Designing Educational Systems to Support Enactment of the Next Generation Science Standards

By Charles W. Anderson, Elizabeth de los Santos, Sarah Bodbyl Roels, Beth Covitt, Kirsten D. Edwards, James Brian Hancock, II, Qinyun Lin, Christie Morrison Thomas, William Penuel, MaryMargaret Welch

Table of Contents

Abstract

The *Next Generation Science Standards* describe goals for three-dimensional learning that represent a broad consensus of science educators. Yet we aspire to achieve these goals in educational systems that are fragmented and resource-poor. Long-term improvements based on *NGSS* will require research and development that attend to *systems* of teaching and learning, including key communities of practice: science classrooms, schools, and professional networks. This paper focuses on a design-based implementation research project that aims to provide material, human, and social resources to these communities of practice. Our research program is organized around four research and development goals: (1) Learning progression frameworks and assessments focused on key scientific practices and crosscutting concepts in an important content domain: carbon cycling from atomic-molecular to global scales; (2) An instructional model and curriculum units that function as tool kits to scaffold students' successful engagement in three-dimensional learning; (3) Professional development to promote responsive and rigorous teaching; and, (4) Working with teachers and key personnel in their schools and districts to understand and change purposes, norms, and obligations in classroom and school communities of practice. Results from the first two years of the project provide evidence that it is possible to measure and achieve three-dimensional learning at scale. However, this accomplishment requires substantial investments in the material, human, and social resources of educational communities of practice.

Keywords: curriculum development, achievement, teacher education, practicing teachers

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. Additional support comes from the Great Lakes Bioenergy Research Center (United States Department of Energy Office of Science BER DE-FC02-07ER64494), and the Dow Chemical Company Foundation]

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

Designing Educational Systems to Support Enactment of the *Next Generation Science Standards*

There are encouraging developments in science education: Communities of science educators have achieved a broad consensus around the *Framework for K-12 Science Education* (National Research Council, 2012) and the *Next Generation Science Standards* (*NGSS*). In particular, the framework of three-dimensional science learning has provided a common language for articulating goals for science learning and discussing how those goals can be achieved. Yet we also face critical problems: Resources and programs for enacting this shared vision are increasingly fragmented and privatized. In the United States, science teachers and school districts rely on a smorgasbord of bits and pieces of curriculum and instructional strategies that compete for teachers' time and attention—individual lessons found on the Internet, workshops on different topics, etc. This fragmentation is one consequence of the nation's failure to invest in the resources that our schools need to achieve ambitious science teaching and learning goals. The mismatch between the coherent goals of *NGSS* and the fragmented means we are using to address them is obvious.

Hiebert and Stigler (2017) argue that long-term improvements in education require attention to *systems* of teaching and learning. Our research program involves understanding how science education systems are currently working, including tensions between current structures and functions and the goal of three-dimensional learning. Furthermore, supporting enactment of *NGSS* requires us to not just understand science education systems but develop ways to *improve* them. In this paper, we report on a design-based implementation research (DBIR; Penuel & Fishman, 2012) project in which we aim to work with school communities and professional networks to build systems that accomplish three-dimensional learning at scale. DBIR studies design and test systems at moderately large scale involving dozens to hundreds of teachers. They differ from very large-scale studies, which can identify factors affecting student success but thus far have rarely documented successful large-scale innovations. They also differ from small scale design-based research studies (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) that often depend on exceptional performances by teachers or small cohesive communities of practice.

The purpose of this paper is to describe *Carbon, Transformations in Matter and Energy*, or *Carbon TIME* from a systems perspective and share what we have learned so far about how to design educational systems to support enactment of *NGSS*. Our paper has three parts:

- In the remainder of the introduction we (a) briefly describe the goals of the project and its enactment through an iterative research and development cycle, and (b) present a framework for describing and analyzing the project in the context of educational systems as communities of practice with material, human, and social resources.
- In the main section, we describe research findings and progress in development to date, using the key goals of the iterative research and development cycle: (1) using learning progressions to define learning goals and develop assessments, (2) developing curricular resources, (3) supporting productive classroom discourse and practices, and (4) supporting school and professional communities of practice.
- In the discussion, we consider the implications of what we have learned for designing educational systems to support enactment of *NGSS* at scale.

The *Carbon TIME* Project

Carbon TIME focuses on improving secondary science learning and has been supported by a series of National Science Foundation (NSF) grants since 2005. Like other design-based research projects (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003), our project seeks both to develop new knowledge in the field of science education and to develop useful products—in this case, tools for improving systems of science education. We have worked through a series of *iterative development cycles* towards the four goals listed above, each of which has a research and a development dimension. In its early stages, the project focused primarily on the first goal (defining student learning goals and developing assessments); over time, it has expanded to develop knowledge and resources for the other three goals. We are currently in the fourth year of a five-year NSF grant that seeks to develop resources for large-scale implementation of the curriculum we developed using a previous grant. In the remainder of this section we briefly describe our current activities around the four goals. Key findings are described in the section on research findings and progress in development.

Goal 1: Using Learning Progressions to Define Learning Goals and Develop Assessments

This project began before *NGSS* and the *Framework for K-12 Science Education* (National Research Council, 2012) were published. In fact, our learning progression research informed the development of the *Framework.* We began with the general goal of *environmental science literacy:* preparing students to use scientific knowledge and practices in their decisions about environmental issues. Our current learning goals and research and development progress are described below.

Carbon TIME focuses on key carbon-transforming processes in socio-ecological systems at multiple scales: cellular and organismal metabolism in plants, animals, and decomposers; energy flow and carbon cycling at ecosystem and global scales; carbon sequestration; and, combustion of fossil fuels. The current imbalance among these processes is a primary driver of global climate change. Our online supplemental materials include tables documenting the middle- and high-school *NGSS* performance expectations addressed in our assessments and teaching materials (see Mapping Supplement in Supplementary Materials). In particular, our work has focused on:

- All eight science practices, organized into two clusters: inquiry (asking questions, planning and carrying out investigations, analyzing and interpreting data, engaging in argument from evidence) and application (developing and using models, constructing explanations, designing solutions).[1]
- Three crosscutting concepts: scale, proportion, and quantity; systems and system models; and, energy and matter: flows, cycles, and conservation.
- Disciplinary core ideas in the life sciences (LS1: From molecules to Organisms: Structures and Processes; LS2: Ecosystems: Interactions, Energy, and Dynamics); Earth and space sciences (ESS2: Earth's Systems; ESS3: Earth and Human Activity); and physical sciences (PS1: Matter and Its Interactions; PS3: Energy).

Research dimension: Learning progression frameworks. The research dimension of this project has focused on the development of learning progression frameworks focused on three practices: (a) explanations of carbon-transforming processes, (b) inquiry and arguments from evidence, and (c) interpreting data, predictions, and explanations of carbon cycling at ecosystem and global scales.

Development dimension: Assessments of three-dimensional performances. The learning progression frameworks provided the foundation for developing assessments to elicit a range of student responses. We developed an assessment system that includes formative assessments teachers can use to facilitate three-dimensional performances and summative assessments that help teachers identify students' responses in terms of learning progression levels.

Goal 2: Developing Curricular Resources

The learning progression frameworks and assessments provided the foundation for the development of six units about carbon cycling and energy flow at multiple scales (http://carbontime.bscs.org/). Four units (*Systems and Scale, Plants, Animals, Decomposers*) focus on investigations of macroscopic-scale processes and explanations using atomic-molecular models. The other two units (*Ecosystems, Human Energy Systems*) focus on ecosystems and global carbon cycling and climate change. Curriculum development is an iterative process: We incorporated revisions and

---

[1] Using mathematics and computational thinking and obtaining and obtaining, evaluating, and communicating information are included in both clusters.

additions each year in a collaborative feedback and research cycle between project *Carbon TIME* staff and teachers.

Research dimension: Developing an instructional model to scaffold three-dimensional learning. The learning progression frameworks and assessments define learning goals and measure students' progress as they complete units, but they do not prescribe the methods by which those goals should be achieved. We have drawn on learning research (e.g., Collins, Brown, & Newman, 1989; Liew & Treagust, 1995; National Research Council, 2007; Sadler, 1989; Quintana et al., 2004) as well as our own experiences to design an instructional model that supports students' engagement in three-dimensional practices using an inquiry-application sequence. The instructional model includes a set of key features such as investigations that provide data for students to engage in argument from evidence, scaffolds for constructing model-based scientific explanations, discourse routines for supporting student sense-making (e.g., eliciting and building on students' ideas), and checking for understanding using three-dimensional formative assessments. The online supplemental materials include a description of the instructional model (see Materials Supplement).

