The Challenge of Assessing "Knowledge in Use": Examples from Three-Dimensional Science Learning and Instruction

James W. Pellegrino (Chair), University of Illinois at Chicago, pellegjw@uic.edu
Brian Gane, University of Illinois at Chicago, bgane@uic.edu
Sania Zaidi, University of Illinois at Chicago, sania@uic.edu
Christopher J. Harris, SRI International, christopher.harris@sri.com
Kevin W. McElhaney, SRI International, kevin.mcelhaney@sri.com
Nonye Alozie, SRI International, SRI International, maggie.alozie@sri.com
Phyllis Pennock, Michigan State University, phyllishpennock@gmail.com
Samuel Severance, Michigan State University, severa18@msu.edu
Knut Neumann, Leibniz-Institute for Science and Mathematics Education, neumann@ipn.uni-kiel.de
David Fortus, Weizmann Institute of Science, david.fortus@weizmann.ac.il
Joseph Krajcik, Michigan State University, krajcik@msu.edu
Jeffrey Nordine, Leibniz-Institute for Science and Mathematics Education, nordine@ipn.uni-kiel.de
Erin Marie Furtak, University of Colorado, Boulder, erin.furtak@Colorado.edu
Derek Briggs, University of Colorado, Boulder, derek.briggs@colorado.edu
Rajendra Chattergoon, University of Colorado, Boulder, Rajendra.Chattergoon@Colorado.edu
William R. Penuel, University of Colorado Boulder, willliam.penuel@colorado.edu
Kerri Wingert, Omaha Public Schools, kerriwingert@gmail.com
Katie Van Horne, University of Colorado Boulder, katie.vanhorne@colorado.edu

Abstract: This symposium includes four papers focused on meeting challenges in the design and use of assessments of science proficiency for which students are expected to demonstrate their ability to explain scientific phenomena and solve problems by integrating disciplinary concepts with science and engineering practices. This view of multi-dimensional integrated science learning is exemplified by the performance expectations articulated in the *Next Generation Science Standards.* The four papers describe work that spans multiple grade levels and includes illustrations of the systematic design of assessments of knowledge-in-use for a range of life and physical science concepts, including a focus on energy. Illustrative tasks are provided together with data on student performance. The papers also consider issues of teacher implementation in classrooms, as well as methods that can be used to help teachers gain a deeper understanding of multi-dimensional science learning goals and effective assessment materials.

Symposium focus and overview

A key challenge in shaping science learning for the 21st century will be to develop new measures of learning that take into account what it means to be proficient in science (Pellegrino, 2013). The emergent view on proficiency, grounded in learning sciences research, emphasizes using and applying knowledge in the context of disciplinary practices. Referred to as *knowledge-in-use*, this perspective on science proficiency is a centerpiece of the National Research Council's (NRC) *Framework for K-12 Science Education* (NRC, 2012), is embodied in the *Next Generation Science Standards* (NGSS Lead States, 2013), and emphasized in the NRC report on developing assessments to measure science proficiency (Pellegrino, Wilson, Koenig, & Beatty, 2014). Central to this view is that disciplinary content and practices should be integrated so that as students apply knowledge to make sense of phenomena and solve problems, they deepen their conceptual understanding of content as well as their understanding of how to do science. This view of the goals of science learning can by juxtaposed with results of international research on science achievement showing that students often possess only fragmented knowledge of isolated science facts and lack the abilities to use their knowledge to explain phenomena or solve meaningful problems (e.g., OECD, 2012).

The shift to integrating science practices with disciplinary core ideas and crosscutting concepts, as emphasized in the *Next Generation Science Standards,* is based upon studies of actual scientific practice and what we currently know about student learning (e.g., NRC, 2007, 2012). This research corpus points to the importance of integrating content (i.e., disciplinary core ideas and crosscutting concepts) and practices by emphasizing that rich science learning requires tight coupling of what students know and what they can do. This presents a different way of thinking about science proficiency in that disciplinary core ideas and crosscutting concepts serve as thinking tools that work together with scientific and engineering practices to enable learners to

solve problems, reason with evidence, and make sense of phenomena (NRC, 2012). It also signifies that measuring proficiency solely as acquisition of core content knowledge is no longer sufficient (see e.g., Pellegrino, 2013).

