Paper 3

Assessing Students' Learning from *Carbon TIME* in the Large and Diverse Education System

Qinyun Lin, Kenneth A. Frank, Charles W. Anderson, Michigan State University

Karen Draney, Shruti Bathia, Berkeley Evaluation and Assessment Research Center

Jay Thomas, ACT

Prepared for the annual meeting of the National Association for Research on Science Teaching,
Portland, OR,
March, 2020

This research is supported in part by a grant from the National Science Foundation: Sustaining Responsive and Rigorous Teaching Based on *Carbon TIME* (NSF 1440988). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Please visit our website: http://carbontime.bscs.org/

Abstract:

Carbon TIME is a design-based implementation research project, which means that we are particularly interested in issues that arise when we enact reform programs in our large, diverse educational system. This paper reports on patterns we have found in the large data set regarding students' learning from *Carbon TIME.* We find that *Carbon TIME* curriculum and professional development has made a big difference in improving overall student achievement, as well as lowering achievement gap within classrooms. Meanwhile, teachers make a significant difference in how much students can learn from *Carbon TIME.* School resources are also significantly related with students' learning from *Carbon TIME* but are less important compared to student' prior knowledge and identities of teachers.

Subject/Problem

This paper reports data from the *Carbon TIME* project, which is designed for middle or high school classrooms focusing on carbon cycling at multiple scales. The project has a goal of supporting environmental science literacy, that is, to support students to use scientific knowledge and practices in their decisions about environmental issues. Importantly, the project aims at supporting classrooms at scale to achieve the three-dimensional learning goals of the Next Generation Science Standards (NGSS). We are aware that the enactment of NGSS at scale needs to deal with the enduring diversity of learning communities in America that involve students, teachers, schools and school districts. As a design-based implementation research (DBIR) project, we are particularly interested in assessing how our program worked in the large and diverse education system. As we argued before (Anderson et al., 2018), "one size fits all" program cannot be responsive to this diversity and we are supposed to pay particular attentions to those underserved schools.

This paper uses quantitative data to study the variations in student learning outcomes in our project. With student learning outcomes and school background information, we use hierarchical linear models (HLM) to investigate how various factors affect students' success in *Carbon TIME*, including school factors, students' prior knowledge and the identities of their teachers. Using longitudinal data, we also investigate how student and teacher success change over time, and how students learn from first unit to third unit in *Carbon TIME.* Our data showed strong evidence that teachers make a significant difference to students' learning. In the end of this paper, we also discuss how we use the data to construct value-added models to provide evidence for evaluating individual teacher's effectiveness. Papers 4 and 5 in this paper set (Morrison Thomas, et al., 2020; Covitt, et al., 2020) further study how the characteristics of classroom discourse, teacher orientations, and contexts are associated with the student learning. However, it is important to note that using single value-added measure to evaluate teacher performance is highly problematic and unreliable.

Research Questions

1. How successful is *Carbon TIME*?
2. What factors affect students' learning from *Carbon TIME*?
3. How do student and teacher success change over time?
4. How do students learn from first unit to last unit in *Carbon TIME*?
5. How can we use these data to construct value-added models that provide evidence about the success of individual teachers?

Methods

Data Sources

The data analyzed for this paper come from 245 classrooms of 133 middle and high school teachers who participated in the project during a four-year period, from the 2015-16 to the 2018-19 academic year. In these classrooms, teachers took *Carbon TIME* surveys and students took *Carbon TIME* assessments. Specifically, this paper analyzes relationships among following two quantitative datasets.

(1) Student learning outcomes. We asked the students to take an overall test at the beginning and end of the school year. These two tests are designed to cover content from six units, and thus are treated as full pretest and full posttest, which are summative assessments of students' learning through *Carbon TIME*. Additionally, students took unit-specific pretests and posttests before and after studying each unit. Most teachers who participated in *Carbon TIME* taught the first three units, including *Systems and Scale* (S&S), *Animals*, and *Plants*. Fewer teachers taught the last three units: *Decomposers, Ecosystems, and Human Energy Systems.*

Validity and reliability. In separate papers and publications, we have presented evidence for the validity and reliability of these tests as measures of three-dimensional learning (Doherty, Draney, Shin, Kim, & Anderson, 2015; Thomas, et al., 2018). The three-dimensional learning performances that we assessed included students' explanations, data analysis, and arguments from evidence. Specifically, there were items that measured three different science practices, including inquiry, explanation, and reasoning about large-scale systems. For this paper, we have chosen to focus on analyses of the inquiry and explanation items from the first three units: *Systems and Scale, Animals,* and *Plants*.

IRT Analysis. With the help of item response theory (IRT), we generated estimates that can be used to evaluate students' three-dimensional learning for all the items included in each test. Although different tests have different items, overlapping items made it possible to calibrate item and test difficulties across years and across test forms on the same scale. That is, an estimate was generated as calculated proficiency, for each student, on each test, and importantly, these estimates across different years and different tests, are on the same scale. The resulting scale units (logits) are a measure of how likely a student of some proficiency is to get an item right or wrong, where 0 represents the overall student mean across all tests.

To give a better sense of the scale in the logit, we also built a link between the logit and learning progression level frameworks (described in Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009; Covitt & Anderson, 2018). We provide thresholds values for which we can claim that a certain logit indicates the student is most likely to be at a certain level in the learning progression framework. For example, if a student's score is below -0.34, the student is most likely at level 2. If the score is between -0.34 and 0.96, the student is most likely at level 3. If a student's score is above 0.96, the student is most likely at level 4. NGSS high school performance expectations generally correspond to Level 4 on the learning progressions. This paper presents findings based on students' test performances on the logit scale from 2015-16 to 2018-19.

(2) School background information. We collected publicly available information for schools of our participating teachers, including percent of free and reduced lunch and percent of marginalized students of color.

Data Analysis

In our analysis, we have applied several two-level hierarchical linear models (as listed in Table 1). The first level is at the student level and the second level is at the classroom level (teacher-year level). We did not have school as the third level because most of the teachers are from different schools. We used the learning gains (the difference between pretest and posttest scores) as our main outcome variable because we think this provides us a reliable and valid measure of students' learning through *Carbon TIME*.

Selection of posttest data. Importantly, we used average unit posttests, rather than overall full posttest, as the measure for student' posttests in our analysis. Although overall full tests were initially designed for evaluating the program, they turned out to have several limitations. Because most teachers did not teach all six units, full tests can include content that students have not studied. In comparison, unit tests are more closely aligned with what students studied. Compared with unit tests, the full posttests showed higher rates of incompleteness and off-topic answers as well as higher proportion of students finishing the tests in unreasonably short times, indicating rapid-guessing behavior (Wise, 2017). Thus, we think unit posttests serve as better measures for students' learning outcomes as they reflect students' best efforts on units that they had studied.

Most teachers taught *Systems and Scale*, *Animals*, and *Plants*. Figure 1 presents the timeline for their students' taking overall full tests and unit tests.

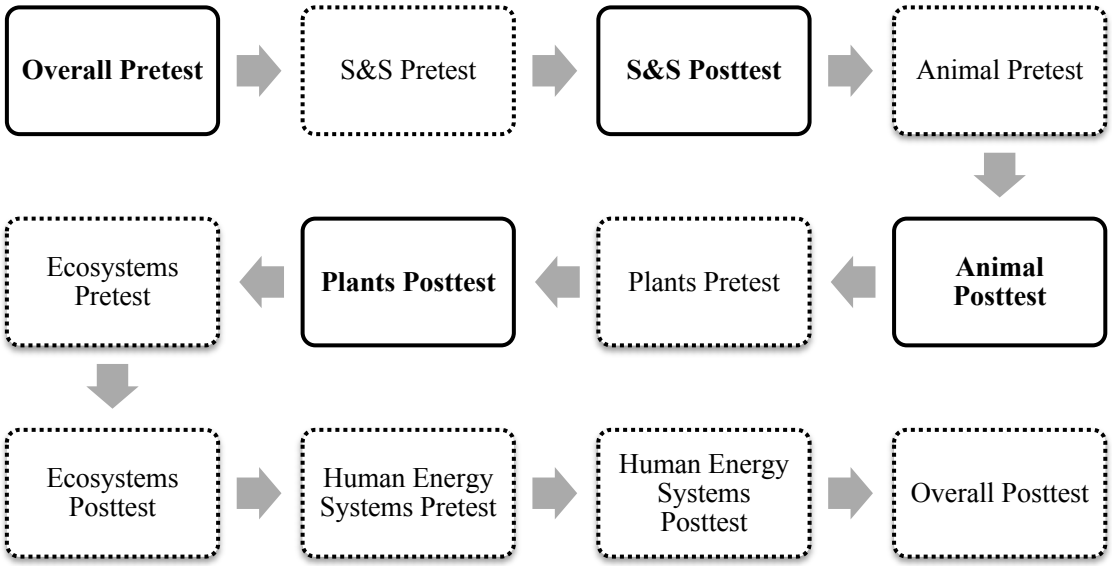

Figure 1. Timeline for students' taking overall full tests and unit tests.

