Tools for Supporting Teachers to Build Quality 3D Assessment Tasks

William R. Penuel[1]

Abraham S. Lo[2]

Jennifer K. Jacobs[1]

April Gardner, Molly A.M. Stuhlsatz, Christopher D. Wilson[2]

[1]*University of Colorado Boulder*

[2]*BSCS Science Learning*

Paper to be presented at NARST Annual Meeting, April 2019

Baltimore, MD

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

Tools for Supporting Teachers to Build Quality 3D Assessment Tasks

The *Framework for K-12 Science Education* (National Research Council, 2012) presents a vision of science proficiency that emphasizes the need for science education to develop students' understanding of disciplinary core ideas, science and engineering practices, and crosscutting concepts in science. The *Framework* proposes integrating these dimensions to make science and engineering more meaningful to students by engaging them in science and engineering practices to develop and apply targeted science ideas (Schwarz, Passmore, & Reiser, 2017). This vision is embodied in the three-dimensional standards of the Next Generation Science Standards (NGSS Lead States, 2013). For all students to meet these new standards, most teachers will need to shift instruction significantly, away from teaching discrete facts and toward supporting the development of student understanding over time though investigating phenomena and solving problems (National Academies of Sciences Engineering and Medicine, 2018).

The shifts in science teaching and learning must also be reflected in assessments used in classrooms. New three-dimensional assessments will require teachers to elicit students' proficiency in using science and engineering practices and applying their understanding of core ideas and crosscutting concepts in the context of explaining phenomena and solving problems (National Research Council, 2014; Pellegrino, 2013). Further, demonstrating proficiency cannot be accomplished through multiple-choice items alone, since these do not require students to actively construct and integrate knowledge through engaging in practices in the ways called for in the NGSS (Lee, Liu, & Linn, 2011; National Research Council, 2014, p. 6; Nehm, Beggrow, Opfer, & Ha, 2012). Needed are multicomponent tasks—that is, tasks with questions linked to a common scenario—focused on having students construct explanations of phenomena or solve

design challenges (National Research Council, 2014). Such tasks are needed to elicit how well students' reasoning and sensemaking approximates aspects of how scientists and engineers really work (Manz, 2015a, 2015b).

A key challenge for the field is preparing teachers to develop three-dimensional assessment tasks for use in their own classrooms. Many teachers' vision for assessment involves assessing isolated facts (or content knowledge) without requiring students to also demonstrate their grasp of practices or understanding of crosscutting concepts (Weidler-Lewis, Penuel, & Van Horne, 2017). Thus, there is a need to support shifts in *vision* with respect to what they expect students to know and be able to demonstrate on assessments. In addition, there is a need to promote shifts in the *form* of assessments toward more multicomponent tasks. This is likely to require significant change, since less than a third of teachers regularly give tests that include constructed response questions (Banilower et al., 2013). Needed are tools that can help teachers develop and score assessments efficiently within the context of an already challenging workload.

This paper reports on a design study investigating the potential of a focused set of tools and designed learning experiences for teachers to help them develop three-dimensional science assessments. In the study, we examined shifts in teachers' vision for science assessment and the qualities of their assessment for two groups of teachers: those who received tools only and those who also participated in a two-day workshop to introduce them to the tools. The tools tested were peer reviewed by experts in the *Framework* and assessment, and they provided scaffolds both for the process of assessment design and for integration of science and engineering practices and crosscutting concepts into the assessments. In our study, we found modest shifts in assessment practice for both groups of teachers, but greater shifts for the group that participated

in the workshop, specifically with respect to the use of scenarios that anchored assessment tasks and integration of the three dimensions.

Theoretical Framework

Many scholars have proposed *assessment literacy* as an important target for teacher professional development over the past two decades (see, Xu & Brown, 2016, for a review). Broadly defined, assessment literacy refers to proficiency with basic concepts and practices of gathering, interpreting, and using evidence of student learning obtained from tasks and activities purposely designed to elicit student understanding (Popham, 2009). Focusing on improving classroom assessment literacy is particularly important, because it can directly shapes both teaching and learning outcomes (Brookhart, 2002). In science education, scholars have proposed that a core component of assessment literacy is an understanding of a vision for how best to support science learning (Abell & Siegel, 2011). However, professional development to develop assessment literacy linked to the vision of the *Framework* is just now beginning to emerge, and much of the research on assessment literacy focuses on assessment literacy goals but not the mechanisms for supporting teacher learning. This study aims to begin to address these gaps.

The tools and workshop that are the focus of the current study are developed from the premise that *targeted scaffolds* are needed to support the development of teachers' capacity to design assessments of three-dimensional science learning. As an act, scaffolding refers to the strategic assistance that a more capable peer or teacher provides to a learner that enables them to perform a task independently that they would otherwise not be able to perform on their own (Wertsch, 1985; Wood, Bruner, & Ross, 1976). Scaffolds can also be embedded in tools and artifacts in ways that support this same general goal, that is, helping individuals perform a

complex task for which they have not been prepared to perform (Quintana et al., 2004; Reiser & Tabak, 2014; Reiser et al., 2001).

The tools we have developed are scaffolds primarily for *adaptation* of existing assessments, though in professional development activities, we also incorporated resources and routines to support the design of new assessments. As adaptation scaffolds, the focus was on supporting teachers in integrating under-assessed dimensions of proficiency into existing assessments, namely practices and crosscutting concepts. In professional development, we provided partially elaborated scenarios to use to support integration, allowing teachers to choose between adapting existing assessments or designing new ones using the adaptation scaffolds provided.

Scaffolds for Developing Assessments for a Given Learning Goal

Over the years, researchers investigating how to improve teachers' assessment literacy have introduced a number of different types of scaffolds to support the design and analysis of assessment tasks. One line of research focuses on scaffolds for using the framework of claim, evidence, and reasoning (CER) to assess students' proficiency in the science practice of engagement in argument from evidence (Hillocks, 2012; McNeill, 2009; McNeill, Katch-Singer, & Pelletier, 2015; McNeill & Krajcik, 2009; McNeill & Krajcik, 2012). Another line of research has investigated the role of conceptual and practical scaffolds to help pre-service and early-career teachers analyze student work products, in order to support teachers' inquiry into the relationship between their instructional practice and student learning (Kang, Thompson, & Windschitl, 2014; Windschitl, Thompson, & Braaten, 2011). Still other research has investigated how teachers can use learning progressions for specific disciplinary core ideas to design tasks and analyze student work (Furtak, 2012).

In each of these lines of work, the approach to supporting teachers involved more than just providing teachers with tools to use on their own. Teachers were introduced to tools through formal professional development workshops (McNeill & Knight, 2013) or worked in professional learning communities that were partly facilitated by researchers (Furtak & Heredia, 2014). Moreover, in some cases, the tools were embedded within specific routines for their use within teacher teams (e.g., Windschitl et al., 2011), and that improvements to assessment literacy are best attributed to participation in these routines, rather than the use of tools alone (Furtak et al., 2016).

In addition, a common feature of these approaches is that the scaffolds embedded assumptions about the kind of disciplinary learning that was to be supported and assessed in the classroom. That is to say, the scaffolds were not simply practical tools; they were conceptual or ideational tools that embed ideas about important learning goals and possible routes for students to attain them (Windschitl et al., 2011). This is significant, because the dual practical-conceptual nature of the tools is critical for helping teachers to solve the perennial "what next" problem in classroom assessment: these tools provide some broad guidance to teachers about how to help teachers support students' development of disciplinary core ideas or science and engineering practices. Evidence from a range of studies suggests that discipline specific-theories of learning underlie effective interventions to improve assessment literacy and student learning (Penuel & Shepard, 2016).

Of course, many well-designed tools "travel" today across the Internet and circulate among teachers in their teacher teams and at professional conferences, and teachers who receive them do not have the benefit of extensive professional development. Not only do such tools travel

widely today, but we know that across the country, about one quarter of teachers report they use free resources from the internet in their teaching (Banilower et al., 2018). Yet, we know little about how those tools are used and whether they can support the kinds of shifts that are required of teachers in contemporary standards, particularly when the shifts entail a significant turn toward disciplinary practices (for a discussion of the practical challenges of the 'practice turn' in science education, see Furtak & Penuel, 2018).

