TRANSFORMING SCIENCE ASSESSMENT: CHALLENGES AND RECOMMENDATIONS FOR STATES

This report serves as an introduction to a series of briefs intended to support state and district efforts to implement high-quality science assessments designed for three-dimensional standards.

Since 2013, 39 states and the District of Columbia have adopted the Next Generation Science Standards (NGSS) or similar standards based on the National Research Council's *A Framework for K-12 Science Education*, signaling a commitment to high-quality and rigorous science education for all students. States' previous science standards[1] took traditional approaches to science, emphasizing students "knowing" disconnected science facts and decontextualized, procedural skills. In contrast to standards that emphasized one aspect of science at a time, states' new science standards are three-dimensional, and integrate disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs) into performance expectations that require students to demonstrate knowledge-in-use as they make sense of real-world phenomena and solve authentic problems. This brief describes some key challenges associated with developing assessments for these new standards, and recommendations for states to consider.

Three-Dimensional Learning and Performance

A hallmark of standards developed based on *A Framework for K-12 Science Education* is three-dimensionality—that each standard comprises aspects of three distinct but complementary strands of science education:

- ***Science and engineering practices (SEPs)*** are what students DO to make sense of phenomena. SEPs are comprised of the skills and behaviors scientists use to make sense of phenomena and address problems as well as the knowledge about what the SEPs are, and how and when to use them. Importantly, while skills and knowledge about the SEPs are important, students need to use them as ways to sense-make and problem solve to truly engage in SEPs. This is in contrast to simply using skills to carry out a procedure.
- ***Crosscutting concepts (CCCs)*** are concepts that hold true across the natural and engineered world. Students can use them to make connections across seemingly disparate disciplines or situations, connect new learning to prior experiences, and more deeply engage with material across the other dimensions. The NGSS require that students use their understanding of the CCCs to make sense of phenomena or solve problems.
- ***Disciplinary core ideas (DCIs)*** are the fundamental ideas that are necessary for understanding a given science discipline. The core ideas all have broad importance within or across science or engineering disciplines, provide a key tool for understanding or investigating complex ideas and solving problems, relate to societal or personal concerns, and can be taught over multiple grade levels at progressive levels of depth and complexity.

Three-dimensional standards are unique because each standard is written as a performance expectation that combines a SEP, CCC, and DCI; subsequently, student proficiency is tied to students being able to use the three dimensions together to make sense of phenomena and solve problems. Phenomena and problems are defined as:

- ***Phenomena*** are observable events that students can use science to explain.
- ***Problems*** are situations that humans wish to change.

Why make this shift?

Standards based on the *Framework* emphasize goals for teaching and learning that are intentionally designed—built on decades of research and expertise—to ensure that **all students** leave their K-12 education experiences ready to be scientifically literate, be critical consumers of information, and able to pursue the full range of opportunities available to them in an increasingly STEM-oriented world. They signal and incentivize teaching and learning that not only *recognizes* that students with diverse backgrounds all deserve a high-quality science education, but *leverages* that diversity to support rigorous, meaningful science experiences that lead to better outcomes for all students.

[1] See the Innovations of the Next Generation Science Standards for more information.

New Standards Need New Assessments

Any time state content standards undergo such a significant shift, states must develop new assessments to measure student progress toward meeting the new standards. Statewide summative assessments designed for the NGSS and similar three-dimensional science standards should measure student performance in ways that help inform, incentivize, and monitor progress of schools and districts as they implement new standards. Validly monitoring progress relative to three-dimensional standards requires assessments that allow all students to demonstrate science proficiency by:

1. Authentically and directly engaging in reasoning about phenomena and problems, and
2. Doing so by integrating the three interacting dimensions of science that each have associated knowledge and application expectations.

Measuring application over rote knowledge, and three dimensions instead of one, requires innovative approaches to alignment and assessment design. If new science assessments are done well, they have the opportunity to shape better teaching and learning in science, but over time can influence assessments in other content areas as well.

Developing New Science Assessments Is Challenging

Under the best circumstances, designing new large-scale assessments that measure fundamentally different standards is hard, and the current environment—both within and outside of the state education agency (SEA)—is particularly challenging. Those within SEAs responsible for developing new science assessments face significant barriers in the form of limited resources for assessment development, an evolving understanding of alignment, and inconsistent policies that shape the timelines, expectations, and quality of new assessments. Some prominent challenges include:

Higher Expectations and Fewer Resources for Assessment Development

High-quality and aligned three-dimensional statewide summative science assessments are costlier to develop and score than previous tests for a number of reasons. They require a significant, if not complete, redesign of both process and products. This includes:

- New assessment designs, specifications, and task formats/approaches;
- Research and new documentation to support and validate decisions about how to assess three-dimensional student performance;
- New approaches to item/task design that include the development of multi-dimensional tasks that involve multiple parts and multiple ways (visual models, written scenarios, simulations, constructed/extended response) for students to engage with the tasks;
- Professional learning for developers, including content leads, psychometricians, and item writers;
- Ongoing and rigorous quality review throughout the development process; and
- Maintenance and improvement plans for future years of administration.