Hierarchical linear models. The models used for our HLM analyses are presented in Table 1, below.

Table 1 Main hierarchical linear models used in this paper	
Unconditional Model	$Level\ 1: Gain_{ij} = \beta_{0j} + r_{ij}$
	$Level\ 2: \beta_{0j} = \gamma_{00} + u_{0j}$
Model 1	$Level\ 1: Gain_{ij} = \beta_{0j} + \beta_{1j} \cdot \left(Pretest_{ij} - \overline{Pretest_j}\right) + r_{ij}$
	$Leve\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \overline{Pretest_j} + \gamma_{02} \cdot FRL_j + \gamma_{03} \cdot GradeBand_j + \gamma_{04} \cdot Marginalized_j + u_{0j}$ $\beta_{1j} = \gamma_{10}$
Model 2	$Level\ 1: Gain_{ij} = \beta_{0j} + \beta_{1j} \cdot \left(Pretest_{ij} - \overline{Pretest_j}\right) + r_{ij}$
	$Leve\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \overline{Pretest_j} + \gamma_{02} \cdot FRL_j + \gamma_{03} \cdot Marginalized_j + u_{0j}$ $\beta_{1j} = \gamma_{10}$
Model 3	$Level\ 1: Gain_{ij} = \beta_{0j} + \beta_{1j} \cdot \left(Pretest_{ij} - \overline{Pretest_j}\right) + r_{ij}$
	$Leve\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \overline{Pretest_j} + \gamma_{02} \cdot FRL_j + \gamma_{03} \cdot Marginalized_j + \gamma_{04} \cdot Indicators\ for\ Academic\ Year_j + u_{0j}$ $\beta_{1j} = \gamma_{10}$
Model 4	$Level\ 1: Gain_{ij} = \beta_{0j} + \beta_{1j} \cdot \left(Pretest_{ij} - \overline{Pretest_j}\right) + r_{ij}$
	$Leve\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \overline{Pretest_j} + \gamma_{02} \cdot FRL_j + \gamma_{03} \cdot Marginalized_j + \gamma_{04} \cdot Indicators\ for\ Experience\ of\ Teaching\ CTIME\ _j + u_{0j}$ $\beta_{1j} = \gamma_{10}$
$Gain_{ij}$ is the learning gain for student *i* taught by teacher *j*; $Pretest_{ij}$ is the learning proficiency for student *i* measured in the Pretest;	

$\overline{Pretest_j}$ is the teacher j's classroom average learning in the Pretest;
FRL_j is the percentage of free and reduced lunch in teacher j's school;
$GradeBand_j$ is whether teacher j is in a high school or middle school;
$Marginalized_j$ is the percentage of marginalized students of color in teacher j's school;
$Indicators\ for\ Academic\ Year_j$ includes three dummy variables indicating whether this classroom is from 2016-17, 2017-18, and 2018-19;
$Indicators\ for\ Experience\ of\ Teaching\ CTIME_j$ includes three dummy variables indicating whether this year is the teacher's 2nd year, 3rd year, or 4th year of teaching *CTIME*.

Within-classroom and between-classroom variation. In the analysis, we started from the unconditional model to decompose the variations of students' learning into within-classroom and between-classroom. The result showed that the intra-class correlation (ICC) is larger than 30%, indicating a considerable amount of between-classroom variation. This is a high ICC compared with other education research (Frank, 1998). [1] Then we fit the most complicated model (Model 1) with all the related variables of interest to study what factors can help explain how much students learn from *Carbon TIME*. These predictors are from both the student level (level 1) and classroom level (level 2).

Students' pretests as predictors. At the student level, we have the predictor $\left(Pretest_{ij} - \overline{Pretest_j}\right)$ that measures how far a student's pretest deviates from the corresponding classroom average pretest. For example, if a student Jack has a pretest of 0.68 and the average pretest in his classroom is 0.5, then for Jack, the value for this predictor is 0.18 (as a result of 0.68-0.5). For another student, Joe, having the same pretest score of 0.68 but from a different classroom where the average pretest was 0.7, his value for this predictor is then -0.02. The -0.02 indicates that he is 0.02 logits lower than his classroom average, in terms of pretest. At

[1] In comparison, the between-year variation only accounts for 3%. That is, the variation in students' learning outcome between academic years is very small relative to the variation within years.

the classroom level (level 2), we first have the average pretest ($\overline{Pretest_j}$). In the previous example for student Jack and Joe, their values are 0.5 and 0.7, respectively.

It is important to note that these two pretest-related predictors both measure students' prior knowledge but are orthogonal to each other. The first part, at the student level, measures within-class variation while the second part, at the classroom level, measures between-class variation. In other words, they represent two independent ways how students' pretest may influence their learning outcomes. Later when we quantify the impact of students' prior knowledge on students' learning gains, the effects of these two parts will be added up to generate an overall impact for students' prior knowledge on their learning outcomes.

School-level predictors. Additionally, there are three school characteristics as potential important predictors at the classroom level (level 2). First is the percentage of free and reduced lunch (FRL_j). Second is the percentage of marginalized students of color (non-White/Asian students) ($Marginalized_j$). We use these two predictors as approximate measures for school organizational resources. Acknowledging the limitations of these measures (Greenberg et al., 2019), they are the best measures we can get to estimate school factors. Previous studies showed that the percent of free and reduced lunch can be a proxy measure for material, social, and human material resources such as students' access to qualified and experienced teachers (Darling-Hammond, 2004; Rice, 2010) and the overall quality of conditions in which teachers work (Johnson et al., 2012).

Another predictor at level 2 is whether the school is high school or middle school ($GradeBand_j$). The results of Model 1 show that Grade Band is not significantly related to students' learning gains. Thus, we excluded this predictor as it did not make significant contributions in explaining the variation in students' learning. As such, we got a more

parsimonious model (Model 2) and used this model to analyze how these predictors affect students' learning.

Cross-year comparisons. We first used Models 1 and 2 to study each year's data separately; each year's data generated consistent and similar findings. Therefore, we combined all four years' data and applied Models 3 and 4, where we added predictors that indicate which academic year (Model 3), or how many years the teacher had taught *Carbon TIME* (Model 4).

Indicators for academic years in Model 3 helps us capture the improvement in *Carbon TIME* as a design-based implementation research (DBIR) project. Over the four-year period, improvements in professional development, curriculum and assessment, and teacher network support have been implemented based on feedback from teachers and students. Model 3 allows us to see if there is any evidence for increase in students' learning gains over the four-year period. Rather than assuming a linear trend in students' growth, we used three binary variables to compare 2016-17, 2017-18, and 2018-19, with the first year 2015-16, respectively.

Similarly, in Model 4, we added three binary variables to capture how many years the teacher had taught *Carbon TIME*. Three binary variables indicate whether this is the 2nd, 3rd, or 4th year of teaching *CTIME*, respectively. That is, the reference group is the 1st year of teaching *CTIME.* The coefficients of these indicators can allow us to tell whether more years of teaching *CTIME* is associated with higher students' learning gains.

Findings

Research question 1: How successful is Carbon TIME?

We report the results for this research question based on four-years' data, from 2015-16 to 2018-19. For better data validity, we excluded teachers for whom fewer than 15 students' data are available or only one unit posttest is available.

Finding 1: Students in the *Carbon TIME* project showed substantial learning gains.

First, we want to figure out the overall effect of the *Carbon TIME* project on students' learning, including curriculum and assessment, professional development and teacher support networks. The project design allows us to examine the overall effect of *Carbon TIME* using two approaches.