Development dimension: Instructional units. The instructional model provides what Roschelle, Knudsen, and Hegedus (2010) refer to as an infrastructure for curriculum and instruction but not the detailed support for teaching particular topics—what they refer to as a *curricular activity system*. Our experience is consistent with the argument that the authors make: Infrastructure alone is insufficient. Much of the development work on this project has gone into developing the units that constitute the *Carbon TIME* curricular activity system. Each unit provides an extensive tool kit for teachers and is organized according to the inquiry-application sequence embedded in the instructional model, including resources such as plans for investigations, online videos and simulations, molecular modeling activities, animated presentation slides, and readings. Each unit includes a set of four Process Tools that scaffold students' engagement in three-dimensional learning through expressing their ideas about phenomena, making predictions about the results of an investigation, using data to engage in argument from evidence, and constructing explanations of macroscopic-scale phenomena using atomic-molecular models.

Goal 3: Supporting Productive Classroom Discourse and Practices

The curriculum provides teachers with an extensive tool kit, but the tools do not use themselves. That is, the instructional materials are designed to work best in classroom learning communities where the discourse is both *rigorous* (students are held accountable for learning) and *responsive* to students' needs and ideas. Thus, the third goal of this project focuses on supporting teachers in scaffolding classroom discourse that has these qualities.

Research dimension: Investigating patterns in classroom discourse. Our research on this goal focuses on analyzing classroom discourse from 17 case study classrooms. Data sources include videos of classroom instruction, interviews with teachers and focus students, and focus students' written work (e.g., Process Tools). We observed different discourse patterns across classrooms and made some progress in understanding their causes and consequences. In some classrooms, the discourse patterns are both *responsive* to students' ideas and interests and *rigorous* in scaffolding three-dimensional learning performances. In other classrooms, teachers use and adapt *Carbon TIME* materials to support task-oriented discourse that keeps students busy but does not necessarily engage them in deeper learning.

Development dimension: A professional development course of study*.* In partnership with local education organizations, we developed a two-year course of study for teachers who participate in our project. Using a cohort model, the professional development includes ten days of face-to-face workshops and online courses throughout the school year. In all phases, teachers share their ideas and learn with and from each other and members of the *Carbon TIME* team, contributing to the iterative refinement of *Carbon TIME* material resources while also empowering teachers as collaborators. The development of the scope and sequence of the course of study is also an iterative process and intended to be responsive to local context and teachers' needs (Henrick, Munoz, & Cobb, 2016).

Goal 4: Supporting School and Professional Communities of Practice

Teachers' schools and professional networks are complex communities with varying norms and expectations. We differentiate between teachers' school communities (e.g., colleagues and administrators) and professional networks because they may or may not overlap. For example, teachers may collaborate with other teachers in their district but not at their school. The teachers who have implemented our curriculum often find that their new *Carbon TIME* tool kit can create opportunities for collaboration with school colleagues. Or, implementing the curriculum can cause conflicts with their established school culture and expectations. In some schools, non-*Carbon TIME* teachers are eager to learn about the resources; other schools are not as open to innovation. The culture of collaboration, conflict, or isolation at a school can affect *Carbon TIME* teachers' opportunities to make progress towards developing responsive and rigorous teaching practices.

Research dimension: Understanding how school and professional communities influence teachers' sensemaking about their implementation of the curriculum. To investigate how teachers' school and professional communities influenced their sensemaking about implementation of the curriculum, we used a quantitative approach to survey all participating teachers and an in-depth qualitative approach to analyze data from selected case study teachers. Our quantitative approach involved using social network analysis methods (e.g., Frank, Kim, & Belman, 2010; Penuel, Fishman, Haugan Cheng, & Sabelli, 2011) to visualize teachers' professional networks (i.e., who they seek help from) over time. Our qualitative approach involved using the concept of organizational sensemaking (Weick, 1995) to explore the uncertainties and conflicts that teachers critically noticed about interactions among the curriculum, their students, and themselves as they implemented the curriculum and engaged in project activities (de los Santos, 2017). Combining both approaches affords us the opportunity to develop a deeper understanding of the challenges posed by implementing the project in particular and three-dimensional learning in general.

Development dimension: Developing professional networks and partnerships. *Carbon TIME* networks are assembled from local or regional groups of teachers and led by project staff members in partnership with local teachers, science coordinators, or other key personnel. These networks function as bridges between communities of practice, connecting teachers who work in classrooms and have obligations to school and professional communities with project staff and the educational research community. Our intent is to build resources and practical knowledge that can be adopted and sustained by classroom, school, and professional communities.

Resources and Communities of Practice in Science Education Systems

Science education currently holds ambitious, coherent goals for an educational system that is fragmented and resource-poor. Like other complex designed systems (e.g., automobiles, computers, or hospitals), science education systems have many interacting parts, each of which are necessary but insufficient for the system to work as a whole. The four goals in this project represent our "bets" about how to improve science education systems, or the areas where new resources will best translate into students' engagement in three-dimensional learning and their ultimate development into scientifically literate citizens.

In taking a systems perspective, we are not simply highlighting the innovative features of our curriculum, assessments, or professional development program. We do not want to encourage the notion that novel features are the "key" that will unlock successful innovation on a large scale. Instead, we see our innovations taking place within educational systems that consist of multiple interacting communities of practice, including classrooms and professional networks. Additionally, we see our DBIR research and development programs as working with those communities for the purpose of positive change.

Material, human, and social resources. Developing educational systems that support ambitious, three-dimensional learning requires material, human, and social resources (Lee, Llosa, Jiang, O'Connor, & Haas, 2016; Spillane, Diamond, Walker, Halverson, & Jita, 2001). *Material* *resources* include time, money, laboratory and classroom space, related equipment, curriculum, and assessments. *Human resources* include “individual knowledge, skills and expertise” of people (Spillane et al., 2001, p. 920), as well as their vision of teaching. *Social resources* include relationships among individuals, which can support the distribution of material and human resources. Additionally, social resources can have enduring qualities, such as purposes, norms, and obligations, or the “culture” of an organization, that extend beyond individual community members. Thus, participation in a community of practice entails what Wenger (1998, p. 81) refers to as “regimes of mutual accountability,” or ways in which individuals are held socially accountable to each other. Targeting the development of and access to material, human, and social resources in specific communities of practice can help build the capacity of the education system as a whole.

Communities of practice in the project. All of these resources exist and are accessed through communities of practice, which are people working together with shared purposes and norms and who hold mutual obligations to each other. Figure 1 represents the key science education communities of practice involved in the *Carbon TIME* project, with numbers indicating the size of each community. The largest communities are classroom communities, with approximately 900 classes over the course of five years. School and professional communities are the next largest, with about 99 schools in 64 districts. And, project staff are the smallest community, with 30 university and school personnel.

Relationships between resources and communities of practice. The communities of practice in Figure 1 are able to influence one another because they have overlapping members (boundary crossers) and exchange material resources (boundary objects). Boundary crossers can carry resources from one community of practice to another, and boundary objects (Star & Griesemer, 1989) are material artifacts that are present in multiple communities but with different meanings and can coordinate work across settings through negotiation of those meanings. For example, researchers and teachers negotiate the meaning of the curriculum materials in the professional development setting but that meaning can change when teachers take those curriculum materials back to their classrooms and negotiate the meaning with their classroom communities. Thus, the project can function only in as much as we can provide boundary objects that enable teachers in classroom and school communities to do their work more effectively and boundary crossers with the understanding needed to act effectively in various communities of practice.

