We face extraordinary promise for the future of science learning, juxtaposed with substantial challenges in achieving the vision of what it means to be proficient in science. Among those challenges are determining how the proficiency of our students will be assessed relative to that vision and doing so in ways that support, rather than inhibit, teaching and learning. Educational assessments ought to be statements about what scientists, educators, policy-makers, and parents want students to learn and become. It is well established that what we choose to assess will end up being the focus of instruction. So, it is critical that science assessments, both external and internal to the classroom, best represent the proficiencies we desire. This Review argues that much of what is needed to effectively assess science learning, either at the classroom level or for purposes of system monitoring, has yet to be created and that design and implementation challenges are substantial. Even so, there are promising cases from which to learn and build.

Shared Perspectives on Proficiency
A disjuncture exists between students’ knowledge of science facts and procedures, as assessed by typical achievement tests, and their understanding of how that knowledge can be applied through the practices of scientific reasoning, argumentation, and inquiry (3, 4). This problem is recognized in reports spanning kindergarten (K) to grade 16+ (K-16+) that simultaneously present a consistent description of what proficiency in science should be (1, 5–11). Seldom has such a consistent message been sent as to the need for 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. The emergent definition of proficiency is perhaps most clearly expressed in three major elements of the U.S. National Research Council (NRC) Framework for K-12 Science Education (1): (i) core or “big” ideas within disciplinary areas, (ii) practices of scientific and engineering reasoning, and (iii) cross-cutting concepts. Collectively they define what it means to know science, not as separate elements but as intertwined aspects of knowledge and understanding [see also (12)]. It is not just the description of each and their intersection that matters but also that the meaning of proficiency is realized through performance expectations about what students at various levels of educational experience should know and be able to do. These statements move beyond vague terms such as “know” and “understand” to more specific statements like “analyze,” “compare,” “explain,” “argue,” “represent,” “predict,” “model,” etc. in which the practices of science are wrapped around and integrated with core content. Educators and researchers are also recognizing that proficiency develops over time and increases in sophistication and power as the product of coherent systems of curriculum, instruction, and assessment.

The virtue of such a view is that science educators are poised to better define the outcomes desired from their instructional efforts, which in turn guides the forms of assessment that can help them know whether their students are attaining the desired objectives, as well as how they might better assist them along the way. It is very important for the science education community, and policy-makers and the public more broadly, to develop a shared perspective on what constitutes high-quality and valid science assessments across K-16+ if assessments are to support teaching and learning and attainment of the desired science education outcomes.

Proficiency, Performance Expectations, and Assessment Design Challenges
The NRC Framework uses the logic of progressions to describe students’ developing proficiency in three intertwined domains—practices, cross-cutting concepts, and core ideas—in a coherent way across grades K through 12. The framework builds in the idea of a 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. The framework also provides sketches of possible progressions for acquiring each practice or cross-cutting concept but does not indicate the expectations at any particular grade level. The Next Generation Science Standards (NGSS) (7) build on these suggestions and include tables that define what each practice might encompass and the expected uses of each cross-cutting concept for students at each grade level. This integrated perspective of what it means to know science suggests that assessment should help determine where a student can be placed along a sequence of progressively more “scientific” understandings of a given core idea that by definition includes successively more sophisticated applications of practices and cross-cutting concepts. This is an unfamiliar idea in the realm of science assessments, which have more often been viewed as simply measuring whether students know particular grade-level content. It means that assessments must strive to be sensitive both to grade-level appropriate performances and to intermediate performances that may be appropriate at somewhat lower or higher grade levels. This is particularly important for the design of assessment materials and resources that can be used in classrooms to support instruction.

The NRC Framework states that assessment tasks must be designed to gather evidence of students’ ability to apply the practices and their understanding of the cross-cutting concepts in the contexts of problems that also require them to draw on their understanding of specific disciplinary ideas. It suggests using a model put forward in Science Standards for College Success (13) by expressing standards in terms of performance expectations. The organization Achieve and its partners in NGSS development have elaborated these guidelines into standards that are clarified by descriptions of the ways in which students at each grade are expected to apply both the practices and the cross-cutting concepts and of the knowledge they are expected to have of the core ideas. The NGSS appear as sets of performance expectations related to a particular aspect of a core disciplinary idea (see the draft example in Fig. 1). Each performance expectation asks students to use a specific practice in the context of a specific element of the disciplinary knowledge relevant to the particular aspect of the core idea. Across the set of expectations at a given grade level, each practice and cross-cutting concept appears with multiple standards.

