21st Century Science Assessment: The Future Is Now

February 2016

SRI Education™
A DIVISION OF SRI INTERNATIONAL

Author

James W. Pellegrino, Learning Sciences Research Institute
University of Illinois at Chicago

This material is based upon work supported by the National Science Foundation under Contract No. NSFDACS12C1299. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

Suggested Citation

Pellegrino, J. W. (2016). *21st Century Science Assessment: The Future Is Now.* (SRI Education White Paper). Menlo Park, CA: SRI International.

Contents

Executive Summary

James Pellegrino's influential research on student learning, instruction, and assessment has helped shape how students learn in the 21st century. In this paper, Pellegrino reflects on the need for substantial change in what we expect students to know and be able to do in science, how science should be taught, and how science competency should be assessed. Pellegrino co-chaired the National Academy of Sciences committee tasked with developing assessments for the Next Generation Science Standards.

Science education in the United States is undergoing dramatic change as a result of the 2012 *Framework for K-12 Science Education*. Drawing on decades of research in science education, this document describes a vision for science learning whereby students gradually deepen their understanding of three core dimensions of science: disciplinary core ideas, scientific and engineering practices, and crosscutting concepts.

For this vision to come to realization, however, new kinds of science assessments must be developed to serve as indicators of progress in providing educational opportunities consistent with today's goals for STEM education. The National Research Council's 2013 report *Monitoring Progress Toward Successful K-12 STEM Education* recommends the collection of data on two indicators related to science assessment:

- States' use of assessments that measure the core concepts and practices of science disciplines
- Inclusion of science in federal and state accountability systems.

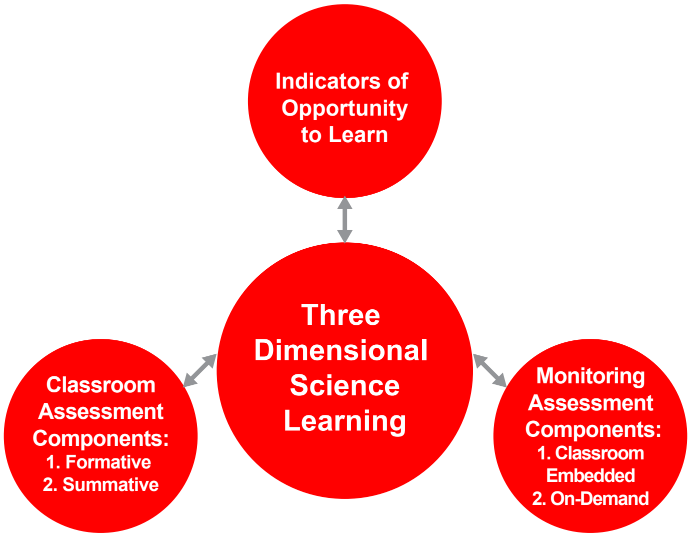
A Systems Approach to Science Assessment

Learning assessments are used by multiple audiences and for different purposes—from classroom teachers to state and national policy makers. To provide the data and information this diverse set of stakeholders requires, a systems approach is necessary to create a science assessment system consisting of the following three parts:

1. **Assessments designed to support classroom instruction.** These include formative assessments that teachers can use to identify areas where students are making progress or struggling so that they can adjust their instruction accordingly, as well as summative assessments to evaluate student learning and assign grades at the end of a course.

2. **Assessments designed to monitor science learning on a broader scale.** These are large-scale assessments used to audit student learning over time and to evaluate the effectiveness of the science education system. Monitoring requires not only state-developed standardized assessments, but also classroom-embedded assessments that fit the instructional sequence of local schools.

3. Indicators that track learning opportunities. In this part of the assessment system is regularly collected information about the quality of classroom instruction to determine whether all students have the opportunity to learn science as described in the *Framework* and to signal whether additional resources and supports are needed.


Indicators of Opportunity to Learn
Three Dimensional Science Learning
Classroom Assessment Components:
1. Formative
2. Summative
Monitoring Assessment Components:
1. Classroom Embedded
2. On-Demand

Assessing all the performance expectations for a given grade level with a single assessment is not possible. Students will need multiple opportunities to demonstrate their competence across the three core dimensions of science outlined in research-based science education reform documents. A focus on assessment design that is carefully aligned with the *Framework* is essential for sending the right signals about what students should know and be able to do when demonstrating competence in science.

How Can We Assess Science Competence?

We assess students to find out what they know and can do, but assessments are not direct pipelines into their minds. Rather, an assessment is a tool for observing students' behavior and producing data that can be used to draw reasonable inferences about what students know. To be reliable tools for assessing student competence, *Framework*-aligned assessments will have the following key design elements:

- **Assessment variety.** Classroom assessment should include various types of evidence about student learning, such as a classroom discussion in which students explore and respond to each other's ideas, a formal test or diagnostic quiz, or the evaluation of artifacts that are the product of classroom activities.
- **Multicomponent tasks.** Central to the *Framework for K-12 Science Education* is the research-based insight that science competence requires the ability to integrate disciplinary core ideas, scientific and engineering practices, and crosscutting concepts. Aligning with the *Framework* therefore means that assessment tasks must be composed of more than one kind of activity or question. Only through multicomponent tasks will students have the opportunity to demonstrate their ability to orchestrate these three dimensions of scientific competence.
- **Connections.** Because science education research emphasizes the importance of the connections among scientific concepts, assessment tasks will need to be designed to provide information about students' capacity to make these connections.
- **Student progress.** Learning is a trajectory along which students gradually progress in the course of a unit, a year, and across the K-12 grades. Thus, it is important that classroom assessments help teachers and students understand where students are relative to expected levels of progress.

Assessment is a key element in the process of educational change and improvement. Done well, it can signify what it is that we want students to know and be able to do and can help educators create the learning environments that support attainment of those objectives. Done poorly, it sends the wrong signals and skews the teaching and learning process toward teaching to tests that have little relationship to the competencies students will need in the future.

Implications for Practice and Policy

A single assessment type cannot serve all the appropriate purposes and needs of stakeholders in the educational system. Thus, **policy makers and state and district leaders need to promote a balanced and coordinated system of multiple assessments that work together with curriculum and instruction to promote science learning.**

Assessments of the three dimensions of science learning are challenging to design, implement, and interpret. Thus, **policy makers should allocate adequate funding for teacher professional development initiatives to support the uptake into classroom practice of assessments aligned with research-based, rigorous standards.**

To provide indicators of progress toward attaining STEM education goals, **state education leaders should provide clear guidelines that define forms of evidence that can be mapped to beginning, intermediate, and sophisticated levels of science knowledge and practice that are expected across grade levels.**

To develop students' skills and dispositions in science and engineering, **teachers should use curriculum materials and assessment tasks that require students to engage in practices that demonstrate their understanding of core disciplinary ideas and crosscutting concepts.**

Students require ongoing feedback about their science learning to succeed and stay motivated. To assist students during the learning process, **teachers need to make use of formative assessment tasks that can guide their instructional decision making in the classroom and provide students with information about which skills and knowledge they need to study further.**

Introduction

At their core, science assessments are statements about what scientists, educators, policy makers, and, indirectly, parents want students to learn and, in a larger sense, become. What we choose to assess in science is what will end up being the focus of instruction. Education research has well established that teachers and students take their cues from large-scale achievement tests and will try to score well on them regardless of the assessment type, especially when high stakes are associated with the outcomes. So it is critical that our science assessments best represent the kinds of learning we want to occur if our students are to achieve the forms of proficiency needed for the worlds of today and tomorrow.

In that regard, we are at an interesting moment for science education in the United States, one that holds extraordinary promise for the future of science learning but is juxtaposed with significant challenges in achieving the vision of what it means to be proficient in science. Among those challenges is determining how to assess the proficiency of our students relative to a vision of competency that has emerged across K-16+ education and how to do so in ways that support teaching and learning rather than inhibit attainment of that vision.

The work of identifying indicators of progress toward attaining the major goals for science, technology, engineering, and mathematics (STEM) education has already begun (National Research Council, 2013). Such indicators include monitoring the extent to which state science assessments measure proficiency in ways envisioned by the National Research Council (NRC) report *A Framework for K-12 Science Education* (NRC, 2012). But much of what is needed to effectively and validly assess science learning, as envisioned in the *Framework*, either at the classroom level or for local and state accountability monitoring, has yet to be created, and the design and implementation challenges are significant (e.g., Pellegrino, 2012, 2013; Pellegrino, Wilson, Koenig & Beatty, 2014; Wilson & Bertenthal, 2005).

The purpose of this paper is to consider what science assessment should look like and what it will take to design and implement a system of science assessments that supports the vision of science teaching and learning derived from research, as described in the NRC's *Framework*. This paper draws on resources that have considered issues related to both the nature of science competency and its assessment. The first section considers the nature of competence in science as reflected in the NRC's *Framework* report as well as the specifics of the *Next Generation Science Standards* (NGSS) derived from that framework.[1] That section then delves into key ideas related to the design of assessments that can validly assess the forms of integrated knowledge that have been highlighted in the *Framework* and describes how critical ideas emanating from an evidentiary reasoning perspective can be applied to design. The section ends with a discussion of examples of classroom and large-scale assessments that begin to approximate what is needed to assess science proficiency as defined by the *Framework* and NGSS. The second section considers assessment

[1] Recognizing that not all states will adopt the NGSS, the NRC's report on monitoring the STEM education system (2013) refers to the NGSS or curricula "solidly grounded in current research on teaching and learning in science and mathematics" (p.18). Similarly, the present paper refers to NGSS for convenience and illustrative purposes. However, the points made about science assessment apply more broadly to assessments aligned with any set of rigorous science standards consistent with integration of the three major components of science proficiency identified in the *Framework*.

development from a larger systems perspective. Drawing on the 2014 NRC report *Developing Assessments for the Next Generation Science Standards* (Pellegrino, Wilson, Koenig, & Beatty, 2014), it considers what state science assessment systems should include, as well as how each system component might be designed and developed. The paper concludes with recommendations for the development and deployment of the system's components and how elements of the assessment system relate to issues of accountability.

The Nature of Competence and Challenges for Assessment

The Changing Nature of Competence: What Do Students Need to Know and Be Able to Do?

