Supporting Coherence Across a System of Assessment for NGSS

Jonathan Osborne and Jill Wertheim

Graduate School of Education, Stanford University

Graduate School of Education

Stanford University

Palo Alto, CA

osbornej@stanford.edu

jwerthei@stanford.edu

Paper to be presented at the annual conference of the American Educational Research Association

Toronto, Ontario

April 5-9, 2019

Abstract

The NGSS are based on a new vision of science learning for U.S. states, and successful implementation of these standards requires a substantial change in how student learning is monitored and evaluated. Indeed, a balanced system of assessment is integral to the success of many high performing education systems around the world. The Stanford NGSS Assessment Project (SNAP) identified three key points of leverage critical to a balanced system that can support a state's implementation of NGSS: the assessment system design, the design of the assessments with particular attention to performance assessments, and capacity-building for using the performance assessments.

SNAP developed resources for each of these three parts of the system including a model balanced assessmet system for NGSS, exemplar three-dimensional performance assessments, and professional development tools to guide the development and use of performance assessments for three-dimensional learning. Each support was designed to model the shifts needed to implement the standards, engage stakeholders in discussing and planning changes needed for those shifts, and to guide stakeholders across a state in developing their own balanced assessment system for NGSS.

The resources were used to work with stakeholders in California, one of the earliest states to implement the NGSS. California adopted a system of assessment that includes a type of performance assessment, though the system falls short of some of the key requirements of a balanced system. The effect of including complex tasks in the statewide summative assessment, however, has sent a clear signal that these tasks are valued, which has led to

widespread local adoption of classroom performance assessments and the use of SNAP's professional learning resources.

The three sets of resources proved to be integral to framing conversations with stakeholders in California. Yet this work also exposed additional areas of need, including professional learning that targets the local education agencies and school leaders, a platform for teachers to share provide feedback on their performance assessments, and modifying grading practices for compatibility with a new assessment system. There are promising signs that other states are developing systems of assessment that will include performance assessments, and the common standards

Supporting Coherence Across a System of Assessment for NGSS

Over the last few decades, assessment systems in the US have depended largely on statewide summative exams that have shown mixed results in increasing student achievement (Center for Education Policy, 2007; Dee and Jacob, 2010, 2011; Lee and Reeves, 2012). These exams have had numerous unintended impacts of narrowing of the curriculum at the expense of science, social studies, and the arts (Center for Education Policy, 2007; Dee and Jacob, 2010; Dee, Jacob & Schwartz, 2013), and increasing some achievement gaps (Center for Education Policy, 2007; Darling-Hammond & Adamson, 2010). Other negative impacts of accountability testing — universal across all subjects — is the overvaluing of knowledge that can be tested easily using a multiple-choice format at the expense of conceptual understanding and writing skills (Jennings & Bearak, 2014). In a recently released report by a 30-member commission of experts on assessment, they pointed to a need for a new vision of how assessments are used in K-12 education:

> "Considerable concern has been expressed in the Commission about the artificiality of 'stand-alone' or 'Drop-in from the Sky' tests. Perhaps more problematic than the isolated character of these examinations is concern with the tendency to treat the data from these tests as independent and sole sources of information concerning the performance and status of students. Some Commissioners argued for the greater use of systems of examinations distributed over time embedded in the ongoing teaching and learning of experiences. It is recommended that assessment in education move

progressively toward the development and use of *diversified assessment systems*[1] for the generation and collection of educational assessment data."

- Gordon Commission Report To Assess, To Teach, To Learn: A Vision for the Future of Assessment, Technical Report Executive Summary, 2013 (p. 24)

In the six years since this report outlined the concerns of assessment professionals, accountability systems, driven by federal policy, remain reliant on single measures from tests developed and delivered by the states. But studies of the highest-achieving education systems in the world (e.g., Hong Kong, Victoria, Singapore, Finland) find that these systems dedicate time and resources primarily to monitoring learning at the classroom level. Linda Darling-Hammond (2010) summarized the following common elements across these assessment systems that are considered integral to the effectiveness of the systems as a whole:

1) common standards that tightly align assessment, curriculum, and teacher development;
2) a balanced assessment system that includes challenging, authentic tasks;
3) involving teachers in development and scoring of assessments (as well as curriculum);
4) use of assessments to guide continuous feedback to teachers, students, curriculum developers, and administrators;
5) timely reporting of results.