Knowledge-in-use learning goals comprise the standards in the NGSS and are articulated as *Performance Expectations* (*PEs*). Each NGSS performance expectation combines a science or engineering practice, disciplinary core idea, and crosscutting concept into a single statement of *what is to be assessed* at the end of a grade level or grade band. It incorporates all three dimensions by asking students to apply disciplinary knowledge and make connections to a crosscutting concept as they engage in a science or engineering practice. An example for middle school physical science is: *Develop a model that predicts and describes changes in particle motion, temperature, and state of a pure substance when thermal energy is added or removed*.

In order to identify factors affecting the development of science proficiency and examine the efficacy of different approaches to supporting students' learning, assessments are needed that can reliably and validly assess students' knowledge-in-use (Pellegrino et al., 2014). This symposium reports results from a set of projects focused on the challenges associated with the design, validation, and use of such assessments to support student learning and classroom instruction. The first paper describes a design approach that starts with a representation of the knowledge underlying each of the dimensions associated with specific performance expectations and then translates that into targeted learning performances and associated assessment tasks that can be used formatively in middle-school classrooms. The second paper applies a similar approach to the design of assessments focused on energy concepts at the middle school level. Results are reported on reliability and validity, including evidence of the assessments sensitivity for showing gains in student performance following instruction that emphasizes three-dimensional learning. The third paper describes work at the high school level on assessment of student understanding of energy together with efforts to build a system of assessments that can fulfill formative as well as summative purposes. The fourth paper considers some of the challenges associated with helping teachers gain an understanding of three-dimensional science learning and assessment. It discusses results from a contrasting cases approach that was used to support teacher learning at multiple grade levels about the properties of assessments of three-dimensional learning and their value for supporting instruction.

Design of next generation science assessments: Measuring what matters

James Pellegrino, Christopher Harris, Joseph Krajcik, Brian Gane, Kevin McElhaney, Phyliis Pennock, Nonye Alozie, and Sania Zaidi

In this paper we overview our systematic and scalable approach for designing assessment items that measure student proficiency with new science learning goals that integrate disciplinary core ideas and crosscutting concepts with scientific practices. The assessment tasks are intended for formative use within classroom instruction. There is tremendous need for such assessment design work, as assessment plays a central role in supporting implementation of the new directions in science education both in the U.S. and internationally (Pellegrino et al., 2014). Our approach to meeting this challenge uses principles of Evidence-Centered Design (ECD; e.g., Almond, Steinberg, & Mislevy, 2002). ECD has been used in wide-ranging assessment design contexts, from the development of large-scale, high-stakes assessments to the design of classroom-based assessments. ECD emphasizes the evidentiary base for specifying coherent, logical relationships among the (1) learning goals that comprise the constructs to be measured (i.e., the claims we want to make about what students know and can do); (2) evidence in the form of observations, behaviors, or performances that should reveal the target constructs; and (3) features of tasks or situations that should elicit those behaviors or performances. We use ECD to systematically unpack science learning goals and synthesize the unpacking into multiple components that we call learning performances. Learning performances are knowledge-in-use statements that guide the development of assessment tasks and rubrics for measuring three-dimensional learning goals such as the performance expectations of the NGSS. Figure 1 overviews our overall design process that follows the logic of ECD and contains 3 distinct phases – **Domain Analysis**, **Domain Modeling**, and **Task and Rubric Development**. While the figure illustrates a linear process, the actual process is very iterative and recursive.

We begin Domain Analysis by first *unpacking core ideas* associated with a bundle of NGSS Performance Expectations at a given grade level or grade band. The process involves elaborating the meaning of key terms, defining expectations for understandings for the targeted student level, determining assessment boundaries for content knowledge; identifying background knowledge that is expected of students to develop a grade-level-appropriate understanding of a disciplinary core idea; and considering research-based problematic student ideas and misconceptions. Next, we *unpack the science practices*. The unpacking involves consideration of the core elements of the practice, intersections with other science practices and the evidence required to demonstrate the practice. This is followed by *unpacking the crosscutting concepts*, which involves identifying

the important elements of the concept and opportunities for intersections with the science practices and with the particular disciplinary core ideas that are the target of the assessment. Using the unpacking documents, we create a modified type of concept map, called an *integrated dimension map.* This map represents connections between the three dimensions (see Harris et al., 2016).

