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

By Charles W. Anderson[1], Elizabeth X. de los Santos[2], Sarah Bodbyl[1], Beth Covitt[3], Kirsten D. Edwards[1], James Brian Hancock, II[1], Qinyun Lin[1], Christie Morrison Thomas[1], William Penuel[4], MaryMargaret Welch[5]

[1] Michigan State University

[2] University of Nevada, Reno

[3] SpectrUM Center, University of Montana

[4] University of Colorado, Boulder

[5] Seattle Public Schools

Table of Contents

Abstract

This article reports on a design-based implementation research (DBIR) project that addresses the question: *How can classrooms be supported at scale to achieve the three-dimensional learning goals of the Next Generation Science Standards?* Inherent in this question are three key design challenges: (a) *three-dimensional learning*—the multidimensional changes in curriculum, assessment, and instruction required for three-dimensional learning, (b) *scale*—the necessity of change at multiple scales in educational systems, and (c) *diversity*—achieving rigor in our expectations with responsiveness to the enduring diversity of our students, classrooms, and schools. We discuss findings from the *Carbon TIME* project, which focuses on teaching carbon cycling and energy transformations at multiple scales. Findings focus on design and knowledge building in three interconnected contexts. (1) *Assessment*—understanding and assessing students' three-dimensional learning. Learning progression frameworks provide insight into students' reasoning and the basis for efficient and reliable classroom and large-scale assessments that have used automated scoring of constructed responses for over 80,000 tests. (2) *Classrooms*—classroom discourse and learning communities. Six *Carbon TIME* units are based on an instructional model that scaffolds students' engagement with phenomena as questioners, investigators, and explainers. The units support substantial learning and reduce the achievement gap between high-pretest and low-pretest students, but with substantial differences among teachers. (3) *Professional communities*—a professional development course of study and research-practice partnerships address issues of organizational resources, conflicting norms and obligations, and building practical knowledge in schools and districts. Project results show continuing advantages for schools with more organizational resources. Overall, results 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

The authors acknowledge the contributions of Jay Thomas of ACT, Inc., prime developer of the *Carbon TIME* automated scoring system, of Ken Frank, Michigan State University, and of Karen Draney, Shruti Bathia, and Jinho Kim of the Berkeley Evaluation and Assessment Research Center, who conducted key analyses of quantitative data.

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*

The research question guiding this article focuses on a design challenge: *How can classrooms be supported at scale to achieve the three-dimensional learning goals of the Next Generation Science Standards (NGSS)?* This question has a design dimension—creating educational systems to accomplish this goal—and a knowledge-building dimension—improving our understanding of learning and teaching at scale.

We begin with a problem statement that focuses on the goals and approaches of the research genre that this study exemplifies: design-based implementation research (DBIR), which involves both design work and knowledge building about learning and teaching at scale. We then present findings addressing *NGSS* performance expectations for carbon cycling and energy flow at the middle- and high-school levels. We discuss design of a curricular activity system and knowledge building across three contexts: (1) *assessment*—understanding and assessing students' three-dimensional learning; (2) *classrooms*—supporting classroom communities where teachers and students share the goal of scaffolding and assessing three-dimensional engagement with phenomena; and (3) *professional communities*—building partnerships that bring teachers, researchers, and school administrators together.

Problem Statement: Research to Support Three-dimensional Learning Goals

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, we note that the performance expectations of the *NGSS* describe goals for science learning as *three-dimensional engagement with phenomena:* Students should use crosscutting concepts and disciplinary core ideas as they engage in scientific practices focused on phenomena in the material world.

Theory of change and associated challenges. The theory of change that guides our knowledge building and design activities focuses on *classrooms* rather than individual teachers or students as the unit of analysis. A productive response to the research question will require changes in the curricular activity systems (Roschelle, Knudson, & Hegedus, 2010) of thousands of classrooms. We point in particular to three challenges associated with this design goal.

Challenge 1: Three-dimensional learning. Our design goal is guided by the integrated learning expectations set forth by the *Framework* and the *NGSS*: Teachers and students should share an understanding of the nature and purposes of their work together: to create classroom learning communities that *scaffold and assess students' three-dimensional engagement with phenomena.*

Our understanding of the nature of this challenge begins with our learning progression research (discussed below), which shows the many ways in which three-dimensional engagement with phenomena as defined by *NGSS* is not "natural" for students. They come to school with approaches to making sense of the world that are reasonable and often productive, but are not based on scientific practices, crosscutting concepts, and disciplinary core ideas. In many classrooms today, teachers and students respond with a kind of compromise: They create curricular activity systems organized around a performance for grade exchange (Doyle, 1984) focused on one-dimensional performances such as remembering facts or practicing skills. In these classrooms *NGSS* performance expectations will require significant changes in classroom practice and student learning.

Challenge 2: Scale. Many studies in the tradition of design-based research document notable successes and admirable practices in classroom teaching and learning—individual classrooms where students' three-dimensional engagement with phenomena is successfully scaffolded and assessed (e.g., Berland et al., 2016; Lehrer & Schauble, 2012; Manz, 2012). Replicating these successes in thousands of classrooms, though, remains an immense challenge. Educational systems work at different scales: individual students, classrooms, schools, school districts, and states. Work at each scale poses its own challenges, so this design goal requires coordinated effort across scales in educational systems.

Challenge 3: Diversity. Large-scale reform efforts sometimes attempt "one size fits all" programs in which fidelity of implementation is a design goal. This fails to account for the enduring diversity of American students, classrooms, and schools, which is one of the great strengths of our country. We have to find ways to be both rigorous in our expectations and responsive to that diversity. This requires work across a range of settings, with particular attention to students, classrooms, and schools that are underserved and resource-poor.

Currently American schools are not prepared to address these challenges. 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—lessons found on the Internet, workshops on different topics, etc. (Opfer, Kaufman, & Thompson, 2016). 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 clear.

Our design goal is to address this mismatch by creating "tool kits" for a curricular activity system that is adaptable to local systems and diverse classrooms. A curricular activity system includes components such as teacher guides, student guides, professional development designs, and formative and summative assessments that fit together to comprise a shared vision for teaching and learning (Roschelle, Knudson, & Hegedus, 2010). In mathematics, there have been powerful proof of concept studies demonstrating that the design of a system can support equity goals at scale, and that such systems can be adapted to local contexts through partnerships (Roschelle, Pierson et al., 2010; Roschelle, Shechtman et al., 2010). In science education, however, such systems have only begun to be developed and examined. This study represents one such attempt.

Design-based implementation research (DBIR) to accomplish three-dimensional learning at scale. DBIR is an approach to research that deliberately addresses issues of coordination and coherence across components and scales of organization in education systems (Dolle, Russell, Gomez, & Bryk, 2013; Penuel & Fishman, 2012; Jackson & Cobb, 2012). The argument for DBIR includes a critique of the assumptions that underlie current compartmentalized research on innovations in science education. We consider the accomplishments and limitations for design-based research in each of our three focus areas.

Learning and assessment: The NRC *Framework* and *NGSS* are based on learning research, especially in the articulation of their three-dimensional framework (practices, crosscutting concepts, and disciplinary core ideas) and in their learning progressions for specific topics (National Research Council, 2005; 2007; 2012). This research helps us to *understand* science learning, but a successful reform effort will also require efficient and reliable assessments that *measure* students' three-dimensional learning at multiple scales, both for classroom teachers and for large-scale research and development.

The National Research Council report, *Developing Assessments for the Next Generation Science Standards* (NRC, 2104) follows earlier NRC assessment reports (NRC, 2001; 2005) in calling for systems of science assessments at multiple scales that are explicitly based on

learning progressions or other models of student thinking. This article describes the challenges inherent in measuring three-dimensional learning, which requires both assessment tasks that engage students in three-dimensional performances and measurement models for analyzing and scoring those performances. In this article, we describe a response to this agenda through using learning progression frameworks as a basis for efficient and reliable classroom and large-scale assessments.

Classroom teaching and design-based research: Design-based research has a well-established tradition of developing innovative classroom practice and theory through the design and testing of innovations (Brown, 1992; Cobb, Confrey, Lehrer, Schauble, & DiSessa, 2003). We recognize the value of the insights design-based research provides into mechanisms of learning and instructional design, including existence proofs showing when and how teachers can make a difference.

Yet there's a problem. As Fishman & Krajcik (2003) point out, instructional programs that work in design-based research settings are difficult to enact in implementation contexts because small design teams are often resource-rich in ways that cannot be scaled up: teacher-researcher teams, extra planning time, exceptional teachers, etc. Strategies that were successfully enacted by teachers with direct support from researchers are much more difficult to carry out in large, fragmented, and resource-poor systems. Thus an essential feature of our research is the design and testing of curricular activity systems in larger numbers of classrooms where teachers are not working directly with researchers. Our system is what Cohen and Ball (1999) might refer to as "highly developed" because the team has sought to "creat[e] the organizational, social, and intellectual resources required to enact" a curricular intervention (p. 19).