Efforts to create balanced systems of assessments that include many of these critical elements grew through the 1980s and 90s (e.g, Connecticut, Kentucky, California), but the strict federal requirements for accountability and reporting imposed by the No Child Left Behind Act (NCLB) in 2001 led states to focus development efforts on the statewide high stakes summative tests

[1] Emphasis added

that would be used for sanctions on schools if insufficient progress was made in individual students' achievement on the state science standards. Changes to accountability requirements in the 2015 Every Student Succeeds Act (ESSA), successor to NCLB, along with lessons learned from numerous studies of the successes of balanced assessment systems around the world (Darling-Hammond & Adamson, 2010; Darling-Hammond, 2017), have created an opportunity to build systems around new common standards that enact these best practices.

California adopted the NGSS in 2013 and planned an ambitious timeline for developing an operational system for 2019 (http://www.cde.ca.gov/pd/ca/sc/ngsstimeline.asp). This timeline placed California in the position of being a leader in NGSS assessment by making it the first state to implement a system of assessment for the ambitious three-dimensional science standards. Not surprisingly, other states were closely observing the decisions California made, and the approaches it has taken to developing an assessment system.

The Stanford NGSS Assessment Project (SNAP) saw California's role as a leader in NGSS implementation as an opportunity to provide leadership on the design of a vertically coherent system around the common elements outlined above. SNAP focused this work in three areas:

1) outlining a balanced assessment system that includes the use of vertically-aligned performance assessments at the state and local level and involves teachers in the development and scoring of the local assessments;
2) developing model performance assessments that illustrate a vision of challenging, authentic tasks developed around the NGSS standards and that offer continuous feedback to teachers and students; and

3) capacity-building activities to support teachers in developing and using performance assessments to make instructional decisions.

The intent of this work was not to develop and implement an entire assessment system, but to focus on key points of leverage within the system: state policymakers and teachers. The policymakers establish the vision, guidelines, and funding priorities for the assessment system. The teachers have the most direct impact on students, so they determine the degree to which the system is effective. The three focii were chosen for their central roles in implementation of the best practices of high-achieving education systems (Darling-Hammond et al., 2013). The resources were designed around the goal of stakeholder engagement; thus each of the three sets of resources were designed to stimulate dialogue among stakeholders across the system, from policymakers to classroom teachers, about how their practices must evolve to implement such a system.

I. A model balanced assessment system for NGSS

The Stanford NGSS Assessment Project (SNAP) model assessment system for NGSS is designed with the goals of balancing the data required to monitor the education system for state and federal policy with the timely and informative data needed to support districts, schools, and teachers as they enhance students' progress toward the new standards (Osborne et al., 2015). The system is also designed around the goal of establishing coherence across each of these elements of the system, as they signal to teachers, students, and administrators the learning outcomes that are valued by policymakers, while at the same time they offer policymakers insight into what is being learned in classrooms (NRC, 2001). In order for assessments to play

this role effectively, however, they must represent what is valued, and must effectively elicit evidence of the achievement of those goals.

The NGSS are based on a new vision of the goals for science learning. To implement these standards successfully they require a substantial change in how learning is monitored and evaluated (Gorin & Mislevy, 2013). In particular, they represent a major shift from testing knowledge and understanding to measuring performance or competence – a reflection of a shift which is happening globally (Koeppen et al, 2008, Rychent and Salganik, 2003). The system SNAP designed to operationalize this vision balances centralized and local assessments, a model that is widely considered essential for monitoring and supporting learning (Darling-Hammond and Adamson, 2010). It is has four components, and uses a combination of short items that contribute to individual student scores, matrix-sampled short performance assessments that contribute to system monitoring during federal testing years, and short and extended performance assessments in classrooms that could be used as often as every year (Table 1).

Grade	**Part 1: External Mandated Tests**		**Part 2: Periodic Classroom Assessments**	
	Component A: Multi-item types • Variety of item types including selected and constructed response • Computer-scored	**Component B: Performance Tasks** • Two short performance tasks • Scored by trained group of teachers	**Component C: Stand-alone Performance Tasks** • Shorter • Optional • State-developed • Teacher-scored • Use is reported, scores are not but may be used by districts	**Component D: Instructionally Embedded Assessment (IEA)** • Longer • Task bank, state curated and controlled • Teacher-scored • Use & scores are reported
1st – 4th			(x)	(x)
5th	x	x	(x)	(x)
6th – 7th			(x)	(x)
8th	x	x	(x)	(x)
9th – 10th			(x)	(x)
11th	x	x	(x)	(x)
12th				

Table 1. Proposed System of Assessment for California Science Assessment (from Osborne et al., 2015). X denotes testing years; (x) denotes recommended testing years.