Domain Modeling involves 3 components. First we *articulate Learning Performances*. We use the integrated dimension maps as resources to help conceptualize and articulate a set of Learning Performances (LP) that constitute a trajectory towards mastery of each performance expectation (PE). Learning Performances are opportunities for students to demonstrate the knowledge-in-use they need at a specific point in the school year to be on track towards mastery of a PE by the end of the school year (Harris et al., 2016). For each PE, we develop multiple LPs. Once we have articulated LPs we move to *specifying task design patterns* in order to identify an "assessment argument" for constructing tasks, as per guidelines of ECD. Our assessment argument builds on the claim articulated in each Learning Performance by constructing a LP-specific design patterns that includes: (a) the focal (and additional) knowledge, skills, and abilities (KSAs) underlying the Learning Performance; (b) evidence statements that articulate how those KSAs can be observed in student performance; and (c) assessment task design features. We construct evidence statements by considering how we can observe student ability in each KSA; these statements are later used to develop assessment tasks and rubrics. Finally, we articulate characteristic and variable features for assessment tasks that help ensure the task can elicit the focal KSAs. In articulating these features, we use a Universal Design for Learning framework to also ensure that our design features result in tasks that are accessible to all students.

The final phase of the design process involves using the information detailed in the assessment argument to *develop assessment tasks and rubrics*. The task design depends on the specification of characteristic and variable task features and allows for assembly of multiple tasks within a "family" where the variations among the tasks can readily reflect intended levels of challenge. The task design process also takes into account the forms of evidence needed to support the learning performance claim and the ways in which that evidence will be scored and evaluated for purposes of rubric development.

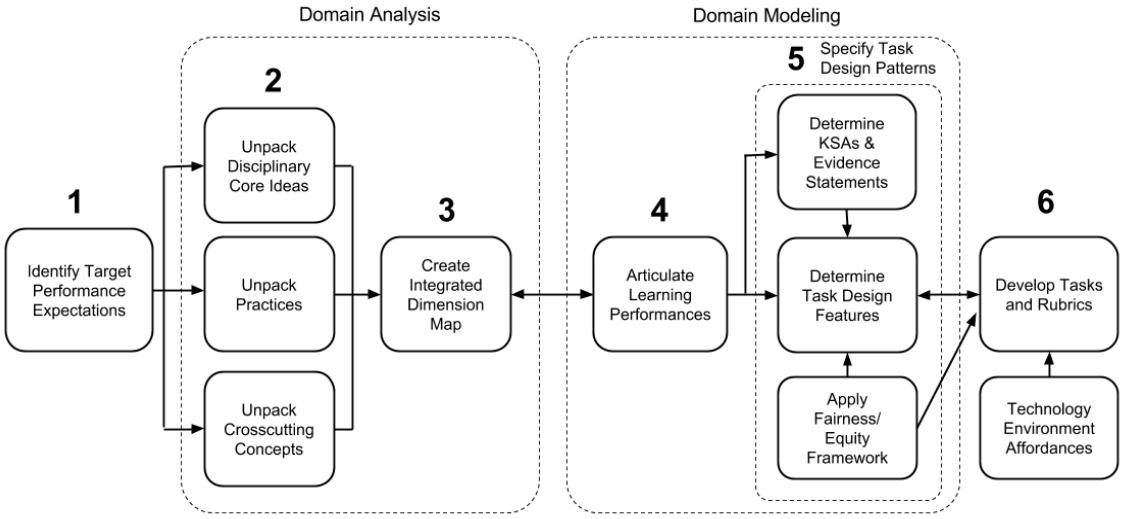

Figure 1. Illustration of the major elements of the assessment design process.

We have used the design process outlined above to unpack 9 PEs from the physical and life science disciplines and have created over 100 assessment tasks. The designed tasks are technology-enhanced (e.g., use of simulation, modeling software, video) and many tasks use non-textual representations to elicit student responses (e.g., through drawing or modeling). Across a series of studies using multiple research methods, we have assembled data indicating that our tasks are functioning as intended, minimize construct-irrelevant variance, and support teachers' classroom practice. For example, classroom observations have shown that teachers use the assessments in a variety of different modes, spanning a range between formative and summative use. Student cognitive lab studies have provided both data on task comprehensibility and identified issues of construct-irrelevant variance. Task performance studies have provided a preliminary set of data on item features (e.g., difficulty) that affect student performance and on the utility of our rubric design in affording partial credit scores based on the presence or absence of FKSAs in the student responses.