- *Approach 1: Comparing Carbon TIME pretests with Carbon TIME posttests.* In the first approach, we compare pretest and posttest from the same students taught by same teachers with *Carbon TIME*. That is, we measure students' proficiency before they learned *Carbon TIME* and then measure their proficiency again after they learned *Carbon TIME*. By comparing their pretests and posttests, we can learn about their progress through learning *Carbon TIME*.
- *Approach 2: Comparing baseline tests (same teachers using other curricula) with Carbon TIME posttests.* In the second approach, we compare posttests from two groups of students taught by same teachers but with different curricula. One group of students studied *Carbon TIME* while the other group studied other curricula. For this second group of students, posttests were given in the classrooms of the same teachers the year before they started using project materials (Anderson, et al., 2018). By comparing how different these two groups of students performed in the posttests, we can study how *Carbon TIME* helped students learn compared to other curricula.

Table 2 summarizes the comparison results in both approaches. Note that in the second approach, we only include students in the *CTIME* group if it is the first year their teachers taught *Carbon TIME* (i.e., either 2016-17 or 2017-18). That is, we exclude students whose teachers had more than one year's experience teaching *Carbon TIME* to reduce the potential confounding

effect of teachers' gaining experience as teaching more *CTIME*. As such, the *CTIME* group in our analysis includes 3191 students from 57 teachers, and the non-*CTIME* group includes another 3615 students from these 57 teachers. In contrast, the "matching" process for the first approach occurs at the student level, which allows us to compare pretest and posttest for 16,195 students taught by 133 teachers across the 4-year period. Both approaches presented robust evidence that overall students in the *Carbon TIME* project showed substantial learning gains.

Table 2 How successful is *Carbon TIME* in general

		N	Mean	Std. Dev.	Min	Max
Approach 1	Pretest	16,195	-1.435	0.813	-4.948	3.886
	Posttest	16,195	0.396	1.286	-3.798	6.061
	Gain	16,195	1.831	1.221	-2.727	8.988
	Paired t test: difference = 1.831, SE = 0.010, $p < 0.001$. Effect size = 1.423.					
Approach 2	non-*CTIME*	3191	-1.274	1.027	-4.983	3.387
	CTIME	3615	0.264	1.268	-3.798	6.061
	Two-sample t test: difference = 1.538, SE = 0.028, $p < 0.001$. Effect size = 1.213.					
Comparing baseline with pretest	*CTIME* pretest	3615	-1.478	0.794	-4.948	3.365
	non-*CTIME* posttest	3191	-1.274	1.027	-4.983	3.387
	Two-sample t test: difference = -0.204, SE = 0.022, $p < 0.001$. Effect size = 0.199.					

Figures 2 and 3 present the results from these two approaches. Figure 2 shows the findings from the first approach, where we compare the pretest and posttest for the students who studied *Carbon TIME.* The pink part shows the distribution of pretest and the light blue part shows the distribution of posttests. If we compare these two parts, we can see that students' proficiency level increased from pretest to posttest. The unit of analysis on the x-axis is logits, an IRT-based measure of overall student proficiency, with 0 as the grand mean for all students. The average increased from -1.435 to 0.396, indicating an overall average gain score of around 1.8 logit (which is about one average learning progression level from level 3 to level 4, effect size = 1.423).

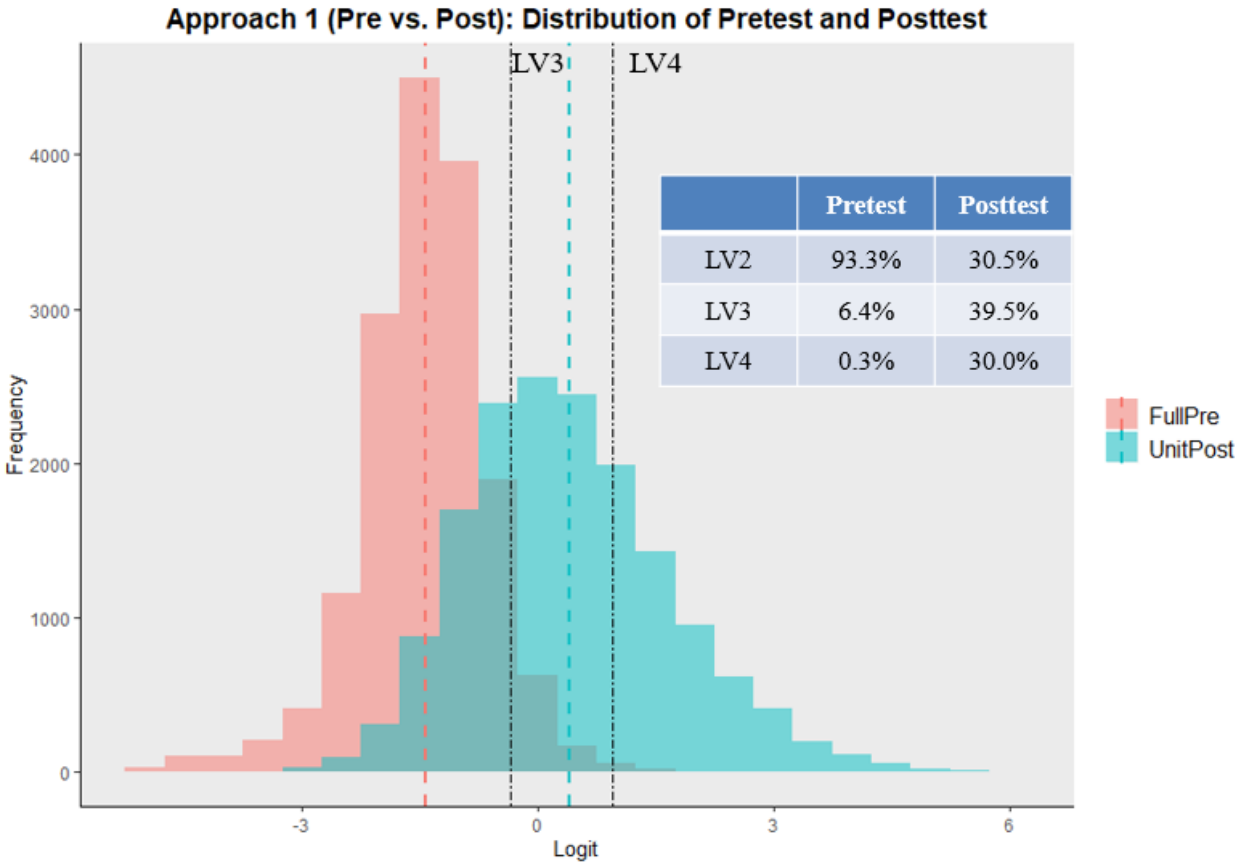

Figure 2. Approach 1: Comparing *Carbon TIME* pretests and posttests.

The vertical black lines indicate average proficiencies for transitions to learning progression levels 3 and 4. (As explained above level 4 corresponds to high-school level NGSS performance expectations.) More students achieved level 3 and level 4 from pretest to posttest: in pretests, 93.3% students were most likely at level 2 while in posttests, only 30.5% students were mostly likely at level 2; in pretest, only 0.3% students were mostly likely at level 4 while in posttest 30.1% students were mostly likely at level 4.

Figure 3 shows the findings based on the second approach, where we compare students who studied *Carbon TIME* (*CTIME* group) with students who studied other curricula (non-*CTIME* group). The blue part shows posttest distribution for students who studied other curricula. The pink part shows the distribution for students who studied *Carbon TIME*. We can

tell that students who studied *Carbon TIME* performed much better in posttests than students who studied other curricula. The average difference is around 1.538 logit (effect size = 1.213). More *Carbon TIME* students achieved level 3 and level 4: in the non-*CTIME* group, only 12.9% students were most likely at level 3 and only 2.5% students were most likely at level 4; in comparison, the *CTIME* group has 39.2% students most likely at level 3 and 26.5% students mostly likely at level 4.