Multiple Interconnected Dimensions of Competence

The need for a sophisticated understanding of science competence can be illustrated by considering student performance on assessment tasks that mimic aspects of scientific investigations (National Center for Educational Statistics [NCES], 2012). Two unique types of activity-based tasks were administered as part of the 2009 National Assessment of Educational Progress (NAEP) science assessment. In addition to more typical paper-and-pencil questions, 4th-, 8th-, and 12th-graders completed hands-on and interactive computer tasks designed to reveal not just what they knew at a factual level, but also how well they were able to reason through complex problems and apply science to real-life situations. While performing the interactive computer and hands-on tasks, the students were asked to manipulate objects and perform actual experiments; these assessment tasks generated rich data on how students respond to scientific challenges. The 2012 NCES report of findings from these assessment tasks noted that students were successful on parts of investigations that involved limited sets of data and making straightforward observations of those data, but they had difficulty with parts of investigations that contained more variables to manipulate or involved strategic decision making to collect appropriate data. The NCES report noted also that significant numbers of students could select correct conclusions from an investigation but *could not explain their results.*

These NAEP results illustrate the disjuncture between students' knowledge of science facts and procedures, as assessed by typical science achievement tests, and their understanding of how that knowledge can be applied through the practices of scientific reasoning and inquiry. Recognition of this science education problem can be found in reports spanning elementary, secondary, and postsecondary education (K-16+). These reports present a consistent description of the nature of competence in science and include NRC reports on K-8 science education in formal and informal learning environments (NRC, 2007; 2009), curriculum and assessment frameworks for Advanced Placement (AP) science courses (e.g., College Board, 2011a, b), and even revisions in the nature of the science knowledge required for entry to medical school and assessed on the Medical College Admissions Test (e.g., American Association of Medical Colleges, 2012). Seldom has such a consistent message been sent across K-16+ as to the need for substantial change in what we expect students to know and be able to do in science, how science should be taught, and how it should be assessed.

This reconceptualization of the nature of science competence is most clearly expressed in the 2012 NRC *Framework* report, which articulates three interconnected dimensions of competence. The first of these dimensions is *Disciplinary Core Ideas*. In

> *Seldom has such a consistent message been sent across K-16+ as to the need for substantial change in what we expect students to know and be able to do in science, how science should be taught, and how it should be assessed.*

reaction to criticisms of U.S. science curricula being "a mile wide and an inch deep" (Schmidt, McKnight, & Raizen, 1997, p. 62) relative to those of other countries, the *Framework* identified a small set of core ideas in four disciplines: (1) life sciences, (2) physical sciences, (3) earth and space sciences, and (4) engineering, technology, and the application of science. In doing this, the *Framework* attempts to reduce the long and often disconnected catalog of factual knowledge that students currently must learn. 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 carry¬ing out investigations; developing and using models; and engaging in argument from evidence.

Although the *Framework's* three dimensions are conceptually distinct, the vision is of coordination in science and engineering education so that the three are integrat¬ed in the teaching, learning, and doing of science and engineering. By engag¬ing in science and engineering practices, 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 competence is realized through performance expectations—what students at various levels of educational experience should know and be able to do. Statements of 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.

> *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.*

The virtue of such a view of the nature of competence and its development is that science educators are poised to better define the outcomes desired from their instructional efforts and that such articulations can guide the forms of assessment that will help educators know whether their students are attaining the desired competencies and how they might better assist them along the way. This is true whether or not a state formally adopts the NGSS or any other rigorous science education standards that are aligned with the *Framework* vision. Thus, it is very important for the science education community and the educational policy community to develop a shared set of perspectives on what constitutes high-quality and valid science assessments across K-16+ if assessments are to serve their desired educational purposes.

From Frameworks to Standards: A Focus on Performance Expectations

To illustrate the instructional and assessment challenges posed by the conception of competence in the *Framework*, we can consider the projected end point of K-12 science education. By the end of 12th grade all students—not just those interested in pursuing science, engineering, or technology beyond high school—should have gained sufficient knowledge and understanding to

1. engage in public discussions of science-related issues such as the challenges of generating sufficient energy, preventing and treating diseases, maintaining supplies of clean water and food, and addressing problems caused by global environmental change;
2. be critical consumers of scientific information related to their everyday lives; and
3. continue to learn about science throughout their lives.

Students should come to appreciate that science as a discipline and current scientific understanding of the world are the result of hundreds of years of creative human endeavor (NRC, 2012, p. 24).

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 in a way designed to meet this ambitious goal. This three-part structure signals an important evolutionary shift for science education and presents 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" (NRC, 2012, 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 they are integrated with. 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.

The *Framework* uses the logic of learning progressions to describe students' developing proficiency in these three intertwined domains in a coherent way across grades K through 12, noting that "If mastery of a core idea in a science discipline is the ultimate educational destination, then well-designed learning progressions provide a map of the routes that can be taken to reach that destination" (NRC, 2012, p. 26). The stress on learning progressions is supported by research on learning described in the 2007 NRC report *Taking Science to School* and in other documents (e.g., Alonzo & Gotwals, 2012; Corcoran, Mosher, & Rogat, 2009). The *Framework* builds in the idea of a developmental progression of student understanding across the grades by specifying grade band end point targets at grades 2, 5, 8, and 12 for each component of each core idea. For the practices and crosscutting concepts, the framework also provides sketches of possible progressions for learning each practice or concept but does not indicate the expectations at any particular grade level. The NGSS, for example, built on these suggestions and developed tables that define what each practice might encompass at each grade level; the NGSS also define the expected uses of each crosscutting concept for students at each grade level.

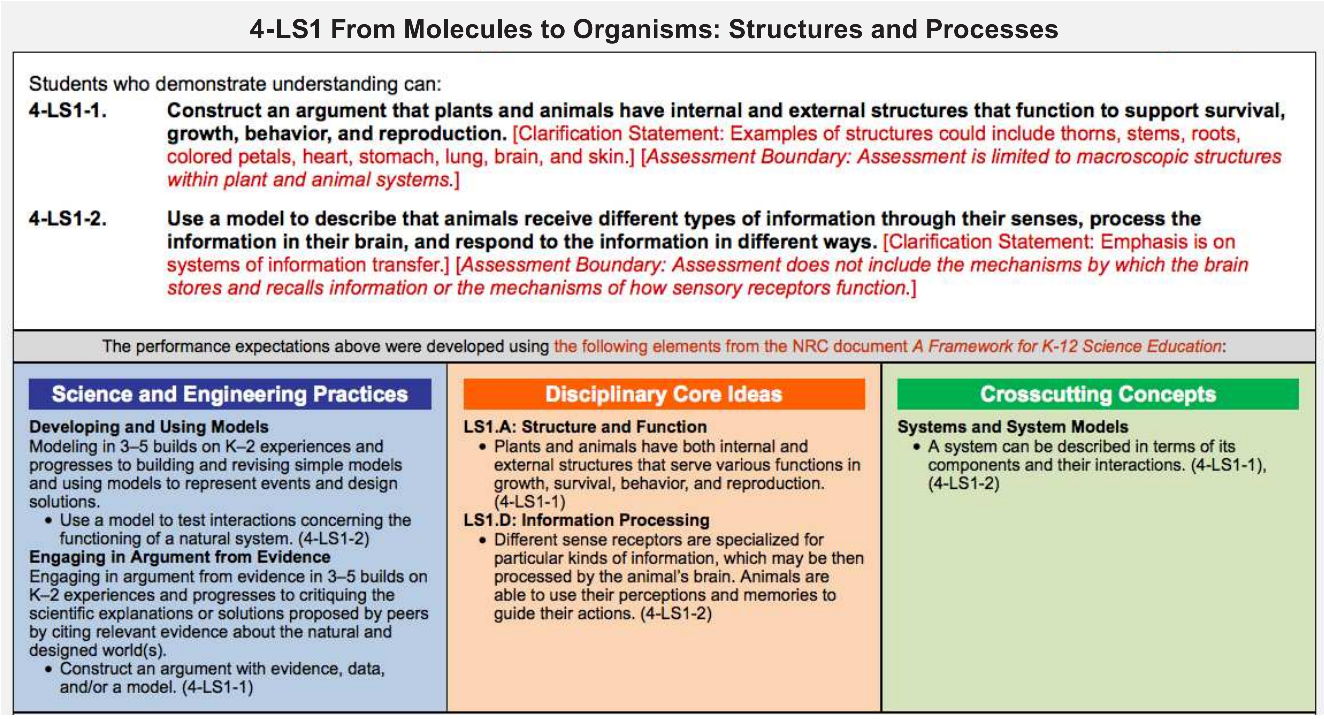
In the context of assessment, the importance of this integrated perspective of what it means to know science is that one should be attempting to assess where a student can be placed along a sequence of progressively more "scientific" understandings of a given core idea and successively more sophisticated applications of practices and crosscutting concepts. This is a relatively unfamiliar idea in the realm of science assessments, which have more often been viewed as simply measuring whether students know or do not know particular grade-level content. It means that assessments must strive to be sensitive both to grade-level-appropriate understanding and to those understandings that may be appropriate at somewhat lower or higher grades. This is particularly important for assessment materials and resources that can be used to support classroom instruction.

To support the integrated approach to science learning the *Framework* explains 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* (NRC, 2012). It recommends using a model put forward in Standards for Success (College Board, 2009) by expressing science standards in terms of performance expectations. The latter describe activities and outcomes that students are expected to achieve in order to demonstrate their ability to understand and apply the knowledge described in the disciplinary core ideas. Performance expectations

> specify what students should know, understand, and be able to do. . . . They also illustrate how students engage in science practices to develop a better understanding of the essential knowledge. These expectations support targeted instruction and assessment by providing tasks that are measurable and observable. (College Board, 2009, p. 21)

In developing the NGSS, Achieve (2013) and its partners elaborated these guidelines into standards that are clarified by descriptions of the ways 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. The NGSS standards appear as clusters of performance expectations related to a particular aspect of a core disciplinary idea (see Exhibit 1). Each performance expectation asks students to use a specific practice and a crosscutting concept in the context of a

Exhibit 1. An example of the NGSS architecture for one aspect of fourth-grade life science



4-LS1 From Molecules to Organisms: Structures and Processes

Students who demonstrate understanding can:

4-LS1-1. **Construct an argument that plants and animals have internal and external structures that function to support survival, growth, behavior, and reproduction.** [Clarification Statement: Examples of structures could include thorns, stems, roots, colored petals, heart, stomach, lung, brain, and skin.] [*Assessment Boundary: Assessment is limited to macroscopic structures within plant and animal systems.*]

4-LS1-2. **Use a model to describe that animals receive different types of information through their senses, process the information in their brain, and respond to the information in different ways.** [Clarification Statement: Emphasis is on systems of information transfer.] [*Assessment Boundary: Assessment does not include the mechanisms by which the brain stores and recalls information or the mechanisms of how sensory receptors function.*]

The performance expectations above were developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices	Disciplinary Core Ideas	Crosscutting Concepts
Developing and Using Models Modeling in 3–5 builds on K–2 experiences and progresses to building and revising simple models and using models to represent events and design solutions. • Use a model to test interactions concerning the functioning of a natural system. (4-LS1-2) **Engaging in Argument from Evidence** Engaging in argument from evidence in 3–5 builds on K–2 experiences and progresses to critiquing the scientific explanations or solutions proposed by peers by citing relevant evidence about the natural and designed world(s). • Construct an argument with evidence, data, and/or a model. (4-LS1-1)	**LS1.A: Structure and Function** • Plants and animals have both internal and external structures that serve various functions in growth, survival, behavior, and reproduction. (4-LS1-1) **LS1.D: Information Processing** • Different sense receptors are specialized for particular kinds of information, which may be then processed by the animal's brain. Animals are able to use their perceptions and memories to guide their actions. (4-LS1-2)	**Systems and System Models** • A system can be described in terms of its components and their interactions. (4-LS1-1), (4-LS1-2)

