Supporting Teachers in Designing Assessments Aligned to the Vision of the Framework: Findings from Two Design Studies

Abraham S. Lo

BSCS Science Learning

William R. Penuel and Kerri Wingert

University of Colorado Boulder

Paper presented at AERA Annual Meeting, April 2022

San Diego, CA

This material is based in part upon work supported by the National Science Foundation under Grant Numbers DRL-1748757 and DRL-2010086. 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.

Supporting Teachers in Designing Assessments Aligned to the Vision of the Framework: Findings from Two Design Studies

The vision of *A Framework for K-12 Science Education* (National Research Council, 2012) outlined new aims for developing all students' proficiency in science, often referred to in shorthand as "three-dimensional (3D) science learning," whereby students develop and use understandings of disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs) to explain science phenomena or solve engineering problems. In contrast to the first generation of science standards that separated content from inquiry goals (National Research Council, 1996, 2000), the *Framework* called for standards that focused on helping students gain a grasp of and use of science and engineering practices to support their learning of ideas core to the disciplines of science and crosscutting concepts that unify different fields of science and engineering. This image of science proficiency is grounded in studies of how people learn science that illustrate ways in which students' understanding of ideas and grasp of practice co-develop and support one another (e.g., Lehrer & Schauble, 2006; Manz, 2012; Windschitl et al., 2008), as well as an extensive body of research on how scientists and engineers develop, share, and critique knowledge and designs (e.g., Bucciarelli, 1994; Knorr-Cetina, 1999; Pickering, 1995).

Implicit in the *Framework's* multi-dimensional framework for science proficiency are two additional latent dimensions, *interest* and *identity* (Bell et al., 2016). A core assumption of the *Framework* (National Research Council, 2012) is that science learning experiences should "connect with their [students'] own interests and experiences" (p. 28), and that "build their [students'] identities as capable learners of science and engineering" (p. 249). Doing so is an important vehicle for realizing another goal in the *Framework,* that of promoting equity, specifically the goal of broadening participation in science (p. 28), insofar as connecting to students' interests and identities could help a broader range of students identify with science and engineering as enterprises (p. 285).

Designing assessments that address all five of these dimensions presents several challenges to assessment developers and classroom teachers alike. For one, they need to elicit and require students to apply their understanding of DCIs, SEPs, and the CCCs (Pellegrino, 2013). Assessment tasks can accomplish this goal when anchored in specific phenomena to be explained (for science) or problems to be solved (for engineering), but choosing phenomena that invite student sensemaking related to key science ideas and practices and that are culturally relevant and personally meaningful to students is not a straightforward process (Lo et al., 2022; Penuel et al., 2022; Penuel, Lo, et al., 2019; Penuel, Turner, et al., 2019). Further, eliciting students' understanding in ways that serve the larger goal of sensemaking (Odden & Russ, 2019) requires the design of a sequence of prompts that elicit relevant student resources for sensemaking as starting points, and building incrementally toward more complete explanations or solutions (National Academies of Sciences Engineering and Medicine, 2019). Finally, developing tasks that allow and encourage students to use and teachers to recognize a wide range of communicative resources that are accessible to emergent multilingual learners require its own set of guidelines to support designers (Bang et al., 2017; Fine & Furtak, 2020; Furtak et al., 2020).

In this paper, we report on conclusions from two design studies conducted over two successive years focused on a professional learning workshop designed to support teachers in developing assessments aligned to the vision of *A Framework for K-12 Science Education*. Our high-level conjecture behind our design studies was that engaging teachers in task adaptation and design activities could support teachers in creating assessments that reflect the vision and provide a context for developing a better understanding of the challenges of developing tasks that connect to students' interests and identities and that require students to apply their understanding of DCIs, SEPs, and CCCs to explain observable events in the natural world. In these design studies, we focused particularly on supporting educators in learning to develop tasks that focused on the three dimensions highlighted in the standards (NGSS Lead States, 2013). However, as we elaborate below, our design studies informed a redesign of the professional learning to support teachers in developing a more robust vision for five-dimensional (5D) science teaching and learning that involve teachers engaging students' interests and identities as two additional dimensions to attend to when designing opportunities for students to use the 3Ds to explain phenomena or solve problems.

Theoretical Framework

Guiding our work were two interrelated layers of ideas about (1) science assessment of the 3Ds of the *Framework* in a way that attends to interest and identity and (2) how task adaptation and design can function as ways to support teacher learning.

Assessing the Three Explicit Dimensions in the Framework

Assessing students' proficiency requires tasks in which students must apply their understanding of all 3Ds (i.e., DCIs, SEPs, and CCCs) (Pellegrino, 2013). If such tasks are to elicit students' ability to do science, such tasks need to engage students in multiple, "connected" practices (National Research Council, 2014, p. 130). Scientific practices are not isolated but are integrated in service of answering specific questions to develop knowledge about the natural world (Kelly, 2012; Manz, 2015). Anchoring tasks in specific phenomena to be explained (for science) or problems to be solved (for engineering) provides a context to elicit this kind of connected use of SEPs, integrated with DCIs and CCCs. Tasks that facilitate sensemaking about phenomena and problems create contexts for students to demonstrate their integrated understanding of the 3Ds of science learning (Achieve, 2018; Penuel, Turner, et al., 2019). In these contexts, students develop and revise models for explaining target phenomena (Windschitl et al., 2008) or for designing solutions to address specific problems (Kolodner et al., 2003).

Beyond being organized around phenomena and problems, assessment tasks must be organized to demand that students use their understanding of the 3Ds to make sense of the scenario presented to them. This will likely require that the scenario point explicitly to gaps in knowledge or needs to be addressed to explain the phenomenon or solve the problem. This goal may be accomplished through an explicit problematizing of the context to highlight DCIs at stake and to motivate students (Engle, 2012; Reiser, 2004). Further, prompts should be designed to be *in the service of* sensemaking about phenomena and problems, such that knowledge of the dimensions is not elicited as isolable knowledge or skills in a manner that would undermine the image of science learning presented in the *Framework for K-12 Science Education* (National Research Council, 2012).

Interest and Identity in Assessment

Learning sciences research in the past decade has increasingly recognized that students' interest and identity are fundamental to learning. Interest facilitates motivation, engagement, persistence, and self-regulation in learning (National Academies of Sciences Engineering and Medicine, 2018). In addition, interest has been shown to be related to learning outcomes, such as text comprehension and problem solving (Renninger & Hidi, 2002; Walkington & Bernacki, 2014). Similarly, students' personal and social identities shape their goals and motivation, as well as how they participate in different kinds of learning environments (Calabrese Barton &

Tan, 2010; Nasir, 2002; Wortham, 2004, 2006). Of particular importance in the practice-focused view of science learning is helping cultivate *practice-linked* identities (Nasir & Cooks, 2009; Nasir & Hand, 2008), that is, a sense of identification with disciplinary practices that grows from an understanding of how they can help realize personal and community values and aims.

There is strong evidence that interest is malleable and can be influenced by the tasks and materials teachers use (Hunsu et al., 2017; Renninger et al., 2002). However, an enduring challenge for teachers is to find practical ways to link students' interests and identities to externally defined goals for learning (i.e., standards). Some methods for supporting individualized pathways are too time intensive, and teachers experience tension when trying to meet standards that link to student interests and community priorities (Kapon & Merzel, 2019; Morales-Doyle et al., 2019). For this reason, our workshop highlights strategies for *group personalization* (Walkington & Bernacki, 2014) that are feasible to implement and consistent with the goal of promoting science learning as a collective enterprise. These include surveys of student interest that can be used to select phenomena and problems that all students will address, as well as visualization tools to help teachers see and act on patterns of student experience in their classrooms (Penuel & Watkins, 2019). The workshop also includes guidance for personalization grounded in social psychological studies showing that asking students to reflect on the relevance of lessons can enhance interest in lessons (e.g., Hulleman & Harackiewicz, 2009).