Performance expectations also may 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, lack sufficient detail to create an assessment.

From NRC Frameworks, Standards, and Performance Expectations to Assessments
The design of valid and reliable science assessments hinges on elements that include but are not...
restricted to what is articulated in disciplinary frameworks and standards such as those illustrated above (14, 15). In the design of assessment items and tasks related to performance expectations, one needs to also consider (i) the kinds of conceptual models and evidence in which we expect students to engage, (ii) grade-level appropriate contexts for assessing performance expectations, (iii) options for task design features (e.g., computer-based simulations or animations, paper-pencil writing and drawing) and which of these are essential for eliciting students’ ideas about the performance expectation, and (iv) the types of evidence that will reveal levels of student understanding and skill.

Assessment involves evidentiary reasoning (14), so it has proven useful to be more systematic in framing assessment development as an evidence-centered design process (ECD) [e.g., (15, 16)]. The process starts by defining the claims that one wants to be able to make about student proficiency—the ways in which students are supposed to know and understand some particular aspect of a domain. Examples might include aspects of force and motion or heat and temperature. The most critical aspects of defining these are to be as precise as possible about what matters and to express this in the form of verbs such as “model,” “explain,” “predict,” etc. In essence, the performance expectations found in the NGSS are claims about student proficiency.

Claims about the student must be linked to forms of evidence that would support those claims. Evidence statements capture features of work products or performances that would give substance to the claims. This includes which features need to be present and how they are weighted—what matters most, least, or not at all. If evidence in support of a claim about a student’s 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 drawing a free-body diagram with all the forces labeled, including their magnitudes and directions.

The precision that comes from elaborating the claims and evidence statements pays off when it is time to design tasks or situations to provide the requisite evidence. Tasks 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 necessary evidence and should allow students to show what they know in ways that are as unambiguous as possible with respect to what the performance implies about student knowledge and skill (17).

Science Assessment Example Cases

Given the relative newness of the NRC Framework, it is no surprise that comprehensive sets of assessment examples that align completely with the NGSS performance expectations do not exist. Many of the tasks that have been used for classroom
assessment, and 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 few exceptions, such assessments do not integrate core concepts and science practices in the ways intended by the NRC Framework or NGSS. Nevertheless, we can draw from example cases to illustrate what is needed, many of which have used an ECD approach to guide assessment design and validation. The examples are diverse in several ways—including the science content and practices represented, age and grade level, intended use (at the classroom, state, national or international level; whether the consequences of student performance have low or high stakes), as well as the innovative use of technology.

**Classroom Instruction and Assessment**

Several projects have developed assessments for use in classroom instruction with a particular emphasis on the integration of core science concepts with one or more science practices such as modeling; evidence-based explanation and argumentation; and/or the design of investigations to test hypotheses, analyze results, and construct explanations from data. Some of the clearest examples can be found in a volume edited by Alonzo and Gotwals (18) focused on learning progressions and in a special issue on assessment of the *Journal of Research on Science Teaching* (19). Additional examples include the SimScientists (20, 21), Science ASSISTments (22), and BioKids projects (23).

Several of these projects illustrate the feasibility of designing tasks and situations, whether in paper-and-pencil format or mediated via simulations embedded in technology, that challenge students to reason with and about core science concepts in life and physical science. They demonstrate ways to obtain forms of evidence that can serve multiple purposes, such as measurement of student proficiency as well as diagnosis of student thinking for instructional improvement. The SimScientists project has shown how assessment situations and tasks involving dynamic simulations of science phenomena can be built from a principled design process that supports classroom formative assessment as well as summative assessment in large-scale state programs (21).

**National and International Large-Scale Assessment**

Much of what students and teachers experience as science assessments is external to regular classroom instruction and comes in the form of large-scale state tests, for example, administered in response to the U.S. No Child Left Behind legislation. Although the quality of such state assessments varies, none approximates the performance expectations discussed in the NRC Framework and NGSS. In contrast, there are two large-scale assessment programs that more closely exemplify aspects of science proficiency that involve science practices: the U.S. National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA).