Component A consists of computer-scored items that are primarily in a selected-response format, with some constructed-response and technology-enhanced items. This component of the system accommodates federal requirements for reporting science scores once in each grade band. The highly-constrained selected formats in Component A are required to meet standards of validity, generalizability, and scaling needed to report individual student scores (Linn & Herman, 1997; Gorin and Mislevy, 2013), and for the time and cost of scoring data for each student (Linn et all, 1991). Component B is designed to be composed of short (20 minute) performance tasks that would evaluate students' progress, using the three dimensions to engage in sustained reasoning with evidence about a phenomenon. This component would elicit evidence of the type of scientific thinking that is the goal of NGSS but is not easily assessed in Component A. This component would not contribute to students' individual scores, but would provide school or district-level scores about more sophisticated reasoning skills to the state. Equally important, this component of the assessment emphasizes that the kind of learning that can be assessed with multiple-choice items is not the *only* kind that is valued by the state.

Components C and D potentially take place as often as each year, and are administered in the classroom. These tasks are included in the system with the intention of providing responsive assessments to teachers that are benchmarked across the district (Component C), and are closely tied to the enacted curriculum (Component D) so they can help drive the annual, incremental changes in instruction necessary to monitor and support NGSS learning (Coffey, 2011; Marion et al., 2018). There are many ways to handle these interim assessments. One is to "loosely-couple" the interim assessments with statewide tests such that they are tied

to the same learning goals and communicate the same vision of learning, but the classroom tasks do not contribute to the state accountability scores (Marion et al., 2018). Districts or states may document completion of the interim assessments, they may collect them and involve teachers in scoring them (E.g. Wyoming Body of Evidence), or they may choose to provide teachers with the tools and training to calibrate themselves in order to evaluate their own students (e.g. New Hampshire's PACE pilot model). The most important outcomes of these components of the assessment system is that teachers, administrators, and students are experiencing tasks that represent the goals of NGSS each year – that is, complex, engaging tasks that communicate the vision of learning underpinning NGSS and provide students, teachers, and administrators with actionable information about students' progress toward that vision.

Each component of this assessment system plays a critical role in maintaining coherence between policy and instruction. Component A supports federal policies for reporting science scores for individual students in grades 5, 8, and in high school. Component B provides policymakers with additional data about progress of schools or districts using performance assessments that are designed to collect information about students' progress through tasks that more closely represent the multidimensional reasoning that is the goal of NGSS. Components C and D provide teachers with tasks that ensure that students are being monitored for progress in each grade — not just the federally-tested years — and are using tasks that well represent the vision of learning embedded in the standards. These components may have elements that are reported to the district for monitoring purposes, but the primary goals are to ensure that the NGSS is being addressed each year (not just in the federally-tested years), to provide assessments that will give teachers and students the information they need

to ensure that science learning is meeting the state's learning goals, and to offer feedback that they can use to address any areas of need.

The SNAP model system for NGSS assessment was not necessarily intended to be adopted and implemented as described, but was designed to initiate a conversation among policymakers, assessment developers, district leaders, and teachers about the changes needed at each level of the education system to implement the new standards effectively. The model system can be used to drive these changes. It can help stakeholders recognize that NGSS requires different types of information about science learning and at different frequencies from previous standards – for example, that assessment systems will need additional information about student performance beyond selected-response items. To ensure that these conversations are grounded in concrete terms, however, the model system needed to be combined with illustrative examples of assessments.

II. Model assessments for a balanced system

The NGSS defines outcomes in terms of what students know and can do, not just what they know and understand. As such it requires a move to a competency-based system of assessment which is better suited to the demands of a schooling system which increasingly demands higher-order competencies of its populace (Baker, 2014). Moreover, the challenge for any assessment system that will support the vision of learning underpinning the NGSS must measure competency-based performances in not one but three dimensions (NRC, 2014).

The second stage of supporting a coherent use of assessment was to illustrate what assessments that represent the vision of NGSS look like. Indeed, if assessments do not closely

reflect the changes in expectations for science teaching and learning, they cannot act as a lever to drive those changes (Sunal & Wright, 2006; Au, 2007). Assessments that illustrate the vision of the standards are critical tools for communication about goals between stakeholders in a system. if policymakers, administrators, district leaders, and teachers have different conceptions of the vision of science learning they are working toward, the system will lack the necessary horizontal coherence across curriculum, instruction, and assessment (NRC, 2006).

SNAP created and piloted a bank of model multidimensional performance assessments, scoring rubrics, and sample student data to be used as resources for communication. These tools anchored conversations with policymakers about systemwide goals for NGSS learning. They were also used by district-level leaders in professional development activities about instructional shifts for NGSS, and by teachers to learn how to make decisions about instruction and assessment in 3-dimensions.

