# 

**Factors Affecting Students’ Learning from a Design-based Implementation Research Project in Diverse Education Systems**

Qinyun Lin1, Kenneth A. Frank2, Shruti Bathia3, Karen Draney3, Jay Thomas4, and Charles W. Anderson5

1 Center for Spatial Data Science, University of Chicago

2 Department of Counseling, Educational Psychology, and Special Education, Michigan State University

3 BEAR Center, University of California, Berkeley

4 ACT, Inc

5 Department of Teacher Education, Michigan State University

**Author Note**

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/](https://urldefense.com/v3/__http:/carbontime.bscs.org/__;!!BpyFHLRN4TMTrA!vQx_ZmnEGIhC65VtvMYRdk60BVy5_8sNUws0yCGCE_5Ck0T6ejcc1HiEg_lURpr95s99$).

**Table of Contents**

[Abstract 3](#_Toc71207800)

[Introduction and Literature Review 4](#_Toc71207801)

[Standards-based Reform Efforts 4](#_Toc71207802)

[Design-based Implementation Research 6](#_Toc71207803)

[*Carbon TIME* as Design-based Implementation Research 10](#_Toc71207804)

[Research Questions 11](#_Toc71207805)

[Methods 13](#_Toc71207806)

[Participants 13](#_Toc71207807)

[Data Collection 14](#_Toc71207808)

[Validity and Reliability of Assessments 15](#_Toc71207809)

[Datasets 18](#_Toc71207810)

[Selection of Students and Teachers 18](#_Toc71207811)

[Selection of Posttest Data 19](#_Toc71207812)

[Data Analysis 20](#_Toc71207813)

[Research Question 1a: Effects of *Carbon TIME* Units 20](#_Toc71207814)

[Other Research Questions: Hierarchical Linear Models 21](#_Toc71207815)

[Robustness Analysis 26](#_Toc71207816)

[Findings 26](#_Toc71207817)

[Findings About the Effects of the *Carbon TIME* Units 26](#_Toc71207818)

[Findings about Factors Influencing Variation in Student Learning Gains 35](#_Toc71207819)

[Findings about Changes in Student Learning over Time 40](#_Toc71207820)

[Conclusion 44](#_Toc71207821)

[References 49](#_Toc71207822)

**Abstract**

This article reports on quantitative findings from a design-based implementation research (DBIR) project focused on teaching carbon cycling at multiple scales, from atomic-molecular to global. We report results from a large data set from diverse schools, including 59,654 student assessments of three-dimensional learning aligned with Next Generation Science Standards (NGSS Lead States, 2013) collected in classrooms of 133 middle- and high-school teachers over a four-year period. We used a subset of these data to investigate factors related to students’ learning in this large and diverse set of classrooms*.* Our findings included: First, the project’s curriculum and professional development were associated with substantial improvements in overall student achievement; within classrooms students with lower pretest scores had higher learning gains. Second, differences among individual teachers accounted for more of the variance in student learning gains than other measures including school demographics and students’ prior knowledge*.* Third, school demographic factors (percent free and reduced lunch and percent marginalized students of color) had statistically significant but small effects on students’ learning. Finally, sustained investment and participation were important; student learning gains improved across years for the project as a whole, and for individual teachers as they gained experience and participated in professional development. These findings contribute to programs of systemic change to support three-dimensional science learning.

*Keywords*: design-based implementation research, curriculum evaluation, achievement, teacher effectiveness, NGSS

# Introduction and Literature Review

This article reports analyses of quantitative data from a design-based implementation research (DBIR) project that aims to support student learning at scale to achieve three-dimensional learning goals in the Next Generation Science Standards (NGSS Lead States, 2013). One key goal of DBIR research is to connect large-scale reform efforts with daily classroom activitiesof teachers and students (Fishman et al., 2013). We begin this article with an overview of some key aspects of large-scale standards-based reform efforts and their connections with daily activities of teachers and students in science classrooms.

## Standards-based Reform Efforts

Cohen and Mehta (2017) have written a synthesis of lessons learned from David Cohen’s long career studying school reform. Reviewing reforms in American schools from the nineteenth century to the present, Cohen and Mehta (2017) distinguish between *system-wide reforms* that affected American education generally and *niche reforms* that “made significant changes in instruction by defining and creating a niche—a protected and bounded educational and political territory—in which they could thrive and survive.” (Cohen & Mehta, 2017, p.676).

One of the reforms that Cohen & Mehta analyze is standards-based reform (SBR). They mention SBR efforts in several different subject areas, then offer this summary of the general pattern:

SBR succeeded in several ways. It acquired elite political support and became the dominant frame for thought and action. It persists, and though it shifts shape among states and over time and has encountered serious problems, it shows few signs of evaporating. All states have the key instruments, and each presents these as ‘‘aligned,’’ which implies operation on several dimensions of schooling. Nearly all states profess commitment to the aims of the reform: strengthen academic quality, improve weak schools, and reduce educational inequality. The reform also brought unprecedented attention to educational inequality.

Despite these successes, SBR thus far seems to have not changed teaching and learning system-wide. There are mixed reports but no evidence of system-wide success. (Cohen & Mehta, 2017, p, 668).

So while new approaches to teaching have taken hold in individual schools and smaller programs, traditional approaches continue to predominate system wide. Therefore standards-based reforms have not led to system-wide improvements in student learning.

So how do the current standards-based reforms in science education, based on the *Framework for K-12 Science Education* (NRC, 2012), and the Next Generation Science Standards (NGSS Lead States, 2013) fit into the general pattern described by Cohen & Mehta? The answer, it seems to us, is “all too well.” System-wide, most states have adopted NGSS-based science curriculum frameworks, are developing statewide assessments based on those frameworks, and express commitment to the NGSS goals of improving science teaching and learning and reducing educational inequality (NRC, 2015). However, we see little evidence that these commitments have yet led to widespread changes in classroom teaching or students’ three-dimensional learning. There is little evidence of large-scale changes in science instructional practices (Banilower et al., 2018, Chapter 5). While current large-scale assessments are not fully aligned with NGSS, they do not show evidence of changes in student learning. For example, the average science scores on the ACT has stayed between 20.6 and 21.0 for the past decade (ACT 2016, 2020), while the percentage of students showing science achievement consistent with college readiness has been between 36 and 38 percent since 2013 when NGSS was widely adopted by states. Similarly, the NAEP twelfth grade science test shows no gains in science achievement from 2009 to 2015 with the majority of students below the proficient benchmark (U. S. Department of Education, 2015).

## Design-based Implementation Research

This context provides a motivating question for DBIR: *How can this standards-based reform be different?* Previous reforms have had deep effects on classroom teaching and learning only in protected niches. This time, how can the small-scale successes documented in the reform literature provide a foundation for system-wide change?

Our answer to this question relies on the framework presented in Figure 1, below. Figure 1 traces what Tekkumru-Kisa, Stein, and Doyle (2020) describe as “channels of influence” through which science educators and science teachers can seek to change the system. We use this figure to clarify how DBIR differs from both large-scale research on educational outcomes and design-based research that focuses on developing more powerful approaches to science teaching and learning.

**Figure 1**

*Task Framework*



*Note.* The figure was adapted from Tekkumru-Kisa et al. (2020).

In their quote above, Cohen & Mehta were describing patterns found in large-scale studies of standards-based reform. These studies have generally focused on statistical relationships between the blue sides of Figure 1: How can curriculum standards, and curricula based on those standards, affect student learning (e.g., Goldman et al., 2019; Grigg et al., 2013; Jacob et al., 2017; Roblin et al., 2018)? In these large-scale studies, the red middle of Figure 1—individual communities, schools, teachers, students, and classrooms—are perceived as statistical noise, weakening the relationship between reform efforts and student learning. The larger and more diverse the educational system, the weaker the relationship becomes, until it fades away in the diversity and complexity of the United States educational system as a whole. Furthermore, larger scale statistical studies have not used credible measures of three-dimensional science learning. Thus these studies have not provided evidence that the ambitious changes in science teaching and learning envisioned by the current reform movement can be enacted on a system-wide scale.

In contrast with large-scale statistical research, design-based research and articles discussing lessons learned from that research (e.g., Dauer, et al, 2014; Furtak & Penuel, 2019; Lehrer & Schauble, 2012; Miller & Anderson, 2017; Penuel & Reiser, 2018; Rosebery et al., 2014) has worked mostly within the red parts of Figure 1, developing and enacting standards-based reforms in small numbers of classrooms, and developing detailed qualitative accounts of how those innovations affect classroom discourse and student learning. These studies have produced “existence proofs” with important insights into how teachers and students can work together to enact and scaffold three-dimensional sensemaking (e.g., Lehrer & Schauble, 2012; National Research Council, 2007, 2015; Rosebery et al., 2014; Thompson et al., 2016). However, as Cohen & Mehta point out, the insights and products of design-based research have worked well in specific niches within our educational systems without having significant effects on the diverse educational system as a whole.

DBIR builds on insights from both quantitative large-scale research and qualitative design-based research. Design-based implementation researchers recognize the special challenges of designing reforms that work at scale, in diverse classrooms, while using insights from design-based research about how the sensemaking efforts of teachers and students in classrooms are much more than statistical noise. DBIR is an approach to research that deliberately addresses issues of coordination and coherence across components and scales of organization in education systems (Dolle et al., 2013; Fishman et al., 2013; Jackson & Cobb, 2013; Penuel et al., 2011; Penuel & Fishman, 2012).

DBIR projects differ from traditional large-scale and design-based research in a number of ways. DBIR projects seek to make strategic investments that will improve both our understanding of diverse educational systems and the ability of those systems to scaffold students’ three-dimensional engagement with phenomena. We describe three essential research and design goals below.

