

An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS

Jill Wertheim Stanford NGSS Assessment Project
Presented at National Association for Research in Science Teaching
Baltimore, 2016

Introduction. The assessment system presented in Paper 1 (and in greater detail in Osborne et al., 2015) would require the development of banks of short answer and short performance tasks for all performance expectations in the NGSS, as well as the curriculum-embedded performance tasks aligned to the broader goals of each subject. Examples of such tasks are critical for signaling the kinds of changes required for implementations of the new standards across the education system. Assessments operationalize the standards and instantiate the expected performances in a way that helps to communicate what competency would look like under the new standards. Examples also can be used to make explicit some of the fundamental shifts that are implicit in the new standards in a manner that no other means can do.

The importance of example assessments of all types and across all the goals of the NGSS, combined with the immediate need for these examples, led SNAP to perform an analysis of the landscape of existing assessments that hold promise for serving as models for NGSS. It was anticipated that there would be few existing assessments that are fully aligned to NGSS, but the purpose of this analysis was to find assessments that are aligned at least in part, and could be used as inspiration for developing a new, fully-aligned assessment. For example, few existing science assessments focus on the practice of data analysis and each of its sub-practices (such as summarizing data using descriptive statistics), but successful approaches to probing this practice can be drawn from statistics assessments (E.g., modelingdata.org) and can be extremely useful in providing ideas for effective ways to probe this practice in a scientific context. Therefore, this project focused on finding model assessments that probe at least one dimension of the NGSS and that span the assessment formats that are part of our proposed assessment system.

The goals of this project were threefold: 1) to characterize the landscape of promising existing assessments, 2) to evaluate the strengths and weaknesses of those assessments for the NGSS, and 3) to identify model assessments that can be used to guide the development of new, NGSS-aligned items.

Methods. SNAP created a task bank of promising assessments by conducting a wide-ranging search drawing on their own knowledge and that of a set of experienced advisors. The task bank includes 203 assessment resources, and each resource contains between one and several hundred assessment items (such as the NAEP item bank). The range of formats of the tasks included in this bank reflect the needs of the proposed assessment system: short response, short performance tasks, and curriculum-embedded performance tasks, and they are designed for paper-based or computer-based test platforms.

The contents of the task bank were summarized quantitatively (Part I) and qualitatively (Part II):

Part I. A sample of roughly 400 tasks from 51 assessment resources drawn from the task bank was evaluated using the *NGSS Assessment Review Criteria*¹ to identify the areas of NGSS that have numerous models and to identify areas where there are gaps in the landscape of existing assessments.

¹ See snapgse.stanford.edu for a full list of the criteria

These review criteria characterize existing assessment resources in terms of:

Characteristics: where they fit in the assessment system (grade, subject, format, timescale, etc.);

Alignment: how the tasks they contain are and are not aligned to the NGSS and other prioritized qualities (e.g., accessibility); and

Evaluation: whether they have tasks that hold promise as models for NGSS assessments.

A small selection of the findings from this analysis are reported here; a detailed discussion of the methods and results can be found on snapgse.stanford.org.

Part II. The entire task bank was used to evaluate general trends in the ways most promising existing assessments do or do not match the goals of NGSS. These trends, and exemplar items, are used to illustrate the changes in approach to item development needed for the new generation of assessments.

Part I: summary of findings from the task bank analysis

With the exception of Science and Engineering Practices *asking questions and defining problems* and *obtaining, evaluating, and communicating information*, the analysis identified numerous existing resources from the task bank that could serve as models for ways the other 6 practices could be probed (Figure 1). For example, the review identified 19 assessment resources (out of 51) that contain tasks aligned to elements of the practice *developing and using models*. These tasks use a variety of formats, from multiple-choice to performance tasks, different content areas, and platforms (e.g. computer-based and paper/pencil), offering diverse examples of ways to probe this practice. When it comes to assessing students' ability to ask questions there were few examples and little diversity among those examples. It is unclear if this a reflection of the lack of imagination and creativity of item writers for how this practice could be assessed, placement of this practice as a low priority for assessment, or a failure to define what is meant by this practice in terms of performance expectations that can be operationalized as assessments.

There were far fewer examples of ways in which each cross-cutting concept has been assessed (Figure 2), although the assessment of *cause and effect: mechanism and explanation* can be found in these assessments about twice as frequently as the other concepts. This finding is unsurprising given that the cross-cutting concepts bring us into uncharted territory that create a mismatch between the NGSS and nearly all existing assessments.