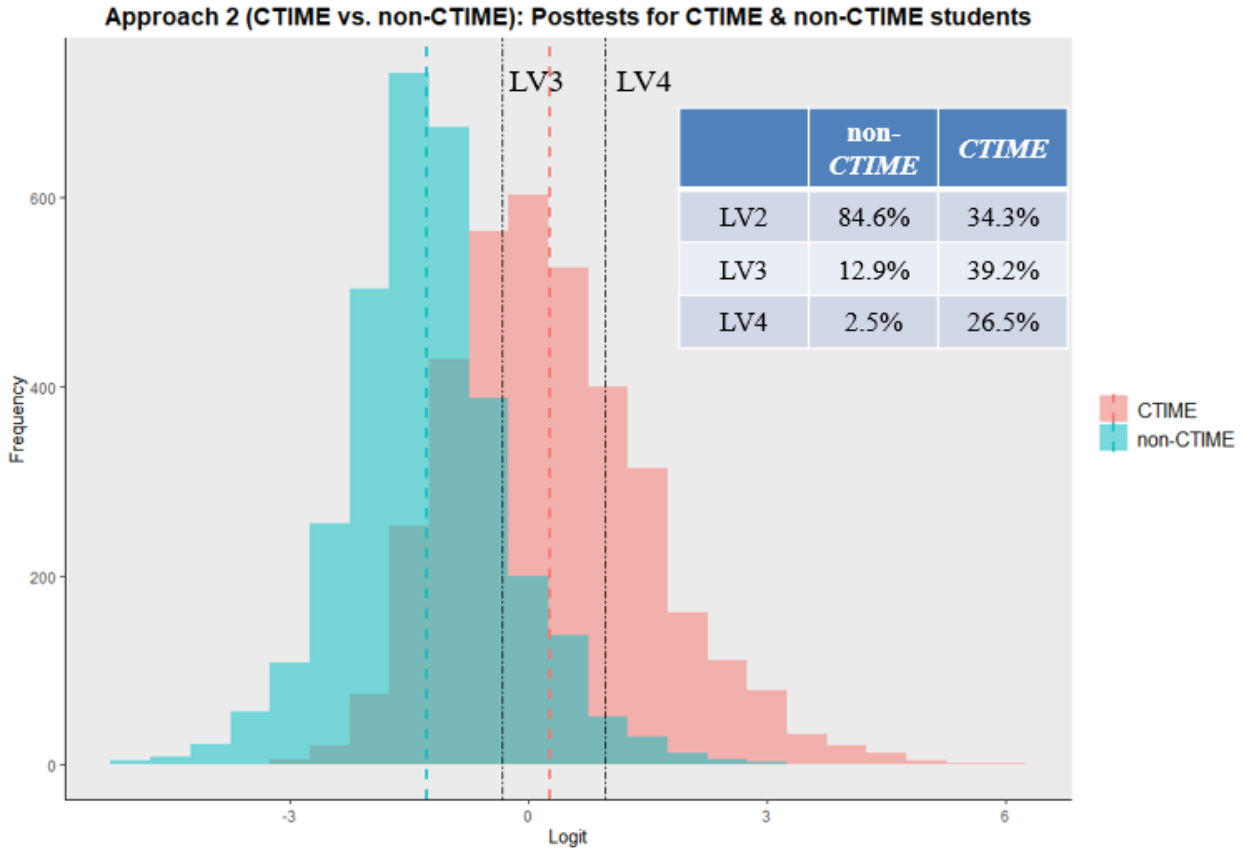

Figure 3. Approach 2: Comparing posttests between *Carbon TIME* and non-*Carbon TIME* students.

The last part of Table 2 compares the average *pretest* of students who learned *Carbon TIME* with the average *posttest* of students who learned other curricula. We control teacher effects by only looking at classrooms where (1) the teachers are at their first year of teaching

Carbon TIME; and (2) the teachers have both pretest of *Carbon TIME* group and posttest of non-*Carbon TIME* group data available. As shown in Table 2, the difference is statistically significant but small in terms of educational significance, which is only 0.204 logit (effect size = 0.199). This indicates that students only performed slightly better after learning other curricula. In comparison, students' performance improved by 1.831 logit (effect size = 1.423) after learning *Carbon TIME.*

Research Question 2: What factors affect students' learning from Carbon TIME?

Table 3 presents key findings from analyses comparing Models, 2, 3, and 4. Students' learning gains were significantly associated with (a) students' prior knowledge, (b) membership in classrooms of individual teachers, (c) school organizational resources measured by percent of free and reduced lunch and percent of marginalized students of color. These findings are discussed in more detail below.

Table 3: Parameter Estimates for Two-level Hierarchical Linear Models 2, 3 and 4			
Outcome: Gain score	Model 2	Model 3	Model 4
Deviation from class average pretest.	-0.507***	-0.507***	-0.507***
	(0.0101)	(0.0101)	(0.0101)
Class average pretest.	-0.135	-0.130	-0.0655
	(0.138)	(0.134)	(0.136)
Percent of free and reduced lunch.	-1.077***	-1.029***	-1.009***
	(0.293)	(0.284)	(0.286)
Percent of marginalized students of color.	-0.544	-0.469	-0.613*
	(0.304)	(0.295)	(0.296)
Whether this is the 2nd year of teaching *CTIME*.		0.257**	
		(0.0885)	
Whether this is the 3rd year of teaching *CTIME*.		0.531***	
		(0.144)	
Whether this is the 4th year of teaching *CTIME*.		0.430	
		(0.250)	
Whether this is the year of 2016-17.			0.204
			(0.147)
Whether this is the year of 2017-18.			0.509***
			(0.144)
Whether this is the year of 2018-19.			0.424**
			(0.154)
Conditional ICC	31.2%	29.8%	29.9%
Constant	2.028***	1.840***	1.784***
	(0.195)	(0.195)	(0.213)
Observations	16,195	16,195	16,195
Number of Teacher_Year	245	245	245

Notes. Standard error in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Finding 2.1: The differences in learning gains in different teachers' classrooms are real and substantial. As discussed earlier, the intra-class correlation (ICC) in the unconditional model illustrates that more than 30% of the variance in students' learning gains is at the classroom level. In Model 2, 3, 4, we added more predictors to explain the variance in students' learning gains. However, the conditional ICC, as shown in Table 3, is still around 30%. This shows that the differences in learning gains in different teachers' classrooms are real and substantial, even after controlling the effects of students' prior knowledge and school factors.

Finding 2.2: Participation in the *Carbon TIME* project reduced student achievement gaps within classrooms. We decomposed the factor of students' prior knowledge (measured by students' pretests) into two parts: within-class and between-class. The within-class component measures the difference between an individual student's pretest and the student's classroom

average pretest. The between-class component measures the class average pretest. The results indicated that within classes, students with lower pretest proficiencies showed significantly higher learning gains, as reflected by the coefficient of -0.507 ($p<0.001$) for deviation from class average pretest in Table 3. The standardized coefficient is around -0.308. This partial correlation is around -0.290 after adjusting for measurement error in pretest and learning gains (Willett, 1988), which is still a strong, educationally significant correlation. In other words, our data show strong evidence that *Carbon TIME* reduced achievement gaps within classrooms.

On the other hand, class average pretest scores were not significantly associated with student learning, so learning gains were not significantly different in classes with lower vs. higher average pretest scores. (As reported above, learning gains also were not significantly different in middle school vs. high school classrooms.)

Finding 2.3: Students' learning was negatively affected by limited school organizational resources, as indicated by the percentage of free and reduced lunch, and the percentage of marginalized students of color. Students in classrooms from schools with higher percent of free and reduced lunch, or higher percent of marginalized students of color, showed smaller learning gains. These are reflected by the coefficient of -1.077 ($p < 0.001$) for free and reduced lunch, and the coefficient of -0.544 ($p = 0.074$) for marginalized students of color, in Table 3. Note that these two proximate measures for school resources are highly correlated with each other (correlation is 0.498, $p < 0.001$), and the collinearity leads the two coefficients to show weaker statistical significance when both predictors were included in one model. Without the percent of marginalized students of color, the percent of free and reduced lunch has a coefficient of -1.41 with a p value of 0.001. Without the percent of free and reduced lunch, the percent of marginalized students of color has a coefficient of -1.149 with a p-value smaller than

0.001. To better understand the effect size, we also checked that the standardized coefficients are -0.168 for free and reduced lunch, and -0.071 for marginalized students of color. These partial correlations are around -0.19 and -0.08, respectively, after adjusting for measurement error in learning gains (using approach from Willett, 1988), which are strong, educationally significant correlations.

