Making Sense of New Science Assessments

New state science standards that integrate core ideas, cross-cutting concepts, and practices cry out for new assessments that do likewise.

by James W. Pellegrino

What we choose to assess in science is what will end up being the focus of instruction. US science standards once treated content and inquiry as fairly separate strands of science learning, with content standards stating what students should know and inquiry standards stating what they should be able to do. In its content coverage, these standards were also deemed to be a "mile wide and an inch deep."[1] Assessments followed suit.

By contrast, the National Research Council in its Framework for K-12 Science Education Standards presents a very different way of thinking about science proficiency.[2] The framework articulates three interconnected dimensions of competence: disciplinary core ideas, crosscutting concepts, and scientific and engineering practices.

The framework focuses on core ideas in four areas: (a) life sciences, (b) physical sciences, (c) earth and space sciences, and (4) engineering, technology and the application of science. In so doing, it attempts to reduce the long, often disconnected catalog of factual knowledge that students have been expected to 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, of which the framework identifies seven that have importance across many science disciplines: patterns, cause and effect, systems thinking, and stability and change, for example. Eight key science and engineering practices—the third dimension—include asking questions, defining problems, planning and carrying out investigations, developing and using models, and engaging in argument from evidence.

The 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.

Adopting this view of science competence also requires a change in how learner proficiency is measured. Figure 1 shows a task that could be used to assess middle school students' understanding of the properties of substances that are associated with chemical identity. It does so in the context of constructing an argument about which substances are the same or different based

Figure 1. A Physical Science Assessment Task for Middle School

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: Adapted from C. Harris et al., "Constructing Assessment Items That Blend Core Ideas, Crosscutting Concepts, and Science Practices for Classroom Formative Applications," unpublished paper (Menlo Park, CA: SRI International., 2015).

on patterns of data. A teacher could use this task to gauge how well her students have understood which properties are associated with chemical identity as well as their ability to use evidence to construct a scientific argument.

The central argument of *A Framework for K-12 Science Education Standards* is that competence is realized through performance expectations about what students should know and be able to do. These statements of performance expectations integrate the three dimensions and move beyond vague terms such as "know" and "understand" to more specific statements—"analyze", "compare", "predict", and "model." Finally, *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.

Each performance expectation asks students to use a specific practice and a crosscutting concept in the context of an element of disciplinary knowledge.

How Will We Know What Students Know?

The framework's three-part structure—practices, crosscutting concepts, core ideas—signals an important shift for science education. It also presents a challenge for the design of both instruction and assessment—finding a way to describe and capture students' developing competence along these intertwined dimensions.

In developing the Next Generation Science Standards (NGSS), Achieve and its partners described how students at each grade are expected to apply the practices and crosscutting concepts to the knowledge of core ideas they are expected to have.[3] The standards appear as clusters of performance expectations related to a particular aspect of a core disciplinary idea. Each performance expectation asks students to use a specific practice and a crosscutting concept in the context of an element of disciplinary knowledge. Across each grade-level set of expectations, each practice and crosscutting concept appears multiple times.

But standards and performance expectations, even as explicated in the NGSS, do not provide sufficient detail to create assessments. The designer of valid, reliable science assessments must consider many things at once: the kinds of conceptual models and evidence that we expect students to engage in; grade-level contexts; options for task design such as computer-based simulation or animation, writing, or drawing; and the types of evidence that will reveal levels of understanding and skill. Further, assessments must be developed for specific purposes, and that intended use will drive their design.[4]

In the classroom context, instructors need formative assessments to make decisions about next steps for instruction, give students feedback about their progress, and motivate them. They need summative assessments to help determine whether a student has attained a certain level of competency after completing a particular phase of education.

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 performs relative to others. But the results seldom help teachers or students make day-to-day or month-to-month decisions about teaching and learning. Another common purpose of these assessments is to help administrators, policymakers, or researchers form judgments about the quality and effectiveness of educational programs and institutions. Just as with individuals, the quality of the measure determines the validity of these decisions.

The purpose of an assessment determines priorities, and the context of use imposes constraints on the design. A one-size-fits-all fallacy is often leading to inappropriate choices of assessments for instructional, evaluation, or research purposes that in turn can lead to invalid conclusions regarding persons, programs, or institutions.