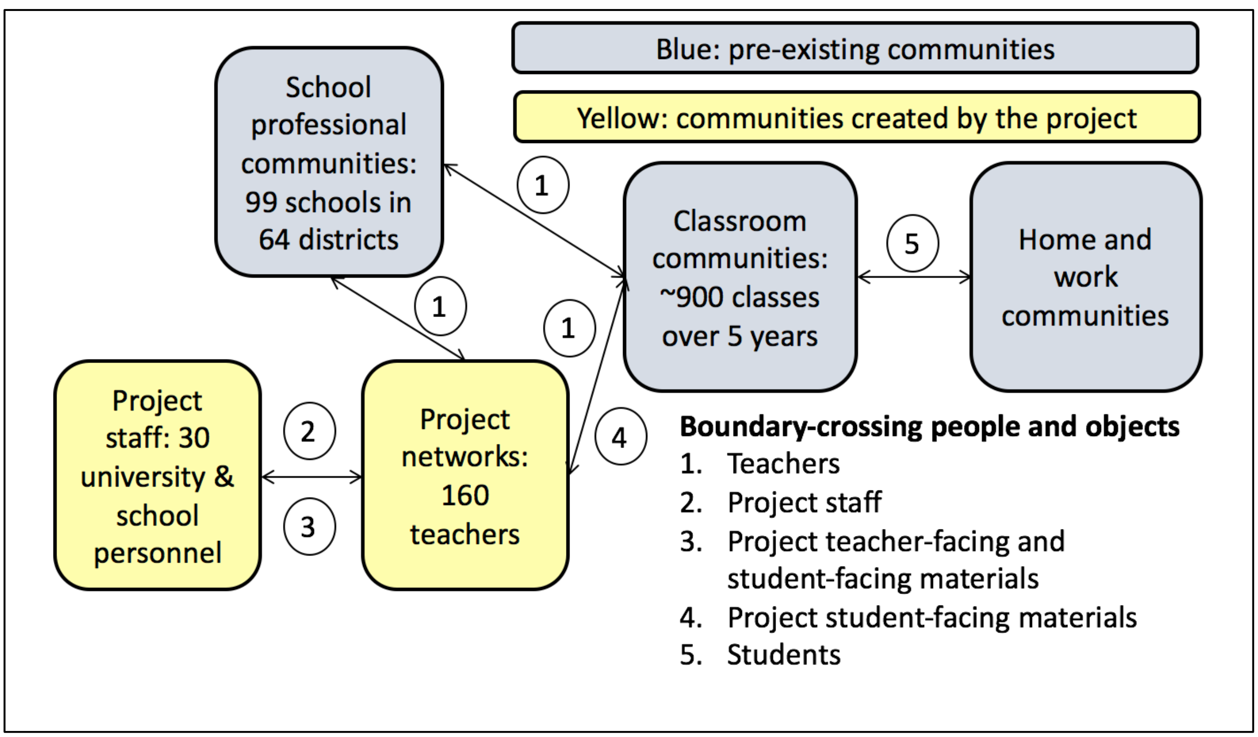

Figure 1: Carbon TIME communities and boundary-crossing people and objects. Numbers are total numbers involved in the current project (2015–19)

As *Carbon TIME* staff, our work involves spending substantial time with science teachers through face-to-face and online professional development in *Carbon TIME* networks. Despite the effort we have put into building project-network communities of practice, we believe they are peripheral to *classroom communities*, where the main business of science teaching and learning actually takes place. Teachers also spend more time with colleagues in their schools and districts—their *school and professional communities*—than in *Carbon TIME* network activities. Thus, our influence on teachers' practices and student learning is limited. However, we believe that our project's strategic integration of material, human, and social resources has developed the capacity of the larger communities of practice in ways that help students meet ambitious, three-dimensional *NGSS* science learning goals and can be a model for deploying resources in effective ways. The next section outlines our research and development progress in each goal area.

Research Findings and Progress Toward Development Goals

In this section, we present a progress report based on current and previous research and development cycles in the *Carbon TIME* project. Data come from the first three years of this DBIR study; the full five years of the study will involve about 160 participating teachers in diverse middle and high school classrooms in three states, with each teacher and their students providing data for two successive years. Table 1 outlines major project data sources and quantities of data collected. In the remainder of this section, we report on progress with respect to each of the four research and development goals. Data analyses come mostly from the first full year of implementation, 2015-2016 including case study classrooms (Yin, 2014)

Table 1. Data Sources for the Project

Data Source	**Baseline Year (2014–5)**	**First Year (2015–6)**	**Second Year (2016–7)**	**Additional Data***
Full Data Set (120 participating middle and high school teachers)				
Participating teachers	17	40	105	
Student tests (8/student)	2,920	21,058	60,878	244
Teacher surveys (3/teacher each year)	104	169	294	
PD videos & field notes (3 days/cohort)	0	52 hrs.	95 hrs.	
Online PD (~10 hours/cohort)**	0	300 hrs.	450 hrs.	
Case Study Data Set (17 cases involving 14 teachers: 5 middle school, 9 high school)				
Participating teachers		8	9	
Student interviews (4 focus students/class)		40	65	52
Teacher interviews (5/teacher)		22	47	
Classroom videos (~10 lessons/teacher, 2 videos/lesson)		195	197	
Student work (~12 examples/focus student)		472	498	
* We also collected some interview and test data from college students for learning progression development. ** We collected video, field notes, assignments, and discussion threads from 3 days of face-to-face and ~10 hours of online PD each year for each teacher.				

Goal 1: Using Learning Progressions to Define Learning Goals and Develop Assessments

Measuring student learning is a complex endeavor, and the *NGSS* provide new challenges for defining and evaluating student learning goals. The *NGSS* performance expectations interweave disciplinary core ideas, science and engineering practices, and crosscutting concepts, but they do not identify relevant phenomena, provide criteria for adequate performance, or identify students' intellectual capacities and barriers to learning. Systems designed to support enactment of *NGSS* require effective ways to develop and refine frameworks and assessments that have these qualities. In this section, we describe how the project has carried out iterative cycles of resource development and refinement, including (1) the development of empirically validated learning progression (LP) frameworks that describe levels of proficiency as students develop scientific knowledge and practice with instructional support, and (2) the development and validation of written assessments that can be used to measure students' three-dimensional learning.

Research dimension: Student learning progressions

In LP research, the development and validation of frameworks and assessments are connected in a cyclical, iterative process built around an "assessment triangle": (a) articulation of a cognitive model, (b) design and implementation of assessments, and (c) interpretation and analysis (National Research Council, 2001; 2005; 2014). Implementing the assessment triangle is challenging in the context of *NGSS* because of the three-dimensional nature of the standards. Models of cognition, assessments, and analyses cannot focus on just a core disciplinary idea, science practice, or crosscutting concept; instead, all three components need to be integrated in performances that engage students with real-world phenomena. This integration means that *NGSS* performances involve students engaged in complex forms of talk and writing, not just selecting the answer among choices in forced-choice questions.

Through more than a decade of iterative assessment cycles, we have developed three interconnected LPs focused on different practices: (a) explanations of carbon-transforming processes at the macroscopic scale, (b) inquiry and arguments from evidence, and (c) interpreting data, predictions, and explanations of carbon cycling at ecosystem and global scales. These LPs describe transformations in students' knowledge and practice that are necessary to master *NGSS* performance expectations (Covitt & Anderson, in press; Dauer, Doherty, Freed, & Anderson, 2014; Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009). We describe our LP frameworks as *discourse-based* in that increasing sophistication involves mastering new forms of talk and writing that represent changes in worldview and student sense-making about phenomena. These new intellectual resources enable students to participate in scientific model-based discourse.

The three LPs involve substantially different practices, and we have evidence that students' progress on the different LPs is not highly correlated. Nevertheless, all three LPs describe a broad shift from colloquial to scientific discourse that is somewhat like learning a second language: Students retain their proficiency in colloquial explanations and arguments while mastering new forms of discourse that are personally and socially valuable. Next, we describe this shift in general terms that include all the LPs.

Colloquial (lower-level) discourse. Students' lower-level performances rely largely on knowledge that they have acquired outside of school and discourse practices that are common in everyday life. For example:

- Explanations of familiar carbon-transforming phenomena (e.g., plant growth, combustion, decay) that rely on force-dynamic discourse involving actors, purposes, and enablers (Pinker, 2007; Talmy, 1988).
- Critiquing arguments from evidence by invoking folk theories or substituting easier questions while viewing alternate hypotheses as compatible "good ideas" (Windschitl, 2004).

- Making predictions about large-scale phenomena based on "covering laws" that generally distinguish between things that are good and bad for the environment without considering mechanisms such as carbon cycling (Braaten & Windschitl, 2011).