Several strategies also have been tested for enhancing students' practice-linked identities, which can be incorporated into both learning and assessment tasks. One key strategy is to make connections to the knowledge, goals, and practices valued by students' communities (Nasir, 2002). This strategy can be effective, because it supports their sense of belonging and helps them see connections between community values and the learning setting. Another strategy is to re-position students as "epistemic agents" within the classrooms, that is, as bearers of knowledge and expertise (Miller et al., 2018), which can help open students to opportunities presented by teachers (Lee & Reeves, 2012). Other strategies include eliciting and connecting to place-based, social, and cultural dimensions of students' identities that are important to them (Esteban-Guitart & Moll, 2014; Kang & Furtak, 2021; Llopart & Esteban-Guitart, 2017).

Supporting Teacher Learning through 5D Task Design

In our professional learning, we prepare teachers to engaged in a process of principled assessment design, as outlined in *Knowing What Students Know* (National Research Council, 2001). Such a process begins by defining the constructs to be assessed identifying the kinds of situations and tasks where evidence of student understanding may be observed, and specifying how such evidence is to be interpreted to draw inferences about student learning (Messick, 1994). Accordingly, the course's flow reflects this process, beginning with an analysis of standards and identification of questions to elicit understanding of students' communities. Once teachers have analyzed the standards, they identify candidate phenomena and design challenges that, if explained or solved by students, would require use of the 3Ds targeted in a single standard, while also potentially connecting to matters of concern to students and their communities, develop and administer a survey to students to identify which are most likely to be interesting to them, and then develop prompts to elicit student understanding of the 3Ds for the phenomenon or challenge they select. Teachers use the analysis of standards to develop scoring guides for prompts, so that they can draw inferences of students' mastery of performance standards, as well as questions that ask students to report directly about the success of a task in connecting to students' lives.

The approach we take to principled assessment design is grounded in a sociocultural theory of learning that is consistent with the model of science learning we are seeking to support. Sociocultural theory views learning as the transformation of participation in valued sociocultural activities that are themselves changing (Holland & Lave, 2009; Lave & Wenger, 1991; Rogoff et al., 1995; Rogoff et al., 2007). Participation in sociocultural activity necessarily involves more than simply acquiring knowledge; it involves processes of identification that, in turn, present opportunities for participants to become certain kinds of people in activity (Lave & Wenger, 1991). A sociocultural theory applied to *assessment design* emphasizes the need to elicit and support students' navigation between disciplinary and everyday ways of knowing and engaging in practice, and calls for tools that can help connect students' interests and identities to curriculum and strengthen relationships between teachers and students (Bell, 2019; Penuel & Shepard, 2016b). A call for such a focus on assessment appear in *Designing Assessments of the Next Generation Science Standards* (National Research Council, 2014), "A potential focus of classroom assessment at the outset of instruction is to elicit students' interests and experiences that may be relevant to the goals for instruction" (p. 127).

A final principle underlying our approach is the idea that engaging teachers in adaptation and design activities is a powerful context for promoting teacher learning. As compared to having teachers design new tasks or sequences of learning from scratch or asking them to implement existing materials "as is," adaptation supports improvements to both practice and student learning (Penuel et al., 2011). Task adaptation in the context of professional learning (PL) is a powerful tool for learning, because it provides for teachers models of tasks that reflect desired qualities, as well as the authority to tailor tasks to local classroom contexts in ways that build teacher buy-in and ownership (Penuel et al., 2017; Remillard, 1999; Voogt et al., 2011). As teachers become more skilled, engaging them in more expansive design—involving in this case the selection of phenomena and problems to anchor tasks, may be beneficial, because it helps them see underlying principles and criteria for selection that they can apply to tasks developed on their own (Penuel, Reiser, et al., 2018; Severance et al., 2016). Further, supporting teacher involvement in design within ongoing research-practice partnerships, as enacted within this project, has the potential to support collaborative learning between researchers and teachers, as they struggle together to overcome challenges related to the task of developing assessments aligned to the *Framework* (Kang & Furtak, 2021).

Methods

We conducted two concurrent design studies over two years to iterate upon a professional learning workshop designed to support teachers in developing assessments that elicit understanding of 3D science standards. Our primary research question across the two design studies was: *How can engaging teachers in task adaptation and design build teachers' capacity to create tasks that reflect the vision of* A Framework for K-12 Science Education?

In each version of our 20-hour workshop series, spread over three days, teachers first adapted tasks by developing prompts (items) to elicit student understanding of the 3Ds for a scenario (a rendering of a phenomenon or problem) we provided them, then had the opportunity to design their own, using a suite of tools for integrating SEPs and CCCs and following an evidence-centered design process we outlined for them. In what follows, we describe the general structure of the professional learning workshop series that took place, as well the iterations that were made between design studies and our method for analyzing growth in teachers' ability to develop assessments aligned to the vision of the *Framework*.

Description of the professional learning workshop Series

The principal aim of the workshop series was to provide teachers with opportunities to use a structured process and set of tools for designing 3D assessments to create their own assessments. Below we outline the general structure of the initial design of the three-day workshop series; later in the paper, after presenting an overview of findings from the first design study, we describe revisions made after the first iteration on the series at the conclusion of the first design study.

The first day of the workshop focused on task analysis and adaptation. First, teachers developed and refined a set of criteria for what makes for good assessments that elicit students' integrated understanding of the 3Ds of science proficiency outlined in the *Framework*. Next, they analyzed a set of tasks that varied with respect to their dimensionality, but that were supposed to target the same performance expectation (Stromholt, 2017). They looked at a separate model 3D assessment that had been peer reviewed using a task screener developed by Achieve, Inc. (2018), and further refined criteria for good 3D assessments. Then, each teacher was given a scenario identified by the research team for educators as a potentially viable means for assessing a standard that teachers had selected to assess, and then they were invited to develop a complete multi-component, 3D task. To support prompt development around the SEPs and CCCs, we provided two scaffolds: the Science and Engineering Practices Task Formats tool (Van Horne et al., 2016) and the Prompts for Integrating Crosscutting Concepts into Assessment and Instruction tool (Penuel & Horne, 2016). They shared their assessments with each other and analyzed the cognitive demand of their assessments using Appendices F and G of the NGSS, which describe grade-band expectations for the practices and crosscutting concepts.

The second day of the workshop focused on developing scoring guides that attended to specific aspects of standards that had been "unpacked" by the research team and by teachers (Krajcik et al., 2014). During the workshop, participants examined an existing task, the task's scoring guide, and student work to iteratively develop claims about how the task could help teachers make conclusions about what students knew and could do and suggest revisions to the provided scoring guide. These activities were intended to help provide a basis for revising the task and developing scoring guides that attended to component ideas from performance expectations that were evident in student responses. Next, participants had the opportunity to work on a new assessment, again using the two key scaffolds for integrating SEPs and CCCs.

The third day of the workshop focused on scenarios, exit tickets, and task analysis. To help teachers learn about phenomenon selection for assessments, we engaged teachers in comparative analysis of different scenarios, and then worked to identify strategies for selecting phenomena for tasks that would require the use of disciplinary core idea components to explain. Next, participants had a chance to practice with our scaffolds to design exit tickets for an upcoming lesson. With respect to exit tickets, educators were introduced to prompts to elicit students' understanding of an individual lesson's DCIs and SEPs, as well as the student experience of the lesson (Penuel & Watkins, 2019). Finally, participants collectively self-assessed their growth in understanding of 3D assessments.