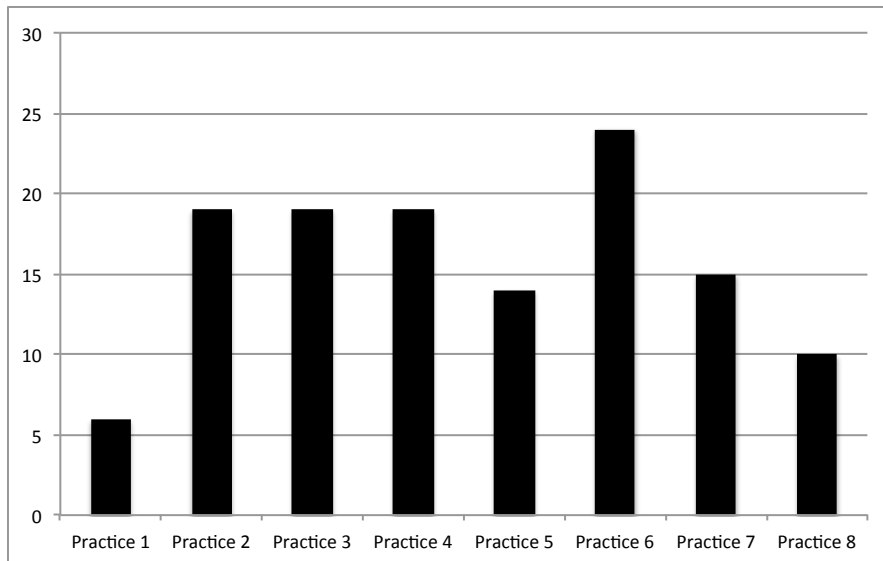


Figure 1. Bar graph showing the number of assessments in the task bank that have at least one task that targets a science and engineering practice. Practice 1: asking questions and defining problems; Practice 2: developing and using models; Practice 3: planning and carrying out investigations; Practice 4: analyzing and interpreting data; Practice 5: using mathematical and computational thinking; Practice 6: constructing explanations and designing solutions; Practice 7: engaging in argument from evidence; Practice 8: obtaining, evaluating, and communicating information. (N=51)

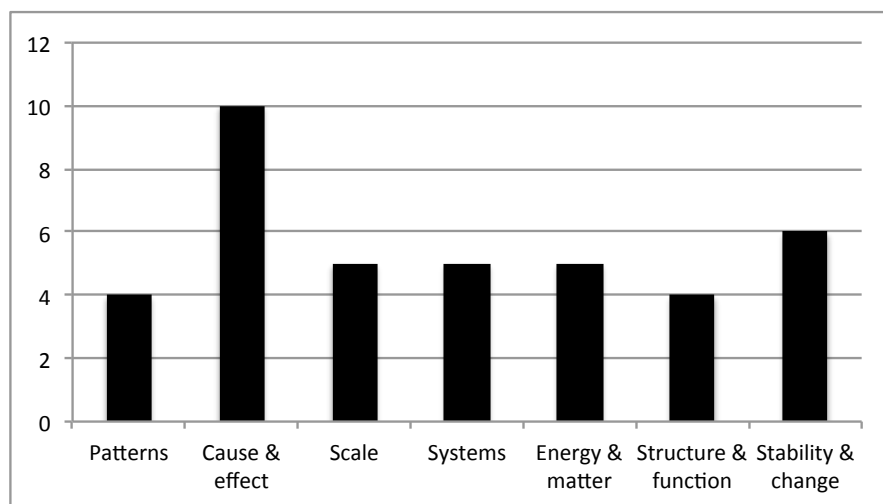


Figure 2. Bar graph showing the number of assessments in the SNAP task bank that have at least one task that aligns to a cross-cutting concept. (N=51)

The separate assessment of content and practices was typically a core design principle for the previous generation assessments. Yet the integration of the dimensions is a primary goal for assessments aligned to the NGSS. Therefore, identifying tasks that model integration of more than one dimension of the NGSS was a high priority for the task bank. Seventy-eight percent (40) of the sampled assessments contained tasks that probe more than one dimension. Out of those 40 assessments, 63% (25) had at least one task that was rated as “strongly integrated,” meaning that students must draw on a DCI to engage in a practice or cross-cutting concept. In contrast, weakly integrated tasks might use a context relevant to a DCI where knowledge of the content would be

helpful but not necessary (see Figure 3). Un-integrated tasks were composed of individual or clusters of items, each of which probed one dimension but no single item was aligned to more than one dimension.