Intermediate-level discourse. Intermediate-level performances include mixtures of colloquial and scientific discourse. Intermediate-level explanations of phenomena may trace matter and energy with errors or rely on inaccurate or incomplete mechanisms. Intermediate critiques of arguments from evidence attend to data but sometimes neglect key patterns in those data and may consider evidence about carbon-transforming processes in light of inaccurate scientific principles (e.g., gases don't have mass). Intermediate predictions of large-scale phenomena reveal capacities to identify trends in non-noisy data while also showing challenges such as making predictions that extend trends in data without invoking associated scientific models.

Scientific (high-level) discourse. High-level performances are consistent with the goals of *NGSS.* Key characteristics include:

- Tracing matter and energy through systems in ways that consistently conform with conservation laws, including the ability to use pool-and-flux models for large-scale systems.
- Connecting system models at different scales, including atomic-molecular, cellular, macroscopic, ecosystem, and global scales.
- Precision in language by consistently distinguishing, for example, matter from energy and atoms from molecules.

A key characteristic of high-level performances is students' ability to use scientific principles incorporated into *NGSS* as crosscutting concepts such as energy and matter; scale, proportion, and quantity; and, systems and system models (Miller & Anderson, 2017). We also note that scientific discourse is not merely an intellectual performance. It involves participation in communities of practice that share purposes and norms for communication. Scientific discourse also requires commitments or habits of mind, such as commitments to rigor and precision in thought and language.

Development dimension: Assessments of three-dimensional performances

Assessments of three-dimensional performances in *Carbon TIME* support both iterative research and development cycles and classroom formative and summative assessment. In order to serve these purposes, the assessment system must meet the criteria of validity and reliability. Content and construct validity evidence includes development of: (1) assessments that elicit three-dimensional performances that require making sense of phenomena, and (2) assessments that elicit responses ranging from colloquial to scientific discourse. We found that items combining forced-choice with an accompanying written explanation provided the evidence necessary to examine how students were making sense of phenomena and what type of discourse they were employing. Closed-ended only items provided insufficient evidence for these purposes. *Carbon TIME*'s assessment system also demonstrated response process validity evidence, which was established through examining levels of consistency between how students talked about phenomena during in-depth clinical interviews and how they responded to written assessment items. Finally, internal structure validity evidence for assessments was established through meeting standards for reliability and item functioning. These forms of evidence are discussed in further detail by Doherty, Draney, Shin, Kim, & Anderson (2015). Analyses based on item response theory (IRT) are included in the online supplemental materials.

One challenge with applying discourse-based and three-dimensional assessment systems to large-scale projects involves the need to code hundreds of thousands of open-ended responses generated by students. In *Carbon TIME*, the scaling up of the assessment system has been made possible through the development of an online assessment system and automated scoring for constructed response explanations. *Carbon TIME* students' online responses generate a large database that includes forced-choice and constructed responses collected at multiple time points before, during, and after their *Carbon TIME* instructional experiences. Students' constructed responses are coded in batches by the automated scoring system. The process of developing and testing the *Carbon TIME* machine learning model for constructed response items has involved several steps, including the development of written exemplar worksheets, human coding of "training response sets," and subsequent machine coding training on the sets using the open-source LightSide platform developed by researchers at Carnegie Mellon University. Using the automated scoring system, most items become machine-scorable with an acceptable Quadratic Weighted Kappa of greater than 0.70. Other items are revised or dropped from the assessment.

Goal 2: Developing Curricular Resources

Our baseline assessments demonstrate that current teaching practices do not lead to mastery of three-dimensional learning. In classrooms using non-*Carbon TIME* materials, less than 10% of student responses are at the highest level in our learning progression—equivalent to the *NGSS* performance expectations (Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009). In order to meet the goals of *NGSS*, teachers need flexible and learning progression-based curricular resources aligned to the three-dimensions of the standards. The research dimension of our work on this goal involves development of what Roschelle, et al. (2010) describe as a curriculum infrastructure: an instructional model and related features that recur in every unit. The development dimension involves the design of units focused on particular systems and processes that provide tool kits for teachers' learning of responsive and rigorous teaching practices.

Research dimension: Developing an instructional model to scaffold three-dimensional learning

Drawing from learning theories and our own learning progression research, we have developed an instructional model as an infrastructure (Roschelle, et al., 2010) for three-dimensional science learning that can be used across multiple science topics. The instructional model is based on theories of learning and research on science teaching, particularly cognitive apprenticeship and the synthesis of research on the teaching and learning of science in *Taking Science to School* (Collins, Brown, & Newman, 1989; National Research Council, 2007).

The instructional model incorporates two intertwined storylines: a science content storyline and a student storyline. The science content storyline focuses on particular systems and carbon-transforming processes. For example, in the *Plants* unit, students learn how plants transform matter and energy by creating organic materials through the process of photosynthesis and then transforming those materials as they grow (biosynthesis) and use energy to function (cellular respiration). Students figure out the scientific story through a sequence of activities that engage them in different practices. The student storyline involves placing students in different roles: (1) students as *questioners*: students express their initial ideas and pose questions around an initial phenomenon (e.g., how radish plants grow); (2) students as *investigators*: students conduct investigations that involve tracing matter and energy through systems, following a PEOE (predict, initial explanation, observe, revised explanation) sequence; they end their investigation by engaging in argument from evidence, striving for consensus conclusions, and identifying unanswered questions; (3) students as *explainers*: students follow a cognitive apprenticeship sequence (Collins, et al., 1989) as they develop explanations that trace transformations of matter and energy in living and Earth systems.

Within each portion of the instructional model, students receive scaffolding (Quintana et al., 2004) that enables them to engage successfully in these complex scientific practices. This scaffolding takes multiple forms, including (a) Process Tools that help students record and organize their writing and thinking, (b) a set of Three Questions that both define the components of model-based explanations (e.g., tracing atoms) and guide the principled reasoning used in such explanations (e.g., conservation of matter), and (c) discourse routines that engage students in sharing ideas and seeking consensus during small-group and whole-class discussions. This infrastructure supports student engagement as well as formative and summative assessment (Sadler, 1989). Both teachers and students can see progress as students figure out the phenomena of interest, engage in scientific practices while adhering to scientific principles, and develop more rigorous, model-based explanations of phenomena.

Development dimension: Instructional units as tool kits for supporting responsive and rigorous teaching

An infrastructure alone provides insufficient support for classroom instruction. For this reason, we have developed curriculum materials for specific content, or a curricular activity system (Roschelle et al., 2010). Our units are content-specific tool kits that teachers can use based on knowledge of their students and requirements of their local context.

We have developed a total of six units, each designed to last about three weeks. Four units focus on systems at the macroscopic scale: *Systems and Scale, Animals, Plants,* and *Decomposers.* These units each follow the instructional model as they focus on particular systems and processes. For example, in the *Plants* unit students (a) express ideas and pose questions about how plants grow, (b) investigate where the mass of plants comes from, (c) use molecular models to model photosynthesis, biosynthesis, and cellular respiration, and (d) write explanations of these processes. We have developed two large-scale units: *Ecosystems* and *Human Energy Systems*. These units still align with the instructional model, but the scale of focus necessitates different content-focused materials. For example, instead of investigations using living organisms, students conduct investigations using online simulations. Both the instructional model and the individual units continue to be iteratively refined based on student assessment data, observations of teacher enactment, and teacher feedback. Of note is how responsive our curriculum development team has been to feedback from teachers and students about what works and does not work in their classrooms. For example, teachers and students found the Three Questions confusing, so the development team changed one of the questions to differentiate better between movement of matter and transformation of molecules.

Effects on student achievement. Our data on student learning have provided strong evidence that our *Carbon TIME* curriculum and support system make a difference in student achievement. Figure 2 compares IRT-based estimates of student pretest and posttest proficiencies with baseline levels (students of the same teacher the year before) for this study and a previous study. (For detailed methods and results see the online supplemental materials and Doherty et al., 2015). The differences are statistically and educationally significant (The error bars on Figure 2 are 95% confidence intervals).