The NAEP 2009 and 2011 assessments were constructed from a framework document that identified specific areas of content in the life, physical, and Earth and space sciences, as well as a set of science practices: (i) identifying science principles, (ii) using science principles, (iii) using scientific inquiry, and (iv) using technological design. Item types fell into two broad categories: selected-response items (such as multiple choice) and constructed-response items (such as short answer). To further probe students’ abilities to combine their understanding with the investigative skills that reflect practices, a subset of the students completed hands-on performance or interactive computer tasks (3, 4, 24, 25). In contrast to NAEP, which is administered to 4th-, 8th-, and 12th-grade students, the PISA assessment is administered only to 15-year-olds. The most recent PISA science assessment results are based on a framework that includes science proficiencies that overlap with the science practices of the NRC Framework and NGSS, as well as aspects of the NAEP framework (26, 27).

What are especially important about both NAEP and PISA are the sets of simple and complex science assessment tasks that demand reasoning about science content as described in the NRC Framework and NGSS. Both assessment programs are a source of examples of the types of performances that align with the descriptions of proficiency discussed earlier. Neither NAEP nor PISA represent static assessment programs. Both undergo major revisions to the framework used to guide assessment design and task development, and both are increasingly moving to incorporate technology as a key aspect of task design and assessment of student performance. It is likely that the NAEP framework will be revised within the next decade, and work is already under way in revising the PISA science framework for 2015. Changes in both will ostensibly move in directions that even more closely align with the NRC Framework. Thus, both might constitute reasonable ways to monitor overall progress of science teaching and learning in U.S. classrooms in ways consistent with implementation of the NRC Framework and NGSS.

**Advanced Placement (AP) Science**

A contemporary approach to rethinking science proficiency can be found in the redesign of the AP courses and assessments for biology, chemistry, and physics (8, 9, 28, 29). The AP program offers college-level curricula to high school students. Starting in 2006, the College Board, which administers AP, with support from the U.S. National Science Foundation, initiated a process that started by redefining the focus, critical content, and science practices that should define proficiency at the end of each AP science course (30). This would then guide development of both a curriculum framework for each course as well as the high-stakes assessment often used by colleges for purposes of granting course credit and/or advanced course placement.

Using the complementary processes of backward design (31) and ECD, a framework was developed for each science discipline that is organized in terms of disciplinary big ideas, enduring understandings, and supporting knowledge as well as a set of seven science practices. This structure parallels that of the core ideas and science practices in the NRC Framework. Similar to what is advocated in the NRC Framework and
realized in the NGSS, performance expectations or learning objectives were defined within each discipline to reflect the blending of core ideas with science practices. Through application of ECD, sets of claim-evidence pairs were elaborated in each science discipline to focus and support course instruction as well as development of assessment tasks for new AP exams.

The first of those exams will be given in May 2013 in biology, with chemistry to follow in 2014 and physics in 2015. To help teachers and students orient to the new course and exam, a wealth of sample assessments were provided (32). To reflect the shift in focus demanded by integration of science practices with core content ideas, new item types were created, and a greater emphasis is being placed on constructed response questions. Figure 2 provides an example of a short constructed response item that involves the integration of conceptual knowledge with aspects of the practices.

AP science redesign is still a work in progress. Much remains to be determined about the quality and impact of the new exams on student learning and classroom instructional practice. But AP science instruction and assessment are changing in ways closely aligned with the perspective on science proficiency described earlier.

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 help educators create learning environments that support attainment of those objectives. Done poorly, it sends the wrong signals and skews teaching and learning. Our greatest danger may be a rush to turn the NGSS into sets of assessment tasks for use on high-stakes state accountability tests before we have adequately engaged in research, development, and validation of the range of tasks and tools needed to get the job done properly. Most especially we must ensure that teachers are given the time, support, and assessment tools to create instructional environments where their students have adequate opportunities to learn what is now expected of them.

References and Notes


17. Although this discussion of a principled approach to assessment design has focused exclusively on specific tasks and situations, it is equally applicable to the design of an assessment of a task—a test or test that might be used for various purposes at the classroom level or beyond. Such a process also requires careful selection and assembly of tasks so that the desired inferences about student proficiency can be supported at the level of the overall instrument or its subparts. Thus, assessment design at multiple levels from separate tasks to larger instruments needs to be approached as a principled design process in which the claims one wants to make about student proficiency are supported by sufficient forms of evidence carefully chosen to meet the intended purpose and use of the assessment results.


30. College Board, From Research to Practice: Redesigning AP Science Courses to Advance Science Literacy and Support Learning with Understanding, final report submitted to the National Science Foundation for award number ES-0525575 (2010).


32. College Board, AP Biology Course and Exam Description, Effective Fall 2012 (College Board, New York, 2012).