Each of these components requires money and human capital. This is compounded by the reality that three-dimensional standards and assessment are new—developers (researchers, educators, and vendors alike) are still figuring out how to translate complex standards into effective assessment targets. This means that not every attempt at item specifications or tasks themselves will be successful, and assessment processes have to be ready not only to create tasks that are more involved than previous science assessment items, but also to create more of them, to account for losing a larger proportion of tasks due to quality and alignment concerns during development than in previous assessment development efforts.

SEAs are tasked with developing these assessments with limited funding allocated for assessments. Limited budgets mean SEAs have to make difficult decisions about where to make tradeoffs, often resulting in lower quality assessments that fail to live up to the promise of new science standards—and ultimately, provide less useful information for parents, students, educators, and policymakers.

Limited Flexibility and Unreasonable Timelines for Assessment Transitions

Several states are pursuing very short timelines between beginning assessment development efforts and having an operational assessment. In most states, this is driven by a combination of two major factors:

1. Federal- and state-level policies, such as federal requirements for states to report on science achievement each year (including comparable individual student scores for the tested grades), and limited assessment budgets that impact the states' ability to develop a new assessment while another assessment is being administered or to develop of transitional assessments; and
2. Decisions made within the SEA based on targeted implementation timelines and values regarding the relationship between assessments and local transitions to new standards (e.g., state leadership that takes the perspective that new assessments will make classrooms transition faster often leads to science assessments that are developed very quickly; conversely, state leadership perspectives that hold that new assessments should not be in place until teachers and students have had sufficient opportunity to transition teaching and learning pursue assessments more slowly, limiting the common monitoring information available for all students).

While faster timelines for developing three-dimensional assessments have important benefits that should be weighed, many states pursue new assessments on unreasonable timelines that do not allow for sufficient expertise to be built among those developing the assessment (e.g., item writers), and careful and intentional design needed to produce high-quality tests with meaningful results. For most states, this means advocating to extend the time between standards adoption and the administration of an operational test; for others, this means investing in assessment development earlier, leaving more time for planning, design, item development, piloting, evolution over time, and quality review throughout the process. In both cases, this means concerted SEA effort to direct support toward other factors that can directly influence classroom progress, such as local assessments that are closer to the classroom, high-quality instructional materials designed for three-dimensional standards, and effective professional learning for all educators.

Opportunity to Learn Considerations

The NGSS and similar standards establish fundamentally different expectations for what and how students should learn science—and districts and schools are still figuring out how to define and operationalize these expectations. Given the state of science teaching and learning in schools, coupled with the current scarcity of high-quality instructional materials and ongoing professional learning designed for three-dimensional standards, it is extremely unlikely that students taking the first generation of new assessments will receive sufficient opportunity to learn to be successful on these tests. Without clear investment in collecting and operationalizing data about opportunity to learn as part of the assessment design and reporting process, states and test developers are in a bind: do they develop the high-quality (expensive, and time-consuming) tasks and tests that are appropriately designed for three-dimensional standards that few students have experienced in their classrooms, or do they design tests that reflect students' current science experiences but do not fully align to new standards?

Insufficient State Capacity for Creating or Procuring Three-Dimensional Assessments

A high-quality, aligned NGSS assessment requires that states, test developers, and item writers have a clear and common understanding of what it means for a student to demonstrate proficiency on the performance expectations that maintains the integrity of the standards and is consistent with how the state is supporting teaching and learning. Regardless of whether the state's role is primarily in developing or procuring the assessment, the overall direction and quality expectations need to be driven by the state; however many SEAs do not have sufficient internal capacity to meet this need. State assessment departments frequently lack internal content expertise and have limited access to in-house expertise housed within teaching and learning divisions. When science teaching and learning expertise is engaged in assessment development processes, there is often a single person responsible for providing science content expertise for all science instructional, professional learning, assessment, and policy activities coordinated or led by the state, all of which require more support in light of new, three-dimensional standards. This net result is insufficient expertise to guide a state's science assessment work, and lack of coherence throughout.