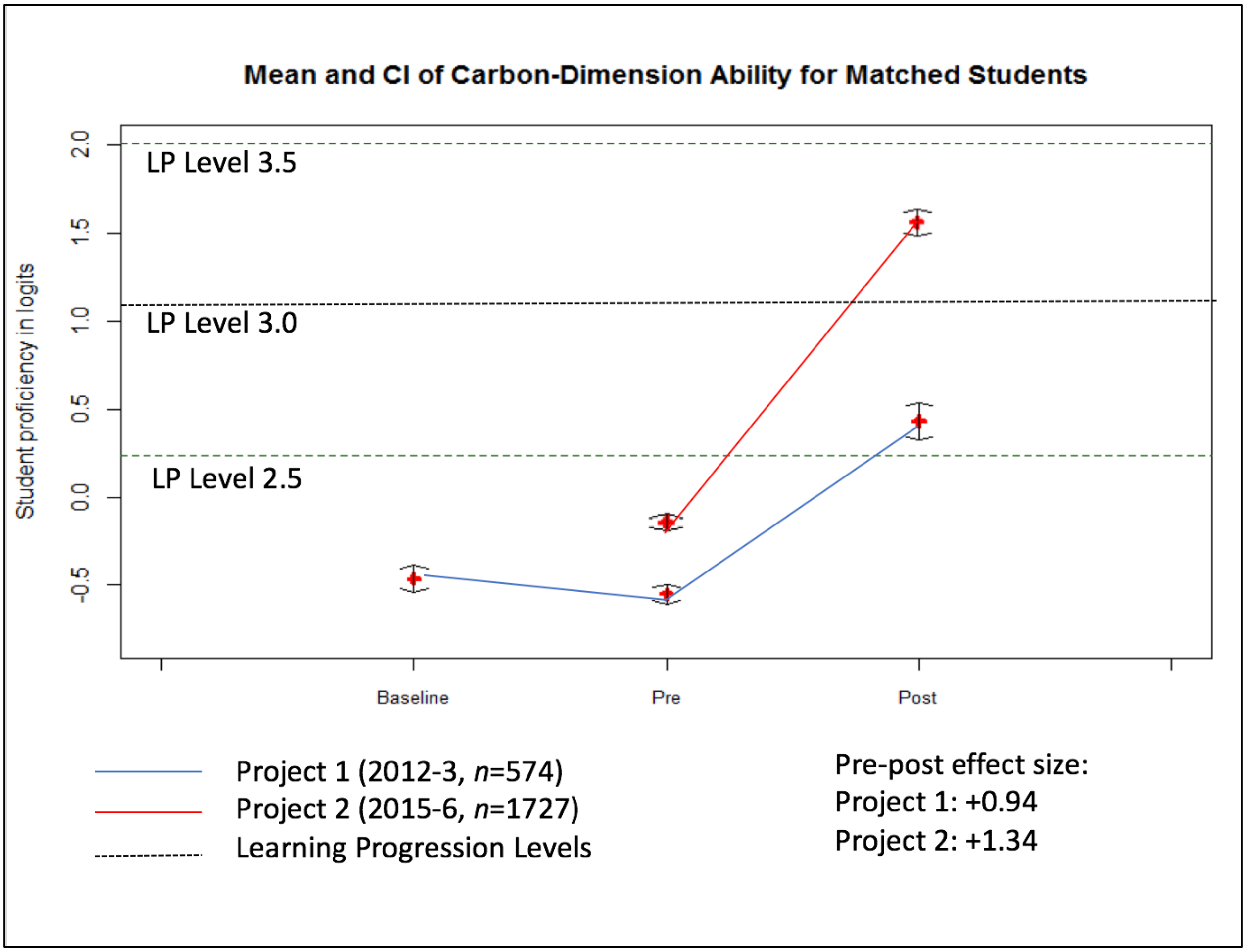

Figure 2: Carbon TIME and baseline learning progression levels

In a series of studies, we have shown that high school students using our project materials show higher proficiency on learning progression-based assessments than college science majors in biology courses or students of project teachers before they joined the program (Doherty et al., 2015; Jin & Anderson, 2012; Rice, Doherty, & Anderson, 2014). But these results also show that most students still fall short of the highest learning progression level that would be equivalent to the *NGSS* performance expectations in this domain.

Goal 3: Supporting Productive Classroom Discourse and Practices

Our data indicate that there is wide variation in the rigor and responsiveness of teachers' instructional practices, which means that, at the level of classroom communities of practice, teachers and students engage differently with the *Carbon TIME* tools described above. These results helped us to identify our primary development goal: creating a professional development course of study that supports teachers in developing rigorous and responsive classroom discourse.

Research dimension: Investigating patterns in classroom discourse

Using classroom video data from case study classrooms, we noticed that the *Carbon TIME* teachers we judged as more successful at scaffolding rigorous and responsive discourse were articulating a classroom purpose for students to "figure out" the phenomena that were presented in the *Carbon TIME* units. In these classrooms, students took up roles as epistemic agents (Stroupe, 2014) and both the teacher and students employed language that illustrated evidence of principled reasoning around core science ideas (Johnson et al., 2017). For these teachers, *Carbon TIME* instructional materials and assessments were leveraged as a tool kit that scaffolds students' engagement in productive classroom discourse and three-dimensional learning. In less successful classrooms, however, the purpose was not to "figure out" phenomena but instead "learn about" science (Achieve, 2016). For these teachers, *Carbon TIME* tools were treated as worksheets and incorporated into systems of task-based discourse that held students primarily accountable for procedural display: Students were considered successful if they could reproduce essential content but were not held accountable for three-dimensional understanding.

Effects of classroom discourse on student achievement. Our quantitative data confirm that teachers and classroom discourse patterns make a difference on students learning outcomes. Figure 3 compares student pre-post learning gains for individual teachers in the 2015–2016 school year and highlights how teachers differ in terms of their students' learning gains. (The online supplemental materials include descriptions of hierarchical linear models (HLM)-based analyses that support Figure 3 and the other conclusions in this paragraph; see Methods Supplement.) Even when we employ different statistical models that control students' posttest scores for their grade levels (middle or high school), affluence of their schools (as measured by percent of students on free and reduced lunch), and prior knowledge (as measured by pretest scores), we find this variation across teachers. More importantly, besides students' pretest scores, the teacher is the most significant factor affecting student learning.

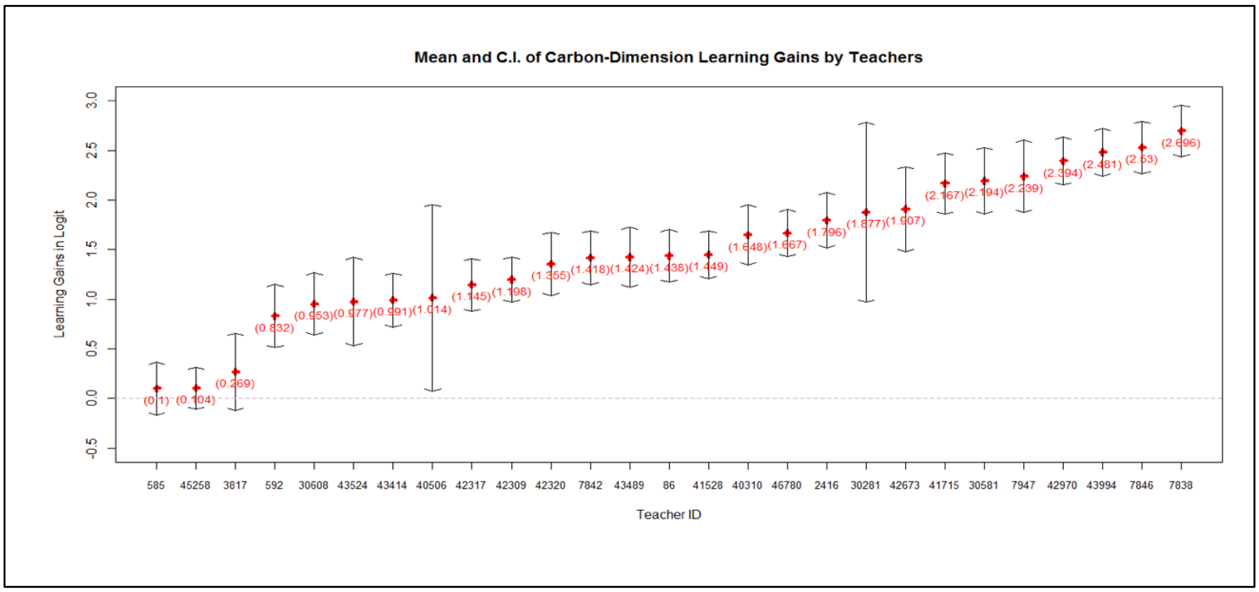

Figure 3: Comparing learning gains for individual teachers for the 2015-6 school year. 0 logits indicates no difference in average pretest and posttest proficiencies. One learning progression level is a gain of about 1.7. logits.