Professional communities and large-scale experimental studies: Another approach to science education research emphasizes the importance of studies in which specifically defined innovations are tested in experimental and control groups of classrooms. These studies typically show large implementation differences among classrooms that lead to inconsistent learning outcomes among their students. Researchers' attempts to achieve uniform implementation in diverse systems are rarely successful and fundamentally flawed: They fail to recognize and account for the necessity of flexible adaptation to the enduring diversity of students and schools. In contrast, some projects using a curricular activity system approach have shown positive results for a wide variety of students (Roschelle et al., 2010).

Even so, such results are difficult to sustain over the long haul (Fishman et al., 2011). Therefore, it is also necessary to create collaborative processes for adapting systems to fit local circumstances. In this article, we describe an approach to working through *research-practice partnerships,* creating a system of research and development that can learn and that is focused as much on continuous improvement as it is on existence proofs for innovation (Peurach, 2016).

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." In this article, we show how designed systems, data, and analysis from the *Carbon TIME* project address the research question above. After an introduction to the project, we describe design work and research results in three contexts—assessment, classrooms, and professional communities—with special attention to how we have addressed the challenges of three-dimensional learning, scale, and diversity.

Introduction to the *Carbon TIME* Project

The *Carbon TIME* project (http://carbontime.bscs.org/) has been supported by a series of National Science Foundation (NSF) grants since 2005. The project began with the general goal of supporting *environmental science literacy:* preparing students to use scientific knowledge and practices in their decisions about environmental issues. Like other DBIR projects we have used an *iterative design cycle* in which (a) goals for student learning are formulated, (b) assessments and instructional systems are designed to achieve those goals, and (c) designed innovations are tested in school settings, producing data that can be analyzed to inform revision of goals and a new cycle (Penuel, 2015; Roy, Fueyo, & Vahey, 2017). Below, we provide a brief overview of pertinent aspects of the *Carbon TIME* project, methods, and results that serve as a context for this article.

Curriculum and assessment system. *Carbon TIME* focuses on 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. Online supplemental materials for this article include tables documenting the middle- and high-school *NGSS* performance expectations addressed in our assessments and teaching materials (see Mapping Supplement in Supplementary Materials). Broadly, our curriculum and assessments focus on:

- All eight science practices, organized into three clusters: (a) asking questions; (b) inquiry (planning and carrying out investigations, analyzing and interpreting data, engaging in argument from evidence); and (c) application (developing and using models, constructing explanations, designing solutions).[1]
- Three crosscutting concepts: (a) scale, proportion, and quantity; (b) systems and system models; and (c) 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).

We have developed six three-week-long teaching units, each including options for use at the middle- or high-school level (http://carbontime.bscs.org/). Four units focus on macroscopic scale systems: *Systems and Scale, Animals, Plants,* and *Decomposers.* Two units focus on large scale systems: *Ecosystems* and *Human Energy Systems* (which focuses on global carbon cycling). All of the units are accompanied by an online assessment system that provides teachers with partially scored responses while simultaneously enabling us to collect and analyze student achievement data at scale.

The development process exemplifies our focus on classrooms as learning communities. Our design process is based on work with teachers in classrooms, and we study classrooms as communities of practice where teachers and students have mutual obligations (Wenger, 1998) and where teachers, students, and curriculum materials combine to enact curricular activity systems.

Our design and research work also connect scales and components of educational systems. DBIR work at each scale in the system (individual students, teachers, classrooms,

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

professional networks) is connected with work at other scales. In particular, these connections are built into the iterative design cycle with data from systems at one scale informing design decisions at other scales (Jackson & Cobb, 2013; Zuiker, Piepgrass, & Evans, 2017).

Data collection. We present analyses of data from the first three years of our current five-year study; the full five years of the study will involve approximately 160 participating teachers working in diverse middle and high school classrooms, with each teacher and their students participating for two successive years (about 900 different classrooms total). The 94 schools participating to date include urban, suburban, and rural schools. There are 26 middle schools and 68 high schools. The percentage of students in a school receiving free and reduced lunch ranges from 3% to 99%, with a mean of 41%. The percentage of underrepresented minority students in the participating schools ranges from 0% to 100%, with a mean of 43%.

Table 1 outlines major project data sources and quantities of data collected in the first three project years; the main years of project data collection are Years 2–5. The results reported in later sections of this article come from different parts of these data sets, as described in the individual results sections.

Table 1. Data Sources for the *Carbon TIME* Project (First Three Years)

Data Source	Baseline Year (2014–5)	First Full Year (2015–6)	Second Full Year (2016–7)	Additional Data*
Full Data Set (120 participating middle and high school teachers in 2014-17)				
Participating teachers	17*	27	83	
Student tests (8/student)	2,920	21,058	60,371	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	

*Participating teachers in the baseline year implemented assessments with their students but did not implement *Carbon TIME* instruction.
** 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 Professional Development (PD) each year for each teacher.

Effects on student achievement. Although we present more data and details of analysis in later sections, we begin here by presenting results to support a basic claim concerning the overall accomplishments of the project: Teachers using *Carbon TIME* materials are enacting instruction that produces three-dimensional learning at scale. Figure 1 compares IRT-based estimates of student pretest and posttest proficiencies with end of school year baseline levels (students of the same teacher the year before) for the first two years of this study. (For detailed methods and results see the section on learning below, the online supplemental materials, and Doherty et al., 2015).

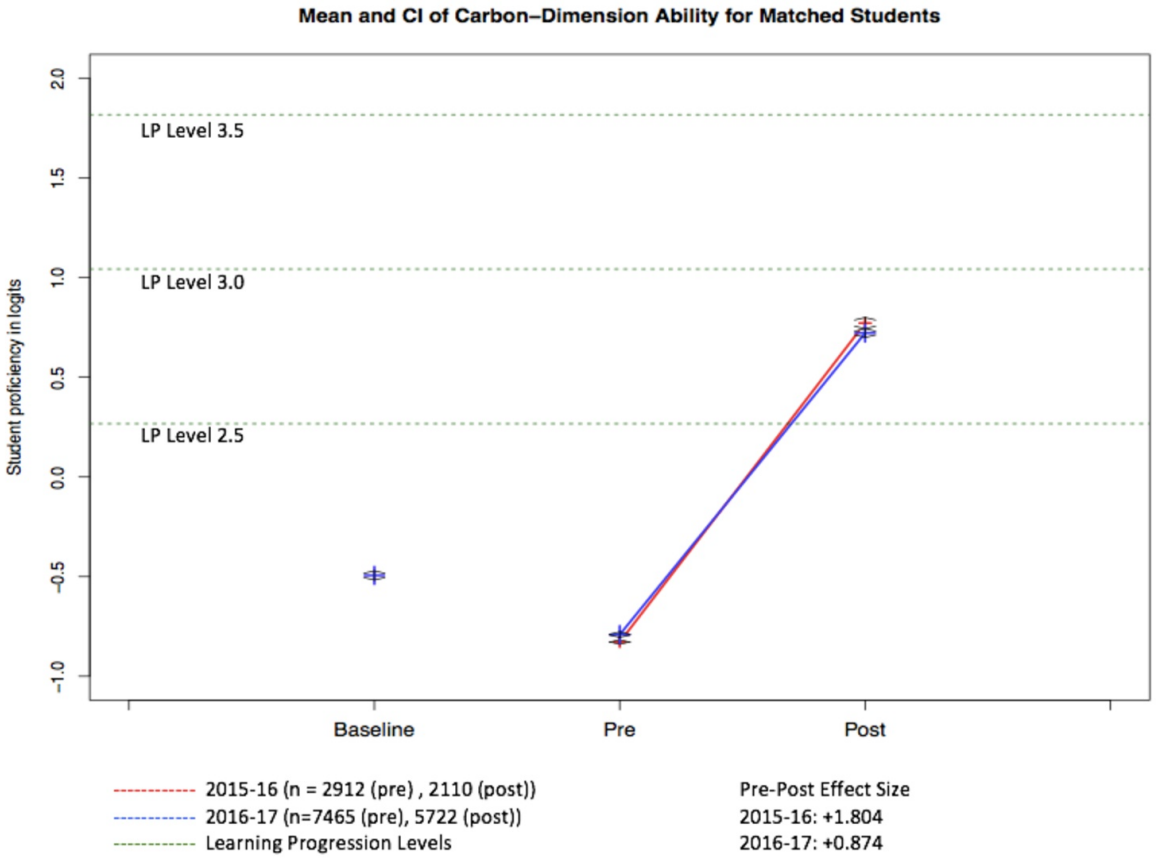

Figure 1: Mean learning progression (LP) levels of students in Carbon TIME and baseline (classes of participating teachers the year before they started using Carbon TIME). Error bars represent 95% confidence intervals. LP Level 4 is equivalent to full achievement of NGSS performance expectations in this domain.

These results can be interpreted from norm-referenced and criterion-referenced perspectives. Students improved significantly compared to both pretest and baseline performances. In other studies we have shown that high school students participating in *Carbon TIME* show higher proficiency on learning progression-based assessments than college science majors in biology courses (Rice, Doherty, & Anderson, 2014). However, from a criterion-referenced perspective, these results also show that most students still fall short of Level 4 of the learning progression, which is equivalent to the *NGSS* performance expectations in this domain.

