

Next Generation Science Standards: Challenges and Illustrations of Designing Assessments that Serve Learning

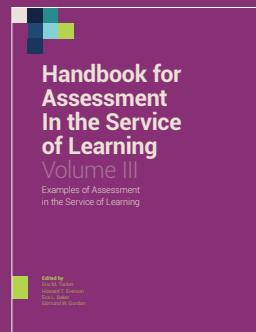
James W. Pellegrino and Howard T. Everson

UMassAmherst

University Libraries

Series Editors:

Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, & Eric M. Tucker





© 2025 by James W. Pellegrino and Howard T. Everson

The Open Access version of this chapter is licensed under a Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License (CC-BY-NC-ND 4.0).

ISBN: 978-1-945764-33-2

Suggested Citation:

Pellegrino, J. W., & Everson, H. T. (2025). Next Generation Science Standards: Challenges and illustrations of designing assessments that serve learning. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning*. University of Massachusetts Amherst Libraries.

Next Generation Science Standards: Challenges and Illustrations of Designing Assessments that Serve Learning

James W. Pellegrino and Howard T. Everson

Abstract

This chapter examines challenges and solutions in designing assessments aligned with the Next Generation Science Standards (NGSS), focusing on the NGSS's multi-dimensional approach to science education, integrating Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts. The chapter describes two major assessment design projects—the Next Generation Science Assessment (NGSA) project which developed classroom-focused assessment tasks for grades 3–8 that support formative assessment, and the Stackable, Instructionally-Embedded, Portable Science (SIPS) assessments which created end-of-unit assessments for grades 5 and 8. Both projects addressed the challenge of assessing integrated knowledge rather than separate dimensions of science learning. Throughout, the emphasis is on the importance of viewing science competence as a multi-dimensional performance that integrates content knowledge with scientific practices. The chapter concludes by discussing the benefits of these projects, including providing models for assessment design, creating ready-to-use resources for educators, and offering students challenging tasks that can better represent their scientific proficiency. While these efforts require further validation evidence with respect to their intended classroom use, the work described represents significant progress in developing assessments that align with contemporary views of science education while acknowledging the ongoing challenges in creating valid, reliable, and instructionally supportive measures of multi-dimensional science learning.

I. **Changing Nature of Science Competence: What Students Need to Know and Be Able to Do**

A. Multiple, Interconnected Dimensions of Competence

The nature of science competence has been reconsidered and the current conceptualization is most clearly expressed in the 2012 NRC report *A Framework for K–12 Science Education*, which articulates three interconnected dimensions of competence. The first of these dimensions are Disciplinary Core Ideas. In reaction to criticisms of U.S. science curricula being “a mile wide and an inch deep” (Schmidt, McKnight, & Raizen, 1997, p. 62) compared to other countries, the Framework identified and focused on a small set of core ideas in four areas: (a) life sciences, (b) physical sciences, (c) earth and space sciences, and (d) engineering, technology, and the application of science. In so doing, the Framework attempted to reduce the long and often disconnected catalog of factual knowledge that students typically had to memorize. Core ideas in the physical sciences include energy and matter, for example, and core ideas in the life sciences include ecosystems and biological evolution. Students are supposed to encounter these core ideas over the course of their school years at increasing levels of sophistication, deepening their knowledge over time. The second dimension is Crosscutting Concepts. The Framework identifies seven such concepts that have importance across many science disciplines; examples include patterns, cause and effect, systems thinking, and stability and change. The third dimension is Science and Engineering Practices. Eight key practices are identified, including asking questions (for science) and defining problems (for engineering); planning and carrying out investigations; developing and using models; analyzing and interpreting data, and engaging in argument from evidence.

While the Framework’s three dimensions are conceptually distinct, the vision is one of coordination in science and engineering education such that the three are integrated in the teaching, learning, and doing of science and engineering. By engaging in the practices of science and engineering, students gain new knowledge about the disciplinary core ideas and come to understand the nature of how scientific knowledge develops. Thus, it is not just the description of key elements of each of the three dimensions that matters in defining science competence; the central argument of the Framework is that the meaning of competence is realized through performance expectations describing what students at various levels of educational experience should know and be able to do. These performance

expectations integrate the three dimensions and move beyond the vague terms, such as “know” and “understand,” often used in previous science standards documents to more specific statements like “analyze,” “compare,” “predict,” and “model,” in which the practices of science are wrapped around and integrated with core content. Finally, the Framework makes the case that competence and expertise develop over time and increase in sophistication and power as the product of coherent systems of curriculum, instruction, and assessment.

B. From Frameworks to Standards: A Focus on Performance Expectations

The Framework uses the three dimensions—the practices, crosscutting concepts, and core ideas of science and engineering—to organize the content and sequence of learning. This three-part structure signaled an important evolutionary shift for science education and presented the primary challenge for the design of both instruction and assessment—finding a way to describe and capture students’ developing competence along these intertwined dimensions. The Framework emphasizes that research indicates that learning about science and engineering “involves integration of the knowledge of scientific explanations (i.e., content knowledge) and the practices needed to engage in scientific inquiry and engineering design” (p. 11). Both practices and crosscutting concepts are envisaged as tools (skills and strategies) for addressing new problems that are equally important for students’ science learning as the domain knowledge topics with which they are integrated. Students who experience use of these tools in multiple contexts as they learn science are more likely to become flexible and effective users of them in new problem contexts.

To support the approach to science learning described above, the Framework states that assessment tasks must be designed to gather evidence of students’ ability to apply the practices and their understanding of the crosscutting concepts in the contexts of problems that also require them to draw on their understanding of specific disciplinary ideas. In developing the Next Generation Science Standards (NGSS), Achieve and its partners elaborated these guidelines into standards that are clarified by descriptions of the ways in which students at each grade are expected to apply both the practices and crosscutting concepts, and of the knowledge they are expected to have of the core ideas (NGSS Lead States, 2013). As shown in Figure 1, the NGSS standards appear as clusters of performance expectations related to a particular aspect of a core disciplinary

idea. Each performance expectation asks students to use a specific practice and a crosscutting concept in the context of a specific element of the disciplinary knowledge relevant to a particular aspect of the core idea. Across the set of such expectations at a given grade level, each practice and crosscutting concept appears in multiple standards. Figure 1 shows the “architecture” of the performance expectations in terms of the underlying knowledge associated with each of the three facets of the Framework—disciplinary core ideas, science and engineering practices, and crosscutting concepts—for the set of three 4th grade performance expectations for the Life Science topic area labelled *From Molecules to Organisms: Structures and Processes*.

Figure 1.
Example of the NGSS Architecture for one Aspect of 4th grade Life Science.

| 4-LS1 From Molecules to Organisms: Structures and processes | | |
|--|---|--|
| <p>4-LS1 From Molecules to Organisms: Structures and processes</p> <p>Students who demonstrate understanding can:</p> <p>4-LS1-a. Use simple models to describe that plants and animals have major internal and external structures, including organs, that support survival, growth, behavior, and reproduction. [Clarification Statement: Examples of structures include thorns, stems, roots, stamens, ovaries, heart, brain, skin, or bones.] [Assessment Boundary: Students are responsible for the overall functions of major structures, but the mechanisms of how they function within a system are not assessed. Students are not expected to memorize different types of structures but should be able to use information given.]</p> <p>4-LS1-b. Design, test, and compare solutions that replace or enhance the function of an external animal structure necessary for survival. [Clarification Statement: Students might compare solutions for mobility based on the strength of different materials used.]</p> <p>4-LS1-c. Construct models to describe that animals' senses receive different types of information from their environment, process the information in the brain, and respond to the information in different ways. [Clarification Statement: Examples of models could be diagrams or analogies.] [Assessment Boundary: Students are not expected to know the mechanisms by which the brain stores and recalls information, nor the mechanisms of how sensory receptors function.]</p> | | |
| <p>The performance expectations above were developed using the following elements from the NRC document <i>A Framework for K-12 Science Education</i>:</p> | | |
| Science and Engineering Practices | Disciplinary Core Ideas | Cross-cutting Concepts |
| <p>Developing and Using Models</p> <p>Modeling in 3–5 builds on K–2 models and progresses to building and revising simple models and using models to represent events and design solutions.</p> <ul style="list-style-type: none"> Develop a model using an analogy, example, or abstract representation to describe a scientific principle or design solution. (4-LS1-c) Identify limitations of models. (4-LS1-c) Use a simple model to test cause and effect relationships concerning the functioning of a proposed object, tool or process. (4-LS1-a) <p>Constructing Explanations and Designing Solutions</p> <p>Constructing explanations and designing solutions in 3–5 builds on prior experiences in K–2 and progresses to the use of evidence in constructing multiple explanations and designing multiple solutions.</p> <ul style="list-style-type: none"> Use evidence (e.g., measurements, observations, patterns) to construct a scientific explanation or design a solution to a problem. (4-LS1-b) Apply scientific knowledge to solve design problems. (4-LS1-b) <p>Obtaining, Evaluating, and Communicating Information</p> <p>Obtaining, evaluating, and communicating information in 3–5 builds on K–2 and progresses to evaluating the merit and accuracy of ideas and methods.</p> <ul style="list-style-type: none"> Compare and/or combine across complex texts and/or other reliable media to acquire appropriate scientific and/or technical information. (4-LS1-a) Use multiple sources to generate and communicate scientific and/or technical information orally and/or in written formats, including various forms of media and may include tables, diagrams, and charts. (4-LS1-a) | <p>LSA.A: Structure and Function</p> <ul style="list-style-type: none"> Plants and animals have both internal and external structures that serve various functions in growth, survival behavior, and reproduction. (4-LS1-a), (4-LS1-b) <p>LS1.D: Information Processing</p> <ul style="list-style-type: none"> Different sense receptors are specialized for particular kinds of information, which may be then processed and integrated by the animal's brain, with some information stored as memories. Animals are able to use their perceptions and memories to guide their actions. Some responses to information are instinctive—that is, animals' brains are organized so that they do not have to think about how to respond to certain stimuli. (4-LS1-c) <p>ETS1.C: Optimizing the Design Solution</p> <ul style="list-style-type: none"> Different solutions need to be tested in order to determine which of them best solves the problem given the criteria and the constraints. (secondary to 4-LS1-b) | <p>Structure and Function</p> <ul style="list-style-type: none"> Structures have shapes and parts that serve functions. (4-LS1-a), (4-LS1-b), (4-LS1-c) <p>-----</p> <p>Influence of Engineering, Technology, and Science on Society and the Natural World</p> <ul style="list-style-type: none"> Engineers improve existing technologies or develop new ones to increase their benefits, decrease known risks, and meet societal demands. (4-LS1-b) |

In contrast to science standards like the NGSS that call for the integration of science practices and content knowledge, the prior generation of U.S. science standards (e.g., NRC, 1996) treated content and inquiry as fairly separate strands of science learning, and assessments followed suit. In some respects, the form the standards took contributed to this separation: content standards stated what students should know, and inquiry standards stated what they should be able to do. Consequently, assessments separately measured the knowledge and inquiry practice components. Thus, the idea of an integrated, multi-dimensional science performance presents a very different way of thinking about science proficiency. 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. Such a view of competence also signifies that measuring proficiency solely as the acquisition of core content knowledge or as the ability to engage in inquiry processes free of content knowledge is neither appropriate nor sufficient.