Assessing students' progression in developing knowledge-in-use for energy

Knut Neumann, David Fortus, Joseph Krajcik, and Jeffrey Nordine

This paper details efforts to develop summative assessments of Middle School students' knowledge-in-use about energy as part of an ongoing effort to compare two different approaches to teaching Energy in Middle School science instruction. Based on the *Next Generation Science Standards* we identified a set of performance expectations for elementary and middle school science to characterize the knowledge-in-use about energy expected of students at the end of middle school (NGSS Lead States, 2013). In authoring tasks, we used an evidence-centered-design approach and followed the procedures discussed in the preceding presentation and suggested by Harris et al. (2016). The procedure began with unpacking the performance expectations to identify major elements of the disciplinary core idea (DCI), the crosscutting concept (CCC), and the science and engineering practice (SEP). From these elements, learning performances were generated by combining one or more elements of the DCI, CCC, and SEP. Each of these learning performances represented a different aspect of a performance expectation, in order to ensure a sufficiently broad, yet concise specification of the construct (see Messick, 1995). For each learning performance, an assessment argument was created which specified the evidence required to conclude that students have met the learning performance together with additional knowledge that students may need, as well as fixed and variable task features. The assessment argument served as a blueprint for the authoring of tasks. We authored a total 24 tasks assessing students' knowledge-in-use about energy (see Figure 2 for a sample task).

Stuntman Felix holds the world record for the fastest speed achieved without an engine - 700 miles per hour. He used a balloon to fly almost 25 miles above the Earth and - wearing a special suit - jumped down.

Develop a model of Felix' jump that describes how Felix had enough energy to travel as fast as he did without an engine.

Figure 2. Sample item assessing students' ability to develop a model to describe Felix Baumgartner's jump from the stratosphere using their knowledge about potential gravitational energy.

In addition to their function in the authoring of tasks, the assessment arguments also served as a basis for the development of scoring rubrics specific to each item to guide the interpretation of students' responses. Assessments of students' knowledge-in-use must not separately target students' ability to engage in a SEP or their knowledge about a DCI or CCC independently, but their ability to engage in a SEP in the context of a DCI and CCC, assessments. Therefore, the scoring rubrics were designed to credit demonstrations of the integration of a SEP, with a DCI and CCC. To obtain insights into how reliably and validly the tasks assess students' progression in developing knowledge-in-use about energy, the tasks were administered to N = 311 students from 8th grade classes at two schools in Midwest USA – prior to and after an instructional unit on energy.

To examine the reliability and validity of our assessments, we first examined scoring quality. About 20 percent of student responses were scored by a second scorer, leading to good to very good agreement, $p > .82$. We then explored the extent to which the tasks, as a whole, functioned as a reliable measure of students' three-dimensional learning about energy. Our analyses yielded reliability of $\alpha = .63$. In a continued effort to examine the validity with which the developed task assess students' knowledge-in-use, we investigated the correlation of students' pre-test scores with their last grades in Science ($r_S = .33$, $p < .001$), Mathematics ($r_M = .19$, $p < .001$) and English ($r_E = .28$, $p < .001$). Finally, we investigated the extent to which the assessment measures students' progression in developing knowledge-in-use about energy in terms of score gains between pre- and post-test. In doing so, we found a significant and strong gain, Cohen's $d = 1.03$.

The results suggest that our approach to developing assessments of students' knowledge-in-use about energy is suitable to yield reliable and valid assessments. The reliability is below a typically applied cut-off value of $\alpha > .80$, but still satisfactory given the complex knowledge-in-use construct. In addition, the correlations of the pre-scores with student grades, and the strong gain in students' scores from pre- to post-test suggest that the assessments indeed represent valid assessments of students' progression in developing knowledge-in-use. We envision our procedure as a blueprint for developing high quality tasks to assess students' learning as a function of instruction and to compare different instructional approaches in terms of their efficacy. The presentation will discuss in greater detail task development procedures, with sample tasks, scoring rubrics and scored students responses, and discuss how the data can be analyzed to obtain reliable and valid information about students' progression in developing knowledge-in-use about a core science concept.