Although assessments in the task bank are mostly paper/pencil-based (70%) and half of those reviewed were in multiple-choice and short-answer formats, the analysis identified a few sources of tasks that utilize less common formats and provide a variety of examples of approaches to probing the three dimensions of NGSS. The Connecticut Department of Education, for example, has been a leader among several states in providing student performance tasks as tools for teaching and learning. A single task draws on multiple science and engineering practices as they are needed to answer questions and solve problems, and performance tasks are able to tap some of the practices rarely found in other formats, such as *asking questions and defining problems*. Some of these examples and a discussion of the ways in which they might serve as models to guide the development of new tasks, are presented in Part II.

Part II. Landscape analysis: implications for NGSS item development

The landscape analysis brought into focus four critical areas where existing assessments fall short of the needs for assessments for the NGSS. Still, these gaps are not complete and the analysis did identify some tasks that model ways of addressing the gaps. In some cases, finding these models required looking beyond the task bank and doing targeted searches of the research literature, but the important message is that there are existing tasks that can be immensely useful resources for guiding the design of new tasks. The following section presents each of these four areas illustrated by example tasks, contrasts the examples with models of what an assessment that better fits the needs of NGSS might look like, and discusses the implications for significant changes in the development of new assessments for NGSS.

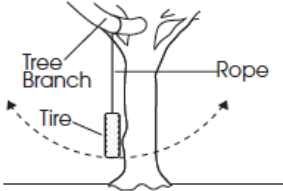
1. Assessment design must integrate multiple dimensions: most existing tasks are aligned to a single dimension; even tasks that appear to tap both content and science practices often require only one of those dimensions to provide a correct response. Indeed, many tasks that probe science practices present data about a scientific phenomenon in a table or graph, but the data can be analyzed or interpreted without the use of any knowledge of a DCI or cross-cutting concept.

Although the task in Figure 3 (Ohio, Grade 8) presents physical science content about the motion of a pendulum in the context of an investigation, the task is not multidimensional. A student could identify a pattern on the data table without knowing about the phenomenon being investigated. Moreover, a student who knew the physical science concept being probed could answer the question without the data table, which would make it single dimensional but drawing only on content knowledge.

There are some instances in which items developed for NGSS might align to a single dimension, such as some items used for formative assessment where a teacher might want to monitor his or her students' knowledge without being conflated by evaluating other dimensions. But where development of assessments that probe just a single dimension was a core design principle for most item development prior to NGSS, multiple dimensions must be blended to evaluate the three-dimensional thinking that is fundamental to the new standards. There exist some models of two-dimensional tasks, but very few models of three-dimensional tasks exist for any subject or grade.

Use the information below to answer question 5.

5. A class investigating the motion of a tire swing collected the data in the table below. The students were able to draw conclusions about the factors that affect the motion of a swing. Two students from the class decide to use the class data to build a different-size tire swing in their backyard. They build the tire swing shown in the diagram.



Tire Swing

Tire Swing Investigation Data			
Swing	Length of Rope (meters)	Mass of Tire (kilograms)	Time it Takes for the Tire Swing to Move Back and Forth Once (seconds)
1	2	10	2.8
2	2	20	2.8
3	4	10	4.0
4	4	20	4.0

After testing the swing, they decide that they want to make it swing faster.

Based on the data from the class investigation, what could the students do to make their tire swing move back and forth faster?

- A. use a shorter rope
- B. use a longer rope
- C. use a less massive tire
- D. use a more massive tire

Figure 3. A test question from the Ohio Grade 8 state science exam.
 Reproduced from education.ohio.gov.

2. Assessment design must focus on the big ideas in science: the NGSS emphasize the big ideas and themes in science; fine details are encouraged only as they are required to making sense of the big ideas. But fine details, often discrete facts, are generally much easier to assess than students' deep conceptual understanding. So it is unsurprising that many existing tasks evaluate these facts, and therefore, evaluate knowledge and skills that are not directly targeted in the new standards.

Many existing tasks probe students' content knowledge that is so narrow in scope that it would be difficult or impossible to infer what the student thought about the broader core principles. For example, a task from TIMSS (2011) (Figure 4) asks 8th grade students about the function of a specific part of the reproductive system for a mammal.