C. Assessing Competence: How Will We Know What Students Know?

As illustrated in Figure 1, the NGSS performance expectations reflect intersections of a disciplinary core idea, science and engineering practices, and related crosscutting concepts, and they may also include boundary statements that identify limits to the level of understanding or context appropriate for a grade level and clarification statements that offer additional detail and examples. But standards and performance expectations, even as explicated in the NGSS, do not provide sufficient detail to create assessments. The design of valid and reliable science assessments is a complex endeavor that hinges on multiple elements that include, but are not restricted to, what is articulated in disciplinary frameworks and standards, such as those illustrated above for K–12 science education (Pellegrino et al., 2001; Mislavy & Haertel, 2006). For example, in the design of assessment items and tasks related to the performance expectations in Figure 1, one needs to also consider: (1) the kinds of conceptual models and evidence that we expect students to engage in; (2) grade-level appropriate contexts for assessing the performance expectations; (3) options for task design features (e.g., computer-based simulations, computer-based animations, paper-and-pencil writing and drawing) and which of these are essential for eliciting students' ideas about the performance expectation; and (4) the types of evidence that will reveal levels of student understanding and skill.

The challenge with standards expressed in this multi-dimensional form is how to design curricular and instructional materials to support acquisition of the important competencies underlying these performance expectations, and how to organize classroom instruction, including the design and use of formative and summative assessments, to promote student attainment of the complex disciplinary objectives embodied by such contemporary STEM standards. As discussed by Pellegrino, Wilson, Koenig, and Beatty in the 2014 NRC report *Developing Assessments for the Next Generation Science Standards*, significant assessment design challenges are posed by these multi-dimensional performance statements, especially when contrasted with previous generations of science assessment tasks that separately tested either disciplinary content knowledge or science "inquiry" (See also Pellegrino, 2013). They argued that considerable research and development was needed to create and evaluate assessment tasks and situations to determine if they can provide adequate and valid evidence of the proficiencies implied by the performance expectations of the NGSS, or any similar multi-dimensional standards derived from the NRC Framework.

Multiple arguments about the assessment design and validation challenges posed by the Framework and NGSS were explicated in some detail (Pellegrino et al., 2014), including the need for a principled design process to guide the work, of which the evidence centered design framework (Mislevy & Haertel, 2006) constitutes one such example. A related and critical argument was that such design and validation work needed to be conducted in instructional settings where students were being provided with adequate learning opportunities to construct the integrated knowledge envisioned by the NRC Framework and NGSS (Pellegrino, 2013; Pellegrino et al., 2014). While work of this type has advanced over the ensuing decade, much still needs to be done across the K–12 grade span and for multiple content domains. In the remainder of this chapter, we provide two examples of such efforts. Both focus on developing assessments and related instructional resources for use in K-8 classrooms. The two projects share an emphasis on supporting teachers as they strive to support students' progress toward developing and demonstrating the proficiencies underlying the performance expectations articulated in the Framework and NGSS. It is our contention that these two projects embody and support each of the multiple *Principles for Assessment in the Service of Learning* as espoused by Professor Edmund Gordon and his colleagues and as described in Volume I of this publication series.

II. The Next Generation Science Assessment (NGSA) Project

A. Introduction

As described above, the *Framework for K–12 Science Education* and the NGSS articulate an ambitious vision for what students should know and be able to do in science. They emphasize that all students must have the opportunity to learn and actively participate in authentic science through using and applying disciplinary core ideas (DCIs) in concert with science and engineering practices (SEPs) and crosscutting concepts (CCCs) to make sense of phenomena or solve problems. Central to this vision is the notion of knowledge-in-use, where students use and apply the three dimensions to build the integrated proficiencies identified in the NGSS Performance Expectations. Many science educators and scientists have embraced the vision described in the Framework and instantiated in the NGSS (e.g., NSTA, 2016), and the vast majority of states, representing more than 75% of the U.S. student population, now have standards influenced by the NGSS and/or the Framework. While this vision holds promise for engaging a broad diversity of students in the learning of science, the opportunity to learn can be realized only if teachers have the tools that can help them examine, reflect on, and improve their science instruction.

Among the most essential tools for teachers are classroom-based assessments. High-quality science instruction requires high-quality classroom-based assessments that can be used formatively and that are aligned with the standards (e.g., Fuhrman et al., 2009; Pellegrino et al., 2014; Pellegrino, 2018). Importantly, assessments provide a necessary picture of how students' science learning is building over time. Yet, many teachers do not feel well prepared to develop their own NGSS-aligned assessments or use them formatively in their classrooms (e.g., Furtak, 2017). Science teachers need purposefully designed assessment tasks for the NGSS that they can readily use in their classrooms. Especially needed are (1) tasks and rubrics that provide just-in-time information about students' progress in building toward the NGSS performance expectations (PEs), (2) resources that support instructional decision-making based on the assessment information, and (3) a delivery system for easy access and use by teachers and students.

The Next Generation Science Assessment project was initiated to address these needs by developing the NGSA System (<http://nextgenscienceassessment.org>). The system consists of innovative NGSS-aligned classroom-focused assessment tasks with rubrics for interpreting student performance and teacher guides for classroom use, all housed on an online portal for flexible administration and scoring

(<https://ngss-assessment.portal.concord.org>). As noted below, the NGSA System resources have been widely used both in the U.S. and internationally.

In the brief descriptions that follow we provide relevant background on the project's overall logic and need, the design team, the assessment design and development approach, validity evidence, and further information on the NGSA Portal's resources including some examples of resources.

B. Need for the NGSA System Resources

The NGSA Project Team pursued development of a technology-enabled assessment system for three important reasons. First, we know from considerable published literature and the wisdom of practice that assessment can be valuable for classroom pedagogy, especially when it is integrated within instruction and used formatively to guide the progress of student learning (e.g., Penuel & Shepard, 2016). But we also know that the NGSS Performance Expectations pose considerable challenges when it comes to designing assessments that support instruction and students' learning (Pellegrino et al., 2014). This creates a compelling reason to provide exemplar tasks and rubrics to teachers and others to illustrate what is expected of students and how to evaluate it.

Second, highly specified and developed resources (Cohen & Ball, 1999) are needed to help teachers integrate formative assessment practices into their instruction so that they can monitor students' progress. Indeed, well-designed assessment tasks are valuable for giving teachers a foothold to determine what their students know and can do—information that is also useful for making informed instructional decisions (Ruiz-Primo & Furtak, 2007; Ruiz-Primo & Furtak, 2024). However, assessment tasks alone are not enough. Enacting assessment tasks for formative use in classrooms presents unique problems of practice for teachers (Sezen-Barrie & Kelly, 2017), and these become even more pronounced when orchestrating science assessment within NGSS instruction (Furtak, 2017). Problems of practice include using tasks in formative ways and supporting students as they engage in tasks; interpreting student work; and determining next steps to advance student learning (e.g., Furtak, 2017; Kang, Thompson, & Windschitl, 2014; Shepard, Penuel, & Pellegrino, 2018). A viable solution is to provide teachers with assessment resources such as practice guides that illustrate how to formatively integrate assessment tasks into instruction over time, thereby making tasks usable and instructionally beneficial to teachers and their students.

Third, classroom assessments should take advantage of the capabilities provided by learning technologies. Technology-delivered assessments have several benefits for teachers and students to engage in regular formative assessment practice (Davies, 2010; Gane, Zaidi, & Pellegrino, 2018; Zhai & Wiebe, 2023). For students, technology enhancements such as video and simulations can expand the phenomena that can be investigated. Various assistive technologies can be used to make assessment materials more accessible to all students; for example, through screen readers that facilitate navigation and reading of text and speech-to-text capabilities that support students in responding to tasks. By providing background drawings, drawing tools, stamps, and/or predetermined model components, technologies can help scaffold students in demonstrating their learning in deeper ways. Moreover, because technology-delivered assessment tasks can enable students to use multiple modalities and representations, students with diverse abilities and language backgrounds may have better opportunities to demonstrate their proficiency than typical print-based assessments (Pellegrino & Quellmalz, 2010). For teachers, technology is well-suited to support implementation by providing scaffolding, data collection, and feedback features needed for effective formative use of assessment. Accordingly, technology-delivered assessments hold tremendous promise for supporting students in demonstrating their learning and for supporting teachers to implement assessments with relative ease and more readily interpret and use assessment information.

In summary, the NGSA project was designed to offer the field critical elements of a technology-supported comprehensive assessment system including a range of assessment tasks that can be used formatively to support science learning for all students.

C. The NGSA Design Team

The NGSA design and development team has been comprised of experts in science education, assessment, psychometrics, and technology from WestEd, the CREATE for STEM Institute at Michigan State University, the Learning Sciences Research Institute at the University of Illinois Chicago, and the Concord Consortium. This group initiated collaborative work in 2013, with an initial focus on developing NGSS-aligned assessment tasks and rubrics for instructionally supportive use in middle-school science classrooms. This was in response to the call for classroom focused assessment development and validation work in the NRC Report on

Developing Assessments for the NGSS (Pellegrino et al., 2014). Since the initial work on middle-school assessment, the collaborative has expanded to include experts from the STEM Education Center at the University of Chicago who have worked with other team members to develop assessment resources for upper elementary grades (3–5) teachers and students.