The analyses leading to Figure 1 are complex and are discussed in more detail in the section on student learning and assessment below. We also discuss the large variations in enactment of the *Carbon TIME* curricular activity system in different classrooms. Nevertheless, we feel that these data provide an important warrant for our claim that, at scale, teachers using *Carbon TIME* materials are enacting instruction that produces three-dimensional learning. In the next three sections we discuss design products and research results at different scales.

Assessment: Understanding Students and Assessing Three-dimensional Learning

The National Research Council report, *Developing Assessments for the Next Generation Science Standards* (NRC, 2104), follows both previous reports and the *NGSS* document itself in advocating for *learning progressions* as measurement models (National Research Council, 2005; 2007; 2012). Unfortunately, empirically grounded learning progressions still are not available for most topics in the *NGSS* (Corcoran, et al., 2009; NRC, 2014).

The *Carbon TIME* project builds on more than a decade of learning progression research describing transformations in students' discourse and practice that are necessary to master *NGSS* performance expectations related to carbon cycling and energy flow at multiple scales. In this section, we describe how the project has carried out iterative cycles of development and refinement to produce (1) empirically validated learning progression frameworks that describe levels of proficiency as students develop scientific knowledge and practice with instructional support, and (2) validated written assessments that can be used to measure students' three-dimensional learning at the classroom scale and the network scale.

Understanding Three-dimensional Learning through Learning Progression Frameworks

Through more than a decade of iterative assessment cycles, we have developed three discourse-based learning progressions focused on interconnected practices of: (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 (Covitt & Anderson, 2018; Dauer, Doherty, Freed, & Anderson, 2014; Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009).

We describe our learning progression frameworks as *discourse-based* because increasing sophistication involves mastering new forms of talk and writing that represent changes in students' language and sense-making about phenomena. Successful learning in *Carbon TIME* enables students to participate in scientific model-based discourse (Covitt and Anderson, 2018). All three *Carbon TIME* learning progressions 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 (Gee, 1991; Pinker, 2007; Talmy, 1988).

Our learning progression frameworks are based in research utilizing both interviews with students and student responses to written questions. We have explored how middle- and high-school students explain and investigate a variety of phenomena related to carbon cycling and chemical changes in carbon compounds. Example phenomena include burning matches, growing oak trees, people exercising to lose weight, decaying apples, deer and wolf populations on an island, and the increase in atmospheric CO_2 concentrations recorded in the Keeling curve.

An example of learning. The learning progression frameworks are explained in detail in the publications cited above. Rather than trying to summarize the details we will follow the example of Fishman, Marx, Best, and Tal (2003) by drawing on one example from our learning progression research as a context for discussing research and design at different scales in the project. Our chosen example concerns how students can develop understanding of two related crosscutting concepts: *Energy and matter: Flows, cycles, and conservation* and *systems and system models,* focusing particularly on the hierarchy of systems at different scales.

Scientific explanations of the phenomena listed above rely on a *sense of necessity* associated with conservation laws and on connections among systems and subsystems at

different scales. For example, students who provide responses representing the upper level of the learning progression understand that the carbon in carbon dioxide that we breathe out *must* have come from somewhere in our bodies. Similarly, if a plant gains mass during an investigation then that mass *must* have come from somewhere outside the plant (Miller, Johnson, Freed, Doherty, & Anderson, submitted, 2017). Tracing carbon through systems requires connecting macroscopic-scale phenomena (such as people breathing or plants growing) with atomic-molecular scale models of chemical change (such as cellular respiration or photosynthesis).

In contrast, students who provide responses representative of the lower levels of the learning progression do not see a need to trace matter through systems or to follow conservation laws. They tend to rely on what Talmy (1988) and Pinker (2007) describe as *force-dynamic* reasoning. For example, most middle school students identify sunlight, air, water, and soil nutrients as enablers—things plants need to grow. But they do not explain that those enablers somehow *become* the plant, or that a change in the mass of the plant necessarily means that materials have moved into the plant from its environment. Similarly, they explain that our bodies convert oxygen to carbon dioxide when we breathe.

Developing Classroom and Large-scale Assessment Systems

While learning progression frameworks afford important insights into student reasoning, our research question requires both classroom assessments and large-scale or monitoring assessments that are valid, reliable, and efficient (Alonzo, Neidorf, & Anderson, 2012; NRC, 2005; 2014). Thus, we have devoted many research and development cycles to creating, testing, and improving *Carbon TIME* assessment systems.

Online testing system. The core of the *Carbon TIME* assessment system is an online testing platform that includes an overall test to be taken by students at the beginning and end of the school year as well as pretests and posttests for each of the six *Carbon TIME* units (http://carbontime.org/Index.php). Since the tests are capable of eliciting student responses across learning progression levels, the same tests are used for both middle- and high-school students. Teachers can download student responses in full test or spreadsheet format. In both formats, forced-choice portions of responses are automatically scored by the system. Anonymized responses are also shared with researchers. Details about how items are developed and scored are available in other reports (Doherty, Draney, Shin, Kim, & Anderson, 2015; Jin & Anderson, 2012).

Classroom formative assessment and grading. Recent scholarship on classroom assessment has generally focused on formative assessment (e.g., Covitt, Gunckel, Caplan, & Syswerda, 2018; Furtak, 2012; Furtak, Morrison, & Kroog; 2014; Gotwals & Birmingham, 2015). Teachers, however, are also legitimately concerned with grading and what Doyle (1983) refers to as “the performance for grade exchange” (de los Santos, et al., 2018). We have worked to design classroom assessment systems that serve both of these purposes, as well as the important purpose of helping students assess the quality of their own work. For example, teachers can print out questions and students’ responses (with forced-choice items scored) or download spreadsheets with responses for all students. Pretests are accompanied by Assessing Tools that help teachers to interpret students’ written responses in terms of learning progression levels. Posttests are accompanied by Grading Tools that suggest how teachers can award points for particular aspects of student responses..

Large-scale assessments with automated scoring. 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 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.

Development and testing of the *Carbon TIME* machine learning model for constructed response items has involved several steps, including construction 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 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 for human-machine reliability. Other items are revised or dropped from the assessment. This system has now been used successfully to score more than 850,000 student written explanations (Thomas, Kim, & Draney, 2018).

Validity evidence. The design and validation of learning progression-based assessments is an active area of research (DeBarger, Penuel, Harris, & Kennedy, C. A., 2016; Todd, Romine, & Cook Whitt, 2017; Yao & Guo, 2018). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) begin their discussion of validity by saying that validity is not a property of tests *per se.* Kane (2013) explains that, "it is the proposed score interpretations and uses that are validated and not the tests or the test scores" (p. 21). In our case we want to make (a) criterion-referenced claims that our assessments measure three-dimensional learning in terms of learning progression levels, and (b) norm-referenced claims comparing learning among classroom groups of students (see Figure 3 below), or for the program as a whole (see Figure 1 above).

We found that items combining forced-choice with an accompanying written explanation provided the evidence necessary to examine how students make sense of phenomena and what type of discourse they employ. Forced-choice only items provided insufficient evidence for these purposes. For example, students at lower learning progression levels often correctly choose "sunlight" in response to a multiple-choice question about where plants get their energy. But when the question is changed to a multiple true-false format, they also choose water and soil minerals as sources of energy for plants. Similarly, students' written explanations often reveal force-dynamic reasoning about plants' needs or conflate forms of matter and energy rather than tracing energy through plant systems. Our current item pool is available in the online supplemental materials.

We have also examined levels of consistency between how students talk about phenomena during in-depth clinical interviews and how they respond 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 detail by Doherty, Draney, Shin, Kim, & Anderson (2015). Analyses based on item response theory (IRT) and an explanation of how we used IRT analyses to estimate learning progression levels are included in the online supplemental materials.

Summary: Addressing the challenges of three-dimensional learning, scale, and diversity. Through multiple cycles of development and revision we have developed learning progression frameworks and assessment systems that provide the foundation for our DBIR research. The learning progression frameworks describe approaches that diverse middle- and high-school students take to investigating and explaining carbon-transforming processes. We have used those frameworks to develop formative and summative assessment systems that are used at classroom and program scales. We have developed validity evidence that these systems measure students' three-dimensional learning. During the 2016-17 school year, the *Carbon TIME* assessment system was used more than 8000 students in the classrooms of 83 teachers. The students took a total of more than 60,000 unit tests and overall tests. Our

automated scoring system assessed more than 500,000 student written explanations. This system is an important tool for classroom teachers, and it has provided essential data for our research and design work.

Classrooms: Supporting Three-dimensional Learning Communities

As described above, classrooms are the key unit of analysis in our theory of change, and our core design goal is to develop a curricular activity system supporting classroom discourse that *scaffolds and assesses three-dimensional engagement with phenomena.*

Our learning progression frameworks and assessments document how difficult this goal is to accomplish—what we label the three-dimensional learning challenge. On our pretests, the numbers of students with learning progression Level 4 responses (indicating successful three-dimensional engagement with phenomena) is vanishingly small. Instead, most students engage phenomena such as plant growth or animal movement in a sensible but entirely different way, describing plants and animals as actors and materials such as food, water, and air as enablers that they use to accomplish their life purposes, such as growth and movement. So the problem isn't just that students don't know *how* to produce three-dimensional explanations that trace matter and energy through systems; students don't realize that these explanations even exist!