SNAP's assessment development process draws on a sociocultural learning framework (Lave & Wegener, 1991) in which students are engaged in tasks that present students with a complex and real question or problem that has no single right or wrong answer, and place students in the role of a scientist who must solve the problem using real science resources. In the case of extended performance assessments, students work in groups to become experts in their 'scientist' role, and in the course of the task multiple expert 'scientist' groups discuss their findings from the perspective of their position. Once students have had opportunities to discuss the problem in small and large groups, solicit peer and teacher feedback, and make revisions, they complete an individual product. The design of the performance tasks shifts the intent of assessment away from tests that identify what students do and don't know, and toward tasks

that provide sufficient opportunity to understand a complex problem by drawing on individual and community sources of knowledge, such that students are prepared to provide the best possible evidence of their progress.

The set of sample assessments were designed to model all four components of the hypothetical assessment system, A-D. Samples were developed from each subject area (earth and space science, life science, physical science, and engineering) and span across grades K-8. Assessment development followed SCALE's design principles for high-quality performance assessment that Darling-Hammond et al. (2013) modified to address NGSS-specific design goals (NRC, 2014):

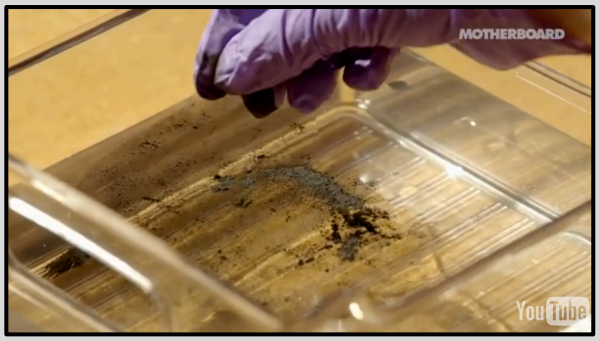
1) Tasks engage students in exploring a question or problem about a real-world phenomenon;
2) Evidence of students' use of all three dimensions of a performance expectation(s) to answer a question or solve a problem is elicited over the course of the task;
3) Scenarios are accessible to all students, such that it is clear to all students why the question or problem is meaningful, important, and scientifically relevant;
4) Prompts and scoring guides evaluate students' progress with evidence-based reasoning from novice to expert across two- (and if appropriate, three-) dimensional constructs;
5) Constructs pair the DCI with multiple different elements of the science and engineering practices throughout the task;
6) Tasks highlight an aspect of the performance expectation or NGSS in general that represents significant shifts from previous standards.

These assessments were designed to model how assessments can support foundational goals for NGSS, such as how to use assessments to move away from evaluating disconnected facts and concepts, and toward evaluating the strength of students' reasoning about phenomena (NRC, 2012). The sample performance assessments exemplify these shifts so that stakeholders across the system can consider the implications for how to prepare for them. In response to these assessments, for example, elementary teachers described concerns about NGSS rooted in their weaknesses in science expertise, and the time it would take away from their other responsibilities. When teachers saw sample elementary performance tasks (e.g., Fig 1), they reported that NGSS expectations seemed very attainable because of the way informational texts, computational thinking, and evidence-based reasoning can be integrated into sensemaking tasks about phenomena. They saw these activities as complementary and compatible with their math and ELA requirements.

In the 3rd grade Physical Science task shown below, students are shown a video of a scientist who is investigating solutions for cleaning oil spills in the ocean. The video shows an experiment he is doing in his lab using magnets and iron filings to collect some oil from water. Students respond to a series of prompts that increase in sophistication through the task and prepare students to propose how the experiment will need to be modified to work on real oil spills in the ocean.

Your teacher will show you a video of a scientist, Dr. Warner, doing an experiment. He is testing the **research question: Can magnets be used to collect oil in water?** He designed an experiment with these steps:

1. Places water in a large plastic tub.
2. Pours oil into the water.
3. Puts black magnetic powder on the oil.
4. Places a large magnet on the side of the plastic tub.



Dr. Warner puts the black powder in a tub with water and oil (left). Then he holds a magnet outside the tub to pull the oil and powder toward it (right).

Students individually answer a series of questions based on this scenario. For example, Questions 2&3 below provide evidence of their progress with the bold parts of these two dimensions.

- **Ask questions that can be investigated based on patterns such as cause and effect relationships.**
- Electric, and **magnetic forces between a pair of objects do not require that the objects be in contact. The sizes of the forces in each situation depend on the properties of the objects and their distances apart** and, for forces between two magnets, on their orientation relative to each other.