Analytical Methods

To analyze the effectiveness of the professional learning workshop on supporting teachers' ability to develop 3D assessments, each study included two conditions: one in which teachers experienced a workshop, and one where they were given tools to support assessment design to use on their own. We had 12 participants in the first design study (6 in workshop condition, 6 On your own condition) and 13 participants (4 in workshop condition, 9 On your own condition) in the second design study. We collected teacher-designed assessments before and after the workshop series and interviewed participants at the end. Teachers submitted assessment tasks and scoring guides prior to and after using the provided scaffolds and/or professional learning. When submitting each assessment task, teachers explained the assessment's goal, which could involve assessing a particular content area or performance expectation. In this paper, we focus on the growth and persistent challenges of teachers who participated in workshops.

We developed a rubric to assess the extent to which teachers designed 3D assessment tasks that involved students using important DCIs, SEPs, and CCCs in an integrated way to explain a phenomenon or solve a problem (see Appendix A and Penuel, Lo, et al., 2019). The rubric has five categories aligned with key features that were the focus of the professional learning:

- **Category 1: Appropriateness of the scenario.** This category assessed the authenticity of the problem or phenomenon for students to figure out and the extent to which students had the opportunity to make sense of information from the scenario and apply what they had learned in class to complete the assessment tasks. In addition, we assessed whether students had the opportunity to demonstrate their understanding using modes other than words. From an equity perspective, using multiple modes to demonstrate understanding enhances the assessment's accessibility to a range of learners (Achieve, 2018).
- **Category 2: Disciplinary core ideas (DCIs).** This category examined the extent to which the assessed content ideas, both stated explicitly by the teacher and those assessed

in the assessment task, were aligned with DCI elements found in the NGSS and were grade-band appropriate.

- **Category 3: Science and engineering practices.** This category examined whether the teacher created grade-band appropriate opportunities for students to engage in the SEPs to complete the tasks that contributed toward achieving the assessment goal.
- **Category 4: Crosscutting concepts.** Similar to Category 3, we assessed opportunities for students to use crosscutting concepts to achieve the assessment goal.
- **Category 5: 3D integration.** This category examined the overall coherence of the assessment task, which included the extent to which the assessment items were connected to explaining the target problem or phenomenon, and whether completion of assessment items required integrated or discrete use of the three NGSS dimensions.

We used our rubric to identify successes and challenges that teachers faced as they designed 3D assessment tasks. In what follows, we synthesize the successes and the challenges across both design studies and discuss how we applied what we've learned to develop a revised 25-hour professional learning course that not only supports teachers in developing assessment tasks that provide evidence of 3D learning, but also makes explicit how to do so in ways that engage student interest and identity.

Findings

In this section, we present a selection of findings from each design study, paying close attention to teachers' design and use of scenarios that present a phenomenon to students, and the design of prompts to allow students to use their 3D science learning to explain the phenomenon presented to them.

Findings from the First Design Study

In the first design study, the degree to which assessments created by teachers and submitted to the research team embodied the vision of the *Framework* improved modestly from pre-workshop to post-workshop. In both areas of focus for the current study—use of phenomena and integration of the 3Ds—scores on our rubric improved.

Growth and Challenges with Respect to Use of Scenarios

An area of growth across the sample of teachers in the workshop was the use of scenarios in which teachers described a phenomenon to be explained. Where before the workshop, teachers presented assessments that were a series of questions presented to students, their revised assessments began with a phenomenon.

At the same time, teachers found it challenging to develop scenarios that supported students in using the DCIs and engaging in the SEPs learned in class to answer prompts on the assessment. For example, a teacher created an assessment whereby students could piece together an explanation using information directly obtained from the assessment. This assessment did not require them to use what they had learned already in conjunction with information presented in the scenario to explain a phenomenon.

In addition, identifying what was “at stake” for students in figuring out phenomena was not always clear; that is, the phenomena were not clearly “problematized” (Reiser, 2004). Teachers often designed scenarios that involved straightforward solutions that did not require students to use the 3Ds to make sense of the evidence or have opportunities to decide between viable solutions. These findings highlighted the need for greater attention to the students’ perspective to consider how to develop scenarios that engage and invite students to wonder and create authentic contexts for students to use the 3Ds to engage in figuring out science ideas and practices.

We also found that even after the workshop, teachers presented scenarios and gave students opportunities to respond to prompts in a limited range of modalities. The most frequent modes were written responses and constructing diagrammatic representations of one's ideas, which included models or graphs.

Growth and Challenges with Respect to Eliciting Use of the 3Ds of Proficiency

One area of growth in the first design study was identifying a 3D performance expectation that was to be the target for the assessment. While at the beginning of the study, few identified a target from the NGSS three of four teachers in the workshop did at the end of the study. Further, we saw evidence that the identification of the performance expectation "bootstrapped" teachers' uses of two (SEPs and CCCs) of the 3Ds in prompts and scoring guides.

Teachers experienced some struggles related to constructing coherent assessments that were connected to the scenario, a key dimension of 3D integration. Most teachers were successful in creating assessment items that were topically related to one another, such as having questions that are related to natural selection or energy. However, there were challenges with creating assessment items that were explicitly connected to the assessment scenario and contributed towards answering the overarching question posed by the scenario. For example, in an assessment that asked students to explain how it was possible for all the human body's elements to come from stars, students were expected to use data to graph the general relationship between the mass of an atom and the temperature required to produce it in a stellar fusion reaction. However, students were not required to use this derived relationship to explain how this relationship would contribute to their understanding of the broader phenomenon.

In addition, teachers struggled with developing prompts and scoring guides linked to the specific elements of the dimensions targeted in the NGSS. Teachers' scoring was generally one-dimensional and focused on the DCIs. Further, the prompts and scoring guides often missed key components of those DCIs. For example, in an assessment explaining why whale size is limited, students explored ideas related to homeostasis (LS1.A) and how particular adaptations helped regulate an organism's body temperature. However, central to the elements of the DCI involved students understanding the role that feedback mechanisms played in this process, an aspect of the DCI element that was not present in the assessment. One possible explanation is that teachers were focused more on identifying performance expectations that were related to the desired content focus rather than using the DCI elements identified by the target PE to guide the assessment design.

Across both pre- and post-workshop assessments, teachers wrote assessments and scoring guides that required use of elements of the 3Ds that were often below grade-band as specified in targeted performance expectations. For example, when using the crosscutting concept systems and system models, high school students were asked to examine components and relationships within systems rather than between systems, as called for in the standards. When using the CCC of cause and effect, students often were not required to distinguish between cause and correlation.

Revisions Made Between First and Second Design Study

In light of what we learned from the first design study, we made modifications that we describe below. Of note, these initial revisions focused principally on improving alignment to targeted standards, which was one of the areas with which teachers struggled the most, along with writing prompts where students had to use all 3Ds to make sense of phenomena presented in scenarios.

Emphasis Given to Standards Alignment

In the revised workshop design, we provided more time for teachers to analyze the key components of a common performance expectation and discuss what it would mean for a student to master it. In addition, they engaged in discussions about where they stood with respect to some key tradeoffs that 3D assessment developers have identified, as well as some common conceptions about the NGSS (see Furtak & Penuel, 2019). Both activities, which took place in the initial hours of the workshop, were intended to signal the importance of standards alignment to teachers and make visible design choices related to standards when developing assessment, which was a challenge we identified as persistent for teachers.

Later in the workshop, we incorporated additional support for teachers in using their performance expectations analyses to develop an initial scoring guide for a provided assessment. Participants first identified prompts might elicit a facet or component of understanding (Minstrell, 1992) from the analysis of the performance expectations and then assigned point values to the desired responses. The goal of this activity was to help teachers map their specific prompts to opportunities to elicit understanding of the performance expectations and, if they found that a prompt was not adequate for eliciting any aspect of the standard, to revise their prompt. We emphasized the importance of scoring for all 3Ds as well, since teachers in the first iteration of the workshop tended to focus on DCI elements only.