So the challenge facing classroom teachers is not so much whether students' three-dimensional engagement with phenomena is desirable; the question is whether this is *possible.* With the curriculum materials being used in most classrooms today, the answer to this question is clear: No; without much better scaffolding students cannot use *NGSS* practices, crosscutting concepts, and disciplinary core ideas to investigate and explain phenomena. So most teachers respond to this dilemma in sensible ways that enable them to fulfill their obligations to students and administrators—by holding students accountable for less difficult performances. Research shows that three-dimensional science learning is rare in current secondary classrooms (NRC, 2007; 2015). More commonly, classroom teaching focuses on students learning one-dimensional facts and skills, with little attention to making sense of phenomena (Banilower et al., 2013; Furtak, Thompson, Braaten, & Windschitl, 2012; Roth & Garnier, 2007). These patterns of practice reflect longstanding traditions in American schooling (Cuban, 1993; Jackson, 1990; McNeil, 1998).

Recent work in design-based research has contributed to the development of alternate models of instruction that respond productively to the three-dimensional learning challenge (Lehrer & Schauble, 2012; McNeill & Knight, 2013; Windschitl, et al., 2012). Some key principles emerging from this work are summarized in the NRC report *Guide to Implementing the Next Generation Science Standards*: Classrooms should become learning communities where diverse students are engaged in three-dimensional practices, with support from instructional scaffolds, ongoing dialogic assessment, and classroom cultures that support productive disciplinary engagement (Engle, 2012; NRC, 2007; 2014; Reiser et al., 2017). Both this research and learning progression frameworks inform the design of *Carbon TIME* units.

Since we are engaged in DBIR, we also have to consider the scale and diversity challenges: How do we design curricular activity systems that support these learning communities in hundreds of diverse classrooms? For us this question implies some important design goals. We have to design curricular activity systems that are *robust* (students can still learn even if the plants die or the teacher hasn't mastered the nuances of three-dimensional discourse) and *educative* (the curricula help teachers learn as well as students; see Davis, Palincsar, Arias, Bismack, Marulis, & Iwashyna,, 2014).

In this section we first describe how we have developed *Carbon TIME* units as "tool kits" to support curricular activity systems in classrooms that enact three-dimensional learning. We then report on the varied success of teachers in using *Carbon TIME* tools and enacting three-dimensional learning. Finally, we discuss how our research on classroom teaching has addressed the core design challenges of three-dimensional learning, scale, and diversity.

Developing Curriculum Materials and Supports for Classroom Discourse

The *Carbon TIME* units, like the assessments, are the products of multiple iterative development cycles. We think of them as tool kits for teachers to use rather than as scripts that teachers must follow. In this section we describe the curricular activity system (Roschelle et al., 2010) that we envision and provide some examples of how *Carbon TIME* materials support enactment of that system.

Instructional model and storylines. Drawing from learning research 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 serves to organize each of the units. The instructional model incorporates two intertwined storylines: a *student storyline* with people (scientists, students) as protagonists and a *science content storyline* focusing on how systems (flames, animals, plants, decomposers, ecosystem, the Earth) transform matter and energy.

The *student storyline* focuses on what people do as they engage phenomena scientifically. An initial reading about a scientist who investigated systems that the unit focuses on is followed by a sequence of activities engaging students in three roles that encompass all eight science practices (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; 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, Brown, & Newman, 1989) as they develop explanations that trace transformations of matter and energy in living and Earth systems. Thus the student storyline provides what Engle et al. (2012) refer to as social framing by scaffolding students' engagement in different roles.

The *science content storyline* traces matter and energy as they are transformed in particular systems and carbon-transforming processes. For example, in the *Plants* unit, students learn how plants (1) create organic materials through the process of photosynthesis, then transform those materials (2) as they grow (biosynthesis) and (3) use chemical energy to function (cellular respiration).

Detailed enactment. The instructional model and storylines alone provide insufficient support for classroom instruction (Roschelle et al., 2010). For this reason, our units are content-specific tool kits that teachers can use based on knowledge of their students and requirements of their local context. 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.

Our learning progression findings informed our curriculum design. Students develop a sense of necessity for matter and energy conservation only after they have recognized these principles and applied them rigorously to multiple phenomena. Therefore a key scaffold in *Carbon TIME* units is the Three Questions (Figure 2). The Three Questions both provide a general framework for explaining carbon-transforming phenomena and help students focus on specific crosscutting concepts. The "Rules to Follow" support students in focusing on matter and

energy conservation, while “Evidence We Can Observe” helps students connect observations of phenomena at the macroscopic scale with explanatory models at the atomic-molecular scale.

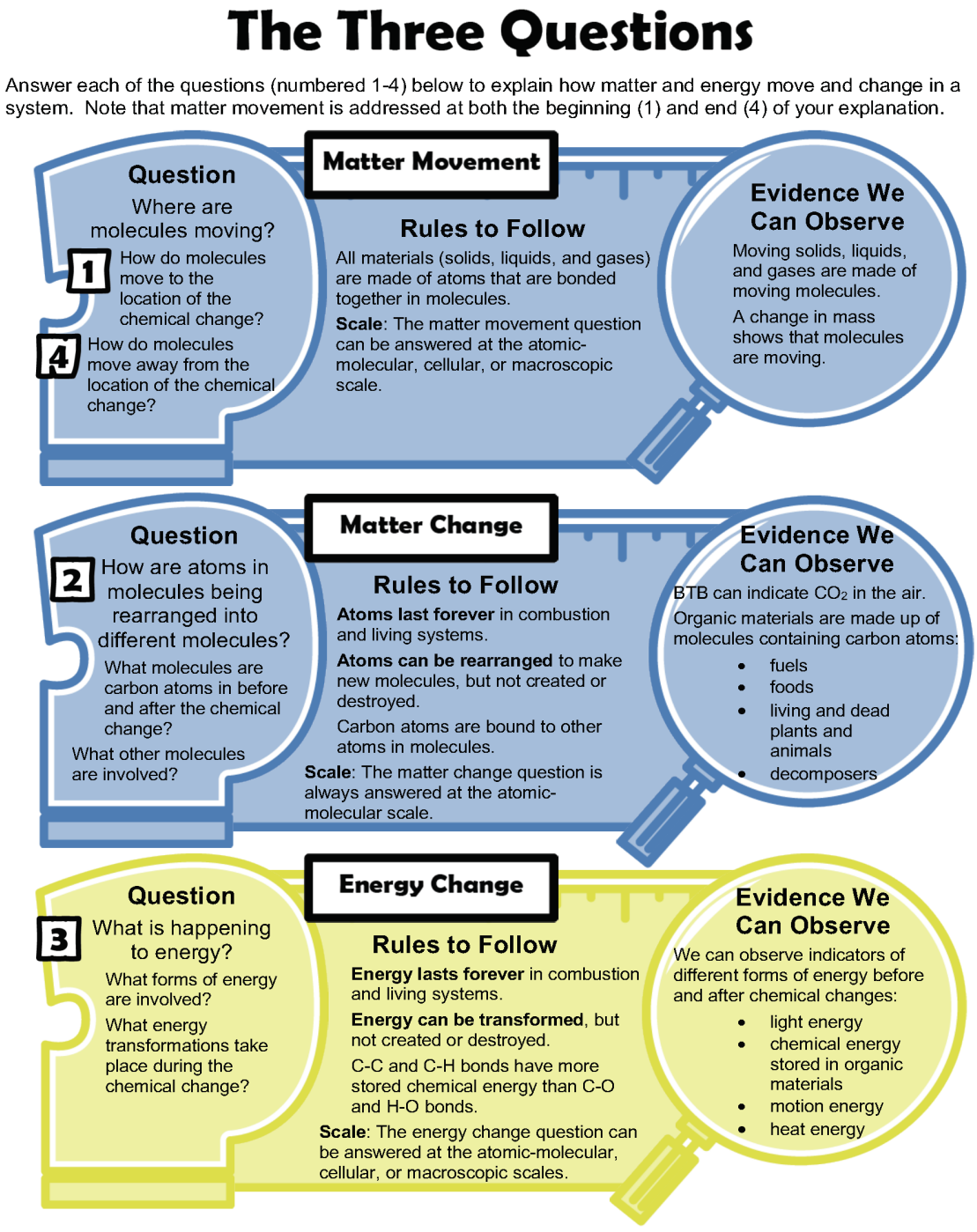

Figure 2: The Three Questions

The Three Questions provide a framework for other tools and activities in the curriculum to scaffold investigations and explanations. These different forms of scaffolding (described in the Instructional Model file in Supplemental Materials) include (a) Process Tools that help students organize their writing and thinking, (b) hands-on investigations and activities that model processes and chemical change, (c) simulations, videos, and other audiovisual supports, and

(d) discourse routines that engage students in sharing ideas and seeking consensus during small-group and whole-class discussions. This tool kit supports student engagement as well as formative and summative assessment (Sadler, 1989).