Previous studies have shown the percentage of free and reduced lunch can provide a proxy measure for a school's material, social and human resources (Darling-Hammond, 2004; Rice, 2010; Johnson, Kraft & Papay, 2012). Similarly, our survey data has showed that teachers from schools with higher percentage of free and reduced lunch had lower scores on items measuring science knowledge and pedagogical content knowledge.

Finding 2.4: Teachers are more important than students' prior knowledge and school organizational resources in explaining variation in students' learning gains. We showed that students' learning gains were significantly associated with (a) students' prior knowledge, (b) membership in classrooms of individual teachers, and (c) school organizational resources. After a series of analysis, we decomposed the variance in students' gain scores into these three important factors so that we can tell how important each factor is in explaining how much students learned from *Carbon TIME*. Because variables may collinear with each other to some extent, we cannot make a clean cut for the contribution of each factor. Instead, we report a range of variance in the gain score that can be accounted for by each factor.

Figure 4 presents the findings, in which a range of percent of variance can be found for each factor. For example, on the very left, the blue bar shows that identities of teachers make at least 27.7% variance to students' learning, and the grey bar shows that teachers make at most 31.4% variance to students' learning. Similarly, the two bars in the middle present that students'

prior knowledge contributes 9.6% to 9.7% in students' learning. On the right, school resources only accounts for less than 4.0% of the total variance.

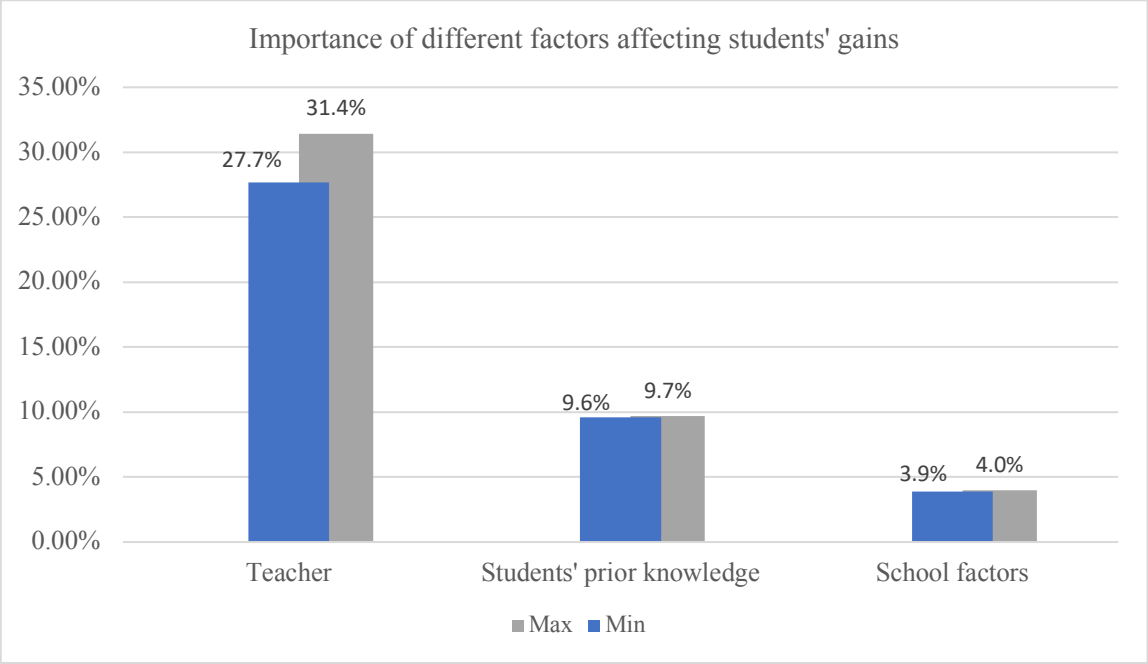

Figure 4. Importance of different factors affecting students' learning gains in *Carbon TIME*.

Research question 3: How do student and teacher success change over time?

We used Models 3 and 4 to compare student learning from different years. Our results from these analyses are summarized below.

Finding 3.1: More teachers achieved level 4 class averages over time. From 2015-16 to 2018-19, there are more teachers who achieved level 4 class averages (corresponding to NGSS high school performance expectations). Figure 7 shows how the mean and 95% confidence interval of class average unit post changed from 2015-16 to 2017-18. The black horizontal dashed lines represent the thresholds for level 3 and level 4, respectively. The red, green, blue, purple horizontal lines represent the year-average unit post scores, respectively. One can tell that the average unit post in all teachers' classrooms increased over the four-year period. More

importantly, there are more teachers who achieved level 4 over time, 7.7% of teachers in 2015-16 to 13.3% in 2016-17 and to more than 22% in 2017-18 and 2018-19.

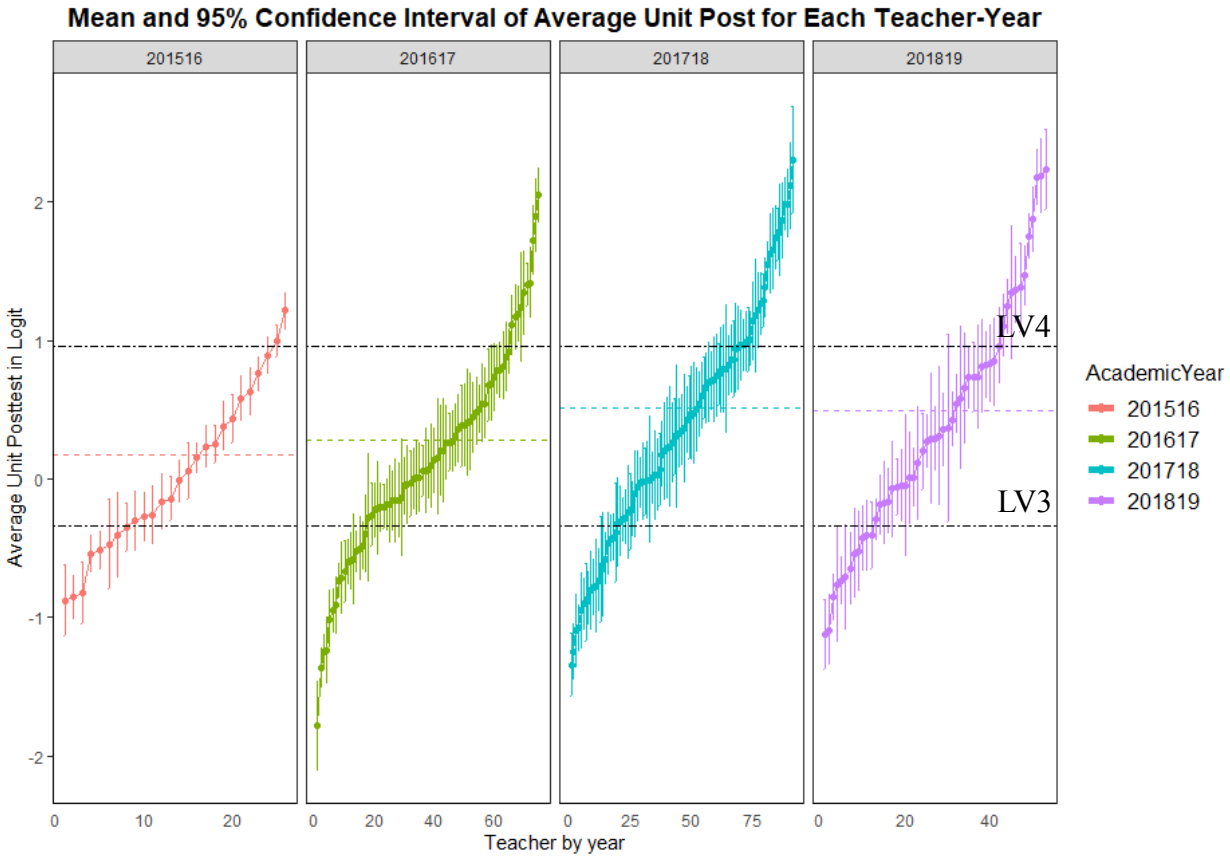

Figure 5. Mean and 95% confidence interval of average unit post for 133 teachers and their 245 classrooms, across academic years from 2015-16 to 2018-19.