**1. Designing curricula as flexible “tool kits” for three-dimensional sensemaking.** The NGSS define goals for science literacy in terms of performance expectations that involve *three-dimensional sensemaking about phenomena:* Ways that students can engage with phenomena through science and engineering practices, crosscutting concepts, and disciplinary core ideas. A substantial body of design-based research focuses on developing curricular resources that support these goals in classrooms (e.g., Furtak & Penuel, 2019; Kang et al., 2014; National Research Council, 2015, 2018; OpenSciEd, 2020; Penuel & Reiser, 2018). These resources are built around an NGSS-based core goal—assessing and scaffolding students’ three-dimensional engagement with phenomena—while enabling individual teachers to respond to their students in rigorous and responsive ways.

**2. Supporting schools as learning organizations.** The schools, teachers, and students in the middle of Figure 1 are not just sources of statistical noise; they represent people who are actively responding to and shaping their local systems. Penuel & Fishman (2012, p. 297) “see [in DBIR] a common commitment to building theory and knowledge within the research community. The object of that theory is learning, but across scales of a system, where ‘learning’ applies not just to students in classrooms, but to individual adult actors (e.g., teachers, principals), organizational units (e.g., schools, curriculum departments in districts), and systems.” Penuel (2019) refers to this process of both building knowledge and redesigning systems at multiple scales as *infrastructuring.*

**3. Two-way learning in research-practice partnerships.**  Finally, we note that while the arrows in Figure 1 all go one way, from curricula to classrooms to student learning outcomes, DBIR research is built around the conviction that what researchers learn from teachers is at least as important as what teachers learn from researchers. In research-practice partnerships there is a two-way street between researchers and practitioners, such that researchers, teachers, and administrators play essential but complementary roles (Anderson, et al., 2018; Tseng et al., 2017). This is the hard work that requires new kinds of understanding— both for teachers and for researchers, a kind of “practice-based evidence” that can develop only through iterative cycles of design research conducted in partnership between researchers and practitioners (Bryk et al., 2015).

## *Carbon TIME* as Design-based Implementation Research

In this article we report analyses of quantitative data from the *Carbon TIME* project (<https://carbontime.bscs.org/>), a DBIR project enacted in 99 schools in multiple states. This project is the culmination of a research program that began in 2004 with a project to develop learning progressions leading to environmental science literacy. Our early learning progression work was used to develop the NRC *Framework,* and is cited in that document (Anderson, 2010; Mohan et al., 2009; Smith et al., 2006).

This learning progression research provided the foundation for subsequent work in instructional design, assessment development, and professional development. Beginning in 2007, we have engaged in ten cycles of curriculum design, implementation, assessment, and revision, always using learning progression-based assessments and close work with participating teachers to inform the revision process (2007, 2008, 2009, 2010, 2011, 2012, 2015, 2016, 2017, 2018). The first six cycles were funded by previous grants for learning progression and design-based research; the last four cycles were part of the current DBIR project. The primary data driving the revision process came from our case studies, in which almost every staff member on our project acted as a case study coach, visiting and videotaping classes of volunteer teachers, interviewing the teachers and their students, and discussing their observations with one another and other project staff. Covitt, et al. (in preparation) reports on a detailed analysis of these case study data; an abstract of this article is included in the supplementary materials.

One important insight from our learning progression and design-based research was that “scaling up” involved much more than doing design-based research on a larger scale. This led to our current DBIR project, which includes a focus on emergent properties of larger educational systems. A previous article in this journal identifies two key design challenges: “scale—the necessity of change at multiple scales in educational systems; and diversity—achieving rigor in our expectations with responsiveness to the enduring diversity of our students, classrooms, and schools” (Anderson et al., 2018, p. 1026). This project has each of the features of DBIR discussed above.

1. *Designing curricula as flexible tool kits.* Through an iterative development process, we developed six instructional units and classroom assessments focused on developing curriculum, assessments, and professional development resources that support middle- and high-school students’[[1]](#footnote-2) three-dimensional sensemaking about carbon cycling at multiple scales, from atomic-molecular to global (see Covitt & Anderson, 2018; Dauer, Doherty, Freed, & Anderson, 2014; Dauer, Miller, & Anderson, 2014; Miller & Anderson, 2017).
2. *Supporting professional and organizational learning:* As described in Anderson et al. (2018), the project included a professional development course of study; work with school partners to build organizational resources in participating schools and classrooms; and quantitative and qualitative data collection on schools and teachers. We discuss how we analyzed these data below.
3. *Improvement through iterative design cycles research-practice partnerships.* The current versions of the curriculum units, assessments, and professional development are the product of many years of iterative development involving both teachers and researchers (Anderson et al., 2018; Doherty et al., 2015; Thomas et al., 2020). In this article we report results from across four years of the study (2015–2019), and we analyze the data for evidence of improvement over time.

The quantitative analyses of student learning in participating classrooms reported in this article focus on these design challenges. Since these analyses were completed after the last revision cycle, they did not contribute directly to the iterative revision process described above. As described below; we do report analyses examining the effects of that revision process on student learning.

# Research Questions

Like other DBIR projects, this project used both qualitative and quantitative methods. This article is one of three articles being prepared for submission at the conclusion of the project. The other two articles, described briefly in the Discussion section below and in the online supplemental materials, focus on analyses of qualitative data from classroom videos and interviews with teachers and students. In this article we report on quantitative analyses of student learning data, including differences among students, teachers, and schools.

In the Methods section below we describe how we developed a system that enabled us to collect a large database on student learning. This database is larger, and represents a more diverse sample of schools, teachers, and students, than any of the design-based research studies cited above. It differs from other large-scale studies in that it uses credible measures of students’ three-dimensional learning, and in that the intervention was based on the three DBIR design goals described above. As described below we use hierarchical linear models (HLM) to address questions related to each of these three design goals.

**Assessing the success of the curricular “tool kit.”** As described above, one core goal of the *Carbon TIME* project was to design a flexible set of curricular resources that teachers and students in diverse schools could use to support three-dimensional learning in their own circumstances. This leads to some very basic questions: Did it work? If so, how well, and for which students? Research Question 1 addresses different aspects of those basic questions.

1. *Effects of curricular resources:* How did teachers’ participation in the *Carbon TIME* project affect student learning?
   1. How does student learning from *Carbon TIME* compare with student learning from curricula previously used by participating teachers?
   2. How is student learning related to pretest scores measuring students’ prior knowledge?
   3. What are the cumulative effects of multiple units of instruction?

**Understanding the effects of diversity in schools, teachers, and students.** We can compare learning data from different schools and classrooms to identify key factors that affected students’ learning and estimate their relative importance. These are key questions that are relevant to the second DBIR design goal above: If we need to support schools as learning organizations, what issues should we address in our infrastructuring efforts? Research Questions 2 and 3 address these factors.

1. *Teachers and school-level variables:* How do differences among teachers and schools affect students’ learning?
   1. How is student learning related to differences among individual teachers and their classrooms?
   2. How can the achievement data be used to estimate differences in student learning across classrooms?
   3. How is student learning related to demographic differences among schools?
2. *Relative importance of different factors:* Which factors had the largest effects on students’ learning?

**Assessing two-way learning through iterative design cycles.** The third DBIR design goal focuses on how both researchers and school partners need to learn through iterative design cycles. The project database includes students learning data from four successive years, and most teachers participated in data collection for two years or more. Our final research question addresses evidence about whether curricular resources or individual teachers improved over time.

1. *Sustained investment:* How did student and teacher success change over time?

# Methods

## Participants

In this article we report analyses of student assessment data from the classrooms of 133 experienced science teachers from public school districts primarily located in three states (Michigan, Colorado, and Washington). The datasets used to address specific research questions are described below. Importantly, these middle and high schools have different policies, curricular contexts, and student body demographics (described in more detail below) enabling us to explore implementation in diverse settings. All of the teachers voluntarily participated in the *Carbon TIME* project. Thus, the teachers were more experienced and more committed to research-based science teaching than a representative sample of US teachers. This is a diverse but not random or representative sample of all science students and teachers in the United States.

## Data Collection

A key part of this project involved systematically collecting pretest, posttest, and unit test data using assessments which aligned with three-dimensional learning goals. These assessments are described in more detail below. Figure 2 presents the timeline for data collection. We analyzed data from three cohorts of teachers. The rows in Figure 2 represent each cohort’s timeline. For example, Cohort 2’s Teacher Year 1 occurred in 2016–2017, its Teacher Year 2 in 2017–2018, and its Teacher Year 3 in 2018–2019. We asked teachers to commit to two years of data collection when they joined the project, though 39 teachers did not complete the second year due to changes in teaching assignment and other factors. Some teachers volunteered to continue participation in data collection for Teacher Years 3 and 4.

The bottom half of Figure 2 depicts the assessments administered during each Teacher Year, using Teacher Year 1 for Cohort 2 as an example.

* Teachers in Cohorts 2 and 3 administered baseline posttests (identical to Full *Carbon TIME* pretests and posttests) to their students the spring *before* their Teacher Year 1.
* Across all cohorts, students completed a full *Carbon TIME* pretest at the beginning of each academic year, which included a matrix sampling of items assessing their prior knowledge on three-dimensional learning associated with all six *Carbon TIME* units.
* During the school year students completed unit pretests and posttests. Most teachers taught the first three units in the suggested order: 1) *Systems and Scale*, 2) *Animals*, and 3) *Plants*. A smaller number of teachers taught the other three units: *Decomposers, Ecosystems,* and *Human Energy Systems.*
* After all the units were finished students completed a full *Carbon TIME* posttest using matrix sampling of the same item set as the pretest.