Question 2. Write a research question that he could investigate to find out how he can use magnets to collect *more* oil.

Question 3. Use what you know about magnets to explain how your question will help Dr. Warner investigate if magnets can be used to collect even more oil.

Figure 1. An excerpt from one of SNAP's short performance tasks for 3rd grade physical science. Find the rest of the task, scoring guides, and student data at https://snapgse.stanford.edu/snap-assessments/short-performance-assessments.

Exemplar NGSS assessments help stakeholders envision the ways science education will need to change to support NGSS learning. For example, state leaders in Nebraska and Oklahoma found that studying one of SNAP's model assessments was central to their processes for engaging stakeholders in planning and development of assessments for NGSS. As state leadership recognizes the gaps between sample performance assessments that allow students to reason about phenomena, and the highly-constrained tasks that vendors develop for their computer-based assessment platforms, they are beginning to coordinate teams across states to develop classroom-based tasks that are better able to meet those goals. Samples of high-quality assessments, however, are not sufficient to support the broad changes in instruction and assessment needed to implement NGSS. Educators and administrators need to know how to use them to support three-dimensional instruction.

III. Capacity-building

An assumption implicit in many assessment systems is that teachers will know how to use the assessment data to inform their instructional practices. But few teachers in the US are trained in effective use of assessment data, though Professional Development (PD) focused on collabortive analysis of student responses has been shown to have some of the strongest effects on student outcomes (Kazemi & Franke, 2004; Gearhart et al., 2006; Heller et al., 2012). In each example of education systems that have used classroom assessments as one of the pillars for improving student outcomes (e.g., Queensland, Australia; Ontario, Canada), there is a commitment to professional development that prepares teachers to develop and use these assessments (NRC, 2003). If assessment is to support implementation of NGSS, all teachers that

are part of this implementation must have the expertise they need to develop assessments that collect and analyze evidence of students' progress, and know how to use this evidence to provide constructive feedback and determine instructional moves. Yet even in states that have built substantial capacity among select teachers in developing and scoring classroom assessments for NGSS, such as Kentucky, few teachers outside these leaders have training on how to integrate these assessments into their instructional practice.

Indeed, a monumental challenge for any state that is considering developing a vertically coherent, high-quality assessment system for science is building capacity for *all* science teachers across their state. SNAP addressed this challenge by drawing on emerging best practices in professional development (Wei et al., 2010), including employing collaborative learning, focusing on activities that are common to teachers, and giving teachers an opportunity to apply, analyze, and evaluate new content in relation to their own students. Research on PD also emphasizes the importance of providing opportunities for practice and immediate feedback.

Few districts offer dedicated time for teachers to collaborate on professional development activities, and the need for PD to be ongoing makes high-quality learning opportunities about performance assessment even more rare. Online PD is one viable solution due to its flexibility to extend over a period of months, increased opportunity for discussion, low cost, and broad accessibility (Dede et al., 2009). Online PD can be completed asynchronously, so teachers can engage in professional learning without interfering with instructional time or other PD priorities for the school or district.

Online PD also addresses the issue of scale. In-person PD is expensive, time-consuming, and rarely meets the need of PD to be ongoing, with frequent feedback as teachers implement the new material. Online PD can reach across an entire state preparing to implement new science standards, giving many teachers immediate access to high-quality professional learning. However, online learning also has additional requirements if it is to be effective, including collaboration, and the need for multiple opportunities for practice (Ronaghi, Saberi, & Trumbore, 2014).

A variation of Massive Open Online Courses (MOOCs) that combine online and in-person learning, called hybrid MOOCs, capitalize on the extended reach, flexibility, and low cost of online learning, with the benefits of collaborative learning among colleagues from in-person activities. Hybrid online courses can be delivered at any time and any pace across a state, making this format particularly appealing for rapid, large-scale implementation of the new science standards. SNAP hybrid MOOCs blend video-based instruction with in-person sessions in which colleagues meet to collaborate on applying the skills they are learning, and discuss how they might adapt and adopt them into their practice. These courses were designed to enable participants to asynchronously learn about 3D performance assessment, and to begin applying their knowledge at their own pace. Valuable in-person time is dedicated to PD practices that are deepened through discourse and community, including analysis of student work and peer-to-peer feedback.

Over 1500 participants have taken the courses since Fall 2017, and the courses are continually revised and updated based on participant feedback. The current iteration of the first course (Course 1) has 319 active participants (413 registered) and will close in June, 2019.