Moreover, analysis of the 2015–2016 data shows important relationships between surveys of teachers' practices and student learning outcomes. We applied a two-level model[2] in which students' posttest achievement is predicted by their pretest scores and teachers' science teaching practices as measured by our survey instruments. The results indicate that teachers' *Carbon TIME* featured practices ($CTFT_j$) have shown a positive effect (with borderline statistical significance) on students' scores, and this effect is significantly greater for students who had low proficiencies on the pretest[3]. These results indicate that the students of teachers who reported more frequent use of key instructional practices supported by the *Carbon TIME* units showed overall higher learning gains, with significant benefits especially for students with lower pretest scores.

[2] See Model S1 and Data S1 in the supplemental materials. We excluded the percent of free and reduced lunch and grade level in this model because these two variables did not add significant contributions to explaining the observed score variation.

[3] See Data S1 in the supplemental materials: the estimate for γ_{11} is negative and statistically significant.

Development dimension: A professional development course of study

Both classroom observations and student learning data support the design of a professional development course of study that addresses the purposes, obligations, and norms of *Carbon TIME* classroom communities. Our goal is to help teachers lead classroom communities that support responsive and rigorous discourse. The *Carbon TIME's* professional development program is a two-year course of study that includes both face-to-face and online components. In addressing the desired shifts in practice, the course works to explicitly discuss the realities of teachers' current classroom communities while also providing rationales, modeling, and support for what classroom communities engaged in three-dimensional science learning can look and sound like. We describe key elements of this course of study in terms of classroom social resources: purposes, obligations, and norms of successful classroom communities.

Purposes. The course of study begins with initial online work that orients teachers to the significant learning goals of sophisticated science reasoning around carbon-transforming processes (Parker et al., 2015). The collaborative development of a vision and purpose for rigorous and responsive teaching is further supported through continuing online and face-to-face learning opportunities. During the course of study, teachers talk through unit storylines, examine student learning goals and work samples, and discuss exemplars of productive classroom discourse, all in the service of developing their own classroom communities around a shared purpose of three-dimensional learning and figuring out phenomena.

As teachers are developing their understanding of the *NGSS* vision for student learning, they are simultaneously learning to use *Carbon TIME*'s instructional materials. Using the instructional model diagram as a reference, teachers first experience lessons as learners before reflecting about these experiences as practitioners. This structure affords teachers repeated opportunities to gain familiarity with the student storyline (students as questioners, investigators, and explainers) and coordinating classroom discourse routines using the Process Tools and other instructional strategies, such as talk moves (Michaels & O'Connor, 2012).

Obligations. Our current course of study focuses on helping teachers support the purposes of three-dimensional learning while recognizing that the traditional obligations of science teaching do not go away. For example, teachers still need to manage time, materials, and students; they still need to evaluate students and give grades. Thus, the obligations of participants (both teachers and students) in these three-dimensional classroom communities are negotiated and codified through assessing and grading practices. For example, in each Carbon TIME unit, students are held accountable for having, sharing, and returning to their ideas and earlier practices across the unit. Teacher-facing guides share examples of student thinking at different points in the unit and provide guidance on when students should be expected to use more sophisticated reasoning and thus be held accountable for their performance.

Norms. Productive classroom discourse requires communication and norms in classroom communities of practice that support three-dimensional science learning. The course of study includes significant attention to Carbon TIME classroom discourse routines. These routines outline important steps for both the teacher and students around each Carbon TIME Process Tool. Through these routines, the teacher establishes a purpose for the intellectual work, students engage in both private and public writing, and students participate in classroom talk that includes both divergent (sharing ideas) and convergent (reaching consensus) components.

For many teachers, these routines constitute a substantial change from traditional task-based discourse, in which students learn about science through following procedural directions and practicing correct answers. Using online and face-to-face opportunities, the course of study allows teachers to examine classroom exemplars and discuss and question their observations with colleagues and professional development leaders. Later in their course, teachers analyze discourse in their own classrooms and reflect on their practice with network colleagues. As teachers experience the *Carbon TIME* course of study (and as they leave the project after their two-year commitment), we notice that some of their project experiences diffuse into what were originally non-*Carbon TIME* communities of practice. This diffusion is evidenced by the growing numbers of schools and districts adopting *Carbon*

TIME materials as well as project teachers who are applying ideas gained from their project work in other content areas (e.g., genetics and evolution).

Goal 4: Supporting School and Professional Communities of Practice

Many of the teachers who chose to participate in our project are the only teachers at their school implementing the curriculum. As members of multiple and overlapping communities of practice, teachers must navigate among different communities that often have competing purposes, norms, and obligations. We recognize that what we are asking them to do can differ from their school communities and professional networks and are therefore interested in understanding how teachers move among three communities:

- *The classroom*: Teachers spend several hours a day with students in classrooms. In these communities, teachers are obligated to plan and deliver instruction, manage the social environment of their classrooms, and grade students' performances fairly.
- *The school*: Teachers spend several hours a week working with other teachers and administrators in school communities that give them access to material, human, and social resources. These communities invoke both informal norms and formal obligations such as those instantiated in teacher evaluation systems.
- *Professional networks*: Teachers spend a few hours a month engaging in professional development with staff and colleagues both in their *Carbon TIME* and non-*Carbon TIME* networks.

One question we had, then, was how teachers would manage their social obligations given that the purposes of *Carbon TIME* may differ from their usual classroom and school norms. We hypothesized that implementing *Carbon TIME* units would disrupt teachers' usual practices and routines, with an opportunity to shift towards three-dimensional science teaching and learning. Additionally, we were hopeful that the social relationships teachers developed in their *Carbon TIME* networks with staff and project colleagues would support changes in their teaching practices.

Research dimension: Understanding teachers' sensemaking in classrooms, school, and professional communities.

Our analyses of case study and survey data have focused in particular on teacher interviews as a way to access teachers' sensemaking. We conducted five in-depth, semi-structured interviews with case study teachers over the course of one year of implementation. Our analyses of these interview data have focused on teachers' *sensemaking* (the issues that they critically noticed and how they reasoned about them) and *agency* (how they felt empowered or constrained) in their classrooms and local professional networks to enact change. The teachers we work with have a lot on their minds besides student learning and classroom discourse: teacher evaluations, common assessments, district initiatives, etc. Some teachers manage to navigate these many demands by seeing connections. Many teachers, though, see them all as different and competing for time and attention.

Using the concept of sensemaking from organizational theory (Weick, 1995; Weick, 2001; Weick, Sutcliffe, & Obstfeld, 2005), we examined teachers' sensemaking about implementation of *Carbon TIME* (de los Santos, 2017). Spillane, Reiser, and Reimer (2002) argued that sensemaking is a crucial dimension of implementation. We defined sensemaking in this project as *critical noticing that leads to action situated in context over time* (de los Santos, 2017), privileging the classroom enactment setting as the primary focus of teachers' time and therefore sensemaking. One affordance of sensemaking is the attention given to examining how teachers' social commitments to various communities, including their obligations to school communities, influence their teaching practices. We defined productive sensemaking as sensemaking that helped teachers make progress towards rigorous and responsive science teaching practices whereas unproductive sensemaking did not. We illustrate this contrast between productive and unproductive sensemaking with examples that show differences in how teachers managed their social commitments and the resulting outcomes for their learning of new practices.

Mr. Ross and Ms. Callahan (pseudonyms) were high school biology teachers with 8 and 13 years of teaching experience, respectively. Both teachers were faced with the dilemma of how to manage competing social obligations among their classroom, school, and project network communities. A theme from Mr. Ross's case was the idea of "trying to fit it all together" because he felt that he had to meet the demands of all his communities (see Figure 4). Mr. Ross was trying to fit *Carbon TIME* with other school and district initiatives, such as using the Claims-Evidence-Reasoning framework for argumentation, using Accountable Talk for productive classroom discourse, and meeting International Baccalaureate program requirements. Mr. Ross's sensemaking often focused on how to use the project to meet his other obligations. His attempts were met with mixed success. For example, he tried using *Carbon TIME* assessments to measure student growth for his teacher evaluation. This was a reasonable action to take, but, in developing an elaborate grading scheme, Mr. Ross spent more time engaged in sensemaking about how to grade students' performance than noticing the nature of students' ideas and how he could support their development.