Finding 3.2: Student learning gains increased during the four-year period. *Carbon TIME* is a design-based implementation research (DBIR) project. Over the four-year period, improvements in professional development, curriculum and assessment, and teacher network support have been implemented based on feedback from teachers and students. Importantly, we also see evidence showing increasing effect of *Carbon TIME* on students' learning gains over time.

Figure 8 shows the average learning gain (i.e., average difference between full pretest and unit posttest) and 95% confidence interval in each teacher's classroom over the four-year period from 2015-16 to 2018-19. Specifically, this figure shows the average learning gain (represented by the black dot) and 95% confidence interval (represented by the error bar) in 245 classrooms. Although all classrooms have positive average learning gains, the difference among classrooms can be substantial, even after taking account into sampling variability.

The dashed horizontal line in each year represents the average learning gain for that particular year. As we can tell, the average learning gain keeps increasing from 2015-16, to 2017-18, and 2018-19 is only a little lower than 2017-18, but still higher than 2016-17. This indicates that on average, the effect of *Carbon TIME* on improving students' learning increases over time.

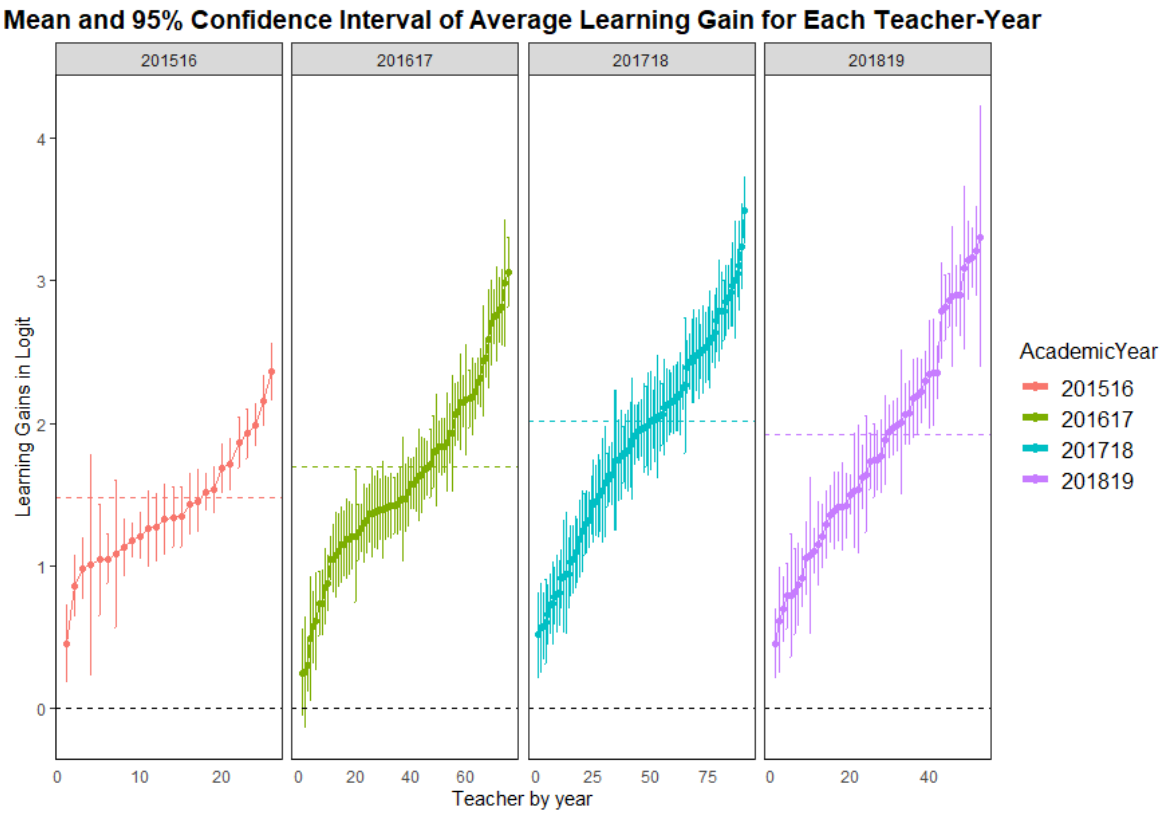

Figure 6. Mean and 95% confidence interval of average learning gain for each teacher from 2015-16 to 2018-19.

Finding 3.3: Class average gain increased as teachers gained more experience in teaching *Carbon TIME*. Teachers gained more experience and learned more from professional development as they taught more *Carbon TIME* in their classes. Figure 7 shows the average learning gain (i.e., average difference between full pretest and unit posttests) and 95% confidence interval across each stage of teachers' experience of teaching *Carbon TIME:* from their first year of teaching *Carbon TIME* to their fourth year of teaching *Carbon TIME*. The horizontal dashed lines represent the average for each stage. Class average learning gains increased as teachers gained more experience, especially in the first three years of teaching *Carbon TIME*.

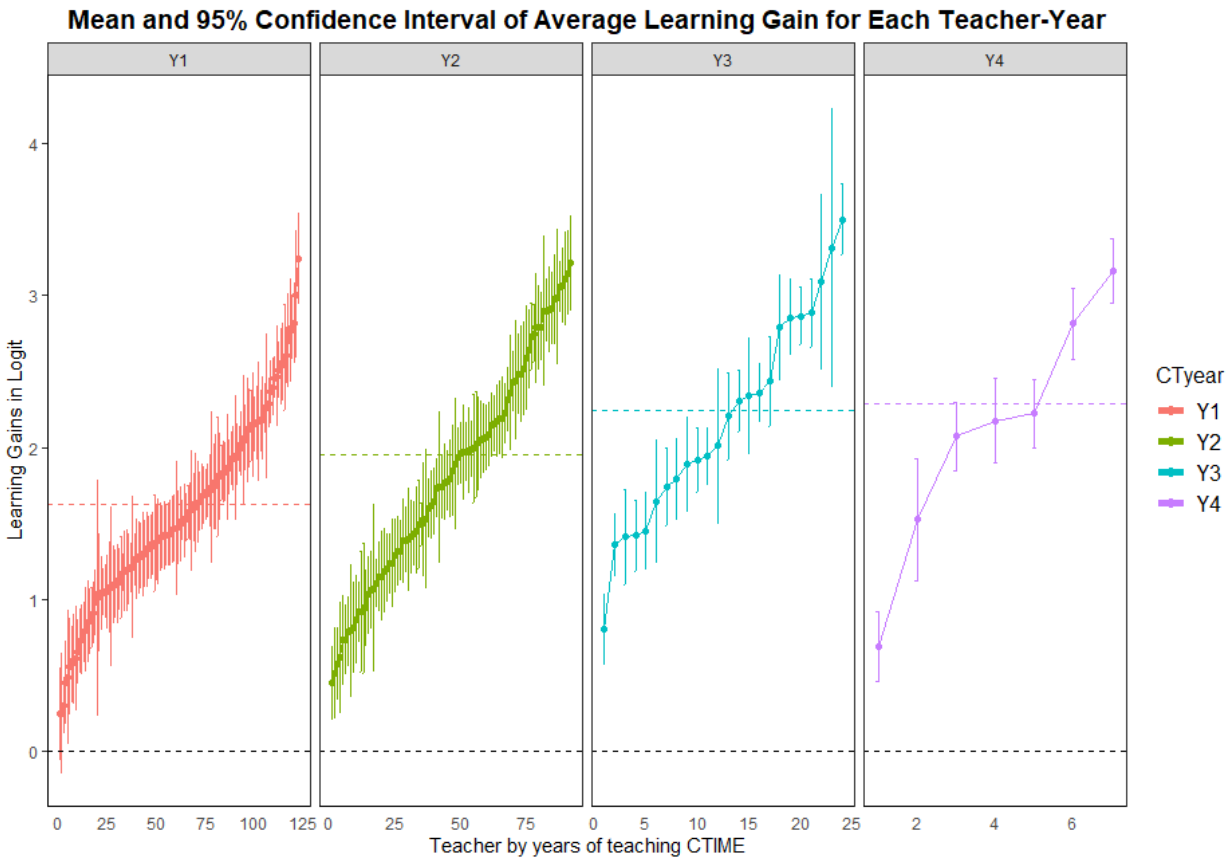

Figure 7. Mean and 95% confidence interval of average learning gain as they gained more experience in teaching *Carbon TIME*.