Eighty-five participants who have finished Course 1 and 2 have submitted complete evaluation surveys (74%). Most respondents (94%) said that the course not only changed their ideas about assessment for NGSS, but also their ideas about instruction. Furthermore, 86% of respondents described specific ways that they are making changes to how they will teach science based on this course: “I plan to spend time redesigning all my activities and units,” and “Feedback will be conducted in a variety of ways and will be a focus of my instruction.” Participants in Course 2 describe outcomes of the course that go beyond learning to develop a performance assessment: “This collaborative process has informed how we are now approaching other curriculum projects as well for other grade levels.”

The utility of the courses in influencing participants’ ideas about NGSS assessment and instruction likely derives from two areas of weakness in existing assessment systems. First, despite the critical role that teachers play in using assessment data to improve student outcomes, professional development for science rarely focuses on assessment development, and even more rarely focuses on how analysis of student work can inform instructional practice. Second, traditional approaches to professional development require quantities of time and money that are not feasible for most states to use to reach all teachers, coaches, and district leaders, so the online courses that provide a structure for flexible local use have the ability to reach far more people more quickly.

The hybrid online format models the central role of collaborative learning in the use of performance assessment to support 3D instruction. Colleagues analyze student data together and discuss strategies for feedback and revision, modeling ways that professional learning communities can use assessment data to anchor decisions about curriculum and instructional

practices. Moreover, in their reflections, administrators who participated in the professional learning communities report developing an appreciation of what it means to implement NGSS. There are drawbacks to hybrid online learning, including limited access to experts and few opportunities for feedback on the quality of the participants' work. But since the growing interest in developing high-quality assessments for NGSS is not balanced with state-level commitments to providing the professional learning to ensure that teachers are prepared to play a central role in these systems, free hybrid online courses can fill a critical gap by structuring local learning and development around performance assessment for NGSS.

The California System

The Stanford NGSS Assessment Project developed a model assessment system and exemplar assessments with the goal of initiating communication across stakeholders about a balanced assessment system for California. In 2016 California committed to adopting a more sophisticated science test than in previous years – one that included a short performance assessment as part of the statewide summative exam. The form of performance assessment that was adopted is computer-based, and differs from the form developed by SNAP, but it does represent a step forward in moving toward two-dimensional competence-based measures of student performance.

Due to cost constraints, the state did not include classroom performance assessments as part of the accountability system. Yet the inclusion of a performance assessment (albeit a limited one) as part of the statewide summative test has sent a clear signal to districts, administrators, and teachers that complex tasks for science are a priority. This message has

driven demand for such tasks at the local level. Schools and districts are seeking performance tasks and professional learning opportunities to help teachers incorporate these tasks into instruction. In response to surveys about SNAP online courses, coaches report taking the course on NGSS performance assessment, for example, "to know how to best support my teachers in delivering and assessing performance of mastery of NGSS," and in some cases because districts are developing classroom tasks for their own monitoring system complementary to the state's system.

The effect of SNAP's three-tiered approach to supporting system change in California resulted in limited changes in the statewide assessment. Funding priorities for statewide assessment remain focused on tests used for federal reporting, and without changes to federal policy few states will be able to dedicate significant funding to classroom performance tasks. But the impacts of this work at the local level are building momentum. Schools and districts are finding information from multiple-choice tests insufficient for supporting their implementation of NGSS and are using their own resources to build and use performance assessments.

The demand for assessments that support teaching and learning extends beyond California. So far, 28 states have joined a coalition of states that are interested in building performance assessments into their assessment systems for science, called the State Performance Assessment Learning Community (SPA-LC). These states are considering a wide range of approaches to systems of assessment, but many are exploring the use of classroom performance assessments that are developed locally by district-led groups of teachers. The growing interest around broadening science assessment systems beyond on-demand selected-response tests at the state and local level must be seen as a positive outcome. These efforts

include many of the elements of high-quality assessment systems described by Linda Darling-Hammond (2010) including alignment of curriculum, assessment, and professional learning around common goals, complex tasks, involvement of teachers, and a focus on timely and rich feedback to teachers and students.

Recommendations to states considering performance assessment for science

The resources that SNAP developed to support each of three "high leverage points" identified in California were integral to driving the use of performance assessment for NGSS across the state. The model assessment system framed discussions with policymakers about the importance of performance assessment as part of a science assessment system; the exemplar assessments communicated a vision of high-quality assessment and served as models for schools and districts; and the assessments and assessment toolkits formed the foundation for professional development about how to use performance assessment. Combined, these resources reinforced the need for collaboration among teachers, and broadened teachers' and administrators' ideas about assessment as a tool to inform teaching and learning. Their value is evident in the early progress seen across California in adopting performance assessments and the widespread use of the SNAP assessments, courses, and toolkits. The use of these resources, however, has revealed additional areas of need that should be addressed by any state that wants to ensure that performance assessments can be used to support NGSS teaching and learning.