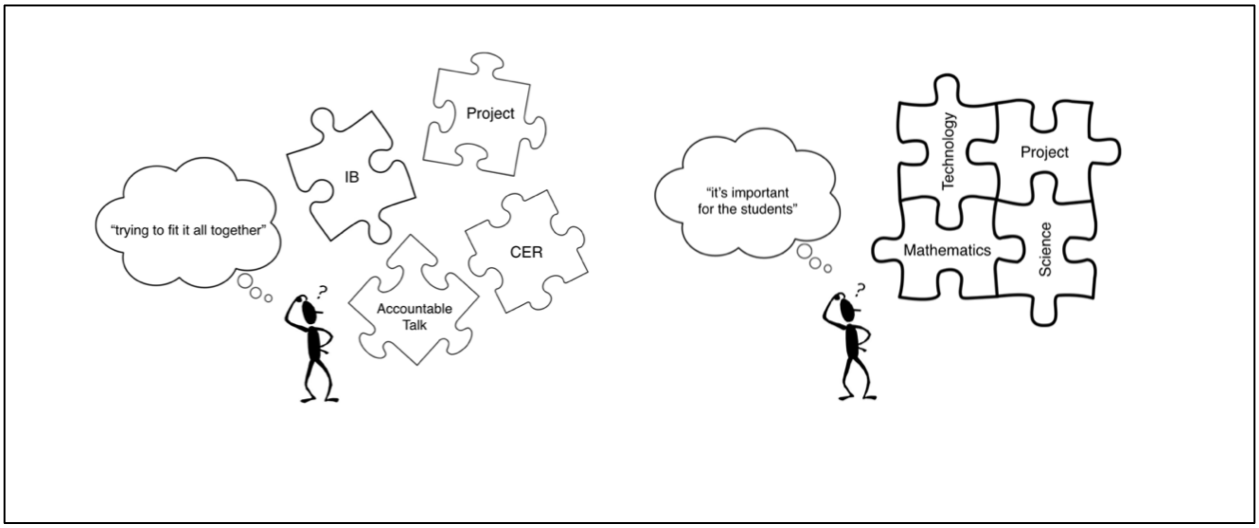

Figure 4: Mr. Ross's and Ms. Callahan's responses to Carbon TIME in the context of their school professional communities

In contrast, Ms. Callahan was in a magnet school context that focused on collaboration between the mathematics, science, and technology teachers to support their high school students in developing advanced research skills. A theme from Ms. Callahan's case was the idea that "it's important for the students" to develop particular research skills and content knowledge. Ms. Callahan's sensemaking was productive because she critically noticed features of *Carbon TIME* that aligned with her school context, which helped her meet her obligations to her school colleagues by conforming to local norms while also focusing on students' development. For example, Ms. Callahan critically noticed the precision of students' language (e.g., matter versus atoms and molecules) and worked to support students' conceptual understanding. Being a part of the *Carbon TIME* community helped shift her teaching practices from more traditional discourse patterns (e.g., Initiation-Response-Evaluation) to more responsive patterns by providing her with particular strategies and differing viewpoints about her teaching practices.

Generally, we found that the degree of fragmentation versus coherence in school communities and professional networks was associated with teachers' actions in their classrooms. We note that, although all teachers exhibited a range of human resources, the teachers who seemed to engage in the most productive sensemaking had strong visions of who they were as teachers and what they wanted their students to accomplish in the classroom, and these visions aligned with those of the curriculum. We found that the productivity of teachers' sensemaking varied depending on their purposes for classroom science teaching. All teachers modified *Carbon TIME* tools through their sensemaking, but some teachers were more focused on having students complete tasks, with less critical noticing of how students were interacting with the curriculum. As the case of Ms. Callahan illustrated, teachers in a school community with a more coherent vision that aligned with the project and *NGSS* seemed to provide teachers with the space to try new teaching practices.

Case study data provided us with a detailed look inside classrooms and how teachers' social commitments to people in their various communities influenced their practices. In contrast, our quantitative social network data provided us with a broad overview of how all the teachers participating in our project were connected to each other and to others outside the project. Visualizations of network structures based on 2015-2016 survey data (see Penuel et al., in press) show that patterns of interaction are stronger when teachers' *Carbon TIME* communities were also their professional communities. In one of the project networks, all of the teachers came from the same large district, and we found that there were more close relationships among teachers and between teachers and the network leader (who was also the science curriculum leader in the district) than in other networks. More data is needed to determine whether this difference in network composition is associated with significant variation in student learning. However, these results point to the important interactions between project networks and enduring local professional networks and communities.

Development dimension: Developing professional networks and partnerships

Variability in teachers' sensemaking and success with implementation indicate a need to design education systems that support innovative teaching practices. Our work shows that tensions among the purposes, norms, and obligations of different communities of practice can be difficult for teachers to resolve, particularly by themselves with little guidance from either curriculum developers or their schools and districts. By developing a research-based curriculum and then involving teachers in large-scale implementation, we can pinpoint specific challenges that teachers faced, including resolving tensions among different communities of practice, shifting teaching practices towards more reform-oriented visions, and navigating social obligations to students and school colleagues. We envision working in partnership with schools and districts in the future to discuss these challenges and how we can work together to support teachers and students. We suggest that reform visions of science education cannot be achieved unless we can work with schools and districts to develop compatible, rather than competitive, social resources for teachers to access through their professional networks.

Discussion: Contributions of this Study

Our assessment results and case studies show that this project has modified educational systems in ways that make a difference. Students show large learning gains on three-dimensional assessments compared with baseline data from the same teachers using other curricula. Case studies show significant changes in classroom discourse. Teachers testify to the importance of what they have learned through participation in the project. Still, most students fall short of fully accomplishing *NGSS* performance expectations, many classrooms are still mostly task-driven, and many teachers are still struggling to navigate among competing norms and obligations.

We could describe the individual challenges and responses above in separate publications. However, we argue that these are not separate problems: each challenge and response provides information that makes research on the others more meaningful. Successful design requires attention to *systems*, not just components of systems. *NGSS* gives us coherent goals for science teaching. We believe science educators can achieve those goals but only if we respond by designing coherent educational systems in partnership with local communities.

We have argued that, as design-based implementation research, this project documents a program to effect systemic change. The project is itself an educational system, but it is not the one that endures in the long run. Enduring change must come in the communities of practice that are at the core of our science education system: classrooms, schools, and professional communities. We have hopefully helped to develop an integrated tool kit that committed teachers, schools, and districts can use as they transform themselves to support responsive and rigorous teaching and three-dimensional science learning. We conclude with brief comments on four issues that we have encountered and that will affect other programs that aspire to enact *NGSS* at scale: *priorities and division of labor, collaborative development, evidence-based development,* and *research and development at scale.*

Priorities and Division of Labor

The order of the goals in this paper is akin to the order of temporal development for this project: We started by developing learning progression frameworks and assessments, added curriculum materials, then added professional development to support curriculum implementation, and finally added work with understanding teachers' classroom, school, and professional communities of practice. Unsurprisingly, there have been more cycles of research and development for the earlier goals. Other DBIR projects, especially those built around research-practice partnerships (Cobern & Penuel, 2016; Penuel & Gallagher, 2017), have tended to approach these goals in a different order: for example, starting first with attention to school and professional communities of practice (social resources) and then developing material and human resources based on work with specific schools and districts.

We do not believe there is necessarily one best order; every project has to start somewhere, and it is usually not possible to start by paying equal attention to all four goals at once. However, we do believe that work on all the goals is necessary because each goal addresses an aspect of educational systems where change is essential if the goals of *NGSS* are to be achieved. We also feel that we have learned something about what kinds of *division of labor* between teachers and developers are practical and possible. For example, practitioners of other professions are not expected to invent their own tools before they can do their work. It is equally unrealistic to expect that professional communities of practice and a curriculum infrastructure will be sufficient; teachers need high-quality curricula and assessments, and they do not have the resources to develop them easily on their own. Thus, we suggest that developers take primary responsibility for developing curricula with support from teachers who volunteer their time and expertise. Teachers, in turn, have practical knowledge (van Driel, Beijaard, & Verloop, 2001) that can inform the design of curricula that will work in classrooms.