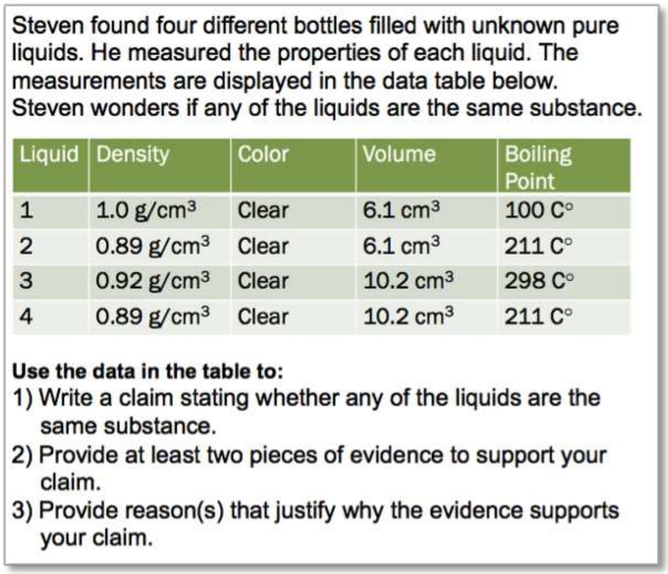
Source: Achieve (2013), Next Generation Science Standards.

specific element of the disciplinary knowledge relevant to the 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.

In contrast to science standards calling for the integration of science practices and content knowledge, the prior generation of U.S. science standards (e.g., NRC, 1996, 2000) 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, *multidimensional 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 is neither appropriate nor sufficient. Exhibit 2 provides an example of a task that could be presented to middle school students to assess their understanding of the properties of substances that are associated with chemical identity, and it does so in the context of constructing an argument about which substances are the same or different based on patterns of data provided in the table. A teacher might use such a task to gauge how well the students have understood which properties are associated with chemical identity as well as their ability to use evidence to construct a scientific argument given a scheme that guides them to present the essential elements of such a data-based argument.

> *The idea of an integrated, multidimensional science performance presents a very different way of thinking about science proficiency.*

Exhibit 2. An example of a physical science assessment task designed for classroom use at the middle school level



Steven found four different bottles filled with unknown pure liquids. He measured the properties of each liquid. The measurements are displayed in the data table below. Steven wonders if any of the liquids are the same substance.

Liquid	Density	Color	Volume	Boiling Point
1	1.0 g/cm^3	Clear	6.1 cm^3	100 C°
2	0.89 g/cm^3	Clear	6.1 cm^3	211 C°
3	0.92 g/cm^3	Clear	10.2 cm^3	298 C°
4	0.89 g/cm^3	Clear	10.2 cm^3	211 C°

Use the data in the table to:

1) Write a claim stating whether any of the liquids are the same substance.
2) Provide at least two pieces of evidence to support your claim.
3) Provide reason(s) that justify why the evidence supports your claim.

Source: Harris, C. McElhaney, K. D'Angelo, C. Krajcik, J., Dahsah, C., Lee. J., Pellegrino, J., DiBello, L., Gane, B., & Damelin, D. (2015). Constructing assessment items that blend core Ideas, crosscutting concepts, and science practices for classroom formative applications. Unpublished paper. Menlo Park, CA: SRI International.

Assessing Competence: How Will We Know What Students Know?

As shown in Exhibit 1, the NGSS performance expectations reflect intersections of a disciplinary core idea, science and engineering practices, and related crosscutting concepts. 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 and the College Board's Standards for College Success, do not provide detail sufficient to create assessments. The design of valid and reliable science assessments is complex, hinging 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 (Mislevy & Haertel, 2006; Pellegrino, Chudowsky, & Glaser, 2001). For example, designers of assessment items and tasks related to the performance expectations in Exhibit 1 also need to consider (1) the kinds of conceptual models and evidence that we expect students to engage with, (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.

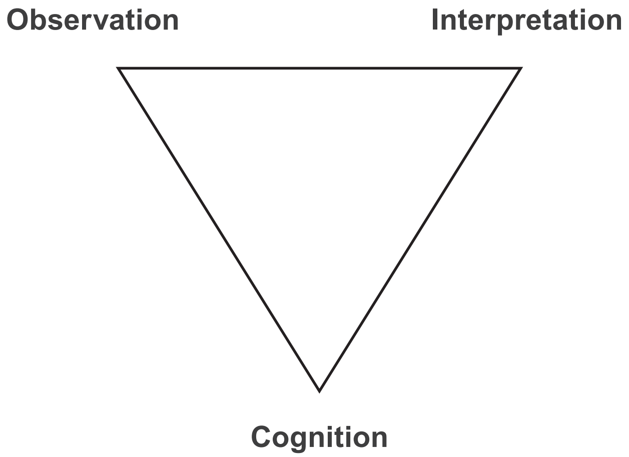
The *Framework* and the NGSS performance expectations raise many questions about what valid science assessment should look like. Addressing them requires serious consideration of some fundamental issues related to the design and use of educational assessments.

Assessment as a Process of Reasoning From Evidence

We assess students to find out what they know and can do, but assessments are not direct pipelines into students' minds. Unlike height or weight, the mental representations and processes educators care about are not outwardly visible. Thus, an assessment is a tool for observing students' behavior and producing data that can be used to draw reasonable inferences about what students know. A chain of reasoning connecting observed behaviors to inferences about what students know is required for all educational assessments, from classroom quizzes and standardized achievement tests to computerized tutoring programs and even to the conversation a student has with the teacher as they work through a science problem or discuss the meaning of part of a science article or text.

The first question in the assessment reasoning process is: evidence about what? Data do not provide their own meaning; they become evidence only when their relevance to a conjecture being considered has been established through some framework for systematic interpretation. What a person perceives visually, for example, depends not only on the data received as photons of light striking the retinas, but also on what the person thinks she might see. In the NRC report *Knowing What Students Know* report (Pellegrino et al., 2001), the process of reasoning from evidence was portrayed as a triad: the assessment triangle (Exhibit 3). The vertices of the assessment triangle represent the three key elements underlying any assessment: a model of student cognition and learning in the domain of the assessment, a set of assumptions and principles about the kinds of observations that will provide evidence of students' competencies, and an interpretation process for making sense of

Exhibit 3. The assessment triangle



the evidence in light of the assessment purpose and student understanding. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented or evaluated without consideration of each.

The *student model* corner of the triangle (the cognition vertex) encompasses theory, data, and a set of assumptions about how students represent knowledge and develop competence in a subject matter domain (e.g., core ideas such as Newton's laws; evolutionary processes, thermodynamic principles, or practices such as using models or constructing explanations). In the case of science, these are derived from the theory and research behind the development and articulation of the *Framework*.

Every assessment is also based on a set of *assumptions and principles* about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates the knowledge and skills in the student model (the observation vertex). The tasks students are asked to respond to must be carefully designed to provide evidence that is linked to the model of learning and to support the kinds of inferences and decisions that will be made on the basis of the results. The observation vertex of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. In the case of science, this is based on key features of tasks related to practices such as design, modeling, explanation, and argumentation.

Every assessment is also based on certain assumptions and models for *interpreting* the evidence collected from observations. The interpretation vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed.

A crucial point is that each of the three elements of the assessment triangle not only must make sense on its own, but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Thus, all three vertices of the triangle must work together in synchrony. Central to this entire process, however, are theories, models, and data on how students learn and what students know as they develop competence for important aspects of the science curriculum.

Evidence-Centered Design

Given that assessment involves evidentiary reasoning, framing assessment design as a systematic evidence-centered design process has proven useful (e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006). The process starts with defining as precisely as possible the claims that the evaluators want to be able to make about student competence—what students are supposed to know and understand—in a particular aspect of a domain, such as aspects of force and motion, heat and temperature, etc. The most critical aspect of defining the claims to be made is to be as precise as possible about the elements that matter and express them as verbs (model, explain, predict, argue) as in the statements of performance expectations related to the science practices. In essence, the performance expectations found in science standards are major claims about student proficiency.

Although the claims to be made or verified are about the student, they must be linked to the forms of evidence—the warrants—that would support them. The evidence statements associated with given sets of claims capture the features of work products or performances that would give substance to the claims. The designer must specify the features that need to be present and how they are weighted in any evidentiary scheme—i.e., what matters most and what matters least or not at all. For example, if the warrant in support of a claim about knowledge of the laws of motion is that the student can analyze a physical situation in terms of the forces acting on all the bodies, then the evidence might be the student's drawing a free body diagram with all the forces labeled, including their magnitudes and directions.

The precision in elaborating the claims and evidence statements associated with a domain of knowledge and skill pays off when the time comes to design assessment tasks or situations that can provide the requisite evidence. Tasks such as the one shown in Exhibit 2 are not designed or selected until it is clear what forms of evidence are needed to support the range of claims appropriate to a given assessment situation. The tasks need to provide all the necessary evidence and should allow students to "show what they know" in as unambiguous a way as possible.

Differentiating Assessment Purposes and Contexts

The specific purposes an assessment will be used for are important to consider in all phases of design. For example, assessments used by instructors in classrooms to assist or monitor learning typically need to provide more detailed information than assessments whose results will be used by policy makers or accrediting agencies. One of the central points of *Knowing What Students Know* was that assessments must be developed for specific purposes and that the nature of their design is driven by their intended interpretive use.

In the classroom, instructors use various forms of assessment to inform day-to-day and month-to-month decisions about next steps for instruction, to give students feedback on their progress, and to motivate students. These situations are referred to as *assessments to assist learning*, or *formative use of assessment* (see Black & Wiliam, 1998; Wiliam, 2007). These assessments provide specific information about students' strengths and difficulties with learning, and teachers can use this information to adapt their instruction to meet students' needs, which may be difficult to anticipate and are likely to vary from one student to another. Students can use this information to determine which skills and knowledge they need to study further and what adjustments in their thinking they need to make.

Another type of assessment is conducted to help determine whether a student has attained a certain level of competency after completing a particular phase of education, whether it be a 2-week curriculum unit, a semester-long course, or 12 years of schooling. This is referred to as *assessment of individual achievement,* or *summative use of assessment.* Some of the most familiar forms of summative assessment are those used by classroom instructors, such as end-of-unit or end-of-course tests, which often are used to assign letter grades when a course is finished. Large-scale assessments—which are administered at the direction of users external to the classroom— also provide information about the attainment of individual students, as well as comparative information about how one individual performed relative to others. Because large-scale assessments are typically given only once a year and involve a time lag between testing and availability of results, the results seldom provide information that can be used to help teachers or students make day-to-day or month-to-month decisions about teaching and learning.