**Figure 2**

*Data Collection Timeline*Diagram

Description automatically generated

*Notes.* The pale blue arrows indicate that all the assessments were optional. The bottom half of the figure (within the red box) demonstrates the assessments administered in each *Carbon TIME* year, using *Carbon TIME* Year 1 for Cohort 2 as an example. The Pre-*Carbon TIME* baseline tests only apply to Cohort 2 and 3.

Teachers used this series of assessments for classroom formative and summative purposes, and these assessments served as this study’s primary data source.

## Validity and Reliability of Assessments

The assessments were developed and validated using an iterative process over a 10-year period, administered online, and scored using an automated scoring system developed in partnership with ACT (Thomas, et al., 2020). The assessments use what Penuel et al. (2019) describe as proximal transfer tasks to assess phenomenon‐based science learning; they required students to investigate or explain phenomena that were similar to, but not the same as, phenomena that they studied in *Carbon TIME* units. The three-dimensional learning performances that we assessed included students’ explanations, analyses of data from investigations, and arguments from evidence. For this article, we chose to focus on analyses of the inquiry and explanation items from the first three units: *Systems and Scale, Animals,* and *Plants*. Two example items and their relationships with NGSS performance expectations are included in the online supplemental materials; the full item set for all assessments is available online (<https://carbontime.bscs.org/assessment-links>).

Most test items combined one or more forced-choice responses with constructed written explanations. Through a partnership with ACT we developed a machine-scoring system that achieved high standards of reliability. For example, the quadratic weighted kappa of the scoring of all items exceeded 0.7; all items were checked for item fit and the few items with poor fit were revised or replaced (Thomas, et al., 2020). While we agree with the conclusion in the National Research Council report on developing assessments for the NGSS (NRC, 2014) that on-demand assessments cannot comprehensively measure students’ mastery of NGSS performance expectations, we have good evidence that these assessments elicit and evaluate students’ three-dimensional performances. In separate papers and publications, we have presented evidence for the validity and reliability of these tests as measures of three-dimensional learning (Authors, 2015; Authors, 2018).

This system scored student performances according to learning progression levels, described in detail in other publications (Covitt & Anderson, 2018; Jin & Anderson, 2012; Mohan et al., 2009) and summarized briefly below.

* *Level 4: Coherent scientific accounts (equivalent to NGSS high school performance expectations):* Students successfully trace matter and energy through carbon-transforming processes at multiple scales in space and time.
* *Level 3: Incomplete or confused scientific accounts:* Students show awareness of important scientific principles and of models at smaller and larger scales, but they have difficulty connecting models at different scales and applying principles consistently.
* *Levels 1[[2]](#footnote-3) and 2: Force-dynamic accounts:*Rather than tracing matter and energy students’ explanations and arguments focus on actors (e.g., animals, plants, people) and enablers (e.g., food, water, sunlight) to which the students attribute agency.

**IRT Analysis.** Using item response theory (IRT), we generated estimates of student ability that can be used to evaluate students’ three-dimensional learning for all the items included in each test. Although different tests had different items, overlapping items made it possible to calibrate item and test difficulties across years and across test forms on the same scale (Kelderman, 1988). Therefore, an estimate was generated as a calculated proficiency for each student, on each test, and these estimates across different years and different tests were calibrated on the same scale.

We report the test reliability coefficient and percentage of person misfit for each test in supplemental materials (see Table S1). The reliabilities range from acceptable (0.7 to 0.8) to good (0.8 to 0.9). The median percentage of persons with misfit values across all tests and years is approximately 13%, and most distributions of person fit statistics show a noticeable right skew. With large samples (e.g., greater than 50,000) parametric techniques are less affected by skew (Lumley et al., 2002).

**Logit Scale Scores.** The resulting scale units (logits) were a measure of how likely a student of a given proficiency was to achieve a particular scoring level on a given item, where 0 represented the overall student mean across all tests. To give a better sense of the logit scale, we also built a link between logits and learning progression level frameworks. We calculated threshold values for which we can claim that a certain logit indicated the student was most likely to be at a certain level in the learning progression framework described above:

* If a student’s logit score was above 0.96, the student’s overall proficiency was most likely at level 4 (equivalent to NGSS high school performance expectations).
* If the logit score was between –0.34 and 0.96, the student’s overall proficiency was most likely at level 3 (incomplete or confused scientific accounts).
* If a student’s logit score was below –0.34, the student’s overall proficiency was most likely at level 2 (force-dynamic explanations and arguments).

## Datasets

This paper presents findings based on students’ test performances on the logit scale from 2015–2016 through 2018–2019. In addition to student learning data, we also collected publicly available information for schools of our participating teachers, including percent of free and reduced lunch and percent of marginalized students of color. Using the online assessment system, we collected a large database: 197,000 student assessments from over 25,000 students across 178 teachers’ classrooms over a 4-year period. To ensure the validity of our analysis, we excluded incomplete and unreliable data, as described below.

### Selection of Students and Teachers

We restricted our samples for analysis to students and teachers for whom we had relatively complete and accurate data. Specifically, we included (1) students who completed both the overall pretests and unit posttests, enabling pre-post analyses for individual students; and (2) teachers whose classroom data included valid overall pretest and unit posttest data for at least 15 students on at least 2 of the 3 focus units (*Systems and Scale, Animals,* and *Plants*). We also excluded students who had extremely low scores (for either pretest or posttest), as the lowest scores were associated with short finish times, unanswered questions, or a large proportion of off-topic answers, indicating a lack of effort. See supplementary materials for more details about how this data exclusion procedure affects sample size, and distributions of overall pretest, unit posttest, and gain scores.

Table 1 shows the composition of the analytical samples of student learning outcome data. After omitting such cases, the final dataset included assessments from students across 245 Teacher Years instructed by 133 teachers who participated in the project from 2015–2016 to 2018–2019 academic years. The percentage of students in these schools receiving free and reduced lunches ranged from 5%–99%, and the percentage of marginalized students of color ranged from 0%–100%; thus, our analytical sample includes data from a diverse sample of teachers and students.

**Table 1**

*Analytical Samples of Student Learning Dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Analytical Sample 1 (Research Question 1a) | | | | | |
| Year | 2015–2016 | 2016–2017 | 2017–2018 | 2018–2019 | Total |
| Student baseline posttests | 1,988 | 1,264 | NA | NA | 3,252 |
| Student full pretests |  | 1,883 | 1,312 | NA | 3,195 |
| Student full posttests |  | 1,883 | 1,312 | NA | 3,195 |
| Teachers providing baseline data | 34 | 22 | NA | NA | 56 |
| Analytical Sample 2 (All Research Questions) | | | | | |
| Year | 2015–2016 | 2016–2017 | 2017–2018 | 2018–2019 | Total |
| Student full pretests | 1,967 | 4,976 | 5,932 | 3,320 | 16,195 |
| Student unit posttests | 5,310 | 13,034 | 16,206 | 8,909 | 43,459 |
| Teachers participating in project | 26 | 75 | 91 | 53 | 133 |

*Notes.* The baseline posttests were administrated in the spring semester. For example, 34 teachers provided baseline data for the school year of 2015–2016 means that these teachers administrated baseline tests in the spring of 2016.

### Selection of Posttest Data

We used different posttests to answer different research questions. When comparing with baseline tests from students who studied other curricula, we used the overall posttests to ensure a fair comparison between *Carbon TIME* and other curricular, even given any limitations of the test. That is, we restricted our sample to compare the test results administered to different groups of students based on the same tests, taught by the same teachers in the same schools; one group of students studied *Carbon TIME* while the other group of students studied other curricula.

However, when comparing the pretest and posttest proficiencies for the same students in our main analyses, we used average unit posttests, rather than overall full posttests, as the post-instruction measures. This is because the overall posttests turned out to have several limitations. Some teachers administered the overall posttests late in the school year and under time constraints, so these tests 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 behaviors (Wise, 2017). The overall posttests also included items based on units were not taught in many classrooms. Thus, we thought unit posttests served as better measures for students’ learning outcomes as they reflected students’ best efforts on units that they had studied.

## Data Analysis

Here we will first describe how we used both analytical samples (Table 1) to answer Research Question 1a. Then we will describe analyses using analytical sample 2 (Table 1) for all other research questions.

### Research Question 1a: Effects of Carbon TIME Units

We applied two approaches to analyze the effects of the curricular resources (Research Question 1a). First, we compared posttests from two groups of students taught by the same teacher in different years (analytical sample 1 in Table 1). As described above (see Figure 2), teachers in Cohorts 2 and 3, who started teaching *Carbon TIME* units in the fall of 2016 or 2017, administered the overall posttest (which we labeled the baseline test in Figure 2) to their students who had studied other curricula while taking the same courses during the spring before they started using *Carbon TIME* units. We used an independent-sample t test to compare baseline posttest proficiencies of students in the previous spring with the overall posttest proficiencies of students in TeacherYear 1*.* Teacher Year 1 in our analyses included 3,195 students from 56 teachers, and the traditional curricula group included another 3,252 students from the same 56 teachers.

Second, we compared pretests and posttests from the same students taught by the same teachers using *Carbon TIME* units. That is, we measured students’ proficiency before and after their classes studied *Carbon TIME* units. Thus the “matching” process for this analytical sample 2 (Table 1) occurred at the student level, which allowed us to compare pretests and posttests for 16,195 students taught by 133 teachers across the 4-year period (245 teacher-years). Accordingly, a paired t-test was conducted for analytical sample 2 in the second approach.