Developing Scaffolds for Task Design That Are Designed to Travel

In science education, few resources designed to travel have been studied systematically, in part because of the difficulty of doing so; however, there exists at least one example of education research where researchers purposefully designed and studied uses of tools designed to support independent assessment task design (Frezzo, Behrens, & Mislevy, 2010). The Cisco Networking Academy is a world-wide program intended to support the development of computer skills for setting up and maintaining computer networks, particularly in communities and regions that are economically disadvantaged. More than 9,000 educational institutions in over 170 countries make use of its resources to offer learning opportunities to students. A team of researchers worked with Cisco and some institutional partners to develop Packet Tracer, a set of tools that could support distributed task design, sequencing of learning opportunities, and assessments across the different educational institutions. The tools included different pre-made resources and formats to help instructors structure assessments that could elicit evidence most important to their local learning goals.

The Cisco Networking Academy project leaders used a process known as evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 1999) to develop these tools for instructors. ECD

is a principled approach to assessment design that provides a framework for developing assessments that are grounded in the idea that assessment is a form of argument from evidence (Mislevy, Haertel, Riconscente, Rutstein, & Ziker, 2017). ECD of assessments begins always with an analysis of the domain to be assessed, that is, an analysis of both the learning goals and known processes for supporting their attainment that relies on available research and on the perspectives of stakeholders (Mislevy & Haertel, 2006). On the basis of domain analysis, assessment developers construct a set of *design patterns* that specify a range of elements, including the features that should be in all assessments of particular learning goals and some features that may be varied across tasks to elicit different kinds of information. In the CISCO Networking Academy, design patterns provided to instructors included activity structures for helping students develop networking concepts, build discrete skills such as assembling the network, as well as design challenges and troubleshooting scenarios that presented more open-ended problems to solve, with guidance as to how to vary scaffolding within tasks for learners with different levels of experience (Frezzo et al., 2010). Thus, these design patterns—as all design patterns do within ECD, embed some conception of learning, as well as ideas about how best to elicit evidence related to learning, within a specific domain (Mislevy, 2003, 2007; Mislevy, Riconscente, & Rutstein, 2009). As all design patterns do, moreover, these design patterns demonstrate the potential for re-use in different contexts, as well as generativity—that is, to create new assessment tasks. Studies of the system suggest it is highly usable for instructors, and that the assessments created by instructors can produce usable and accurate descriptions of networking skills (Frezzo, DiCerbo, Behrens, & Chen, 2014).

Notably, the Packet Tracer tools were meant to be used *independently* by instructors in a wide variety of contexts, unlike ways that tools for assessment design have been used in studies in science education. They provide a useful existence proof for the possibility of tools based on design patterns that carry conceptual (learning ideas) and practical (how to structure assessment tasks) guidance being used effectively at scale. It is an open question, however, whether such tools could be used in our field, and especially under conditions when design patterns may reflect significantly new images of teaching and learning to help teachers design assessments.

Processes for Professional Development: Noticing, Analyzing, and Adapting

In our study, we sought to understand the added value of a two-day professional development workshop designed to support teachers' effective use of the scaffolds for assessment design we provided them. The structured activities we designed to support teacher learning within the workshop relied on mechanisms of teacher learning that have been studied in the past: noticing key features of tasks, analyzing tasks, and adapting tasks.

Noticing. Learning research also suggests that key to helping teachers notice the distinctive features of tasks and their enactment will be to present them with a set of "contrasting cases," that is, a set of tasks that vary with respect to key features. According to Bransford & Schwartz (1999), "experiences with contrasting cases can affect what one notices about subsequent events and how one interprets them, and this in turn can affect the formulation of new hypotheses and learning goals" (p. 70). Noticing can also affect participation in and use of tools in future activity (Sinha et al., 2010). Therefore, when selecting tasks, teachers need to be presented with a range of possible opportunities that allow them to discern salient features, such as the use of practices to explain core ideas and prompts for students to reflect on crosscutting concepts.

Noticing—supported through various means such as video clubs and analysis of student work—has been a key mechanism for promoting teacher development in mathematics for supporting closer attention to student thinking (e.g., Sherin, Jacobs, & Philipp, 2010) and for attending to equity (e.g., van Es, Hand, & Mercado, 2017).

In science education, Lo (2017) investigated how engaging with teachers in cogenerative dialogues (Tobin, 2006) to analyze one's classroom practice, using classroom video and reflections from classroom observations, and support teachers' learning to notice epistemic dimensions of scientific practice that were important for supporting students' meaningful engagement in scientific practice -- characterized by students understanding of what they are doing and how their actions and decisions will help them achieve their scientific goals (Berland et al., 2016). Central to this work involved teachers cast an ideal vision for what they desired science teaching and learning to look like in their classroom (Hammerness, 2006), identifying ways in which actual classroom practice aligned with this vision, and cogenerating with stakeholders ways to refine and make progress in realizing this vision. The epistemic features served as lenses through which the teacher and researcher co-constructed understandings of classroom events and used our interpretations to make decisions that enhanced her students' knowledge-building role and the meaningfulness of their engagement in scientific practice. Embedded in this professional learning involved fading scaffolds that supported teachers' ability to notice, but maintained the teacher's independence to make classroom decisions.

A recent study by Penuel, Wingert, and Van Horne (2018) found that professional development activities focused on helping teachers notice key features of three-dimensional assessment tasks could help them identify new qualities of tasks. The study involved 99

elementary and secondary teachers who participated in six days of professional development over the course of a year, including ones designed specifically to support teacher noticing. The study found modest shifts in the kinds of features teachers noticed: more focused on features such as phenomena and on attributes they thought differentiated good from weak assessments, notably accessibility and interest to students. It found greater shifts, though, in teachers' assessment practice: a significantly higher percentage of teachers reported assessing a practice other than explanation at posttest than at pretest.

Task analysis. There is strong evidence that professional learning experiences organized around task analysis can shift what teachers notice about the affordances of particular tasks presented to students. Task analysis can help teachers discern, for example, the level of cognitive demand of tasks and how these relate to student learning opportunities (Boston, 2013). Analyzing tasks can also help teachers discern opportunities for students to engage in disciplinary practices while solving problems and attune to the language demands of tasks (Johnson, Severance, Penuel, & Leary, 2016). Task analysis features strongly in professional development provided by Windschitl and colleagues (2011) described above. In that professional development, teachers focus on analysis of student work, but they also discuss ways that the work students produce reflects the kinds of tasks students are assigned and teachers' moves to support student learning. Similarly, in professional development with teacher teams developed by Furtak and colleagues (Furtak & Heredia, 2014; Furtak, Morrison, & Kroog, 2014), task analysis features strongly, as does task adaptation, as a mechanism for supporting learning.

Task adaptation. A number of studies point to the value of adapting tasks that have already been designed as a tool for supporting 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, Gallagher, & Moorthy, 2011). Task adaptation in the context of professional development 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 (DeBarger et al., 2017; Remillard, 1999; Voogt et al., 2011).

The Current Study

In the current study, we address four different research questions:

1. To what extent does the use of tools for task adaptation support shifts in teachers' vision for science teaching and learning?
2. What is the added value of professional development in supporting shifts in teachers' vision for science teaching and learning?
3. To what extent does the use of tools for task adaptation result in assessments that are likely to elicit responses that allow teachers to draw inferences about students' 3D science learning?
4. What is the added value of professional development for improving the quality of teachers' assessments?

Here, we report on the first phase of a design-based research study, that is, a study of an initial version of our tools and professional development. We are currently in the second phase of this study, having iterated upon both our tools and professional development workshop on the basis of study findings reported here, as well as on the basis of *micro-cycles* of design and

revision (Gravemeijer & Cobb, 2013) supported by reflective activities of the members of the research team who led or observed teacher activities.

Population and Sample

Six teachers took part in face-to-face PD workshops. These participants were all currently teaching high school science in a school district that has partnered with our research team for the past several years on the development and implementation of science curriculum. The teachers all had experience teaching this curriculum to their biology or earth science students. Over the course of the study, only four of the teachers were able to complete their participation by attending both workshops and taking part in all of the data collection described below. Therefore, we discuss the results only from these four teachers.

An additional six teachers took part in the study but did not have the option to attend PD workshops. These teachers were recruited from a cross the United States via a message sent from the first author's On their Own account. The message explained the study and offered high school teachers currently teaching biology or earth science the opportunity to participate. Seven teachers expressed an interest, and six were selected to participate. These teachers, as well as the teachers who took part in the PD, were all provided with a stipend for their participation.