Another common purpose of assessment is to help administrators, policy makers, or researchers formulate judgments about the quality and effectiveness of educational programs and institutions. Instructional evaluation can be considered formative in nature when used to improve the effectiveness of instruction. Summative uses of assessment for evaluation are incorporated increasingly in making high-stakes decisions not only about individuals, but also about programs and institutions (e.g., Linn, 2013). For instance, public reporting of state assessment results by schools and districts can influence the judgments of parents and taxpayers about the quality and efficacy of schools and affect decisions about resource allocations. Just as with individuals, the quality of the measure is critical in the validity of these decisions.

The purpose of an assessment determines priorities, and the context of use imposes constraints on the design. Thus, it is essential to recognize that one type of assessment does not fit all purposes or contexts of use. A persistent mistake is to assume that an assessment is appropriate and interpretable for a particular context without determining whether evidence exists regarding the validity of the assumptions in that context. The one-size-fits-all fallacy is especially frequent and problematic because it produces inappropriate choices of assessments for instructional, evaluation, or research purposes that in turn can lead to invalid conclusions regarding persons, programs, or institutions.

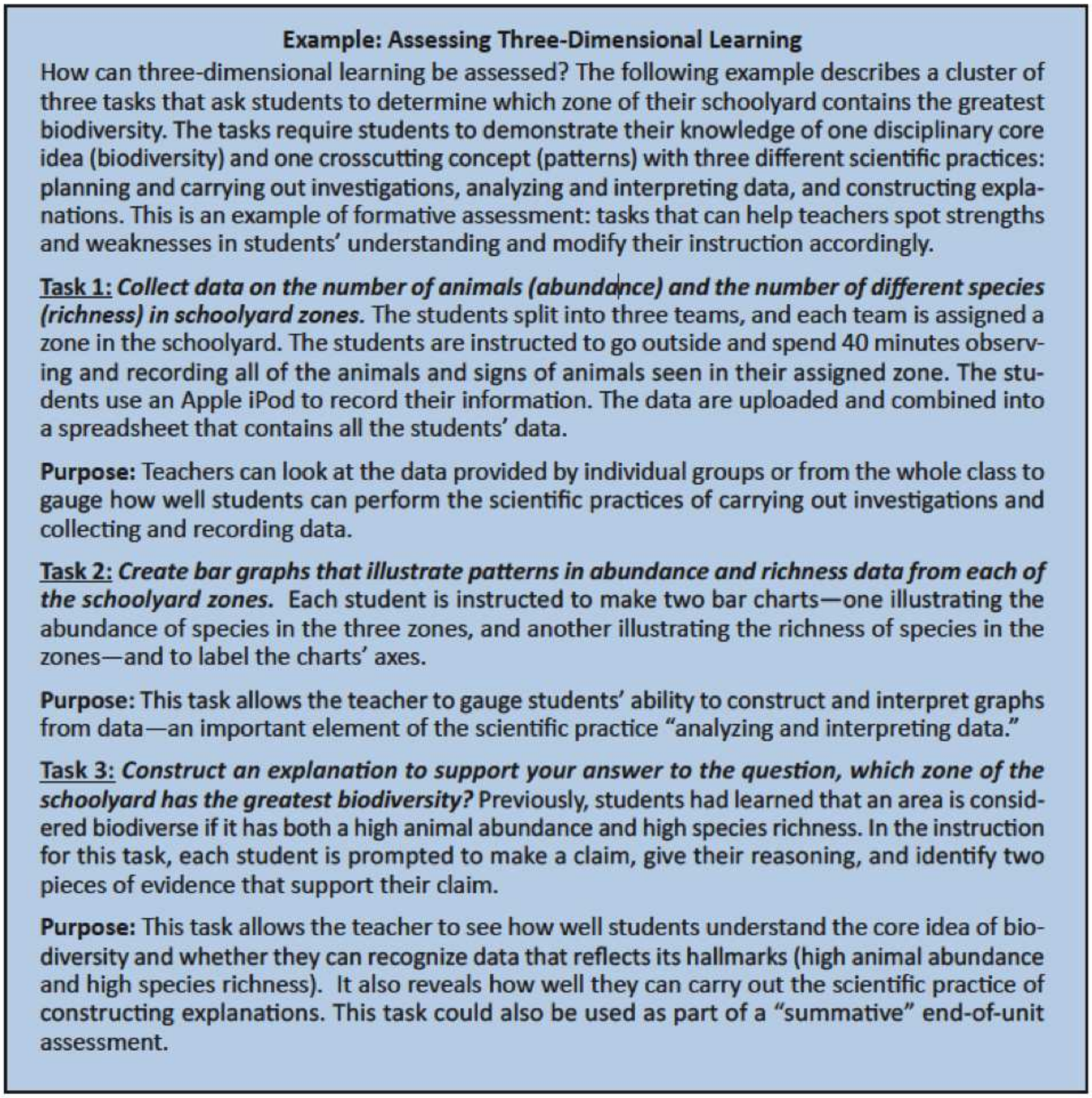
Challenging But Not Impossible: Examples of the Possible

Given the relative newness of the NRC *Framework* and *Framework*-inspired science standards, it should come as no surprise that we have no comprehensive sets of examples of the types of assessments that align completely with the performance expectations in these standards. Many of the science assessment tasks that have typically been used for classroom assessment, as well as those found in large-scale state, national, and international tests, focus primarily on science content or on aspects of scientific inquiry separate from content. With relatively few exceptions, such assessments do not integrate core concepts and science practices in the ways intended by the *Framework*. Fortunately, some of what we now know about the science and design of educational assessments has been productively used to develop science assessments that approximate the types of tasks and situations called for by the K-12 *Framework*. Although not plentiful, there are cases to draw from to illustrate forms of science assessment that approximate what is needed. Several of these were presented and discussed in the NRC report on *Developing Assessments for the Next Generation Science Standards* (Pellegrino et al., 2014). The examples are diverse in several ways, including the science content and practices represented, age and grade level, whether the assessments are delivered using technology, whether the consequences of student performance have low or high stakes, and scale of use (classroom, state, national, or international level). In many of the cases, a principled approach such as evidence-centered design was used to guide assessment design and validation.

Classroom Instruction and Assessment

Several research and development projects have focused on developing assessments for use in classroom instructional contexts with a particular emphasis on the integration of core science concepts with one or more of the science practices such as modeling, evidence-based explanation and argumentation, or the design of investigations to test hypotheses, analyze results, and construct explanations from data. Several of the clearest examples can be found in a volume on learning progressions edited by Alonzo and Gotwals (2012) and in a 2012 special issue of the *Journal of Research on Science Teaching*. Exhibit 4 is a brief description of one example set of tasks for formative classroom use. It is discussed in detail in Pellegrino et al. (2014). The available examples demonstrate the feasibility of designing tasks and situations, whether in paper-and-pencil format or mediated via technology, that challenge students to reason with and about core science concepts in life and physical science. They illustrate ways to obtain evidence related to student proficiency, including diagnosis of student thinking for instructional planning.

Exhibit 4. Description of a set of classroom formative assessment activities



Example: Assessing Three-Dimensional Learning

How can three-dimensional learning be assessed? The following example describes a cluster of three tasks that ask students to determine which zone of their schoolyard contains the greatest biodiversity. The tasks require students to demonstrate their knowledge of one disciplinary core idea (biodiversity) and one crosscutting concept (patterns) with three different scientific practices: planning and carrying out investigations, analyzing and interpreting data, and constructing explanations. This is an example of formative assessment: tasks that can help teachers spot strengths and weaknesses in students' understanding and modify their instruction accordingly.

Task 1: ***Collect data on the number of animals (abundance) and the number of different species (richness) in schoolyard zones.*** The students split into three teams, and each team is assigned a zone in the schoolyard. The students are instructed to go outside and spend 40 minutes observing and recording all of the animals and signs of animals seen in their assigned zone. The students use an Apple iPod to record their information. The data are uploaded and combined into a spreadsheet that contains all the students' data.

Purpose: Teachers can look at the data provided by individual groups or from the whole class to gauge how well students can perform the scientific practices of carrying out investigations and collecting and recording data.

Task 2: ***Create bar graphs that illustrate patterns in abundance and richness data from each of the schoolyard zones.*** Each student is instructed to make two bar charts—one illustrating the abundance of species in the three zones, and another illustrating the richness of species in the zones—and to label the charts' axes.

Purpose: This task allows the teacher to gauge students' ability to construct and interpret graphs from data—an important element of the scientific practice "analyzing and interpreting data."

Task 3: ***Construct an explanation to support your answer to the question, which zone of the schoolyard has the greatest biodiversity?*** Previously, students had learned that an area is considered biodiverse if it has both a high animal abundance and high species richness. In the instruction for this task, each student is prompted to make a claim, give their reasoning, and identify two pieces of evidence that support their claim.

Purpose: This task allows the teacher to see how well students understand the core idea of biodiversity and whether they can recognize data that reflects its hallmarks (high animal abundance and high species richness). It also reveals how well they can carry out the scientific practice of constructing explanations. This task could also be used as part of a "summative" end-of-unit assessment.

Source: Pellegrino et al., 2014.

Advanced Placement Science

The second example is in the redesign of the AP courses and assessments for biology, chemistry, and physics (College Board, 2011a, b; Huff, Steinberg & Matts, 2010; Wood, 2009). Starting in 2006, the College Board, with support from the National Science Foundation (College Board, 2010), initiated a process of redefining the focus of each AP science course in terms of the critical content and the science practices that should define competence at the end of the course. This redefinition then guided the development of a curriculum framework for each course and the high-stakes assessment often used by colleges for granting course credit and/or advanced course placement. The first of the new AP science exams was given in spring 2013 in biology, with chemistry following in 2014 and physics in 2015. To help teachers and students orient to the new course and exam demands, a wealth of materials, including sample assessments, were provided in advance on the College Board website (e.g., College Board, 2012). While the AP science redesign is a work in progress and much remains to be determined about the quality and impact of the new framework and exams on student learning and classroom instructional practice, AP science instruction and assessment clearly are changing in ways closely aligned with the perspective on science competence and proficiency described above.

National and International Large-Scale Assessment[2]

Much of what students and teachers experience as science assessments is external to regular classroom instruction and is in the form of large-scale state tests such as those administered in response to the No Child Left Behind (NCLB) legislation. While NCLB has since been replaced with the passage of the Every Student Succeeds Act (ESSA) in December 2015, these large-scale state tests are likely to continue. Whereas the quality of state science assessments varies, none of the state assessments used in 2014 approximate the performance expectations discussed in the NRC *Framework* and NGSS. In contrast, two large-scale assessment programs more closely approximate the *Framework's* conception of student competencies, the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA).