Across time, the group has worked closely with science teachers from multiple states and districts to develop usable and instructionally beneficial assessment tools that can help teachers better grasp the Framework and NGSS vision and more adeptly plan instruction to move students forward in their science learning. Final products developed by the team include teacher-tested and classroom-ready assessment tasks and rubrics that highlight learning in all three dimensions; guides to help teachers administer and interpret the assessment tasks and results; and an online platform that is searchable and enables teachers to assign tasks to students (individually or groups), monitor and obtain reports of student work, and access various support materials. The NGS System is an open education resource housed in an online platform freely available to schools and districts with the explicit goal of promoting easy access and rapid adoption and use.

D. Development of the NGS System's Resources

The current NGS System was initially developed under the NSF-funded project, Collaborative Research: Designing Assessments in Physical Science Across Three Dimensions (DRL-1316903, 1903103, 1316908, & 1316874). In this project, the collaborative team developed a transformative approach for designing classroom-based assessment tasks that can provide teachers with meaningful and actionable information about students' progress toward achieving the NGSS PEs (See Harris, Krajcik, Pellegrino, & DeBarger, 2019). The approach follows the evidentiary reasoning logic of evidence-centered design (Mislevy & Haertel, 2006) and provides a systematic method for developing a variety of tasks that fulfill the important requirements for NGSS-designed assessment. Central to the design approach is the generation of sets of Learning Performances that establish targets to assess student progress towards mastery of the knowledge and competencies required by the PEs (Harris et al., 2018; McElhaney et al., 2016). The design approach is described in more detail in the following section.

The team used the design approach to iteratively develop tasks and rubrics aligned with a selected set of physical science PEs for the middle-school grade band. They also created the online task portal prototype through which the technology-based tasks could be delivered and used. In this initial work, the team also conducted task performance studies involving over 800 middle-school students (Gane et al., 2018) while also examining classroom use (Pennock & Severance, 2018; Zaidi et al., 2018; Gane et al., 2019). Subsequently, with funding support from the Gordon and Betty Moore Foundation and the Chan-Zuckerberg Initiative, the team completed the development of tasks and rubrics for all the physical science PEs. They also carried out early development work for some PEs in life science (tasks for four of the 21 life science PEs). All told, the team has produced an online bank of nearly 200 tasks designed to align with the middle-school PEs in the physical science and life science domains with accompanying resources. Most recently, with support from another NSF funded project—Collaborative Research: Improving Multi-dimensional Assessment and Instruction: Building and Sustaining Elementary Science Teachers' Capacity Through Learning Communities (Award #1813737 and #1813938), members of the NGSa team from UIC and STEM educators from the University of Chicago developed similar sets of resources for Performance Expectations spanning grades 3–5, including over 45 assessment tasks with accompanying rubrics and other resources.

E. Assessment Development: Design and Validation

NGSA Assessment Design Approach. The NGSa Project's approach to assessment design and development draws from evidence-centered design (ECD; Mislevy & Haertel, 2006). ECD emphasizes the evidentiary base for specifying coherent, logical relationships among the (a) learning goals that comprise the constructs to be measured (i.e., the claims articulating what students know and can do); (b) evidence in the form performances that should reveal the target constructs; and (c) features of tasks to elicit those performances. Using ECD, the design team created a principled approach for developing classroom-based science assessment of tasks that integrate the three dimensions (Harris et al., 2019). This approach allows for systematic derivation of a set of Learning Performances (LPs) from a single PE or bundle of PEs. LPs constitute knowledge-in-use statements that incorporate aspects of DCIs, SEPs, and CCCs that students need to be able to integrate as they progress toward achieving PEs. A single LP is smaller in scope and partially represents a PE. Taken collectively, a set of LPs describes the proficiencies that

students need to demonstrate to meet a PE. The project uses the LPs to guide the development of assessment tasks, evidence statements, and rubrics. Figure 2 presents a screenshot from the Portal showing the resources available to teachers for the Chemical Reactions topic area in middle school. Listed at the top are the three middle-school performance expectations that were bundled together under the Physical Science 1 middle-school topic area given their conceptual interrelationships to create the set of seven Learning Performances listed. Each of the seven Learning Performances covers a part of the conceptual space associated with the performance expectations for chemical reactions and each is stated as a three-dimensional expectation. Next to each Learning Performance is a button that expands to show the descriptions of two or more specific assessment tasks aligned to that specific Learning Performance. Teachers can then preview the sample tasks and find further information about them including rubrics that can be used for scoring student work.

Figure 2.

Illustration of Some Portal Resources for the Middle School Topic of Chemical Reactions.

Chemical Reactions

MS-PS1-2. Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred.

MS-PS1-5. Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved.

MS-PS1-1. Develop models to describe the atomic composition of simple molecules and extended structures.

LP C01: Students analyze and interpret data to determine whether substances are the same based upon characteristic properties.



LP C02: Students construct a scientific explanation about whether a chemical reaction has occurred by using patterns in data on properties of substances before and after the substances interact.



LP C03: Students evaluate whether a model explains that different molecular substances are made from different types and/or arrangements of atoms.



LP C04: Students evaluate whether a model explains that a chemical reaction produces new substances and conserves atoms.



LP C05: Students use a model to explain that in a chemical reaction atoms are regrouped and why mass is conserved.



LP C06: Students develop a model of a chemical reaction that explains new substances are formed by the regrouping of atoms, and that mass is conserved.

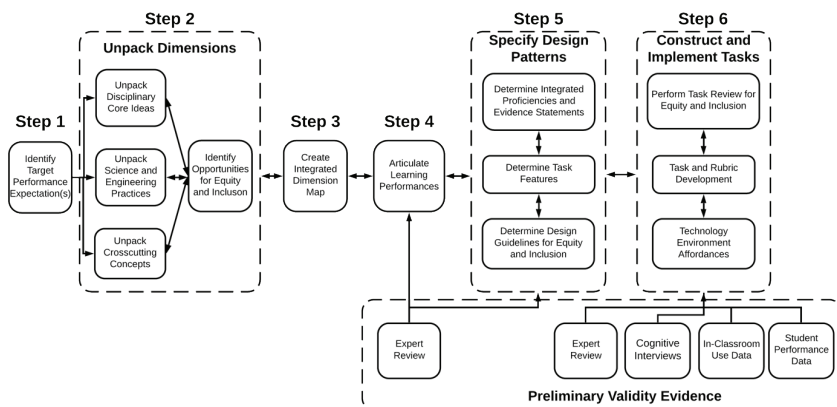


LP C07: Students evaluate whether a model explains that a chemical reaction produces new substances and conserves mass because atoms are conserved.



Figure 3 overviews the six-step design approach that was used to develop the actual tasks (for further information see Harris et al., 2019). Steps 1–3 are a domain analysis that entails unpacking the three NGSS dimensions of a PE(s). For the case illustrated in Figure 2, doing so involves consideration of the three PEs listed for chemical reactions. Unpacking the dimensions of the target PE(s) provides the anchors constituting each dimension and reveals a clear focus for what should be assessed. Integrated dimension maps are then created that provide a visual representation of the target PE(s). Steps 4 and 5 involve constructing Learning Performances such as those shown in Figure 2 and specifying design patterns for tasks associated with them. The integrated dimension map is used to articulate and refine a set of LPs that serve as claims, as they specify what students are expected to demonstrate for evidence that they have achieved one or more aspects of a PE. From each LP, design patterns are derived that include elements to ensure that the tasks elicit evidence of proficiency for the PE, notably evidence statements that articulate the observable features of student performance, equity and fairness considerations for characteristic task features, aspects common to all tasks, and variable task features, such as levels of scaffolding that vary from task to task. The final step in the design process, Step 6, involves using the design patterns to create tasks and accompanying rubrics.

Figure 3.
Overview of the NGS Design Process



NGSA Validation Activities. In parallel with the design and development work, attention is given to the validation of the design products via multiple forms of evidence obtained during the design and implementation process as shown in Figure 3 (See Pellegrino et al., 2016). Detailed discussions of specific validation activities and results for the middle-school physical science and life science assessments can be found in several papers (e.g., Alozie et al., 2018; Gane et al., 2018, 2019; McElhane et al., 2018; Zaidi et al., 2018).

Each stage in the process involves an independent review of products by science and science education experts. They review the integrated dimension maps, and the LPs derived from them. These same experts review the tasks designed to align with each LP and corresponding design pattern. Throughout the process we conduct an equity/fairness review to minimize bias. Once tasks have been through the expert review phases, they are further refined using several steps, including cognitive interviews with students that examine whether tasks are comprehensible and whether they elicit the target performance, collection of classroom performance data to determine applicability and reliability of scoring rules using the rubrics, and classroom studies with teachers who provide design feedback on tasks and help us consider strategies for formative use.

Equity and Inclusion are critical elements that are woven throughout the design and validation process, beginning with (a) the initial domain analysis of the PEs, and continuing through (b) the development of tasks, rubrics, and teacher guides; (c) recruitment of teacher and student participants; and (d) data analyses for validation. Moreover, by conducting the development work with teachers in districts across states that have adopted the NGSS, each serving distinct student populations, the project has been able to further ensure that the tasks and overall system are usable in diverse classroom settings and for broad access and participation.

F. Key Features of the NGSA System

As noted earlier, the NGSA System consists of a library of NGSS-designed tasks, teacher resources for implementing a formative assessment approach, and an online platform for task delivery and access to resources. What follows is some further information on the tasks, the teacher resources, and the open access portal.

NGSS-designed assessment tasks and teacher resources. Each task, anchored in a phenomenon and contextualized within a brief scenario, requires anywhere from 5 to 15 minutes to complete, depending on the requirements of the task. The shorter task duration balances the desire to engage students in authentic science practices with the need for teachers to use the tasks flexibly during instruction and to get timely information from the tasks for formative purposes. Because the task authoring system is web-based it is possible to integrate computational models, which students can manipulate to explore phenomena and generate data. Videos of phenomena, a drawing tool, a system modeling tool, and data analysis tools are also embedded in tasks, providing innovative ways for students to use and apply SEPs, DCIs, and CCCs.