Toward a system of classroom assessments for three-dimensional secondary science learning: The case of the Aspire study

Erin Marie Furtak, Derek Briggs, and Rajendra Chattergoon

This presentation will draw examples from the ongoing work of the Aspire Research-Practice Partnership, a long-term, mutualistic collaboration (Penuel et al., 2011) between researchers at the University of Colorado Boulder and teachers and science curriculum coordinators in a culturally, linguistically, and socioeconomically diverse school district near a large city in the Western US. Our partner district has adopted sets of exit statements that are either aligned to or fully verbatim facsimiles of the *Next Generation Science Standards* and professional learning experiences are supporting changes in classroom practice. However, the secondary science teachers in the district are only beginning to realize these changes in their instruction.

We have taken the approach of developing, in partnership with the district and teams of high school chemistry, biology, and physics teachers, a system of classroom assessments aligned with this three-dimensional vision. To focus our work, we have dedicated our assessment design efforts to the disciplinary core idea and crosscutting concept of Energy cycling within systems. Consistent with the district's focus on three-dimensional assessment, this has also included a focus on model-based explanation. We take as the centerpoint of this process of assessment design a model of cognition (Pellegrino, et al., 2001) based upon a hypothesized learning progression for energy (Neumann et al., 2013) that articulates the development of student understanding of energy forms, transfer/transformation, dissipation, and conservation.

Our design activities have started with tracing energy as a crosscutting concept using the original format of the Neumann et al. (2013) learning progression, which has involved determining how the energy concepts it articulates apply in chemistry and biology. At the same time, we have worked toward creating three coordinated forms of classroom assessment linked to this learning progression: a *pre-post summative* assessment that will allow us to model student growth within and across grade levels; a *modeling performance task* that engages students in model-based explanation of an energy phenomenon across disciplinary contexts; and sets of *embedded formative assessments* that engage students in modeling and explanation of energy cycling in individual instructional units. We describe each in the following paragraphs.

Building on pre-existing, multiple-choice and constructed-response item sets from Neumann et al. (2013), Park & Liu (2016) and Opitz (2016), we have created clusters of items that map onto the energy learning progression, including those that focus on energy as it manifests in content-specific disciplinary core ideas, as well as sets of linking items that create opportunities for students to demonstrate their understanding of energy as a crosscutting concept that bridges their learning experiences in physics, chemistry, and biology. Our initial pilots of these item sets have indicated that items associated with the upper end of the learning progression are more difficult for students than those associated with the lower levels. Furthermore, the factor structure of the items is dependent on whether students were enrolled in a biology course or a physics course.

Acknowledging that the pre-post summative assessment focuses primarily on student understanding of energy as a disciplinary core idea and a crosscutting concept, we also are seeking to assess students' engagement in model-based explanation through a performance task. Working with the example of ethanol-fueled engines, this task asks students to trace how energy from the sun is used to power a bus as it drives up a hill, tracing energy transfer and transformation from corn to ethanol production to powering a piston in a bus engine. Our ultimate goal is to score the performance task on the same scale we use for the summative pre-post assessment. Initial efforts to pilot this item indicated that high school students weren't even sure where to begin to think about this problem, since it was so far outside their learning experiences in school. Furthermore, we have received surprising feedback from scientists, who self-consciously admitted they were unsure of how to trace energy in the contexts outside their areas of expertise. This raises questions as to the validity of claims we might make about high school student learning when it pushes on the boundaries between disciplines acknowledged and enforced by scientists themselves.

In parallel, we have collaborated with high school science teachers to co-design and embed formative assessments in units of instruction in which energy is explicitly taught: potential and kinetic energy in 9th grade Physics, chemical reactions in 10th grade Chemistry, and matter and energy cycling in 11th grade Biology. This process of co-design works within a pre-existing professional learning model to support teachers in designing, enacting, and determining next instructional steps for formative assessments (Furtak & Heredia, 2014), and engages teachers in adapting pre-existing scaffolds to engage students in modeling (e.g., Kang et al., 2016).