The NAEP 2009 and 2011 science assessment was constructed from a framework document that identified specific areas of science content in the life, physical, and earth and space sciences as well as a set of science practices. To probe students' abilities to combine their understanding of core ideas with the investigative skills that reflect practices, a subset of the students completed hands-on performance or interactive com¬puter tasks (see NCES, 2008, for details). Earlier, we described students' performance on the latter tasks in the 2009 assessment. The PISA assessment

> focuses on things that 15-year-old students will need in the future and seeks to assess what they can do with what they have learned—reflecting the ability of students to continue learning throughout their lives by applying what they learn in school to non-school environments, evaluating their choices and making decisions. The assessment is informed, but not constrained, by the common denominator of national curricula. (Organization for Economic Cooperation and Development, 2009)

The most recent PISA science assessment results are based on a 2006 framework that includes science competencies that overlap with the science practices

[2] TIMSS is not included in this discussion. Instead, the discussion focuses on large-scale assessments more closely aligned to the *Framework* and NGSS.

of the NRC *Framework* as well as aspects of the NAEP framework.

What is especially important about both NAEP and PISA are the sets of simple and complex science assessment tasks that demand the types of reasoning about science content described in the NRC Framework. Both assessment programs thus are a source of examples of performances that align with the descriptions of competency and proficiency discussed earlier. Furthermore, neither program is static, with both undergoing periodic revisions to the framework that guides their assessment design and task development, and both are moving to incorporate technology as a key aspect of assessing student performance. The NAEP framework will most likely be revised within the next decade, and work has already been done to revise the PISA science framework for the 2015 test administration. Changes in the assessment frameworks and operational tests of both programs are ostensibly moving in directions that even more closely align with the NRC *Framework*. Thus, both assessment programs might constitute reasonable ways to monitor the overall progress of science teaching and learning in U.S. classrooms in ways consistent with implementation of the *Framework*.

Designing and Implementing an Assessment System

A Balanced System of Assessments

One form of assessment does not and cannot serve all the appropriate purposes and needs of various actors in the educational system. Thus, it is inevitable that multiple assessments will be required to serve the varying science assessment needs of different audiences, ranging from classroom teachers to state and national policy makers. A multitude of different assessments are already used in schools, and it is not surprising that educators are often frustrated when such assessments appear to have conflicting achievement goals and to yield inconsistent results. Sometimes such discrepancies can be meaningful and useful, such as when assessments are explicitly aimed at measuring different school outcomes. More often, however, conflicting assessment goals and feedback cause much confusion for educators, students, parents, and policy makers. Thus, it is critical that there be a vision for a balanced and coordinated system of multiple assessments that work together and along with curriculum and instruction to promote effective science teaching and learning.

> *It is inevitable that multiple assessments will be required to serve the varying science assessment needs of different audiences.*

> *An assessment system is needed that exhibits three properties: comprehensiveness, coherence, and continuity.*

The current educational assessment environment in the United States clearly reflects the considerable value placed on external, large-scale assessments of individuals and programs relative to classroom assessments designed to assist learning. The resources invested in producing and using large-scale tests in terms of money, time, research, and development far outweigh the investment in the design and use of effective classroom assessment. A lesson can be learned from the investment made (via the U.S. Department of Education's Race to the Top program) in large-scale assessments developed by the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced (SBAC) state consortia for the Common Core State Standards in English language arts (ELA) and mathematics. Experience with this Race to the Top effort suggests that to better serve the goals of learning, the investment in assessment research, development, professional development, and training should be shifted toward the classroom, where teaching and learning occur.

In addition, there is ample evidence that the large-scale assessments negatively affect classroom instruction and assessment practices. Teachers feel pressure to teach to the state-mandated test, which often results in a narrowing of instruction. They also model their own classroom tests after less-than-ideal standardized tests (Linn, 2000; Shepard, 2000). These kinds of problems suggest that beyond striking a better balance between classroom and large-scale assessment, what is needed are coordinated systems of assessments that collectively support a common set of learning goals rather than working at cross-purposes. To this end, an assessment system is needed that exhibits three properties: comprehensiveness, coherence, and continuity.

By *comprehensiveness*, we mean that multiple measurement approaches are used to provide a variety of evidence to support educational decision-making. Multiple measures take on particular importance when important, life-altering decisions (such as high school graduation) are made about individuals. No single test score can be considered a definitive measure of a student's competence. Multiple measures enhance the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence. Multiple measures can also be used to provide evidence that improvements in test scores represent real gains in learning, as opposed to score inflation due to teaching narrowly to one particular test (Heubert & Hauser, 1999).

For the assessment system to support learning, it must also have the quality of coherence. One dimension of coherence is that the conceptual base or models of student learning underlying the various external and classroom assessments within a system are compatible. While a large-scale assessment might be based on a model of learning that is at a more macro level than that underlying the assessments used in classrooms, the conceptual base for the large-scale assessment should be a broader version of one that makes sense at the finer grained level (Mislevy, 1996; Pellegrino et al., 2001). In this way, the external assessment results will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment. As long as the underlying models and targets of learning are consistent, the assessment results at different levels of the system will complement each other rather than promote conflicting goals for learning and contradictory judgments about student competence.

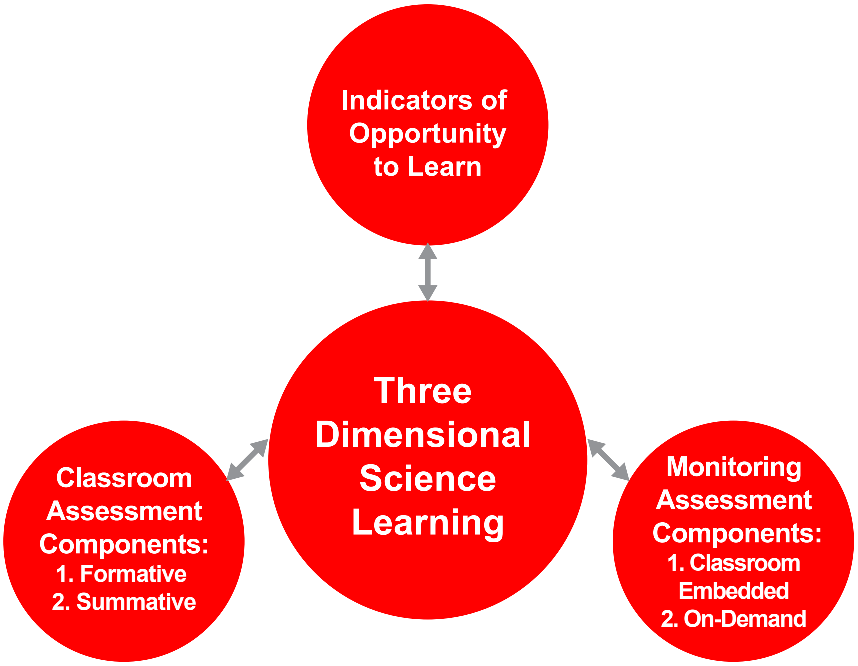
> *Assessments should measure student progress over time, more akin to a videotape record than to the snapshots provided by most current tests.*

Finally, an ideal assessment system would be designed to be continuous. That is, assessments should measure student progress over time, more akin to a videotape record than to the snapshots provided by most current tests. To portray progress over time, multiple sets of observations made at different times must be linked conceptually so that change can be observed and interpreted. Models of student progress in learning should underlie the assessment system, and tests should be designed to provide information that maps back to the progression.

Arguments for balanced and coherent systems of assessments have been made in a number of reports (e.g., NRC, 2003; Pellegrino et al., 2001; Wilson & Bertenthal, 2005). The recent NRC report Developing Assessments for the Next Generation Science Standards (Pellegrino et al., 2014) recommended a systems approach to science assessment that included a balance among three components designed to support and complement one another. Exhibit 5 illustrates the coordination of these components through a common emphasis on the *Framework's* vision for three-dimensional science learning.

- assessment designed to support classroom instruction;
- assessments designed to monitor science learning on a broader scale; and
- a series of indicators to ensure that students are provided with adequate opportunity to learn science in the ways laid out in the *Framework* (see also NRC, 2013).

Exhibit 5. The multiple components of a system of assessments aligned with a vision of three-dimensional science learning


Indicators of Opportunity to Learn
Three Dimensional Science Learning
Classroom Assessment Components:
1. Formative
2. Summative
Monitoring Assessment Components:
1. Classroom Embedded
2. On-Demand

Source: Pellegrino et al., 2014.

The first two components may seem obvious, given the prior discussion about the differing goals and purposes of educational assessment (classroom teaching and learning versus system monitoring purposes). For each system component, new sets of assessments will need to be designed to fulfill the intended goals and purposes of that component. Both assessment components should be aligned with the *Framework's* vision of science competency and its expression via performance expectations specified by standards. The discussions that follow consider issues related to the development of each of these first two components of a balanced and coherent system of assessments.

The one system component that is perhaps not so obvious is the series of indicators of opportunity to learn. Such indicators make it possible to evaluate the equity of students' opportunity to learn science in the ways envisioned by the *Framework*. Such knowledge will be necessary to adequately and appropriately interpret assessment results obtained at each level of the system. This is especially critical if the assessment results are to be used for purposes of accountability (see NRC, 2013 for an extended discussion of such a system of indicators).

Development of the Classroom Assessment System Components

Importance

Classroom assessment is integral to instruction and learning and should include tasks that are specifically designed for formative purposes and separate tasks specifically designed for summative purposes. The kind of instruction that will be effective in teaching science in the way the *Framework* envisions will require students to engage in scientific and engineering practices in the context of disciplinary core ideas and to make connections across topics through the crosscutting ideas. To develop the skills and dispositions to use scientific and engineering practices, students need to experience instruction in which they (1) use multiple practices in developing a particular core idea and (2) apply each practice in the context of multiple core ideas. Effective use of the practices often requires that they be used in concert with one another, such as in supporting explanation with an argument or using mathematics to analyze data.

Assessment activities will be critical supports for such instruction. Students will need guidance on what is expected of them and opportunities to reflect on their performance as they develop proficiencies. Teachers will need information about what students understand and can do so they can adapt their instruction. Instruction that is aligned with the *Framework* and associated standards will naturally provide many opportunities for teachers to observe and record evidence of student thinking, such as when students develop and refine models; generate, discuss, and analyze data; engage in both spoken and written explanations and argumentation; and reflect on their own understanding. The richness of the products of such instruction is a natural link to the characteristics of classroom assessment that aligns with the *Framework*.