The resources available to teachers include scoring rubrics for pinpointing areas for student feedback and instructional support, strategies for effectively using the assessment tasks in classrooms, and practical guidance for using the NGSA online system. Accompanying each task is a rubric that differentiates levels of proficiency and that includes exemplar responses.

Figure 4 provides an example of a life science task that involves a model for an experiment related to photosynthesis. The middle-school performance expectation is MS-LS1-6. Construct a scientific explanation based on evidence for the role of photosynthesis in the cycling of matter and flow of energy into and out of organisms. The related Learning Performance is Students evaluate how well a model shows that plants and other photosynthetic organisms use energy from the Sun to drive the production of food (sugar) and oxygen.

Figure 4. Illustrative Task Related to the Topic of Photosynthesis

Carmen's leaf model (ID #090-04-p04)

Tap text to listen 

Carmen's class is growing plants. The class wanted to investigate the role of oxygen and sugar in plants. They followed the steps below.

1. On each plant, they chose one leaf and covered it up halfway with foil to block the light it receives. The rest of the leaves on each plant were not covered with foil.
2. They then placed the plants in the sunlight for 1 day.
3. After a full day of sunlight, the foil was removed from the partially covered leaves.
4. All the leaves of the plant were soaked in iodine. Iodine shows that sugar molecules are stored and will turn a purplish-black color.
5. The plant leaves were observed and compared.



Results of Experiment: The image below shows the results of the experiment.

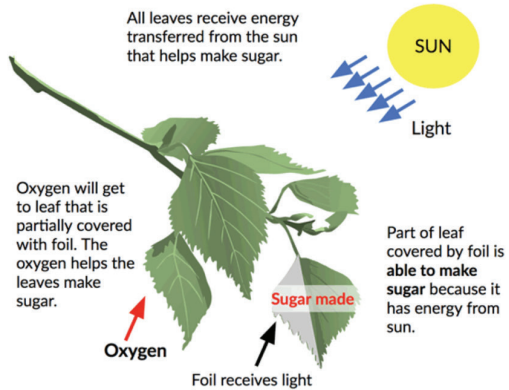
The model that Carmen and her classmates came up with to show their results is shown to the right.

Question #1

Carmen thinks there are errors in how the model explains what happens with sugar and oxygen on the half of the leaf that was covered with foil.

Desc *Interactive content* in the model. Explain how Carmen should improve the model. In your response, use what you know about sunlight and oxygen and sugar in plants.

Please type your answer here.



Task Portal. The online portal (<https://ngss-assessment.portal.concord.org>) houses the current task library and teacher resources and includes a range of features for practitioners and researchers. Teachers can set up classes, assign tasks, receive reports of student work, and gain access to the resources linked to each task. As students work through tasks, their progress can be monitored in real-time. Teachers can review student responses and provide feedback via the portal using rubric-based responses, written notes, or scores. The portal also supports research activities, allowing tasks to be earmarked for research use and can even be tagged for specific research cohort designations.

The NGSA System's assessment tasks and supporting instructional resources for elementary and middle school have been in use in classrooms around the U.S. for several years. The online portal currently has more than 11,000 registered teacher accounts and over 85,000 registered student accounts. Registering an account enables teachers to directly assign tasks to students, access teacher guides, and collect and organize student work. However, to make it convenient for users, the use of the portal and its tasks alone does not require registration, so there is also a substantial "unregistered" user base. Overall, most users are from the U.S., with participation from every state, as well as some international interest with visitors from 126 countries. The user base continues to grow and team members are contacted regularly by teachers and districts with requests to expand the task library to include tasks covering more of the NGSS' elementary and middle grade PEs.

In addition to all the resources contained on the Portal, the team has published a book that serves as a guide for teachers and other educators to develop and use the design process to create similar types of tasks for use in their own classrooms. The volume is published by NSTA Press and titled *Creating and Using Instructionally Supportive Assessments in NGSS Classrooms* (Harris, Krajcik, & Pellegrino, 2024). Finally, the NGSA team has developed an open access website designed to support an ongoing Virtual Learning Community (VLC) for educators interested in the design and use of science assessments for classroom formative use. (<https://www.upinscience.org>). The VLC contains a variety of resources related to the formative assessment process and the use and interpretation of some of the tasks currently found on the Portal.

III. The Stackable, Instructionally-Embedded, Portable Science (SIPS) Assessments Project

In this section we review the rationale and goals of the SIPS project (hereafter the Project) and provide a brief summary of the pilot study that was conducted to test out key ideas for designs for assessing science learning in middle school as discussed in earlier Sections of this paper. We begin by describing the overall design thinking that guided the Project with selected illustrations, and then describe in broad strokes the multi-state pilot study we implemented to demonstrate a proof of concept that end-of-unit assessments could be developed and used by science teachers in their classrooms.

A. Rationale and Goals of the SIPS Project

As noted earlier, release of the NRC Framework and the NGSS standards shifted the focus to emphasize how well students can apply their science knowledge and this in turn has major implications for how assessments should be designed and developed to assess students' science learning (Pellegrino, 2013; Pellegrino et al., 2014). The Project was funded by the US Department of Education under the Competitive Grants for State Assessments Program, CFDA 84.368A. It brought together six states, five educational research organizations, and a panel of experts to address states' growing need for large-scale science assessments, as well as the needs of educators, parents, and students for resources that could support science learning throughout the school year. To meet this challenge the Project set out to build a bank of innovative science assessment tasks designed to measure students' learning that were carefully aligned with curricular and instructional resources to support ongoing instruction over the course of a school year. The term stackable in the Project's title indicates that the assessments can be used together sequentially or in varying orders across the academic year depending on the varying structure and sequence of local science instruction. They were designed to be embedded in the flow of instruction across the year with administration of the assessments proximal to the completion of each of a set of coherent instructional units. And they are portable because they can be used with a variety of science curricula and in a variety of instructional settings in and out of the classroom. The Project focused on grades five and eight as a proof of concept because these are the grades most often targeted in statewide science assessment systems.

To carry out the Project's research and development plan, a collaboration of educational researchers and representatives from departments of elementary and secondary education from six states was organized to carry out the Project. The six states included Nebraska, Alabama, Alaska, Montana, New York, and Wyoming. The educational research team included learning scientists, curriculum and instruction experts, assessment designers, and measurement experts from edCount LLC, the Learning Sciences Research Institute (LSRI) at the University of Illinois Chicago, SRI International, the National Center for the Improvement of Educational Assessment, and the Creative Measurement Solutions group.

B. Approach to Curriculum-Instruction-Assessment Design

The design team was charged with producing a wide range of science assessment resources for public access and use that are coordinated and aligned across all parts of a standards-based system for teaching and learning science that emphasized the interplay of curriculum, instruction, and assessment. The Project was grounded by the idea that to achieve coherence, the Curriculum-Assessment-Instruction (Pellegrino, 2010) connections ought to be balanced among our expectations and plans for student learning, how we carry out science instruction in classrooms, and how we assess students' science learning. With coherence as the guiding principle, the Project identified meaningful bundles of Next Generation Science Standards (NGSS) performance expectations for both grades 5 and 8 and created four instructional unit maps (i.e., instructional frameworks) that covered those expectations. An eighth-grade unit bundle of performance expectations for Force and Energy for grade 8 is shown in Figure 5.

Figure 5.

Eighth Grade Unit Bundle of Performance Expectations

| NGSS Grade 8 Unit 1: Forces and Energy |
|--|
| Bundle 1 |
| MS-PS2–2. Plan an investigation to provide evidence that the change in an object's motion depends on the sum of the forces on the object and the mass of the object. |
| MS-PS2–1. Apply Newton's Third Law to design a solution to a problem involving the motion of two colliding objects. |
| MS-PS3–1. Construct and interpret graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object. |
| MS-PS2–4. Construct and present arguments using evidence to support the claim that gravitational interactions are attractive and depend on the masses of interacting objects. |

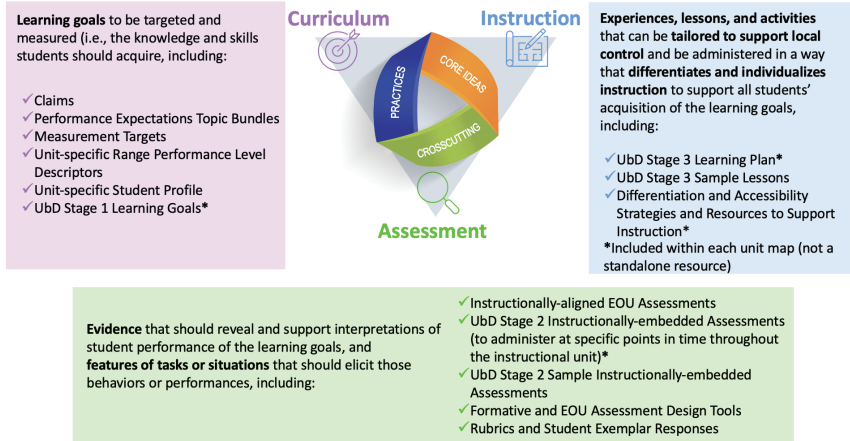
For each unit, a unit map was created, and it encompassed a suite of interconnected and coherent curriculum, instruction, and assessment resources, all designed to support high-quality, three-dimensional science teaching and learning along a year-long instructional pathway. Figure 6 provides an overview of the design logic and lists the design elements and products generated under each of the Curriculum, Instruction, and Assessment components of the Unit design process. Figure 7 provides an illustration of the specific sets of resources created for the eighth-grade unit on Forces and Energy. Similar resources were created for all four eighth-grade units and all four fifth-grade units. All resources for each unit at each grade level can be accessed at the SIP Project website.

(<https://sipsassessments.org/resources/>).

Figure 6.

Overview of the Sets of Resources Created for Each Instructional Unit.

