co-teachers.

TABLE A.2 | COMPARING STUDENTS IN PARTICIPATING CLASSROOMS TO OTHERS IN DISTRICT

	District A		District B		District C		District D		District E	
	In Study	Other	In Study	Other	In Study	Other[a]	In Study	Other	In Study	Other
Achievement (SDs)[b]										
Prior year ELA	-0.78	-0.66	0.16	-0.12	-0.31	---	-0.20	-0.54	0.45	-0.11
Prior year math	-0.80	-0.68	0.19	-0.11	-0.34	---	-0.20	-0.49	0.69	0.34
Growth in ELA[c]	-0.01	0.00	0.02	0.00	0.06	-0.01	0.06	0.00	0.01	-0.02
Growth in math[c]	-0.01	0.00	-0.07	0.00	-0.05	0.01	0.00	0.00	0.01	-0.04
Demographics (%)										
White	8%	20%	32%	25%	30%	11%	2%	2%	86%	80%
Black	58%	47%	15%	13%	44%	34%	7%	6%	10%	15%
Latinx	15%	19%	44%	54%	21%	51%	89%	91%	4%	5%
Asian	18%	9%	3%	3%	2%	3%	1%	-0%	-0%	-0%
Receives FRL	88%	78%	57%	62%	43%	80%	84%	91%	37%	42%
ELL	23%	15%	33%	37%	3%	15%	13%	19%	2%	3%
Special Education	9%	20%	8%	11%	11%	21%	7%	10%	12%	14%

Some students were in multiple participating classrooms, but all summary statistics in the table count each student equally, regardless of the number of classrooms in which they participated. Because districts provided anonymous student data, we were unable to connect demographic and achievement information individually to students but could connect district-provided data to specific classrooms. Thus, all students in participating classrooms are included in the table's results, not just those who were eligible for student surveys or work collection.

a: District C's (the CMO) schools were each located in a separate district, and thus comparisons to an "Other" district are based on publicly available district statistics from the Elementary and Secondary Information System (ELSI) from the 2015-2016 school year. Because the three districts that housed a CMO school varied in size, we averaged the three surrounding districts' information together, weighting them based on the number of participating students in the CMO school in that district. One CMO school did not provide student-level FRL, ELL, or special education information, and thus its surrounding district is not included in the averages for these values.

b: All achievement results represent student-level standard deviation units. Prior achievement was standardized against the state distribution. Positive values imply that students tended to perform better than the average student in the state. Because the CMO's surrounding districts did not always provide average scaled scores, we did not calculate a standardized value for the "Other" group in District C.

c: See later section for discussion of how we calculated student growth. We were only able to calculate student growth for one school in District C because one school did not provide appropriate student achievement data, and in the other school we worked with all classrooms in the school, so without reference data from other students in the district, the average "growth" would mathematically be zero.

COLLECTING AND CALCULATING MEASURES OF STUDENT EXPERIENCE

CLASSROOM AND STUDENT WORK

OVERVIEW AND PROCESS

We asked participating teachers to photograph all work completed by their six focal students during three unique data collection weeks during the 2016-2017 school year. This typically included one week in the fall, one week in the winter, and one week in the spring. We asked teachers to interpret assignments broadly, and photograph anything students needed to write to complete. Assignments could include problem sets in math, grammar worksheets or essays in ELA, reading comprehension responses in social studies, and much more. And assignments could come from a variety of sources, such as questions from a textbook, a worksheet from the teacher's district curriculum, or a set of questions the teacher created or found herself. Because we used photographs of written work as the means by which teachers shared their assignments, we were unable to capture any assignments that did not require writing, like classroom conversations, though we observed this type of instruction when we visited, as is discussed in the "Classroom Instruction and Lessons" section later in this Appendix.

During these weeks, teachers photographed the same six students' work each day and tagged images with unique student identifiers, so that we could track students' progress and connect their work to other measures, like student surveys. We excluded assignments that were not intended to represent a substantial portion of the lesson – such as warm-ups, math fluency sprints, and exit tickets – so that they would not affect our measure of the quality of the typical assignment in each classroom. Teachers uploaded student work images daily, and for each assignment completed a brief survey indicating the source of the assignment and on which days in class the assignment was used. For each submitted assignment, teachers were required to choose one of the following assignment sources: state-developed curriculum or materials, district-adopted curriculum/textbook, district-developed curriculum or materials, other teachers in the district, self-made and not used prior to this school year, self-made and used at least once prior to this school year, website, and other.[4] Assignments connected to any of the sources except the first three were considered teacher-created or teacher-selected. Research team members were present in the participating schools every day during data collection windows to help teachers photograph their students' work and upload the images.

Each assignment and its associated samples of student work was assigned to a research team member to rate. All assignments were assigned to raters randomly at the classroom-by-week level. For example, Rater X would be randomly assigned to all assignments in Ms. Doe's 2nd period Algebra 1 class during our second visit. We found that allowing raters to rate the entire week of assignments made it easier for raters to identify situations where teachers submitted incomplete or duplicate versions of the same assignment.

Raters were subject matter experts for their grade-level and all were former teachers. Seven core raters conducted most of the ratings, though by the end of the project, fifteen individuals had contributed some ratings. Table A.3 shows the number of raters for each subject area and grade level. Some raters rated multiple subjects and/or grade levels.

In total, we collected 4,674 primary assignments representing 21,993 samples of student work. Table A.3 shows the number of assignments collected in each subject and grade level.

TABLE A.3 | **NUMBER OF ASSIGNMENTS AND NUMBER OF RATERS BY SUBJECT AND GRADE LEVEL**

	ELA	Math	Science	Social Studies	CTE
Number of Assignments					
K-2	670	394	6[a]	1[a]	---
3-5	745	568	12[a]	1[a]	---
6-8	336	405	231	190	2
9-12	309	345	218	173	68
Number of Raters					
K-2	8	7	2[a]	1[a]	---
3-5	6	8	2[a]	1[a]	---
6-8	5	8	5	3	2
9-12	3	4	5	3	2

a: Some participating elementary classrooms in District C were used for math, ELA, science, or social studies, depending on the day. So while we targeted mostly math and ELA classrooms at the elementary level, there were some exceptions.

Raters scored two elements of each assignment:

Assignment quality. First, raters evaluated the extent to which assignments gave students the opportunity to meaningfully engage in worthwhile grade-level content by using a TNTP-developed rubric to score each assignment on three domains:

- **Content:** Does the assignment align with the expectations defined by grade-level standards?
- **Practice:** Does the assignment provide meaningful practice opportunities for this content area and grade level?
- **Relevance:** does the assignment give students an authentic opportunity to connect academic standards to real-world issues and/or contexts?[5]

Raters gave each domain a rating of 0, 1, or 2, representing "No Opportunity", "Minimal Opportunity", and "Sufficient Opportunity" respectively. The criteria for each domain rating were subject-specific. See our online resources at tntp.org/student-experience-toolkit for each assignment's quality rubric and examples of how the rubric was applied to assignments.

Student Performance. Raters also reviewed the extent to which each student answered the assignment's questions correctly. This meant applying the assignment's scoring key provided by the teacher to each sample of student work submitted, or if there was no scoring key, determining the proportion of questions correctly answered. Raters determined that a student successfully completed an assignment if the student earned at least 80% of the assignment's potential points or answered at least 80% of the questions correctly, as 80% typically represents the minimum score needed to earn a B on assignments or in a class.

In addition to determining whether students successfully completed assignments, raters also determined if the students' work on the assignment met the requirements of grade-level content standards. For math and ELA, raters referenced [redacted]; for other subjects, raters used the literacy standards from CCSS and the Next Generation Science Standards (in science), the College, Career, and Civic Life (C3) Framework (in social studies), and the Common Career Technical Core (in CTE). Raters first determined whether the assignment earned a rating of "Sufficient Opportunity" on the content domain, thus giving students a chance to meet the demands of grade-level content standards. Raters then applied the assignment's scoring key to determine if the student had earned at least 80% of the assignment's potential points or answered at least 80% of the questions correctly. If the student met this 80% bar on a "Sufficient Opportunity" assignment, the student met the requirements of grade-level content standards. If the student did not achieve this 80% bar on a "Sufficient Opportunity" assignment or if the assignment received a rating of "No Opportunity" or "Minimal Opportunity" in the Content domain, then the student did not fully meet the requirements of grade-level content standards.

All raters received at least two hours of training on each subject-specific student work protocol that they used. This training consisted of an orientation to the protocol they would be using, as well as some practice assignments they rated and then discussed with the project's training team. After this training, all raters were required to pass a norming exercise in each subject area in which they were assigned to rate assignments. Norming exercises consisted of at least five assignments per subject area. To pass the norming exercise, raters' domain ratings needed to exactly match at least 60% of the master ratings and be within one category of the master ratings 100% of the time. Additionally, raters must have exactly matched at least 75% of the master ratings for student performance. Raters were required to pass the norming exercise before they could begin rating assignments in our study.

In the case where an individual rater did not pass the norming exercise on the first try, they received feedback on their initial norming exercise and further training. They were then given another chance to complete another norming exercise using a different set of five assignments. Raters who did not meet the norming bar in their second attempt were not allowed to rate student assignments or work.

RATER AGREEMENT AND RELIABILITY

We assigned approximately 35% of assignments and student work samples to two different raters so that we could test rater agreement and reliability. In most cases (about 30% of assignments) we assigned assignments to two different official raters, but we also had some assignments rated by both an official rater and a master rater, the two most senior experts on the content rubrics. In all cases, raters reviewed their assignments independently and did not discuss ratings.

Table A.4 provides agreement rates for all domains for all rater pairs. Across all subjects, each assignment domain had exact agreement rates of at least 70%, though agreement on the Content and Practice domains in math was lower than other subjects. Binary ratings of whether students successfully completed their assignments had agreement rates above 80%. Table A.4 also reports kappa and tau statistics which represent, respectively, the probability that two raters will give the same rating adjusting for chance agreement, and the correlation between raters' ratings. Our raters' agreement rates and reliability statistics were similar to other studies examining student assignments and work.[6] And raters were within one point on the combined sum of domain ratings, which we use for the majority of analyses, 80% of the time. We determined from these results that assignment raters demonstrated an acceptable level of reliability.

TABLE A.4 | RELIABILITY OF ASSIGNMENT RATINGS

	Exact Agreement	Partial Agreement	Cohen's Ð (weighted)	Kendall's Ð (correlation)
Content				
ALL SUBJECTS	74%	98%	0.75	0.71
CTE	86%	100%	0.82	0.83
ELA	75%	98%	0.76	0.72
Math	68%	97%	0.64	0.58
Science	79%	100%	0.62	0.61
Social Studies	82%	99%	0.71	0.71
Practice				
ALL SUBJECTS	71%	94%	0.65	0.61
CTE	86%	100%	0.86	0.82
ELA	73%	96%	0.67	0.63
Math	62%	90%	0.56	0.51
Science	86%	100%	0.63	0.59
Social Studies	82%	99%	0.62	0.52

TABLE A.4 | **RELIABILITY OF ASSIGNMENT RATINGS**

	Exact Agreement	Partial Agreement	Cohen's Đ (weighted)	Kendall's Đ (correlation)
Relevance				
ALL SUBJECTS	78%	98%	0.75	0.72
CTE	64%	100%	0.55	0.56
ELA	75%	99%	0.72	0.69
Math	84%	96%	0.79	0.78
Science	75%	100%	0.69	0.60
Social Studies	67%	100%	0.65	0.62
Total Score (Sum of domains: 7-point scale)				
ALL SUBJECTS	50%	81%	0.77	0.67
CTE	61%	96%	0.91	0.81
ELA	52%	81%	0.79	0.69
Math	44%	76%	0.70	0.59
Science	56%	92%	0.76	0.64
Social Studies	54%	89%	0.78	0.69
Student Work Ratings (Success on assignment? – yes/no)				
ALL SUBJECTS	82%	--	0.58	--
CTE	75%	--	0.52	--
ELA	81%	--	0.50	--
Math	82%	--	0.63	--
Science	84%	--	0.51	--
Social Studies	89%	--	0.70	--
Student Work Ratings (Did work demonstrate content standards? – yes/no)				
ALL SUBJECTS	85%	--	0.52	--
CTE	100%	--	--	--
ELA	87%	--	0.53	--
Math	78%	--	0.47	--
Science	96%	--	0.24	--
Social Studies	98%	--	0.58	--

Partial agreement represents the percent of responses off by no more than 1 category (i.e. a 0 and 1 are partial matches). Cohen's Kappa employs "squared" weights to differentiate between disagreements that were farther apart: disagreements are weighted according to their squared distance from perfect agreement. Because student work ratings were binary, weighted Kappa is no different than unweighted Kappa for these results. Pairs of ratings from "Master" raters and official raters are included.

DEFINING GRADE APPROPRIATE ASSIGNMENTS

Assignments that had a total domain score of at least a 4 out of 6 were considered "grade appropriate". This definition required that any assignment labeled grade appropriate had to score at least a 1 ("Minimal Opportunity") on all three domains and a 2 ("Sufficient Opportunity") on at least one domain. Alternatively, an assignment could score a 2 ("Sufficient Opportunity") on two domains and a 0 ("No Opportunity" on the other). Our definition of grade appropriate was based on our rubric and belief that a grade appropriate assignment must provide students some opportunity to engage in grade-level content, practices, and have an appropriate connection to the broader world, or must provide a full opportunity in at least two of these domains. See our online resources at tntp.org/student-work-library for examples of the differences between grade-appropriate assignments and other assignments.

CLASSROOM INSTRUCTION AND LESSONS

OVERVIEW AND PROCESS

We observed at least two full lessons in nearly all participating classrooms. For classrooms that had co-teachers, we focused at least two observations on each teacher. We coordinated all lesson observations with participating schools and teachers, so that participating teachers always knew in advance when we would be observing them. This allowed us to schedule days when teachers' instruction would not be shortened by, for example, tests or quizzes, school assemblies, or other events that altered the schedule. In a few cases (n = 31), teachers were unexpectedly absent on scheduled observations or some other conflict disallowed us from observing them on a scheduled day, and we were only able to visit their classrooms once.

We conducted observations during the same three weeks in which teachers photographed classroom assignments. We scheduled observations so that we would not observe the same teacher in the same classroom more than once in the same week, but in some situations (n = 17) this occurred to make up for observations that had to be canceled in previous weeks.

In all, we rated 942 lessons and 422 classrooms had at least two observations.

Observers rated each lesson using a subject specific rubric. See our online resources at tntp.org/student-experience-toolkit for copies of these rubrics. Observers rated the following five domains:

- **Culture of Learning:** Are all students engaged in the work of the lesson from start to finish? Do they follow behavioral expectations?
- **Content:** Does the content of the lesson reflect the key instructional shifts required by college and career ready standards?
- **Reading Foundations (K-2 ELA only):** Does instruction develop skills in service of comprehension?
- **Instructional Practices:** Does the teacher employ instructional practices that allow all students to learn the content of the lesson?
- **Student Ownership:** Are students responsible for doing the thinking in the classroom?

These domains were based on TNTP's Core Teaching Rubric and the Student Achievement Partners' Instructional Practice Guides.[7] Observers gave a 4-point categorical rating on each domain: 0 "Not Yet", 1 "Somewhat", 2 "Mostly", and 3 "Yes". In some lessons, it was impossible to judge all domains. Specifically, when students spent most of the lesson focused on narrative writing in ELA, or on data analysis and experiment in science or CTE, observers did not rate content, instructional practices, or student ownership. These lessons were excluded from all analyses (2% of all lessons). All other lessons were rated (n = 921), and 405 classrooms had at least two non-excluded observations.

In addition to using the domains to assess different aspects of lesson quality, observers also tallied the number of minutes students spent in different activities, (e.g., partner or small group work, introductions and warm-ups, whole class discussions). They also indicated the number of minutes that students spent on activities unrelated to class content, assignments, or activities, though they did not specify what students were using this unrelated time to instead do.

Like assignment raters, observers were experts in their subject area and all were former teachers. Given the number of observations required, we were unable to randomly have multiple observers rate the same classroom to check inter-rater reliability. However, all observers went through content-specific training for each observation protocol they used, including virtually led sessions that oriented them to the observation protocols and practice video observations that allowed them to apply the tool. Observers needed to exactly match at least 60% of the master ratings and be within one of the master ratings at least 85% of the time to be considered normed.

Observers practiced applying the rubric to at least five video lessons in each subject they would be observing. Observers needed to rate at least five other videos successfully in order to be allowed to officially observe the classroom in the study. After observers were certified but before starting observations in the field, they continually engaged in a set of practice observations (reviewing video footage of instruction) and calibrated on ratings through team discussion. Additionally, practice sessions occurred throughout the observers' first three months in the field to continue to verify that all observers were normed. Master raters also reviewed scripted notes and evidence from observations to check for normed ratings. During the first two site visits, master raters reviewed all scripted notes, and reviewed 2-3 random observation scripts in each of the following site visits. Observers received written feedback on any observation rating that was not normed, and occasionally engaged in debrief conversations.

DEFINING STRONG INSTRUCTION

We considered lessons where the average domain rating was at least a 2 (out of 3) to represent strong instruction, as it implied that the average domain scored at least "Mostly". Table A.5 displays the characteristics of lessons with strong instructions compared to other lessons. As expected and by definition, lessons with strong instruction were substantially more likely to be rated "Mostly" or "Yes" in each of the four observation domains. Nearly all lessons with strong instruction had strong cultures of learning, focused on the right content, and had highly rated instructional practices. Though "Student Ownership" tended to be the lowest rated domain across all lessons, most lessons with strong instruction earned high ratings on this domain while nearly zero lessons without strong instruction had this quality.

There were also differences in the time students spent on things unrelated to class, only 3% of class time in lessons with strong instruction, but 14% in other lessons. And when we asked raters to determine, holistically, whether the lesson represented the type of instruction and content called for by rigorous content standards, raters of nearly all (92%) lessons with strong instruction said "Yes" or "Yes, but only in some areas", while only 5% of other lessons elicited this response.

TABLE A.5 | DESCRIPTION OF LESSONS WITH STRONG INSTRUCTION AND OTHER LESSONS

	Lessons with strong instruction	Other lessons
Percent of lessons rated as "Mostly" or "Yes" on…		
Culture of Learning	97%	52%
Content	99%	41%
Instructional Practices	87%	3%
Student Ownership	64%	1%
Percent of time students spent on activities unrelated to class	3%	14%
Percent of lessons reflecting the demands of the standards	92%	5%
N	142	779

Note that "Percent of lessons reflecting the demands of the standards" was based on a 4-category holistic rating that raters assigned to every lesson. Raters responded "No", "Not really, but there were some promising practices", "Yes, but only in some areas" or "Yes" to the prompt "Overall, did this lesson reflect the demands of the standards and/or the instructional shifts the standards require?" Percentages represent "Yes" or "Yes, but only in some areas."