Key Design Elements

Assessment variety. Because *Framework*-aligned instruction will involve a range of activities, classroom assessment that is integral to instruction will also need to elicit a variety of types of evidence about student learning. Indeed, the distinction between instructional activities and assessment activities may be blurred, particularly when the assessment purpose is formative. A classroom assessment may be based on a classroom discussion or a group activity in which students explore and respond to each other's ideas and learn as they go through this process. Science and engineering practices lend themselves well to assessment activities that can provide this type of evidence. For instance, when students are developing and using models, they should be given the opportunity to explain their models and to discuss them with classmates, thus providing the teacher with an opportunity for formative assessment reflection. Similarly, student discourse can give the teacher a window into students' thinking and help to guide lesson planning. A classroom assessment may also involve a formal test or diagnostic quiz. Or it may be based on artifacts that are the products of classroom activities, rather than on tasks designed solely for assessment purposes. These artifacts may include student work produced in the classroom, homework assignments (such as lab reports), a portfolio of student work collected over the course of a unit or a school year (which may include both artifacts of instruction and results from formal unit and end-of-course tests), or activities conducted using computer technology. A classroom assessment may occur in the context of group work or discussions, provided the teacher ensures that all the students who need to be observed are in fact active participants. Summative assessments may also take a variety of forms, but they are usually intended to assess each student's independent accomplishments.

Multicomponent tasks. Performance expectations blend a practice and a crosscutting idea with an aspect of a particular core idea. As noted, performance expectations provide a start in defining the claim or inference that is to be made about student proficiency. However, it is also important to determine the observations (the forms of evidence in student work) that are needed to support the claims and then to develop tasks or situations that will elicit the needed evidence. To provide such evidence, assessments aligned with the performance expectations will need to be composed of more than one kind of activity or question. They will need to include tasks in which students have opportunities to engage in practices as a means of demonstrating their ability to apply them. For example, the task shown in Exhibit 2 is designed to elicit evidence that a student can develop an argument about the identity of substances by using data linked to key aspects of a core idea, the chemical properties of substances. Tasks may require that students articulate a claim about selected structure-function relationships, develop or describe a model that supports the claim, and provide a justification that links evidence to the claim (such as an explanation of the mechanism described by the model). A multicomponent task may include some short-answer questions, possibly some carefully designed selected-response questions, and some extended-response elements that require students to demonstrate their understandings (such as tasks in which students design an investigation or explain a pattern of data). For making an appraisal of student learning, no single piece of evidence is likely to be sufficient; rather, the pattern of evidence across multiple components will be needed to provide a sufficient indicator of student understanding.

Connections. The *Framework* emphasizes the importance of the connections among scientific concepts. Not surprisingly then, the performance expectations for one disciplinary core idea may be connected to performance expectations for other core ideas, both within the same scientific domain or in other domains, in multiple ways: One may be a prerequisite for understanding another, or a performance expectation may be linked to more than one practice and more than one core idea. *Framework*-aligned assessment tasks will need to be constructed so that they provide information about how well students make these connections. For example, a task that focused only on students' knowledge of a particular model would be less revealing than one that probed students' understanding of the kinds of questions and investigations that motivated the development of the model. Tasks that do not address these connections will not fully capture or adequately support the kind of science learning called for in the *Framework*.

Information about progress. The *Framework* and related standards address the process of learning science. They make clear that students should be encouraged to take an investigative stance toward their own and others' ideas, to be open about what they are struggling to understand, and to recognize that struggle as part of the way science is done, as well as part of their own learning process. Thus, revealing students' emerging capabilities with science practices and their partially correct or incomplete understandings of core ideas is an important function of classroom assessment. The *Framework* also postulates that students will develop disciplinary understandings by engaging in practices that help them to question and explain the functioning of natural and designed systems. Although learning is an ongoing process for both scientists and students, students are emerging practitioners of science, not scientists, and their ways of acting and reasoning differ from those of scientists in important ways. The *Framework* discusses the importance of seeing learning as a trajectory along which students gradually progress in the course of a unit, a year, and across the whole K-12 span and of organizing instruction accordingly. Thus, uncovering students' incomplete forms of practice and understanding is critical: Assessments will need to clearly define the forms of evidence associated with beginning, intermediate, and sophisticated levels of knowledge and practice expected for a particular instructional sequence. A key goal of classroom assessments is to help teachers and students understand what has been learned and what areas will require further attention. Such assessments will also need to identify likely misunderstandings and interim goals for learning.

Examples

In the chapter on Classroom Assessment, the Developing Assessments report (Pellegrino et al., 2014) provides six examples of assessment tasks to be embedded in classroom instruction. In these tasks, the kinds of activities that are part of high-quality instruction are deployed in ways to yield actionable assessment information. A description of one of those examples is presented in Exhibit 4. The six examples reveal how classroom work products and discussions can be used as formative assessment opportunities, and several of the examples include summative assessments as well. In each case, the evidence produced provides teachers with information about students' thinking and their developing understanding that is useful for guiding next steps in instruction. Moreover, the time students spend in doing and reflecting on these tasks should be seen as an integral part of instruction rather than as a stand-alone assessment task. The example assessment tasks also produce a variety of products and evidence that can be scored. Some include illustrations of typical student work, and some include the construct map or scoring rubric used to guide data interpretation.

The examples are drawn from different grade levels and assess knowledge related to different disciplinary core ideas. Evidence from their use documents that, with appropriate prior instruction, students can successfully carry out such tasks. Further discussion of any of the six cases is beyond the scope of this white paper. Yet one conclusion is that the six examples constitute existence proofs in support of the claim that it is possible to design tasks that elicit students' thinking about disciplinary core ideas and crosscutting concepts by engaging them in scientific practices and that students can respond to such tasks successfully. It goes without saying that much more research and development need to be done if we are to have classroom assessments that teachers can use to monitor student learning toward attainment of the competencies implied by the NGSS and other college- and career-ready science performance expectations. The work must be guided by the conceptual frames and principled design processes discussed earlier.

Multidisciplinary Development

Developing assessment tasks of the type needed requires the participation of several different types of experts. For the tasks to accurately reflect science ideas, scientists must be involved. Experts in science learning will also be needed to ensure that knowledge from research on learning is used as a guide to what is expected of students. Assessment experts will be needed to clarify relationships among tasks and the forms of knowledge and practice that the items are intended to elicit. Practitioners will need to be involved to ensure that the tasks and interpretive frameworks linked to them are usable in classrooms. And this multidisciplinary group of experts will need to include people who have knowledge of and experience with population subgroups, such as students with disabilities and students with varied cultural backgrounds, to ensure that the tasks are not biased for or against any students for reasons irrelevant to what is being measured.

For teachers to incorporate tasks of the type needed into their practice and to design additional tasks for their classrooms, they will need to have worked with many good examples in their curriculum materials. Because many of the classroom assessments used by teachers are derived from curriculum materials, it is especially critical that the assessment materials found in curricula, textbooks, and other resources, such as digital content, also reflect the characteristics discussed above. Thus, curriculum developers and others who are creating resource materials to align with the *Framework* and associated standards should ensure that the assessment activities included (such as, mid- and end-of-chapter activities, suggested tasks for unit assessment, and online activities) require students to engage in practices that demonstrate their understanding of core ideas and crosscutting concepts.

Professional Development

In classroom assessment, significant adaptation will be asked of teachers, and they will need support from other levels of the system including school principals and district administrators. For example, teachers will need systematic opportunities to learn how to use classroom discourse as a means to elicit, develop, and assess student thinking. Professional development will need to include opportunities for teachers to learn how to orchestrate classroom discussion of the core disciplinary ideas that are integrated with the use of various practices such as modeling, explanation, and argumentation. Eliciting student thinking through skillful use of discussion is not enough, however. Assessment tasks and teacher questions also must successfully elicit and display students' appropriate and problematic ways of reasoning about disciplinary core ideas and appropriate and problematic aspects of their participation in practices. They must also elicit students' interests and experiences so that instruction can build on them. This is part of the larger process of integrating teaching and assessment in alignment with the *Framework* and associated standards. Thus, both teachers and assessment developers need to be aware of typical student ideas about a topic and the various alternative conceptions that students are likely to hold. In addition, teachers need guidelines for interpreting students' responses to tasks or questions. Such guidelines should be intelligible and usable in practice: they cannot be so elaborate that teachers find them difficult to use in real time as they try to understand their students' thinking during instruction.

Development of the Monitoring (Large-Scale) System Components

A Multiplicity of Questions and Design Options

This system component is used to monitor or audit student learning over time, and it is often referred to as external or large-scale assessment. Such assessments can be used to answer a range of important system-level questions about student learning, and Exhibit 6 shows examples of the variety of questions that such assessments might be designed to answer. As implied by this range of questions, monitoring is complex and can take multiple forms depending on policy concerns and the intended interpretive uses of the results.

In the United States, the data currently used to answer monitoring-related questions about science learning are obtained predominantly through assessments that use one of two types of test administration strategies. One is a fixed-form test, in which all students on a given testing occasion take the same or comparable forms of the test.[3] The science assessments that states used to comply with the No Child Left Behind (NCLB) Act are examples of this strategy: Each public school student at the tested grade level in a given state took the full test. According to NCLB requirements, these tests were given to all students in the state at least once in each of three grade spans (K–5, 6–8, 9–12). Fixed-form tests of all students (census tests) are designed to yield individual-level scores, which are used to address the questions about student-level

Exhibit 6. Possible questions that might need to be addressed by a science monitoring assessment