The teachers who received PD were, on average, less experienced (average = 7.5 years, range = 5-12 years) than the teachers who did not attend PD (average = 12.2 years, range = 2-18 years). None of the teachers in the PD group had previously taken part in professional development based on designing 3D assessments, whereas three of the teachers in the non-PD group already had such professional learning opportunities. However, all of the teachers reported

taking part in PD over the last three years related to crosscutting concepts and planning instruction or designing materials aligned to the Next Generation Science Standards.

Description of the Tools

All teachers in the study received new versions of two tools for adapting assessment tasks to be 3D: (1) a tool for structuring tasks so as to engage students in science and engineering practices to explain a given phenomenon or solve a problem and (2) a tool for integrating questions related to relevant crosscutting concepts into assessments. The tools' design are grounded in work conducted in an earlier evaluation study that used an evidence-centered design approach to specify a set of *design patterns* for developing assessments that reflect a specified model of learning (Mislevy & Haertel, 2006), focusing on the characteristic and variable task features linked to practices and crosscutting concepts. Both sets of tools are intended to be used with tasks that present students with a phenomenon represented in a brief scenario about which they are asked a series of open-ended questions. They are published through the STEM Teaching Tools website, where they have been downloaded thousands of times by teachers since first published in 2016.

Tool for integrating science and engineering practices into assessment tasks. This tool consists of a set of design pattern components (i.e., characteristic and variable task features) for each of the eight science practices and two engineering practices (defining problems, designing solutions). Each practice includes multiple design patterns for task design described at a high level. For each practice, there are simpler and more complex blueprints aligned to levels on progressions specified in the *Framework*. For example, a simpler Asking Questions format reads, "Present students with a scientific phenomenon to be explained, then Ask students to formulate a

scientific question to investigate the phenomenon." The tool also includes a 3D assessment example.

Tool for integrating the crosscutting concepts into assessment tasks. This tool consists of a set of partially and fully developed prompts that are associated with each of the seven crosscutting concepts in the *Framework.* The prompts also provide broad direction as to when they might be used. As an example, for the concept of *Patterns,* after presenting data from an experimental study focused on isolating causal variables as part of the scenario, questions might ask: "What does the pattern of data you see allow you to conclude from the experiment?" or "Does the pattern in the data support the conclusion that _____ is caused by _____? Why or why not?"

Description of the Professional Development Workshop Series

Teachers who participated in professional development took part in a three-day workshop series in early 2018. The principal aim of the workshop series was to provide teachers with opportunities to use the scaffolds for designing three-dimensional assessments to create their own assessments. Over the course of the series, teachers had the opportunity to analyze existing tasks, practice using the scaffolds with existing scenarios that presented phenomena to be explained, and then develop their own assessments and scoring guides for assessments.

The first day focused on task analysis and adaptation. During this workshop, teachers developed and refined a set of criteria for what makes for good 3D assessments. Next, they analyzed a set of tasks that varied with respect to their dimensionality, but that were supposed to target the same performance expectation. They looked at a model three-dimensional assessment, and further refined criteria for good 3D assessments. Then, they were given a phenomenon and

scenario to use to develop a complete multi-component task using the Science and Engineering Practices Task Formats scaffold and the Crosscutting Concepts scaffold. 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. During the workshop, participants examined an existing task, the Swallows Task, developed for use with the Inquiry Hub high school biology curriculum. They looked first at the task prompts without a scoring guide and developed "claim statements" for the task, that is, a claim about what they imagined the task would allow them to conclude about what students know and can do from completing the task. Next, the participants analyzed the scoring guide and refined their claims. Finally, after looking at student work, participants refined their claims again and made suggestions for how to refine the scoring guides. These activities were intended to help provide a basis for developing scoring guides that attended to component ideas from performance expectations that were evident in student responses. Following this set of analyses, we allowed teachers to work on their own scoring guides or revise the guide for the Swallows Task to better align with both the performance expectation and with student work. Next, participants had the opportunity to work on a new assessment, again using the two key scaffolds for integrating practices and crosscutting concepts into the tools. The day concluded with a discussion of how to develop brief, 3D exit tickets.

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. Finally, participants completed the task analysis with which we started the workshop series, as a way to collectively self-assess the group's growth in understanding of 3D assessments.

Analysis

Teachers submitted assessment tasks and scoring guides prior to and after using the provided scaffolds and/or professional learning. We refer to the assessment tasks that were submitted by both groups before providing the scaffolds and/or PD as pre-assessments and after providing the scaffolds and/or PD as post-assessments. When submitting each assessment task, teachers explained the assessment goal, which could involve assessing a particular content area or performance expectation, and a brief rationale for why they submitted the assessment for feedback. We assessed the assessment goal's alignment with the NGSS, such as whether the assessment task addressed a particular performance expectation (PE), or whether it assessed content knowledge alone.

Rubric design. We iteratively designed a rubric to assess the extent to which teachers designed 3D assessment tasks that involved students using important disciplinary core ideas (DCIs), science and engineering practices (SEPs), and cross-cutting concepts (CCCs) in an integrated way to achieve a particular goal, such as explaining a phenomenon or solving a problem. The rubric has five categories aligned with key features that were the focus of the professional learning (see Appendix for rubric).

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 analyze data from the scenario and use and apply learned knowledge 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. To perform this analysis, we examined the DCIs related to the targeted content focus and used the relevant grade band endpoints from the *Framework*. In addition, we assessed the extent to which the body of assessed items were in service of achieving the targeted assessment goal. For example, a task assessing ideas related to Newton's 2nd Law of Motion, which relates a force applied on an object, its mass, and the object's acceleration, should not include items related to an object's density. From a student's perspective, it is not obvious how and why ideas related to density are important for helping students demonstrate an understanding of Newton's 2nd Law of Motion and would be an example of an idea that would distract from the identified assessment goal.

Category 3: Science and engineering practices. This category examined whether the teacher created opportunities for students to engage in the SEPs to complete the tasks that contributed toward achieving the assessment goal. For example, did students' data analysis and

interpretation support an argument for how and why a phenomenon occurred as it did (assessment goal) or was the goal simply to create the graph? Thus, we examined connections between students' engagement in the SEPs with the assessment goal. To assess whether the students' expected use of the SEPs was aligned with the NGSS and grade-level appropriate, we examined the relevant SEP elements for a particular performance expectation and the learning progressions found in Appendix F of the NGSS.

Category 4: Crosscutting concepts. Similar to Category 3, we assessed opportunities for students to use crosscutting concepts to achieve the assessment goal. We examined relevant CCC elements for a particular PE and the learning progressions found in Appendix G of the NGSS to assess whether students' expected performance was aligned with the NGSS and grade-band appropriate.

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 dimensions (DCIs, SEPs, and CCCs). Evidence of the latter came from analysis of the items themselves and the teacher's scoring guide, which made explicit the teacher's plan for assessing student understanding. The analysis of the scoring guide also examined whether teachers were looking for the right answer or whether teachers considered levels of performance related to each dimension.

Assessment Scoring. For each category, we identified elements and levels of performance for each element. For example, when assessing opportunities for students to engage in the SEPs to demonstrate understanding, example levels of performance included *no*

opportunities, *some* opportunities, and opportunities *throughout* the assessment. Whole point values were associated with each level of performance with each row having a maximum value of 1 or 2 points.

A point was awarded if teachers made explicit attempts to align their assessment task with the NGSS by identifying a target PE or identifying target DCIs, SEPs, or CCCs. If teachers identified a target PE, scorers used the foundational DCI, SEP, and CCC elements for that particular PE. If teachers did not identify a relevant PE or dimension, scorers attempted to identify opportunities for students to use the SEPs or CCCs in the assessment and used them for the basis for scoring. At a minimum, teachers reported a content focus for the assessment task (e.g., homeostasis or energy). In those cases, teachers would not receive points for intending to align the assessment task with the NGSS. However, scorers would subsequently perform a closer examination of the assessed ideas to ascertain whether the targeted scope was alignment with ideas found in the NGSS and grade-level appropriate.