Preparing teachers to notice key dimensions of next generation science assessment tasks

In this paper, we report on a strategy for helping teachers shift their vision for assessment. The strategy entails analyzing sets of multi-component assessment tasks for their adequacy in eliciting students' science proficiency. The aim is to shift teachers' attention toward the power of tasks to elicit each of three dimensions of proficiency emphasized within the *Framework for K-12 Science Education* (NRC, 2012): their understanding of disciplinary core ideas and crosscutting concepts, as well as their grasp of science and engineering practices. The conjecture we explored in this design study is whether analyses of assessment tasks can support teachers in noticing key dimensions of proficiency that are either present or absent. In this paper, we present evidence related to this conjecture, and we consider the kinds of tasks that make it easier or harder for teachers to discern the dimensions of proficiency tasks are intended to elicit.

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). Learning research also suggests that key to helping teachers notice the distinctive features of assessment 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). 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.

Participants in this study were a total of 99 teachers from a single US state, organized into groups of three by grade level. There were 12 groups analyzing tasks targeting fifth grade students, 5 groups analyzing tasks targeting first grade students, eight groups focused on middle school tasks, and eight focused on high school tasks. We presented each group with a set of five or six tasks, and we told the teachers that each task was intended to elicit student understanding of a "three-dimensional" standard or learning goal, that is, one that could assess students' integrated understanding of core ideas, science practices, and crosscutting concepts. We created each task set to purposefully include 1-2 tasks that had great potential to elicit three-dimensional proficiency (e.g., multicomponent tasks that presented students with a phenomenon in which they had to use science practices and their understanding of core ideas and crosscutting concepts to explain), as well as 1-2 tasks that elicited only factual recall. Tasks were selected by researchers knowledgeable both about assessment and the specific demands of assessment outlined in the National Research Council report, *Developing Assessments for the Next Generation Science Standards* (Pellegrino, et al., 2014).

We asked the teachers to work in their groups to rate each task within their set on a scale from 1 (completely inadequate) to 5 (completely adequate) with respect to the prompts' ability to elicit student understanding of the learning goal, which was printed at the top of the task for teachers to review. In addition, to entering their ratings for each task using a Google Form, groups entered reasons for their judgments. Later, we presented the summary ratings of tasks to the group and discussed them as a whole.

We calculated mean ratings for each task of the group, along with standard deviations for each, and compared them to an expert rater's ratings for the task, a research analyst who had no part in assembling or creating the task but who had extensive experience in designing assessments to elicit three-dimensional science proficiency. This analyst also used an open coding approach (Charmaz, 2000) to identify reasons for teachers' ratings. We found that teachers were able to distinguish tasks with respect to their three-dimensionality and that the relative rankings of tasks were similar to those of the expert coder overall. Each task set exhibited the full range of ratings for tasks (1-5), as intended, though the standard deviation of ratings varied across the tasks. The most extreme range was for the set of first grade tasks, where the standard deviation of ratings ranged from 0.45 to 1.34. An analysis revealed that the standard deviations were lower for each end of the scale. The correlation between the standard deviation and distance from the middle of the rating scale (3) was $r = -.803$. In addition, when their ratings diverged most from the expert rater's, the farther the average rating was from the middle of the rating scale ($r = -.718$). Put together, these two correlations suggest teachers were better able to discern when three-dimensionality was starkly absent (e.g., on closed-ended tasks eliciting factual recall) or strongly present (e.g., in multicomponent tasks requiring students to explain a natural phenomenon).

The open coding also revealed that teachers did in fact attend to the dimensions as intended, but teachers also attended to concerns sometimes raised by experts about assessment (Table 2). Of the focal three

dimensions, the science and engineering practices were attended to most. Teachers also attended to cognitive complexity, as evident in language they used, such as "Depth of Knowledge" and "Bloom's Taxonomy." Both are terms in education and assessment that have been taken up widely in practice (Schneider, 2014).