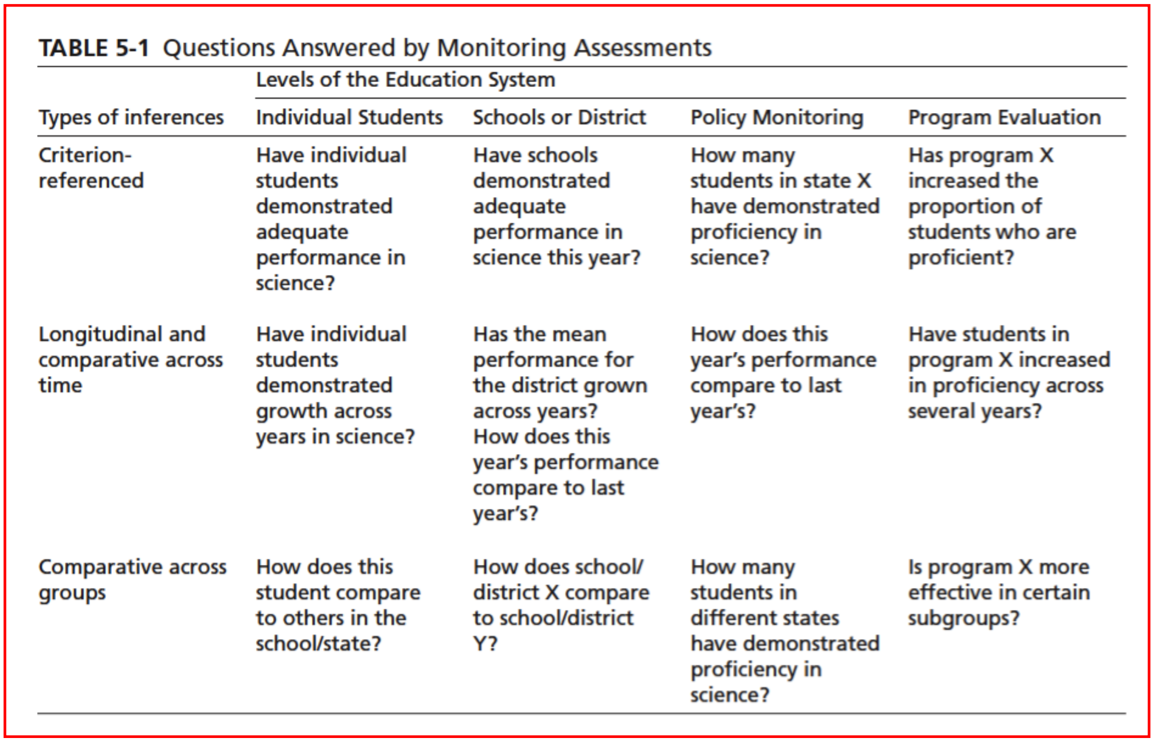


TABLE 5-1 Questions Answered by Monitoring Assessments

	Levels of the Education System			
Types of inferences	Individual Students	Schools or District	Policy Monitoring	Program Evaluation
Criterion-referenced	Have individual students demonstrated adequate performance in science?	Have schools demonstrated adequate performance in science this year?	How many students in state X have demonstrated proficiency in science?	Has program X increased the proportion of students who are proficient?
Longitudinal and comparative across time	Have individual students demonstrated growth across years in science?	Has the mean performance for the district grown across years? How does this year's performance compare to last year's?	How does this year's performance compare to last year's?	Have students in program X increased in proficiency across several years?
Comparative across groups	How does this student compare to others in the school/state?	How does school/district X compare to school/district Y?	How many students in different states have demonstrated proficiency in science?	Is program X more effective in certain subgroups?

Source: Adapted from Pellegrino et al., 2014.

[3] With the passage of the Every Student Succeeds Act to replace NCLB, testing requirements will likely allow for computer adaptive testing and such an approach is being used as part of the Smarter Balanced assessment design for ELA and mathematics.

performance shown in the second column of Exhibit 6. The scores are also aggregated as needed to provide information for the monitoring questions about school-, district-, and state-level performance shown in the three right-hand columns.

Matrix sampling, the other type of test administration strategy, is used when the primary interest is group- or population-level estimates (i.e., schools or districts) rather than individual-level estimates. No individual student takes the full set of items and tasks. Instead, each of the tasks is completed by a sample of students that is sufficiently large and representative to yield valid and reliable scores for schools, states, or the nation. This method makes it possible to gather data on a larger and more representative collection of items or tasks for a given topic than any one student could be expected to complete in the time allocated for testing. In some applications, all students from a school or district are tested (with different parts of the whole test). In other applications, only some students are sampled for testing but in sufficient number and representativeness that the results will provide an accurate estimate of how the entire school or district would perform. Such a test can provide data to answer some of the monitoring questions in Exhibit 6 but not the questions in the second or fifth columns. When individual student results are not required, matrix sampling is a powerful, economical, and relatively straightforward option. Matrix-sampling approaches have not generally been possible in state testing in the last decade because of the requirements of NCLB for individual student reporting.

These two types of administration strategies for external assessments can be combined to answer different monitoring questions about student learning. Both approaches can be combined in a single test: For example, a test could include both a fixed-form component for estimating individual performance and a matrix-sampled component used to estimate a fuller range of performance at the school level. This design was used by several states before the implementation of NCLB, including Massachusetts, Maine, and Wyoming. Such hybrid designs can be constructed to include a substantial enough fixed or common portion of the test to support individual estimates, with each student taking one of multiple matrix forms to ensure broad coverage at the school or district level.

Regardless of the form that a monitoring assessment takes, the tasks used must have the same basic characteristics discussed earlier to align with the *Framework* and associated standards: they need to address the progressive nature of learning, include multiple components that reflect three-dimensional science learning, and include an interpretive system for the evaluation of a range of student products. In addition, assessments for monitoring need to be designed so that they can be given to large numbers of students, are sufficiently standardized to support the intended monitoring purpose, cover an appropriate breadth of the standards, and are cost-effective.

Fulfilling the Monitoring Function via Multiple Components

A number of key issues affect the design of a valid assessment to fulfill the monitoring function. First, it will not be possible to cover all the performance expectations for a given grade (or grade band) during a typical single testing session of 60–90 minutes. To obtain a sufficient estimate of students' proficiency with the performance expectations, multiple testing sessions will be necessary. Even with multiple testing sessions, however, assessments designed for monitoring purposes cannot fully cover the set of performance expectations for a given grade. One implication of this is that *Framework*-aligned assessments for monitoring should include some combination of tasks given at a time mandated by the state or district (**on-demand assessment components**) and tasks given at a time that fits the instructional sequence in the classroom (**classroom-embedded assessment components**). These two designs should not be viewed as either-or options. Rather, they can be creatively and selectively combined, with varying weighting, to produce a monitoring assessment that appropriately and adequately reflects the depth and breadth of college- and career-ready standards compatible with the *Framework* (Pellegrino et al., 2014).

Second, assessments for monitoring, like assessments used for instructional support in classrooms, must be composed of multiple types of tasks. The assessments themselves, as well as the individual tasks that comprise them, will be in varied formats—some that require actual demonstrations of practices, some that make use of short- and extended-constructed responses, and some that use carefully designed selected-response (multiple-choice) questions. Use of multiple assessment task components will help cover the performance expectations more completely than any assessment that uses only one format.

Third, the use of technology holds promise in addressing some of the practical challenges of such a mixed-format assessment. For example, technology can be useful in scoring multiple aspects of students' responses on performance tasks, and technology-enhanced questions (e.g., those using simulations or data display tools) can be useful and may be essential for giving students efficient ways to demonstrate their proficiency in some of the practices. Nevertheless, technology alone is unlikely to solve problems of score reliability or of equating, among other challenges.

Finally, we need to assume that matrix sampling will be important in the design of assessments for monitoring purposes to ensure proper coverage of the entire *Framework*. Matrix sampling as a design principle may be extremely important, even when individual scores are needed as part of the monitoring process. This would include hybrid designs in which all students respond to the same core set of tasks that are mixed with matrix sampled tasks to ensure representativeness of the full *Framework* for monitoring purposes (making inferences about student learning at higher levels of aggregation, columns, 2–4 in Exhibit 6).

Options for Developing and Implementing the On-Demand Components

The on-demand assessment component should be composed of sets of multicomponent tasks. To the extent possible, these tasks should include, as a significant and visible aspect of the assessment, multiple performance-based questions. When appropriate, computer-based technology should be used to broaden and deepen the range of performances demanded on these assessments. The on-demand component might be administered in one or more sessions toward the end of a given academic year. Such an assessment would be designed to cover multiple aspects of the *Framework* and associated standards and might typically be comprised of mixed-item formats including written constructed responses and/or performance tasks. The revised AP science exams are examples of mixed-item format assessment tasks that include both selected-response items and free-response questions. The AP free-response questions include both short-answer and extended- constructed responses. Two current state-level assessment programs, the New England Common Assessment Program and New York's science assessments, have a mixed-item format with performance activities. Performance events could be a set of tasks that center on a major science question. This task set could include assessment questions in a variety of formats, such as some short-answer questions and some short constructed-response items, all of which lead to producing an extended response for a complex performance task. The short-answer questions could help activate prior knowledge to provide scaffolds for the more complex tasks for a broad range of students.

Ideally, three or four of these performance assessments would be administered during the academic year, which would allow the task sets to cover a wider breadth of topics. The use of multiple items and multiple response types would help to address the reliability concerns that are often associated with the scores reported for performance-based tasks (see Davey et al., 2015; Dunbar, Koretz, & Hoover, 1991). Use of multiple task sets also opens up other design possibilities, such as using a hybrid task sampling design: In this design, all students at a grade level receive one common performance task, and other tasks are given to different groups of students using matrix sampling. This design allows the common performance task to be used as a link for the matrix tasks so that student scores could be based on all the tasks they complete.

Options for Developing and Implementing the Classroom-Embedded Components

The second proposed component of a monitoring system would be classroom-embedded tasks and performances administered at different points in a given academic year to align with the completion of major units of instruction. These instructional units and assessments would be targeted at various sets of standards, such as those associated with one or more core ideas in the life sciences. Such a classroom-embedded assessment would be designed to cover more selective aspects of the science standards and would be comprised of tasks that require written constructed responses and/or performance activities. The classroom-embedded assessment component could take various forms, three of which are briefly described here.

One option involves the use of replacement units. These are curricular units that have been approved centrally by the district or state for use throughout its jurisdiction and made available to schools. These units would cover material or concepts that are already part of the curriculum, and they would be designed to teach the material in a way that promotes deeper learning (see NRC, 2012). The replacement units would not add topics to the curriculum but would be exemplary substitutes for existing units. Replacement units would be designed to be used locally as meaningful examples to support implementation of the state's science standards. The end-of-unit assessment in the replacement unit could include performance tasks and perhaps shorter constructed-response tasks (similar to those described in the previous two sections). The assessments could be scored locally by teachers or via a central or regional scoring mechanism.

Replacement units could be designed by state consortia, regional labs, commercial vendors, or other groups on a high-priority topic for a given grade level. Each unit would include instructional supports for educators and formative assessment probes. The supports embedded in the replacement units would serve as a useful model of how to improve classroom assessment practices at a relatively large scale.

A second option would be for a state or district (or its contractors) to design standardized performance tasks that would be available for teachers to use at the appropriate time. Classroom teachers could be trained to score these tasks, or student products could be submitted to the district or state and scored centrally.

A third option would be for a state or district to provide criteria and specifications for a set of performance tasks to be completed and assembled as work samples at set times during the school year. The tasks might include assignments completed during a school day or homework assignments. The state or local school system would determine the scoring rubric and criteria for the work samples. Classroom teachers could be trained to score these tasks, or they could be submitted to the district or state and scored centrally.

Implementing any of these options for using classroom-embedded assessments for monitoring purposes leaves a number of important decisions to the district and/or school. While this can have many positive consequences, quality control procedures would be essential so that these assessments meet appropriate technical standards (e.g., AERA/APA/NCME, 2014).

Building a Coherent State Science Assessment System: Implementation and Accountability

State education leaders and policy makers need to understand and plan for the development and implementation of new science assessment systems in stages, over a span of years. We know that a number of innovative assessment programs floundered in the 1990s in part because they were implemented far too rapidly (perhaps to meet political exigencies). In many cases, the developers were not given sufficient time to implement what were major changes or to make modifications as they learned from experience (McDonnell, 2004). Some have cited this rush to implement at scale as a key factor in the lack of sustainability of many such efforts (see NRC, 2010).