Coherent Sets of C-I-A Resources were created for each of 4 NGSS-aligned Instructional Units at each of Grades 5 and 8



<https://sipsassessments.org/resources/>




Figure 7.
Illustration of the Resources Created and Available for the 8th Grade Unit on Forces and Energy.

Grade 8 Unit 1: Forces and Energy

The Grade 8 Unit 1 topic, "Forces and Energy," organizes the Next Generation Science Standards performance expectations with a focus on helping students develop an understanding of the motion of objects and how interactions between objects can be explained and predicted.

Grade 8 Unit 1 Curriculum, Instruction, and Assessment Resources

Unit Map / Instructional Framework ([.docx](#), [.pdf](#))

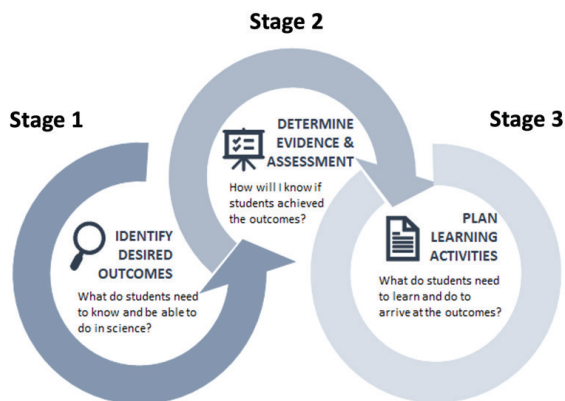
| | |
|--|---|
| Curriculum  | <ul style="list-style-type: none"> • Claim, Measurement Target, and PE Bundle (.docx, .pdf) • Storyline Overview (.pptx, .pdf) • Student Profile (.docx, .pdf) • Policy and Range Performance Level Descriptors (.docx, .pdf) • Stage 1 Learning Goals* (See Unit Map / Instructional Framework) |
| Assessment  | <ul style="list-style-type: none"> • Stage 2 Instructionally-embedded Assessments* (See Unit Map / Instructional Framework) • Designing Equitable Assessments for Diverse Learners (.docx, .pdf) • Sample Instructionally-embedded Assessments: <ul style="list-style-type: none"> • Segment 3: "Kinetic Energy vs. Mass/Speed Investigation" <ul style="list-style-type: none"> • Task Specification Tool (.docx, .pdf) • Task (.docx, .pdf) • Segment 4: "Designing Solutions to a Problem Involving a Collision" <ul style="list-style-type: none"> • Task Specification Tool (.docx, .pdf) • Task (.docx, .pdf) • End-of-Unit Assessment (.docx, .pdf) <ul style="list-style-type: none"> • Assessment Scoring Guide (.docx, .pdf) • Design Tools: <ul style="list-style-type: none"> • Unpacking Tool (.docx, .pdf) • Design Pattern (.docx, .pdf) • Task 1 Specification Tool: "Storing Grocery Carts" (.docx, .pdf) • Task 2 Specification Tool: "Barriers on the Highway" (.docx, .pdf) • Task 3 Specification Tool: "Roller Coaster Thrills" (.docx, .pdf) |
| Instruction  | <ul style="list-style-type: none"> • Stage 3 Learning Plan* (See Unit Map / Instructional Framework) • Differentiation Strategies and Resources (.docx, .pdf) • Sample Lessons: <ul style="list-style-type: none"> • Segment 1: "Newton's Third Law" (.docx, .pdf) • Segment 2: "Getting to the Bottom of Newton's Second Law" (.docx, .pdf) |

*Embedded within Unit Map / Instructional Framework

To move forward with this integrated design framework, the Project team drew on two heretofore and largely distinct approaches—a curriculum and instruction development approach known as Understanding by Design (UbD) (Wiggins & McTighe, 2005) and the principled assessment design framework called Evidence Centered Design (ECD) discussed earlier and developed by Robert Mislevy and his colleagues (e.g., Mislevy, Haertel, Riconscente, Rutstein & Ziker, 2017).

Understanding by Design (UbD). The Project partners developed a prototype science curriculum framework based on the Understanding by Design (UbD) model of curriculum design. UbD uses a multi-stage method of backward planning that begins with a statement or vision of the desired results—the learning goals—and works backward to identify the assessment evidence needed to support inferences of student learning (See Figure 8). UbD calls for careful planning of the curriculum sequence and pedagogical tools and activities to achieve those stated learning goals. The UbD approach ensures that teachers are deliberately planning their lessons with a focus on the expected learning objectives and performance expectations of each of the science instructional units. Furthermore, UbD provides a framework for aligning the assessment design with the taught curriculum and the sources of evidence of student learning. A more complete description of UbD is beyond the scope of this chapter and the interested reader can find a richer description of this approach in Wiggins and McTighe, 2005.

Figure 8.
Simplified Representation of the three Stages of the Understanding by Design Framework



Source: Adapted from Wiggins, G.P. & McTighe, J. (2005).

Evidence Centered Design (ECD) and End-of-Unit Assessments. To design end-of-unit (EOU) assessments in a way that ensures alignment with the curricular frameworks and the relevant instructional resources the design team adapted a principled assessment design approach, i.e., ECD, to design and develop each of the Grade 5 and Grade 8 assessments (Mislevy & Haertel, 2006; Mislevy, Haertel, Riconscente, Rutstein & Ziker, 2017). Like the approach described earlier for the NGSA project, the team addressed these three key design questions: 1) what constructs do we want to measure; 2) what evidence is needed to make inferences about students' ability related to those constructs; and 3) how can tasks be designed to collect the desired evidence? Other explicit design criteria included the need to administer the EOUs at the end of completion of each of four instructional units—approximately every 10–12 weeks of science instruction; and they had to be administered by teachers within one 50-minute class session. Again, a more detailed description of the ECD methodology is beyond the scope of this chapter. The interested reader can find more thorough descriptions of this approach in the early work of Mislevy and Haertel (2006) and Mislevy & Riconscente (2006).

The ECD approach led us to compose each EOU assessment as a set of three sub-tasks, each containing multiple prompts (i.e., test items). The component tasks were designed to measure well-defined science constructs based on a clearly articulated theory of science learning. The aim was that any given assessment would produce evidence of students' science learning in terms of the NGSS performance expectations (PEs) that were the focus of the associated instructional unit. They were meant to provide a summative characterization of student learning as an outcome of the immediate prior instructional unit, as well as to inform the content and focus of subsequent instructional units. The evidence produced by the EOUs, by design and following the NGSA system described earlier, would support inferences about students' proficiency in integrating Scientific and Engineering Practices (SEPs) with important Disciplinary Core Ideas (DCIs) and Cross-cutting Concepts (CCCs) to scientifically investigate and understand natural phenomena and solve important science and engineering design problems. To make the multi-dimensional assessment design feasible, the design team defined proficiency and determined bundles of PEs that could be taught and measured together and that would meaningfully represent the scope of an instructional unit.

Each EOU assessment measured the key knowledge, skills, and abilities (the KSAs) as represented by a thorough unpacking of the PEs within the associated

instructional unit bundle identified during the UbD analysis process. Each PE was a combination of three dimensions: the disciplinary core ideas (DCI), science and engineering practices (SEPs), and cross-cutting concepts (CCC). Each of these dimensions was not unique to a given PE (e.g., the same scientific practice appears in multiple PEs), but the PE uniquely defines one combination of the three dimensions.

Another key step in the process required the design team to collaborate with the science teachers to develop a set of performance level descriptors (PLDs). These descriptors organized multi-dimensional statements into levels representing different levels of student performance. The PLDs provided statements that are at a finer grain size than the overall claim and provided further insight into what is to be measured on the assessment. Once the PLDs were developed, the design team created task design patterns for each PE in the instructional unit bundle.

In practice the design patterns provided task designers with a menu of options to use when designing tasks aligned to the PEs. The design patterns and PLD documents provided guidance on what should be measured, as the PLD statements and the KSAs describe the measured concepts related to the bundle of PEs. The design patterns also provided information on what evidence is needed to measure these concepts (through the demonstration of learning). Once the design team established the design patterns, the next step was to determine how to measure these concepts.

Like all educational assessments, the assessments developed in this Project had constraints on their design; specifically, they needed to be able to be completed in approximately one class period, and they needed to be administered as paper/pencil tasks. With these constraints in mind, each EOU assessment consisted of three tasks, each using one scenario and/or phenomenon, and a set of questions related to that phenomenon. Another critical design feature for measuring three-dimensional science standards is to engage students in a chain of sense-making. Therefore, the set of prompts within each task required students to engage with different aspects of the scenario and meet the expectation of increasing the complexity of the required response. The design team anticipated that each individual task would take students 10 to 15 minutes to complete, and consequently, determined that each EOU assessment would consist of three tasks.

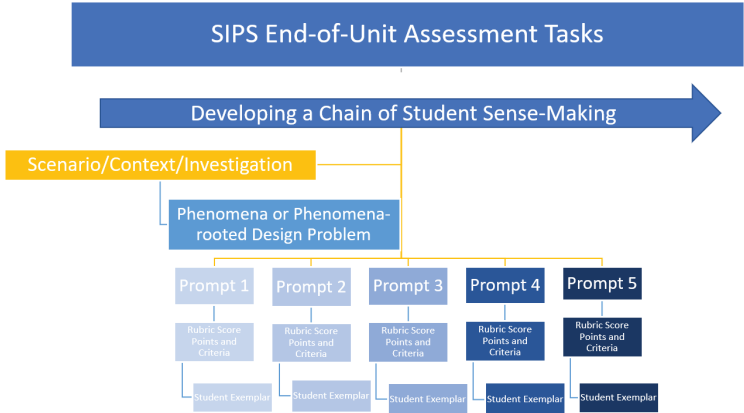
As noted previously, each EOU assessment consisted of three tasks. To provide further specifications for each task as part of an ECD approach, the design team created task specifications. Each task specification tool provides specification for the following:

- List of performance expectations covered in the task (each task covers one to two PEs);
- Information on the phenomenon or phenomenon-rooted design problem: Each task is rooted in a phenomenon or design problem related to the PEs;
- Scenario: Each task requires a scenario or situation which would make sense to students, be coherent and understandable to students, and provide enough context to allow students to engage meaningfully with the task;
- Variable Features: A list of features (or decision points) that could be modified to shift the complexity and/or focus of the task while still measuring the PEs;
- Chain of Sensemaking: An overview of the flow of the task, including the alignment of different sections to the KSAs;
- KSAs: A list of the KSAs that are targeted by the task, including any additional (not from the original set of design patterns) KSAs that are a cross between two PEs;
- Student Demonstration of Learning: A list of the expectations of students taken from the design patterns;
- Work Products: A list of the physical responses that students might produce;
- Application of Universal Design for Learning-based Guidelines: A set of guidelines to promote equity and inclusion in the task design; and
- SIPS Complexity Framework Components: A description of how the prompts for the task are designed to align with the degrees of sophistication represented by the complexity framework.