### Other Research Questions: Hierarchical Linear Models

We used hierarchical linear models (listed in Table 2) with analytical sample 2 in Table 1 to answer the other research questions. Because students *within* the same class are not independent from each other, applying hierarchical linear models accommodates the nested nature of the data of students within classrooms, and allowed us to model simultaneously effects at multiple levels (see Raudenbush & Bryk, 2002). Analyses done only at a single level (e.g., the student level) do not recognize dependencies (e.g., among students within the same classroom) and, therefore, compromise standard errors and corresponding significance tests (Raudenbush & Bryk, 2002). Thus, we created a two-level model. Our outcome variable was students’ learning gains (the difference between students’ pretest and posttest proficiencies), measured as logit scores. The level-1 model specifies a linear regression at the student level, and level-2 is at teacher-year level. Our data could not adequately differentiate classrooms within teachers, so we use teacher-year (see Figure 2 above) for level-2.[[3]](#footnote-4) For the remainder of the manuscript, we will refer to teacher-year as classroom. We did not include a third level, modeling the effect of *Carbon TIME* at the school level, because most schools contained only a small number of teachers (71% of schools had only 1 teacher). The models we estimated are presented in Table 2.

**Table 2**

*Hierarchical Linear Models*

|  |  |
| --- | --- |
| Unconditional Model |  |
|  |
| Model 1 |  |
|  |
| Model 2 |  |
|  |
| Model 3: program development |  |
|  |
| Model 4: teacher experience |  |
|  |
| is the learning gain for student *i* taught by teacher *j*;  is the learning proficiency for student *i* measured in the Pretest;  is the teacher *j*’s classroom average learning in the Pretest;  is the percent of free and reduced lunch in teacher *j*’s school;  is whether teacher *j* is in a high school or middle school;  is the percent of marginalized students of color in teacher *j*’s school;  includes three dummy variables indicating whether this classroom is from 2016–2017, 2017­–2018, and 2018­–2019;  includes three dummy variables indicating whether this year is the teacher’s 2nd year, 3rd year, or 4th year of teaching *Carbon TIME*. | |

**Within-classroom and between-classroom variation.** In the analyses, we started from the unconditional model in Table 2 to decompose the variation of students’ learning gains into within-classroom and between-classroom components. The result showed that the intra-class correlation (ICC)[[4]](#footnote-5) was larger than 30%, indicating that 30% of the variation in student learning gain was *between* classrooms. This was a high ICC compared with other education research (Frank, 1998).[[5]](#footnote-6) Ultimately, we fit a full model (Model 1 in Table 2) with all the related variables of interest to study what factors help explain how much different students learned from *Carbon TIME*. These predictors were from both the student level (level 1) and classroom level (level 2).

**Students’ pretests as predictors.** At the student level, we included the predictor that measured how far a student’s pretest deviated from her classroom’s average pretest score. For example, if Adrienne had a pretest of 0.68 and her classroom’s average pretest was 0.5, then, the value of her predictor would be 0.18 (0.68 - 0.5 = 0.18). If another student, Elizabeth, had the same pretest score of 0.68 but her classroom’s average pretest was 0.7, then the value of her predictor would be –0.02. The –0.02 indicates that Elizabeth is 0.02 logits lower than her classroom average. At the classroom level (level 2), we included the average pretest (). In the previous examples of students Adrienne and Elizabeth, their classroom’s average pretest values were 0.5 and 0.7, respectively. Therefore, if the coefficient associated with the individual score () is positive, that means a student with higher pretest relative to their classroom mean (0.18 vs –0.02) has higher learning gains. If the coefficient associated with the classroom average score () is positive, this means that students from classrooms with higher average pretests (0.5 vs 0.7) have higher learning gains on average.

It is important to note that these two pretest-related predictors both measured students’ prior knowledge, yet they are orthogonal to each other. The first predictor, at the student level, measured *within-class* variation, while the second predictor, at the classroom level, measured *between-class* variation. In other words, these predictors represent two independent ways that students’ pretests may be related to their learning outcomes. One can think about this as a two-step process for estimating a student’s learning gain. First, we estimated students’ classroom average learning gains. Second, we estimated how different any given student’s learning gain is from her classroom average. Later, when we quantified the impact of students’ prior knowledge on students’ learning gains, the effects of these two parts were added up to estimate the overall impact of students’ prior knowledge on their learning outcomes.

**School-level predictors.** Three school characteristics were analyzed as potentially important predictors at the classroom level (level 2). First was the percentage of students eligible for free and reduced lunch (FRL). Second was the percentage of marginalized students of color (non-White/Asian, marginalized students). We used these two school-level predictors as approximate measures for school factors that might impact learning outcomes. Acknowledging the limitations of these measures (Greenberg et al., 2019), they were nevertheless the best measures we had to estimate school demographic factors. Previous studies have shown that the percent of free and reduced lunch can be a proxy measure for material, social, and human material resources such as overall school funding (EdBuild, 2019), 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). Therefore, if the coefficient of percent of free and reduced lunch () and the coefficient of percent of marginalized students of color () in Model 1 are negative, that means students from schools with more resources have higher learning gains.

A third predictor at level 2 was whether the school was a high school or middle school (GradeBand). The results of Model 1 showed that grade band was not significantly related to students’ learning gains when students’ prior knowledge was included in the model. We therefore used a more parsimonious model (Model 2) to analyze how the other predictors affected students’ learning. The Q-Q Plot of the level-2 residuals showed that the residuals fall along a roughly straight line at a 45-degree angle, indicating that the residuals are roughly normally distributed, supporting parametric assumptions for making inferences

**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’ of data and applied Models 3 and 4, where we added predictors that indicated the academic year (Model 3), or how many years the teacher had taught *Carbon TIME* (Model 4).

Indicators for academic years in Model 3 helped us analyze 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 were implemented based on feedback from teachers and students (Anderson et al., 2018). Model 3 allowed us to see if there was 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 2016–2017, 2017–2018, and 2018–2019, with the first year 2015–2016.

Similarly, in Model 4, we added three binary variables to capture how many years the teacher had taught *Carbon TIME*. The three added binary variables indicated whether a teacher was in her second, third, or fourth year of teaching *Carbon TIME*. That is, the reference group was the first year of teaching *Carbon TIME.* The coefficients of these indicators allow us to tell how years of experience teaching *Carbon TIME* were associated with student learning gains.

### Sensitivity and Robustness Analysis

To evaluate how sensitive our conclusions are to our data selection procedure, we report how estimates and inference were affected in each step of the procedure (see supplementary materials). We also recognize that our estimates may be biased because teachers (and indirectly their students) self-selected into the *Carbon Time* curriculum. To minimize potential bias, we included pretests and other important covariates in our models, which extensive research has shown reduced omitted variable bias by 60–90% when comparing with the “Gold Standard” of randomized experiments (Shadish et al., 2008; Wong et al., 2017). Additionally, we conducted robustness analysis to evaluate how robust our inference is with respect to potential bias, as the ultimate goal of DBIR research is to inform practical decision making based on inferences drawn from research in the field. Specifically, we discuss how different the sample needs to be or how strong a confounder needs to be to change our inferences (Frank, 2000; Frank et al., 2013). The results of our robustness checks are included in the Findings section, below.

# Findings

We report our findings from the analyses for each research question.

## Findings About the Effects of the *Carbon TIME* Units

**Finding 1.a. Students studying *Carbon TIME* units showed significantly higher achievement than students studying curricula previously used by participating teachers.** The project design compared student end-of-year proficiencies (full *Carbon TIME* posttests in Figure 2B) from *Carbon TIME* with student end-of-year proficiencies from curricula previously used by participating teachers (baseline tests in Figure 2B), using analytical sample 1 (Table 1). The first part of Table 3 and Figure 3A summarize the comparison results. In Figure 3A purple shows the full posttest distribution for students who studied other curricula, and yellow shows the distribution for students who studied *Carbon TIME*. Students who studied *Carbon TIME* performed much better on the posttests than students who studied other curricula. The average difference was 1.25 logits (effect size = 0.81, *p* < 0.001).[[6]](#footnote-7) The vertical black lines indicate the cut scores for transitions to learning progression levels 3 and 4. (As explained above, level 4 corresponds to high-school level NGSS performance expectations.) More *Carbon TIME* students achieved level 3 and level 4:

* In the non-*Carbon TIME* group, 14% of the students most likely achieved level 3 and 3% of students most likely achieved level 4.
* In the *Carbon TIME* group, 32% of the students most likely achieved level 3 and 24% of students mostly likely achieved level 4.

**Table 3**

*How Successful is Carbon TIME*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | N | Mean | Std. Dev. | Min | Max |
| Analytical Sample 1: compare baseline with posttest | non-*CTIME* | 3,252 | –1.251 | 1.049 | –4.983 | 3.387 |
| *CTIME* | 3,195 | –0.003 | 1.546 | –4.907 | 8.524 |
| Two-sample t test: difference = 1.247, SE = 0.033, *p* < 0.001. Effect size = 0.807. | | | | | |
| Analytical Sample 2: compare pretest with posttest | 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. | | | | | |
| Analytical Sample 1: compare baseline with pretest | *CTIME* pretest | 3,195 | –1.454 | 0.795 | –4.948 | 1.549 |
| non-*CTIME* posttest | 3,252 | –1.251 | 1.049 | –4.983 | 3.387 |
| Two-sample t test: difference = –0.203, SE = 0.023, *p* < 0.001. Effect size = 0.194. | | | | | |

**Figure 3**

**A**

Chart, histogram

Description automatically generated

Comparing posttests between *Carbon TIME* and non-*Carbon TIME* students

B

Chart, histogram

Description automatically generated

Comparing *Carbon TIME* pretests and posttests

*Note.* Both figures show consistent evidence of the *Carbon Time* effect when compared with baseline defined by pre-*Carbon Time* experience or pre-test.