Unprepared Vendors

Given limited SEA capacity in science content expertise in assessment development, many states depend more heavily on their vendors to produce high-quality and aligned assessments. The success of this approach hinges on vendor expertise in three-dimensional science teaching, learning, and performance, and a commitment to developing new approaches, designs, and tasks for three-dimensional assessments. Most vendors have limited three-dimensional science expertise; their science content experts are often well-versed in traditional approaches to science teaching and learning, but not deeply involved in work around new standards, which is a quickly evolving field. While vendors may recruit NGSS experts in advisory roles, the majority of the individuals engaged in the design, psychometrics, and item writing have at best a superficial understanding of three-dimensional standards and their implementation. This impacts every step of the process and leads to low-quality items that comprise ineffective assessments that do not embody the standards.

Inconsistencies in State Policies for Science Education

The NGSS and similar standards specify learning goals for K-12 science that include substantial science at every grade in elementary and middle school, at least three years of science in high school, and the assumption that science learning defined in previous grades is necessary to be successful in later ones. When states adopt these standards, they are making a statement to all stakeholders in their states: all standards are required for students to be considered "college and career ready" in science. However, there are major inconsistencies in the other signals provided by the state, including:

1. Disconnect between standards and graduation requirements. While most states with three-dimensional standards have three years of high school science across four domains in their standards, their graduation requirements do not reflect this expectation—few, if any, states require that all students take courses that would cover all standards. This forces those responsible for the single high school assessment to decide what expectations assessments should be designed to meet, with difficult tradeoffs associated with each decision. For example, some states have decided to shift from an end-of-course exam to a comprehensive high school test because of the domains included in high school science in the NGSS and similar standards. However, most students who are taking these tests will not have had instruction in all domains, presenting both interpretation and design challenges. Conversely, some states are planning to continue current practice and administer a single-domain end-of-course assessment (e.g., high school biology) as the high school assessment, but are facing major pushback from stakeholders because the assessment does not reflect the breadth of state standards.
2. Challenges measuring depth and breadth of the standards. The Every Student Succeeds Act continues to require that science assessments be administered at least once per grade band while mathematics and English language arts (ELA) assessments are administered annually in grades 3-8. Accordingly, most states administer science assessments once per grade band, but do so without clarifying how expectations shift for assessments administered less frequently. This means that science tests face pressure to cover more content (both across grade bands as well as across dimensions) in less time than those assessments administered annually. This would have been a challenge with any set of content standards, but is particularly difficult given the depth and breadth of standards like the NGSS. This is not to suggest that more testing equals better testing, but it often leads to assessments that are trying to do too much, and failing to meet those expectations.
3. Limited role in accountability. Stakeholders often take tests, and the associated subject matter, seriously based on what those results are going to mean for school, district, and state evaluations. Science assessments play a limited role in accountability systems relative to mathematics and ELA—and as a result, science is generally not high on decisionmakers' lists of priorities for time and resource allocation.

Recommendations for States

While standards implementation efforts are underway, educators, students, and parents are largely in the dark about how students—and the programs being implemented to support them—are progressing toward new expectations because state science assessments have not yet caught up. For better or worse, what is tested gets taught; statewide assessments shape what is happening in classrooms by providing signals about what kinds of student performance are valued, what it means to be proficient, and whether students are meeting those expectations.

As states develop and administer statewide science assessments, it is important that they ensure that the context within which these development efforts happen is setting students up for success. State assessments need to be as reflective and informative of student expectations and progress as possible to support effective classroom, school, and district practices. While some existing policies may make assessment efforts in science more difficult, there are several steps SEAs can take to support better science assessments. These include:

1. **Collaborate and partner to increase capacity.** Limited budgets are often immutable realities for most SEAs, and while those charged with developing new science assessments should advocate for the resources they need to create effective assessments, they should also explore other ways of increasing capacity. This can include:
   - Collaborating within the SEA to bring science expertise from teaching and learning divisions to both support the content of the assessment as well as ensure the assessment connects to standards implementation timelines, instructional strategies, and messaging regarding new expectations;
   - Engaging a range of decisionmakers within the state in strategic parts of the assessment development process to increase leadership buy-in;
   - Leveraging expertise and resources that are available through within-state partners, including higher education, informal education, and state science and STEM organizations; and
   - Connecting with cross-state and national partners to share assessment resources designed for similar standards.

2. **Invest in systems of assessment to cover a wider range of depth and breadth.** Three-dimensional assessments have a lot to contend with. States should consider investing in systems of assessment—either state-led or locally-led and state-supported—to distribute the expected scope and use of any given assessment. Assessment systems can provide stakeholders with more information on a range of timescales, and can also provide the opportunity to consider innovative designs and approaches in lower stakes environments.