For example, the Three Questions (Figure 2 above) play different roles as students progress through the *Plants* unit. At the beginning of the unit the Three Questions are not mentioned at all as students express ideas and pose questions in whatever ways are natural to them. Students' investigations of growing radish plants are informed by the Three Questions: As they collect data and construct evidence-based argument, they are reminded that a good explanation of plant growth needs to answer the Three Questions. They identify both conclusions warranted by the evidence and unanswered questions (especially the Matter Change question) that are not fully answered by their investigations. For the concluding cognitive apprentice sequence, the Three Questions structure the Explanation Tool as well as rubrics that students and teachers use to evaluate the quality of their explanations.

Evidence from student work and student post-tests suggests that students who follow the rules of the Three Questions when completing their *Carbon TIME* process tools can subsequently provide post-test performances that trace matter through systems in highly sophisticated ways such as accounting for missing mass in an animal growth data set by explaining how that missing mass must have left the animal through cellular respiration and breathing (Edwards, Scott, & Anderson, 2018).

Studying Classrooms as Learning Communities and Curricular Activity Systems

Our design work, described above, focused on designing curricular tool kits for teachers that are robust and educative. Our research work focuses on how those tool kits are working in the diverse classrooms of teachers participating in the project—27 teachers in 2015-16 and 83 teachers in 2016-17. We observe complex classroom activity systems as teachers work to balance their continuing obligations (around curriculum, classroom management, student engagement, and grading) with new expectation of three-dimensional learning. We have studied these different activity systems in two ways: (a) through quantitative analyses using pretest and posttest data, and (b) through more detailed case studies of 17 classrooms with diverse teachers and students.

Quantitative analyses of pretest and posttest data. Our quantitative data confirm that teachers and classroom discourse patterns make a difference in students' learning outcomes (Lin, Kim, Holste, Bathia, Draney, & Frank, 2018). Figure 3 compares student pre-post learning gains for 58 individual teachers in the 2016–17 school year. (The online supplemental materials include descriptions of hierarchical linear models (HLM)-based analyses that support Figure 3 and the other conclusions below; see Methods Supplement.) The differences among classrooms are both statistically and educationally significant. Students gained an average of 1.41 logits (about 0.91 learning progression levels on a scale from Level 2 to Level 4). In the most successful classrooms learning gains were about twice the average; in the four least successful classrooms students did worse on the posttest than on the pretest. (See the supplemental materials for a detailed discussion of HLM analysis methods and checking data quality.)

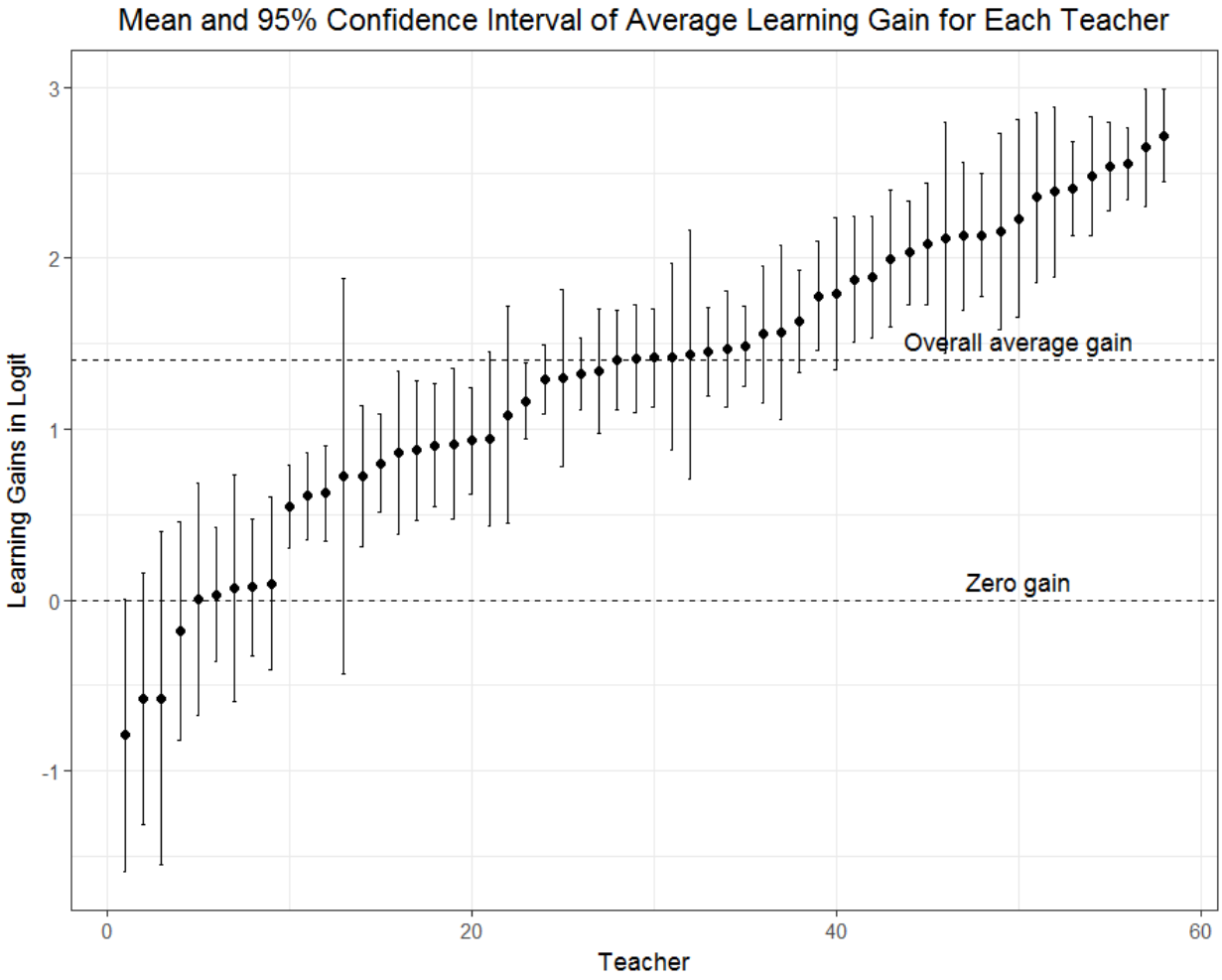

Figure 3: Comparing learning gains for macroscopic explanations (carbon dimension) for individual teachers for the 2016–17 school year. All teachers (N = 58) who had at least 20 students taking both pretest and posttest are included. Zero logits indicates no difference in average pretest and posttest proficiencies. One learning progression level is a gain of about 1.55 logits. The average learning gain for all students of these teachers was 1.41 logits.

Additional HLM analyses investigated associations between student learning gains and other variables associated with diversity in students and schools. We applied a two-level model in which students' learning gains are predicted by (a) how much individual students deviate from their class-average pretest proficiencies, (b) class-average pretest proficiencies and (c) the school's percent of free and reduced lunch. Separate analyses of 2015–16 and 2016–17 data show consistent patterns:

- *Carbon TIME reduced the achievement gap between higher-pretest and lower-pretest students within classrooms.* Within classes, students with lower pretest proficiencies showed significantly higher learning gains.
- *Carbon TIME was less successful in higher-poverty schools with fewer organizational resources.* The school percentage of free and reduced lunch was negatively associated with class-average learning gain. That is to say, classrooms from schools with higher percent of free and reduced lunch benefit less from implementing *Carbon TIME*. We discuss this finding in more detail below; we interpret it as evidence that schools with more organizational resources are more successful in implementing *Carbon TIME*.

Previous studies have shown the percent of free and reduced lunch can be a proxy measure for material, social, and human material resources such as students' access to qualified and experienced teachers (Darling-Hammond, 2004; Rice, 2010) and the overall quality of conditions in which teachers work (Johnson, Kraft & Papay, 2012).

- *Other variables were not significantly associated with student learning gains.* We also investigated other variables, including grade band (middle school vs. high school), racial composition of students, and class average pretests. None of these variables added significantly to the predictive power of models that included the three key variables above: individual teachers, student pretest, and school percentage of free and reduced lunch.

Qualitative analyses of case study data. Figure 3 shows that there were significant differences in how effective *Carbon TIME* units were in different classrooms. But the quantitative data do not reveal *how* or *why* the classrooms differed. To address these questions we turned to our case study data. We conducted case studies in 17 classrooms during 2015–16 and 2016–17, developing an extensive database, including videotaped classroom lessons, semi-structured interviews with teachers and students, and samples of student work. We chose case study teachers to represent a range of experience levels, grade levels, and school placements. (See Table 1 above for an inventory of case study data.) We have reported analyses to date in a series of dissertations and presentations (Covitt, et al., 2018; de los Santos, 2017; de los Santos, et al., 2018; Johnson, 2017). Some key patterns from these analyses are summarized below.

Observed classroom differences. The classroom video data from case study classrooms show teachers taking different approaches to using the *Carbon TIME* curriculum and assessment tools. The *Carbon TIME* teachers we judged as more successful voiced 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), both the teacher and students employed language that illustrated evidence of principled reasoning around core science ideas (Johnson et al., 2017), and students were active participants in using science practices and discussion to come to consensus around three-dimensional science performances (Covitt, et al., 2018). For these teachers, *Carbon TIME* instructional materials and assessments scaffolded students' engagement in productive classroom discourse and three-dimensional learning.