Table 1: Dimensions of Tasks Teachers Noticed

	# of Responses Noticed	% of All Reasons
Science and Engineering Practices	80	32
Specific Elements of standard (nonspecific)	41	17
Cognitive Complexity	40	16
Concern for Clarity of the Task	32	13
Disciplinary Core Ideas	24	10
Cross-cutting Concepts	21	8
Concern for Equity	10	4

We found some support for our conjecture that task analysis can facilitate teachers' noticing of features of next generation science assessment tasks in ways that align with experts' ratings. As an entry point for teacher learning, teachers readily discerned tasks that clearly had little potential to elicit students' integrated understanding of core ideas, practices, and crosscutting concepts, as well as those tasks that clearly did. At the same time, the analysis revealed teachers did not as easily notice features of tasks that were partly flawed. Though this may simply have been an artifact of the tasks themselves, the fact they were not in agreement on these tasks and diverged from the expert's ratings suggests otherwise. Their reasons, overall, revealed a wider range of concerns about assessment than just the three dimensions. We envision this type of activity as an entry point to a longer learning trajectory for teachers, in which they become proficient in designing high-quality multi-component tasks. As others have found, engaging teachers in designing assessment tasks is both promising and also challenging (e.g., Furtak et al., 2016; Shepard, 1997), in part because of the broad knowledge base required in both science and assessment. To judge whether task analysis can bootstrap teachers' learning toward more productive design, then, we need to explore the effects of longer sequences of teacher learning, which will also require scaffolding that supports their knowledge building in science and assessment.

Symposium summary and implications

This symposium brings together cutting-edge work in the design, piloting, and use of multiple forms of assessment, all aligned with the purpose of assessing knowledge-in-use. Taken together, they represent elementary, middle, and high school science learning; classroom formative assessments, as well as tasks being designed for more proximal and distal uses; and international partnerships. They also raise a number of questions critical for the field to consider, including:

- While performance expectations often follow a thread of unfolding disciplinary core ideas or scientific practices, the NGSS also create opportunities to follow crosscutting concepts as they unfold across multiple grade levels. What does it mean to assess a crosscutting concept, and what new challenges for assessing knowledge-in-use are presented with a focus on crosscutting concepts?
- What are the different types of task formats being developed, and how do these change as we move from activities that are proximal to classroom instruction toward large-scale assessments?
- How do similarities and differences in international visions for science knowledge-in-use influence the ways we think about and design assessments, both at-scale, and in classrooms?
- What constraints and affordances can be identified when assessment development activities are conducted with teachers, schools, and school districts?

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1* (5).

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

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.
Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with interesting implications. *Review of Research in Education, 24*, 61-101.
Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 509-535). Thousand Oaks, CA: Sage.
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., & Heredia, S. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching*, *51*(8), 982–1020.
Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K. W. (2016). *Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications.* Menlo Park, CA: SRI International.
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., Windschitl, M., Stroupe, D., & Thompson, J. (2016). Designing, launching, and implementing high quality learning opportunities for students that advance scientific thinking. *Journal of Research in Science Teaching*, *53*(9), 1316–1340.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.
National Research Council (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.
National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Research Council.
Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188.
NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.
OECD (2013). *PISA 2012 Results*. Paris: OECD.
Opitz, S.T. (2016). *Students' progressing understanding of the energy concept: An analysis of learning in biological and cross-disciplinary contexts*. Doctoral Dissertation, Leibniz-Institute for Science Education, Kiel, Germany.
Penuel, W. R., Fishman, B. J., Haugan Cheng, B., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*(7), 331–337.
Park, M., & Liu, X. (2016). Assessing understanding of the energy concept in different science disciplines. *Science Education, 100*(3), 483–516.
Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*, 320-3.
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment.* Washington D.C.: National Academies Press.
Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A (Eds.) (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.
Schneider, J. (2014). *From the ivory tower to the schoolhouse: How scholarship becomes common knowledge in education*. Cambridge, MA: Harvard University Press.
Shepard, L. A. (1997). *Insights gained from a classroom-based assessment project*. CSE Technical Report 451. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
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*(5), 941–967.

Acknowledgments

The research and development activities discussed in these 4 papers were supported by the following grants from the U.S. National Science Foundation: DRL–1316903, DRL-1316908, DRL-1316874, DRL-1748757, DRL-1748757, and DUE-1431725; and Grant #4482 from the Gordon and Betty Moore Foundation. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Gordon and Betty Moore Foundation.