3. **Consider smart summative assessment designs.** No single NGSS statewide summative assessment—designed within current and reasonable constraints on timing and cost—is going to be able to cover the full depth and breadth of the targeted grade-band standards. Given this, states should consider innovative approaches to summative assessment design, such as specific claims that can guide assessment design that reflect state values and priorities, a variety of targets at different grain sizes (e.g., a topic bundle of standards as opposed to individual standards), the inclusion of a range of task types, partial matrix sampling, and compartmentalizing assessments such that different segments of the assessment contribute different kinds of information for varying purposes.

4. **Focus on quality.** Within reasonable breadth targets, states should prioritize high-quality processes and products related to their assessments. This should include:
   - Anchoring expectations in high-quality criteria. States should push to ensure that their design, documentation, tasks, and reports are consistent with the expert-developed criteria for statewide summative science assessments. This can include incorporating relevant components of the criteria into requests for proposal (RFPs) for assessment, building item and documentation development processes based on the expectations laid out in the criteria, and using the criteria to support internal and external quality review processes.
   - Developing and using high-quality tasks. The assessment tasks students engage with should be grounded in rich and meaningful scenarios that are driven by problems and phenomena, and they should elicit student thinking and reasoning via the science and engineering practices, disciplinary core ideas, and crosscutting concepts. Ensuring that these tasks are of high quality and truly embody the standards should be a focal point of assessment development processes, with careful and intentional design, appropriate research and cognitive labs, and regular quality control embedded.

5. **Invest in professional learning for those developing the assessments.** High-quality tasks and assessments will be easier to develop if those developing them are deeply knowledgeable about the *Framework for K-12 Science Education*, the NGSS, and assessment design. This overlap in expertise is hard to find, but can be cultivated, and states that invest in this early are likely to see effective returns on their investment. This can and should include professional learning for test developers, and states should consider including this expectation as part of RFPs and contracts with vendors.

6. **Intentionally consider opportunity to learn.** While opportunity to learn considerations have not always been a priority for traditional science assessments, they are particularly important to address during this transition period. States should embed data collection and analysis of opportunity to learn factors (e.g., educator and student surveys, monitoring of instructional materials being used and time devoted to science instruction, classroom observations) into their assessment plans, and this should happen early so that this information can be used when interpreting pilot and field test results and making needed modifications for operational tests.

7. **Follow timelines that allow for the high-quality assessments that are worth students' time.** All of these efforts require that states allow sufficient time in their assessment development processes to engage in rigorous and iterative design, evaluation, and refinement.

8. **Value science in transparent and internally aligned ways.** States should make efforts to align their policies to drive better assessments, and ultimately better science education, for their students. As an interim step, states should be transparent about what the assessment will target and what that means for students—both in terms of what experiences they should be getting and how assessment results should be interpreted and used. Since it will take some time for states to science tests of reasonably high quality, states might consider phasing in any accountability stakes as the quality improves.

9. **Invest in building capacity for high quality 3D instruction locally.** Even the best statewide summative assessments can only do so much; they can tell stakeholders whether students are meeting new and rigorous standards, but play a limited role in providing mechanisms for continuous improvement. States should consider investing in those efforts that will help make continuous improvement in the classroom a reality, such as local assessments (as part of a system), high-quality professional learning and networked improvement communities, and effective instructional materials. Absent those investments, high-quality assessments alone won't get the job done.

State Examples

Across the country, states are implementing these recommendations to improve science outcomes for their students. Some examples include:

- ***13 states*** are partnering with researchers through the ACESSE (Advancing Coherent and Equitable Systems of Science Education) Project to develop resources to support instructional, professional learning, and assessment systems that can be used across all states and districts implementing three-dimensional standards.
- ***27 states*** are considering using performance assessments as part of innovative assessment system designs through the State Performance Assessment Learning Collaborative (SPA-LC).
- ***Kentucky, Delaware, and Nebraska*** are implementing innovative state-led assessment systems that leverage different kinds of assessment (instructionally-embedded, periodic through-course or external assessments, and statewide summative assessments) to reduce the scope of any one assessment while providing stakeholders with coherent and more nuanced feedback across the full range of science expectations.
- ***California*** is pursuing *an innovative three-part summative assessment design* that balances individual student reports for 5th grade science standards with comprehensively assessing the 3-5 grade-banded expectations.
- ***Washington*** included the Criteria for Procuring and Evaluating Statewide Summative Assessments in Science as part of their alignment study RFP to ensure that their statewide summative assessments are aligned to three-dimensional standards and the intent of the Framework.
- ***Louisiana*** conducted a state review of science instructional materials and is incentivizing schools and districts to use the best materials in their classrooms.