STUDENT ENGAGEMENT AND PERCEPTIONS OF WORTH

OVERVIEW AND PROCESS

To capture how students perceived their day-to-day classroom experiences we aimed to repeatedly measure students' experiences as they occur in their natural setting and in real time. Our approaches focused on allowing students to tell us how they felt while, or as close as possible to, interacting with daily class activities.

For grade 6-12 students, we used the Experience Sampling Method (ESM) pioneered by Mihaly Csikszentmihalyi and colleagues.[8] ESM is a strategy by which participants are randomly signaled throughout their day in order to complete a brief survey about what they are doing, thinking, and how they feel about it. During the entire week of our second and third site visits, all students with parental consent were provided a vibrating watch and a survey at the beginning of class. At six points during class, a handful of watches would vibrate. When a student's watch vibrated, it was his or her signal to complete the survey about their current activity and perceptions. In this way, all eligible students completed one survey per class per day, and we could capture experiences throughout class instead of at one distinct point in time.[9]

Grade 3-5 students did not receive a watch, but instead completed a survey in the last five minutes of an instructional period (like math or language arts) about their experiences and activities during the just-finished period. This approach is more akin to a daily diary than the signal contingent ESM approach used for secondary students. Yet completing surveys immediately after instruction increases the level of detail with which we can ask questions and reduces memory bias over one-time retrospective surveys.[10]

Students completed surveys separately in every class that participated in the study. In several cases, this meant the same student was signaled and provided daily surveys in two, three or even four classes. Because all surveys were pre-coded with student and classroom IDs created by the research team, we could track the same students' responses in different classes. In all, we collected 28,575 ESM and daily diary responses; 3,133 students completed at least one ESM or daily diary survey.

In addition, each student completed a one-time survey in each participating class that asked broader questions about their educational and career aspirations, as well as questions about how they viewed their teacher and his/her instruction and beliefs. We collected 3,926 of these background surveys representing 2,973 students.[11]

USING STUDENT SURVEYS TO MEASURE ENGAGEMENT AND WORTH

All daily student surveys included a collection of questions meant to measure students' engagement. (See our online resources at tntp.org/student-experience-toolkit for a copy of the student survey.) Engagement is a broad concept that encompasses multiple constructs.[12] Drawing on prior research, we defined engagement along three dimensions: enjoyment, interest, and concentration, and asked students multiple questions connected to each domain (see Table A.6)[13]. Our definition of engagement is not synonymous with "being on task" or "paying attention;" it encompasses students' emotional and cognitive reactions to what they're being asked to learn. These types of reactions are not often visible to external observers – students who are on task, for example, could still very well be doing little thinking about their work. And these reactions are not always easily generalized – asking students if they are bored now is different than asking students if their class this year is boring. Our measurement of engagement takes advantage of the in-the-moment and repeated measures of our survey process.

In addition to engagement, each survey also used several questions to measure the extent to which students viewed what they were doing in class as worthwhile (see Table A.6). Though ESM research in schools has tended to focus more on engagement, we adapted three items from Uekawa, Borman, and Lee (2007) that asked students the importance of their experience for different outcomes as the basis for our construct measuring students' perception of worth. [14]

For both constructs, we combined all the individual questions into separate, single measures of engagement and worth using a Rasch measurement process. A Rasch measurement process uses the patterns and frequencies of responses to individual survey items to create an overall scale. At any point on this scale we know the probability of responding a certain way to each question. A Rasch process allowed us to generate a numeric "engagement" and "worth" score from each collected survey that was rooted in the probability of endorsing the various questions connected to these constructs. Survey responses that scored higher on these scales were more likely to agree with the statements and thus showed higher levels of engagement and perceptions of worth, respectively. Importantly, the scales created through Rasch processes are linear, and can account for the possibility that a response of "A little true" is farther from a response of "Not true" than it is from "Mostly true," and thus have a wider distance on the scale for the former.

We conducted Rasch processes separately for secondary ESM and elementary daily diary responses. This meant we created two separate scales for each construct. To make these scales easier to understand, we converted both to a range of 0 to 10, where 0 represents the strongest possible disagreement on all items and 10 represents the strongest agreement on all items. Only surveys that had responses to at least half the items in a given construct were eligible to be part of the Rasch scale creation.

TABLE A.6 | RASCH CHARACTERISTICS OF ENGAGEMENT AND WORTH CONSTRUCTS

	Elementary Daily Diary		Secondary ESM	
	Middle Threshold	Infit	Middle Threshold	Infit
ENGAGEMENT	**Reliability: 0.77**		**Reliability: 0.83**	
Class was about something interesting. [a]	4.9	0.95	5.4	0.86
I felt excited about learning. [b]	3.5	0.85	5.6	0.95
I really liked what we were doing in class. [a]	4.7	0.85	5.4	0.84
I felt bored. [b, c]	3.1	0.92	3.7	1.05
I wish I was doing something else. [b, c]	3.5	1.13	4.4	1.04
I was thinking more about class than anything else. [a]	5.1	1.34	4.7	1.21
I felt focused. [b]	2.6	1.02	3.0	1.08
WORTH	**Reliability: 0.78**		**Reliability: 0.82**	
Class was about something interesting. [a]	5.0	0.96	5.1	0.94
Class was about something I can use outside of school. [a]	5.0	1.21	6.0	1.14
Class was about something important to my future. [a]	4.6	0.93	4.8	0.91

Because secondary students responded to questions in the moment, they were asked to report how they felt the moment their watched vibrated. Elementary students were asked to think about how they felt in class that day. Reliability represents the proportion of variance between responses that is not due to error (often referred to as person separation reliability) and is similar to Cronbach's Alpha. Middle threshold is the location on our transformed 0-10 scale where the probability of responding Yes (for binary items) or Mostly True or Very True (for 4-point items) is 50%. Infit is the unstandardized mean square infit value.

a: Questions on a four-point scale: Not true, A little true, Mostly true, Very true

b: Questions on two-point scale: No, Yes.

c: Reverse coded item.

Table A.6 provides two key Rasch metrics for each survey item used in the engagement and worth constructs. The first is the item's middle threshold location, which represents the point on our 0-10 scales where a student would have a 50% probability of responding "Yes" or at least "Mostly true." Larger values imply that students had more difficulty agreeing with an item. For example, secondary students were more likely to say they felt focused than that they felt excited to learn. The second key Rasch metric is infit, which represents whether the responses on the survey item tended to match what we expect given that survey response's overall engagement or worth score. Reasonable infit values for a survey like ours are 0.6 to 1.4.[17] All our items' infit values fall within this range.

We also used the fact that at any point on the Rasch-based scale we can determine the probability of a specific response to create four meaningful categories within each scale. For both engagement and worth, we used the lowest point on the scale where the most likely response for each question was marking "Very true" or "Yes" as the threshold that above which responses were placed into the highest category. Similarly, we used the highest point on the scale where for each question the most likely response was "Not true" or "No" as the threshold that below which responses were placed in the lowest category. To create the middle two engagement categories, we used the point on the scale that represented where a response was equally likely to respond "A little true" as "Mostly true" on the most difficult 4-choice Likert item. This implies that for each survey item, responses categorized as "Engaged" were most likely to have a response of at least "Mostly true". We followed the same approach for worth but used the least difficult item instead of the most difficult item so that responses categorized as "Worthwhile" had at least one item where the most likely response was at least "Mostly true", as it was more theoretically meaningful for students to see at least one type of worth in what they were doing for the experience to be considered worthwhile.

Table A.7 shows the response breakdown for each category. By design, there are clear differences between student experiences classified as engaged or highly engaged compared to disengaged or minimally engaged. For example, 47% of the survey responses we classified as "Minimally Engaged" indicated they were bored, while only 4% of responses classified as "Highly Engaged" did so.

TABLE A.7 | **AGREEMENT RATES FOR ENGAGEMENT AND WORTH ITEMS BY CATEGORY**

	ALL Survey Responses	BY SURVEY RESPONSE CATEGORY			
		Disengaged	Minimally Engaged	Engaged	Highly Engaged
ENGAGEMENT					
Class was about something interesting.	55%	- 0%	20%	79%	99%
I felt excited about learning. [a]	58%	4%	31%	76%	97%
I really liked what we were doing in class.	56%	- 0%	20%	83%	-100%
I felt bored. [a]	30%	81%	47%	13%	4%
I wish I was doing something else. [a]	36%	82%	55%	19%	9%
I was thinking more about class than anything else.	59%	3%	36%	76%	98%
I felt focused. [a]	75%	21%	67%	87%	95%
N		3194	9400	8637	6726
Elementary scale range		0 - 2.2	2.2 - 5.1	5.1 - 6.9	6.9 - 10
Secondary scale range		0 - 2.3	2.3 - 5.6	5.6 - 8.4	8.4 - 10
WORTH					
Class was about something…					
Important to my life right now.	50%	0%	18%	84%	98%
I can use outside of school.	44%	0%	14%	66%	97%
Important to my future.	54%	0%	24%	88%	99%
N		5308	8445	7991	5755
Elementary scale range		0 - 2.5	2.5 – 4.6	4.6 – 6.7	6.7 -10
Secondary scale range		0 - 1.9	1.9 - 4.8	4.8 – 8.5	8.5 - 10

All questions on a four-point scale: Not true, A little true, Mostly true, Very true, except questions marked with an (a), which are on a two-point scale: No, Yes. All percentages represent percent of "Yes" responses for the latter and Mostly True or Very True for the former. Ns represent total number of experiences in each category, though in some cases students did not respond to all questions, so Ns for each item are slightly smaller. All survey responses also include responses where students did not answer enough questions to be classified into one of the engagement/worth categories.

STUDENTS' CAREER AND EDUCATIONAL AMBITIONS

The background survey asked all students to name the job they hoped to have when they were an adult. 2,854 students provided us with their job aspiration. We coded all aspired-to jobs using the Bureau of Labor Statistics' 2010 Standard Occupation Classification codes but added an additional category for students who said they were unsure. Additionally, for each job we classified whether it required a college degree by considering the entry-level education requirements from the Bureau of Labor Statistics' Occupational Outlook Handbook as well as researching the typical entry requirements for less common choices.[18] Because secondary students completed the background survey in each participating class, some students had multiple chances to indicate their career ambitions. In cases where the same student responded differently, we took the modal response, and if there was no mode, randomly selected an entry.

The background survey also asked students about their expected educational attainment; students were asked to state whether they expected to finish high school, finish high school but not attend college, attend college, complete college, and attend graduate school. The survey did not, however, differentiate between types of colleges, such as 2-year versus 4-year programs, or provide a choice for technical or career training not housed in a college or university, such as apprenticeships.

TEACHERS' PERCEPTIONS OF CONTENT STANDARDS AND EXPECTATIONS FOR STUDENT SUCCESS AGAINST THE STANDARDS

Nearly all participating teachers completed a one-time survey that asked questions about their experiences, beliefs, and knowledge of their state's content standards (n = 252; response rate = 99%). In addition to obtaining basic teacher background characteristics, like years of teaching experience, and race/ethnicity, we used a set of eleven questions to create two constructs used in multiple analyses: Support for state standards, and expectations for student success against the standards. Because the standards are unique to each subject area, teachers responded to these questions separately for each subject in which they had a participating class – for example, self-contained elementary teachers responded separately for math and ELA. All items were on a 6-point scale: Strongly disagree, Disagree, Somewhat disagree, Somewhat agree, Agree, and Strongly agree. See Table A.8 for a list of all eleven items.

Though informed by recent teacher surveys on the applicable state standards, most existing surveys had focused questions on implementation of the standards or general job satisfaction.[19] We developed these eleven items specifically for this research project because we also wanted to learn more about the extent to which teachers' state content standards aligned to their beliefs about teaching and learning, satisfaction with the day-to-day work of teaching, and their views of the appropriateness of these standards for students. An exploratory factor analysis revealed two separate groupings of items. We thus split items based on their factor loadings and what they represented conceptually. After splitting items by construct, the first principal component explained 60% of the support construct and 68% of the expectations construct.

Like engagement and worth from the student surveys, we used a Rasch measurement process to create these constructs, and similarly put them on a 0-10 scale to aid interpretation. Table A.8 provides a list of all items making up the "support" and "expectations" constructs, as well as their infit and the location of the threshold where responders were at least 50% likely to at least somewhat agree with the item. Both scales demonstrate suitable reliability and nearly all items were properly fitting.[20]

We also similarly created four categories within each construct based on where the overall score fell on our 0-10 scale. Because the Rasch process allows us to estimate the probability of responses to each item at any point on the scale, we used the highest point on the scale where the most likely response for each question was "Disagree" or "Strongly Disagree" (or "Agree" or "Strongly Agree" in reverse-coded items) as the threshold that below which responses were placed into the lowest category. Similarly, we used the lowest point on the scale where for each question the most likely response was "Agree" or "Strongly Agree" to mark off the top category. To split the middle two categories, we took the median threshold representing a 50% chance of responding "Somewhat Agree" or higher. Table A.8 shows how we created four categories out of each scale, as well as the agreement rates to each item given one's categorical placement.

TABLE A.8 | **THRESHOLDS, INFIT, AND AGREEMENT RATES FOR SUPPORT AND EXPECTATIONS CONSTRUCTS**

SUPPORT FOR STANDARDS *Reliability = 0.89*						
			AGREEMENT RATES			
	Middle Threshold	Infit	Oppose	Moderately Oppose	Moderately Support	Support
The standards reflect my beliefs about the content students should be focusing on.	3.7	0.97	0%	22%	83%	98%
Teaching and learning that is aligned to the standards gives students a deep understanding of the subject area.	3.4	0.90	0%	39%	88%	100%
Teaching and learning that is aligned the standards make class more engaging for students.	4.2	0.94	0%	19%	72%	98%
The standards reflect my beliefs about good teaching.	3.6	0.89	17%	18%	85%	100%
The standards make teaching less enjoyable. [a]	4.9	1.64	100%	83%	45%	6%
Teaching and learning that is aligned the standards provides students with lifelong skills.	3.1	0.95	0%	43%	90%	100%
Teaching and learning that is aligned to the standards prepares students for their future.	3.1	0.76	0%	31%	95%	100%
N			6	54	210	62
Scale range			0 – 1.7	1.7 – 3.6	3.6 – 6.3	6.3 - 10

EXPECTATIONS FOR STUDENT SUCCESS *Reliability = 0.84*						
			AGREEMENT RATES			
	Middle Threshold	Infit	Low	Moderately Low	Moderately High	High
The standards make it difficult for students to learn basic skills in this subject. [a]	4.5	0.99	98%	65%	12%	2%
Students are overburdened by the demands of the standards. [a]	5.5	0.90	100%	91%	31%	4%
My students need something different than what is outlined in the standards. [a]	5.6	1.19	100%	81%	53%	4%
The standards are too challenging for my students. [a]	4.5	0.98	100%	65%	13%	0%
N			44	150	83	55
Scale range			0 – 2.4	2.4 – 5.0	5.0 – 6.8	6.8 - 10

All questions on a six-point scale: Strongly disagree, Disagree, Somewhat disagree, Somewhat agree, Agree, Strongly agree. Agreement rates represent percent of Somewhat Agree or higher responses. Reliability represents the proportion of variance between responses that is not due to error (often referred to as person separation reliability) and is similar to Cronbach's Alpha. Middle threshold is the lowest location on our transformed 0-10 scale where the probability of responding Somewhat Agree is at least 50%. Infit is the unstandardized mean square infit value.

[a] Reverse coded item.

CREATING CLASSROOM-LEVEL MEASURES OF ASSIGNMENTS, OBSERVATIONS, ENGAGEMENT/WORTH, AND TEACHER PERCEPTIONS

Most of our analyses are at the classroom level, as prior research has pointed to substantial variation between classrooms within the same school in the amount of time spent on academic content and its rigor.[21] In addition, participating districts provided only anonymous student-level data, meaning we could not connect any of our student-level metrics like engagement or assignment success to the individual student characteristics districts provided. The smallest possible unit of analysis for most of our analyses was the classroom. We wanted to know, for example, whether classrooms serving certain types of students tended to have higher-rated assignments than others, or whether classrooms that tended to have higher-rated lessons had better achievement results.

As expected, classrooms' scores on each metric were not identical from assignment to assignment or student-reported experience to experience. Table A.9 shows how much of the variation in each measure could be attributed to different factors, including our primary unit of analysis: classrooms. We estimated variation in three ways:

- **Model 1.** First, we combined data from all core classes (math, ELA, science, and social studies) and partitioned the variance into every grouping level of interest by using an unconditional multi-level model and including random effects for each group. We indirectly accounted for differences due to grade level (K-2, 3-5, 6-8, and 9-12) and subject by standardizing all metric scores on both dimensions prior to modeling. For metrics with subdomains, we performed an identical process on each domain. Notably, Model 1 includes random effects for both teachers and classrooms meaning we are estimating differences between teachers as well as between classrooms of the same teacher.

- **Model 2.** Second, we used the same approach as Model 1 but excluded teacher random effects in order to better represent the variation a typical student experiences. Because students are most often only in one class of a given subject, a student's experience is simultaneously affected by teacher and classroom effects. We chose to focus on classroom variation in these models because our research questions are generally focused at the classroom level. For most metrics, the fuller results from Model 1 demonstrated meaningful variation between teachers so it's important to interpret the percent of variation due to classrooms in these models as also including variation due to differences between teachers.

- **Model 3.** Finally, in addition to excluding teacher effects, we also excluded school and district random effects in order to make classrooms the highest-level grouping and better show the full range of variation between the classrooms in our study. Excluding the (on most metrics) relatively small variation between schools and districts allows us to interpret the variation due to classrooms in these models as an estimate of how much classrooms across our study varied, rather than an estimate for how much classrooms varied within the same school and district.

For all models, we converted variance estimates to the percent of total variance. Using the results from Model 3, we also estimated the reliability of taking the mean of all assignments, observations, and survey results in a classroom given a set of minimum sample sizes. Below, we interpret the variance decomposition results in Table A.9 and discuss what they mean for the reliability of our classroom measures.