The total possible number of points for a particular assessment task was 35: 8 points for Scenario, 7 points for each NGSS dimension, and 6 points for 3-D Integration. To assess pre to post changes in teachers' performance, the three points awarded for the intent to align with the NGSS were removed from the totals to investigate pre to post changes in teachers' use of the three dimensions independent of the identification of a target PE. Thus, the findings in the next section have a maximum point value of 32. Scores for each category and element were averaged together for PD and On their Own teachers and compared pre to post. Due to the sample size, we do not wish to overextend claims about the significance of the observed pre and post assessment changes. However, we use these analyses to identify the successes and challenges

that teachers faced as they designed 3D assessment tasks, consider the role that the professional learning and tools played in supporting some of these changes, and identify common issues that both groups encountered that could be areas of future support for teachers.

Most teachers submitted multiple pre and post assessment tasks. Scorers chose a single pre-assessment task and a single post-assessment task that best demonstrated the desired features of 3-D assessment. Within the assessment feedback for teachers, we provided rationales for why a particular assessment task was scored and provided brief remarks about the other assessment tasks. Two coders independently scored each assessment task and reached consensus on the final scoring. In addition, we provided formative feedback to the teachers for each element as well as overall comments. Due to the iterative rubric development, participating teachers did not receive the feedback on their pre-assessments before submitting their post-assessments. However, a subsequent cohort will receive feedback on their pre-assessments using the rubric as part of the professional learning.

Findings

Teachers submitted a range of types of pre-assessments. There were several examples of "traditional" assessment tasks that included discrete multiple choice or short answer questions. Others were scenario-based tasks that either required students to figure something out, meaning they used what they've learned to solve a different problem or explain another phenomenon, or provided a context for students to demonstrate their understanding. Still other assessment tasks involved the creation of a product that was intended to elicit students' understanding, such as creating a comic book explaining what happened when pathogens try to attack the body. These

tasks differed from scenario-based tasks in that students were demonstrating their understanding in another form, but not necessarily applying it to explain another situation.

Table 1 shows the range and average pre and post assessment scores for each group. On average, teachers' pre-assessment scores for both groups were similar (see Figure 1). Although both sets of teachers' total scores improved, PD teachers (23.5/32) had a slightly higher average post-assessment score than the On their Own teachers (21.17/32).

Table 1. Average and range of scores for On their Own and PD teachers

	On their Own (n=6)			PD (n=4)		
	Min	Max	Average	Min	Max	Average
Pre	13	25	19.17	14	27	19.5
Post	15	27	21.17	19	28	23.5

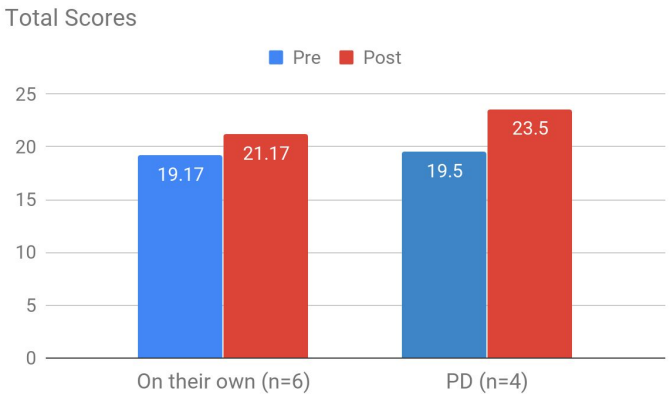


Figure 1. Comparison of average pre and post-assessment scores for On their Own and PD teachers.

Figure 2 disaggregates the findings by category. Figures 2a and 2b examine pre to post changes in each group's scores, whereas 2c and 2d compare each group's pre and post-assessment scores by category. Both groups experienced increases in the scenario, DCI, and CCC categories. The PD group's scenario scores increased more than the On their Own

group's scenario score (2.75 vs. 1.33 point gains). The On their Own group's DCI score (4.67/6) was slightly higher than the PD group's DCI score (4/6). Although the PD group had slightly higher SEP and CCC scores than the On their Own group, the On their Own group's scores increased, particularly in the CCC category, so that their performance was comparable to the PD group. There were differences in observed scores for 3-D integration category: the On their Own group's average score decreased, while the PD group's average score slightly increased. In what follows, we attempt to explain the observed trends and identify issues that teachers encountered as they sought to design coherent, scenario-based assessments that required integrated use of the three dimensions.

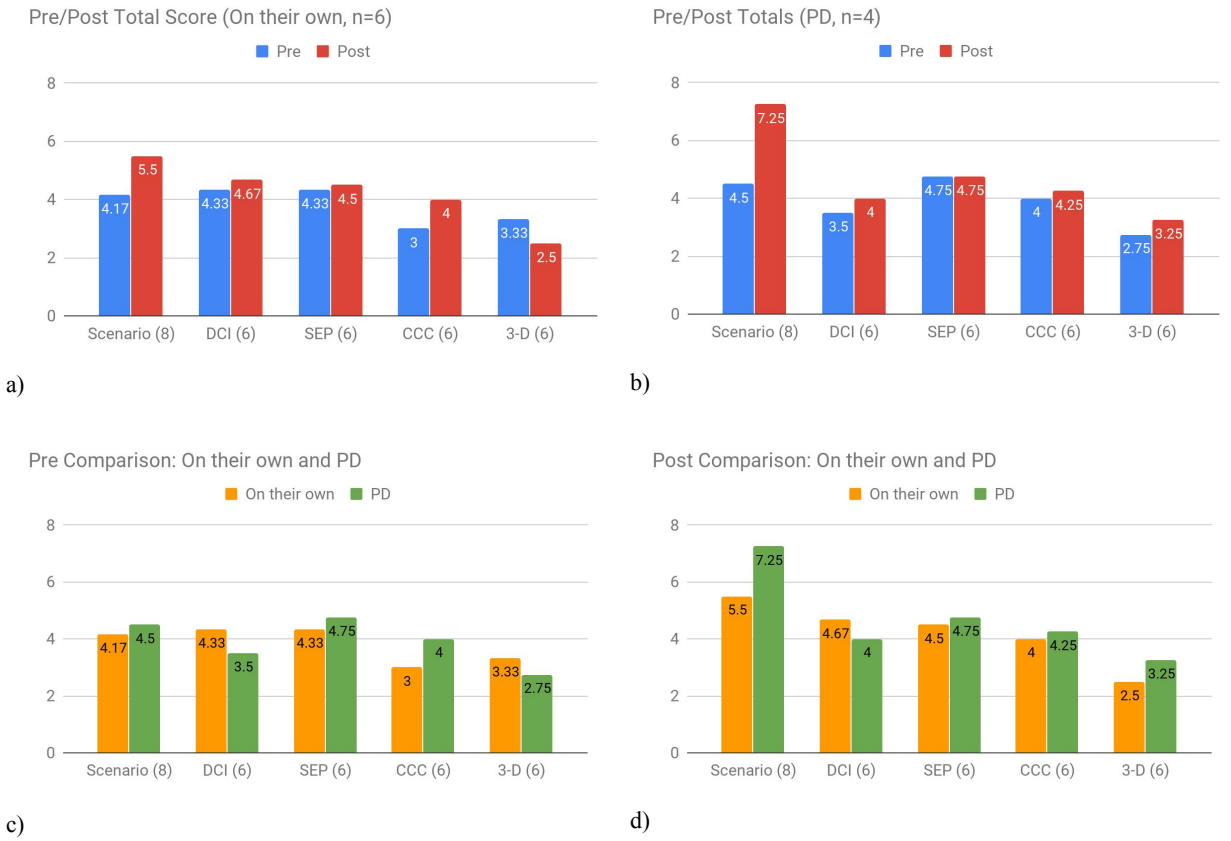


Figure 2. Comparison of pre and post-assessment scores for each category. a) and b) show changes in each group's pre and post assessment scores, whereas c) and d) compare each group's pre and post-assessment scores.

Scenario and Assessment Coherence

It was not surprising that there were increases in this category, as the identification of a phenomenon or problem to solve was a central feature of the tools that were provided to both groups and the professional learning. Both groups made greater use of real or scientific data in their scenarios. Teachers encountered issues with developing scenarios that supported students in figuring things out and that did not have an obvious solution. For example, a teacher created an assessment whereby students could piece together an explanation using information directly obtained from the assessment rather than using what they've learned to analyze and interpret information from the scenario to construct an explanation or solve a problem. The PD group had better success in identifying scenarios where students had to figure something out (average score 2.0/2.0, n=4) than the On their Own group (average score 1.0/2.0, n=6).