The task specification tool described the design elements of the task and provided guidance to task developers. This information was used to further develop the tasks. Each task is aligned to one or two PEs and is situated in a given phenomenon or design problem. The phenomenon was situated in an overall scenario and scaffolded such that students were provided a foundational context, the context is then problematized, and then students engage with the context through a series

of prompts or questions. The scenario had to make sense to students, be coherent and understandable, and provide enough context to allow students to engage meaningfully with the task. Again, leaning on the UbD approach, each task included rubrics that clearly defined what was required of students and how evidence from students could be evaluated. Figure 9, below, shows the components of an EOU assessment task.

Figure 9.
Illustration of the Components of an EOU Assessment Task



The EOU development process described above was used to produce eight prototype EOU assessments—four each at grade 5 and grade 8, all of which were intended to be administered after approximately 8 to 10 weeks of instruction (i.e., following each of the SIPS instructional units in each grade). Each assessment contains three multi-part tasks which are scenario/phenomena based and are designed in a way that students engage with sense-making as they move through the task.

To the extent possible, the task scenarios were based on a phenomenon or design problem that occurred outside of the classroom and has local or global relevance. However, given variation in curricular and instructional resources used across states and districts, SIPS partners acknowledge that tasks address phenomena

or phenomena-rooted design problems that may or may not have been addressed through instruction.

The tasks designed for each EOU were meant to be illustrative examples of (1) PE bundles and (2) task scenarios. Additional tasks can be designed using the SIPS design process to support use with other SIPS unit sequences or other curricula. While the EOUs were designed to be administered in the recommended order of the SIPS instructional units, if educators taught the instructional units in a different order then the assessments may be administered in the sequence that best aligns with instruction. Scoring for these assessments would be the same regardless of the order in which they are administered.

While not every prompt had to cover every dimension in the PE cluster, every dimension within the unit's PE bundle had to be aligned to at least one item on one task on the EOU assessment. Once tasks were developed, the design team reviewed the tasks for alignment against the task specification tool, ensuring coverage of the KSAs specified in the tool. Tasks were also reviewed for clarity, sense-making, accessibility and fairness, and the degree to which they require sense-making. Feedback was obtained from teachers as well as from outside experts and included reviews of the tasks as well as the scoring rubrics (described below). The Project design team applied revisions to the tasks based on this feedback.

Rubric Development. Scoring rubrics for each task were developed in conjunction with our science teacher partners to highlight aspects of the student responses that demonstrate understanding of the concepts. The scoring rubrics included evaluative criteria to support the evaluation of evidence for each prompt (or a set of sub-prompts) within each task and were developed based on the student demonstration of learning from the task specification tool. The number of score points possible for each prompt or set of sub-prompts varied from one to four points depending on the expectations of students.

Rubrics were designed with the expectation that teachers would be the primary users of the rubrics. Each score point was defined to provide clear guidelines of the differences between student responses that fall in each score point. Rubrics also cover the range of possible student responses and are specific to the given prompts as this allows for more guidance for scorers. Once the rubrics and tasks

were developed, the SIPS team aligned them back to the PLD descriptors, ensuring that the tasks and rubrics are focused on aspects of the PLDs that are deemed important and that the set of tasks as a whole cover the critical aspects of the PLDs. The SIPS team applied revisions to either the tasks or the PLDs (as concepts of the PLDs changed throughout the development process).

C. Pilot Study Overview and Results

To collect evidence about the validity and utility of the EOU assessments, a small pilot study was designed to focus on three overarching research questions: (1) to what degree do the EOU assessments, generally, provide evidence of students' three-dimensional science learning?; (2) how well do latent variable measurement models fit the empirical EOU assessment data?; and (3) overall, what do the EOU assessment results tell us about students' science learning? To address these issues, we recruited at least five classrooms of students from each state—aiming for a mix of grade 5 and grade 8 classrooms. See Table 1 for an overview of the teachers and students who participated in the pilot study.

Table 1.
Number of Educators and Students Included in the Pilot Study by EOU Assessment

| EOU Assessment | Number of Teachers | Number of Students |
|----------------|--------------------|--------------------|
| Grade 5 Unit 1 | 23 | 341 |
| Grade 5 Unit 2 | 28 | 473 |
| Grade 5 Unit 3 | 19 | 341 |
| Grade 5 Unit 4 | 26 | 417 |
| Grade 8 Unit 1 | 14 | 151 |
| Grade 8 Unit 2 | 10 | 189 |
| Grade 8 Unit 3 | 13 | 258 |
| Grade 8 Unit 4 | 4 | 51 |

The main requirement for educators to participate was teaching a curriculum aligned to three-dimensional science standards (e.g., NGSS standards or similar). In the end, the Project recruited 121 educators from across four states that expressed initial interest in participating in the pilot. Of those 121 educators, 63 educators representing three of the six partner states participated in the study by administering one or more EOU assessments.

Summary of Findings from Pilot Study. It is important to note that the study was designed as a pilot of a limited set of initial prototypes of each of the four end-of-unit (EOU) assessments administered to samples of 5th and 8th graders. We organized our findings around the three research questions that animated the general design of the pilot study. Our goal throughout was to collect information related to each of the guiding research questions to support, ultimately, revisions to the prototypes and to learn more about how three-dimensional end-of-unit tasks could be used in practice by teachers.

Our first research question focused on the utility of the EOU assessments for providing evidence of students' three-dimensional science learning. We collected information related to whether it was appropriate to use the EOU assessments for measuring students' science learning. What we found, briefly, is that while students were able to demonstrate science knowledge, there were some issues with the initial versions of the prototype assessments. Given that our plan was for each EOU to be administered in one class period, we discovered that substantial revisions to the tasks were needed because most tasks took students more than 20 minutes to complete, which meant, for the most part, students could complete only two of the three EOU tasks in a class period. While we expected to see some degree of missing responses from students, the number of missing responses by prompt (i.e., test item) was often much higher than we expected. Some of this may be because students simply ran out of time. We also found that several full classrooms skipped certain prompts or tasks within an EOU, suggesting that there were certain science topics that students were not familiar with or were not able to engage with on the assessment as intended.

Overall, the prototype EOUs were challenging for students in our study. While there were two assessments for which students were able to achieve the highest possible points, for most assessments, students fell short. The prototype EOUs did provide information about where students stood with respect to the rubrics scoring scheme used, and they also allowed us to measure variation in students' achievement as we

found prompts, tasks and EOU scores distributed across a range of performances. Importantly, based on the data obtained the Project has subsequently made adjustments to the timing and difficulty levels of the prototypes.

Further study will be needed to determine how well the end-of-unit assessments were able to reflect students' opportunities to learn. Throughout the pilot study teachers reported on whether they taught a particular topic, but there was no information on how deeply they went into a topic or how the topic was taught. While we found some evidence of differences in scores based on if teachers indicated they taught a given concept or not, these differences did not always favor the students who received instruction related to this concept. However, this could be due to differences in the organization of classrooms, or to the degree or depth to which the concept was taught.

Finally, while teachers were able to provide scores on student work, further study is needed to determine the reliability of these scores, particularly if the goal is to compare students across classrooms. While data on scores from different teachers on the same set of students were collected, these data were limited, and we saw differences in the overall reliability of scoring depending on the prompt or task being scored. While the limited pilot study data indicate we were able to see differences between and among students, and that some students were able to demonstrate their science knowledge, further information on how future iterations of the assessments will be used in the classroom need to be gathered to guide additional explorations into the design and use of the assessment tasks.

Our second research question asked if we could develop latent variable measurement models that fit the empirical EOU assessment data. Each of the prototype EOUs was scaled separately using the Rasch model, i.e., a one parameter IRT model. This modeling approach produced reasonable estimates of the items' difficulty parameters and student ability estimates. When using the Rasch model, item (or prompt) fit statistics were estimated which, in turn, proved useful for evaluating the measurement quality of the EOU prompts. Further, these fit statistics offered insights into the relationships among students' abilities and their responses to specific EOU prompts. More specifically, the fit statistics generated by the Rasch model measured the appropriateness of a prompt's difficulty relative to the students' abilities. Lower than expected values indicated that the prompt may have been too easy for our sample of students, leading to a high probability of correct

responses. Conversely, a higher-than-expected value suggested that the prompt may be too difficult. This model fit information was shared with the designers of the prototypes as they worked to improve the measurement quality for the next iteration of the EOU assessments.

The Rasch model fit statistics allowed us to evaluate the fit of a prompt or task in a more general sense, i.e., reflecting how well a prompt performs across the entire student ability spectrum. The use of latent variable models, like the Rasch model, allowed us to identify prompts that performed erratically suggesting that students' performance on the prompt may have been influenced by factors other than the students' abilities, such as guessing or simply misunderstanding the prompt. With this approach we were also able to flag prompts that were too predictable and, therefore, did not discriminate sufficiently among students with different abilities. In sum, our approach to latent variable modeling provided rich information about the measurement characteristics of the prototype EOUs. Unlike typical statewide assessment programs used for accountability purposes, IRT derived scale scores did not play a major role in this pilot, and thus were not computed based on a theta to scale score conversion formula.