Use of Different Tasks in Task Analysis

In the second design study, we asked participants to analyze two extended tasks that had been rated and annotated as part of Achieve, Inc.'s Task Annotation Project in Science (TAPS; Achieve, 2018). As part of this project, a group of educators and researchers developed and applied a rubric for rating the quality of 3D assessment tasks to a set of tasks submitted by several research groups from across the country. The tasks chosen for our professional learning workshops both met initial screening criteria for alignment to the vision, including the fact that both were scenario-based, multi-component tasks. At the same time, the tasks varied in quality, specifically with respect to their alignment to targeted standards.

We chose these tasks for three key reasons. First, in contrast to the sets of tasks used in the past that included mostly "one-dimensional" assessments, both of these provided models that teachers could emulate to show how students might use the 3Ds to explain a phenomenon presented in an assessment scenario. Second, the variability in standards alignment facilitated a discussion of standards alignment, which as noted above, we sought to emphasize more in the professional learning workshops. Third, the tasks could be reviewed on their own, and because they included annotations from experts, teachers could see what others noticed as salient features of tasks. In this way, we sought to leverage findings from other studies of teacher noticing that suggest a key aspect of noticing is learning how to make specific connections between classroom tasks and broader principles or ideas (Sherin & van Es, 2005).

Formalizing Feedback as a Component of the Intervention

As part of the first design study, we provided feedback on assessments in both conditions. The feedback to participants took the form of qualitative comments that were based on the research rubric created to analyze assessment quality for the study. In addition, the professional learning group received feedback during the workshop on assessments in progress from members of the research team, in the form of suggestions for improving different aspects of their tasks, from alignment to elicitation of science and engineering practices.

Findings from the Second Design Study

As in the first design study, the quality of assessments improved for both groups, particularly in the use of scenarios and success in developing assessments that integrated all 3Ds. And while teachers were responsive to changes made, some challenges persisted among the teachers, particularly with respect to the use of elements of targeted performance expectations of all 3Ds in task prompts and scoring guides.

Growth and Challenges with Respect to Use of Scenarios

As in the first design studies, there was significant growth among teachers who participated in the workshop with respect to their use of scenarios to anchor assessments. One teacher, for example, submitted as a pre-workshop assessment a test that consisted entirely of multiple-choice questions and constructed response items that were not connected to any scenario. Both post-workshop assessments this teacher submitted included scenarios, and each made clear especially what was "at stake" for students to address. In one task, which targeted students' ability to plan and carry out an investigation using information about the structure and properties of matter, as well as electrical charges at the atomic scale (HS-PS1-3), students were asked to help a group of engineers design a robot that could investigate pollution in underwater pipes by "gliding" across the water like water striders, without breaking the surface of the water. The scenario explained that students were going to plan an investigation after considering how the properties of both water and the striders' legs could help them stay on top of the water and evaluating whether other liquids with different kinds of bonds might be able to support the robots. Just as with this teacher, the scenarios made clear to students what was important to figure out and included authentic data. One assessment adapted data from a published scientific study that they had found (Puinean et al., 2010).

A challenge that persisted from pre-workshop to post-workshop—as in the first design study—was that most participants did not present the scenario or invite student responses using multiple modalities. Three of four teachers presented scenarios using words only. These suggested to us the need for more explicit attention to multimodality in assessments.

Growth and Challenges with Respect to Eliciting Use of the 3Ds of Proficiency

As in the first design study, there was an increase in the use of all three NGSS dimensions within teachers' tasks from pre-workshop to post-workshop. More teachers were able to design tasks that were tied explicitly and in a manner that could be coherent to students to the overall scenario. For example, in the above assessment about water striders, students had to generate explanations of macro-scale, observable phenomena presented (surface tension that allows striders to float on the water) in terms of intermolecular forces present in the water. In addition, they were asked to reflect on whether the temperature of the water would matter for whether the robot would work, using what they know about how a change in temperature can cause a change in surface tension. In these prompts, students were asked to engage in using the 3Ds in a way that was connected to the larger design problem they were tasked with solving.

At the same time, some challenges with respect to use of the 3Ds that we identified that teachers faced in the first design study persisted. Teachers still struggled to match assessments to grade-level expectations and to targeted pieces or elements of the 3Ds in targeted performance expectations. As an example, in a post-workshop assessment focused on natural selection, students were asked to reason about how insecticide resistance affects the survival of peach potatoes. Students examine data from a scientific study as part of the assessment, and the data shows variation in the thickness of cuticles, along with a pattern showing a correlation between cuticle thickness and insecticide penetration. But the assessment does not ask students to consider how natural selection may be at work in the scenario, nor do the data allow students to distinguish between cause and correlation. Nor are students asked to engage in argumentation where students defend an argument against alternative explanations. All three of these would have been necessary aspects of student understanding to elicit, to align with the targeted performance expectation.

Further Re-Design of the Professional Learning Experience

Our study findings suggest the need for further refinements to our professional learning workshop, to support the larger vision of supporting educators in developing a practice of 5D teaching and assessment. For one, we identified persistent challenges in presenting phenomena in ways that enable students to use different modalities for expressing what they know and can do. Broadening the range of ways students can express their ideas in assessments is key for developing assessments that connect to students' identities in meaningful ways (Fine & Furtak, 2020; Taylor, 2022). More generally, we recognized the need to interweave more explicitly throughout the professional learning experience how engaging student interest and practice-linked identities can be a vehicle for supporting more equitable assessment practices. Second, we saw the need for more support to help teachers design assessments and scoring guides that considered all 3Ds in an integrated fashion, and in ways that focus on targeted elements at the appropriate grade levels. That is, we see persistent challenges to alignment that are always present when designing assessments, and that require explicit methodologies to address (Herman & Webb, 2007). Third, we see the need to provide more explicit scaffolds to help teachers select their own phenomena and problems for assessment, if they are to develop greater autonomy in designing assessments for their classrooms. Our conjecture is that by making explicit this process of phenomenon selection—which typically requires a deep "unpacking" of standards, generation of multiple ideas, and gathering of data from students about how compelling a phenomenon is to them (Penuel, Bell, et al., 2018)—could help not only develop their skill further, but also address some of the persistent challenges of developing assessments aligned to standards.

Research on formative assessment in science and other disciplines (e.g., Black & Harrison, 2001; Kluger & deNisi, 1996; Penuel & Shepard, 2016a; Ruiz-Primo & Furtak, 2007) suggests that our professional learning workshops also needed to go further in considering how assessments could inform teaching and support student learning. For example, to the extent that we want improved assessment practice to help catalyze changes to teaching, we need to help teachers conceptualize the purpose and role that performance-based assessments played in classroom teaching. In addition, we need to provide more guidance on how to provide feedback to students on how to improve their students' understanding. Although it has always been an important goal for our work, we recognized the opportunity to use this course to promote broader shifts in classroom practices -- to consider how we can use professional learning grounded in assessment to drive shifts in instruction.

Noting the ambitious nature of our work and with new support from the National Science Foundation (DRL-2010086), we transformed our 3-day workshop series into a 25-hour online professional learning course that took place during nine, 2.5-3 hour sessions spread across three months. The course was first taught during Fall 2022 with eleven rural, secondary science teachers: three middle school science teachers and eight high school science teachers. The course was designed to be online to allow teachers, who may not have access to in-person professional learning, to have access to high-quality professional learning. Teaching the course after school allowed teachers the ability to practice what they learned from the course in their own teaching and provide opportunities for teachers to receive feedback from instructors and peers. Teachers were grouped by grade level and content area to allow teachers to collaborate and provide feedback on one another's work.