Teachers experienced some struggles related to constructing coherent assessments that were connected to the scenario. 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. There was a decrease in the On their Own group's pre to post assessment coherence scores (1.83 to 1.33; max score 2.0) because all but one teacher had a common assessment task that brought coherence to all items for the pre-assessment. However, these common assessment tasks did not necessarily involve a scenario. Thus, the rubric was sensitive enough to tease apart challenges involving assessment coherence and the challenges that teachers now faced with creating coherent, scenario-based assessments.

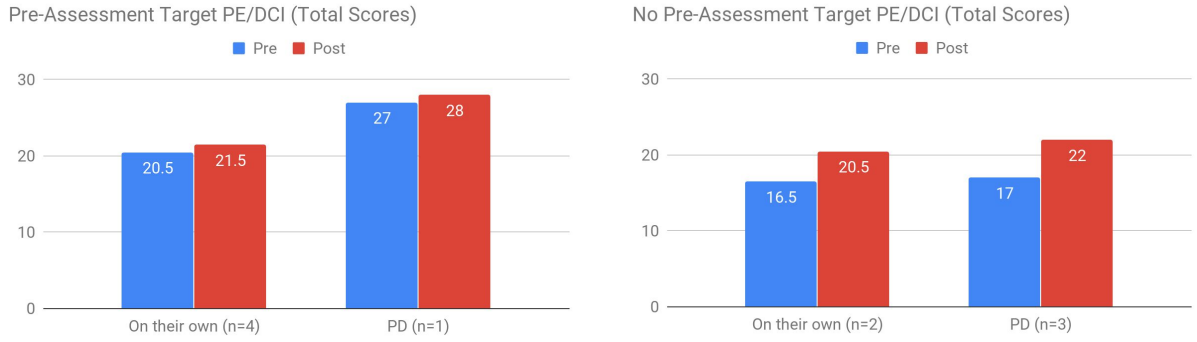
Although the scenario-based tasks provided an authentic context through which to construct knowledge, there were differences between groups in the extent to which students needed to apply their existing knowledge in new contexts, thus creating an opportunity for students to assess their learned knowledge. All of the PD post-assessments required students to apply their knowledge to new contexts (average score, 1.0/1.0). In contrast, half the On their Own group developed performance-based assessments that required students to construct explanations, but these tasks did not necessarily require students to apply their previously learned knowledge. Thus, the designed tasks were authentic in nature, but not necessarily the best venue for assessing students' existing understanding or learned knowledge.

Creating opportunities for students to demonstrate their understanding using multiple modes was an area of struggle for both groups. The most frequent mode was written responses and constructing diagrammatic representations of one's ideas, which included models or graphs. This area was not a focal feature of the PD.

Identifying Target PEs Key for Supporting 3D Alignment

We argue that teachers' identification of target PEs for their assessment contributed to increases in content alignment with the NGSS and opportunities for students to engage in the CCCs. Teachers that did not identify a target PE often identified non-NGSS standards or content foci that were not completely aligned with the NGSS or grade-band appropriate. Three of four PD teachers did not choose target PEs for their pre-assessment, whereas four of six On their Own teachers chose one. All but one participant included either a target PE or identified DCIs that were the focus of the post-assessment. The single outlier was a teacher who cited a state standard rather than a NGSS standard. When we looked at the aggregate data, we observed modest

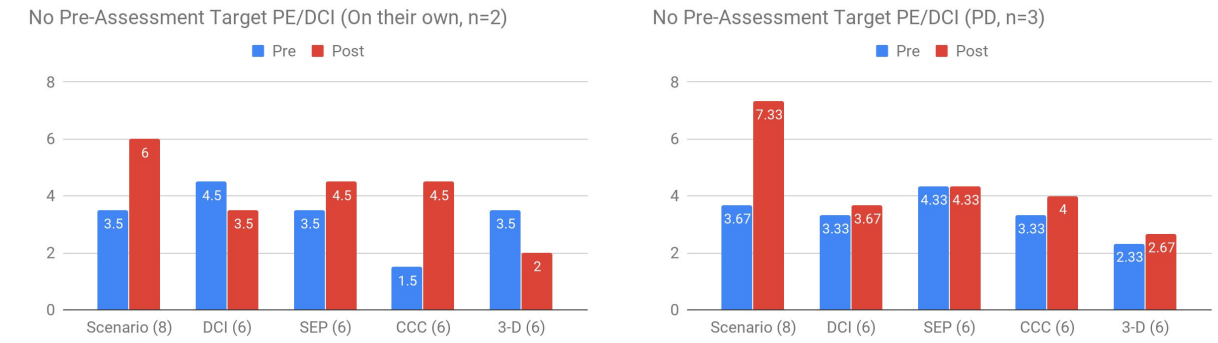
changes in teachers' DCI and SEP scores and a greater shift in teachers' CCC scores (see Figure 2). However, when we separately examined pre to post gains for teachers who identified or did not identify a target PE for their pre-assessments, there were minimal changes for teachers that had identified pre-assessment target PEs (see a Figure 3a) as compared to teachers who did not (see Figure 3b). Similar patterns occurred regardless of treatment condition. As a reminder, these scores did not include the 3 potential points for intention to assessment the three dimensions.



a) b)

Figure 3. Comparison of pre to post assessment scores for teachers that identified target PEs for their pre-assessments (a) compared to teachers that did not (b).

When we disaggregate the data by category, we see that identifying a target PE contributed to consistent increases in teachers' use of CCCs and moderate increases in use of SEPs for On their Own teachers (see Figure 4). There was an observed decrease in the On their Own group's DCI score and no changes in the PD group's average SEP score. In what follows, we examine the areas of growth and challenge that teachers faced with regards to supporting 3D assessment.



a) b)

Figure 4. Comparison of pre to post score changes by category for teachers that did not identify target PE for their pre-assessments.

DCIs. In general, the identification of one or more targeted PEs contributed to increased content alignment with the NGSS. Teachers that did not choose target PEs often assessed content that was often below grade-band or targeted ideas that did not address the full scope of a disciplinary core idea. For example, one teacher designed an assessment that examined the features of the immune system and the mechanisms involved in attacking pathogens. The assessment focus foregrounded details about a specific system and backgrounded DCIs related to the interactions between system components and how they work together to perform a given function.

Teachers had challenges with designing assessment items that thoroughly addressed all aspects of the identified DCI elements. Often, there were aspects of DCI elements that were not assessed. 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 DCI elements 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 explanation is that teachers focused more

on identifying PEs that were related to the desired content focus rather than using the DCI elements identified by the target PE to guide the assessment design. In the whale example, the teacher reported that she found PEs to be too limiting and instead chose DCIs related to the content focus of the assessment. Thus, more support might be needed to help teachers attend to the targeted DCI elements to understand the scope of the assessed content and design assessment items that adequately address these ideas. An elemental examination or unpacking of the DCI elements might help teachers the depth of knowledge required to demonstrate an understanding of a particular DCI and notice the nuances for how DCIs build on previously learned knowledge and become more sophisticated across grade-bands.

On the other hand, some teachers identified more target DCI elements or PEs than could be adequately addressed in assessment or included DCI elements that were tangentially related to the stated assessment goal. For example, the teacher who designed the whale also identified LS4.C: Adaptation as a targeted PE. However, issues related to how the external environment plays a role in leading to changes in a population were not evident in this assessment. More support might be necessary to help teachers be clear about the assessment goal and ensure that all of the assessed ideas and tasks are related to achieving that goal. Identifying one or more target PEs might be helpful for addressing this challenge, as teachers will see how the targeted DCI elements are related to the assessment goal of achieving a particular PE.

SEPs. In general, teachers were attentive to creating opportunities for students to engage in the SEPs to complete tasks, with small observed increases in opportunities for students to use SEPs with the On their Own group. We argue that increased opportunities are likely associated with increased use of coherent scenarios that require students to figure something out. However,

teachers experienced challenges with connecting the product of students' SEP with contributing to the overall assessment goal. For example, students had frequent opportunities to analyze and interpret data in both pre and post assessments. However, teachers did not always connect what students' figured out with the broader problem or phenomenon. For example, in an assessment that asked students to explain how it was possible for all of 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.