- **Assignments.** The majority of variation was attributable to differences in assignment quality given to students in the same classroom. This is seen in the variation due to assignments, which was near 50% for all outcomes. Some classroom assignments were better than others, which reinforces our study's decision to collect assignments over multiple weeks: A single assignment is a noisy measure of the typical classroom experience. Teachers composed a meaningful proportion of variance, though we also saw meaningful difference between classrooms of the same teacher. Only a small fraction of the variation was due to differences between schools or districts. Overall, only a small proportion of variation was due to consistent differences between raters, though these rater effects were higher for content and practice ratings. The remaining variation, known as residual variation, represents variation due to raters disagreeing about the ratings to give assignments. When we let classrooms be the highest unit of analysis (i.e., Model 3), we can reliably represent differences between classrooms with only 8 assignments and 2 raters (reliability = 0.68). Across all core subjects, 84% of classrooms had enough assignments to have an estimated reliability above 0.65.

- **Instruction.** Because all observations were in-person, we are unable to partition the variation in instruction scores beyond classrooms, schools, districts, and raters. This means that the residual variation, which accounts for most of the variation represents differences in instruction scores due to different ratings of the same classroom on different days, as well as other sources of error. Like Kane & Staiger (2012), we found substantially more variation between teachers than between two classrooms of the same teacher. And like these authors we too found that between classroom variation was highest for domains related to classroom management (i.e., Culture of Learning). Like assignments, there was more variation between teachers in the same school than between schools or districts. School variation was highest for Culture of Learning and district variation was highest for Student Ownership. This suggests that instruction might be more sensitive to the types of students in the classrooms being observed. Though most of the non-residual variation was between teachers, most classrooms did not have sufficiently reliable estimates of their instruction given only two observations were conducted: three observations were required to have a reliability of over 0.65, and only 16% of classrooms had this. This does not mean we can't reliably use instruction scores in the analysis that includes many classrooms, just that our overall instruction scores are a noisy reflection of a single classroom on its own.

- **Student Surveys.** Differences between students represented the biggest source of non-residual variation on engagement and worth. That means even within the same class and on the same day of instruction, some students were more engaged or had higher perceptions of worth than others. Additionally, the larger residual variation implies that students were not always engaged or perceiving worth: their responses varied survey-to-survey. These sources of variation far surpassed differences in engagement and worth due to different days in the same class or due to differences between classrooms or teachers themselves. Schools and districts represented a minimal proportion of variation, too. Fortunately, our survey design captured as many students in a classroom as possible and collected multiple surveys from the same student throughout the study. Thus, a classroom with 10 students, each surveyed once across five days had reliabilities near 0.65, and when we separately calculated the reliability for each classroom given the variance decomposition, 79% and 75% of classrooms had enough students and surveys to have reliabilities above 0.65 for engagement and worth respectively.

- **Teacher Perceptions.** Expectations and support were based on a one-time teacher survey that each teacher completed separately for every subject they taught in the classes they used for the study, Most secondary teachers only completed one survey as they only taught one subject. We thus focused the variance decomposition at the teacher level and did not estimate the proportion of variation between classrooms of the same teacher. This latter source of variation is instead part of the residual term. For both support and expectations, most of the variation is between teachers, though there was some meaningful variation on expectations between districts. Notably, District C's teachers had significantly higher expectations ($p<0.001$) than all other districts except District B, after controlling for grade level and class subject.

The results from Table A.9 show that most classrooms had reliable estimates on most measures. But instead of just averaging scores to the classroom level we took the extra precaution of accounting for different sources of variability in these measures as well as differences in sample sizes between classrooms. Below, we describe how we created a single summary measure on each metric for each classroom. These summary measures were used in our analyses connecting our metrics to student characteristics as well as to student achievement.

- **Assignments.** We excluded any classrooms that submitted fewer than five days' worth of assignments and for the remaining classrooms used a multi-level linear model to obtain "shrunk" estimates of the mean overall assignment score (the sum of all three assignment domains) for each classroom. A shrunk estimate for a classroom is a compromise between simply averaging all the assignments used in a given classroom and the average of all assignments across all classrooms. Each classroom's raw mean gets shrunk toward the overall mean based on how much variation there is between classrooms – more variation means less shrinkage – and based on how many assignments were collected in a class – more assignments means less shrinkage. This process is conceptually based on the notion that a simple classroom mean, especially one based on only a few assignments, could overstate differences between classrooms, and so we can use what we know about typical assignment scores across all classrooms and about how different classrooms tend to be to get a better estimate for each classroom. Shrunk means are commonly used in educational research.[22] We ran separate multi-level models for each subject area so that classrooms were shrunk to their subject specific means, and we included a rater random effect to account for differences in which raters rated which classrooms.

  We weighted assignments by how much class time students spent on them. Therefore, each assignment was weighted so that on any given day in a classroom the sum of these weights was one. This implies that assignments used on days when teachers used several other assignments were weighted less than an assignment that was the only one used in a day, because ostensibly students had a lower fraction of class time to spend on them. We assumed assignments given on the same day required an equal amount of class time.23 We also weighted double-rated assignments so that each counted as one half of an assignment.

- **Instruction.** We excluded classrooms that had fewer than two lesson observations. For the remaining classrooms we used a similar multi-level linear model-based shrinkage process on the total instruction scores (the sum of all four domains) across all observed lessons, similarly including a rater random effect. Because most classrooms had an equivalent number of observations – two – and because in our predictive analyses we standardized all classroom metric scores, the use of shrunk estimates here had only a minor effect. Like assignments, all classrooms were shrunk to their subject-specific means.

- **Student Surveys.** For each student survey metric, we excluded classrooms that had fewer than 20 survey responses.[24] For the remaining classrooms we again applied the modeling process described above, shrinking each classroom to its subject-specific mean. Prior to running the multi-level models, however, all student survey responses were first standardized by grade, with all high school grades collapsed together. This allowed us to include elementary and secondary surveys in the same analyses and control for the relationship between grade level and engagement and worth in subsequent analyses. The final shrunk classroom means, therefore, represented the extent to which students in the class were more or less engaged or perceived more or less worth than other students in the same grade.

- **Teacher Surveys.** Most classrooms (n = 429) had one teacher who completed the teacher survey once. Thus, we simply used these values to represent each classroom. When classrooms had co-teachers (n = 14), we averaged the expectations and support constructs together.

TABLE A.9 | DECOMPOSING VARIANCE IN MEASUREMENT SCORES FOR MATH AND ELA

ASSIGNMENTS	Assignment	Class	Teacher	School	District	Rater	Residual	Reliability[a]
Model 1								
Overall score	56%	5%	13%	2%	3%	4%	17%	
Content	45%	8%	8%	4%	1%	15%	19%	
Practice	43%	8%	5%	3%	2%	14%	26%	
Relevance	57%	4%	15%	-0%	3%	5%	16%	
Model 2 - Overall score	57%	17%	---	3%	3%	4%	17%	
Model 3 - Overall score	57%	22%	---	---	---	4%	17%	0.68
INSTRUCTION		**Class**	**Teacher**	**School**	**District**	**Rater**	**Residual**	**Reliability[a]**
Model 1								
Overall score		-0%	29%	5%	8%	-0%	57%	
Culture of Learning		10%	29%	20%	-0%	3%	38%	
Content		-0%	18%	2%	4%	3%	72%	
Instructional Practices		-0%	21%	1%	9%	2%	66%	
Student Ownership		-0%	16%	3%	11%	-0%	69%	
Model 2 - Overall score		27%	---	7%	8%	-0%	58%	0.57
Model 3 - Overall score		41%	---	---	---	1%	59%	
ENGAGEMENT	**Day**	**Class**	**Teacher**	**School**	**District**	**Student**	**Residual**	**Reliability[a]**
Model 1	5%	1%	6%	1%	1%	39%	48%	
Model 2	-0%	6%	---	1%	1%	38%	52%	
Model 3	-0%	8%	---	---	---	39%	53%	0.63
WORTH	**Day**	**Class**	**Teacher**	**School**	**District**	**Student**	**Residual**	**Reliability[a]**
Model 1	2%	1%	7%	2%	-0%	44%	45%	
Model 2	-0%	6%	---	2%	-0%	44%	47%	
Model 3	-0%	8%	---	---	---	45%	47%	0.60
EXPECTATIONS[b]				**Teacher**	**School**	**District**	**Residual**	**Reliability[a]**
Model 1				70%	-0%	19%	11%	
Model 3				88%	---	---	12%	0.88
SUPPORT[b]				**Teacher**	**School**	**District**	**Residual**	
Model 1				77%	-0%	8%	15%	
Model 3				85%	---	---	15%	0.85

We used a multi-level unconditional model to estimate all variances and proportion of total variance. For all models, we used the lme4 package in R, which is suitable for the nested and crossed random effects in our models. See: Bates, D. Maechler, M., Bolker, B., & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

a: Reliability calculated using Model 3 estimates with the following assumptions: For assignments, 2 reviewers with 8 assignments and 12 total assignment ratings; For instruction, 2 observations with only 1 reviewer; For engagement and worth, 10 students on 5 different days with 50 total responses; For expectations and support, 1 teacher in 1 classroom. Reliability calculated as the percent of variation attributable to classrooms divided by the sum of the remaining variation, each divided by the appropriate number of instances. For example, for assignments: %Class / (%Class + %Assignments/(# of assignments) + %Rater/(# of raters) + %Residual/(# of assignment ratings). See Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation.*

b: With the minor exceptions of some co-teaching classes, there was only one teacher perception score per class, thus making it impossible to estimate the variation within classrooms. Because we surveyed each teacher on each subject, the expectations results are more strongly connected to the teacher than the classroom, so we excluded between classroom variation. Model 2, therefore would be identical to Model 3.

MEASURES OF STUDENT EXPERIENCE FROM DISTRICT ADMINISTRATIVE DATA

COURSES COMPLETED AND GRADES EARNED

Participating school districts provided us records on every course taken by every student in the district – not just those in participating schools. These data included whether students earned credit in the course and their final course grade. The four public school districts provided these data from the 2012-2013 school year through the 2016-2017 school year.[25] In all, this amounted to over a million student-by-course records. Districts provided a mapping of their courses to the appropriate SCED code and/or state course code.[26]

Districts varied in how they reported students' grades and within some districts, some schools used letter grades, some used numeric grades, while still others had standards-based grades. Table A.10 shows the common types of grades students received by district and grade level. In students' core courses, most students received numeric grades on a scale of 0-100.

We converted all other grades to numeric by assigning a grade of 98 for an A+, 95 for an A, 91 for an A-, 88 for a B+, and so on, giving a score of 50 for an F. Nearly all standards-based grades, which were only used in elementary and middle schools, had a single 4-category holistic component meant to capture overall performance against the standards; we assigned the top category an "A" (95), the second a "B" (85), the third a "C" (75) and the lowest a "D" (65).

For each student in each school year, we calculated a single grade per course code. In most cases, this meant taking the mean grade across semesters, trimesters, or quarters. In standards-based grading contexts, the overall holistic grade was only provided at the end of the school year, and thus no averaging was needed.[27] Table A.10 shows the median grade earned in core courses by grade level and district.

TABLE A.10 | **TYPES OF GRADING SYSTEMS USED AND MEDIAN GRADE BY DISTRICT AND GRADE LEVEL IN 2015-2016 AND 2016-2017 CORE COURSES**

	District A	District B	District C[a]	District D[b]	District E
GRADES 3-5					
% Numeric grades	98%	0%	---	100%	100%
% Letter grades	2%	0%	---	0%	0%
% Standards-based grades	0%	100%	---	0%	0%
Median numeric grade	83	85	---	84	80
GRADES 6-8					
% Numeric grades	94%	0%	0%	100%	100%
% Letter grades	6%	51%	100%	0%	0%
% Standards-based grades	0%	49%	0%	0%	0%
Median numeric grade	78	83	78	81	87
GRADES 9-12					
% Numeric grades	93%	0%	0%	100%	100%
% Letter grades	7%	100%	100%	0%	0%
% Standards-based grades	0%	0%	0%	0%	0%
Median numeric grade	76	80	78	81	87

Only math, ELA, science, and social studies included. Median grade based on numeric imputation of letter and standards-based grades.

a: Only one school in District C provided historical grade information

b: District D provided both numeric and letter grades, but we used the numeric grades in all cases as they provided more precise information about a student's grade.

STUDENT STANDARDIZED TEST SCORES

Participating public school districts provided district-wide student-level test results from 2012-2013 through 2016-2017, including mandated state tests, college admissions test like the ACT and SAT (in some districts), and Advanced Placement results when available. The CMO schools in District C provided recent state test results on all students in just their school, not the surrounding district as well. Because participating districts were in different states, it was not always possible to directly compare test scores. To interpret results across our participating districts, we used three strategies to create a standardized measure of performance on state assessments:

- **We used each state's definition of meeting grade-level expectations.** All participating districts were in a state that converted raw test scores to four or five performance levels. In each case, one level represented the score needed to have been considered to meet the expectations for the grade and subject. Though what type of performance is required to meet this bar varies by state, we used states' own definitions to classify each test score in our data as having met the grade-level expectations or not. Some states' policies were more focused on how many students obtained the level just below meeting grade-level expectations (often some version of "approaching" expectations). However, we focused specifically on test scores that represented meeting grade-level expectations, regardless of which category in the state had more accountability factors tied to it.
- **Standardized test scores against all other students in the district.** For each school year, test subject, and test grade (or class for end-of-course exams) we standardized student-level results by converting raw scale-scores to the number of standard deviations away from their respective district mean. We used these district standardized scores for all analyses predicting student growth. Schools in the CMO did not have access to their student-level district data and so consequently we used the publicly available means from their surrounding districts and imputed the district-wide standard deviation (see below).
- **Standardized test scores against the average student in the state.** Because our participating districts varied in overall test performance, we used the same standardization approach, but compared each test score to the average score in the same grade, subject, and school year among all test takers in the respective state. This allowed us to situate students' performance against a broader sample and compare average performance across districts.[28] We did not have access to statewide student-level data, but some states provided us the necessary means and standard deviations to perform this standardization accurately. For others, we used publicly available state means and imputed the standard deviations by identifying the standard deviation that optimized the likelihood of getting the actual distribution of statewide performance categories given the fixed state mean, fixed scale-score range, and an underlying skew-normal distribution.[29]

We used raw AP and SAT/ACT scores more directly. For AP, we considered a score of 3 out 5 passing. Though some participating districts offered many AP courses, we focused exclusively on Calculus AB, Statistics, English Language and Composition, and English Literature and Composition, as these were some of the most common courses across all districts and matched the project's heavier emphasis on math and ELA. The results were qualitatively similar when we included additional core subject AP courses. District C did not provide AP test results and District E did not have AP courses.

For the ACT and SAT we used these tests' college readiness benchmarks, which are scores that have been empirically linked to having a 75% probability of earning at least a C in a first-year credit-bearing college course of the same subject.[30] Only districts B and E provided ACT or SAT test results.

Scores on both tests, therefore, were converted to binary outcomes: pass/fail for AP and ready/not ready for ACT/SAT.

DATA ON STUDENT CHARACTERISTICS

Data on student characteristics came from two sources:

1. All students who were eligible to complete student surveys were asked to self-report their race/ethnicity, gender, and whether a language other than English was primarily spoken at home on the one-time background survey. Though this data allowed us to connect demographics to individual students, it represents only grade 3-12 students and only those who completed the background survey.

2. Districts provided data on student race/ethnicity, English Language Learner (ELL) status, Special Education status, whether the student receives free or reduced-price lunch (FRL), and students' historical achievement records for all students in the district. This data was anonymous, however, so we could not connect district provided student data directly to individuals in our study. But districts did indicate for each student which class they belonged to, so we could connect district demographic data to the classroom measures of student experience we created.

We used both sets of demographic data to compare student characteristics to our measures of academic experience in our analyses.

DEFINING RACIAL/ETHNIC MATCH BETWEEN STUDENTS AND TEACHERS

We created two sets of variables representing the type of racial/ethnic match between teachers and students:

1. **Broad matches.** We classified classrooms by whether a) a majority of students were students of color and their teacher was a teacher of color, b) a majority of students were students of color and their teacher was white, and c) a majority of students were white. (We did not have a sufficient number of classrooms that were majority white and taught by a teacher of color (n = 5)).

2. **Specific matches.** We classified classrooms by whether a) a majority of students were students of color and the teacher had the same race/ethnicity as the majority of students in the class, b) a majority of students were students of color and the teacher did not have the same race/ethnicity as the majority of students in the class, and c) the majority of students were white. This definition differs from the previous one in that classes that are, for example, majority Black, only get classified as having a teacher racial/ethnic match if the teacher is also Black. In the previous definition, any teacher of color would count as a match.

ANALYSIS METHODS AND RESULTS

RESEARCH QUESTION 1: DESCRIBING STUDENTS' DAY-TO-DAY AND ACCUMULATED EXPERIENCES

Our first research question was entirely descriptive and most analyses simply tallied the proportion of experiences – assignments, lessons, surveyed experiences, etc. – that met our definitions of a grade appropriate or engaging. Table A.11 shows the distribution of metrics' ratings by district. Results in Table A.11 are not weighted by classroom, meaning classrooms that submitted more assignments or had more students completing surveys, for example, are represented more heavily. This differs from the approach we took to estimate the amount of time spent with high quality assignments, lessons, or engaged, which is detailed below.

TABLE A.11| **DISTRIBUTION OF METRICS BY DISTRICT - UNWEIGHTED**

	District A	District B	District C[a]	District D[b]	District E	ALL
ASSIGNMENTS & STUDENT WORK						
Percent grade appropriate	24%	26%	38%	10%	27%	25%
Mean raw score [a]	2.20	2.30	2.89	1.37	2.27	2.18
Mean standardized score [b]	-0.01	0.06	0.34	-0.28	-0.01	0.00
Overall success rate	69%	67%	67%	74%	74%	71%
Success rate on grade-level assignments	61%	61%	56%	54%	67%	62%
Assignment Sources						
State developed	29%	4%	10%	5%	4%	11%
District developed	9%	6%	5%	8%	3%	6%
District adopted	14%	38%	14%	14%	40%	26%
Self-made and new	17%	19%	38%	25%	17%	21%
Self-made and used before	10%	14%	15%	14%	7%	11%
Other teachers	3%	4%	3%	5%	4%	4%
Website	13%	9%	7%	20%	18%	14%
Other	5%	6%	8%	9%	7%	7%
INSTRUCTION						
Percent of lessons with strong instruction	11%	18%	24%	2%	28%	15%
Mean raw score [a]	0.96	1.15	1.24	0.65	1.38	1.05
Mean standardized score [b]	-0.24	0.15	0.18	-0.37	0.41	0.00
ENGAGEMENT						
Mean standardized score [b]	-0.10	-0.04	-0.22	0.08	0.17	0.00
Percent of responses in each category:						
Disengaged	12%	12%	13%	12%	9%	11%
Minimally engaged	34%	36%	37%	37%	26%	34%
Engaged	32%	32%	29%	34%	27%	31%
Highly engaged	21%	20%	21%	18%	37%	24%

a: Raw assignment score is on a scale of 0-6, raw observation scores are on a scale of 0-3, raw expectations and support scores are both on scales of 0-10

b: Assignments, instruction, expectations and support standardized against all other responses in the same subject and grade-band. Student surveys standardized against all other responses in the same grade. We accounted for assignments that were rated by multiple reviewers by assigning them a lesser weight, so that all assignments were weighted equally.