We designed three sets of lessons to address lessons learned from our design studies. Each lesson set was guided by the investigation of questions and use of tools to support teacher learning and address challenges identified our design study: 1) What does 5D teaching look like?, 2) How can we use phenomena to frame instruction and assessment?, and 3) How can we develop and use tasks to assess student understanding? Because a key goal was to address the need for a closer integration of assessment with teaching, we chose to devote the first two lesson sets to developing a 5D vision for designing phenomenon-driven instruction *and* assessment opportunities before pivoting explicitly in the third lesson set to strategies for designing assessments. We intentionally used examples from freely-available, high-quality phenomenon-driven instructional materials to allow teachers who desired to implement such an approach in their classroom the ability to do so. Although we are not explicitly supporting how to use these materials to support shifts in classroom instruction, we are collecting data to analyze what shifts in classroom instruction do occur as a result of our work in motivating the need for phenomenon-driven instruction that is guided by their understanding of the 3D standards and student interest and identity. This first lesson set also allowed us to provide teachers with alignment methodologies that focused not only on identifying elements of each of the three science-specific dimensions of the NGSS, but also interest and identity as co-equal dimensions that required their attention when designing instructions and assessment opportunities.

As in the previous professional learning, teachers practiced what they had learned to develop an assessment that they administered to their students. In the revised course, the development of this assessment was scaffolded throughout the course, involving teachers unpacking the targeted 3D standard for their assessment and going through the steps of developing components of their assessment throughout the course before administering it to their students. Thus, more time and opportunities for feedback were provided to teachers throughout the assessment development process. Table 1 describes the questions that guided each lesson set in our revised design, the learning goals, and how the lesson set was designed to address identified challenges. We conclude with a discussion of how these findings support a vision of equitable assessment practices.

Table 1
Course Revisions and Rationales

Questions	Learning Goals	How addressed challenges
Lesson Set 1: What does 5D teaching look like?		
● What does meaningful science learning and performance look like? ● How do we currently attend to the 5Ds in our instruction? ● What guidance do the standards provide for designing 5D instruction and assessment opportunities?	● Develop an explicit understanding of components of the 5D vision ● Understand how phenomenon-based instruction can support our vision of meaningful, 5D science learning ● Identify and understand how to use resources to inform design of 5D learning opportunities	● Develop integrated vision for use of 3Ds and engaging student interest and identities to explain a phenomenon. ● Unpack performance expectations to the elemental level to examine what is distinct at grade-band and what it means for students to demonstrate understanding of these key understandings.
Lesson Set 2: How can we use phenomena to frame instruction and assessment?		
● How do we choose phenomena to frame instruction and assessment?	● Identify criteria for choosing productive phenomena for 5D learning opportunities.	● Design assessment scenarios that present authentic situations that problematize the need to

		use the 3Ds to make sense of the phenomenon and apply what they have learned in class.
Lesson Set 3: How can we develop and use tasks to assess student understanding?		
● How do we write prompts that create meaningful opportunities for students to use the 5Ds to figure something out? ● How can we make sense of what students have learned and provide meaningful feedback? ● How can we support diverse sensemaking using assessments?	● Understand how to use criteria and tools to develop accessible prompts for explaining assessment scenarios. ● Understand how to use student work to identify the incremental build of students' 3D understanding. ● Identify ways to make assessments more accessible and better support diverse sensemaking	● Teachers develop tasks that include opportunities to integrate students' use of the 3Ds. ● Develop rubrics to provide feedback aligned with desired 3D outcomes. ● Teachers explicitly consider ways to enhance the accessibility of their assessments.

Lesson Set 1: What does 5D teaching look like?

In this lesson set, participants co-develop a vision for meaningful science teaching and learning that involves students engaged in "figuring out" authentic science situations that are relevant to students' lives. They engage in a series of activities designed to iterate upon this vision and motivate features of the 5D vision that can be made explicit to teachers later in the lesson set. First, teachers evaluated the suitability of various scenarios, such as preventing wildfires or identifying sustainable ranching and farming practices, for supporting their vision for meaningful science teaching and learning. Next, teachers engaged in learner hat experiences and watched videos of actual classrooms to see how phenomenon or problem-drive instruction could support their co-developed vision for meaningful science learning. Teachers used the language from the *Framework* to outline a vision for 5D teaching and learning and consider how the 5Ds were evident in the classroom video that they watched. By providing learner experiences that help teachers to recognize the value of attending to these dimensions before naming them, we hoped to better support meaningful use of these dimensions so that teachers could not only describe what these dimensions were, but also how and why attending to them in instruction and assessment could enhance student learning. For teachers who were new to the NGSS or *Framework*-aligned instruction, these activities provided the opportunity for teachers to explore the value of phenomenon-based instruction and how it could be used to develop the targeted 3D understandings in ways that engage student interest and identity.

We motivated the need for teachers to explicitly collect information about their students so that they can design instruction that can engage student interest and develop practice-linked identities in students when designing instruction and assessment opportunities to meet the targeted 3D learning goals. Through the distribution and analysis of the Student and Community Interest Inventory (see Appendix B), teachers solicited information about their students' communities and students' interests and ideas about using and doing science, both in school and in their everyday lives. This tool was not designed to limit teachers' design choices to students' immediate interests, but to provide them with insight about their students to develop potential entry points for motivating interest to issues that move beyond individual or local relevance, and to consider the relevance of regional or global issues.

Having analyzed input from students' interests, the class shifted to analyzing the 3D standards in the NGSS or their state's standards that would ground the development of students' 5D learning opportunities. This sequence was intentional to ensure that teachers were thinking about their students' interests, identities, and experiences when conceptualizing what it would mean for students to demonstrate these targeted understandings and how they might build upon students' prior experiences. In light of the challenges that teachers faced with developing assessments that attended to the standards at the elemental level and provided grade-level opportunities for students to use the DCIs, SEPs, and CCCs, we developed an unpacking tool (see Appendix C) to support teachers in conceptualizing what it would mean for students to engage in each dimension both individually, but also in the context of its use with the other dimensions when demonstrating the performance expectation. To help teachers to better attend to the distinctions of what is expected at the appropriate grade-band level, we created spaces for teachers to use the NGSS Appendices and the *Framework* to describe what the dimension or element looked like at earlier and later grade-bands to help teachers clarify the aspects of DCIs, SEPs, and CCCs that they are responsible for developing and assessing. We model how to engage in the unpacking process in an iterative way, both through unpacking common PEs before unpacking a PE of their choice, to ensure that teachers develop a common understanding of why unpacking is important and the considerations that are needed when unpacking.

In addition, teachers considered students' prior experiences, the questions their students might have about the targeted understandings or examples of real-world phenomena or problems that they might leverage when designing instruction and assessment opportunities to ensure that they are continuously thinking about engaging student interest and identity as dimensions that are interwoven and co-equal with the NGSS dimensions. Taken together, this approach to developing instruction and assessment opportunities is potentially more equitable, as we are building upon what students have learned in the past to ensure grade-level appropriate understandings, while thinking about productive contexts for building upon students' prior experiences.

Lesson Set 2: How can we use phenomena to frame instruction and assessment?

Although teachers have experienced phenomena in Lesson Set 1, we shifted towards co-developing explicit criteria that can be used when choosing phenomena that can not only support students in developing the target 3D learning, but also in ways that engage and sustain student interest and create authentic context for students to use the 3Ds to figure out a phenomenon. To do so, we clarified our understanding of what phenomena are and what they are not and brainstormed criteria that are important to consider when choosing phenomena that will engage our students' interests and support them in figuring things out. Throughout this lesson set, we used examples of phenomena from a high school biology unit that allowed us to see the characteristics that phenomena need to embody if they are used to either anchor a unit of instruction, support students in investigating key idea, and provide an opportunity for students to demonstrate what they had learned in the context of an assessment. For example, all phenomena should involve puzzling observations and invite students to use the 3Ds to explain the phenomena. However, the time needed to authentically make sense of the provided assessment data should be scaled back, yet still invite wonder as to ensure that students feel like they are still authentically engaged in using the 3Ds to figure out the phenomenon.