Any new assessment system has to evolve alongside other elements that are also changing. It will take time for the changes to curriculum, instruction, professional development, and the other components of science education envisioned in the *Framework* to be developed and implemented. Coordinating new modes of assessment with those changes will be necessary, both because what is needed has to be embedded into curriculum and instruction and because there is little value in assessing students on material and kinds of learning that they have not had the opportunity to master. With regard to opportunity to learn, many schools and districts have reduced the amount of science instruction, particularly in the early grades, in response to the accountability demands of NCLB that will likely change under ESSA. Many jurisdictions will need to reintroduce science in the early grades and review and revise policies that have limited the time available for it if they are to effectively implement the new standards. Often, schools that serve the most disadvantaged student populations are those where the opportunity to learn science has been most reduced. Even in schools and districts that have maintained strong science programs at all grade levels, neither students nor teachers have had experience with instruction that involves applying the practices as envisioned in the *Framework*.

Given the magnitude of the change needed across multiple aspects of science education, policy makers would do well to adopt an orientation toward assessment systems development that is bottom up (i.e., grounded in the classroom) rather than top down (grounded in such external needs as monitoring, accountability, and/or teacher evaluation). Such an approach is most likely to yield the evidence needed to support instruction and learning that are aligned with the *Framework's* goals. Although monitoring and accountability are important functions of an assessment system, placing the initial focus on developing high-quality, valid assessments that are as close as possible to the point of instruction will be the best way to identify successful strategies for teaching and assessing science in ways that promote deep learning. Such assessment strategies can then serve as the basis for developing assessments for purposes such as monitoring and accountability.

A bottom-up orientation to developing a system of science assessments should not be construed as avoidance of the need for the monitoring components of that system or the role a monitoring component might play in accountability. There is little doubt that assessments developed to measure science proficiency will be used for accountability purposes, so it is important to consider how accountability policies might affect the ways the assessments operate within the system. The incentives that come with accountability can serve to support or undermine the goals of improving student learning (Koretz, 2009; NRC, 2011). Most likely whoever is held accountable within a school system will make achieving higher scores a major goal of science teaching. In practice, accountability policies often result in teaching to the test, so that testing tends to drive curriculum and instruction, even though the avowed intention may be for curriculum and instruction to drive testing (Koretz, 2005; 2009). Too often, the result of test-based accountability has been a narrowing of the curriculum to match the content and format of what is tested, which has led to coverage of superficial knowledge at the expense of understanding and reasoning practices that are not assessed (Dee, Jacob, & Schwartz, 2013). Schools and classrooms serving students with the greatest educational needs are often those presented with the most ambitious challenges for improvement and thus also face the greatest pressure to teach to the test. Thus, it is extremely important that the assessments used for monitoring and accountability purposes measure the learning that is most valuable.

One implication of this is that continued use of the large-scale science assessments that states developed under NCLB is neither appropriate nor advisable. Such monitoring instruments are not aligned with the *Framework* or college- and career-ready science standards, so they will not support the changes desired in teaching and learning. Interim solutions will be needed that can simultaneously satisfy federally mandated testing requirements and allow the space for change in classroom practice. The recent passage of ESSA, with its increased flexibility on the use of assessment results, may provide an opportunity in this area. As discussed, the three-dimensional learning described in the *Framework* cannot be well assessed without some use of more extended engagements with multipart science assessment tasks. We also emphasized that the assessments used for monitoring purposes will need to include both on-demand and classroom-embedded assessment components. Thus, if accountability policies are part of the science education system, they must incorporate results from a variety of types of assessments. When external on-demand assessments predominate in an assessment system and are the sole basis for monitoring and accountability, curriculum and instruction are most likely to become narrowed to reflect only the material and testing formats that are represented on these assessments (Koretz, 2005; 2009).

In summary, developing and implementing new state assessment systems will require a transition period, just as the implementation of college- and career-ready science standards for curriculum and instruction will require a gradual and strategic approach. A gradual approach will ease the transition process and strengthen the resulting system, both by allowing time for development and phasing in of curriculum materials aligned with the *Framework* and by allowing all participants to gain familiarity and experience with new kinds of instruction and assessment that address the three dimensions of the *Framework*. Ideally, the transition period for full system design and implementation might be 5 years or more, but this need for transition time is juxtaposed with the realization that many states will face political pressures for much shorter timelines for implementation. Even so, a balanced approach to system design and implementation is still warranted.

Final Thoughts: The Road Ahead

Assessment is a key element in the process of educational change and improvement. Done well, it can signify what we want students to know and be able to do and can help educators create the learning environments that support attainment of those objectives. Done poorly, it sends the wrong signals and skews the teaching and learning process toward teaching to tests that have little relationship to the competencies students will need in the future. In the case of science assessment, we have an opportunity to rethink and redesign our approach to assessment so that it more closely aligns with the vision of competence in science in which the practices of scientific reasoning are intimately connected with the understanding and application of core disciplinary ideas and crosscutting concepts. Defining the nature of such knowledge and understanding and developing valid ways to assess its attainment present a substantial design and implementation challenge. That said, there are tools, methods, and technologies available that make these design and engineering tasks possible, especially if we are willing to invest in the effort and provide time and opportunity for assessment to be well integrated with curriculum and instruction. The greatest danger may be a rush to turn any *Framework*-associated standards into sets of assessment tasks for use on high-stakes accountability tests before we have adequately engaged in the needed research, development, and validation. Hopefully, we have learned enough from our experience with implementing the Common Core State Standards for ELA and math and the Race to the Top assessment programs that the teaching, learning, and assessment of science can profit from hindsight and a bit of foresight.

There is very limited evidence that accountability policies to date, which focus largely if not solely on data derived from external large-scale assessments, have led to improved student achievement (NRC, 2011). In contrast, the positive relationship between classroom assessment and student learning outcomes is well established (Black & Wiliam, 1998; Kingston & Nash, 2011; NRC, 2007). Assessment that closely aligns with curriculum and instruction and that engages students in the kinds of science learning described in the *Framework* will return the focus to what is most important—the direct support of students' learning.

References

Achieve (2013). *Next Generation Science Standards.* Retrieved from http://www.nextgenscience.org

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progression in science: Current challenges and future directions.* Rotterdam, Netherlands: Sense.

American Association of Medical Colleges. (2012). *MR5 fifth comprehensive review of the Medical Colleges Admission Test (MCAT): Final MCAT recommendations*. Retrieved from https://www.aamc.org/download/273766/data/finalmr5recommendations.pdf

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–73.

College Board. (2009). *Science: College Board standards for success.* New York, NY: Author.

College Board. (2010). *From research to practice: Redesign AP science courses to advance science literacy and support learning with understanding.* Final Report submitted to the National Science Foundation, Award # ESI-0525575. Retrieved from http://www.nsf.gov/awardsearch/showAward?AWD_ID=0525575

College Board. (2011a). *The AP biology curriculum framework 2012–2013*. New York, NY: Author.

College Board. (2011b). *The AP chemistry curriculum framework 2013–2014.* New York, NY: Author.

College Board. (2012). *AP biology: Course and exam description effective fall 2012.* New York, NY: Author.

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* New York, NY: Columbia University, Teachers College, Consortium for Policy Research in Education, Center on Continuous Instructional Improvement.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment.* Retrieved from http://www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf

Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis, 35*(2), 252–279.

Dunbar, S., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4*(4), 289–303.

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation.* Committee on Appropriate Test Use; Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academies Press.

Huff, K. Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310–324.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Koretz, D. (2005). *Alignment, high stakes, and inflation of test scores.* CSE Report 655. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, Graduate School of Education & Information Studies, University of California, Los Angeles.

Koretz, D. (2009). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L. (2013). *Test-based accountability.* Princeton, NJ: Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/ publications_reports/ assessment_education.html

McDonnell, L. M. (2004). *Politics, persuasion, and educational testing.* Cambridge, MA: Harvard University Press.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379–416.

Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25,* 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: Erlbaum.

National Center for Educational Statistics (2008). *Science framework for the 2009 National Assessment of Educational Progress.* Washington, D.C.: National Assessment Governing Board.

National Center for Educational Statistics. (2012). *The nation's report card: Science in action: Hands-on and interactive computer tasks from the 2009 Science Assessment (NCES 2012-468).* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Research Council. (1996). *National science education standards.* National Committee for Science Education Standards and Assessment. National Committee on Science Education Standards and Assessment, Board on Science Education, Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning.* S. Olson & S. Loucks-Horsley (Eds.), Committee on the Development of an Addendum to the National Science Education Standards on Scientific Inquiry, National Research Council. Washington, DC: National Academy Press.

National Research Council. (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment.* Washington, DC: National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grade K-8.* Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A., Schweingruber, & A. W. Shouse (Eds.), Washington DC: National Academy Press.

National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits.* Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A., Schweingruber, & A. W. Shouse (Eds.), Washington DC: National Academy Press.

National Research Council. (2010). *State assessment systems: Exploring best practices and innovations: Summary of two workshops.* A. Beatty, Rapporteur, Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

National Research Council. (2011). *Incentives and test-based accountability in education.* M. Hout & S. W. Elliott (Eds.), Committee on Incentives and Test-Based Accountability in Public Education, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing?* Committee on the Evaluation Framework for Successful K-12 STEM Education, Board on Science Education and Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on a Conceptual Framework for New K-12 Science Education Standards, Board on Science Education. Washington, DC: National Academies Press.

Organization for Economic Cooperation and Development. (2009). *PISA 2009 Assessment framework: Key competencies in reading, mathematics and science.* Paris, France: Author.

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831–841.

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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the Next Generation Science Standards.* Washington, DC: National Academies Press.

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.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Wiliam, D. (2007). Keeping learning on track: Formative assessment and the regulation of learning. In F. K. Lester, Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age.

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2005). *Systems for state science assessments.* Washington DC: National Academies Press.

Wood, W. B. (2009). Revising the AP biology curriculum. *Science, 325,* 1627–1628.

SRI Education™

SRI Education, a division of SRI International, is tackling the most complex issues in education to identify trends, understand outcomes, and guide policy and practice. We work with federal and state agencies, school districts, foundations, nonprofit organizations, and businesses to provide research-based solutions to challenges posed by rapid social, technological and economic change. SRI International is a nonprofit research institute whose innovations have created new industries, extraordinary marketplace value, and lasting benefits to society.

Silicon Valley
(SRI International headquarters)
333 Ravenswood Avenue
Menlo Park, CA 94025
+1.650.859.2000
education@sri.com

Washington, D.C.
1100 Wilson Boulevard, Suite 2800
Arlington, VA 22209
+1.703.524.2053

www.sri.com/education

Stay Connected