The second part of Table 3 summarizes the findings when we compared full pretests and average unit posttests from the same students taught by the same teachers with *Carbon TIME* (the second approach, using analytical sample 2 in Table 1). Figure 3B displays the results, where pink shows the distribution of the pretests and light blue shows the distribution of posttests. The average student proficiency increased from –1.44 logits on the pretest to 0.40 logits on the posttest on average, indicating an overall average gain score of around 1.83 logits (effect size = 1.42, *p* < 0.001). Most students achieved level 3 or level 4 on the unit posttests.

* On the pretest 6% of the students most likely achieved level 3 and 0.3% of students most likely achieved level 4.
* On the unit posttests 39% of the students most likely achieved level 3 and 30% of students mostly likely achieved level 4.

Additionally, we compared the average full *pretest* of students who learned *Carbon TIME* with the average *baseline posttest* of students who learned other curricula (using analytical sample 1). The difference was 0.20 logits (effect size = 0.19, see the last part of Table 3): statistically significant but small in terms of educational significance. This finding indicates that the more traditional curricula previously used by the teachers were minimally effective in helping students master the three-dimensional performances included in the *Carbon TIME* assessments*.*

**Finding 1.b: Within classrooms, students with lower pretest scores had higher learning gains.** As discussed in the methods section, we decomposed the factor of students’ prior knowledge (measured by students’ pretests) into two parts: within-class and between-class. The within-class component measured the difference between an individual student’s pretest and the student’s classroom’s average pretest. The between-class component measured 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 the class average pretest in Table 4. For example, if a student has a pretest that is 0.1 logits below their classroom average, then their value for this variable is –0.1. When multiplied by the coefficient of –0.507, we get a positive value of 0.0507, indicating this student’s expected gain score is 0.0507 logits higher than an average student’s gain score in the classroom.

**Table 4**

*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 | –0.0655 | –0.130 |
| (0.138) | (0.136) | (0.134) |
| Percent of free and reduced lunch. | –1.077\*\*\* | –1.009\*\*\* | –1.029\*\*\* |
| (0.293) | (0.286) | (0.284) |
| Percent of marginalized students of color. | –0.544 | –0.613\* | –0.469 |
| (0.304) | (0.296) | (0.295) |
| Whether this is the year of 2016–2017. |  | 0.204 |  |
|  | (0.147) |  |
| Whether this is the year of 2017–2018. |  | 0.509\*\*\* |  |
|  | (0.144) |  |
| Whether this is the year of 2018–2019. |  | 0.424\*\* |  |
|  | (0.154) |  |
| Whether this is the 2nd year of teaching the project units. |  |  | 0.257\*\* |
|  |  | (0.0885) |
| Whether this is the 3rd year of teaching the project units. |  |  | 0.531\*\*\* |
|  |  | (0.144) |
| Whether this is the 4th year of teaching the project units. |  |  | 0.430 |
|  |  | (0.250) |
| Constant | 2.028\*\*\* | 1.784\*\*\* | 1.840\*\*\* |
|  | (0.195) | (0.213) | (0.195) |
| Observations | 16,195 | 16,195 | 16,195 |
| Number of Teacher\_Year | 245 | 245 | 245 |
| Level 2 residual variance | 0.419 | 0.394 | 0.392 |
| Level 1 residual variance | 0.909 | 0.909 | 0.909 |

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

We conducted robustness analyses to evaluate our inference that participation in *Carbon TIME* reduced differences among student test scores *within* classrooms. These analyses indicated that to invalidate the inference that students with lower pretest scores showed higher learning gains, more than 96% of the estimate would have to be due to bias, based on the conventional statistical significance at a 0.05 level (Frank et al., 2013). Reinterpreted, to invalidate this inference, one would need to replace 96% of the students in our analytical sample with students whose pretests had no effect on their learning gains. Thus, this is a highly robust inference compared with the studies reviewed in Frank et al (2013).

Because we found a negative relationship between pre-test and gains, another threat to the validity of the inference comes from potential measurement error in test scores. We checked for this possibility. The standardized coefficient was –0.308. After adjusting for measurement error in pretest and learning gains (Willett, 1988), the partial correlation between pretest and learning gains became –0.290, which was still a strong and both statistically and practically significant correlation. There could also be concerns about omitted variables. We calculated that an omitted confounder would need to correlate with both the pretest and learning gains at 0.598 (conditioning on other observed covariates) to invalidate the inference. These are very high correlations, typically on the order observed for pre-tests, indicating that our data showed strong, robust evidence that *Carbon TIME* reduced differences among student test scores *within* classrooms.

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. In other words, studying *Carbon TIME* reduced achievement differences within classrooms, but not between classrooms with different average pretest scores. (Also, as reported above in the Method section, learning gains were not significantly different in middle school vs. high school classrooms.)

Note that the negative association we described above is between student *learning gain* and students’ prior knowledge, where student learning gain is measured as the change from the full pretest to the average unit posttest. Students’ pretest scores were positively associated with posttest scores, indicating that the achievement gap was not reduced to zero.

**Finding 1.c: Student learning increased from the first unit to the third unit.** We also found evidence that as students moved from the first unit (*Systems and Scale*) to the third unit (*Plants*), their posttest performance improved, 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 4 shows the distribution of learning gains in these three units. The differences among learning gains for three units are statistically significant (see Table S2 in supplementary materials).

**Figure 4**

*Distribution of Learning Gains in Different Units*

Chart, histogram

Description automatically generated

## Findings about Factors Influencing Variation in Student Learning Gains

**Finding 2a, b: The differences in learning gains across teachers’ classrooms (controlling for student pretests and school demographic variables) were both statistically and educationally significant.** As discussed earlier, the intra-class correlation (ICC) in the unconditional model illustrated that more than 30% of the variance in students’ learning gains was between classrooms. In Model 2, we added predictors that were statistically significant (including school-level variables and students’ pretests) to explain the variance in students’ learning gains. The residual variance at the teacher–year (level 2) only decreased a little, from 0.473 (unconditional model) to 0.419 (Model 2), a change of about 11%. This showed that the differences in learning gains in different teachers’ classrooms were real and substantial, even after controlling for the effects of students’ prior knowledge and school demographic factors.

*Value-added measures of teacher effectiveness.* We used Model 2 (Table 4) as our value-added model to estimate teacher effectiveness, also known as value-added measures (VAMs). Using Model 2 (see Table 2, results reported in Table 4), the expected students’ gain score was calculated based on the students’ pretest, their school’s 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 was then regarded as a teacher’s value added or teaching effectiveness. In other words, if a teacher had a positive value here, it means that their students performed better than expected on the posttests. Figure S1 in supplementary materials presents the VAMs and 95% confidence interval for 133 teachers’ 245 classrooms. These are our best estimates of overall effectiveness for individual teachers. The VAMs show that variation in teachers’ effectiveness is substantial, and not attributable to student pretest or school demographic variables.

*Limitations of VAMs.* Rooted in education production literature (e.g., Ben-Porath, 1967; Hanushek, 1979), VAMs are intended to capture teachers’ individual contributions to students’ achievement. VAMs are essentially the “deflections” between students’ expected test scores and their actual ones (Murphy, 2012). Proponents of value-added models cite research that shows teachers’ considerable and long-lasting influences on student achievement (e.g., 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 VAMs (Hanushek & Rivkin, 2010).

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 (e.g., Guarino et al., 2014; Harris, 2009; Raudenbush, 2015). *We agree with the skeptics*. The data in our project showed important differences among teachers, but ranking individual teachers based on the value-added measures was not reliable: only a small amount of potential bias is needed to alter the rankings. Additionally, analyses focusing on individual teachers, such as those reported by Covitt, et al. (2020) and summarized in the supplementary materials (Figure S2), show that rankings across years for the same teachers often change.

Thus, we conclude that the VAM analyses are useful as a research tool for investigating patterns of difference in student learning that are associated with patterns of difference in classroom discourse, particularly when we compare classrooms with substantial differences in learning gains. Our qualitative analyses (Covitt, et al., 2020; Morrison Thomas, et al., 2020) examined how teachers with different VAMs differed in terms of classroom discourse and teacher commitments and perceptions. The classrooms with higher VAMs were characterized by teaching that was both rigorous and responsive to students’ ideas and interests. The teachers in these classrooms successfully assessed and scaffolded their students’ three-dimensional engagement with phenomena. These analyses are summarized in the online supplemental materials. However, even with careful control over curricula and assessments, the value-added measures are not sufficiently precise or reliable to inform high-stakes comparisons among individual teachers.

**Finding 2c: Students’ learning gains were negatively related to school-level demographic variables including percent free and reduced lunch and percent marginalized students of color.** Students in schools with higher percent of free and reduced lunch, or higher percent of marginalized students of color, showed smaller learning gains. These were reflected by the coefficient of –1.077 (*SE* = 0.293, *p* < 0.001) for free and reduced lunch, and the coefficient of –0.544 (*SE* = 0.304, *p* = 0.074) for marginalized students of color, in Table 4. Note that these two measures were highly correlated with each other (correlation is 0.498, *p* < 0.001), and the collinearity led the two coefficients to show weaker statistical significance when both predictors were included in one model. When we model separately the percent of free and reduced lunch had a coefficient of –1.361 with a *p* value of 0.001 (*SE* = 0.247). Without the percent of free and reduced lunch, the percent of marginalized students of color had a coefficient of –1.149 with a *p* value smaller than 0.001 (*SE* = 0.262). 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. Therefore, in our analyses of decomposition of variance (Finding 3 below) we combined these two variables into a single school demographic factor.