To iterate upon these characteristics, we used them to evaluate the suitability of example *assessment scenarios* for supporting 5D learning opportunities using information from their student and community interest inventories and the alignment of potential explanations to the target 3D standards. Assessment scenarios involve the presentation of the phenomena or problems to be explained, which may involve information or data and the question that motivates students to use the 3Ds to make sense of the phenomenon. Follow-up prompts would then scaffold and support the explanation of that phenomenon or problem. Teachers worked in small groups to consider revisions to these assessment scenarios to better align them with our 5D vision. Preparing teachers to adapt existing assessment scenarios is an important step to allow them to use existing resources as sources for phenomena.

Our design study data showed that teachers find it challenging to choose appropriate phenomenon and problems for sensemaking, to not only ensure that the desired 3D learning outcomes align at the element level and grade level expectations, but also that the phenomenon is presented in such a way that students are invited to wonder about the phenomenon and create authentic contexts for which students use the 3Ds to figure out the phenomenon, rather than have a straightforward response requiring little deliberation. To better prepare teachers for choosing appropriate phenomenon for both instruction and assessment, we developed a Choosing Phenomenon Tool (see Appendix D) that formalized the use of criteria for vetting candidate phenomena and choosing appropriate data or information to make visible the observations to be explained and how students will use the targeted 3D elements to make sense of the phenomenon. This tool supported meaningful use of the criteria because the course activities supported teachers in going through these steps, so teachers were familiar with how to do it and why they were important. Teachers practiced using the tool with common examples before using it to develop new assessment scenarios for a PE of their choice.

It is important to note that most of what is learned in the first two lesson sets could be applied to help teachers choose phenomena to develop 5D learning opportunities for both assessment *and* instruction. Although the final product that is developed through this course will be an assessment, we hope that teachers will recognize the importance of needing to revise their classroom instruction to align with the expected performance on their assessments. In addition, we hope this will better support teachers in recognizing how the designed assessments need to create opportunities for students to apply what they had learned in class to make sense of the assessment scenarios rather than creating an authentic 3D performance task.

Lesson Set 3: How can we develop and use tasks to assess student understanding?

Lesson Set 3 is focused on developing the prompts that will scaffold students' use of the 3Ds to make sense of the assessment scenarios developed at the end of Lesson Set 2. In our design study, we recognized that teachers often had challenges using the provided tools to develop prompts that invited students to use the 3Ds in integrated ways. Oftentimes, they would use the provided SEP Task Formats and CCC Prompts tools to develop unidimensional prompts that resulted in opportunities to use the 3Ds individually, but not necessarily together to reflect the true vision of the 3D standard.

In the revised course, we invited teachers to revisit the targeted 3D elements and engaged in visioning work to consider how these dimensions could work together to support students in explaining the phenomenon or solving the problem. Teachers then used a newly developed Prompt Development Tool (see Appendix E) that invited teachers to use the previously developed tools (SEP Task Formats and CCC Prompts tools) to provide example language that teachers could build upon in their assessments. Teachers worked in small groups to develop integrated prompts for example scenarios with which teachers had experienced in previous lesson sets before applying what they learned to develop assessment prompts for an assessment that they would administer to their students. It is important to note that it was intentional for teachers to engage in this 3D visioning work *before* introducing the tools. Since the tools provided uni-dimensional prompts, we did not want teachers to use the tools in isolation without considering the bigger picture for how these prompts could be synthesized or used in concert with one another to develop assessment opportunities that supported integrated use of the 3Ds and engaged student interest and identity.

As before, teachers developed assessments that they administered to their students. There was a month between sessions to ensure that teachers had sufficient opportunity to develop and administer the assessment to ensure that the assessment was framed in a unit of instruction that was aligned with the targeted 5D vision. In addition, teachers had the opportunity to receive feedback from a course facilitator during this time. To complement the student and community interest inventory that teachers administered earlier in the year, teachers administered a Student Assessment Experience Survey (see Appendix F) to their students to ascertain the extent to which students felt prepared to complete the 3D assessment and whether they were engaged and felt as if they were positioned to be a knower, doer, and user of science.

After administering the assessment to their students, teachers brought examples of student work and the results of the Student Assessment Experience Survey to evaluate how well their designed assessments aligned with the 5D vision and met the target 3D learning goals. Teachers engaged in a process of developing a rubric based on student work to examine desired responses with respect to the 3Ds and to consider how the rubric could be used to provide 3D feedback to students. Teachers then applied what they learned to revise their assessments and scoring guides. It is important to note that teachers felt that this rubric discussion was too late, and desired guidance for scoring their designed assessments before administering them. Subsequent versions of the course will include opportunities to examine example student work from a common example to support teachers in developing their own rubrics before administering them to their students.

Although equity has been an explicit focus throughout this course, we culminate the course by thinking about how we can operationalize this equity vision in the context of our assessment design. Up to this point, we have considered one aspect of accessibility in that we use information about our students to develop prompts that scaffold students' use of the 3Ds to make sense of phenomena that are engaging to them. Using our student work, we revisit another aspect of our assessment design to see the extent to which we can better support opportunities for diverse sensemaking. Using a universal design lens (Rose & Meyer, 2002), we use a Supporting Diverse Student Sensemaking checklist to evaluate the extent to which we have created diverse opportunities to engage students and represent and express student ideas. After evaluating their assessments, teachers consider how they might revise their assessments to make them more equitable. Although teachers reported that they wished they had this checklist earlier, the underpinnings of this checklist had been interwoven in the course. In addition, our prior design work showed us that the task of developing 3D assessments, let alone 5D assessments, is a sufficient challenge, so we felt that introducing the checklist at the end of the course, after they had developed and administered their assessments, would create a more opportune moment for teachers to consider this aspect of equity in their assessments. At the end of the course, teachers consider how they might apply what they had learned to enhance their instruction and assessment practices.

Discussion

Our study findings are consistent with many other studies focused on supporting teachers in developing their assessment practice, in that we found both areas of growth and persistent struggle. For example, we found that engaging teachers in the design of assessments in accord with a particular vision for learning could improve the quality of teachers' designed tasks (cf., Furtak et al., 2018). At the same time, as others have found, the alignment of assessments designed by teachers does not always address each of the substantive disciplinary ideas and practices targeted in standards or learning progressions (Coffey et al., 2011; Furtak et al., 2018).

In addition, just as we supported greater attunements to the importance of grounding assessments of student learning in meaningful contexts, other projects have demonstrated some success in this area as well, while also noting how challenging it can be to recognize and affirm students' diverse modes of sensemaking (Coffey et al., 2005; Kang & Furtak, 2021). In contrast to previous studies (e.g., Furtak et al., 2016; Penuel et al., 2017), we did not examine impacts of changed practice on student learning outcomes, so we cannot compare our professional learning design to those studies.

Inspired by ways that our own professional learning was less comprehensive than other designs that seek to support not only the design but also use of assessments to inform teaching, we sought to revise our own designs a third time for our study. Originally, our argument had been that a brief workshop could address a practical need for a workshop that could be offered within the typical constraints of schools and districts. Indeed, we found that our workshop could shift the quality of teachers' assessment designs, but alignment remained an issue for teachers, and teachers were not given the opportunity to design a task from start to finish. Nor were they invited to consider interest and identity as co-equal dimensions to the 3Ds of science proficiency emphasized in the standards. Thus, when given the opportunity to redesign our course, while we remained concerned about the practicalities of the workshop—by offering it online, in smaller-sized chunks of time—we opted to ground the redesign in a more expansive view of assessment in relation to teaching and student interest and identity that was consistent with our own commitments and prior research.