Our third and final research question had to do with what the EOU assessment results tell us about students' science learning? As part of the investigation into this research question we examined the relationship between student scores and additional variables, including gender, prior ELA and Math learning, and curricular materials. We found that three out of the eight EOU assessments had statistically significant differences based on gender (in favor of females), but the sample size for this was low and so further study is needed to draw more solid conclusions. We also found that scores on the assessment tended to increase as prior ELA and mathematics levels increased. While this could indicate a dependency between ELA and math ability and the science assessment, there is often overlap between the science practices and ELA skills (e.g., communicating information) as well as the science practices and mathematical practices (e.g., problem solving). Therefore, more exploration is warranted to determine if there is too much of a dependency among and between skills.

Our analysis found statistically significant differences between students who used different curricular materials at the 5th grade (and for the Grade 8 EOU 2 assessment). However, without further investigation of the differences among the different curricula materials it was not clear how to interpret these differences. Further investigation to determine if the differences are due to desirable

characteristics (e.g., if different curricula cover different aspects on the assessment, we would naturally expect different scores) or to characteristics we would want to address in the assessment (e.g., if different curricula use different representations and the assessment is too closely aligned to one specific representation).

Cross-EOU Growth. The pilot study sample was modest—not all students in a grade took all four EOUs. Nonetheless, 64 5th graders and 21 8th students took all four EOUs. Based on these limited data we found that an increase in performance level from EOU to EOU reflected growth in students' learning because (a) each EOU had a unique set of performance level descriptors (PLDs) that form the basis for the task-PLD alignments and score estimations and (b) each level of each EOU's PLDs reflected a common expectation for student performance relative to the EOU's instructional unit. For example, PLD level 3 reflected the minimal performance expected of all students following each instructional unit. Thus, each level was qualitatively comparable across the four EOUs. In summary, the calibration of each level of the PLDs to a common goal relative to the instructional unit supports the measurement of cross-EOU growth. The current study had a limited number of cases from which to evaluate the efficacy of the proposed growth metric—change in performance level from EOU to EOU. It is recommended that the efficacy of this approach be further evaluated when a more robust data set is available.

Reporting of the EOU results. In the case of the pilot study, teachers scored their own students, and thus had access to student level data. However, no additional data were reported back to teachers about their students, and additional guidance on how this information could be used to inform subsequent units of instruction were not provided. Nevertheless, the pilot results suggest EOUs scores could be used to report back to teachers. We explored whether two different reporting metrics might be used to summarize individual student performance for each EOU and aggregated across EOUs.

Students could receive a reportable performance level based on each administered EOU. These performance levels, for example, may be used for reporting individual student results from multiple EOUs. Profiles can be summarized at the individual student level by reporting performance level profiles in both tabular and graphical formats. Performance level results can also be reported at the group level for each EOU. Group level performance level results are typically reported as the percentage of students in the group attaining each level. Multiple EOU administrations can be reported at the group level by reporting the percentage of students in the group

achieving each level on each EOU in both tabular and graphical (e.g., stacked bar chart) formats. Performance level reports for multiple EOU administrations over the course of the year can be supported via Performance Level Profiles. For example, a rubric may be adopted that links students' four EOU performance level profiles with an overall performance level.

It is important to note that we did not have a common scale across EOUs in a grade. However, performance-level based scores can be reported for each EOU and aggregated across EOUs to support within-grade, cross-EOU score interpretation based on the following rationale: Each EOU has a unique set of PLDs that form the basis for the Task-PLD alignments and cut score estimation and each EOU's PLD level reflects a common expectation for student performance relative to the EOU's instructional unit. PLD-based scores can be averaged on individual student reports to summarize multiple EOU administrations. Group level scores can be reported as an average of the individual students' PLD-based scores.

Educators may use the PLDs to inform subsequent units of instruction. That is, educators are able to review the descriptor for a student's current level of performance on an EOU—this tends to describe the range of performance for students achieving that level. However, by examining the next higher level, the educator can observe the skills the student needs to acquire to advance to that higher level. While the subsequent unit of instruction may be quite different, the information obtained from such a review may provide insight into students' strengths and weaknesses to inform the next unit of instruction—see below for a brief description of the subsequently funded CASCIA Project's interpretive resources that were developed for each revised EOU.

D. Summary of SIPS' Accomplishments

SIPS was an ambitious project in pursuit of multiple goals, primary among them is integration of science curriculum, instruction, and assessment resources for multiple instructional units at each of two grade levels. Among its accomplishments was the integration of two major conceptual and principled design frameworks—Understanding by Design and Evidence Centered Design—to guide the creation of Curriculum-Instruction-Assessment Unit materials and Design & Development Tools together with a multitude of specific resources for each C-I-A element of eight science learning units. Despite its limitations, the Pilot study data collection was sufficient for determining the quality and variability

of student performance on challenging, multi-dimensional science assessment tasks. The data collection also proved sufficient for providing evidence regarding: (a) teacher capabilities for reliable task administration and scoring, (b) challenges students face in task completion time and comprehension, (c) guidance for EOU task revision and scoring for subsequent use and validation, (d) EOU basic measurement properties, (e) exploration of alignment of performance with claims associated with embedded standard-setting processes, and (f) suggesting ways to evaluate year-long performance.

Since the completion of SIPS, a follow-on project called CASCIA, also funded by the U.S. Department of Education and involving some of the original SIPS partners, has pursued EOU assessment revision based on the SIPS pilot study results together with the development of interpretive guides and resources for each of the revised EOUs. It is beyond the present chapter to describe the work being done in the CASCIA project to validate the EOUs and interpretive resources, as well as what they are learning about classroom implementation of the instructional units and EOUs. However, it is useful for present purposes to provide an illustration of the types of interpretive resources that have been created to support multiple stakeholders for understanding and using results from the EOUs. Figure 10 is an illustration of the types of interpretive resources CASCIA has designed and is making available, who they are directed towards, and their intended interpretive use. Further information about these resources and other findings regarding their use should be directed to members of the CASCIA Project team via edCount LLC.

Figure 10.

Examples of the Reporting Mechanisms Developed by the CASCIA Project.

| Reporting Mechanism | Audience | Proposed Purpose / Uses |
|--|--|--|
| Individual Score Report (ISR) | <ul style="list-style-type: none"> • Students • Parents/Guardians • Educators | <ul style="list-style-type: none"> • Summarize individual student performance on the end-of-unit assessment that can be used to monitor student progress and plan meaningful learning opportunities to ensure students are on track to achieve end-of-year learning goals in science. |
| Classroom Roster Report (CRR) | <ul style="list-style-type: none"> • Educators • Administrators | <ul style="list-style-type: none"> • Summarize student performance by classroom on the end-of-unit assessment and offer information about students' instructional needs levels that educators can use to inform a variety of individualized, small, and whole group learning opportunities and make timely and meaningful adjustments to instruction. |
| Interpretive Guidance and Instructional Strategies | <ul style="list-style-type: none"> • Educators | <ul style="list-style-type: none"> • Provide information to help educators understand their students' performance on the end-of-unit assessment and offer instructional strategies and resources for planning and adjusting instruction to help students learn. |
| Family Guidance and Learning Resources | <ul style="list-style-type: none"> • Parents/Guardians • Students | <ul style="list-style-type: none"> • Provide information to help families understand their student's performance on the end-of-unit assessment and offer resources recommendations for engaging their student in science learning home. |
| Task Interpretation Guide | <ul style="list-style-type: none"> • Educators | <ul style="list-style-type: none"> • Provide information to help educators understand the assessment tasks and prompts, their features, and the evidence they are designed to elicit about student learning, and to reflect on prior |

IV. Lessons Learned and Implications for Future Work on Assessments to Support Teaching and Learning in Science

We began this chapter with a description of the changes in expectations for student knowledge and learning in science as signaled by the 2012 NRC *Framework for K–12 Science Education* report and the derivative 2013 *Next Generation Science Standards*. In addition to describing multiple dimensions of knowledge—Disciplinary Core Ideas, Crosscutting Concepts, and Science and Engineering Practices—these reference documents specified ways of knowing in the form of multi-dimensional performance expectations requiring their integration. The goal is to have knowledge capable of explaining scientific phenomena, solving problems, and designing solutions to challenges posed by the natural and designed world in which we live. The ensuing decade has seen multiple efforts to articulate the instructional and assessment challenges posed by this contemporary framing of science proficiency. The two projects we have overviewed in this chapter represent some of the many attempts to address these challenges with a particular focus on assessment design, implementation, and interpretation for students in grades K–8. What follows are some reflections on what has been learned and issues that remain to be addressed by the science education research, development and practice communities.

A. Challenges of Multidimensional Science Assessment Design

Early on, the challenges of multi-dimensional science assessment design were duly noted, and recommendations were made that developing valid and reliable assessments for formative or summative use in classrooms and for large-scale assessment at state levels would require application of a principled approach to assessment design. The NGSAs and SIPS projects are illustrations of the benefits that accrue from following such advice, emphasizing application of the Evidence-Centered Design framework articulated by Mislevy and his colleagues. The assessments designed within each project have well specified claims as to what knowledge and skills are being assessed and what evidence is required in student responses to support proficiency. The design patterns and item specifications are transparent allowing for the tasks to be reviewed by experts as to their validity and the interpretability of student performance. By following a principled design process, the stages of which have been articulated in both projects for their respective tasks, others can use these design tools to develop new tasks aligned to multiple aspects of the Framework or NGSS for various grade levels and content areas.

B. Challenges of Interpreting and Scoring Multidimensional Science Performance

One thing that we have not focused on in our discussion of the assessments developed under each project is the issue of how best to interpret performance on the types of multi-dimensional tasks developed by each project. Given that the tasks and performances are supposed to be multi-dimensional, many educators and assessment designers advocated for the production of “separate” scores for each of the dimensions represented in the task. For example, a score for the disciplinary content and a score for the science and engineering practice. We, however, have viewed such an approach as inappropriate and antithetical to the presumption of integrated knowledge that is useable. Thus, in both projects, the interpretation of student performance focuses on evidence of integrated proficiencies that vary in their sophistication relative to the target proficiency for the given task. This avoids sending a message to educators that instruction should focus on the dimensions as separable targets and maintains an instructional focus on dimensional integration during instruction. Based on our experience with teachers using our tasks, we continue to believe that this approach to interpretation and scoring is far more meaningful and useful for both formative and summative interpretive uses.