TABLE A.11| **DISTRIBUTION OF METRICS BY DISTRICT - UNWEIGHTED**

	District A	District B	District C[a]	District D[b]	District E	ALL
WORTH						
Mean standardized score[b]	-0.07	-0.03	-0.18	0.10	0.11	0.00
Percent of responses in each category:						
Not worthwhile	22%	19%	22%	17%	17%	19%
Minimally worthwhile	29%	33%	35%	32%	26%	31%
Worthwhile	29%	30%	26%	35%	25%	29%
Highly worthwhile	20%	17%	17%	17%	32%	21%
EXPECTATIONS						
Mean raw score[a]	3.78	5.55	6.04	4.82	4.22	4.79
Mean standardized score[b]	-0.51	0.39	0.68	-0.00	-0.31	0.00
Percent of responses in each category:						
Low	21%	4%	0%	15%	16%	12%
Moderately low	59%	39%	31%	42%	50%	45%
Moderately high	16%	28%	33%	26%	27%	25%
High	4%	29%	36%	18%	7%	17%
SUPPORT						
Mean raw score[a]	4.76	5.43	6.00	5.25	4.65	5.14
Mean standardized score[b]	-0.25	0.19	0.47	0.13	-0.33	0.00
Percent of responses in each category:						
Oppose	3%	0%	0%	0%	4%	2%
Moderately oppose	19%	15%	7%	16%	19%	16%
Moderately support	67%	64%	57%	60%	68%	64%
Support	11%	21%	36%	24%	9%	19%
GRADES AND TEST SCORES[c]						
Mean ELA test score relative to state	-0.69	-0.11	-0.33	-0.58	0.24	-0.39
Mean Math test score relative to state	-0.69	-0.10	-0.27	-0.59	0.37	-0.38
Percent of students meeting state test expectations:						
D or lower in class	2%	2%	8%	2%	7%	2%
C in class	4%	6%	22%	5%	13%	6%
B in class	18%	46%	55%	26%	48%	35%
A in class	53%	79%	87%	71%	84%	71%
HS COURSE TRAJECTORIES[d]						
Below standard	33%	12%	---	15%	0%	17%
Standard	31%	25%	---	9%	17%	20%
Mid-level	25%	45%	---	54%	83%	45%
Rigorous	11%	17%	---	23%	0%	17%

c: Math and ELA classes only. Grades 3-8 only. 2015-2016 and 2016-2017 school years. All students in district are included, not just participating classrooms.

d: The graduating classes of 2016 and 2017 were used in all districts, though District E also included the class of 2015. District E did not offer a third-year foreign language class and so no student had the opportunity to be classified in the Rigorous trajectory. All students in districts are included, not just participating classrooms. District C only provided grades for one of the three schools and not enough years of data to run a course trajectory analysis.

HOURS AND MONTHS SPENT ON GRADE APPROPRIATE, ENGAGING, OR WORTHWHILE EXPERIENCES

For each classroom meeting our minimum data requirements (see previous section for minima in each metric), we calculated the percent of values that met our definition for the experience. We then took the mean of these classroom-level values to obtain an overall value for our study.[31] This approach weights classrooms equally instead of giving increased weight to classrooms that submitted more assignments or collected more student surveys. This approach is meant to show the experiences students had in an average classroom. Table A.12 displays the means and percentages using this approach.

We converted the overall study percentages into hours of class time by assuming that a single class (or subject in a self-contained classroom) lasts one hour each day for a 180-day school year.[32] And we converted these percentages to months out of the school year by assuming a single school year is nine months long.

TABLE A.12 | AVERAGE CLASSROOM MEAN SCORE ON EACH METRIC BY DISTRICT

	District A	District B	District C	District D	District E	ALL
ASSIGNMENTS & STUDENT WORK [c]						
Percent grade appropriate	26%	30%	39%	10%	30%	26%
Mean raw score	2.33	2.55	2.95	1.43	2.48	2.29
Mean standardized score	0.04	0.16	0.37	-0.29	0.06	0.04
Overall success rate	65%	67%	67%	74%	72%	69%
Success rate on grade-level assignments	56%	60%	56%	48%	60%	57%
INSTRUCTION						
Percent of lessons with strong instruction	11%	18%	24%	2%	31%	16%
Mean raw score	0.97	1.18	1.27	0.62	0.44	1.06
Mean standardized score	-0.20	0.16	0.22	-0.41	0.40	-0.00
STUDENT SURVEYS						
Mean standardized engagement score	-0.03	0.05	-0.20	0.08	0.17	0.01
Percent engaged	54%	51%	49%	51%	63%	54%
Mean standardized worth score	-0.00	0.02	-0.12	0.07	0.10	0.02
Percent worthwhile	50%	48%	46%	50%	56%	50%
EXPECTATIONS						
Mean raw score	3.87	5.57	5.93	4.80	4.23	4.81
Mean standardized score	-0.52	0.40	0.68	-0.00	-0.31	0.01
Percent high expectations	22%	58%	67%	43%	34%	43%
SUPPORT						
Mean raw score	4.71	5.36	6.00	5.24	4.65	5.13
Mean standardized score	-0.25	0.20	0.54	0.13	-0.33	0.02
Percent high support	76%	86%	95%	83%	77%	83%

Assignments, instruction, expectations and support standardized against all other responses in the same subject and grade-band. Student surveys standardized against all other responses in the same grade. Raw assignment score is on a scale of 0-6, raw observation scores are on a scale of 0-3, raw expectations and support scores are both on scales of 0-10

STUDENTS' HIGH SCHOOL COURSE TRAJECTORIES

The four participating public school districts each provided at least four years of course records for all students in the district. We identified all students who attempted at least 2.5 credits each semester in a participating district high school for the four consecutive years prior to and including their 12th grade year. Because we did not have transcript-level data, we had to exclude students who moved into or out of the district during their high school years as we were unable to determine which courses they took elsewhere. This includes excluding students who dropped out of school.

Students who had four years of high school course-taking data were classified into one of four categories based on the number of credits they earned in specific subjects and classes. These categories are adapted from NAEP's 2009 High School Transcript Study, though we made three important changes.[33] First, because we had access to completed courses rather than full transcripts, we focused on whether students reached a certain level within a subject's hierarchy instead of only counting credits. For example, succeeding in Algebra II implies succeeding in Algebra I. This helps prevent us from misclassifying students who took key courses (like Algebra I) in middle school. Second, our mid-level trajectory requires Algebra II and NAEP requires only Algebra I or Algebra II. Third, NAEP required at least two of Biology, Chemistry, or Physics and all three courses in the mid-level and rigorous trajectories respectively. We found that several students completed Biology in 8th grade - outside the purview of our data window – and thus dropped the biology requirement, assuming that earning credit in chemistry or physics implied already completing biology. Table A.13 shows the requirements for each course trajectory.

TABLE A.13 | **MINIMUM REQUIREMENTS FOR EACH HIGH SCHOOL COURSE TRAJECTORY**

	STANDARD	MID-LEVEL	RIGOROUS
ELA	4 credits or credit in at least one senior ELA course (like English IV or AP English)	Same as standard trajectory	Same as standard trajectory
MATH	3 Credits	3 Credits and credit in Algebra II and Geometry	3 Credits and credit in Pre-Calculus or higher.
SCIENCE	3 Credits	3 Credits and credit in either Chemistry or Physics	3 Credits and credit in Chemistry and Physics
SOCIAL STUDIES	3 Credits	3 Credits	3 Credits
FOREIGN LANGUAGE	0 Credits	1 Credit	3 Credits or credit in a level III course or higher

RESEARCH QUESTION 2: EXPLAINING THE ASSOCIATIONS BETWEEN STUDENT CHARACTERISTICS AND ACADEMIC EXPERIENCES

CONNECTING MEASURES OF CLASSROOM EXPERIENCE TO STUDENT AND CLASSROOM CHARACTERISTICS

We wanted to know whether certain types of students tended to have more access to better rated academic experiences.

Except for the student survey constructs ("engagement" and "worth"), our metrics of interest cannot be disaggregated by student and are instead measured at the classroom level. We therefore used a series of simple linear regression models predicting the classroom metrics defined in the previous section.

Model 1
Model 1 is a simple linear regression that includes sets of classroom and teacher variables

$$y_{ctd}=X_{ctd}\beta+W_{td}\delta+\varepsilon_{ctd}$$

Where y_{ctd} represents the outcome for classroom c, of teacher t, in district d. X_{ctd} represents a vector of classroom characteristics that include subject area, grade-level, course-type (i.e., "advanced", "normal", or "remedial"), percent of students in the class who were female, percent who received free or reduced lunch, percent of students who were white, an indicator if more than 25% of a classroom's students were English Language Learners, and an indicator if more than 25% of a classroom's students had an Individualized Education Plan (IEP). W_{td} is a vector of indicator variables representing different ranges of teacher experience.

Model 2
Model 2 is identical to Model 1 except we added a district fixed effect, α_d:

$$y_{ctd}=X_{ctd}\beta+W_{td}\delta+a_d+\varepsilon_{ctd}$$

This addition accounts for the overall effect of a district on a given metric outcome. Including a_d allows us to look within districts to see whether, after accounting for districts' overall connections to the metric, the various classroom and teacher characteristics were associated with the outcome of interest.

Model 3
Model 3 is identical to Model 1 except that we removed the course-type variables (i.e., "advanced", "normal", or "remedial"), from the vector of classroom characteristics and replaced it with the classroom mean of students' prior math and ELA achievement standardized against the average student in the state. In order to maximize the number of classrooms in these models we used all years of data provided by districts to identify the most recent prior year end-of-grade test score in math and in ELA. In some cases, like a 6th grade classroom, this is the previous academic year, but for other classrooms, like a 10th grade math classroom, this could be two or more years ago. In all cases, we separately standardized the prior math and ELA test scores against the average student in the state in the same academic year, and then averaged these two values together. Because only grade 4-9 students and some grade 10-12 students had a recent test score in math and ELA, Model 3 is based on a smaller sample size.

Model 4
Model 4 is actually a set of models where we ran a linear regression with only one classroom or teacher characteristic variable as well as controls for grade and subject area. Because some of the classroom metrics are highly correlated—for example, the correlation between percent students of color and percent receiving FRL is 0.87—we wanted to also show the associations between each outcome and each demographic variable separately.

For all models, only classrooms with at least 10 students with each demographic variable in the model are included. Because of some co-teaching situations, we used weighted least squares regression in all models, with weights based on how many teachers worked in the classroom. This ensured each classroom was equally weighted in the model. To make the coefficients more interpretable, we standardized all classroom outcome metrics at the classroom-level. Reported coefficients from these models therefore represent changes in classroom-level standard-deviation units.

Results
Table A.14 shows the coefficient on each variable for each of our models. Below we briefly summarize the findings:

- **Subject area and grade level.** There were stark differences between subjects across all four metrics. Whereas math and ELA classes tended to earn better assignment and observation scores, students tended to be more engaged in science and social studies. There were fewer difference by grade level, except for observations where observers tended to score 3-5 classes most highly, and K-2 teachers had higher expectations than teachers of classrooms with older students.
- **Course level.** Advanced courses were often similar to non-advanced courses, but remedial courses tended to receive lower-rated assignments, lessons, and had teachers with lower expectations. Students in remedial courses did tend to, however, have higher levels of engagement.
- **Student demographics.** The gender composition of a classroom was weakly connected to all metrics except instruction, where classrooms with more female students tended to have higher rated lessons. Students' free-reduced lunch eligibility was consistently connected to lower ratings across all four metrics, though this negative relationships was sometimes smaller when we controlled for recent prior achievement. Classrooms with a 25-percentage point increase in its FRL student population tended to receive metric scores that were a tenth to a half a standard deviation lower depending on the model. Likewise, classrooms with more students of color also tended to have lower ratings on assignments and instruction, as well as lower engagement, but these associations mostly disappeared, or in some cases became positive when we controlled for FRL status as well. There was a strong relationship between the

percent of students in a class receiving FRL and the percent who were students of color, so the negative results on the SOC variable in Model 4 clearly indicate that students of color, overall, tended to receive worse opportunities on our metrics. Yet, Models 1-3 also show that when comparing classrooms with similar proportions of FRL students – a broad proxy for family income – having more students of color was not additionally associated with lower quality opportunities. The proportion of students of color in a classroom had no association with expectations overall (Model 4), and when controlling for other variables tended to be positively associated.

The associations with other demographic variables were relatively small. Classrooms with at least a quarter ELL students tended to have better assignments and instruction once other variables were controlled. When pooling districts together, these classrooms also tended to have higher expectations. Classrooms with at least a quarter of students with an IEP had on average significantly lower teacher expectations, but higher engagement. The relationships between the proportion of students receiving special education and average assignment and instructional quality were closer to zero and never significant.

- **Student achievement.** Classrooms with initially higher performing students tended to get better assignments, better instruction, were more engaged, and had teachers with significantly higher expectations. Though the relationship between recent prior achievement and classroom metrics were closer to zero when we also controlled for other student demographic variables, on its own (Model 4) prior achievement was significantly positively associated with all metrics.
- **Teacher experience.** Teachers with at least 10 years of teaching experience tended to have better-rated assignments, instruction, and higher expectations. However, students were more engaged in classrooms with a teacher in their first five years.

USING INDIVIDUAL STUDENT CHARACTERISTICS TO MODEL ENGAGEMENT

Our student survey process allowed us to further connect student engagement to individual students and their characteristics. Because students completed surveys multiple times during our study, we used a multi-level model with responses clustered within students clustered within classrooms to predict engagement:

Response Level: $y_{rsc} = \pi_{0sc} + e_{rsc}$

Student Level: $\pi_{0sc} = \beta_{00c} + \beta_{01c}\,(Gender)_{sc} + \beta_{02c}\,(Race/Ethnicity)_{sc} + \beta_{03c}\,(Language\ Status)_{sc} + r_{0sc}$

Classroom level: $\beta_{00c} = \gamma_{000} + \gamma_{001}\,(Subject)_{c} + \gamma_{002}\,(High\ IEP)_{c} + u_{00c}$

Where y_{rsc} represents the engagement score (standardized by grade level) for survey response r, from student s, in classroom c. In addition to the three student-level variables derived from the student survey, we also controlled class subject and an indicator for a high population of students with IEPs. We additionally ran the same model with district fixed effects (Model 2) and with a control for recent prior achievement (Model 3).

Table A.15 shows the results for these three models on the student-level variables – the classroom-level coefficients are suppressed for parsimony. We found that compared to white students, students of all other race/ethnicities tended to have lower engagement, though results were most strongly and consistently negative for students who categorized themselves into multiple races or into a race/ethnicity not listed in the table. Female students tended to be more engaged than males[34] and students where another language besides English was primarily spoken at home also tended to be more engaged.

TABLE A.14 | **REGRESSION ESTIMATES FROM MODELS PREDICTING STANDARDIZED CLASSROOM-LEVEL OUTCOMES**

	ASSIGNMENTS				LESSONS			
MODELS	1	2	3	4	1	2	3	4
SUBJECT (COMPARED TO ELA)								
Math	0.35***	0.34***	0.25		0.03	0.05	-0.21*	
Science	-0.95***	-0.85***	-0.88***		-1.19***	-1.14***	-1.26 ***	
Social Studies	-0.73***	-0.74***	-0.74***		-1.08***	-1.07***	-1.16***	
GRADE-LEVEL (COMPARED TO K-2 OR 3-5 IN MODEL 3)								
3-5	0.13	0.12	---		0.29 *	0.27 *	---	
6-8	0.13	0.22	0.11		-0.19	-0.19	-0.50 ***	
9-12	0.16	0.12	0.16		-0.26	-0.32*	-0.48 ***	
COURSE TYPE [a]								
Advanced course	-0.05	0.12	---	- 0.07	-0.02	0.06	---	0.06
Remedial course	-0.28 *	-0.36 **	---	-0.31*	-0.16	-0.19	---	-0.28 *
STUDENT DEMOGRAPHICS								
Percent female [b]	-0.01	-0.01	-0.02	-0.03	0.09*	0.07 *	0.06	0.08 *
Percent FRL [c]	-0.28**	-0.21*	-0.32**	-0.21***	-0.15*	-0.20*	-0.04	-0.23 ***
Percent SOC [c]	0.03	0.05	0.08	-0.12***	-0.09	0.05	-0.13	-0.17 ***
> 25% ELL	0.28 *	0.17	0.21	-0.00	0.14	0.11	0.11	-0.12
> 25% Special Ed	0.07	0.11	-0.00	-0.01	0.03	-0.02	0.19	-0.05
PRIOR ACHIEVEMENT [d]								
Mean math and ELA test score	---	---	0.01	0.17*	---	---	0.15	0.28 ***
YEARS TEACHING (COMPARED TO 1ST – 4TH YEAR TEACHERS)								
5th – 9th Year	-0.23	-0.30*	-0.07	-0.18	-0.02	-0.02	0.02	-0.05
10th+ Year	0.16	0.08	0.29 *	0.12	0.07	0.05	0.14	0.04
R^2	0.33	0.39	0.31		0.49	0.50	0.50	
N	351	351	244		349	349	246	