1. **Professional learning for local education leaders.** Most state leaders acknowledge the important role high-quality classroom tasks have in supporting and monitoring learning for NGSS, but in most cases development and implementation of the tasks will fall to regional and local education units. Some have initiated district-wide efforts to develop and use performance assessments for NGSS, but most districts have not planned any changes to their use of assessments in science. The result is that there is very uneven use of performance assessments in districts, and even in schools. Small groups of teachers scattered across the states choose to learn to use performance assessments. Many of these teachers describe challenges with administrators and other teachers in their school who are not willing to dedicate time and resources to using performance assessments to engage in collaborative analysis and decision making about instruction. To help these early adopters get the support they need, and to encourage others to join them, district and school leaders need professional learning to help them understand the role they can play in driving development and use of performance assessments for science. These leaders are responsible for creating the opportunities for teachers to collaborate in the analysis and discussions about evidence of students' progress, as well as to use this information to identify areas of need across subjects and grades, and plan instructional moves. Without commitment from leadership, teachers are unable to find time or willing colleagues to engage in this collaborative work, limiting the use and effectiveness of the assessments.
2. **A platform for collaboration**. Many teachers are developing their own performance assessments using SNAP's tools, the Council of State Science Supervisors' ACCESSE

project, and other resources. A frequent request from groups involved in this work is for a central place to submit, review, and retrieve these resources. A central bank of resources-in-development would enable these teachers to share with and learn from their colleagues. This platform would allow those that are leading the efforts to use performance assessment to show others what they have done. These examples can be used to build interest among groups that are more reluctant to begin. Prior task banks have shown, however, that at least one person needs to be responsible for adding resources and communicating to users. Without a dedicated caretaker, these banks rarely fulfill their promise.

3. **Address the scoring vs grading dilemma**. Systems of assessment are in an intermediate stage: numerous policymakers acknowledge the importance of performance assessment in an effective science education system, but many of the structural changes to the system that are required to use performance assessments have not been made. For example, in many schools teachers are expected to record letter or numerical grades weekly, or even multiple times a week for their students. Teachers under these constraints feel unable to dedicate time to assessments that will not provide grades. But grades are not entirely compatible with performance assessments. Assessments designed *for* learning provide insight for teachers and students into the students' progress with the dimensions being assessed. If the student is to learn about how they are moving toward proficiency and how they might continue to progress, they need to get feedback that highlights both of these features of their work. Grades obscure this rich information by reducing feedback to a single indicator of achievement. Teachers in

schools that require frequent grades often find that they need to either grade the performance assessments instead of providing descriptive feedback, or provide an additional "traditional" assessment that they can easily grade, which creates an untenable amount of work for themselves and students. Select districts have begun making changes to their evaluation system such as converting to standards-based grading. This system records students' proficiency toward a specific objective for a course (such as a performance expectation), instead of a general achievement score on a test. Teachers in districts that have made this change find that it is much more compatible with performance assessments because scoring translates more easily to standards-based evaluation and because the focus on feedback enables the use of efficient scoring techniques (e.g. evaluating only the most informative questions in a task). Teachers routinely describe the need to grade their students as a barrier to their ability to focus their efforts on providing effective and efficient feedback to students from performance assessments. Communication to school and district leaders about the effects of grading policies on teachers' ability to commit to using performance assessments could begin removing one of the most persistant barriers to the use of performance assessments.

Beyond California

The three elements of the Stanford NGSS Assessment Project's work were designed to engage stakeholders in California, but they are quickly being adopted by other states. The common standards have enabled many other states to take advantage of the resources as part of their own version of a balanced assessment system. In fact, the local education agencies, such as

district offices, are in many cases leading the efforts to develop assessments that fit into SNAP's Components C and D, and they are using the model assessments and capacity-building resources to frame and align discussions about NGSS assessment across networks of policy-makers, administrators, teachers, and Professional Development providers. State and district leaders who plan to leverage these resources in developing a balanced science assessment system need to follow some initial steps to lay the groundwork for their use:

1. The SNAP model assessment system offers one option for a balanced assessment system, but leaders need to set a vision for the system that they will adopt. Decisions about how performance assessments will be integrated into school, or how the data from the performance assessments will be used, set crucial parameters for the specifications of the assessments that will be developed.
2. Leaders need to make decisions about their priorities for information they want to be able to get from these assessments. This information will determine how the teachers write task specifications (for example, decisions about the length and frequency of performance assessments, depth or breadth of each task, etc) and how they evaluate students' opportunities to learn prior to the assessment.
3. Leaders need to communicate regularly with key stakeholders (district and school leaders, coaches, and teachers). Frequent staff turnover can lead to gaps in knowledge and awareness of resources available to inform assessment development evenly across a schoo, district, or state.