The uterus (womb) is part of the reproductive system in mammals.
 Name one function of the uterus.

Figure 4. An open-response task from TIMSS, 2011.
 Reproduced from nces.ed.gov/timss.

The shift in expectations away from students learning isolated pieces of knowledge toward assembling a coherent view of science is built into the fabric of the *Framework*. To illustrate this change, consider the middle school DCI statement in NGSS relevant to this targeted content for the item in Figure 4:

In multicellular organisms, the body is a system of multiple interacting subsystems. These subsystems are groups of cells that work together to form tissues and organs that are specialized for particular body functions. (LS1)

And the corresponding Performance Expectation:

MS-LS1-3 Use argument supported by evidence for how the body is a system of interacting subsystems composed of groups of cells.

The focus on big ideas in science means that even much of the content in NGSS has a different emphasis than the content in previous standards. The change is so profound that many topics have no strong models for ways to probe the content. Indeed, the misalignment between the new disciplinary core ideas and tasks aligned to previous content standards explored above using Figure 4 is common across all content areas in NGSS.

3. Assessment design must incorporate the full range of science and engineering practices: The NGSS present the practices of science and engineering differently from previous standards. This means that there are some entire practices in the NGSS that were not in previous standards, and therefore there are few existing assessments that target them. Moreover, only narrow segments of practices that were seen in prior standards were assessed, leaving some aspects of those practices with abundant examples of promising approaches for assessment and some aspects with virtually none.

Some practices are defined differently in the NGSS compared with previous standards, but other practices in the NGSS are effectively new to standards, such as developing and using models, and engaging in argument from evidence. These practices have been explored in the research literature (e.g., Kuhn, 1993; Grosslight et al., 1991; Sandoval and Reiser, 2004; Sampson and Clark, 2008; Schwartz et al., 2009), but have rarely been assessed beyond that realm. Only 19% of US state and national tests examined in our landscape study had any items related to modeling, and 13% for argumentation.

An example of a task that deeply explores a foundational concept is shown in Figure 5. This excerpt of a task is from the *Assessment of Argumentation in Science* (scientificargumentation.stanford.edu). In this task, students clarify an argument about what makes bubbles in boiling water. Students also combine content with elements of argumentation by articulating the underlying reasoning of an argument and constructing a counter-argument using evidence.

Although tasks that are aligned to the content goals described in the NGSS do exist, the amount of content goals for which we found no good models, and the number of promising assessments in our task bank that were not well aligned to the new content goals, were some of the most alarming findings of the study.

Brian and Joe are looking at the water boiling in the pan on the stove.



Brian says that the bubbles are made of air that gets pushed out of the water when the water gets hot. He argues that he knows there is air dissolved in water because fish are able to breathe the oxygen in the water.

Joe says that the bubbles are made of water that has turned into a gas -- water vapor.

Joe agrees with Brian that fish are able to breathe oxygen in the water. But the pan has been boiling for 10 minutes and it is still bubbling just as much as it was at the beginning. If Brian was right, wouldn't the air be gone by now?

What idea is Joe arguing for? _____

What is the reason Joe gives to convince Brian he is right?

- a. Fish are able to breathe the oxygen in the water.
- b. Bubbles are made of air.
- c. The pan has been boiling for 10 minutes and it is still bubbling.
- d. Hot water boils

Brian says that he knows that water is made of hydrogen and oxygen. The bubbles are caused by the water breaking down to produce hydrogen and oxygen that are both gases. These form bubbles like the gas in soda.

Joe is unconvinced. He remembers observing that the saucepan lid became covered in water drops as the water continued to boil.

How could he use this observation to convince Brian that he's wrong? _____

Figure 5. An excerpt of a task probing students' use of their knowledge of states of matter to engage in argumentation. Reproduced and adapted with permission from Assessment of Argumentation in Science (Scientificargumentation.edu).

4. Assessment design must include a variety of task formats: A primary goal in building the SNAP task bank was to include as much variety in task formats as we could find. Yet 43% of the assessments reviewed contained multiple-choice items. Most short-response tasks probe a very narrow scope of content or a practice, and though they can offer important psychometric properties that enable generalizability (Yen, 1993), they provide limited insight into what students think and what they can do. Well-designed performance tasks can probe much more deeply into students' reasoning and their ability to draw on knowledge and skills as they are needed to investigate questions and solve problems. Potential solutions to challenges around time and reliable scoring systems for performance tasks are being tested and provide compelling evidence for the feasibility of wider use of this format (Darling-Hammond & Adamson, 2010).