Previous studies have shown that percent of free and reduced lunch and percent of marginalized students of color can be a proxy measures for material, human, and social 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). For example, EdBuild (2019) reported that predominantly White school districts got $23 billion more in funding than predominantly non-White districts in 2016, even for comparable numbers of children served in the districts. Additionally, our survey data showed that teachers from schools with higher percentages of free and reduced lunch had lower scores on items measuring science knowledge and pedagogical content knowledge.

An alternative explanation for the negative association between students’ learning and the school-level variables (percent of free and reduced lunch, and percent of marginalized students of color) is that schools and teachers continue to implicitly but systematically discriminate against poor students and marginalized students of color. Like other reform curricula, *Carbon TIME* relies on teachers to scaffold students’ sense-making by eliciting and responding to students’ expressions of their own ideas, using their lifeworld social languages (Gee, 2005). An abundance of research shows how middle-class White teachers working with poor students and students of color struggle to respond is ways that fully recognize their students’ lifeworld experiences and funds of knowledge (Heath, 1983; Ladson-Billings, 2009; Parsons & Carlone, 2013; Schwarz et al., 2020). Our quantitative data do not allow us to distinguish between these two explanations.

**Finding 3: Teachers were more important than students’ prior knowledge and school demographic factors in explaining variation in students’ learning gains.** We showed that students’ learning gains were significantly associated with (a) students’ prior knowledge, (b) the teachers and classrooms to which they were assigned, and (c) school factors. Through a series of analysis (see online supplementary materials for more details), we decomposed the variance in students’ gain scores into these three important factors in order to estimate how important each factor was in explaining differences in student learning gains. Because variables are collinear with each other to some extent, we cannot clearly identify the effects of each factor separately. Instead, we report a range of variance in the gain score that can be accounted for by each factor.

Figure 5 presents the findings, showing a range of percent of variance for each factor. For example, on the very left, the blue bar shows that differences among teachers accounted for at least 27% variance of students’ learning, and the grey bar shows that differences among teachers accounted for at most 31% variance of students’ learning. Similarly, the two bars in the middle show that students’ prior knowledge accounted for 9.6% to 9.7% of students’ learning. On the right, school demographic factors accounted for less than 4% of the total variance. While school factors were statistically significant, they did not make a large contribution to student learning compared with teachers and students’ prior knowledge.

**Figure 5**

*Importance of Different Factors Affecting Students’ Learning Gains in Carbon TIME*

## Findings about Changes in Student Learning over Time

**Finding 4: Student learning increased (a) as *Carbon TIME* units and professional development improved over time, and (b) as teachers gained experience and learned from professional development.**

*Changes in student learning across school years.* Figure 6 shows how student achievement and student learning gains changed across the four school years of the *Carbon TIME* project. Figure 6A shows posttest scores: how the mean and 95% confidence interval of class average unit post changed from 2015–2016 to 2017–2018. The black horizontal dashed lines represent the thresholds for level 3 and level 4, respectively. The average unit posttest score in all teachers’ classrooms increased over the four-year period. There were also more teachers who achieved learning progression level 4 averages (corresponding to NGSS high school performance expectations) over time, 7.7% of teachers in 2015–2016 to 13% in 2016–2017 and to more than 22% in 2017–2018 and 2018–2019.

**Figure 6**

**A**

**Graphical user interface, chart, line chart

Description automatically generated**

Mean and 95% Confidence Interval of Average Unit Post for 133 Teachers and Their 245 Classrooms, across Academic Years from 2015–2016 to 2018–2019

**B**

**Chart, histogram

Description automatically generated**

Mean and 95% Confidence Interval of Average Learning Gain for Each Teacher from 2015–2016 to 2018–2019

Figure 6B shows the average learning gain (i.e., average difference between full pretest and unit posttests, represented by the dot) and 95% confidence interval (represented by the error bar) in 245 classrooms. Although all classrooms had positive average learning gains, the differences among classrooms were 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 shown, the average learning gain increased from 2015–2016, to 2017–2018, and 2018–2019 was slightly lower than 2017–2018, but still higher than 2016–2017. This indicated that, on average, the learning of students participating in the *Carbon TIME* program increased over time.

*Changes in student learning as teachers gained experience.* Figure 7 shows changes in student learning for individual teachers as they gained experience and participated in professional development. Specifically, Figure 7 shows the average learning gain (i.e., average difference between full pretest and unit posttests) and 95% confidence interval across each year of teachers’ experience of teaching *Carbon TIME.* Most teachers participated in data collection for two years; some volunteered to continue providing data for a third and fourth year. 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**

**Graphical user interface, chart, histogram

Description automatically generated**

Mean and 95% Confidence Interval of Average Learning Gain as They Gained More Experience in Teaching the Project Units

*Notes.* We see the same pattern for VAMs as the raw learning gains, but we chose to show the raw learning gains, as the estimation of standard errors for VAM is controversial (Bibler et al., 2014).

We applied two-level hierarchical linear models (Models 3 and 4 in Table 2) to quantify the effect on students’ learning gains (controlling for school demographic factors) from: (a) the cohort year to estimate how the *Carbon TIME* units and professional development changed over time; and (b) the teacher’s year to estimate how teachers changed with more experience with *Carbon TIME*. The results are presented in the last two columns in Table 4 (Model 3 and Model 4). From Model 3, we find evidence for the effect of improvement in *Carbon TIME* units and professional development on students’ learning: from the year of 2015–2016 to 2017–2018, there was a significant increase of 0.51 logits in students’ learning gains. From Model 4, we can tell students’ learning gains increased significantly (0.53 logits) as teachers gained more experience from their first year to third year of teaching *Carbon TIME*.

We cannot statistically distinguish which of these two factors—teachers gaining experience or improvement in *Carbon TIME* units and professional development—contributed more to the increases in student learning gains. These two predictors were collinear, as we would hope for in a DBIR project: the research-practice partnership enabled teachers to improve their teaching practices and developers to improve the units and professional development. Table S3 in online supplementary materials 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. Teachers gaining experience and improvement in *Carbon TIME* occurred at the same time. Additionally, fewer teachers participated in data collection after their first two years, making it difficult to draw statistical inferences about how effects of *Carbon TIME* changed over time.

# Conclusion

We began this article with a basic question arising from Cohen and Mehta’s review of educational reforms: *How can this standards-based reform be different* from all the previous standards-based reforms that have transformed teaching and learning only in protected niches in our educational system. Like other DBIR projects, we believe that the response to this challenge will require long-term, patient, iterative work that emphasizes research-practice partnerships and adaptations to local communities. In this article we report analyses of student learning data from a project that had these characteristics. In the paragraphs below we first note important limitations of this study, then summarize some key findings:

**Limitations of the study.** Traditional large-scale quantitative research has sought conclusions that are generalizable to large populations, such as the general population of secondary science students and teachers in the United States. Like other DBIR research, this study has not led to those kinds of conclusions. The sample of teachers and students who provided data for this study was large and diverse, but not representative of American secondary schools as a whole. The units that we focused on addressed a small part of the science curriculum (around nine weeks of teaching in most classrooms). The measures of student learning were on-demand tests that inevitably could not assess the full extent of students’ three-dimensional reasoning. Other interpretations are always possible for the analyses we report.

We also note that the findings of this article are limited to analyses of quantitative data from student tests in a larger DBIR project. Other data from this project are reported in prior publications and two other articles in development (Covitt et al., in preparation; Morrison Thomas et al., in preparation; see abstracts in online supplemental materials). Our analyses of factors associated with differences in student learning gains lead to qualitative questions about what was happening in those classrooms. Our analyses addressing some of those qualitative questions are addressed in the paragraphs on additional questions to be investigated, below.

While we acknowledge these limitations, we also believe that the analyses we report in this article include robust and important findings relevant to three design goals that we share with other DBIR projects. We address the implications of those findings below.

**Assessing the success of the curricular “tool kit.”** Our baseline data confirm a common finding of other studies: most science curricula being used in schools today do not support students’ three-dimensional learning. However, students showed substantial learning gains after studying *Carbon TIME*. Students’ learning from *Carbon TIME* units was both educationally and statistically significantly larger compared with the curricula that teachers had used previously: Ninety-one percent of students who studied *Carbon TIME* scored higher on the overall posttest than the median of students who studied the teachers’ previous curricula.

Additionally, *Carbon TIME* helped reduce the achievement gap within classrooms; students with lower pretest scores showed significantly higher learning gains. There is also evidence that students’ learning gains were cumulative over time: Students showed higher learning gains in units that they took later in the sequence.

**Understanding the effects of diversity in schools, teachers, and students.** Both our qualitative studies (in preparation) and the quantitative measures reported in this study show that “fidelity of implementation” is not a useful measure of how different schools and teachers used the *Carbon TIME* curricular tool kit. Our classroom observations and teacher interviews show how teacher used the units differently, in response to their own beliefs and commitments and to differences in their students and classroom circumstances. Our findings show that those differences were consequential: Students in different classrooms varied substantially in how much they learned from *Carbon TIME.* The differences among teachers were large, educationally significant, and not attributable to other factors such as students’ prior knowledge or the racial composition of schools.

School demographic factors (percent free and reduced lunch and percent marginalized students of color) also made a difference in students’ learning but accounted for much less of the variance in student learning than teachers. This suggests that *demography is not destiny;* working with schools to improve their curricular and assessment resources as well as their social infrastructures—policies, practices, and norms for professional communities—can support sustained improvement in teachers’ practices and in student learning.