C. Challenges of Integrating Curriculum, Instruction and Assessment in Science

What educators need to advance their own instructional practice and their students learning in the ways demanded by the Framework and NGSS are coherent and integrated curricular, instructional and assessment materials and resources. Unfortunately, the vast array of science education resources available to teachers since the appearance of the Framework and NGSS are curricular materials with weak and inadequate assessment materials for formative and/or summative classroom use. The development of assessments for most curricular products is largely an afterthought with little to no attention to assessment development using a principled approach such as ECD. One of the major contributions of the NGSA and SIPS projects is bringing curriculum, instruction and assessment together to achieve greater coherence in the classroom. In the NGSA project this has come about by working with teachers to integrate the various tasks into their curriculum and instructional unit materials by providing explicit guidance as to what is being assessed and where it fits with respect to a progression of learning anchored against the NGSS performance expectations. The SIPS project has directly taken on the coherence and integration challenge by bringing together the Understanding by Design curriculum and instruction design framework with the Evidence-Centered assessment design framework. Thus, while SIPS does not claim to provide a complete curriculum, instruction and assessment “package”—a so called “shrink wrapped” solution—it does provide a wealth of resources that teachers can adapt to their contexts and needs as well as tools and examples for how this can be done for other units of instruction at varying grade levels. We cannot underscore the degree of challenge that the SIPS project encountered in bringing these design frameworks together and the benefits that have accrued in terms of the materials and models that have resulted.

D. Benefits of the Work

No project in the science education field can begin to address all the issues related to the teaching, learning and assessment of science proficiency as it has now been envisioned. Each of the two projects described here have limitations with respect to scope of the problems addressed and degree of contribution. Nevertheless, we offer the perspective that much has been accomplished for multiple audiences and stakeholders.

For the Assessment Design and Development field writ large, and in science education specifically, much has been learned about how to conceptualize and execute the design process for multi-dimensional tasks. Models have been developed in both projects that can be deployed by others and modified as needed to create new tasks whether they be multi-dimensional tasks requiring the integration of mathematics practices and content as required by contemporary mathematics standards, or for additional tasks and task types for science education use, including those that can be used on large scale state assessment and/or for classroom or state performance assessments.

For Educators, including State education and assessment leadership teams, District C-I-A leadership teams, and Classroom teachers, both projects provide specific resources that are ready for deployment as well as models and practice guides to support professional learning and additional resource development. We know that the NGSA resources are being used by thousands of teachers as part of their classroom practice and many are using the design guidance to develop new tasks and interpretive tools. We also know that educators in multiple states, including the lead state of Nebraska, are using the SIPS resources for ongoing instruction and as professional development resources with multiple districts.

Finally, one of the most important benefits of the work of both projects is for Students in our K–12 classrooms across the country. Students (and their teachers) now have challenging tasks that can help them develop an understanding and appreciation of what is expected of them with respect to science proficiency. When our assessments are used wisely with constructive feedback from their teachers, students can gain proficiency and confidence in their science learning. Hopefully, they can come to appreciate more fully the elegance of science as a disciplinary activity that goes beyond memorization of facts and procedures and see it as a way to understand their world and guide their personal decision making in many facets of life.

E. What's Needed and What's Next?

We have alluded to some of the many things needed in the field of science assessment and for these two projects. Perhaps the best way to sum up and consider what's next is with respect to concerns regarding validity. Any science assessment effort, whether it be the NGSA tasks designed for classroom formative use, or the SIPS EOU assessments designed for classroom and potential large-scale state use, a primary concern is evidence regarding the intended interpretive use of the resources.

While each project obtained various forms of evidence related to their validity arguments, much remains to be done. The evidence needed is of multiple forms and goes well beyond traditional quantitative measurement or psychometric results (e.g., Pellegrino, DiBello, & Goldman, 2016). While the latter are needed as part of the validation argument, far more of the desired evidence will come from the world of practice. In particular, we need to know far more about how and how well educators can use the NGSA and SIPS resources to impact their practice and consequentially the learning of their students. We are hopeful that future projects making use of the NGSA and SIPS resources will provide many aspects of that evidence.

References

- Alozie, N., Haugabook Pennock, P., Madden, K., Zaidi, S., Harris, C. J., & Krajcik, J. S. (2018, March). *Designing and developing NGSS-aligned formative assessment tasks to promote equity*. [Paper presentation]. Annual conference of the National Association for Research in Science Teaching, Atlanta, GA.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement*. (CPRE Research Report Series No. RR-43). Consortium for Policy Research in Education, University of Pennsylvania.
- Davies, S. (2010). *Effective assessment in a digital age: A guide to technology-enhanced assessment and feedback*. JISC Innovation Group.
- Fuhrman, S. H., Resnick, L., & Shepard, L. (2009). Standards aren't enough. *Education Week*, 29(7), 28–29.
- Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education*, 101(5), 854–867.
- Gane, B. D., McElhaney, K. W., Zaidi, S. Z., & Pellegrino, J. W. (2018, March). *Analysis of student and item performance on three-dimensional constructed response assessment tasks*. Paper presented at the 2018 NARST Annual International Conference, Atlanta, GA.
- Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multi-dimensional learning. *European Journal of Education*, 53(2), 176–187.
- Gane, B. D., Zaidi, S. Z., McElhaney, K. W., & Pellegrino, J. W. (2019, April). *Design and validation of instructionally-supportive assessment: Examining student performance on knowledge-in-use assessment tasks*. Paper presented at AERA Annual Meeting, Toronto, ON, Canada.

- Gorin, J. S., & Mislevy, R. J. (2013, September). Inherent measurement challenges in the next generation science standards for both formative and summative assessment. In *Invitational research symposium on science assessment*. Educational Testing Service.
- Harris, C. J., Krajcik, J. S., & Pellegrino, J. W. (2024). *Creating and using instructionally supportive assessments in NGSS classrooms*. National Science Teachers Association.
- Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2), 53–67.
- 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.
- Lewis, D., & Cook, R. (2020). Embedded standard-setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, 39(1), 8–21.
- Lewis, D. M., Mitzel, H. C., Mercado, R., & Schulz, M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*, (2nd ed.). Lawrence Erlbaum.
- McElhane, K., Vaishampayan, G., D. Angelo, C., Harris, C. J., Pellegrino, J. W., & Krajcik, J. (2016, June). Using learning performances to design science assessments that measure knowledge-in-use. In C. K. Looi, J. L. Polman, U. Cress, & P. Reiman (Eds.), *Transforming learning, empowering learners: Proceedings of the 12th international conference of the learning sciences (ICLS) 2016*, Vol. 2 (pp. 1211–1212). International Society of the Learning Sciences.
- McElhane, K. W., Zaidi, S., Gane, B. D., Alozie, N., & Harris, C. J. (2018, March). *Designing NGSS-aligned assessment tasks and rubrics to support classroom-based formative assessment*. Paper presented at the NARST Annual International Conference, Atlanta, GA.
- 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., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Lawrence Erlbaum.
- Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D.W., & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design: A suite of research-based design patterns*. Springer.
- National Research Council (1996). *National science education standards*. National Academies Press.
- National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
- National Science Teachers Association (2016). *NSTA position statement: The Next Generation Science Standards*. <https://www.nsta.org/about/positions>
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.
- Pellegrino, J. W. (2010). The design of an assessment system for the Race to the Top: A learning sciences perspective on issues of growth and measurement. In P. Forgione & N. Doorey (Eds.), *Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda*. Center for K–12 Assessment & Performance Management, Educational Testing Service.
- Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, 340, 320–323.
- Pellegrino, J. W. (2018). Assessment of and for learning. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 410–421). Routledge-Taylor & Francis.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, 51(1), 59–81.

- Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education* 43(2), 119–34.
- Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2014). *Developing assessments for the next generation science standards*. National Academies Press.
- Pennock, P. H., & Severance, S. (2018, March). *Comparative analysis of three-dimensional research-based and classroom-based rubrics for formative assessment*. Paper presentation at NARST Annual International Conference, Atlanta, GA.
- Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching*, (5th ed., 787–850). American Educational Research Association.
- Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, 44(1), 57–84.
- Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom assessment system to support ambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.
- Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Boston, MA: Kluwer Academic.
- Sezen-Barrie, A., & Kelly, G. J. (2017). From the teacher's eyes: Facilitating teachers' noticing on informal formative assessments (IFAs) and exploring the challenges to effective implementation. *International Journal of Science Education*, 39(2), 181–212.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34.

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Zaidi, S.Z., Ko, M., Gane, B.D., Madden, K., Gaur, D., & Pellegrino, J. W. (2018, March). Portraits of teachers using three-dimensional assessment tasks to inform instruction. Paper presented at the NARST Annual International Conference, Atlanta, GA.

Zhai, X., & Wiebe, E. (2023). Technology-based innovative assessment. In C. Harris, E. Wiebe, S. Grover, and J. W. Pellegrino (Eds.), *Classroom-based STEM assessment: Contemporary issues and perspectives*, (pp. 99–126). *Community for Advancing Discovery Research in Education (CADRE)*. Boston: Education Development Center.

Notes

The work described in this chapter spans a period of time from 2014–2023, with funding from multiple organizations including the National Science Foundation, the Moore Foundation, the Chan Zuckerberg Initiative, and the U.S. Department of Education. It is the product of many individuals representing multiple organizations.

7 KH IROORZLQJ LQGLYLGXDOV PDGH VLJQLdFDQW FRQWU
Joseph Krajcik, Christopher Harris, Daniel Damelin, Brian Gane, Sania Zaidi, Diksha Gaur, Kevin McElhaney, Nonye Alozie, Phyllis Pennock, Sam Severance, Krystal Madden, Angela DeBarger, Carla Strickland, Debbie Leslie, Jeanne DiDomenico, Diksha Gaur, Samuel Arnold, and Elizabeth Lehman.

7 KH IROORZLQJ LQGLYLGXDOV PDGH VLJQLdFDQW FRQWU
Ellen Forte, Erin Buchanan, Bill Herrera, Charlene Turner, Rhonda True, Daisy Rutstein, Allison Kaczmariski, Donald Wink, Brian Gane, Sania Zaidi, Mon Lin Ko, Mary Nyaema, Daniel Lewis, and Nathan Dadey.