Collaborative Development

This project brings together participants from multiple communities of practice, including university researchers, classroom teachers, district administrators, and even teachers' unions. The combination of these diverse perspectives enhances the quality of both our research and development. As researchers, this project gives us an opportunity for deep explorations of how teachers negotiate the different norms and practices of the classroom and professional communities where they work and of how participation in those communities both constrains and enables successful professional learning. We conclude that this kind of collaboration among people who play multiple roles is essential for successful DBIR projects and for improving educational systems at scale. Along with this collaboration there are important issues. For example: (1) investment in different types of resources: What material, human, and social resources do teachers need? and, (2) division of labor: What roles do we need in development and enactment, and what qualifications are essential for the people who play those roles?

From our research, we know that teachers need resources to effect meaningful change in classroom discourse. However, different teachers may need different resources, and one challenge for us is how to determine which teachers need which resources and how we can best support them in accessing those resources. In a sense, we are arguing for differentiated professional development that targets teachers' specific professional learning needs.

Evidence-based Development

"Evidence-based" is sometimes used to describe narrowly defined practices that privilege student testing data at the expense of other kinds of evidence. We argue that development of educational systems needs to be evidence-based in a deeper sense. We need evidence about student achievement—specifically, about achievement related to *NGSS* three-dimensional performance expectations. Beyond that, we need evidence about key communities of practice that affect student achievement, including classrooms and school communities. Furthermore, it is the *combination* of data sources and analyses that enables us to analyze and improve relationships among the practices in different communities and their consequences for student learning. However, this still leaves us with an important question: What kinds of empirical data are essential for development, formative assessment, and summative assessment that will build material, human, and social resources for three-dimensional learning?

Research and Development at Scale

The problems of enacting *NGSS* are inherently problems of scale. Design-based research in a few classrooms can produce examples of virtuoso teaching and learning, but standards-based reform demands that we consider how larger educational systems can and should change. Thus, we need to consider questions about the appropriate scale for development and enactment: What is the right balance between approaches to reform that are “too big to succeed” because resources are spread too thin to have substantial effects and fragmentation in which every teacher is an independent actor choosing from a smorgasbord of individual lessons and units? This research addresses issues of scale in three different aspects: (1) scale in numbers of participating teachers and students, as described in Figure 1 and in Table 1; (2) scale in time of participation, with each teacher committing to two years of involvement; and, (3) scale in curricular impact. We are studying consequential learning: how students develop proficiency in practices that they will need in their personal lives and as informed citizens. It takes time for students to master these practices, so teachers must commit commensurate instructional time to supporting students in developing them. As our research and development program shows, teachers need material resources in the form of high-quality curricula aligned with *NGSS*, human resources in the form of long-term professional development that seeks to support teacher learning of rigorous and responsive science teaching practices, and social resources in the form of coordination among norms and obligations to different communities of practice. Although we have discussed these resources separately, we want to emphasize that they should be in alignment and integrated in order to effectively support the development of coherent educational systems.

We began this paper with a challenge: enacting the aspirational goals of *NGSS* in educational systems that are fragmented and resource-poor. This challenge is worth addressing because these are goals we believe in. The *Carbon TIME* project provides evidence that this challenge is not insurmountable: It is possible to measure and achieve three-dimensional learning at scale. Achieving these possibilities, however, will require substantial investments in the material, human, and social resources of educational communities of practice.

References

Achieve, Inc. (2016). Using phenomena in NGSS-designed lessons and units. Retrieved from http://www.nextgenscience.org/sites/default/files/Using%20Phenomena%20in%20NGSS.pdf

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. Science Education, 95, 639-669.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. Educational Researcher, 32(1), 9–13.

Coburn, C. E., & Penuel, W. R. (2016). Research-practice partnerships in education: Outcomes, dynamics, and open questions *Educational Researcher, 45*(1), 48-54.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 453-494). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Covitt, B. A., & Anderson C. W. (in press, 2017). 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. London: Routledge (Taylor & Francis).

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.

de los Santos, X. (2017). *Teachers' Sensemaking about Implementation of an Innovative Science Curriculum Across the Settings of Professional Development and Classroom Enactment* (Doctoral dissertation, Michigan State University).

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

Frank, K. A. (1998). Quantitative methods for studying social context in multilevels and through interpersonal relations. Review of Research in Education, 23, 171-216.

Frank, K. A., Kim, C., & Belman, D. (2010). Utility theory, social networks, and teacher decision making. In A. J. Daly (Ed.), Social network theory and educational change (pp. 232-242). Cambridge, MA: Harvard University Press.

Henrick, E., Munoz, M. A., & Cobb, P. (2016). A better research practice partnership. Phi Delta Kappan, 98(3), 23-27.

Hiebert, J., & Stigler, R. (2017). Teaching versus teachers as a lever for change: Comparing a Japanese and a U.S. perspective on improving instruction. Educational Researcher, 46(4), 169–176.

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

Johnson, W. R, Miller, H. K., & Anderson, C. W. (2017). Curiosity and Principles in *Carbon TIME* classrooms. Poster presented at the annual meeting of the National Association for Research in Science Teaching, San Antonio.

Lee, O., Llosa, L., Jiang, F., O'Connor, C., & Haas, A. (2016). School resources in teaching science to diverse student groups: An intervention's effect on elementary teachers' perceptions. Journal of Science Teacher Education, 27(7), 769-794.

Liew, C. W., & Treagust, D. F. (1995). A predict-observe-explain teaching sequence for learning about students' understanding of heat and expansion of liquids. Australian Science Teachers' Journal, 41(1), 68-71.

Michaels, S., & O'Connor, C. (2012). Talk science primer. Cambridge, MA: TERC.

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.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academies Press.

National Research Council. (2005). Systems for State Science Assessment. Washington D.C.: National Academies Press.

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

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academies Press.

National Research Council. (2014). Developing assessments for the Next Generation Science Standards. Washington D.C.: National Academies Press.

Penuel, W. R., & Fishman, B. J. (2012). Large-scale intervention research we can use. *Journal of Research in Science Teaching, 49*(3), 281-304.

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., & Gallagher, D. (2017). *Creating research-practice partnerships in education*. Cambridge, MA: Harvard Education Press.

Pinker, S. (2007). The stuff of thought: Language as a window into human nature. New York, NY: Viking.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., ... & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. The journal of the learning sciences, 13(3), 337-386.

Rice, J., Doherty, J. H., and Anderson, C. W. (2014). Principles, first and foremost: A tool for understanding biological processes. *Journal of College Science Teaching, 43(2),* 78-86*.*

Roschelle, J., Knudsen, J., & Hegedus, S. (2010). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In Designs for learning environments of the future (pp. 233-262). New York, NY: Springer.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional science, 18(2), 119-144.

Spillane, J. P., Diamond, J. B., Walker, L. J., Halverson, R., & Jita, L. (2001). Urban school leadership for elementary science instruction: Identifying and activating resources in an undervalued school subject. Journal of Research in Science Teaching, 38(8), 918-940.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. Review of Educational Research, 72(3), 387-431.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology. Social Studies of Science, 19(3), 387-420.

Stroupe, D. (2014). Examining classroom science practice communities: How teachers and students negotiate epistemic agency and learn science-as-practice. Science Education, 98, 487-516.

Talmy, L. (1988). Force Dynamics in Language and Cognition. Cognitive Science, 12(1), 49–100.

van Driel, J. H., Beijaard, D., Verloop, N. (2001). Professional development and reform in science education: The role of teachers' practical knowledge. Journal of Research in Science Teaching, 38(2), 137-158.

Weick, K. E. (1995). Sensemaking in organizations. Thousand Oaks, CA: SAGE Publications.

Weick, K. E. (2001). Making sense of the organization. Malden, MA: Blackwell.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. Organization Science, 16(4), 409-421.

Wenger, E. (1998). Communities of practice: Learning, meaning, and identity. New York: Cambridge University Press.

Windschitl, M. (2004). Folk theories of "inquiry": How preservice teachers reproduce the discourse and practices of an atheoretical scientific method. Journal of Research in Science Teaching, 41(5), 481-512.

Yin, R. K. (2014). Case study research: Design and methods. Thousand Oaks, CA: SAGE.

Supplementary Materials: Mapping Supplement, Materials Supplement, Model S1, Data S1, Methods Supplement.