TABLE A.14 CONTINUED

	ENGAGEMENT				EXPECTATIONS			
MODELS	1	2	3	4	1	2	3	4
SUBJECT (COMPARED TO ELA)								
Math	-0.25 *	-0.22	-0.35 *		-0.13	-0.13	-0.12	
Science	0.33	0.31	0.32		0.27	0.21	0.26	
Social Studies	0.49 *	0.53 **	0.47 *		-0.05	-0.00	-0.02	
GRADE-LEVEL (COMPARED TO 3-5 FOR ENGAGEMENT AND IN MODEL 3; COMPARED TO K-2 OTHERWISE)								
3-5	---	---	---		-0.64***	-0.54***	---	
6-8	-0.03	-0.08	-0.00		-0.49**	-0.52**	0.06	
9-12	-0.05	-0.09	0.01		-0.19	-0.12	0.48**	
COURSE TYPE [a]								
Advanced course	-0.05	-0.13	---	0.02	0.10	-0.03	---	0.15
Remedial course	0.16	0.22	---	0.23	-0.40**	-0.40**	---	-0.51***
STUDENT DEMOGRAPHICS								
Percent female [b]	-0.01	-0.05	-0.02	-0.06	0.01	0.02	-0.01	0.03
Percent FRL [c]	-0.19	-0.49***	-0.18	-0.16**	-0.35***	-0.02	-0.14	-0.14***
Percent SOC [c]	0.04	0.33 *	0.14	-0.14**	0.19**	-0.00	0.19*	-0.01
> 25% ELL	-0.20	-0.00	-0.20	-0.32*	0.23*	-0.08	0.32*	0.02
> 25% Special Ed	0.32	0.30	0.46*	0.28	-0.29*	-0.31	0.01	-0.46**
RECENT PRIOR ACHIEVEMENT [d]								
Mean math and ELA test score	---	---	0.14	0.17 *	---	---	0.46 ***	0.29 ***
YEARS TEACHING (COMPARED TO 1ST – 4TH YEAR TEACHERS)								
5th – 9th Year	-0.26	-0.19	-0.13	-0.42 **	0.13	0.17	0.12	0.20
10th+ Year	-0.20	-0.06	-0.01	-0.29 *	0.17	0.30 **	0.42 **	0.15
R^2	0.13	0.21	0.15		0.21	0.32	0.21	
N	294	294	242		386	386	272	

Model 1 is a weighted least squares model including all variables in table, with weights accounting for co-teaching situations so that each classroom is weighted equally. Model 2 adds district fixed effects. Model 3 adds a control for recent prior achievement. Model set 4 represents the coefficients from separate WLS regressions with only the variable listed in the row and controls for subject and grade level. For all models, classroom-level outcomes have been standardized against all classrooms with sufficient sample sizes. Thus, estimated coefficients represent associated change in standard deviation units. *, **, and *** represent estimates significantly different from zero at the 0.05, 0.01, and 0.001 levels respectively. All student demographic information provided by participating districts. Only classrooms with at least 10 students with each district demographic variable in model are included. One school in District C only provided student race data and so their classrooms are only included in the percent SOC results for model set 4.

a: Advanced courses are courses teachers indicated were AP, Honors, or Dual-Enrollment. Remedial courses are courses teachers labeled Remedial or Intervention.

b: Values represent estimated change associated with a 10-percentage point increase in the proportion of female students.

c: Values represent estimated change associated with a 25-percentage point increase in the proportion of students of color (SOC) or students receiving free or reduced-price lunch (FRL). Students of color are any student not categorized as white, including multi-racial students.

d: Recent prior achievement was calculated by taking, for each student, the mean of their most recent math and ELA grade 3-8 test scores (standardized against the state), and then computing a classroom mean.

TABLE A.15 | **MULTI-LEVEL MODEL PREDICTING ENGAGEMENT**

MODELS	1	2	3
STUDENT RACE/ETHNICITY (COMPARED TO A WHITE STUDENT)			
Asian	-0.09	-0.04	-0.08
Black	-0.10 *	-0.04	-0.03
Latinx	-0.10 *	-0.08	-0.03
Other	-0.28 ***	-0.23 ***	-0.12
Multiple races	-0.16 ***	-0.13 **	-0.12 *
STUDENT GENDER (COMPARED TO MALE STUDENTS)			
Female	0.05 *	0.05 *	0.02
STUDENT LANGUAGE			
Other language besides English primarily spoken at home	0.07 *	0.08 *	0.08 *
N (Responses)	20,244	20,244	16,051
N (Students)	3,439	3,439	2,856
N (Classes)	322	322	266

Model 1 is a multi-level model including all variables in table plus controls for subject, an indicator for whether at least 25% of students in the class had an IEP. Model 2 adds district fixed effects. Model 3 adds a control for recent prior achievement. Recent prior achievement was calculated by taking, for each student, the mean of their most recent math and ELA test scores (standardized against the state), and then computing a classroom mean.

CONNECTING THE MATCH BETWEEN STUDENT AND TEACHER RACE/ETHNICITY TO CLASSROOM OUTCOMES

To explore the role of student and teacher race/ethnicity matching on our different classroom metrics, we repeated all four models from Table A.14 but separately replaced the percent SOC variable with our two sets of matching indicators (see earlier in the Technical Appendix for these definitions). Table A.16 shows the coefficients just on these match variables, though the models contain the same controls as Table A.14

Classrooms that were mostly composed of students of color had similarly rated assignments and lessons, regardless of their teacher's race/ethnicity, especially when we accounted for district effects. But when these classrooms were taught by a teacher of color, students were significantly more engaged, and teachers had significantly higher expectations about students meeting the standards than when taught by a white teacher.

THE RELATIONSHIP BETWEEN EXPECTATIONS AND OTHER CLASSROOM METRICS

We also wanted to know the extent to which teachers' expectations for students' success on the standards was associated with the types of assignments they used or the instruction they provided. Table A.17 shows the coefficient on the standardized expectations variable when we added it to the same four models used previously (i.e., Table A.14). On its own (Model 4), expectations were significantly associated with both assignments and instruction. Though the association remained positive in all models, some of the relationship between expectations and instruction and assignments was explained by student characteristics, and the coefficients in Models 1-3 were smaller.

TABLE A.16 | **REGRESSION ESTIMATES FOR TEACHER MATCH VARIABLES FROM MODELS PREDICTING STANDARDIZED CLASSROOM-LEVEL OUTCOMES**

	ASSIGNMENTS				INSTRUCTION			
MODELS	1	2	3	4	1	2	3	4
BROAD MATCH COMPARED TO CLASSROOMS WITH WHITE TEACHER AND MAJORITY STUDENTS OF COLOR								
Teacher of color and majority students of color	-0.13	0.22	-0.27	-0.18	0.07	0.24	-0.11	0.03
All classrooms with majority white students	-0.18	-0.04	-0.29	0.29 **	0.16	-0.15	0.13	0.53 ***
R^2	0.36	0.42	0.37	0.26	0.49	0.51	0.50	0.46
SPECIFIC MATCH: COMPARED TO CLASSROOMS WITH MAJORITY STUDENTS OF COLOR BUT TEACHER IS NOT THE SAME RACE/ETHNICITY								
Majority students of color and teacher is the same race as majority of students	-0.13	0.11	-0.23	-0.12	-0.00	0.08	-0.20	0.02
All classrooms with majority white students	-0.20	-0.02	-0.33	0.31 **	0.14	-0.14	0.07	0.53***
R^2	0.36	0.41	0.36	0.26	0.49	0.50	0.51	0.46
N	337	337	233	352	331	331	232	349
	ENGAGEMENT				EXPECTATIONS			
BROAD MATCH COMPARED TO CLASSROOMS WITH WHITE TEACHER AND MAJORITY STUDENTS OF COLOR								
Teacher of color and majority students of color	0.44 **	0.28	0.34 *	0.45 **	0.60 ***	0.55 ***	0.67 ***	0.41**
All classrooms with majority white students	0.14	-0.37	-0.12	0.56 ***	-0.24	0.15	-0.33	0.15
R^2	0.15	0.20	0.17	0.11	0.26	0.35	0.31	0.11
SPECIFIC MATCH: COMPARED TO CLASSROOMS WITH MAJORITY STUDENTS OF COLOR BUT TEACHER IS NOT THE SAME RACE/ETHNICITY								
Majority students of color and teacher is the same race as majority of students	0.44 **	0.21	0.34 *	0.54 **	0.49 ***	0.46 ***	0.43 **	0.43 **
All classrooms with majority white students	0.22	-0.31	-0.06	0.53 ***	-0.19	0.26	-0.32	0.12
R^2	0.15	0.19	0.17	0.12	0.23	0.34	0.26	0.11
N	281	281	231	293	367	367	257	385

Model 1 is a weighted least squares model including all variables in table, with weights accounting for co-teaching situations so that each classroom is weighted equally. Model 2 adds district fixed effects. Model 3 adds a control for recent prior achievement. Model set 4 represents the coefficients from separate WLS regressions with only the variable listed in the row and controls for subject and grade level. For all models, classroom-level outcomes have been standardized against all classrooms with sufficient sample sizes. Thus, estimated coefficients represent associated change in standard deviation units. *, **, and *** represent estimates significantly different from zero at the 0.05, 0.01, and 0.001 levels respectively. All student demographic information provided by participating districts. Only classrooms with at least 10 students with each district demographic variable in model are included. One school in District C only provided student race data and thus their classrooms are only included in model set 4. See text for other variables in model but not shown in table.

TABLE A.17 | **REGRESSION ESTIMATES FOR TEACHER EXPECTATION VARIABLE PREDICTING STANDARDIZED CLASSROOM-LEVEL OUTCOMES**

	ASSIGNMENTS				INSTRUCTION			
MODELS	1	2	3	4	1	2	3	4
Expectations (standardized)	0.04	0.07	0.06	0.14 **	0.06	0.07	0.05	0.11 **
R^2	0.33	0.39	0.31	0.23	0.49	0.50	0.50	0.42
N	351	351	244	368	349	349	246	369

Model 1 is a weighted least squares model including all variables in table, with weights accounting for co-teaching situations so that each classroom is weighted equally. Model 2 adds district fixed effects. Model 3 adds a control for recent prior achievement. Model set 4 represents the coefficients from separate WLS regressions with only the variable listed in the row and controls for subject and grade level. For all models, classroom-level outcomes have been standardized against all classrooms with sufficient sample sizes. Thus, estimated coefficients represent associated change in standard deviation units. *, **, and *** represent estimates significantly different from zero at the 0.05, 0.01, and 0.001 levels respectively. All student demographic information provided by participating districts. Only classrooms with at least 10 students with each district demographic variable in model are included. One school in District C only provided student race data and thus their classrooms are only included in Model set 4. See text for other variables in model but not shown in table.

CONNECTING STUDENT CHARACTERISTICS TO STUDENT SUCCESS ON ASSIGNMENTS

Table A.18 shows how successful different types of classrooms were on their assignments and in demonstrating the demands of the standards through their assignments. Because our primary means of connecting demographic information to students was at the classroom level, the first panel of Table A.18 shows average classroom success rates. To estimate the effect of a given demographic variable on classroom success, we used separate linear regression models predicting classroom success rates given demographic classifications and also controlling for grade level and subject. These models are similar to Model 4 from the previous section but with classroom success rate as the outcome of interest.

We restricted all analysis in Table A.18 to classrooms that provided at least five days of assignments and submitted at least five samples of student work. Though this latter analysis rule adds no further restrictions when examining success (on the assignment or against the standards) on all assignments (i.e., the first two sets of columns in Table A.18), it does substantially restrict which classrooms are included when we compare success on grade-level assignments. Many classrooms never provided students a grade-level assignment – i.e., an assignment that earned the highest rating on the "Content" domain – and this was especially true for classrooms with mostly students of color and classrooms with mostly students receiving free or reduced-price lunch. Consequently, the values in the third pair of columns in Table A.18 are based only on a fraction of classrooms participating in the study, and it's difficult to compare these rates directly to the other values in the table.

For most characteristics, overall classroom assignment success rates did not differ significantly by demographic, except for free-reduced lunch status: students in classrooms with fewer free and reduced lunch eligible students tended to have significantly more success on their assignments. This difference was exacerbated when we compared classrooms' rates of demonstrating the demands of the standards. Classrooms that began the year higher achieving also tended to have more success on assignments and against the standards. Classrooms with fewer students of color also tended to demonstrate the standards more frequently.

Most 3rd-12th grade students told us about their racial/ethnic background, gender, and whether another language was primarily spoken at home when they completed the student surveys. The second panel of Table A.18 shows the success rates by individual characteristics, not classrooms, for those who completed student surveys. We used a multi-level linear probability regression model with work samples nested within students nested within classrooms to estimate the effect of individual characteristics on the probability of success:

Work Sample Level: $y_{rsc} = \pi_{0sc} + e_{rsc}$

Student Level: $\pi_{0sc} = \beta_{00c} + \beta_{01c}\,(Demographic\ Variable)_{sc} + r_{0sc}$

Classroom level: $\beta_{00c} = \gamma_{000} + \gamma_{001}\,(Subject)_c + \gamma_{002}\,(Grade\ Level)_c + u_{00c}$

TABLE A.18 | **AVERAGE CLASSROOM SUCCESS RATES AND REGRESSION ESTIMATES FOR EFFECT OF DEMOGRAPHIC VARIABLES ON CLASSROOM SUCCESS RATES**

	Successfully completed the assignment		Successfully demonstrated the standard among all assignments		Successfully demonstrated the standard among grade-level assignments	
CLASSROOM CHARACTERISTICS	Average Class Success Rate	Estimated Effect[a]	Average Class Success Rate	Estimated Effect	Average Class Success Rate	Estimated Effect
Gender						
> 65% Female	73%	0.48	9%	-1.51	55%	2.92
< 35% Female	70%	2.15	14%	1.75	50%	2.60
35-65% Female	69%		16%		54%	
FRL						
> 75% FRL	70%	0.07	12%	-1.77	53%	-1.21
< 25% FRL	77%	7.27 *	33%	14.55 ***	69%	11.17 *
25-75% FRL	67%		13%		49%	
Race/Ethnicity						
> 50% Students of color	68%	-3.66	13%	-4.97 **	51%	-5.07
> 50% White students	71%		15%		54%	
ELL						
> 25% ELL	66%	-3.99	13%	-3.10	51%	-3.70
< 25% ELL	70%		15%		56%	
Special Education						
> 25% with IEP	69%	0.90	17%	0.91	58%	4.97
< 25% with IEP	70%		15%		53%	
Recent Prior Achievement						
Mean prior < −0.25 SDs	66%	-1.12	9%	-1.84	45%	-2.11
Mean prior > 0.25 SDs	72%	4.97	12%	2.44	52%	4.71
Mean prior within 0.25 SDs	68%		10%		47%	
STUDENT CHARACTERISTICS (Grades 3-12 Only)	Average Class Success Rate	Estimated Effect[a]	Average Class Success Rate	Estimated Effect	Average Class Success Rate	Estimated Effect
Gender						
Female	70%	3.82 ***	14%	1.37 **	55%	5.47 ***
Male	66%		12%		49%	
Race/Ethnicity						
Asian	69%	-0.22	10%	-0.81	50%	-2.24
Black	65%	5.86 ***	14%	-1.24	51%	-3.31
Latinx	70%	-3.30 *	7%	-1.69	49%	-2.16
Other	66%	-7.82 ***	13%	-1.80	45%	-6.45
Multiple races	70%	-1.14	15%	0.19	57%	1.38
White	70%		15%		54%	
Language at Home						
Second language at home	69%	-0.21	11%	-0.24	51%	-0.25
No second language	68%		15%		52%	

*, **, and *** represent estimates significantly different from zero at the 0.05, 0.01, and 0.001 levels respectively. Grade-level assignments are those with the highest possible rating on the content domain. Estimated effects based on OLS regression model predicting classroom success rate with controls for class subject and grade level. Each set of demographic variables was modeled separately. Blank cells in the table were the comparison groups in the OLS models. Estimated odds ratios based on multi-level logistic regression with work samples nested within students nested within classrooms. Models also control for classroom subject and grade level. Only grade 3-12 students who completed student surveys are included in multi-level models.

Only core academic subjects included. Student success rate in second panel represents the success rate for each unique student-by-class combination. For analyses on classroom characteristics, only classrooms with at least 5 days of submitted assignments and at least 5 samples of work that meets the column's requirements are included. For analyses on student characteristics, only students with at least 5 samples of student work submitted on their behalf are included. Because of the low rate of grade-level assignments given to students, all students with at least 5 total student work submissions and at least 1 grade-level assignment were included in the success rate on grade-level assignments.

a: Estimated effects represented in percentage points. For example, an estimated effect of 0.48 represents 0.48 percentage points, not 48%.

Where y_{rsc} is a binary variable with 1 representing success and 0 representing no success.[35] This model was run separately for each of the three sets of demographic variables available in the student surveys: gender, race/ethnicity, and home language status.

From these results, we found that female students tended to have more success than males, and students of color tended to have less success on all assignments than white students, but for many groups of students, success on grade-level assignments was more equal.

CONNECTING STUDENT GRADES AND TEST SCORES TO STUDENTS CHARACTERISTICS

Table A.19 shows three sets of analysis connecting student characteristics to course-taking and course grades. The first panel shows the proportion of students in the top two high school course-taking trajectories. The second panel compares the average student grade earned in core academic classes by student characteristic.

The third panel shows estimates from an OLS model predicting state test scores in grades 3-8 math and ELA (standardized among all students in the district) controlling for students' course grades, demographic characteristic, the interaction between course grades and the demographic, and additional controls for subject and grade-level:

Table A.19 shows that in nearly all of our participating districts, students of color, low-income, English Language Learners, and students with IEPs were significantly less likely to have taken a Mid-Level or Rigorous series of courses, earned significantly lower grades in their core classes, received significantly lower test scores for earning the same course grade, and tended to be in classes that had grading scales less aligned to state tests – compared to their comparison groups, their test scores improved significantly less for improving their course grades. District E was one exception, especially for differences by race: white and Black students had similar course patterns, course grades, and grade-test score connections.

$$y_{ik}=\beta_0+\beta_1\,(\textit{Course Grade-85})_{ik}+\beta_2\,(\textit{Demographic})_{ik}+\beta_3*(\textit{Course Grade-85})_{ik}*\textit{Demographic}_{ik}$$

Where y_{ik} represents the standardized test score for student i in subject k. We centered students' grades on 85 (out of 100). The estimates for β_2 represent the average difference in test scores between, for example, a B student receiving free or reduced-price lunch and a student earning the same grade but not receiving FRPL. By including an interaction term, we also estimated the extent to which the alignment between students' course grades and test results varied by student group. We rescaled course grades so that coefficients on the interaction effect (β_3, represented by the "slope" column in the table) represent how much more or less test scores tended to change for every 10 points by which a student increased their course grade. We ran the above models separately for each district.