A Balanced System

Since one form of assessment cannot serve all the needs of the various actors in an educational system, multiple assessments will inevitably be required. Given the many assessments already used in schools, it is not surprising that educators are often frustrated when such assessments appear to have conflicting goals and yield inconsistent results. Such discrepancies can be meaningful and useful, as when assessments are explicitly aimed at measuring different school outcomes. More often, however, conflicting goals and feedback from the tests shed more heat than light. Thus it is critical that state education policymakers develop a vision for a balanced, coordinated system of assessments to promote effective science teaching and learning that takes into account both classroom assessments and large-scale monitoring assessments.[5]

Table 1. Questions Answered by Monitoring Assessments

Types of Inferences	Levels of the Education System			
	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	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?

Reprinted with permission from Developing Assessments for the Next Generation Science Standards, 2014 *by the National Academy of Sciences, Courtesy of the National Academies Press, Washington, DC.*

The Classroom Component. Classroom assessments are integral to instruction and learning and will be both formative and summative. Assessment activities will tell students what is expected of them and give them opportunities to reflect on their performance. Teachers will get information with which they can adapt their instruction. Instruction that is aligned with the NRC 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 Monitoring Component. This system component is used to answer a range of important system-level questions about student learning (see table 1). In the United States, there are two predominant types of data collection. The first is a fixed-form test, in which all students take the same form of the test on a given occasion. The science assessments that states employ to comply with No Child Left Behind (NCLB) and now the Every Student Succeeds Act (ESSA) are examples: Each public school student at the tested grade level in a given state takes the full test. ESSA requires that these tests be given to all students in the state at least once in each of three grade spans (K-5, 6-8, 9-12). They address questions about student-level performance (first column of table 1). The scores are also aggregated as needed to provide information for the monitoring school-, district-, and state-level performance (the three right-hand columns).

A second type of test administration makes use of matrix sampling, which is used when the primary interest is group- or population-level estimates (i.e., schools or districts) rather than individual-level estimates (the middle two columns of table 1). No individual student takes the full set of items and tasks. Instead, a sufficiently large, representative sample of students completes each task. This method makes it possible to gather data on a larger, 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.

Matrix sampling is a powerful, economical, and relatively straightforward option that has not generally been possible in the context of

James W. Pellegrino co-directs the Learning Sciences Research Institute, University of Illinois at Chicago. He co-chaired the National Research Council's Committee on Developing Assessments of Science Proficiency in K-12 and was also on the committee that produced its report, *A Framework for K-12 Science Education.*

Challenging but Not Impossible

Given the relative newness of the NRC *Framework* and the NGSS, it should come as no surprise that comprehensive sets of examples of assessments that align completely with all performance expectations in all grades do not exist. 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 is now known 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 science assessments that approximate what is needed. Several are presented and discussed in the National Research Council's 2014 report, *Developing Assessments for the Next Generation Science Standards.* They include diverse examples of content, practices, 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—at the classroom, state, national, or international level.

state testing in the last decade because of the requirements of NCLB for individual student reporting.

Both fixed-form and matrix sampling approaches can be combined in a single test. For example, one test could include a fixed-form component for estimating individual performance and a matrix-sampled component to estimate performance at the school level. States such as Massachusetts, Maine, and Wyoming used such a hybrid design before NCLB was implemented.

Regardless of the form that monitoring takes, the tasks must address the progressive nature of learning, include multiple components that reflect three-dimensional science learning, and include an interpretive system for the evaluation of student products. Such assessments also must be designed so 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.

Opportunities to Learn. Another system component—perhaps less obvious than the classroom and monitoring components—is a series of indicators of opportunity to learn. Such indicators will enable states to evaluate the equity of students' opportunity to learn science that is aligned to high-quality standards. Without these indicators, there can be no way to adequately and appropriately interpret the assessment results obtained at each level of the system—especially critical if the results are to be used for accountability.