In less successful classrooms the purpose focused more on "learning about" science facts and skills (Johnson, 2017; see also Achieve, 2016; Schwarz, Passmore, & Reiser, 2017). In these classrooms *Carbon TIME* tools were treated as worksheets and incorporated into systems of task-based discourse that held students primarily accountable for one-dimensional practices: Students were considered successful if they could reproduce essential content but were not held accountable for three-dimensional understanding (Bloome, Puro, & Theodorou, 1989; Covitt, et al., 2018; Johnson, et al., 2017).

Why classrooms varied. Our observations and our interviews with teachers help us to understand some of the reasons for these differences. Classrooms are *communities of practice* that existed before the teachers started using *Carbon TIME* units and continued after. These classroom communities have shared *purposes, norms, and obligations* that shape the "culture" of each classroom. Thus, participation in a community of practice entails what Wenger (1998) refers to as "regimes of mutual accountability" (p. 81), or ways in which individuals are held socially accountable to each other—a shared sense of "how things are supposed to be" in the classroom (Michaels, O'Connor, & Resnick, 2008).

These classroom norms and obligations were the basis for observed differences between "good fit" classrooms and "misfit" classrooms. In the "good fit" classrooms, teachers'

and students' sense of how things are supposed to be—their shared purposes, norms, and obligations—aligned well with the *Carbon TIME* instructional model. They often modified *Carbon TIME* instructional materials in ways that enhanced their usefulness. For example, one high school teacher modified the phenomena in the *Animals Unit* from a boy growing to a panda growing in order to better engage her particular students' interest (de los Santos, 2017).

In "misfit" classrooms, on the other hand, teachers struggled to fit *Carbon TIME* materials into a classroom culture organized around different, and sometimes incompatible, purposes, norms, and obligations. These teachers often modified instructional materials in ways that made them less supportive of the intended science student and content storylines. For example, one teacher aligned her instructional practices with her school's goals by modifying the Evidence-based Arguments Tool to fit a Claims-Evidence-Reasoning (CER) framework. This was sensible, but in practice made it more difficult for students to notice and interpret key patterns in their data (de los Santos, 2017).

Summary: Addressing the challenges of three-dimensional learning, scale, and diversity. These and other findings from our classroom case study research have played a critical role in the iterative design cycle at all scales. They have led to changes in the instructional materials, as well as assessment design and research practice-partnerships. For example, in terms of assessment design, we have learned from teachers that the "fairness" of an assessment is particularly important—an issue we discuss in more depth in the section on research-practice partnerships below.

The findings from case study research have also affected our research-practice partnerships. Most of our network leaders were also case-study coaches, and their experiences visiting *Carbon TIME* classrooms have played a crucial role in developing the professional development (PD) course of study (described below). The case studies have enabled communication between classroom teachers and project leaders *based on what's really happening in classrooms.* It requires a substantial investment of time and resources to make this happen, but we are convinced that the investment is essential for developing an effective answer to our research question.

Our data show progress in addressing the core DBIR challenges, as well as important ways in which we continue to fall short of our goals. There was both productive and unproductive diversity in the ways that teachers used the *Carbon TIME* curriculum and assessments. Some teachers tailored instruction to scaffold their students' three-dimensional engagement with phenomena; others used *Carbon TIME* tools in ways that reflected classroom norms more aligned with one-dimensional learning. In most classrooms, the *Carbon TIME* curricular activity system reduced the achievement gap between higher-pretest and lower-pretest students. But *Carbon TIME* was less successful in higher-poverty schools with fewer organizational resources.

Professional Communities: Working with Schools in Research-Practice Partnerships

The previous sections have described the multiple challenges associated with scaffolding and assessing three-dimensional learning at scale. In most classrooms the transition from one-dimensional to three-dimensional learning is long and difficult. Teachers require professional development and the support of their colleagues and administrators to make this transition completely.

The *Carbon TIME* project includes a professional development (PD) course of study that is conducted in networks formed around research-practice partnerships between the *Carbon TIME* project and a variety of local education agencies (LEAs). Over the five-year project about

160 teachers will participate in five networks in three different states, ranging in size from 17 to 58 teacher participants, with each teacher committing to two years of participation. The networks involve different LEA partners. One partner is a large urban school district. Partners for other networks include smaller school districts, a statewide teachers' union, and a network of rural schools connected to a scientific research field station. One online network does not have an LEA partner.

Understanding Professional Networks and Research-Practice Partnerships

Our case study data, as well as our experiences and data from face-to-face and online PD and other communications with teachers, have deepened our understanding of the challenges identified at the beginning of this article. These data, in addition to our reading of relevant research literature, led us to identify three core community-scale issues: organizational resources, professional norms and obligations, and practical knowledge.

Organizational resources. Professional networks and classroom communities that support three-dimensional learning require *material, human, and social resources* (Gamoran et al., 2003; 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 organizational cultures (norms, purposes, obligations) and relationships among individuals, which can support the distribution of material and human resources (Cohen, Raudenbush, & Ball, 2003). All of our LEA partners faced constraints on their material, human, and social resources; in some cases these constraints limited the abilities of teachers to use *Carbon TIME* tools effectively. As we reported above, our quantitative data suggest that classrooms in higher-poverty schools with fewer organizational resources tended to have lower learning gains.

Professional norms and obligations. Our case studies have led us to think more deeply about the obligations that science teachers take on when they agree to be teachers. We have looked both at scholarly literature (e.g., Jackson, 1990; Kennedy, 2016; Ingersoll, 2003; Spillane & Burch, 2006) and at teacher evaluation systems commonly used in schools (Center for Educational Leadership, 2016; Danielson, 2014). These sources all organize the work of teaching around a list of categories including curriculum, student engagement, management, assessment, and communication. Teachers', schools', and students' interpretations of these widely accepted norms of schooling and professional obligations of teaching can be in tension with the *Carbon TIME* goal of assessing and scaffolding three-dimensional science learning (cf., Allen & Penuel, 2015).

Practical knowledge. Our case studies provide evidence about the gap between the daily work of secondary science teachers in traditional classrooms and the principles outlined in the *Guide to Implementing the Next Generation Science Standards* (NRC, 2015). Most of the teachers we work with are professionals dedicated to improving their teaching, but the *NGSS* and accompanying reform documents do not translate their visions into what van Driel, Beijaard, and Verloop, (2001) describe as *practical knowledge:* the beliefs, formal knowledge, and experiential knowledge that teachers use to enact the day-to-day work of teaching. 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).

An example: grading student work fairly. Our interviews with teachers show that one driver sustaining one-dimensional is teachers' concerns about *fairness.* It isn't fair to hold students accountable for three-dimensional performances that they cannot do, but what is the

alternative? Consider a student who writes the following response to a question about plants use sunlight when they grow: "*Photosynthesis is when plants make sunlight into food."* Our learning progression framework would classify this as a lower-level explanation with an implicit matter-energy transformation. The Three Questions framework shows that this is not a satisfactory response because it conflates the Matter Change Question and the Energy Change Question. The *Carbon TIME* materials encourage teachers to ask questions like this for formative assessment purposes.

But that's not the end of the story. Doyle (1983) writes about "the performance for grade exchange" as a core aspect of classroom life. We are working with many teachers who are expected by their school districts to produce evidence that their students are learning. This leads to a pair of complex issues:

- When and how is it fair to ask this question—explaining the observed phenomenon that plants need light to grow?
- How is it fair to grade the student's response?

We have learned that these issues require attention. Some teachers, for example, experienced conflict when their course or district common assessments focused on one-dimensional fact-based questions, which are significantly different from our units' multidimensional assessment questions. We see this pattern repeatedly: Both informal norms and formal evaluation systems create obligations for teachers that are in tension with three-dimensional teaching and learning (de los Santos et al., 2018). Addressing these concerns productively requires attention to each of the issues described above. Teachers need (a) high-quality material resources to scaffold and assess students' three-dimensional performances, (b) help in asserting the importance of three-dimensional learning relative to other conflicting obligations, and (c) the practical knowledge to effectively use three-dimensional curriculum and assessment resources.

Developing Partnerships and Supports for Teacher Learning

Our response to these three core issues—organizational resources, professional norms and obligations, and practical knowledge—has two dimensions: a two-year PD course of study and the development of research-practice partnerships that include teachers, researchers, and school administrators. We briefly describe each below.

PD course of study. Both classroom observations and student learning data supported the design of a two-year course of study that includes twelve days total of face-to-face (six days) and online (six days) PD. We found this to be a frustratingly short time for the challenges involved, yet it was also more than most school districts could afford without grant support. The course responds to the realities of teachers' current classroom communities while providing rationales, modeling, and support for what classroom communities that scaffold and assess three-dimensional science learning can look and sound like.

The course of study addresses each of the community-level issues above—limited resources, conflicting norms and obligations, and practical knowledge—but its main focus is the third issue: practical knowledge. It was not just teachers who needed to learn. As researchers and developers we are also still struggling to develop the practical knowledge it takes to understand students and enact the *Carbon TIME* instructional model.