TABLE A.19 | STUDENT COURSE GRADES AND TRAJECTORIES BY DISTRICT AND STUDENT DEMOGRAPHICS

	District A		District B		District C [a]		District D		District E	
	Mid-Level	Rigorous	Mid-Level	Rigorous	Mid-Level	Rigorous	Mid-Level	Rigorous	Mid-Level	Rigorous
Percent of students in each 4-year high school course trajectory [b]										
Asian	22%	15% **	50%	26%	---	---	---	---	---	---
Black	24%	9% ***	48%	8% ***	---	---	54%	12%	81%	0%
Latinx	20%	8% ***	49%	8% ***	---	---	54%	23%	---	---
Other	29%	10%	43%	22% ***			---	---		
White	31%	19%	36%	40%	---	---	55%	20%	83%	0%
Receives FRL	22%	8% ***	49%	6% ***	---	---	52%	21% ***	81%	0%
Does not receive FRL	32%	18%	41%	31%	---	---	55%	31%	83%	0%
ELL	---	---	50%	8% ***	---	---	53%	14% ***	---	---
Non-ELL	25%	11%	42%	24%	---	---	54%	24%	83%	---
Has IEP	8%	1% ***	23%	1% ***	---	---	38%	1% ***	---	---
Does not have IEP	28%	13%	48%	19%	---	---	56%	25%	90%	0%
Mean numeric grade (0-100) in core courses by student demographic and district [c]										
Asian	83.1 ***		83.9 ***		80.9		87.3 ***		---	
Black	75.5 ***		77.6 ***		74.1 ***		80.7 ***		74.0	
Latinx	76.4 ***		78.6 ***		75.5 ***		81.6 ***		72.0	
Other	79.1 ***		80.7 ***		83.4		83.0		---	
White	80.2		84.7		80.1		83.3		75.2	
Receives FRL	76.7 ***		78.4 ***		73.8 ***		81.3 ***		74.3	
Does not receive FRL	80.5		83.8		80.3		84.2		75.4	
ELL	77.3 **		79.2 ***		72.0 **		81.0 ***		72.6	
Non-ELL	77.5		81.0		76.7		81.7		75.5	
Has IEP	74.3 ***		74.9 ***		70.6 ***		79.1 ***		69.2 ***	
Does not have IEP	78.3		81.0		77.4		81.9		76.3	
Estimated differences in district grade 3-8 Math and ELA standardized test scores for a B student, and differences in in the alignment (slope) between grades and test scores by student demographic [d]										
	B grade	Slope	B grade	Slope	B grade	Slope	B grade	Slope	B grade	Slope
Asian	-0.39 ***	0.04 *	-0.28 ***	-0.05 ***	---	---	-0.28 ***	0.28 ***	---	---
Black	-0.50 ***	-0.16 ***	-0.80 ***	-0.29 ***	-0.59 ***	-0.17	-0.38 ***	-0.35 ***	-0.09	0.26 **
Latinx	-0.57 ***	-0.15 ***	-0.74 ***	-0.29 ***	-0.43 ***	-0.10	-0.29 ***	-0.24 ***	---	---
Other	-0.11 ***	-0.04 *	-0.25 ***	-0.09 ***	---	---	-0.17 ***	-0.24 ***	---	---
Receives FRL	-0.51 ***	-0.17 ***	-0.73 ***	-0.29 ***	-0.29 ***	-0.20 ***	-0.23 ***	-0.16 ***	-0.14 **	-0.06
ELL	-0.88 ***	-0.25 ***	-0.44 ***	-0.19 ***	---	---	-0.14 ***	- 0.00	---	---
Has IEP	-0.71 ***	-0.23 ***	-0.97 ***	-0.33 ***	-1.54 ***	-0.46 ***	-0.70 ***	-0.30 ***	-0.78 ***	-0.06

*, **, and *** represent statistical significant differences at the 0.05, 0.01, and 0.001 levels respectively. In all panels, each race variable was tested against white students, students who received FRL were tested against those who did not, etc. Only combinations with at least 25 records were shown.

a: District C only provided grades for one of the three schools and not enough years of data to run a course trajectory analysis.

b: Significance tests represent 2-sample test for equality of proportions of the proportion of students in the mid-level or rigorous trajectory, even though the asterisks were only marked in the rigorous column. The graduating classes of 2016 and 2017 were used in all districts, though District E also included the class of 2015. District E did not offer a third-year foreign language class and so no student had the opportunity to be classified in the Rigorous trajectory. Only demographic groups with at least 25 students with classified trajectories are displayed.

c: Statistical tests represent simple t-tests with the comparison group. No other controls were included. Math, ELA, Science, and Social studies classes in grades 3-12 included. Across all district, both the 2015-2016 and 2016-2017 school years included. Only demographic groups with at least 25 students with classified trajectories are displayed.

d: Estimated obtained from a simple linear model controlling for students' course grade (centered on a grade of 85), an indicator for the given demographic and their interaction. We also included controls for subject and grade-level. All test scores were standardized against the district average. The course grade variable was re-scaled so that the slope estimate represents the association with a change in grade of 10 points (on a 0-100 scale). Math and ELA courses included only. Across all district, both the 2015-2016 and 2016-2017 school years included.

RESEARCH QUESTION 3: PREDICTING STUDENT OUTCOMES

We used our classroom-level measures of students' academic experiences as predictors of two student outcomes: state test scores and student-reported engagement. Though we consider student engagement a measure of academic experience – and thus an input to predict student test results – we wanted to explore the extent to which teachers' assignments, instruction, and expectations were associated with student perceptions.

PREDICTING STATE STANDARDIZED TEST RESULTS

Following the work of Kane & Staiger (2012),[36] we sought to estimate the correlation between our classroom metrics and the extent to which students' test scores in the year of our study were better or worse than expected given how students had scored on prior state tests. Commonly known as "value-added," we estimated a value for each math and ELA classroom in our participating districts that represented the mean difference between how students in the class actually scored on their end-of-year state tests and how other students in the district with similar demographic characteristics and similar prior test scores tended to score.

Specifically, using every student in each participating district, we used an OLS regression model to predict students' subject specific end of year test result (standardized against all other students in the district): $y_{iskt} = X_i\beta + \beta * f(y_{i,math,t-}) + \beta * f(y_{i,ELA,t-1}) + \bar{X}_{kt}\,\delta + \delta * f(\bar{y}_{math,k,t-1}) + \delta * f(\bar{y}_{ELA,k,t-1}) + \varepsilon_{iskt}$

Where y_{iskt} represents the test score of student i in subject s in school k, in year t. To predict these values, we controlled for students' prior year test scores in both math ($y_{i,math,t-1}$) and ELA ($y_{i,ELA,t-1}$), entering them both into the model with cubic polynomial functions (f()) to account for potential non-linear relationships between prior and current test scores.[37] We also included a vector of student characteristics (X_i), which included students' FRL status, race/ethnicity, gender, ELL status, and IEP status. Finally, we included school means of all demographic and prior achievement variables (e.g., $\bar{X}_{kt}$).

We ran the above model for all grades and courses in a district where students had prior year test results. This includes all grade 4-8 classes, but also some high school courses in states that had End-of-Course (EOC) exams. We constructed separate models for each district and for different combinations of prior achievement information. For example, some students who took the Algebra 1 EOC had 8th grade math as their prior year test scores but others (who took Algebra I in 8th grade) had 7th grade math as their prior. These students were entered into separate models so that students were only being compared to other students who took identical prior year tests. We only modeled combinations of test scores with at least 50 students.

From these models, we extracted a residual for each student – the difference between how the student actually scored on the test and what the model predicted they would score – and then averaged these residuals for all students in a class. We were not provided official teacher-student allocation data but using the data on students' course taking we could identify the appropriate math and ELA class for students, especially those in our study. Some students were in multiple math and ELA courses in the same school year, so we used a weighted average of residuals to allow us to, for example, allocate a "half" a student to one teacher and the other half to another. We only kept value-added estimates for classrooms with at least 10 students, but all students were included in the model to create residuals.

The resulting value-added estimate represents the mean difference, in standard deviation units, between how students scored on their state tests and what was expected of them given their demographics and prior achievement. Like Kane & Staiger (2012), we loosely translated these units to months of learning using the conversion that nine months of learning is associated with a 0.25 standard deviation increase in achievement.

Obtaining value-added estimates for all eligible math and ELA classrooms in our participating districts allowed us to compare classrooms participating in the study to those that did not. Table A.20 displays how the distribution of partner classrooms compared to the rest of the district. In most districts and subjects, participating classrooms did not differ significantly from other classrooms. Because Districts C and E had substantially fewer classrooms and fewer teachers per grade-level, we excluded both districts when comparing value-added results to classroom metrics.

Table A.21 shows the correlations between the classroom metrics we collected and value-added, as well as the mean difference in value-added estimates between classrooms rated in the top and bottom quartiles on our metrics. Quartiles in the latter were based exclusively on classrooms that had both a value-added estimate and a classroom metric score so that the number of classrooms in each quartile were relatively equal. All classroom metrics were standardized by subject, though we did this in two ways. The first pools classrooms across all districts. The

TABLE A.20 | **MEAN DIFFERENCE IN VALUE-ADDED AND NUMBER OF CLASSROOMS BY DISTRICT AND STUDY PARTICIPATION**

	District A	District B	District C	District D	District E
Mean difference in:					
ELA	0.079	-0.014	0.041	0.028	0.091
Math	0.012	-0.100 **	---	-0.019	0.070
Number of classrooms:					
In study	24	34	2	34	17
Not in study	1373	3340	26	2840	29

*, **, and *** represent significantly different mean value-added estimates at the 0.05, 0.01, and 0.001 levels respectively.

second standardizes classroom metric scores separately within each district to account for potential district effects. Both standardization approaches are displayed in Table A.21.

We found that most metrics in both types of correlations demonstrated at least some positive association. With sample sizes near 70, we were not seeking to validate each measure, but instead wanted to explore whether the associations were directionally aligned. Notably, the correlations with assignments and expectations were at least 0.2 and for expectations, significantly different from zero.

We also wanted to know whether the relationships with our metrics held among classrooms with students who tended to be furthest behind their grade-level peers and classrooms who began the year above the average student in the state. We isolated all classrooms that began the year with an average prior achievement score of 0.5 standard deviations below the state average or lower – the rough equivalent of starting at least two years behind the state – as well as classrooms that started the year at least 0.5 standard deviations above the state. Table A.21 shows the correlations among these classrooms. Because the sample size is small, we did not perform within district standardizations, and we split classrooms into top and bottom half groups rather than quartiles.

When we isolated classrooms in which the average student was substantially behind the state average, we tended to find stronger relationships between value-added results and most measures. Most notably, instruction, assignments, and expectations all demonstrated large correlations. The opposite was somewhat true for classrooms that began the year far ahead of the state mean – correlations between VAM and our assignment and instruction metrics were negative, though correlations with expectations and engagement remained positive.

TABLE A.21 | RELATIONSHIPS BETWEEN VALUE-ADDED AND CLASS METRICS

All Classrooms				
	Pooled Standardization		Within-District Standardization	
	Correlation	Quartile difference	Correlation	Quartile difference
Assignments	0.20	0.05	0.17	0.10
Instruction	0.05	0.01	0.04	0.04
Engagement	0.09	0.07	0.13	0.07
Worth	0.01	-0.04	0.09	0.04
Support	0.02	0.04	0.11	0.06
Expectations	0.24 *	0.13 *	0.36 **	0.22 **
Classrooms with prior achievement substantially BELOW the state average				
	Correlation	Difference between top and bottom half		
Assignment score	0.39	0.20 *		
Instruction score	0.62 **	0.17		
Engagement	0.13	0.03		
Worth	- 0.00	-0.02		
Support	0.19	-0.06		
Expectations	0.35	0.22 *		
Classrooms with prior achievement substantially ABOVE the state average				
	Correlation	Difference between top and bottom half		
Assignment score	-0.12	0.03		
Instruction score	-0.74 **	-0.21		
Engagement	0.21	0.18		
Worth	0.12	0.08		
Support	-0.02	-0.03		
Expectations	0.50	0.18		

*, **, and *** represent values significantly different from zero at the 0.05, 0.01, and 0.001 levels respectively. Only Districts A, B, and D are included. Classrooms with prior achievement substantially below the state average are classrooms with mean prior achievement (standardized against the state) of -0.5 or lower. Ns varied by metric but were typically near 70 for all classrooms and near 20 for classrooms with substantially lower prior achievement.

a: Split was based on whether the class was above or below the median on metric among all classes in analysis

PREDICTING ENGAGEMENT

We connected students' perceptions of engagement to the remaining classroom metrics in three ways. First, using a separate multi-level model with survey responses nested within students nested within classrooms for each metric, we modeled engagement as an outcome controlling for student characteristics (gender, race and language status), classroom characteristics (classroom percent IEP, percent FRL, and subject area)[38] and included the classroom metric of interest. We also employed the same model but included both classroom measures of assignments and observations and their interaction to see if there were meaningful differences when classrooms had higher scores on both metrics.

Next, we leveraged the fact that students completed their surveys on multiple days and used a student fixed effects model to compare engagement on days with better or worse assignments and instruction. In a student fixed effects model, students act as their own control such that we're able to analyze whether a student is more engaged than they normally are on days when they had a better assignment. All assignment and instruction scores were standardized by subject area so coefficients reported for this model represent the change in standardized student-level engagement scores for a one standard deviation increase in the daily assignment or instruction score.

The results to all three models are displayed in Table A.22. Across all models instruction was positively and significantly related to engagement: classrooms that had better instruction scores tended to have students who were more engaged. And the same students tended to be more engaged on days with better instruction. On the other hand, while engagement tended to be minimally, though positively, related to the overall assignment quality of a classroom, when we followed the same student on different days, they tended to be less engaged on days with higher quality assignments.

ANALYSIS AND DATA SOFTWARE

We used R for all quantitative analyses and data preparation: [39] We heavily used the R packages from the overarching tidyverse, particularly dplyr.[40] All multi-level models were based on functions from the lme4 package[41], and all fixed-effects linear models were based on functions from the plm package[42]; We used the TAM package[43] to perform all Rasch-based analyses and the irr package to test the interrater reliability of assignment and student work ratings.[44]

TABLE A.22 | **ESTIMATED RELATIONSHIPS BETWEEN EACH METRIC AND ENGAGEMENT AS AN OUTCOME**

All Classrooms			
	MODEL SET 1 Separate multi-level models	MODEL 2 One multi-level model with assignment and instruction interaction	MODEL SET 3 Student fixed effect models
Assignments	0.04	-0.01	-0.04 ***
Instruction	0.11 ***	0.09 *	0.07 *
Assignment X Instruction	---	0.04	---
Expectations	-0.01	---	---

*, **, and *** represent significantly different mean value-added estimates at the 0.05, 0.01, and 0.001 levels respectively. Model set 1 is a separate multi-level model for each classroom metric predicting engagement. Controls include student ender, student race/ethnicity, student language status, classroom subject, classroom grade level, and percent of students in class receiving IEPs and percent receiving FRL. Model 2 is a single version of the same multi-level model but with both assignments and instruction included, as well as their interaction. Model set 3 is two separate student-fixed effects models using each metric (assignments and instruction) measured on a given day in class compared to students' engagement on that same day. Only students who had at least two days' worth of assignments or two days' worth of instruction in our data were included.

STATE AND DISTRICT POLICY ANALYSIS

In addition to our quantitative data analysis we also conducted a qualitative analysis of state and district-level policies. As part of this policy review, we interviewed teachers and school leaders to learn how they experience these policies.

STATE-LEVEL ANALYSIS

In order to understand the state-level policies that influenced the decisions made in classrooms, schools, and districts, we reviewed state legislation – including what the state regulates and funds – related to learning standards, instructional materials, course and graduation requirements, intervention, and assessments. We also considered expectations the state set regarding serving students with disabilities and English language learners.

To conduct this analysis, a policy analyst reviewed the state code and state Board of Education policies in each state where we had a district participating in the study. We then used the findings from this review to inform our district-level interview protocols, as well as our teacher focus group protocols and principal interview protocols, both described below.

TEACHER AND ADMINISTRATIVE INTERVIEWS

District-level policy interviews

We conducted virtual and in-person interviews with district-level staff. Questions focused on each district's approach to adopting instructional materials and setting expectations for their usage, lesson planning, assessments, intervention, graduation requirements, and course scheduling, including advanced course-taking enrollment expectations. We also considered expectations each district set for serving students with disabilities and English language learners.

In each district, we customized the district-level policy interview protocol using the state-level policy analyses referenced above. We then scheduled interviews with as many district staff as needed to fully answer our research questions.

School leader interviews

We conducted at least one in-person interview with each of the school leaders in the sample to answer questions regarding the school's approach to selecting instructional materials and setting expectations for their usage, lesson planning, assessments, intervention, and course scheduling, including advanced course-taking enrollment expectations. We also considered expectations each school set for serving students with disabilities and English language learners.

Teacher focus groups

At each school, we conducted at least one hour-long in-person teacher focus group with a subset of the teachers participating in the study to learn about teachers' approach to using instructional materials, lesson planning, assessments, and intervention. We also asked teachers about their approaches to serving students with disabilities and English language learners. At each school, we selected participating teachers by identifying 5-7 teachers per school, attempting to get a representation across grade levels for both math and ELA.

Tables A.23, A.24, and A.25 display the interview protocols used for district leaders, school leaders, and teachers.

After completing teacher focus groups, school leader interviews, and district-level policy interviews, we compared answers to identify policies or practices that were implemented differently at the teacher or classroom level than expected by the school leaders or district leaders.

TABLE A.23 | **ESTIMATED RELATIONSHIPS BETWEEN EACH METRIC AND ENGAGEMENT AS AN OUTCOME**

TOPIC	QUESTIONS
Adopted Materials at the District Level	**What are the state requirements for curricular adoption in your state?** • What process does your district use when adopting curriculum, given those requirements? Do all of your core adopted materials come from the state-approved list? • (If not) How do you ensure that self-selected materials are high quality and aligned to standards? Do you have to receive a waiver/approval to use these materials? **Are schools or teachers ever allowed to select their own curriculum?** • Are there any processes or guidelines they have to follow when choosing their own curriculum? • How do you know if their curriculum meets state standards? **What core curricular resources have you adopted in literacy across the grade levels?** • How effective and well-aligned to state standards are each of these curricula? How do you determine alignment? • Which are most useful? Which are least useful? Why? • Has the district created any supplements (like a pacing guide) that would support teachers in using these materials? (If yes – can we see them?) **What core curricular resources have you adopted in mathematics across the grade levels?** • How effective and well-aligned to state standards are each of these curricula? • How do you determine alignment • Which are most useful? Which are least useful? Why? • Has the district created any supplements (like a pacing guide) that would support teachers in using these materials? (If yes – can we see them?) **What computer or web-based instructional materials has the district adopted for literacy and math? (e.g., iReady)** • How effective area each of these resources? Which are most useful? Which are least useful? Why? **Are there any materials or resources you wish you had access to that you do not currently have access to? If so, what are they?** **What training have teachers received on using any of these curricular resources?**
Daily Lesson Planning at the School Level (as perceived by the district)	**Where do your math & ELA teachers' daily lesson plans come from? E.g. Do they create them themselves? Are they expected to use district-provided lesson plans?** • Do teachers spend more time creating lesson plans or preparing to use already created plans? • Is there a process for reviewing and approving teachers' lesson plans? How do leaders ensure lesson plans fully address instructional standards? • What daily lesson planning materials, if any, has the district created? (If yes – can we see them?) • What additional materials has the district purchased for schools? • Are certain instructional materials mandated for use? If so, what expectations has the district set for usage of those materials? • If from other sources, where?