A Coherent System

State boards of education and state education agencies need to understand and plan for development and implementation of new science assessment systems in stages, over a span of years. Many innovative assessment programs floundered in the 1990s in part because they were implemented far too rapidly (perhaps to meet political exigencies). In many cases, developers were not given sufficient time to implement major changes or to modify assessments as they learned from experience.[6] Some have cited this rush to implement at scale as a key factor in the lack of sustainability of many such efforts.[7]

Any new assessment system has to evolve alongside other elements that are also changing: curriculum, instruction, professional

development, and the other components of science education. Coordination is needed not just because assessments have to be embedded in curriculum and instruction but also because students ought not be assessed on material and kinds of learning they have not had the opportunity to master.

With regard to the opportunity to learn, many schools and districts reduced the amount of science instruction, particularly in early grades, partly in response to the accountability demands of NCLB. If they are to implement the new standards effectively, many jurisdictions will need to reintroduce science in the early grades and review and revise policies that limited the time available for science. Frequently, schools that serve the most disadvantaged students are those in which the opportunity to learn science has been most reduced. Even in schools and districts that 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 NRC's *Framework*.

Given the magnitude of the change needed across multiple aspects of the science education environment, policymakers would do well to adopt a "bottom up" orientation toward assessment systems development—that is, one grounded in the classroom—rather than grounded in needs for monitoring, accountability, and teacher evaluation ("top down"). 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 strategies can lay the groundwork for subsequently developing assessments for monitoring and accountability.

One follow-on implication is that continued use of large-scale science assessments that states developed under NCLB is neither appropriate nor advisable. These instruments are not aligned to college- and career-ready science standards and thus will not support the desired changes in teaching and learning.

Interim solutions will be needed that can simultaneously satisfy federally mandated testing requirements and allow space for change in classroom practice. 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 narrowly reflect only the material that is tested.[8]

The Road Ahead

It is worth noting that there is very limited evidence that accountability policies driven by large-scale assessment have led to improved student achievement.[9] In contrast, the positive relationship between classroom assessment and student learning outcomes is well established.[10] Assessment that closely aligns with curriculum and instruction and that engages students in the kinds of science learning described in NRC's *Framework* will return the focus to what is most important—the direct support of students' learning. ■

If they are to implement the new standards effectively, many jurisdictions will need to reintroduce science in the early grades.

[1]W. S. Schmidt, Curtis C. McKnight, and S. Raizen, *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education* (Boston, MA: Kluwer Academic Publishers, 1997), p. 62.
[2]National Research Council, *A Framework for K-12 Science Education Standards: Practices, Crosscutting Concepts, and Core Ideas* (Washington, DC: National Academies Press, 2012).
[3]Achieve, Next Generation Science Standards (Washington, DC, 2013), http://www.nextgenscience.org/.
[4]See James Pellegrino, Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment* (Washington, DC: National Academies Press, 2001).
[4]See, e.g., National Research Council, *Assessment in Support of Learning and Instruction: Bridging the Gap between Large-Scale and Classroom Assessment* (Washington, DC: National Academies Press, 2003); M.R. Wilson and M.W. Bertenthal, eds., *Systems for State Science Assessments* (Washington DC: National Academies Press, 2006); J. W. Pellegrino, M. Wilson, J. Koenig, J., and A. Beatty, eds., *Developing Assessments for the Next Generation Science Standards* (Washington, DC: National Academies Press, 2014).
[6]L. M. McDonnell, *Politics, Persuasion, and Educational Testing* (Cambridge, MA: Harvard University Press, 2004).
[7]National Research Council, *State Assessment Systems: Exploring Best Practices and Innovations: Summary of Two Workshops* (Washington, DC: National Academies Press, 2010).
[8]D. Koretz, "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, 2005); D. Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2009).
[9]National Research Council, *Incentives and Test-Based Accountability in Education*, M. Hout and S.W. Elliott, eds. (Washington, DC: The National Academies Press, 2011).
[10]P. Black and D. Wiliam, "Assessment and Classroom Learning," *Assessment in Education* 5, no. 1 (1998): 7–73; N. Kingston and B. Nash, "Formative Assessment: A Meta-Analysis and a Call for Research," *Educational Measurement: Issues and Practice* 30, no. 4 (2011), 28–37; National Research Council, *Taking Science to School: Learning and Teaching Science in Grade K-8,* in R. A. Duschl, H. A. Schweingruber, and A.W. Shouse, eds. (Washington, DC: National Academies Press, 2007).