Therefore, our course of study needed to support joint learning by both researchers and teachers. Weick and others refer to this joint learning as *organizational sensemaking* (Weick, 1995; Weick, 2001; Weick, Sutcliffe, & Obstfeld, 2005; de los Santos, 2017). Spillane, Reiser, and Reimer (2002) argued that sensemaking is a crucial dimension of implementation. 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. A core goal of our PD was to engage teachers and PD leaders in productive sensemaking that helped teachers make progress towards rigorous and responsive science teaching practices.

Research-practice partnerships. The final component of our project involved development of research-practice partnerships to support sustained engagement by teachers, researchers, and school administrators. A key advantage of partnerships is that they provide a means for researchers and practitioners to work together to solve problems of implementation (Penuel & Gallagher, 2017). It is in the context of partnerships and networks conducting DBIR that disparate bodies of knowledge relevant to problems of scale—about student and teacher learning, organizational change, and scale—come together (Russell, Jackson, Krumm, & Frank, 2013). In such partnerships, as illustrated above, there is a two-way street between researchers and practitioners, such that researchers, teachers, and administrators play essential but complementary roles (Tseng, Easton, & Supplee, 2017). For example, grading rubrics developed by teachers provided a basis for improvements in the Three Questions and strategies for giving students credit for three-dimensional learning.

Summary: Addressing the challenges of three-dimensional learning, scale, and diversity. Our experiences with PD and networks reinforce for us the basic insight that coherent and ambitious reform will require changes in school resources and organizational norms, as well as enduring research-practice partnerships. The issues that we encounter at the scale of professional communities—organizational resources, professional norms and obligations, and practical knowledge—require partnerships with school administrators and instructional leaders as well as teachers. Then teachers can focus on developing practical knowledge while they rely on material, human, and social resources from their schools as well as from the *Carbon TIME* project. Administrators and teachers together can create professional norms and evaluation systems that value and reward three-dimensional science learning. It is harder for teachers to develop practical knowledge when their schools lack organizational resources or have conflicting norms and obligations.

Conclusion

We began this article with a research question: *How can classrooms be supported at scale to achieve the three-dimensional learning goals of NGSS?* Inherent in this question are three key design challenges: (a) *three-dimensional learning*—the multidimensional changes in curriculum, assessment, and instruction required for three-dimensional learning, (b) *scale*—the necessity of change at multiple scales in educational systems, and (c) *diversity*—we have to find ways to be both rigorous in our expectations and responsive to the enduring diversity of our students, classrooms, and schools.

We have advocated design-based implementation research—DBIR—as a research approach that addresses these issues, and in this article we report on the *Carbon TIME* project as an example of this approach. *Carbon TIME* shares with other DBIR projects a focus on both design of educational systems and knowledge building, as well as a commitment to some core design principles: (1) An *iterative design cycle* is necessary for researchers' and teachers' learning and in order to develop and revise all the components of a complex educational system. (2) A *focus on classrooms as learning communities* has kept our attention on helping teachers to do their jobs; the difference between this focus and a focus on implementing reforms is subtle, but critically important. (3) Finally, this must be done by research-practice partnerships that *connect scales and components of educational systems*, from individual students to school districts and states. These connections across scales play a key role in the iterative design cycle, as what we learn at one scale affects design at other scales.

The main sections of this article are organized around interconnected goals and activities in three different domains: assessments, classroom teaching, and professional networks. We summarize some key accomplishments and suggest ongoing issues in each domain.

Assessment: Understanding students and assessing three-dimensional learning. We report on more than 15 years of iterative development cycles leading to learning progression frameworks that are the basis for an online assessment system used for both classroom and large-scale assessment. Our findings and hypotheses related to the three DBIR issues are summarized below.

- *Three-dimensional learning:* Learning progression research reveals the size and nature of the intellectual challenges that three-dimensional performance expectations pose for students and guides the development of written assessment questions. Three-dimensional performances cannot be measured with forced-choice questions alone, but we report significant progress with automated scoring of students' written explanations.
- *Scale:* Our assessment system is working at scale—over 60,000 student tests administered and over 500,000 student written responses (as well as forced-choice responses) scored in 2016-17. We see it as an existence proof for the viability of large-scale assessments of three-dimensional learning, but only at the conclusion of multiple topic-specific iterative development cycles.
- *Diversity:* Students who are diverse with respect to academic success and understanding of science can respond meaningfully to *Carbon TIME* assessment questions. The assessments are useful to many teachers for classroom formative assessment. The challenge of how to reconcile the expectation of three-dimensional performances with many teachers' approaches to classroom grading and concerns about fairness will require substantial attention if our goal of large-scale change is to be achieved.

Teaching: Supporting classrooms as learning communities. Classrooms are communities of practice that impose obligations on their members—teachers and students. Meeting those obligations—around management, engagement, curriculum, and assessment, and communication—is necessary for successful classroom work on three-dimensional science learning. We report on the iterative development of curricular activity systems that include storylines, instructional models, assessments, and scaffolds for successful student performance. Our findings and hypotheses related to the three DBIR issues are summarized below.

- *Three-dimensional learning:* Our results confirm that three-dimensional learning is a heavy lift for most teachers and students. Developing curricular activity systems that assess and scaffold three-dimensional engagement with phenomena is intense, demanding work that must be done through partnerships including both teachers and researchers, and including both material supports and professional development. Our evidence shows that *Carbon TIME* tools are robust and educative, but not universally successful.
- *Scale and diversity:* Overall, *Carbon TIME* tools increase student learning and reduce the achievement gap between high-pretest and low-pretest students. But teachers make a difference. We observe large differences among classroom communities, and those differences are consequential for learning. Our direct work with teachers helps to describe both the nature and the reasons for those differences, as well as the continuing challenges posed by the scale and diversity of our educational systems.

Professional communities: Working with schools in research-practice partnerships. Accomplishing successful change at scale will require changes in school

organizations and partnerships that bring researchers into ongoing direct work with school-based professionals—teachers and administrators.

- *Three-dimensional learning:* Our findings reveal the multiple issues that schools and research-practice partnerships must resolve. They will need to (a) build material, human, and social resources, (b) resolve potential conflicts between school-based norms and obligations for teachers and the demands of three-dimensional learning, and (c) develop the practical knowledge necessary to enact three-dimensional learning in classrooms.
- *Scale and diversity:* Schools make a difference. Both our large-scale data analyses and our case studies show that when schools had more organizational resources, and when their professional norms and obligations supported three-dimensional science learning, teachers were able to use *Carbon TIME* tools more effectively.

We began this paper with a design challenge: enacting the aspirational goals of *NGSS* in educational systems that are fragmented and resource-poor. 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 knowledge-building around the connected challenges of three-dimensional learning, scale, and diversity; and in supporting improvements for the material, human, and social resources of educational communities of practice.

References

Allen, C. D., & Penuel, W. R. (2015). Studying teachers' sensemaking to analyze teachers' responses to professional development focused on new standards *Journal of Teacher Education, 66*(2), 136-149.

Alonzo, A. C., Neidorf, T., & Anderson, C. W. (2012). Using learning progressions to inform large-scale assessment. In *Learning progressions in science* (pp. 211-240). SensePublishers, Rotterdam.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Banilower, E. R., Smith, P. S., Weiss, I., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). Report of the 2012 National Survey of Science and Mathematics Education. Chapel Hill, NC: Horizon Research, Inc.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching, 53*(7), 1082-1112.

Bloome, D., Puro, P., & Theodorou, E. (1989). Procedural display and classroom lessons. *Curriculum Inquiry*, *19*(3), 265-291.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, *2*(2), 141-178.

Bryk, A. S. (2009). Support a science of performance improvement. *Phi Delta Kappan, 90(8)*, 597-600.

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.]

Center for Educational Leadership (2016). The 5D+ rubric for instructional growth teacher evaluation. University of Washington.

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., & Geil, K. (2013). *Research-practice partnerships at the district level: A new strategy for leveraging research for educational improvement*. Berkeley, CA and Boulder, CO: University of California and University of Colorado.

Cohen, D. K., & Ball, D. L. (1999). Instruction, capacity, and improvement. Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Cohen, D. K., Moffitt, S. L., & Goldin, S. (2007). Policy and practice: The dilemma. *American Journal of Education, 113*(4), 515-548.

Cohen, D. K., Peurach, D. J., Glazer, J. L., Gates, K., & Goldin, S. (2013). Improvement by design: The promise of better schools. Chicago, IL: University of Chicago Press.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 1-24.

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.

Corcoran, T., Mosher, F.A., and Rogat, A. (2009). *Learning progressions in science*. Philidelphia, PA: Consortium for Policy Research in Education.

Covitt, B. & Anderson, C. (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., Gunckel, K. L., Caplan, B., & Syswerda, S. (2018). Teachers' use of learning progression-based formative assessment in water instruction. *Applied Measurement in Education*, *31*(2), 128-142.

Covitt, B. A., Morrison Thomas, C., Bodbyl, S., Lin, Q., Hancock, J. B., Kohn, C., de los Santos, E., & Anderson, C. W. (2018, March). Characterizing Discourse Patterns of Assessing and Scaffolding with Evidence from *Carbon TIME* Classroom Video. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890-1990*. Teachers College Press.