**Assessing two-way learning through iterative design cycles.** The third DBIR design goal focuses on how both researchers and school partners need to learn through iterative design cycles. We report on the nature of those iterative design cycles and what we learned from them in other publications. Our findings indicate that student learning improved over time, though we cannot statistically distinguish between the effects of improvements in curriculum and PD (i.e., researcher learning) and learning from experience by individual teachers. Average learning gains for all students improved significantly across the first three years of the project. Individual teachers also improved their students’ learning across years as they gained experience through their second and third years of participation in the project. These data provide evidence that “learning…across scales of a system” (Penuel & Fishman, 2012) can occur, and that the learning takes place through sustained, collective engagement in specific contexts.

We note that administrators often value “flexibility” in teaching assignments, with teachers moving among different subjects or courses in response to student demand or administrative priorities. Similarly, teachers value freedom to choose materials and activities that they like. Curriculum developers commonly work on tight timelines, with limited opportunities for observation, feedback, and revision. These data indicate that flexibility, freedom, and speedy development come with a price. Students, teachers, and organizations all benefit when they can work collectively with other professionals, when curricula and teaching assignments are stable, and when iterative revision cycles enable cumulative learning.

**Additional questions to be investigated.** These quantitative findings lead to qualitative questions: How did teachers use the units and assessments to assess and scaffold their students’ engagement with phenomena? Why were some classrooms much more successful than others? How did teachers’ personal goals and perceptions of their students affect their use of *Carbon TIME*? How were teachers influenced by their schools’ professional norms, obligations, and resources?

In this article we describe the development of value-added measures that provide an important tool for these qualitative investigations. Although these measures do not support precise distinctions between individual teachers, the measures are statistically robust and can enable studies that look for patterns of difference between classrooms with higher and lower student learning gains.

Other articles being prepared for publication report our findings from these investigations. Analyses of classroom video data show how the more successful teachers were able to use the project curriculum and assessments as tools for rigorous and responsive teaching (Covitt et al., 2020). Analyses of teacher interview data showed how the practices of the more successful teachers are based on their commitments and perceptions of themselves, their students, and the project. The more successful teachers also described how their commitments and teaching skills improved over time, through their stable teaching assignments and sustained involvement in the project (Morrison Thomas et al., 2020). Abstracts of these articles are included in the online supplemental materials.

Both additional analyses of data from this project and analyses from other DBIR projects will be necessary to achieve the design and knowledge-building goals of this research and to support effective and equitable three-dimensional learning in our diverse educational systems. We hope that these findings will contribute to this important effort.

# References

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. <https://doi.org/10.1086/508733>

ACT. (2016) The Condition of College and Career Readiness 2015. Retrieved from: <https://www.act.org/content/dam/act/unsecured/documents/Condition-of-College-and-Career-Readiness-Report-2015-United-States.pdf>.

ACT. (2020). The Condition of College and Career Readiness 2019. Retrieved from: https://www.act.org/content/dam/act/unsecured/documents/National-CCCR-2019.pdf.

Anderson, C.W. (March, 2010). Learning Progressions for Environmental Science Literacy. Paper prepared for the National Research Council Committee to Develop a Conceptual Framework to Guide K-12 Science Education Standards. Available: <http://www7.nationalacademies.org/bose/Anderson_Framework_Paper.pdf>

Anderson, C. W. (2020, March). Introduction to the paper set: Teaching for science literacy at scale. Prepared for presentation at the annual meeting of the National Association for Research in Science Teaching.

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>

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). Report of the 2018 NSSME+. Horizon Research, Inc. <http://horizon-research.com/NSSME/2018-nssme/research-products/reports/technical-report>

Ben-Porath, Yoram. (1967). The Production of Human Capital and the Life-Cycle of Earnings. Journal of Political Economy, 75(4), 352–365.

Bibler, A., Vosters, K. N., Guarino, C., Wooldridge, J. M., Reckase, M. (2014). Precision for Policy: Calculating Standard Errors in Value-Added Models. Association for Public Policy Analysis & Management. <https://appam.confex.com/appam/2014/webprogram/Paper11279.html>

Bryk, A. S.,Gomez, L., M.,Grunow, A., &LeMahieu, P., G. (2015). Learning to improve: How America's schools can get better at getting better. Cambridge, MA: Harvard University Press.

Chetty, R., Friedman J. N., & Rockoff, J. E. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood (No. w17699). National Bureau of Economic Research. Retrieved from: http://www.equality-of-opportunity.org/assets/documents/teachers\_wp.pdf

Cohen, D. K. & Mehta, J. D. (2017). Why reform sometimes succeeds: Understanding the conditions that produce reforms that last. American Educational Research Journal, 54(4), 644­–690. https://doi.org/10.3102/0002831217700078

Covitt, B. A., & Anderson, C. W. (2018). Assessing scientific genres of explanation, argument, and prediction. In A. L. Bailey, C. Maher, & L. Wilkinson (Eds.) Language, literacy, and learning in the STEM disciplines: How language counts for English learners, pp. 206-230. New York, NY: Routledge.

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](https://carbontime.bscs.org/sites/default/files/research/conference-presentations/4_CovittetalNARST2020.pdf). Annual meeting of the National Association for Research in Science Teaching, Portland, OR. (Conference canceled)

Covitt, B. A., Morrison Thomas, C., Lin, Q., de los Santos, E. X., & Anderson, C. W. (in preparation). Carbon TIME classroom discourse and its connections to student learning. A previous version of this article was prepared for presentation at the annual meeting of the National Association for Research in Science Teaching.

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>

Dauer, J., Doherty, J., Freed, A., and Anderson, C. W. (2014). Connections between student explanations and inquiry for plant photosynthesis and cellular respiration. CBE: Life Sciences Education, 13, 397–409.

Dauer, J., Miller, H. & Anderson, C. W. (2014). Conservation of energy: An analytical tool for student accounts of carbon-transforming processes. In R. Chen, A. Eisenkraft, D. Fortus, J. Krajcik, K. Neumann & A. Scheff (Eds.), Teaching and Learning of Energy in K-12 Education, pages 47–66. New York: Springer.

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>.

Dolle, J. R., Gomez, L. M., Russell, J. L., & Bryk, A. S. (2013). More than a network: Building professional communities for educational improvement. National Society for the Study of Education Yearbook, 112(2), 443–463.

EdBuild Report. (2009). $23 billion. Retrieved from: <https://edbuild.org/content/23-billion/full-report.pdf>.

Fishman, B., Penuel, W., Allen, A. R., Cheng, B., & Sabelli, N. O. R. A. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *Teachers College Record*, *115*(14), 136–156.

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>

Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. (2013). What would it take to change an inference? Using Rubin’s causal model to interpret the robustness of causal inferences. Education, Evaluation and Policy Analysis, 35, 437–460. [https://doi.org/10.3102/0162373713493129](https://doi.org/10.3102%2F0162373713493129)

Frank, K. A. (2000). Impact of a Confounding Variable on the Inference of a Regression Coefficient. Sociological Methods and Research, 29(1), 147–194. [https://doi.org/10.1177/0049124100029002001](https://doi.org/10.1177%2F0049124100029002001)

Furtak, E. M., & Penuel, W. R. (2019). Coming to terms: Addressing the persistence of “hands‐on” and other reform terminology in the era of science as practice. Science Education, 103(1), 167–186. https://doi.org/10.1002/sce.21488

Gee, J. P. (2005). Language in the science classroom: Academic social languages as the heart of school-based literacy. In Establishing Scientific Classroom Discourse Communities: Multiple Voices of Teaching and Learning Research (pp. 19–37). Routledge.Goldman, S. R., Greenleaf, C., Yukhymenko-Lescroart, M., Brown, W., Ko, M. L. M., Emig, J. M., George, M., Wallace, P., Blaum, D.& Britt, M. A. (2019). Explanatory modeling in science through text-based investigation: Testing the efficacy of the Project READI intervention approach. American Educational Research Journal, 56(4), 1148–1216. [https://doi-org.proxy.uchicago.edu/10.3102/0002831219831041](https://doi-org.proxy.uchicago.edu/10.3102%2F0002831219831041)

Greenberg, E., Blagg, K., & Rainer, M. (2019). Measuring Student Poverty: Developing Accurate Counts for School Funding, Accountability, and Research. Research Report. Urban Institute.

Grigg, J., Kelly, K. A., Gamoran, A., & Borman, G. D. (2013). Effects of two scientific inquiry professional development interventions on teaching practice. Educational evaluation and policy analysis, 35(1), 38–56. https://doi-org.proxy.uchicago.edu/10.3102/0162373712461851

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can value-added measures of teacher performance be trusted? Education Finance and Policy. 10(1), 117–156. <https://doi.org/10.1162/EDFP_a_00153>

Hanushek, Eric A. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. The Journal of Human Resources, 14(3), 351–388.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. The American Economic Review, 100(2), 267–271. <https://doi.org/10.1257/aer.100.2.267>

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. <https://doi.org/10.1002/pam.20464>

Heath, S. B. (1983). Ways with Words: Language, Life and Work in Communities and Classrooms. Cambridge University Press.

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. [https://doi.org/10.3102/0002831210387916](https://doi.org/10.3102%2F0002831210387916)

Jacob, B., Dynarski, S., Frank, K., Schneider, B. (2017) Are Expectations Alone Enough? Estimating the Effect of a Mandatory College-Prep Curriculum in Michigan. Educational Evaluation and Policy Analysis. 39(2), 333–360. https://doi:10.3102/0162373716685823

Jackson, K., &Cobb, P. (2013). Coordinating professional development across contexts and role group. In M. Evans (Ed.), Teacher education and pedagogy: Theory, policy and practice (pp. 80–99). New York, NY: Cambridge University Press.