Similar to the challenges faced in the DCI category, there were challenges related to teachers designing assessment items that required below grade-band use of the SEPs or or did not require all aspects of the target SEP element. For example, high school students might be asked to engage in argumentation, but not evaluate multiple arguments, or communicate information to others without considering the *purpose* for clearly communicating information to others as the means for facilitating the persuasion of others about the reliability and validity of presented ideas. During the professional learning, teachers consulted resources, such as Appendix F in the NGSS, to see how expected student use of the SEPs should vary and build in sophistication in higher grade-bands. However, the findings of this research indicate that more in-depth activities might be needed to help teachers notice grade-band distinctions and incorporate those elements in their assessments. One potential activity might involve teachers identifying the SEP elements that contributed to a given PE so that teachers could consider how the target SEP element could be useful in service of demonstrating the target understanding and

ensure that they designed opportunities for students to engage in all aspects of the SEP element. Another activity might involve teachers peer-reviewing one another's assessments to co-construct assessments of whether students' expected use of the SEPs was grade-band appropriate and to take action to resolve identified issues.

CCCs. Although there were increased pre to post assessment opportunities for students to use CCCs to demonstrate their understanding, the expected use of CCCs was generally below grade-band in both groups. For example, when using CCCs related to systems and system models, students often examined components and relationships within systems rather than between systems. When using the CCC of cause and effect, students often were not required to distinguish between cause & correlation. Thus, the professional learning and tools might have contributed to increased awareness of the need to use CCCs, yet still require additional support to identify what level of sophistication use is required at each grade-band. As before, additional activities, such as those described in the previous sections, are likely needed to support teachers in considering the usefulness of the CCC elements for each PE and using Appendix G in the NGSS to design grade-level appropriate assessments.

Integration of 3-D. Of the five categories, this category was the lowest for each group. Although the PD group had pre to post assessment gains in their ability to design 3-D integrated tasks (average score 1.25 to 1.50/2.0), teachers continued to experience challenges in this area. Teachers' scoring guides provided a window into teachers' visions for what expected performance should look like and how they might assess it with respect to the three dimensions. Teachers' scoring was generally one dimensional and focused on DCI-related performance. There was a focus on whether students had the right answer or the correct components of an

explanation rather than considering different levels of performance. For example, scoring guides might expect students to identify connections between systems without considering how the presence or absence of these connections might contribute to a more complex understanding of the phenomenon. Thus, more support might be needed to help teachers broadly think about how each dimension works together to demonstrate a particular understanding and provide opportunities to analyze exemplar scoring guides that explicitly assess levels of performance rather than just the right answer. Thus, the analysis and iterative feedback on scoring guides might help teachers refine their vision for 3D learning and assessment and support their ability to design 3D assessments.

Teachers' Reflections on their Learning

We conducted interviews with teachers in both the PD and On their Own groups at the conclusion of the study. As part of these interviews, we asked teachers to reflect on their use of the scaffolds to design assessments and how they supported change to their assessment practice.

What teachers implemented that they found helpful. In interviews conducted after their participation in the project concluded, the four teachers who took part in the PD workshops all reported having a positive experience and anticipating using more 3D assessments in their classrooms. Two of the teachers pointed to the question stems or "sentence starters" provided on the tools as especially useful for writing various types of assessment questions. Several teachers explained that talking about and developing assessments during the workshops helped them to better understand and clarify aspects of the NGSS, for example when considering how particular questions align to specific performance expectations. Additionally, the workshops helped them to understand what it means to incorporate all three dimensions into assessment tasks, rather than

simply focusing on content. Looking at model assessment tasks in the PD was especially helpful in this regard. Most of the teachers noted that developing scoring guides or rubrics that relate back to the NGSS is still very challenging, and they would be interested in more guidance in that area.

The six teachers who did not attend the PD workshops generally reported that although they created assessment tasks based on the tools, they had not yet used those assessments with their students. Several of those teachers planned to share their newly created assessments with other teachers to get feedback and continue improving them. In other words, these teachers saw their participation largely as an initial step towards learning about and generating 3D assessments, but they also recognized that they still had a large part of the journey ahead of them. Like the teachers who attended the PD, this group found the question stems and ideas for "prompts" particularly valuable. As one teacher explained, "You look at a PE and you're like, 'Ok so I have to make an assessment that incorporates all these things. How do I do that?' So I think those [tools] are really, really beneficial to help come up with questions."

How involvement with the project changed teachers' thinking. All of the teachers who attended the workshops said that the PD solidified their understanding of 3D assessment and strengthened their belief in the importance of 3D science instruction. Most of the teachers came the workshops with rather extensive prior knowledge of NGSS and were convinced of the importance of 3D instruction and assessment. In general, these teachers felt validated in their beliefs about science teaching and learning, and felt that the workshops helped extend their knowledge of NGSS into the assessment domain. As one teacher who attended the PD workshops explained, "I feel I now have a very strong understanding of what 3D assessments

are. What I need now is just practice. I need to write them, see them, take them myself so I can understand how a student would interpret them. I really feel like I'm in a position now where I can take control of my own education and move forward."

The teachers who did not attend the workshops reported that simply using the tools was a valuable exercise, but most said the experience did not change their thinking about 3D instruction and assessment very much. However, a few teachers noted that the experience boosted their confidence, particularly in terms of creating and using 3D assessments. Other teachers talked about "making progress" in their assessment practices, especially in moving beyond traditional ways of assessing and grading their students. Several suggested that it would have been helpful to receive additional resources, such as more examples of assessment tasks and sample student work.

Discussion and Conclusion

One important finding from this study involved the important role that identifying target PEs played in helping teachers incorporate the three dimensions in their assessments. Attention to the elemental features of each dimension using resources, such as NGSS appendices, in the context of designing and analyzing assessments and student work might help address identified issues related to alignment with expected performance with grade-band dimensions with the NGSS. Although some teachers found aligning assessments with PEs to be limiting, this research suggests that PEs can provide a vision for how each dimension can work together to help students figure something out and demonstrate important understandings. Understanding how the three dimensions could be used in an integrated way might help teachers to make important decisions about how they might modify the target PE, such as foregrounding different

SEPs or CCCs, in the context of a particular assessment task. In addition, this research highlights a challenge that teachers face with designing three-dimensional scoring guides that assess students' use of all three dimensions, rather than just identifying the right answer or assessing performance related to the DCIs. The goal is not for teachers to consider each dimension as discrete entity to be assessed, but rather foregrounds the need to support and provide feedback on students' engagement in each dimension and how each dimension contributes in important ways towards explaining the target phenomenon or solving the target problem. This research demonstrates the need to provide additional examples of 3-D assessments and scoring guides to support teachers in crafting a vision for what integrated, 3-D assessment looks like and how they can incorporate these features in meaningful ways when designing their own assessments.

Increased growth in teachers' use of scenarios was linked to PD teachers participation in the professional learning, where they received support in identifying suitable scenarios and designing assessment items connected to that scenario. On their Own group teachers also experienced growth using available resources, such as STEM Teaching Tool 29. This study identifies the needs to refine such resources to include more explicit process descriptions for not only identifying productive scenarios that support students in figuring out, but also how to create three dimensional assessment items that are coherent and connected with this scenario.

Study Limitations

Our rubric was designed to accommodate teachers that might not have identified target PEs or DCIs for their assessment. In doing so, we were able to assess the alignment of their content focus with the NGSS independently from other categories, such as whether the assessed

ideas aligned with the teacher's stated assessment goal or whether the assessed items addressed all of the specified content or DCI elements. Thus, it was possible for teachers who reported a broad content focus, such as "energy," to get a higher score than teachers who identified DCI elements because it is more challenging to construct assessment items that appropriately addressed the DCI elements and contribute towards achieving the assessment goal.