From our perspective, it is an open question how valuable an extended professional learning experience focused on the *design* of tasks can be. To address the question of efficacy, we have just concluded recruitment for an experimental study to test the efficacy of our redesigned course against a “business as usual” condition, where half the teachers will be randomly assigned to a workshop condition, and half will have access to tools that are at their own disposal to design 5D assessments, but not receive professional learning until a later time. As in previous studies, we will examine impacts on the quality of teachers’ assessments, but we will also look for impacts on instruction and also on how engaged students are by the assessments created by their teachers.

Another open question for us is the larger purpose of professional learning that prepares teachers to design assessments that address all five dimensions of science learning we have identified. Teachers have remarked to us how intensive the process of design is, particularly steps relating to the analysis of standards and choosing phenomena (Lo et al., 2022). It is certainly not reasonable to expect teachers to follow the evidence-centered approach experienced during the course in full each time they need produce an assessment; rather, we want to use the learner experience and tools to help teachers appreciate and understand the vision for learning in the *Framework* and attend to important considerations when developing, selecting, or adapting assessments. Of course, such a goal demands that there be more such assessments available to choose from, and the supply remains limited. Thus, there is demand for the courses like ours to develop teachers’ *pedagogical design capacity* to create and adapt assessments that align with the 5D vision, develop teachers’ practices, and improve student learning (Brown, 2002).

Conclusion

Projects like ours reveal just how ambitious the vision of *A Framework for K-12 Science Education* (National Research Council, 2012). Where previous standards in science—and nearly every other discipline's contemporary standards—have focused principally on key ideas, the NGSS focus on three dimensions that are expected to be supported by teachers in an integrated fashion. All 3Ds are also intended to be assessed in a way so as to allow students to use the dimensions to make sense of phenomena and problems that are meaningful to them and important to their communities (National Research Council, 2014). Assessment is but one aspect of practice that must shift to realize the vision of the *Framework,* though it is an important one. Research should continue to explore how best to support educators and researchers working together to figure out just how to bring about these shifts.

References

Achieve. (2018). *Science task screener*. Author.

Bang, M., Brown, B. A., Calabrese Barton, A., Rosebery, A., & Warren, B. (2017). Toward more equitable learning in science: Expanding relationships among students, teachers, and science practices. In C. Schwarz, C. Passmore, & B. J. Reiser (Eds.), *Helping students make sense of the world using next generation science and engineering practices* (pp. 33-58). NSTA.

Bell, P. (2019). Infrastructuring teacher learning about equitable science instruction. *Journal of Science Teacher Education*, *30*(7), 681-690. https://doi.org/10.1080/1046560X.2019.1668218

Bell, P., Stromholt, S., Penuel, W. R., & Van Horne, K. (2016, March). *How to make science instruction compelling to students: Designing formative assessments to build on learners' interests and knowledge* National Science Teachers Association (NSTA) Annual Meeting, Nashville, TN.

Black, P., & Harrison, C. (2001). Feedback in questioning and marking: The science teacher's role in formative assessment. *School Science Review*, *82*(301), 55-61.

Brown, M. W. (2002). *Teaching by design: Understanding the intersection between teacher practice and the design of curricular innovations* [doctoral dissertation, Northwestern University]. Evanston, IL.

Bucciarelli, L. L. (1994). *Designing engineers*. MIT Press.

Calabrese Barton, A., & Tan, E. (2010). 'We be burnin'! Agency, identity, and science learning. *Journal of the Learning Sciences*, *19*(2), 187-229.

Coffey, J., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, *48*(10), 1109-1136. https://doi.org/10.1002/tea.20440

Coffey, J. E., Sato, M., & Thiebault, M. (2005). Classroom assessment up close – and personal. *Teacher Development*, *9*(2), 169-184. https://doi.org/10.1080/13664530500200246

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). Routledge.

Esteban-Guitart, M., & Moll, L. (2014). Funds of Identity: A new concept based on the Funds of Knowledge approach. *Culture & Psychology*, *20*(1), 31-48.

Fine, C., & Furtak, E. M. (2020). A framework for science classroom assessment task design for emergent bilingual learners. *Science Education*, *104*(3), 393-420. https://doi.org/10.1002/sce.21565

Furtak, E. M., Circi, R. K., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results of a four-year study. *Applied Measurement in Education*, *31*(2), 143-156. https://doi.org/10.1080/08957347.2017.1408624

Furtak, E. M., Kang, H., Pellegrino, J. W., Harris, C. J., Krajcik, J. S., Morrison, D., Bell, P., Lakhani, H., Suárez, H., Buell, J. Y., Nation, J., Henson, K., Fine, C., Tschida, P., Fay, L., Biddy, Q., Penuel, W. R., & Wingert, K. (2020). Emergent design heuristics for three-dimensional classroom assessments that promote equity In M. Gresalfi & I. S. Horn

(Eds.), *The interdisciplinarity of the learning sciences* (Vol. 3). International Society of the Learning Science.

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de Leon, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: findings from a four-year intervention study. *Instructional Science*, *44*(3), 267-291. https://doi.org/doi.org/10.1007/s11251-016-9371-3

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

Herman, J. L., & Webb, N. M. (2007). Alignment methodologies. *Applied Measurement in Education*, *20*(1), 1-5. https://doi.org/10.1080/08957340709336727

Holland, D., & Lave, J. (2009). Social practice theory and the historical production of persons. *Actio: An International Journal of Human Activity Theory*(2), 1-15.

Hulleman, C., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(1410-1412). https://doi.org/10.1126/science.1177067

Hunsu, N. J., Adesope, O., & Van Wie, B. J. (2017). Engendering situational interest through innovative instruction in an engineering classroom: what really mattered? *Instructional Science*, *45*(6), 789-804. https://doi.org/10.1007/s11251-017-9427-z

Kang, H., & Furtak, E. M. (2021). Learning theory, classroom assessment, and equity. *Educational Measurement: Issues and Practice*, *40*(3), 73-82. https://doi.org/10.1111/emip.12423

Kapon, S., & Merzel, A. (2019). Content-specific pedagogical knowledge, practices, and beliefs underlying the design of physics lessons: A case study. *Physical Review Physics Education Research*, *15*(1), 010125.

Kelly, G. J. (2012). The social bases of disciplinary knowledge and practice in productive disciplinary engagement. *International Journal of Educational Research*, *64*, 211-214.

Kluger, A. N., & deNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284. https://doi.org/10.1037/0033-2909.119.2.254

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kolodner, J. L., Gray, J. T., & Fasse, B. B. (2003). Promoting transfer through case-based reasoning: Rituals and practices in Learning by Design classrooms. *Cognitive Science Quarterly*, *3*(2), 119-170.

Krajcik, J. S., Codere, S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the Next Generation Science Standards. *Journal of Science Teacher Education*, *25*(2), 157-175. https://doi.org/10.1007/s10972-014-9383-2

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Harvard University Press.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and pesources: State NAEP 1990–2009 Reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209-231. https://doi.org/10.3102/0162373711431604

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 371-388). Cambridge University Press.

Llopart, M., & Esteban-Guitart, M. (2017). Strategies and resources for contextualising the curriculum based on the funds of knowledge approach: a literature review. *Australian Education Research Journal*, *44*, 255-274.

Lo, A. S., Glidewell, L., O'Connor, K., Allen, A., Herrmann-Abell, C. F., Penuel, W. R., Wingert, K., & Lindsay, W. (2022). Promoting shifts in teachers' understanding and use of phenomena in instruction and assessment. International Conference of the Learning Sciences, Hiroshima, Japan.

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education*, *96*(6), 1071-1105. https://doi.org/10.1002/sce.21030

Manz, E. (2015). Representing argumentation as functionally emergent from scientific activity. *Review of Educational Research*, *85*(4), 553-590. https://doi.org/10.3102/0034654314558490

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Miller, E. C., Manz, E., Russ, R. S., Stroupe, D., & Berland, L. (2018). Addressing the epistemic elephant in the room: Epistemic agency and the Next Generation Science Standards. *Journal of Research in Science Teaching*, *55*(7), 1053–1075. https://doi.org/10.1002/tea.21459

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In F. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110-128). IPN.