Finding 3.4: We cannot statistically distinguish between the effects of (a) teachers gaining experience and learning from professional development, and (b) improvements in *Carbon TIME* units and professional development. Previous discussions are based on the raw gain scores (i.e., unadjusted difference between average unit post and full pretest). Knowing that school factors and students' prior knowledge can have significant effects on students' learning, we again applied the 2-level hierarchical linear model (Model 4) to statistically model and better quantify the effect on students' learning gains from: (a) teachers gaining experience and learning from professional development; and (b) improvements in *Carbon TIME* units and professional development.

The results are presented in the last two columns in Table 3 (Model 3 and Model 4). From Model 3, we can tell students' learning gains increased significantly (0.53 logits) as teachers gained more experience from their first year to second year and third year of teaching *Carbon TIME*. From Model 4, we can also find evidence for the effect of improvement in *Carbon TIME* units and professional development on students' learning: from the year of 2015-16 to 2017-18, there is a significant increase of 0.51 logits in students' learning gains.

However, we cannot statistically distinguish between these two effects. That is, we cannot tell how much increase in learning gains is due to teachers' gaining experience and how much is due to improvement in *Carbon TIME*. These two predictors are collinear with each other. Table 4 illustrates this collinearity by presenting how many teachers were in the project for each year and how many years of experience of teaching *Carbon TIME* the teacher had for each particular year. As teachers gained one more year of teaching experience, the improvement in

Carbon TIME also occurred at the same time. Additionally, we have very few teachers in Y3 and Y4, making it difficult to draw statistical inference.

Table 4. Collinearity between teachers' gaining experience and improvement in *Carbon TIME* over time

Notes. Number of teachers are presented in each cell.		Teachers' gaining experience			
		Y1	Y2	Y3	Y4
Improvement in *Carbon TIME*	2015-16	26			
	2016-17	56	19		
	2017-18	40	42	9	
	2018-19		31	15	7

Research question 4: How did students learn from first unit to last unit?

We also find evidence that as students move from the first unit (*Systems and Scale*) to the third unit (*Plants*), their posttest performance improves, indicating a cumulative effect of *Carbon TIME*. Focusing on classrooms where teachers taught the three main units in the same order (first *Systems and Scale,* then *Animals* and finally *Plants*), Figure 8 shows the distribution of learning gains in these three units. As summarized in Table 5, students' learning improved significantly from the first unit to the third unit.

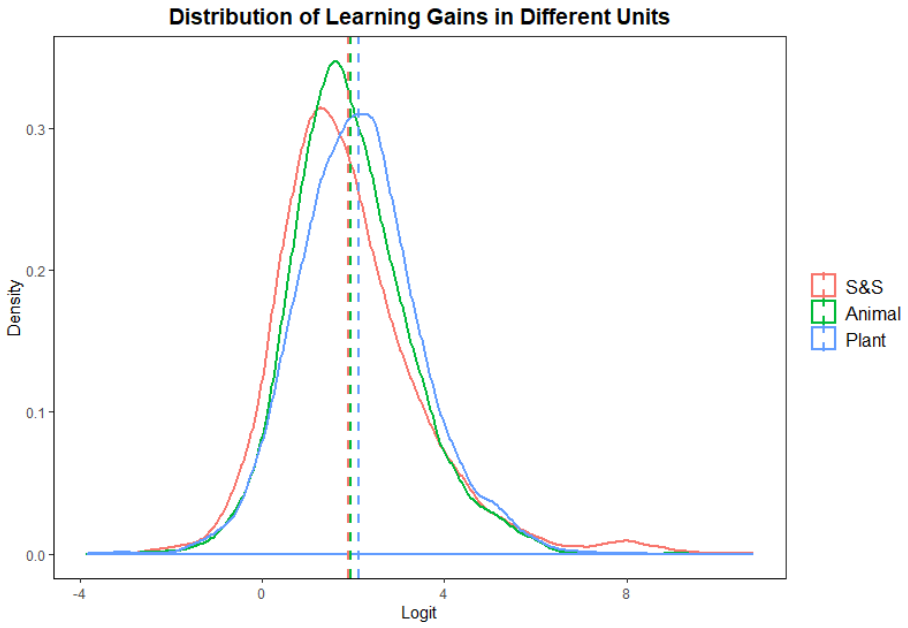

Figure 8. Distribution of learning gains in different units.

Table 5: Comparing learning gains after different units					
	N	Mean	Std. Dev.	Min	Max
S&S gain	10,832	1.897	1.601	-3.675	10.751
Animals gain	10,832	1.953	1.297	-3.839	9.073
Plants gain	10,832	2.129	1.345	-3.509	8.141
Paired t test between S&S and Animal: difference = 0.056, SE = 0.014, $p < 0.001$.					
Paired t test between Animal and Plant: difference = 0.176, SE = 0.011, $p < 0.001$.					
Paired t test between S&S and Plant: difference = 0.232, SE = 0.014, $p < 0.001$.					

Research Question 5: How can we use these data to construct value-added models that provide evidence about the success of individual teachers?

Value added measures (VAM) are essentially the "deflections" between students' expected test scores and their actual ones (Raudenbush & Bryk, 2002). Proponents of value-added models cite research that shows teachers' considerable and long-lasting influences on student achievement (Chetty et al., 2011; Hill et al., 2011; Rivkin et al., 2005). They argue that there is important variation in teachers' effectiveness (Aaronson et al., 2007) that can be better identified by VAM (Hanushek & Rivkin, 2010). We developed VAM for *Carbon TIME* teachers; our results are summarized below.

Finding 5.1: Although the differences in learning in different teachers' classrooms are real and substantial, we should be cautious in reporting value added scores as effectiveness measures for individual teachers. Figure 9 presents the value-added measures and 95% confidence interval for 133 teachers' 245 classrooms. Using Model 2 in Table 1, the expected students' gain score is calculated based on the students' pretest, their schools' percent of free and reduced lunch as well as percent of marginalized students of color. The average difference between the expected gain scores and the actual gain scores is then regarded as a

teacher's value added or teaching effectiveness. In other words, if a teacher has a positive value here, it means that his/her students performed better than expected on the posttests.

These are our best estimates of overall effectiveness for individual teachers. As mentioned before, the student learning data come from assessments of NGSS-aligned three-dimensional performances, backed by extensive evidence for validity and reliability presented in separate papers and publications (Doherty, Draney, Shin, Kim, & Anderson, 2015; Thomas, et al., 2018). Additionally, we only used data from units that the teachers actually taught and assessments that are aligned with those units. We also restricted our data samples in several ways to further ensure validity: (1) we only included data from students for whom both pretest and at least 2 unit posttests are available; (2) we only included teachers for whom we have at least 15 students' test scores that satisfy the condition (1).

The five teachers whose cases studies are analyzed in Paper 4 (Covitt, et al., 2020) and Paper 5 (Morrison Thomas, et al., 2020) of this set are identified by colored error bars in Figure 9. We can tell that variation in teachers' effects is substantial. For example, in both 2015-16 and 2016-17, teacher Harris has a significantly lower effectiveness than teachers Callahan and Eaton. Papers 4 and 5 provide in-depth analysis regarding how these teachers differ in terms of classroom discourse, teacher orientation, and school professional communities.

Some proponents argue that value-added measures can be used for high-stakes teacher evaluations: By selecting or deselecting teachers based on value-added we can improve teacher quality and increase student achievement and long-term outcomes (e.g., Winters & Cowen, 2013a; Gordon, Kane, & Staiger, 2006; Winters & Cowen, 2013b; Chetty et al., 2011). However, skeptics have raised important concerns about the validity and reliability of value-added

measures as a basis to inform teacher evaluation, including test unreliability, missing data, and model specifications (Guarino et al., 2014; Harris, 2009; Raudenbush, 2015).

We agree with the skeptics. The data in our project show important variations in teacher effectiveness, but ranking individual teachers based on the value-added measures is not reliable. From Figure 9, we can tell that for most teachers their effectiveness at a particular year has overlap with others and it is hard to rank one over another. That is, only a small amount of potential bias is needed to alter the ranking. For example, the difference between Eaton and Callahan in 2016-17 is so small that in a counterfactual thought experiment the ranking can be reversed by replacing 2 students out of Callahan's 73 students with average students (Lin, 2019). Additionally, we can observe how the rankings of Ms. Callahan, Ms. Eaton and Mr. Gilbert changed from 2016-17, to 2017-18, to 2018-19. That is, the ranking is not stable across years for the same teachers.