Greater pre to post assessment changes in students' use of SEPs or CCCs might not be evident in cases where teachers did not identify a target PE, SEPs, or CCCs for their pre-assessments. When teachers were not explicit about what SEPs or CCCs they were targeting for assessment, we examined evidence of teachers creating opportunities for students to engage in the SEPs or CCCs to complete tasks. In those instances, students' pre-assessment SEP use might be better tied to the assessment goal than in cases where teachers were trying to assess SEPs that were specified by a target PE. Similar comments could be made about students' pre-assessment use of CCCs. Thus, there are limitations involved in the interpretation of observed pre to post assessment score differences by total score or category without appropriately examining shifts in the elements that inform the total score. With further data, weighing the scores assigned to NGSS alignment and grade-band appropriate use of the DCIs, SEPs, and CCCs might be one way of addressing this issue. However, the rubric was able to detect the challenges that teachers faced when trying to design assessment items that appropriately addressed each NGSS dimension at the elemental level and was coherent with an authentic problem or phenomenon. From a professional learning perspective, identifying these core challenges are likely to be more important for promoting the desired instructional shifts in classrooms.

References

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do? In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Dordrecht, the Netherlands: Springer.

Achieve, Inc. (2018). *Science task screener*. Washington, DC: Author.

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M., L. (2018). *Report of the National Survey of Science and Mathematics Education*. Chapel Hill, NC: Horizon Research, Inc.

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*. Retrieved from Chapel Hill, NC:

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. doi:10.1002/tea.21257

Boston, M. D. (2013). Connecting changes in secondary mathematics teachers' knowledge to their experiences in a professional development workshop. *Journal of Mathematics Teacher Education, 16*(1), 7-31.

Brookhart, S. M. (2002). What will teachers know about assessment, and how will that improve instruction. In R. W. Lizzits & W. D. Schafer (Eds.), *Assessment in educational reform: Both means and ends* (pp. 2-17). Boston, MA: Allyn & Bacon.

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.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2010). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology, 19*(2), 105-114.

Frezzo, D. C., DiCerbo, K. E., Behrens, J. T., & Chen, M. (2014). An extensible micro-world for learning in the data networking professions. *Information Sciences, 264*, 91-103.

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., & Heredia, S. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching, 51*(8), 982-1020.

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.

Furtak, E. M., Morrison, D., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education, 98*(4), 640-673.

Furtak, E. M., & Penuel, W. R. (2018). 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. doi:https://doi.org/10.1002/sce.21488

Gravemeijer, K., & Cobb, P. (2013). Design research from a learning design perspective. In J. van den Akker, K. Gravemeijer, S. E. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 73-112). New York, NY: Routledge.

Hammerness, K. (2006). Seeing through teachers' eyes: Professional ideals and classroom practices. New York: Teachers College Press.

Hillocks, G. (2012). *Teaching argument writing, grades 6-12: Supporting claims with relevant evidence and clear reasoning*. Portsmouth, NH: Heinemann.

Johnson, R., Severance, S., Penuel, W. R., & Leary, H. A. (2016). Teachers, tasks, and tensions: Lessons from a research-practice partnership. *Journal of Mathematics Teacher Education, 19*(2), 169-185.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education, 98*(4), 674-704.

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115-136.

Lo, A. S. (2017). *Epistemic aims, considerations, and agency: Lenses for helping teachers analyze and support students' meaningful engagement in scientific practices.* (doctoral), Northwestern University, Evanston, IL.

Manz, E. (2015a). Representing argumentation as functionally emergent from scientific activity. *Review of Educational Research, 85*(4), 553-590.

Manz, E. (2015b). Resistance and the development of scientific practice: Designing the mangle into science instruction. *Cognition and Instruction, 33*(2), 89-124.

McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education, 93*(2), 233-268.

McNeill, K. L., Katch-Singer, R., & Pelletier, P. (2015). Assessing science practices: Moving your class along a continuum. *Science Scope*, 21-28.

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.

McNeill, K. L., & Krajcik, J. S. (2012). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, reasoning framework for talk and writing*. Boston, MA: Pearson Education, Inc.

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk, 2*, 237-258.

Mislevy, R. J. (2007). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 257-305). Portsmouth, NH: Greenwood.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., Haertel, G. D., Riconscente, M. M., Rutstein, D. W., & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design*. Dordrecht, the Netherlands: Springer-Verlag.

Mislevy, R. J., Riconscente, M. M., & Rutstein, D. W. (2009). *Design patterns for assessing model-based reasoning*. Retrieved from (Large-Scale Assessment Report 6). Menlo Park, CA:

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design: A brief overview* (Vol. 2002). Princeton, NJ: Educational Testing Service.

National Academies of Sciences Engineering and Medicine. (2018). *Science and engineering for grades 6-12: Investigation and design at the center*. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Research Council.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Retrieved from Washington, DC:

Nehm, R. H., Beggrow, E., Opfer, J., & , & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher, 74*(2), 92-98.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Retrieved from Washington, DC:

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*, 320-323.

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.

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

Penuel, W. R., Wingert, K., & Van Horne, K. (2018). *Preparing teachers to notice key dimensions of next generation science assessment tasks.* Paper presented at the 13th International Conference of the Learning Sciences, London, UK.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*(1), 4-11.

Quintana, C., Reiser, B., Davis, E., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D. C., Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337 - 386.

Reiser, B. J., & Tabak, I. (2014). Scaffolding. In R. K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 44-62). New York, NY: Cambridge University Press.

Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 263-305). Mahwah, NJ: Lawrence Erlbaum.

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

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

Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (2010). *Mathematics teacher noticing: Seeing through teachers' eyes*. New York, NY: Routledge.

Sinha, S., Gray, S., Hmelo-Silver, C. E., Jordan, R., Honwad, S., Eberbach, C., Rugaber, S., Vattam, S. Goel, A. K. (2010). Appropriating conceptual representations: A case of transfer in a middle school science teacher. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences* (pp. 834-841). Chicago, IL: International Society of the Learning Sciences.

Tobin, K. (2006). Learning to teach through coteaching and cogenerative dialogue. *Teaching Education, 17*(2), 133-142.

van Es, E., Hand, V., & Mercado, J. (2017). Making visible the relationship between teachers' noticing for equity and equitable teaching practice. In E. O. Schack, M. H. Fisher, & J. A. Wilhelm (Eds.), *Teacher noticing: Bridging and broadening perspectives, contexts, and frameworks* (pp. 251-270). Dordrecht, the Netherlands: Springer.

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.

Weidler-Lewis, J., Penuel, W. R., & Van Horne, K. (2017). *Developing a measure of teachers' vision for equitable science teaching and learning*. Paper presented at the NARST Annual Conference, San Antonio, TX.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Windschitl, M., Thompson, J., & Braaten, M. (2011). Ambitious pedagogy by novice teachers: Who benefits from tool-supported collaborative inquiry into practice and why? *Teachers College Record, 113*(7), 1311-1360.

Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89-100.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162.

Appendix: 3D Assessment Rubric (Revisions 01-28-19)

TEACHER/PRE-POST Assessment: Total score (35)

Which grade level, unit and standards was the assessment was designed to assess?	
Target PE(s): What is the specific evidence outcome, NGSS performance expectation, or lesson level performance expectation that this assessment was designed to address?	
Write one sentence about why you chose the assessment to share with us.	
Write one thing you would change about this assessment if you were to reuse it.	
Scoring Guide. If you do not have the scoring guides or rubrics to upload, describe how you developed a grade for the assessment.	

Overall Comments:

Category 1: Appropriateness of the Scenario (Phenomenon to explain or Problem to solve)		0	1	2	Score	Comments/Evidence
Describe Scenario:						
1a-i	**Authenticity of scenario** - There is something for students to figure out (complex, options for Ss to reconcile)	Not a real problem for students to figure out. Nothing to explain or solve. Solution is obvious.	There is something for students to figure out. There is one viable solution path.	There is something for students to figure out. There are multiple viable solution paths for students to reconcile.		
1a-ii	- Extent to which the scenario uses real and scientifically accurate data, artifacts, or scientific models or simulations	Scenario is not real and/or does not include scientifically accurate ideas.	Scenario is made-up, but leverages scientifically accurate ideas or findings from real situations.	Scenario is real and scientifically accurate.		
1b	**Supports sensemaking** - Assesses the extent to which evidence from the scenario is needed to solve problem or explain phenomenon.	Students do not need to use information from the scenario to complete task(s).	Students need to use information from the scenario to complete task(s).	Students need to analyze or interpret information from the scenario to complete task(s).		
1c	**Application of understanding** Task requires students to **use and apply** their understanding to explain the phenomenon or solve the target problem.	Task does not require students to use or apply their learned knowledge. Task can be completed by restating learned knowledge as declarative knowledge.	Task requires students to use and apply their learned knowledge to explain or make predictions about a new phenomenon or solve a new problem			

1d	**Multiple modes for demonstrating understanding**	Students only have one way to demonstrate understanding	Students can use multiple ways to demonstrate understanding (e.g., diagrams, words)			
				Category 1 Total Score (8)		

Category 2: Disciplinary Core Ideas		0	1	2	Score	Comments/Evidence
Target DCI element(s):						
Alignment of content focus with the NGSS						
2a-i	**Intention to align assessed content with ideas in the NGSS** - Did the teacher specify one or more target PEs or DCI elements for assessment?	Teacher identified content foci for assessment, but did not specify one or more target PEs or DCI elements. or Teacher did not specify content focus for assessment.	Teacher specified one or more target PEs or DCI elements for assessment.			
2a-ii	**Alignment of assessed content with the NGSS** - Was the assessed content aligned with ideas found in the NGSS?	Assessed content is not aligned with any DCI elements found in the NGSS.	Assessed content includes some ideas that are aligned with DCI elements found in the NGSS and some that are not.	All of the assessed content is aligned with DCI elements found in the NGSS.		