Danielson, C. (2014). *The Framework for Teaching Evaluation Instrument*, *2013 Edition*. The Danielson Group. Retrieved from www.danielsongroup.org.

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.

Davis, E. A., Palincsar, A. S., Arias, A., Bismack, A., Marulis, L., & Iwashyna, S.* (2014). Designing educative curriculum materials: A theoretically and empirically driven process. Harvard Educational Review, 84 (1), 24-52.

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). ProQuest:10605710.

de los Santos, E. X., Marshall, S., Hancock, J. B., Bodbyl, S., Forsyth, A., Lin, Q., Penuel, W., & Anderson, C. W. (2018, March). Teachers' sensemaking about accountability and assessment. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta. http://carbontime.bscs.org/conference-presentations

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2016). Building an assessment argument to design and use next generation science assessments in efficacy studies of curriculum interventions. *American Journal of Evaluation*, *37*(2), 174-192.

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. In B. J. Fishman, W. R. Penuel, A.-R. Allen, & B. H. Cheng (Eds.), Design-based implementation research: Theories, methods, and exemplars. National Society for the Study of Education Yearbook. (pp. 443-463). New York, NY: Teachers College Record.

Doyle, W. (1983). Academic work. *Review of Educational Research*, *53*(2), 159-199.

Doyle, W., & Carter, K. (1984). Academic tasks in classrooms. *Curriculum Inquiry*, *14*(2), 129-149.

Engle, R. A. (2012). The productive disciplinary engagement framework: Origins, key concepts, and developments. In Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 161-200). New York, NY: Routledge.

Engle, R. A., Lam, D. P., Meyer, X. S., & Nix, S. E. (2012). How does expansive framing promote transfer? Several proposed explanations and a research agenda for investigating them. *Educational Psychologist*, *47*(3), 215-231.

Fishman, B. J., & Krajcik, J. (2003). What does it mean to create sustainable science curriculum innovations? A commentary. *Science Education*, *87*(4), 564-573.

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and teacher education*, *19*(6), 643-658.

Fishman, B. J., Penuel, W. R., Hegedus, S., & Roschelle, J. (2011). What happens when the research ends? Factors related to the sustainability of a technology-infused mathematics curriculum. *Journal of Computers in Mathematics and Science Teaching, 30*(4), 329-353.

Furtak, E. M. (2012). Linking a Learning Progression for Natural Selection to Teachers' Enactment of Formative Assessment. *Journal of Research in Science Teaching. 49(9)*, 1181-1210.

Furtak, E. M., *Morrison, D. L., & *Kroog, H. (2014). Investigating the Link Between Learning Progressions and Classroom Assessment. *Science Education, 98(4)*, 640-673.

Furtak, E. M., Thompson, J., Braaten, M., & Windschitl, M. (2012). Learning progressions to support ambitious teaching practices. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 405-433). Rotterdam, The Netherlands: Sense Publishers.

Gamoran, A., Anderson, C. W., Quiroz, P. A., Secada, W. G., Williams, T. and Ashmann, S. (2003). *Transforming teaching in math and science: How schools and districts can support change.* New York, NY: Teachers College Press.

Gee, J. P. (1991). What is literacy? In C. Mitchell and K. Weiler (eds.), *Rewriting literacy: Culture and the discourse of the other.* New York: Bergin & Garvey, pp. 3-12.

Gioia, D. A., & Chittipeddi, K. (1991). Sensemaking and sensegiving in strategic change initiation. *Strategic Management Journal, 12*(6), 433-448.

Gotwals, A.W., & Birmingham, D. (2015). Eliciting, identifying, interpreting and responding to students' ideas: Teacher candidates' growth in formative assessment practices. *Research in Science Education*. DOI 10.1007/s11165-015-9461-2

Ingersoll, R. M. (2003). *Who controls teachers' work? Power and accountability in America's schools*. Cambridge, MA: Harvard University Press.

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.

Jackson, P. W. (1990). *Life in classrooms*. Teachers College Press.

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

Jin, H., & Anderson, C. W. (2012b). Development and validation of assessments for a learning progression on carbon cycling in socio-ecological systems. In A. Alonzo and A. Gotwals (eds), *Learning progressions in science: Current challenges and future directions,* pp. 151-182. Boston: Sense Publishers.

Johnson, S. M., Kraft, M. A. & Papay, J. P. (2102). 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*, [need page numbers]

Johnson, W. (2017). Framing classroom activities as figuring out phenomena versus learning about science: The role of curiosity-driven discourse in three-dimensional science learning. Unpublished doctoral dissertation, Michigan State University.

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.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1-73.

Kennedy, M. (2016). Parsing the practice of teaching. *Journal of Teacher Education, 67(1)*, 6-17.

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.

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education, 96*(4), 701-724.

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education, 96*(6), 1071-1105.

McNeil, L. M. (1998). *Contradictions of control: School structure and school knowledge*. New York: Routledge.

McNeill, K. L., & Knight, A. M. (2013). Teachers' pedagogical content knowledge of scientific argumentation: The impact of professional development on K–12 teachers. *Science Education, 97*(6), 936-972.

McNeill, K. L., & Krajcik, J. (2009). Synergy between teacher practices and curricular scaffolds to support students in using domain-specific and domain-general knowledge in writing arguments to explain phenomena. *The Journal of the Learning Sciences, 18*(3), 416-460.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in the Philosophy of Education, 27*, 283-297.

Miller, H. K., Johnson, W. R., Freed, A. W., Doherty, J. H., & Anderson, C. W. (submitted, 2017). Crosscutting concepts for re-orienting science education. Submitted to *Science Education.*

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 Academy 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.

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

Opfer, V. D., Kaufman, J. H., & Thompson, L. E. (2016). Implementation of K-12 state standards for mathematics and English Language Arts and literacy: Findings from the American Teacher Panel. Santa Monica, CA: RAND.

Penuel, W. R. (2015). *'Infrastructuring' as a practice for promoting equity and transformation in design-based implementation research*. Paper presented at the International Society for Design and Development in Education (ISDDE) 15, Boulder, CO.

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

Penuel, W. R., Phillips, R. A., & Harris, C. J. (2014). Analysing curriculum implementation from integrity and actor-oriented perspectives. *Journal of Curriculum Studies, 46*(6), 751-777.

Peurach, D. J. (2016). Innovating at the nexus of impact and improvement: Leading educational improvement networks. *Educational Researcher, 45*(7), 421-429.

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.

Reiser, B. J., Michaels, S., Moon, J., Bell, T., Dyer, E., Edwards, K. D., & Park, A. (2017). Scaling up three-dimensional science learning through teacher-led study groups across a state. *Journal of Teacher Education*, 1-17.

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.

Roschelle, J., Pierson, J., Empson, S., Shechtman, N., Dunn, M., & Tatar, D. (2010). Equity in scaling up SimCalc: Investigating differences in student learning and classroom implementation. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences* (Vol. 1, pp. 333-340). Chicago, IL: International Society of the Learning Sciences.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing

middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47*(4), 833-878.

Roth, K. J., & Garnier, H. (2007). How five countries teach science. *Educational Leadership*, *64*(4), 16-23.

Roy, G. J., Fueyo, V., & Vahey, P. (2017). Supporting middle grades mathematics teachers and students: A curricular activity system used in an urban school district. *Research in Middle Level Education Online, 40*(2), 1-15.

Russell, J. L., Jackson, K., Krumm, A. E., & Frank, K. A. (2013). Theories and research methodologies for design-based implementation research: Examples from four cases. In B. J. Fishman, W. R. Penuel, A.-R. Allen, & B. H. Cheng (Eds.), *Design-based implementation research: Theories, methods, and exemplars. National Society for the Study of Education Yearbook.* (pp. 157-191). New York, NY: Teachers College Record.

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

Schwarz, C. V., Passmore, C. M., & Reiser, B. J. (2017). Moving beyond "knowing" science to making sense of the world. In C. Schwarz, C. Passmore, and B. Reiser (Eds.) *Helping students make sense of the world using next generation science and engineering practices*, 3-21. Arlington, VA: NSTA Press.

Spillane, J. P., & Burch, P. (2006). The institutional environment and instructional practice: Changing patterns of guidance and control in public education. In H.-D. Meyer & B. Rowan (Eds.), *The new institutionalism in education* (pp. 87-102). Albany: State University of New York Press.

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.

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.

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.

Todd, A., Romine, W. L., & Cook Whitt, K. (2017). Development and Validation of the Learning Progression–Based Assessment of Modern Genetics in a High School Context. *Science Education*, *101*(1), 32-65.

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.

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, meanings, and identity*. New York: Cambridge University Press.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878-903.

Yao, J. X., & Guo, Y. Y. (2018). Validity evidence for a learning progression of scientific explanation. *Journal of Research in Science Teaching*.

Zuiker, S. J., Piepgrass, N., & Evans, M. D. (2017). Expanding design research: From researcher ego-systems to stakeholder ecosystems. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1-28). Cham: Springer International Publishing.