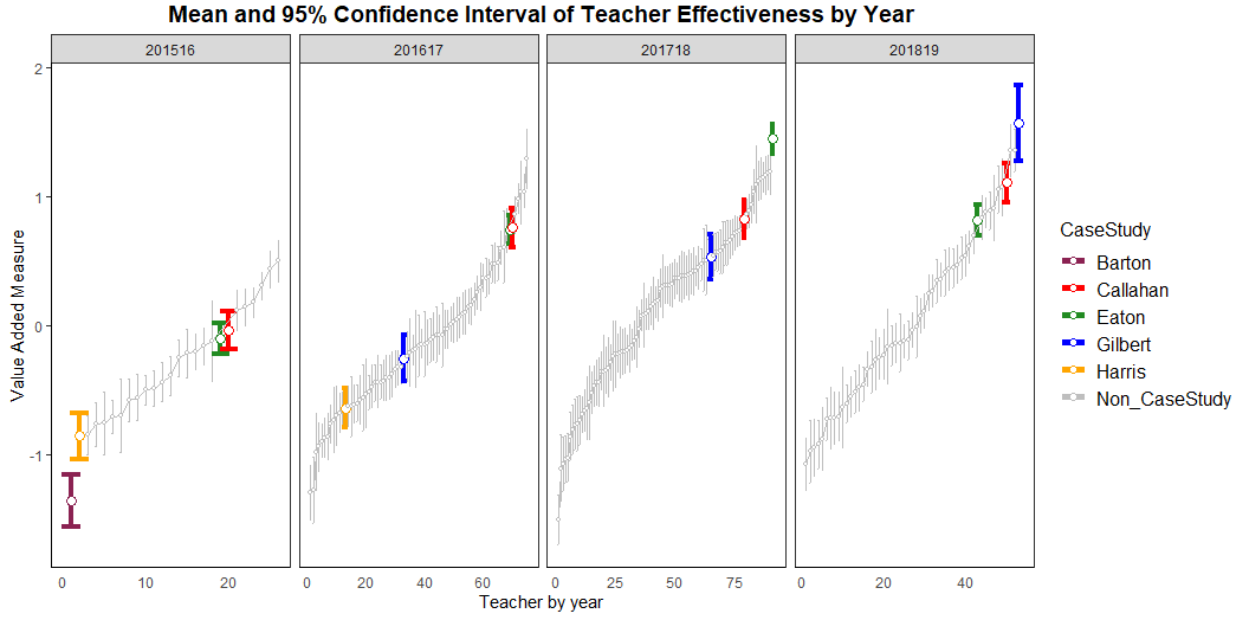

Figure 9. Mean and 95% confidence interval of value-added measures for 133 teachers and their 245 classrooms.

As a final note for this part, we have carefully designed our value-added model, using valid measures, to measure differences among teachers' classrooms. We would like to emphasize that even such efforts do not mean the value-added measures can be used to inform high-stake decision making for individual teachers. The analysis here is only used as a research tool for understanding classroom differences in students' learning gains that might be attributable to teachers and to characteristics of classroom discourse managed by teachers, particularly when we compare classrooms with substantial differences in learning gains.

Discussion

Carbon TIME is a design-based implementation research project with the goal of three-dimensional science learning for all students. We have examined pretest and posttest performances on three-dimensional assessments from a large sample of students in diverse classrooms and schools. Our findings provide evidence for some important conclusions.

Curricula make a difference. Students showed substantial learning gains after studying *Carbon TIME*. The effect of *Carbon TIME* on students' learning is both educationally and statistically significantly larger compared with other curricula: Ninety-one percent of students who studied *Carbon TIME* scored higher than the median of students who studied other curricula. Additionally, *Carbon TIME* helped reduce the achievement gap within classrooms: students with lower pretest scores showed significantly higher learning gains.

Teachers make a difference. We also observe that students in different classrooms vary substantially in how much they learned from *Carbon TIME.* Among all the important factors, teachers make the most difference. School factors (percent free and reduced lunch and percent marginalized students of color) also make a difference in students' learning but account for much less of the variance in student learning than teachers.

Questions to be investigated. This leads us to consider why some classrooms are much more successful than the others. Paper 4 and 5 will analyze how the differences among classrooms and teachers may help explain their differences in students' learning.

Reference

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock, J. B., Lin, Q., Morrison Thomas, C., Penuel, W. R., & Welch, M. M. (2018). Designing educational systems to support enactment of the Next Generation Science Standards. *Journal of Research in Science Teaching*, *55*(7), 1026–1052. https://doi.org/10.1002/tea.21484.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood* (Working Paper No. 17699). National Bureau of Economic Research. https://doi.org/10.3386/w17699.

Covitt, B. A., Morrison Thomas, C., Lin, Q., de los Santos, E. X., & Anderson, C. W. (2020, March). *Carbon TIME* classroom discourse and its connections to student learning. Annual meeting of the National Association for Research in Science Teaching, Portland, OR. (Conference canceled) https://carbontime.bscs.org/conference-presentations.

Darling-Hammond, L. (2004). Inequality and the Right to Learn: Access to Qualified Teachers in California's Public Schools. *Teachers College Record*, *106*(10), 1936–1966. https://doi.org/10.1111/j.1467-9620.2004.00422.x.

Doherty, J. H., Draney, K., Shin, H. J., Kim, J. H., & Anderson, C. W. (2015). Validation of a learning progression-based monitoring assessment.Michigan State University: http://media.bscs.org/carbontime/files/CarbonTIMEAssessmentValidation.pdf.

Frank, K. A. (1998). Chapter 5: Quantitative Methods for Studying Social Context in Multilevels and Through Interpersonal Relations. *Review of Research in Education*, *23*(1), 171–216. https://doi.org/10.3102/0091732X023001171.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Brookings Institution Washington, DC.

Greenberg, E., Blagg, K., & Rainer, M. (2019). *Measuring Student Poverty*. 35.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*. http://www.mitpressjournals.org/doi/abs/10.1162/EDFP_a_00153

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management*, *28*(4), 693–699.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

Jin, H., and Anderson, C. W. (2012). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching, 49(9),* 1149–1180.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*.

Lin, Q. (2019). *Quantifying strength of evidence in education research: accounting for spillover, heterogeneity, and mediation* (Doctoral dissertation, Michigan State University). Retrieved from https://d.lib.msu.edu/etd/48339/datastream/OBJ/View/.

Mohan, L., Chen, J., and Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46 (6),* 675-698.

Morrison Thomas, C., Covitt, B. A., Lin, Q., Hancock, J. B., Marshall, S., & Anderson, C. W. (2020, March). *Carbon TIME* teacher orientations and contexts: Making connections to classroom discourse and student learning. Annual meeting of the National Association for Research in Science Teaching, Portland, OR. (Conference canceled). https://carbontime.bscs.org/conference-presentations.

Thomas, J., Kim, J., & Draney, K. (2018, March). Machine scoring and IRT analysis. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta.

Raudenbush, S. W. (2015). Value added: A case study in the mismatch between education research and policy. *Educational Researcher*, *44*(2), 138–141.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rice, J. K. (2010). *The Impact of Teacher Experience: Examining the Evidence and Policy Implications. Brief No. 11*. National Center for Analysis of Longitudinal Data in Education Research. https://eric.ed.gov/?id=ED511988.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Willett, J. B. (1988). Chapter 9: Questions and Answers in the Measurement of Change: *Review of Research in Education*, 345–422. https://doi.org/10.3102/0091732X015001345.

Winters, M. A., & Cowen, J. M. (2013). Who Would Stay, Who Would Be Dismissed? An Empirical Consideration of Value-Added Teacher Retention Policies. *Educational Researcher*, *42*(6), 330–337. https://doi.org/10.3102/0013189X13496145.

Winters, Marcus A., & Cowen, J. M. (2013). Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies: Value Added Teacher Retention Policies. *Journal of Policy Analysis and Management*, *32*(3), 634–654. https://doi.org/10.1002/pam.21705.

Willett, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. Review of research in education, 15(1), 345-422.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52-61.