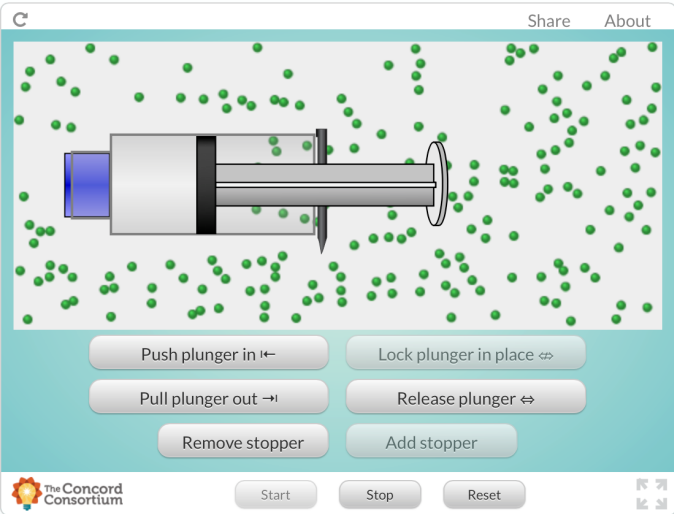
Some computer-based assessments offer a very different approach to performance tasks than those seen in paper/pencil tasks. One example is from the Concord Consortium's [Interactions](#)² project, which embeds simulations into activities that blend learning about the behavior of atoms and molecules with assessment questions that ask students to use evidence from the simulations to perform scientific practices and explain macroscopic phenomena. In this task students explore a phenomenon on their own through the simulation, then follow guided steps to help them draw from their observations and their knowledge of the behavior of atoms and molecules to provide evidence-based reasoning about the phenomenon. Performance tasks can be brief, as in Figure 6, or can be projects that take days or weeks. Their ability to probe how students build coherent ideas about phenomena by drawing on multiple practices, and in some cases, areas of content, is needed to engage students in the kind of reasoning that is the goal of NGSS. While some promising models of curriculum-embedded performance tasks can be found far more are needed, and there exist few models of short performance tasks of the type that would become a substantial part of the proposed assessment system described in Paper 1.

This simulation assumes that gases are made of tiny particles. Set up the model in various ways, simulating what you just did with the real syringe, to see how well a particle model might explain your observations.

Question #11

Revisit your initial model of a gas (the first question of this activity). Do the components of your initial model explain your observations of gas being compressed in the syringe? If not, what revisions would you make to your model?

Type answer here



Question #12

Write a scientific explanation that answers the question, *How is it possible to compress a given amount of air into a smaller space?* In your explanation be sure to include the following:

- Claim - your answer to the question
- Evidence - observations or data
- Reasoning - thinking that includes ideas the class has agreed on and connects your evidence to your claim

Figure 6. An excerpt from an Interactions module in which students explore what happens to the molecules that make up a gas the gas is compressed.

Reproduced with permission from: <http://mw.concord.org/nextgen/>.

² Interactions is a collaboration between the Concord Consortium, the CREATE for STEM Institute at Michigan State University, and the University of Michigan.

Please note that the sources that provided these items do not necessarily share the views expressed in this paper.

References

- Achieve, Inc. (2013). *Next Generation Science Standards*. Achieve, Inc.
- Darling-Hammond, L., & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. *Stanford Center for Opportunity Policy in Education (SCOPE), Stanford University, School of Education*. Retrieved from <http://edpolicy.stanford.edu>.
- Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, 28(9), 799–822.
- Hannaway, J., & Hamilton, L. (2008). Performance-based accountability policies: Implications for school and classroom practices. *Washington: Urban Institute and RAND Corporation*.
- Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, 77(3), 319–337.
- Osborne, J.; Pecheone, R.; Quinn, H.; Holthuis, N.; Schultz, S.; Wertheim, J.; and Martin, M. (2015). A System of Assessment for the Next Generation Science Standards in California: A Discussion Document. Retrieved from snapgse.stanford.edu December 1, 2015.
- Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, 88(3), 345–372.
- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, 92(3), 447–472.
- Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., ... Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, 46(6), 632–654.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, 30(3), 187-213.