2a-iii	**Grade / grade band appropriateness** - Evidence from grade-band endpoints from NRC Framework and Assessment Boundaries	Foregrounded assessment topics, as evidenced by questions and scoring guide, are not grade-level appropriate. They either focus on information that should have been assessed previously or should be assessed in a later grade bands.	Foregrounded assessment topics, as evidenced by questions and scoring guide, are grade-level appropriate and build upon previously learned knowledge. Previous grade-band knowledge is not the primary focus of this assessment task.			
Extent to which targeted content area or DCI elements are assessed and aligned with assessment goal						
2b-i	**Items address <u>targeted</u> DCI elements/content focus** - Are the items assessing targeted DCI elements OR reported content focus?	Assessment items do not address any targeted DCI element or content foci.	Assessment items partially address targeted DCI element(s) or content foci.	Assessment items address all targeted DCI element(s) or content foci.		
2b-ii	**Assessed DCI elements or content align with assessment goal** - Assessment goal could include demonstrating PE or answering overall question. - Do assessed DCI elements/content align with assessment goal? - Are there assessed ideas that distract from the assessment goal?	Assessment includes some items that assess DCI elements or content that distract from or are not aligned with assessment goal.	Assessment only includes items that assess DCI elements or content that are appropriate and align with assessment goal.			
					Category 2 Total Score (7)	

Category 3: Science and Engineering Practices		0	1	2	Score	Comments/Evidence
Target SEP element(s):						
Alignment with the NGSS						
3a-i	**Identification of target PEs, SEPs, or SEP elements** - Targeted SEPs can be identified through PE. - Did the teacher specify one or more target PEs or SEP elements for assessment?	Teacher did not specify any target PEs, SEPs, or SEP elements for assessment.	Teacher specified one or more target PEs, SEPs, or SEP elements for assessment.			
3a-ii	**Alignment of expected performance with the SEPs** - As students demonstrate their understanding, are students expected to use practices in ways that are aligned with the NGSS? - Are the items assessing targeted SEP elements OR reported use of SEPs?	Students are expected to demonstrate their understanding in ways that are *not* aligned with the SEPs. Assessment items do not address any targeted SEP(s) or SEP element(s).	Students are expected to demonstrate their understanding in ways that are *somewhat* aligned with the SEPs. Assessment items partially address targeted SEP(s) or SEP element(s).	Students are expected to demonstrate their understanding in ways that are *completely* aligned with the SEPs. Assessment items address all targeted SEP(s) or SEP element(s).		
3a-iii	**Grade / grade band appropriateness** - Reference NGSS Appendix F and DPS Science Competencies	Expected use of science and engineering practices, as evidenced by questions and scoring guide, are not grade-level appropriate. The focus of the expected performance is either well above or well below targeted grade bands.	Expected use of science and engineering practices, as evidenced by questions and scoring guide, is grade-level appropriate and does not include elements that are well above or well below grade-band.			

Extent to which targeted SEP elements are assessed and aligned with assessment goal						
3b-i	**Opportunities for students to use SEPs to complete task(s)** - Are there opportunities for students to use the SEPs to demonstrate their understanding?	Assessment does not include opportunities for students to use the SEPs to complete task(s).	Assessment includes some opportunities for students to use SEPs to complete task(s).	Assessment includes opportunities for students to use SEPs to complete task(s) throughout the assessment.		
3b-ii	**Assessed SEPs or SEP elements align with assessment goal** - Assessment goal could include demonstrating PE or answering overall question. - Do assessed SEPs or SEP elements align with assessment goal? - Are there assessed SEPs or SEP elements that distract from the assessment goal?	Assessment includes some items that assess SEPs or SEP elements that distract from or are not aligned with assessment goal.	Assessment only includes items that assess SEPs or SEP elements that are appropriate and align with assessment goal.			
Category 3 Total Score (7)						

Category 4: Integration of Crosscutting Concepts		0	1	2	Score	Comments/Evidence
Target CCC element(s):						
Alignment with the NGSS						
4a-i	**Identification of target CCCs or CCC elements** - Targeted CCCs can be identified through PE. - Did the teacher specify one or more target PEs or CCCs elements for assessment?	Teacher did not specify any target PEs, CCCs, or CCC elements for assessment.	Teacher specified one or more target PEs, CCCs, or CCC elements for assessment.			
4a-ii	**Alignment of expected performance with the CCCs** - As students demonstrate their understanding, are students expected to use crosscutting concepts in ways that are aligned with the NGSS? - Are the items assessing targeted CCC elements OR reported use of CCCs?	Students are expected to demonstrate their understanding in ways that are *not* aligned with the CCCs. Assessment items do not address any targeted CCC(s) or CCC element(s).	Students are expected to demonstrate their understanding in ways that are *somewhat* aligned with the CCCs. Assessment items partially address targeted CCC(s) or CCC element(s).	Students are expected to demonstrate their understanding in ways that are *completely* aligned with the CCCs. Assessment items address all targeted CCC(s) or CCC element(s).		
4a-iii	**Grade / grade band appropriateness** - Reference NGSS Appendix G	Expected use of crosscutting concepts, as evidenced by questions and scoring guide, are not grade-level appropriate. The focus of the expected performance is either well above or well below targeted grade bands.	Expected use of crosscutting concepts, as evidenced by questions and scoring guide, is grade-level appropriate and does not include elements that are well above or well below grade-band.			

Extent to which targeted CCC elements are assessed and aligned with assessment goal						
4b-i	**Opportunities for students to use CCCs to complete task(s)** - Are there opportunities for students to use the CCCs to demonstrate their understanding?	Assessment does not include opportunities for students to use the CCCs to complete task(s).	Assessment includes some opportunities for students to use CCCs to complete task(s).	Assessment includes opportunities for students to use CCCs to complete task(s) throughout the assessment.		
4b-ii	**Assessed CCCs or CCC elements align with assessment goal** - Assessment goal could include demonstrating PE or answering overall question. - Do assessed CCCs or CCC elements align with assessment goal? - Are there assessed CCCs or CCC elements that distract from the assessment goal?	Assessment includes some items that assess CCCs or CCC elements that distract from or are not aligned with assessment goal.	Assessment only includes items that assess CCCs or CCC elements that are appropriate and align with assessment goal.			
Category 4 Total Score (7)						

Category 5: 3-D Integration		0	1	2	Score	Comments/Evidence
5a	**Assessment coherence**	Assessment tasks are discrete and have no explicit connection with the scenario.	Some, but not all, of the assessment tasks are explicitly connected to the scenario.	Assessment tasks are explicitly connected to the scenario.		
5b	**Assessment tasks integrate 3-D**	Completion of assessment task involves attention to 0-1 dimension. If 2 or more dimensions are present, there is no evidence of integrating the dimensions. Each dimension is assessed separately.	Completion of assessment task requires integrated attention to 2 of the three dimensions.	Completion of assessment task requires integrated attention to all 3 dimensions.		
5c	**Scoring guide assesses students' mastery of PE**	Scoring guide assesses student performance related to one dimension or whether the response is right/wrong.	Scoring guide assesses student performance related to two dimensions.	Scoring guide assesses student performance related to their understanding of the DCI(s) and use of SEP(s) and CCC(s).		
Category 5 Total Score (6)						