Morales-Doyle, D., Price, T. C., & Chappell, M. J. (2019). Chemicals are contaminants, too: Teaching appreciation and critique of science in the era of Next Generation Science Standards. *Science Education*, *103*(6), 1347-1388. https://doi.org/10.1002/sce.21546

Nasir, N. i. S. (2002). Identity, goals, and learning: Mathematics in cultural practice. *Mathematical Thinking and Learning*, *4*(2&3), 213-247.

Nasir, N. i. S., & Cooks, J. (2009). Becoming a hurdler: How learning settings afford identities. *Anthropology and Education Quarterly*, *40*(1), 41-61.

Nasir, N. i. S., & Hand, V. (2008). From the court to the classroom: Opportunities for engagement, learning, and identity in basketball and classroom mathematics. *Journal of the Learning Sciences*, *17*(2), 143-179.

National Academies of Sciences Engineering and Medicine. (2018). *How people learn II: Learners, cultures, and contexts*. National Academies Press.

National Academies of Sciences Engineering and Medicine. (2019). *Science and engineering for grades 6-12: Investigation and design at the center*. The National Academies Press. https://doi.org/doi:10.17226/25216

National Research Council. (1996). *National Science Education Standards*. National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards*. National Academy Press.

National Research Council. (2001). *Knowing what students know*. National Academies Press. https://doi.org/10.17226/10019

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. National Academies Press. https://doi.org/10.17226/1840

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. National Academies Press.

Odden, O. B., & Russ, R. S. (2019). Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science Education*, *103*(1), 187-205. https://doi.org/10.1002/sce.21452

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*, 320-323. https://doi.org/10.1126/science.1232065

Penuel, W. R., Allen, A.-R., Henson, K., Campanella, M., Patton, R., Rademaker, K., Reed, W., Watkins, D. A., Wingert, K., Reiser, B. J., & Zivic, A. (2022). Learning practical design knowledge through co-designing storyline science curriculum units. *Cognition and Instruction*, *40*(1), 148-170. https://doi.org/10.1080/07370008.2021.2010207

Penuel, W. R., Bell, P., Neill, T., Morrison, D., & Tesoriero, G. (2018). *ACESSE Resource E: Selecting compelling anchoring phenomena for equitable science teaching*. STEM Teaching Tools, University of Washington.

Penuel, W. R., DeBarger, A. H., Boscardin, C. K., Moorthy, S., Beauvineau, Y., Kennedy, C., & Allison, K. (2017). Investigating science curriculum adaptation as a strategy to improve teaching and learning. *Science Education*, *101*(1), 66-98. https://doi.org/10.1002/sce.21249

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in Earth science: A comparison of three professional development programs. *American Educational Research Journal*, *48*(4), 996-1025. https://doi.org/10.3102/0002831211410864

Penuel, W. R., & Horne, K. V. (2016). Prompts for Integrating Crosscutting Concepts Into Assessment and Instruction. http://stemteachingtools.org/brief/41

Penuel, W. R., Lo, A. S., Jacobs, J. K., Gardner, A., Stuhlsatz, M. A. M., & Wilson, C. (2019, April). *Tools for supporting teachers to build quality 3D assessment tasks* NARST Annual Meeting, Baltimore, MD.

Penuel, W. R., Reiser, B. J., Novak, M., McGill, T., Frumin, K., Van Horne, K., Sumner, T., & Watkins, D. A. (2018, April). *Using co-design to test and refine a model for three-dimensional science curriculum that connects to students' interests and experiences* Annual Meeting of the American Educational Research Association, New York, NY.

Penuel, W. R., & Shepard, L. A. (2016a). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching* (pp. 787-851). AERA.

Penuel, W. R., & Shepard, L. A. (2016b). Sociocognitive and sociocultural models of assessment. In A. A. Rupp & J. P. Leighton (Eds.), *Cognitive Models of Assessment* (pp. 146-173). Wiley.

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

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *Annals of the American Academy of Political and Social Science*, *683*, 201-216. https://doi.org/10.1177/0002716219843249

Pickering, A. (1995). *The mangle of practice: Time, agency, and science*. University of Chicago Press. Publisher description http://www.loc.gov/catdir/description/uchi052/94044546.html

Table of contents http://www.loc.gov/catdir/toc/uchi051/94044546.html

Puinean, A. M., Foster, S. P., Oliphant, L., Denholm, I., Field, L. M., Millar, N. S., Williamson, M. S., & Bass, C. (2010). Amplification of a cytochrome P450 gene is associated with resistance to neonicotinoid insecticides in the aphid *Myzus persicae*. *PLOS Genetics*, *6*(6), e1000999. https://doi.org/10.1371/journal.pgen.1000999

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, *13*(3), 273-304. https://doi.org/10.1207/s15327809jls1303_2

Remillard, J. T. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry*, *29*(3), 315-342.

Renninger, K. A., Ewen, L., & Lasher, A. K. (2002). Individual interest as context in expository text and mathematical word problems. *Learning and Instruction*, *12*(4), 467-490.

Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 173-195). Academic Press.

Rogoff, B., Baker-Sennett, J., Lacasa, P., & Goldsmith, D. (1995). Development through participation in sociocultural activity. In J. Goodnow, P. Miller, & F. Kessel (Eds.), *Cultural practices as contexts for development*. Jossey-Bass.

Rogoff, B., Moore, L., Najafi, B., Dexter, A., Correa-Chavez, M., & Solis, J. (2007). Children's development of cultural repertoires through participation in everyday routines and practices. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of socialization: Theory and research* (pp. 490-515). Guilford Press.

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal Design for Learning*. ASCD.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, *44*(1), 57-84. https://doi.org/10.1002/tea.20163

Severance, S., Penuel, W. R., Sumner, T., & Leary, H. (2016). Organizing for teacher agency in curriculum design. *Journal of the Learning Sciences*, *25*(4), 531-564. https://doi.org/10.1080/10508406.2016.1207541

Sherin, M. G., & van Es, E. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education*, *13*(3), 475-491.

Stromholt, S., Van Horne, K., Bell, P., Penuel, W. R., Neill, T., & Shaw, S. (2017). How to assess three-dimensional learning in your classroom: building assessment tasks that work.

OER Professional Development Session from the ACESSE Project. http://stemteachingtools.org/pd/sessionb

Taylor, C. S. (2022). *Culturally and socially responsible asssessment: Theory, research, and practice* Teachers College Press.

Van Horne, K., Penuel, W. R., & Bell, P. (2016). Integrating Science Practices Into Assessment Tasks. *STEM Teaching Tools: Teaching Tools for Science, Technology, Engineering and Math (STEM) Education*. http://stemteachingtools.org/brief/30

Voogt, J. M., Westbroek, H., Handelzaltz, A., Walraven, A., McKenney, S. E., Pieters, J. M., & de Vries, B. (2011). Teacher learning in collaborative curriculum design. *Teaching and Teacher Education*, *27*, 1235-1244.

Walkington, C., & Bernacki, M. L. (2014). Motivational interventions as motivating students by "personalizing" learning and individual interests: A consideration of theory, design, and implementation issues. In *Advances in motivation and achievement* (pp. 139-176). Emerald Group Publishing Limited.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, *92*(6), 941-967. https://doi.org/10.1002/sce.20259

Wortham, S. E. (2004). The interdependence of social identification and learning. *American Educational Research Journal*, *41*(3), 715-749.

Wortham, S. E. (2006). *Learning identity: The joint emergence of social identification and academic learning*. Cambridge University Press.