4. Leaders need to create regular opportunities for teachers to collaborate so that they all teachers able to enact effective practices around performance assessment development and use, not just the particularly motivated early adopters.

We are at a critical juncture for assessment systems. Educators and leaders are recognizing that assessment systems that utilize only multiple choice and technology-enhanced items are insufficient for assessing science learning. Instead, multiple measures are needed, and performance assessments can play an important role in supporting and monitoring implementation of the NGSS. SNAP's resources were developed to advance the dialogue in California around building a vertically-coherent and balanced assessment system, but the interest in developing such systems for NGSS outside of California has led to the use of these materials in many more states (and countries). This renewed focus on performance assessment is promising. Yet there is still a tremendous amount of work to be done in creating balanced assessment systems, from creating the policy and systemic structures that pave the way for the effective use of performance assessments, to training teachers across entire districts or states to use them, not just select early adopters.

References

Au, W. (2007). High Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, 36*(5), 258-267.

Baker, D. P. (2014). *The Schooled Society*. Stanford: Stanford University Press.

Center on Education Policy. (2007). Has student achievement increased since 2002? State test score trends through 2006–07

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, *48*(10), 1109–1136.

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, *40*(3), 291–309.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Palo Alto, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., … Hakuta, K. (2013). Criteria for high-quality assessment. *Stanford Center for Opportunity Policy in Education (Online). Https://Edpolicy. Stanford. Edu/Sites/Default/Files/Publications/Criteria-Higher-Quality-Assessment_2. Pdf [24 Jan 2017]*.

Dee, T. S., Jacob, B. A., Hoxby, C. M., & Ladd, H. F. (2010). The impact of No Child Left Behind on students, teachers, and schools [with Comments and Discussion]. *Brookings Papers on Economic Activity*, 149–207.

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

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., … Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K–12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment*, *11*(3–4), 237–263.

Gitomer, D. H., & Duschl, R. A. (2007). Establishing Multilevel Coherence in Assessment. In *Yearbook of the National Society for the Study of Education* (Vol. 106, pp. 288–320).

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the next generation science standards for both formative and summative assessment*. Presented at the Invitational Research Symposium on Science Assessment, Princeton, NJ.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333–362.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389.

Kazemi, E., & Franke, M. L. (2004). Teacher learning in mathematics: Using student work to promote collective inquiry. *Journal of Mathematics Teacher Education*, *7*(3), 203–235.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Journal of Psychology, 216*(2), 61-73.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation* (Vol. 521423740). Cambridge university press Cambridge.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209–231.

Linn, R. L., & Herman, J. L. (1997). *Standards-led assessment: Technical and policy issues in measuring school and student progress*. National Center for Research on Evaluation, Standards, and Student Testing ….

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21.

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, D. (2018). A TRICKY BALANCE: THE CHALLENGES AND OPPORTUNITIES OF BALANCED SYSTEMS OF ASSESSMENT. Presented at the RILS, NH.

National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academy Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. (J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty, Eds.). Washington, DC: National Academies Press.

National Research Council. (2015). *Guide to Implementing the Next Generation Science Standards*. Washington, DC.

Osborne, J., Quinn, H., Pecheone, R., Schultz., S., Holthuis, N., Wertheim, J. (2015). *A System of Assessment for the Next Generation Science Standards in California: A Discussion Document.* Palo Alto, California.

Ronaghi, F., Saberi, A., & Trumbore, A. (2014). NovoEd, A social learning enviornment. In P. Kim (Ed.), Massive Open Online Courses: The MOOC Revolution (pp. 96-105). New York, NY: Routledge.

Rychen, D. S., & Salganik, L. H. (Eds.). (2003). *Definition and Selection of Key competencies: Executive Summary*. Göttingen, Germany: Hogrefe.

Sunal, D., W., & Wright, E. (Eds.). (2006). *The Impact of State and National Standards on K-12 Science Teaching*. Greenwich, Connecticut: Information Age Publishing.

Wei, R. C., Darling-Hammond, L., & Adamson, F. (2010). *Professional development in the United States: Trends and challenges* (Vol. 28). Dallas, TX: National Staff Development Council.

Wilson, M. (2005). *Constructing Measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.