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. <https://doi.org/10.1002/tea.21051>

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. 114(10), 1–39.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. Science Education, 98(4), 674–704. <https://doi.org/10.1002/sce.21123>

Kelderman, H. (1988). Common Item Equating Using the Loglinear Rasch Model. Journal of Educational Statistics, 13(4), 319–336. <https://doi.org/10.3102/10769986013004319>

Ladson-Billings, G. (2009). The dreamkeepers: Successful teachers of African American children. John Wiley & Sons.

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. Science Education, 96(4), 701–724. <https://doi.org/10.1002/sce.20475>

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual Review of Public Health, 23, 151–169.

Mikeska, J., Anderson, C. W., and Schwarz, C. V. (2009), Principled reasoning about problems of practice. *Science Education, 93(4),* 678–686.

Miller, H. K., & Anderson, C. W. (2017). Using NGSS crosscutting concepts as a tool for climate change and citizenship education. In D. Shepardson, A. Roychoudhury, and A. Hirsch (Eds.), Teaching and learning about climate change: A framework for educators (pp. 181–193). New York: Routledge.

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. <https://doi.org/10.1002/tea.20314>

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>.

Morrison Thomas, C., Covitt, B. A., Lin, Q., Marshall, S., Hancock, J. B., & Anderson, C. W. (in preparation). Carbon TIME teacher orientations and contexts: Making connections to classroom discourse and student learning. A previous version of this article was prepared for presentation at the annual meeting of the National Association for Research in Science Teaching.

Murphy, D. (2012). Where is the value in value-added modeling. (Pearson White Paper). Retrieved from the Pearson Education website: [http://images.pearsonassessments.com/images/tmrs/Where\_is\_the\_Value\_in\_Valu e-Added\_Modeling.pdf](http://images.pearsonassessments.com/images/tmrs/Where_is_the_Value_in_Valu%20e-Added_Modeling.pdf)

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. National Academies Press.

National Research Council. (2014). Developing assessments for the next generation science standards. Washington, DC: National Academies Press.

National Research Council. (2015). Guide to Implementing the Next Generation Science Standards. Washington, DC: National Academies Press.

National Research Council (2018). Science and engineering for grades 6-12: Investigation and design at the center. Washington, DC: National Academies Press.

NGSS Lead States. (2013). Next Generation Science Standards: For states, by states. Washington, DC: National Academies Press. <https://www.nextgenscience.org/>

NGSS Website (September, 2016). Using phenomena in NGSS-designed lessons and units. Retreived from <http://nextgenscience.org/sites/default/files/Using%20Phenomena%20in%20NGSS.pdf>

OpenSciEd. (2020). OpenScied Design Specifications. <https://www.openscied.org/design-specifications/>

Parsons, E. C., & Carlone, H. B. (2013). Culture and science education in the 21st century: Extending and making the cultural box more inclusive. Journal of Research in Science Teaching, 50(1), 1–11. <https://doi.org/10.1002/tea.21068>

Penuel, W. R. (2019). Infrastructuring as a practice of design-based research for supporting and studying equitable implementation and sustainability of innovations. Journal of the Learning Sciences, 28(4–5), 659–677. <https://doi.org/10.1080/10508406.2018.1552151>

Penuel, W. R., & Fishman, B. J. (2012). Large-scale intervention research we can use. Journal of Research in Science Teaching, 49(3), 281–304. <https://doi.org/10.1002/tea.21001>

Penuel, W. R., Fishman, B. J., Haugan Cheng, B., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. Educational Researcher, 40(7), 331–337.

Penuel, W. R., Turner, M. L., Jacobs, J. K., Van Horne, K., & Sumner, T. (2019). Developing tasks to assess phenomenon‐based science learning: Challenges and lessons learned from building proximal transfer tasks. Science Education, 103(6), 1367–1395. <https://doi.org/10.1002/sce.21544>

Penuel, W. R. and Reiser, B. (2018). Designing NGSS-aligned curriculum materials. Paper commissioned for the report Engaging Middle and High School Students in Science and Engineering: New Approaches to Investigation and Design from the Board on Science Education, National Academies of Science, Engineering, and Medicine.

Raudenbush, S. W. (2015). Value added: A case study in the mismatch between education research and policy. Educational Researcher, 44(2), 138–141. <https://doi.org/10.3102/0013189X15575345>

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. <https://doi.org/10.1111/j.1468-0262.2005.00584.x>

Roblin, N. P., Schunn, C., & McKenney, S. (2018). What are critical features of science curriculum materials that impact student and teacher outcomes?. Science education, 102(2), 260–282. https://doi.org/10.1002/sce.21328

Rosebery, A. S., Warren, B., & Lee, C. D. (2014). Learning as a cultural process: Achieving equity through diversity. In The Cambridge Handbook of the Learning Sciences, Second Edition (pp. 686–706). Cambridge University Press.

Schwarz, C. V., Braaten, M., Haverly, C., & de los Santos, E. X. (2020). Using Sense-Making Moments to Understand How Elementary Teachers’ Interactions Expand, Maintain, or Shut Down Sense-making in Science. Cognition and Instruction, 1–36. <https://doi.org/10.1080/07370008.2020.1763349>

Shadish, W. R., Clark, M. H., Steiner, & P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. Journal of the American statistical association, 103(484), 1334–1344. <https://doi.org/10.1198/016214508000000733>

Smith, C., Wiser, M., Anderson, C. W., and Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4(1-2),* 1–98*.*

Smith, D. C., and Anderson, C. W. (1999). Appropriating scientific practices and discourses with future elementary teachers. *Journal of Research in Science Teaching, 36(7)*, 755–776.

Tekkumru-Kisa, M., Stein, M. K., & Doyle, W. (2020). Theory and Research on Tasks Revisited: Task as a Context for Students’ Thinking in the Era of Ambitious Reforms in Mathematics and Science. Educational Researcher, 49(8), 606–617. https://doi.org/10.3102/0013189X20932480

Thomas, J. (2020, March). Using machine learning to interpret NGSS tasks at scale. Prepared for the annual meeting of the National Association for Research in Science Teaching, Portland, OR. (Conference canceled).

Thomas, J., Anderson, C. W., Draney, K., Frank, K., & Bathia, S. (2020). Using machine learning to assess and interpret NGSS tasks at scale. *Carbon TIME* project, Michigan State University.

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.

Thompson, J., Hagenah, S., Kang, H., Stroupe, D., Braaten, M., Colley, C., & Windschitl, M. (2016). Rigor and responsiveness in classroom activity. *Teachers College Record, 118*.

Tseng, V., Easton, J. Q., & Supplee, L. H. (2017). Research-practice partnerships: Building two-way streets of engagement. Social Policy Report, 30(4), 3–16.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009 and 2015 Science Assessments. retrieved from https://www.nationsreportcard.gov/science\_2015/#scores/chart\_loc\_1?grade=12

Willett, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. Review of research in education, 15(1), 345–422. <https://doi>.org/10.3102/0091732X015001345

Wise, S. L. (2017). Rapid‐guessing behavior: Its identification, interpretation, and implications. Educational Measurement: Issues and Practice, 36(4), 52–61. <https://doi.org/10.1111/emip.12165>

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. Journal of Research on Educational Effectiveness, 10(1), 207–236. <http://dx.doi.org/10.1080/19345747.2016.1164781>

1. Middle- and high-school teachers used the same units, with some simpler activities designed primarily for middle-school students and some more complex activities designed primarily for high school students. Our learning progression research showed that virtually none of the high-school students had mastered middle-school NGSS performance expectation (see Results for Research Question 1 below), so some high-school activities addressed NGSS middle-school performance expectations. [↑](#footnote-ref-2)
2. Our learning progression frameworks distinguish between level 1 (simple force-dynamic accounts) and level 2 (elaborated force-dynamic accounts), but our assessments for secondary students were not designed to assess this distinction. The differentiation between level 1 and level 2 occurs primarily in the upper elementary grades. Since this study focuses on middle and high school students only; we developed items for these assessments that reliably distinguish levels 2 through 4. [↑](#footnote-ref-3)
3. This means if a teacher taught two classrooms in one year, that two classrooms were combined into one cluster. Teachers were inconsistent in how they designated classes in the online assessment system; thus, our data were not adequate for an analysis that including an additional level for different classes taught by the same teacher in the same year. This likely contributes to underestimation of standard error, generating conservative inferences. [↑](#footnote-ref-4)
4. The ICC was calculated as in the unconditional model to measure how much of the total variance was at level 2 (i.e., classroom or teacher-year level in this paper). [↑](#footnote-ref-5)
5. In comparison, the between-year variation only accounts for 3%. That is, the variation in students’ learning outcomes between academic years is small relative to the variation within years. [↑](#footnote-ref-6)
6. Since teachers were identified in the second semester of year before they joined the cohort, a baseline pretest was not feasible. However, when we compared Teacher Year 1 pretests with Teacher Year 2 pretests for the Analytical Sample 1 teachers, we found that mean pretest scores were ­–1.46 for Year 1 and –1.53 for Year 2, with a difference of only –0.07 logits. All this said, we agree that there still could be concerns about. Therefore, we quantified the strength of evidence for the inference: more than 94% of the estimate would have to be due to bias to change the inference. This is more robust than all the studies reviewed by Frank et al. (2013). [↑](#footnote-ref-7)