TABLE A.23 CONTINUED

TOPIC	QUESTIONS
Assessments and Data	**In your opinion, are the state's summative assessments a good reflection of your curricula and what students should be learning? Are they a good reflection of state standards?** **How are your students doing on state assessments? To what do you attribute their performance?** **How does student performance on state assessments factor into district, school, and teacher ratings? How do they factor into student grades or promotion and graduation?** **What benchmark assessments do you use for literacy and math by grade band?** • How well-aligned are these benchmarks to your curricula and state standards? How do you know? • Can you send us copies of your benchmarks? **How much time do you spend on benchmark assessment during the year?** • Are there any district-wide expectations for what teachers are expected to do with benchmark assessment results? (e.g. review with students; create plans for individual students who are behind, etc.) **How do you use assessment data to adjust district instruction in relation to standards?**
Intervention	**How does the district determine which students will receive intervention (especially intervention as distinct from special education-related supports)?** **What policies does the district set related to intervention time? (e.g. required 30 min/day of reading on top of Tier 1 instruction)** • When are students supposed to receive intervention during the day? E.g. are they pulled out of certain classes? **How does the instruction that those students in intervention receive differ from the instruction that others receive?** **Does the district use specific instructional programs (e.g. iReady) for intervention purposes? Which ones?** • How well aligned to state standards are those programs? How do you know? **By what measures do you assess the effectiveness of intervention across the district? Is intervention effective in your district?** • Do you know what % of intervention students move off of intervention or achieve proficiency on state exams?
Classroom Structure and Time Usage	**What do the literacy and math blocks/classes look like in your schools?** • How much time do students spend in math and literacy relative to other subjects in elementary, middle and high school? **Are there district-wide expectations for how time is used or the structure of these blocks/classes?** • What expectations has the district set for these blocks? • How is time allocated during the literacy block? • How is time allocated during the math block? • How closely do teachers follow these expectations?
Standards	**How are your teachers generally doing with implementing new state standards: where are they excelling and where are they struggling?** **Has the state provided any support in better understanding and implementing state standards?** • How effective or helpful has that support been? What has been most or least helpful? **What sort of support has the district provided teachers in understanding and implementing state standards?**

TABLE A.23 CONTINUED

TOPIC	QUESTIONS
School Enrollment/Zoning	**What is the district's school zoning and enrollment policy outlining how students get assigned to or choose their schools? Do students have to go to their neighborhood school or do they have choice in where they go?** **Is there any sort of application/selection process for students who want to choose a school outside of their neighborhood?** **What percent of students go to their neighborhood school vs a choice school?** • What are the main barriers that keep more students from choosing schools outside of their neighborhood? What supports does your district offer? **Does the district have any information about how students fare at choice schools when they leave neighborhood schools?** • What about those students who are left behind?
Graduation Requirements & Master Schedules at the District Level	**What is the length of the school year? The school day?** • What expectations do you set at the district level for teacher common planning time? • What expectations do you set at the district level for time teachers spend in district-level PD? **What are the state-level graduation requirements that influence how your district approaches creating course schedules for students?** • Have you created any district-level course or graduation requirements? **How do you expect schools to approach making sure that all students are enrolled in the courses they would need to meet those graduation requirements?** • What do you do at the district level to monitor students' course enrollments? **How do schools approach creating master schedules?** **What are the expectations for how students are assigned to classes? For instance, which students get assigned to Algebra I as freshmen vs a lower level math course?** **What is your district graduation rate? How has that rate changed, if at all, across the past few years?** **What programs (e.g., credit recovery programs, summer school) does the district use to increase graduation rates?** • How do you ensure that those programs continue to be aligned to state standards?
Advanced Programming at the District Level	**What advanced programming opportunities exist in your secondary schools?** • Do any of your schools offer Advanced Placement courses for students? If so, which schools? • Do any of your schools offer International Baccalaureate programming for students? If so, which schools? • Do any of your schools offer dual enrollment courses for your students? If so, which schools? • Are there any other honors-designated courses at your schools? **Do any advanced programming opportunities (i.e. honors courses, gifted and talented programs) exist at the elementary level? If so, what are those?** **How do you determine which students are able to enroll in the advanced courses offered at your schools?** • Are there district-level requirements students have to meet to be able to enroll in these advanced courses? **How do you measure the success of your advanced programming?** **What are your pass rates on any advanced programming assessments?** • Do all or only some students enrolled in advanced coursework take the advanced coursework assessments?

TABLE A.23 CONTINUED

TOPIC	QUESTIONS
Subpopulation Policies & Expectations at the District Level: Students with Individualized Education Plans (IEPs)	**What expectations does the district set for how students with IEPs must be served? E.g. What sort of educational services and modifications must be made available to them?** • How does the instruction for students with IEPs differ from the rest of the student population? • What model do you use to serve students with IEPs (i.e. inclusion, self-contained classrooms, etc.)? • (If students are pulled out) What courses are students with IEPs pulled from? How often are they pulled out? • Are there any specific academic programs or curricula you use with students on IEPs to support their learning? (e.g. Read180) **Are students with IEPs held to the same academic standards as other students?** • Are they required to take/pass the same courses to graduate? • Are they required to take the same standardized assessments? • What, if any, modifications are provided them? **How does evaluation differ for teachers of students with IEPs?** • Are student assessments factored into their evaluations? • Into school/district evaluations?
Subpopulation Policies & Expectations at the District Level: English Language Learners	**How many ELL students are in the district?** **Do you have a tiering system for ELL students that determines what type of services those students receive?** **Do you have pull out classes for ELL students, or are ELL students included in general education classes? (If pull-out: What courses are ELL students pulled from?)** **Are ELL students held to the same academic standards as other students?** • Are they required to take/pass the same courses to graduate? • Are they required to take the same standardized assessments? • What, if any, modifications are provided for them? **How does evaluation differ for teachers of ELL students? Are student assessments factored into their evaluations?**

TABLE A.24 | **SCHOOL LEADER INTERVIEW PROTOCOL**

TOPIC	QUESTIONS
Adopted Curricular Materials at the School Level	**What core curriculum do you use in literacy across the grade levels?** • How effective are each of those curricula in your opinion? • Are they aligned to state standards? How do you know? **What core curriculum do you use in mathematics across the grade levels?** • How effective are each of those curricula in your opinion? • Are they aligned to state standards? How do you know? **Have you adopted any additional computer or web-based curricula for literacy or mathematics? (e.g., iReady)** **Did you choose these curricular resources, or were they chosen at the district level?** • (If chosen by the principal/school) why did you choose these specific resources? **Are math and ELA teachers required to use district/school adopted curricula? Or do they have discretion to find/use their own?** • (If teachers have discretion) How do you ensure that math and ELA teachers' curricula are aligned to state standards? **How often do you do school-level development related to improving teachers' understanding of state literacy and math standards and any district/school-adopted curriculum?** • Has the district or state provided any support to you or your teachers in understanding standards or implementing specific math or ELA curricular resources? How helpful has that support been?
Lesson Planning Materials at the Classroom Level (as perceived by the school leader)	**Where do your ELA and math teachers' daily lesson planning materials come from (e.g. Do they develop their own? Follow a scripted curriculum?)** • In general, would you say that your math and ELA teachers spend more time creating lesson plans or preparing to deliver already developed lesson plans? **What math and ELA lesson planning materials, if any, has the district/school created?** • What is the quality of district/school-developed lesson plans? • Are math and ELA teachers expected to use district/school-created lesson planning materials or do they have discretion in creating their own? **(If teachers are allowed to create their own lesson plans) Are there specific sources teachers are going to in order to get lesson plans (e.g. Teachers Pay Teachers)** • How do you ensure that teacher-developed lesson plans fully address state standards?
Assessments and Data	**What are the main literacy and math benchmark assessments that your school uses to assess student performance throughout the year?** • How did you choose/develop these assessments? • Are they aligned to standards? To your curriculum? How do you know? • How often do you use these assessments to assess student progress? • Are they helpful in identifying student instructional progress and outstanding needs? **How are teachers using assessment data in your school to make instructional adjustments?** **How are your students doing on state literacy and math assessments?** • Are there any particular school-based factors to which you attribute their performance?

TABLE A.24 CONTINUED

TOPIC	QUESTIONS
Intervention	**How does the school determine which students will receive intervention supports (intervention as distinct from special education)?** **What policies does the district set related to intervention time? (e.g. required 30 min/day of reading on top of Tier 1 instruction)** • When do students receive intervention during the day? E.g. are they pulled out of certain classes? **How does the instruction that students in intervention receive differ from the instruction that others receive?** • What additional supports do students receive? **What intervention programs (if any) do you use at the school level?** • Are those programs district-mandated or specific to your school? • Are those programs aligned to state standards? How can you tell? **How effective are your school's intervention process and programs? How do you measure intervention effectiveness?** • Do you know what % of students on intervention move out of intervention? What % achieve proficiency on state assessments?
Graduation Requirements/ Master Scheduling (High Schools Only)	**What are the state requirements for graduation?** • Does your school offer all courses required to meet state graduation requirements? **How does your school approach creating course schedules for students that ensures they meet graduation requirements?** • How do you monitor student course enrollment to ensure students remain on track for graduation? **How do you assign individual students to individual courses?** • How do you factor student ability into course scheduling? **What programs (e.g., credit recovery programs, summer school) does your school use to increase graduation rates?** • Are those programs aligned state standards? How do you know?
Advanced Programming at the School Level	**What advanced programming opportunities exist in your school, if any? E.g. AP, IB, dual enrollment, honors.** **How do you determine which students are able to enroll in the advanced courses offered at your school?** • Who sets those requirements? **(For high schools only) How do you measure the success of your advanced programming?** • What are your pass rates on any advanced programming assessments? • Do all or only some students enrolled in advanced coursework take the advanced coursework assessments? What %?

TABLE A.24 CONTINUED

TOPIC	QUESTIONS
Subpopulation Policies & Expectations at the School Level: English Language Learners	**What expectations do you have for how teachers serve English Language Learners? What modifications/interventions must ELL students receive, for example?** • (For high schools only) How do you determine which courses ELL students are enrolled in? Are they required to take the same courses to graduate as other students? **How do you measure the effectiveness of your school's ELL instruction?** • How effective would you say your school's ELL instruction is? • To what do you attribute those results?
Subpopulation Policies & Expectations at the School Level: Students with Individualized Education Plans (IEPs)	**What expectations do you have for how teachers serve students with IEPs? What modifications/interventions must students with IEPs receive, for example?** • (For high schools only) How do you determine which courses students with IEPs are enrolled in? **How do you measure the effectiveness of your school's Special Education instruction?** • How effective would you say your school's Special Education instruction is? • To what do you attribute those results?
Subpopulation Policies & Expectations at the School Level: Gifted & Talented Students	**How do you identify gifted and talented students?** • How do you determine which courses gifted and talented students are enrolled in?
Standards	**How would you rate your own understanding of state standards and the instructional shifts they require?** **In your opinion, is your teachers' instruction in math and ELA generally meeting the expectations called for in state standards?** **How do you ensure that your teachers' instruction is aligned with standards?**
Classroom Structure and Time Usage	**How much time do students spend in math & literacy classes relative to others?** **How are literacy and math blocks/classes structured in your school?** **Are there school-wide expectations for how time is used or the structure of these blocks/classes? (e.g. twenty minutes must be in whole group).** • What expectations has the school set for these blocks? • How is time allocated during the literacy block? • How is time allocated during the math block? • How closely do teachers follow these expectations?

TABLE A.25 | **TEACHER FOCUS GROUP PROTOCOL**

TOPIC	QUESTIONS
Standards	**How do you take state standards into account when planning your lessons and longer term plans?** **In your opinion, does your instruction generally meet the expectations called for by state standards?**
Adopted Materials at the School Level	**What core ELA curriculum do you primarily use throughout the year?** **What are the core math curricula you use?** **Is your curriculum mandated by the district or school?** • If yes, how effective is the curriculum in your opinion? Does it align to state standards? • If yes, is there any curriculum you would prefer to use? Which one? Why? • If not mandated, why did you choose this curriculum? • If not If not mandated, do you have to demonstrate to your school that your curriculum meets standards in any way? **Does your school or district provide you with any support in understanding state standards and implementing your curriculum effectively?** • What sort of support have they provided? • How helpful has that support been, in your opinion? **Do you use any supplemental curricular materials in math and ELA, whether print or computer or web-based? (e.g., iReady) If so, which ones?**
Daily Lesson Planning at the Classroom Level	**When it comes to daily lesson planning, do you more frequently develop your own materials or do you use materials provided by the district, school, or your adopted curriculum? Why?** **When developing your own lesson plans, what specific sources do you turn to to find materials? Please be as specific as possible (e.g. Teachers Pay Teachers).** • Why do you use those sources? • How do you ensure alignment of your self-created lesson plans to state standards?
Classroom Structure and Time Usage	**What do your literacy and/or math blocks/classes look like?** • How do you allocate time during these blocks? **(For self-contained elementary teachers) How much time do you devote to math and literacy instruction each day on average? How does that compare to other content areas you teach?** • Why do you use this breakdown of time spent on math and literacy?
Assessments and Data	**Do your students take a summative state assessment?** **If so, how did you class last year perform on state assessments?** • To what do you attribute their performance? **What are the main literacy and math benchmark assessments that you use (e.g. on a quarterly basis)?** • Why do you use these benchmarks? • Are they aligned to standards? To your curriculum? How do you know? **How helpful are your assessments in providing you with information about your students?** • How are you using assessment data to adjust your instruction in relation to state standards? **How do student assessment scores factor into your own evaluation?** • Has this changed how you teach at all?

TABLE A.25 CONTINUED

TOPIC	QUESTIONS
Intervention	What role do you play in helping the school determine which students will receive academic intervention supports (intervention as distinct from special education supports)? How does the instruction that students in intervention receive differ from the instruction that others receive? • What additional supports do they receive? Is intervention support at your school effective? How do you know?
Subpopulation Policies & Expectations at the School Level: English Language Learners	How do you adjust your instruction for English Language Learners? How do you measure the success of your English Language Learners? Are your ELL students successful by your standards? How are you held accountable for the success of English Language Learners in your class?
Subpopulation Policies & Expectations at the School Level: Students with Individualized Education Plans (IEPs)	How do you adjust your instruction for students with IEPs? How do you measure the success of your students with IEPs? Are your students with IEPs successful by your standards? How are you held accountable for the success of students with IEPs in your class?
Subpopulation Policies & Expectations at the School Level: Gifted & Talented Students	How do you adjust your instruction for gifted and talented students in your classroom?

CURRICULUM/INSTRUCTIONAL MATERIALS QUALITY REVIEW

MATH

We reviewed districts' instructional materials for their alignment to CCSS standards in both Math and ELA at grades 1, 4, 7, and 10 using the Instructional Materials Evaluation Tool (IMET)[45]. The IMET in mathematics [46] is designed to help determine whether instructional materials are aligned to the shifts and major features of the applicable state standards. In mathematics, those shifts are:

- **Focus:** Focus strongly where the Standards focus.
- **Coherence:** Think across grades and link to major topics within the grade.
- **Rigor:** In major topics, pursue conceptual understanding, procedural skill and fluency, and application with equal intensity.

To determine whether materials meet those three key shifts, the K-8 IMET in Mathematics[46] and High School IMET in Mathematics[47] require that materials are first rated on "non-negotiables." If the set of provided instructional materials meet those non-negotiables, they are then also rated against three "alignment criteria." The non-negotiables are:

- **Non-Negotiable 1**[48]**:** Freedom from Obstacles to Focus : Materials reflect the basic architecture of CCSS by not assessing topics before they are intended by the standards.
- **Non-Negotiable 2:** Focus and Coherence: Materials focus coherently on the Major Work/Widely Applicable Prerequisites for College and Career[49] of the grade in a way that is consistent with the progressions in CCSS, and a large majority of instruction is focused on Major Work/Widely Applicable Prerequisites for College and Career.

And the alignment criteria are:

- **Alignment Criterion 1:** Rigor and Balance: Materials reflect the appropriate balance in CCSS between conceptual understanding, procedural fluency, and mathematical application.
- **Alignment Criterion 2:** Standards for Mathematical Practice: Materials authentically connect content- and practice standards, ensuring that students have ample opportunity to engage in the CCSS standards for mathematical practice (e.g. persevere in solving challenging problems).
- **Alignment Criterion 3:** Access to Standards for All Students: Materials provide supports for English Language Learners and other special populations to access CCSS at their grade level and demonstrate their mathematical understanding independently.

Table A.26 displays the IMAT mathematics ratings for each district and grade. Materials that were rated as Not Aligned typically did not focus sufficiently on the major work of the grade.

TABLE A.26 | **IMET RATINGS BY DISTRICT IN MATHEMATICS**

	District A	District B	District C	District D	District E
1st Grade Math	Aligned	Aligned	Aligned	Not Aligned	Not Aligned
4th Grade Math	Aligned	Aligned	Aligned	Not Aligned	Not Aligned
7th Grade Math	Aligned	Not Aligned	Aligned	Not Aligned	Not Aligned
10th Grade Math	Partially Aligned	Aligned	Partially Aligned	Not Aligned	Partially Aligned

LITERACY

The Instructional Materials Evaluation Tool (IMET) in ELA/Literacy is designed to help determine whether instructional materials are aligned to the shifts and major features of the applicable state standards. In ELA/literacy, those shifts are:

- **Complexity:** Regular practice with complex text and academic language.
- **Evidence:** Reading, writing, and speaking grounded in evidence from literature and informational text.
- **Knowledge:** Building knowledge through content-rich non-fiction.

To determine whether materials meet those three key shifts, the K-2 IMET in ELA/Literacy[50] and 3-12 IMET in ELA/Literacy[51] require, like math, that materials are first rated on non-negotiables, then alignment criteria. The non-negotiables are:

- **Non-Negotiable 1:** High Quality Texts: Anchor texts are high quality and worthy of student attention, of the appropriate quantitative and qualitative grade-level complexity and comprised of both informational texts and literature.
- **Non-Negotiable 2:** Evidence-Based Discussions and Writing: At least 80% of questions and tasks are text-dependent, requiring students use textual evidence. Materials include frequent opportunities for evidence-based discussion and writing
- **Non-Negotiable 3:** Build Knowledge: Materials provide a sequence of texts organized around a variety of topics at each grade level that systematically build knowledge and vocabulary through reading, writing, speaking, and listening.
- **Non-Negotiable 4:** Foundational Reading Skills (K-2 only): Materials include instruction across foundational skills concepts (phonics, concepts of print, etc.) and a variety of reading material with frequent foundational skills practice, helping students use foundational skills to make meaning from reading.

And the alignment criteria are:

- **Alignment Criterion 1:** Range and Quality of Texts: Materials reflect grade level text complexity and the proper distribution of genres and text types (50/50 informational/literature split in 3-5 and a more substantial focus on high-quality non-fiction in 6-12). Materials provide frequent opportunities for developing reading fluency with grade level materials.
- **Alignment Criterion 2:** Questions, Tasks, and Assignments: Materials support students in building reading comprehension, finding and producing textual evidence, and developing grade-level academic language.
- **Alignment Criterion 3:** Building Knowledge with Texts, Vocabulary, and Tasks: Materials build students' knowledge across topics and content areas and include frequent research projects and opportunities to engage with academic vocabulary.
- **Alignment Criterion 4:** Access to the Standards for all Students: Materials are designed to provide thoughtful supports/scaffolds to support all students in accessing the Standards at their grade level.

Table A.27 displays the IMET literacy ratings for each district and grade. Materials that were rated as Not Aligned typically did not feature texts that were sufficiently rigorous for the grade or that did not adequately build relevant content knowledge. Many "Not Aligned" materials also did not feature questions and tasks that were appropriately text-dependent.

TABLE A.27 | **IMET RATINGS BY DISTRICT IN LITERACY/ELA**

	District A	District B	District C	District D	District E
1st Grade ELA	Not Aligned	Not Aligned	NA-Teacher Created	Not Aligned	Partially Aligned
4th Grade ELA	Not Aligned	Aligned	NA-Teacher Created	Not Aligned	Not Aligned
7th Grade ELA	Aligned	Aligned	Aligned	Not Aligned	Not Aligned
10th Grade ELA	Aligned	Partially Aligned	NA-Teacher Created	Not Aligned	Not Aligned

ASSESSMENT QUALITY REVIEW

We reviewed districts' benchmark assessments for their alignment to CCSS standards in both Math and ELA at grades 1, 4, 7, and 10 using the Assessment Evaluation Tool (AET)[52]. Like the IMET, we first reviewed for the AET's listed non-negotiables, and then alignment criteria. Assessments were rated Aligned if they met all the non-negotiables and alignment criteria. They were rated partially aligned if they meet the non-negotiables but not all the alignment criteria. They are rated not aligned if they did not meet all the non-negotiables. Alignment criteria are not rated if the assessment does not meet the non-negotiables.

MATH

The AET math non-negotiables are:

- **Non-Negotiable 1:** Focus on Major Work—The large majority of points in each grade K–8 are devoted to the Major Work of the grade, and the majority of points in each high school course are devoted to Widely Applicable Prerequisites.
- **Non-Negotiable 2:** Freedom from Obstacles to Focus—No item assesses topics directly or indirectly before they are introduced in the CCSSM.
- **Non-Negotiable 3:** Coherence of the Standards—Test items elicit direct, observable evidence of the degree to which a student can independently demonstrate the targeted Standard(s), reflecting the coherence of the CCSSM.

And the alignment criteria are:

- **Alignment Criterion 1:** Rigor and Balance—The Standards set expectations for attention to all three aspects of rigor: conceptual understanding, procedural skill and fluency, and application. Thus, assessments must reflect the balances in the Standards and help students meet the Standards' rigorous expectations.
- **Alignment Criterion 2:** Emphasize the Progressions—Assessments reflect the grade-by-grade progressions in the Standards.
- **Alignment Criterion 3:** Standards for Mathematical Practice—The Standards require mathematical practices to be connected with mathematical content. Thus, assessments should demonstrate authentic connections between content Standards and practice Standards.
- **Alignment Criterion 4:** Supporting Focus—The assessment program supports the focus of the Standards by connecting concepts and presenting score report information in a manner that highlights the emphasis of the grade or course.

Across all four grades, and in all five districts, all math assessments were rated "Not Aligned" on the AET.

LITERACY

The AET literacy non-negotiables are:

- Non-Negotiable 1: Complexity and Quality of Text—Texts are worthy of student time and attention; they have the appropriate level of complexity for the grade, according to both quantitative and qualitative analyses of text complexity.
- Non-Negotiable 2: Text-Dependent and Standards-Based Questions—High-quality reading test questions are text-dependent and Standards-based; they require students to read closely, find the answers within the text, and use textual evidence to support responses.

And the alignment criteria are:

- **Alignment Criterion 1:** Range of Texts—Texts reflect the distribution of text types and genres required by the reading Standards.
- **Alignment Criterion 2:** Assessing Vocabulary—Because of the importance of vocabulary acquisition and use to college and career readiness, vocabulary questions comprise a significant part of ELA/literacy assessments, assess tier 2 words in context, and focus on central ideas in the text.
- **Alignment Criterion 3:** Aligned Use of Item Types—A variety of item types is used to appropriately and strategically assess the Standards.
- **Alignment Criterion 4:** Test Blueprints and Score Reports—Test blueprints and the corresponding score reports reflect the focus of the Standards.
- **Alignment Criterion 5:** Writing to Sources—Writing tasks reflect the writing types named in the Standards and require students to write to sources.
- **Alignment Criterion 6:** Language—Test questions assessing conventions and writing strategies focus on the specifics of the Standards and reflect actual practice to the extent possible.
- **Alignment Criterion 7:** Speaking and Listening—Test questions assessing speaking and listening reflect true communication skills required for college and career readiness.

Table A.28 displays the AET literacy ratings for each district and grade. Assessments that were rated as Not Aligned typically did not feature sufficiently challenging texts for the grades and/or questions that were appropriately text-dependent.

TABLE A.28 | **AET RATINGS BY DISTRICT AND GRADE**

	District A	District B	District C	District D	District E
1st Grade ELA	Not Aligned	Not Aligned	Not Aligned	Not Aligned	Not Aligned
4th Grade ELA	Not Aligned	Not Aligned	Not Aligned	Not Aligned	Not Aligned
7th Grade ELA	Partially Aligned	Not Aligned	Not Aligned	Not Aligned	Not Aligned
10th Grade ELA	Not Aligned	Partially Aligned	Not Aligned	Not Aligned	Not Aligned

ENDNOTES

[1] Though the CMO schools in our study are all located in different public school districts, throughout the Technical Appendix we refer to the CMO schools collectively as a single district.

[2] A few additional teachers (less than five) who taught subjects for which we did not have tools (e.g., world languages) or in types of classes out of the purview of our study and tools (e.g., pull-out special education classes) participated in our study so that they could receive formative feedback. These classrooms were not included in any analyses.

[3] We did not collect any student-level information on students who did not return a consent form and thus did not track the exact return rate. However, we compared the number of consent forms received to the total number of students enrolled in the course derived from district-provided data (see later in the appendix for more information on this data). Percent in text weights each student equally, but the average classroom rate was similar, at 58%.

[4] Choices adapted from Kane, T.J., Owens, A.M., Marinell, W.H., Thal, D.R., & Staiger, D.O. (2016). *Teaching higher: Educator's perspectives on Common Core implementation.* Boston, MA: Center for Education Policy Research at Harvard University.

[5] While the first two domains focus more heavily on the extent to which the assignment aligns to grade-level expectations and gives students meaningful opportunities to engage them, the third domain focuses more heavily on the authenticity of what the assignment asks students to do. An assignment's "value beyond school" was one of three domains used by Newmann and colleagues in their study finding significant associations between student assignments and student achievement. See Newmann, F. M., Lopez, G., & Bryk, A.S. (1998). *The quality of intellectual work in Chicago schools: A baseline report.* Chicago: Consortium on Chicago School Research.

[6] For example, Newmann and colleagues reported exact agreement rates of approximately 70%, though their domains had four categories, not three. See Newmann, F.M., Lopez, G., & Bryk, A.S. (1998). *The quality of intellectual work in Chicago schools: A baseline report.* Chicago: Consortium on Chicago School Research. Additionally, Clare and Aschenbacher (2001) reported kappa coefficients ranging from 0.27 to 0.59 depending on the domain for their tool assessing language arts assignments, though they too had domains on a 4-point scale. Our range of unweighted kappa values across all subjects was 0.51 – 0.63, and 0.54 – 0.62 for ELA. See Clare, L. & Aschbacher, P. R. (2001). Exploring the technical quality of assignments and student work as indicators of classroom practice. *Educational Assessment*, 7(1) 39-59.

[7] See TNTP Core Teaching Rubric (https://tntp.org/publications/view/tntp-core-teaching-rubric-a-tool-for-conducting-classroom-observations) and Achieve the Core Instructional Practice Guide (https://achievethecore.org/page/1119/coaching-tool) respectively.

[8] For an overview of ESM, see Hektner, J.M., Schmidt, J.A., & Csikszentmihalyi, M. (2006). *Experience Sampling Method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage Publications. See also Mehl, M. R. & Conner, T. S. (Eds.) (2012). *Handbook of Research Methods for studying daily life*. New York: The Guilford Press.

[9] Watches were programmed so that they did not vibrate until at least 10-15 minutes of class time had transpired so teachers could hand out materials and begin instruction.

[10] In summarizing existing ESM research Zirkel, Garcia, and Murphy (2015, page 9) note though ESM data cannot eliminate all bias, "… research suggests that by asking people to report on their activities, affect, and actions in situ and on many small occasions, we may be able to get a more accurate picture than when we ask participants to reflect backward over a period of time." Zirkel, S., Garcia, J. A., & Murphy, M C. (2015). Experience-sampling research methods and their potential for education research. *Educational Researcher*, 44(1), 7-16.

[11] Spanish versions of both the daily and background surveys were available in districts, schools, or classrooms that required or requested them.

[12] Fredericks and McColskey (2012) highlight three types of student engagement: behavioral (Does the student participate academically and socially in school? Do they follow school rules and norms?); emotional (Do students have positive reactions to teachers and classmates? Do they have a sense of belonging?); and cognitive (Are students invested in learning?) Fredericks, J.A. & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S.L. Christenson et al. (eds.) *Handbook of research on student engagement* (pp 763 – 782). New York: Springer.

[13] The interest, enjoyment, and concentration approach to measuring engagement is based on Shernoff, D.J., et al. (2003). Student engagement in high school classrooms from the perspective of Flow Theory. School Psychological Quarterly, 18(2), pp 158 – 176. Engagement survey items also adopted from Uekawa, K., Borman, K. & Lee, R. (2007). Student engagement in U.S. urban high school mathematics and science classrooms: Findings on social organization, race, and ethnicity. *The Urban Review*, 39(1), pp 1 – 43.

[14] *Ibid* Endnote 13

[15] We used a Rasch Partial Credit model that allows the differences between estimated thresholds to vary by item. Principal component analyses suggested that the items making up each construct were well represented by a single construct, with a leveling off of eigenvalues after the first, and the first principal component accounting for 62% – 76% of the variation depending on the grade level and construct. All Rasch estimation conducted using Marginal Maximum Likelihood Estimation with the TAM package in the statistics language software, Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules. R package version 2.2-49. https://CRAN.R-project.org/package=TAM. The mean and standard deviation of the posterior distributions were used to obtain Individual estimates of engagement or worth and their standard errors respectively (i.e., expected a posteriori, EAP).

[16] Items where more disagreement represents more engagement – "I feel bored," for example – were reverse coded before they were entered into the Rasch process. The conversion to a 10-point scale is based on the process used in the Chicago Consortium for School Research: http://ccsr.uchicago.edu/downloads/9585ccsr_rasch_analysis_primer.pdf

[17] Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*, second edition. New York: Rutledge.

[18] See https://www.bls.gov/soc/ for the Standard Occupation Classification codes and https://www.bls.gov/ooh/ for Occupational Outlook Handbook.

[19] For example, Kane, T.J., Owens, A.M., Marinell, W.H., Thal, D.R., & Staiger, D.O. (2016). *Teaching higher: Educator's perspectives on Common Core implementation.* Boston, MA: Center for Education Policy Research at Harvard University; Opfer, V.D., Kaufman, J.H., & Thompson, L.E. (2016). *Implementation of K–12 state standards for mathematics and English language arts and literacy.* Santa Monica, CA: RAND; Markow, D., Macia, L., & Lee, H. (2013). The MetLife survey of the *American teacher: Challenges for school leadership.* New York, NY: Metropolitan Life Insurance Company.

[20] One item in the support scale – "The standards make teaching less enjoyable" - had a large infit value, implying that many individuals did not respond to this question as expected, but we retained it because it represented teachers' beliefs about how the standards affected their day-to-day work.

[21] See, for example, Phelps, G., Corey, D., DeMonte, J., Harrison, D., & Loewenberg Ball, D. (2012). How much English language arts and mathematics instruction do students receive? Investigating variation in instructional time. *Educational Policy*, 26(5), 631-662. Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *The Elementary School Journal*, 105(1), 75-101. Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal*, 105(1), 103-127.

[22] For example, Newmann et al. (2001) similarly use shrunk estimates to describe classroom assignment quality. See Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Improving Chicago's Schools*. Chicago: Consortium on Chicago School Research. For more information about shrunk estimates derived from multi-level linear models, see Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/*

hierarchical models. New York: Cambridge University Press.

[23] Results were qualitatively similar if we simply weighted each assignment-by-day equally, regardless of how many other assignments were used on the same day.

[24] 92% of classes that administered surveys had at least 20 responses.

[25] Schools in District C provided as much course historical data as possible but did not have access to the same school years as some of the historical data was maintained by their home district.

[26] School Courses for the Exchange of Data (SCED), see https://nces.ed.gov/forum/SCED.asp. For the rural district, we mapped the courses to SCED codes directly as there were fewer than 300 courses total, and these mappings were not provided by the district.

[27] There were some cases where a holistic rating was not included, especially at the middle school level. In these cases, we took the final mark (1, 2, 3, or 4) on each standard and calculated the mean. We then converted this mean standards-grade to a numeric grade with the formula: 55 + 10*(mean standards-grade).

[28] To be clear, participating districts were in states whose average scores vary on nationally representative tests, like NAEP. We did not adjust state-standardized scores to reflect differences between states.

[29] We adapted this approach from Reardon, S.F., Kalogrides, D., & Ho, A. (2017). Linking U.S. School District Test Score Distributions to a Common Scale (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09. We used the bbmle package in R to perform the optimizations: Bolker, B. (2017). Maximum likelihood estimation and analysis with the bbmle package. We used the same approach to impute the district-wide standard deviations in the CMO districts, swapping out the state mean and performance distribution for the district equivalents. One school in District C provided raw test results rather than scores on the same scale used to make performance categories so we used the school's student-level standard deviation in the standardization process.

[30] For ACT benchmarks, see: https://www.act.org/content/act/en/college-and-career-readiness/benchmarks.html. For SAT benchmarks, see: https://collegereadiness.collegeboard.org/about/scores/benchmarks . (Accessed May 7, 2018.)

[31] We additionally weighted assignments such that the sum of weights on a given day during the study was always 1. The classroom percentages and means, therefore, better represent the percent of time spent on each assignment. See the previous section's description of our assignment metrics for more details.

[32] Most states require 180 days of instruction; see https://nces.ed.gov/programs/statereform/tab5_14.asp. (Accessed May 7, 2018.)

[33] See Nord, C., Roey, S., Perkins, R., Lyons, M., Lemanski, N., Brown, J., & Schuknecht, J. (2011). *The nation's report card: America's high school graduates* (NCES 2011-462). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/studies/2011462.pdf

[34] Students who identified as neither gender were excluded from the models.

[35] Given this binary outcome, we also ran multi-level logistic models and obtained qualitatively similar results. We have chosen to show the linear probability model instead because the interpretation on the coefficients is more similar to the results from the classroom-level models.

[36] See *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Research Paper. MET Project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://eric.ed.gov/?id=ED540960

[37] For end-of-course assessments, we only used the prior achievement information in the same subject, as there was not always a consistent prior test in the other subject.

[38] Engagement scores were already standardized by grade level so we did not need to add a control for that variable.

[39] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[40] Wickham, H. (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse. Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A grammar of data manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr

[41] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

[42] Croissant, Y. & Millo, G (2008). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, 27(2), pp. 1–43.

[43] Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. R package version 2.12-18. https://CRAN.R-project.org/package=TAM

[44] Gamer, M., Lemon, J., & Singh, P. (2015). Various coefficients of interrater reliability and agreement. R package version 0.84. https://CRAN.R-project.org/package=irr

[45] Achieve the Core. Instructional Materials Evaluation Tool (IMET). Retrieved from https://achievethecore.org/page/1946/instructional-materials-evaluation-tool

[46] Achieve the Core. IMET: Mathematics, Grades K-8. Retrieved from https://achievethecore.org/content/upload/Updated%20K8%20Math%20IMET_11.14.pdf

[47] Achieve the Core. IMET: Mathematics, High School. Retrieved from https://achievethecore.org/content/upload/Updated%20HS%20Math%20IMET_v4%202017.pdf

[48] This is not rated in the high school IMET, given the design of the applicable state standards.

[49] Achieve the Core. Mathematics: Focus by Grade Level. Retrieved from https://achievethecore.org/category/774/mathematics-focus-by-grade-level

[50] Achieve the Core. IMET: ELA/Literacy, Grades K-2. Retrieved from https://achievethecore.org/content/upload/IMET%20ELA%20K-2%20Final%20Draft%20revised.pdf

[51] Achieve the Core. IMET: ELA/Literacy, Grades 3-12. Retrieved from https://achievethecore.org/content/upload/IMET%20ELA%203-12%20Final%20Draft%20revised.pdf

[52] Achieve the Core. Assessment Evaluation Tool. Retrieved from https://achievethecore.org/page/1825/assessment-evaluation-tool

ADDITIONAL INFORMATION

CONTACT

Adam Maier (adam.maier@tntp.org) for more information about any of the rubrics, surveys, tools, or protocols discussed in the Technical Appendix.