

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

Prepared by the Stanford NGSS Assessment Project Team (SNAP)
January, 2016.

Jill Wertheim, Jonathan Osborne, Helen Quinn, Ray Pecheone,
Susan Schultz, Nicole Holthuis, and Paolo Martin

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

Introduction	3
Background	4
Synopsis of the landscape review: characteristics of promising existing assessments	9
Gap Analysis: implications for the development of assessments for the NGSS .	14
1. Alignment and integration of the three dimensions	14
2. Focus on big ideas in science.....	21
3. Probe the full breadth of science and engineering practices	26
4. Assessments for the NGSS require the use of a variety of formats.....	32
Summary and conclusions.....	42
Item sources	43
References	44

Introduction

The fundamental change in priorities for K-12 science education initiated by the release of the *Framework for K-12 Science Education* (NRC, 2012) and the *Next Generation Science Standards* (NGSS) (Achieve, 2013) raises the question of what kinds of evidence students can produce to demonstrate that they have met the new expectations, and how that evidence could be elicited. With the new standards, mastery of science concepts cannot be demonstrated just by a display of the correct knowledge of scientific concepts. Instead, the new standards require the ability to reason about phenomena using scientific evidence, draw on general principles to solve problems, and to reflect on common themes that underpin big scientific ideas. In particular, the outcomes of students' learning are expressed as a set of expectations of the kinds of performances or competencies that students will be expected to display. Decades of development of assessments designed to evaluate content knowledge separately from science skills or practices has left teachers, education administrators, and assessment developers with few resources suitable for use in this new paradigm (Hannaway & Hamilton, 2008, NRC, 2011, 2014; Pellegrino, 2013).

Preparation for California's adoption of NGSS has underscored the need for an assessment toolbox that can assist teachers in helping students to cultivate the kinds of performances established by the new standards—standards that are a synthesis of a set of disciplinary core ideas, eight scientific practices, and seven cross-cutting concepts. Such a toolbox will require formative and summative resources for classroom assessment aligned to all these three dimensions across each topic and grade band. Developers of large-scale assessments also need to be given well-defined specifications for the external summative assessments they need to develop. The development of these assessments can, however, build upon existing resources that, to varying degrees, assess performance on one or more of these three dimensions. The Stanford NGSS Assessment Project (SNAP) team¹ has conducted a review of existing assessments to identify which aspects of NGSS have robust examples of assessment tasks and which aspects need more attention. The review also points to the specific characteristics of each assessment that align to NGSS and the characteristics that are not well matched to the new standards. The outcomes of this work show how new assessments will need to differ and identifies existing assessments that could be used as models for development of new assessments that are fully aligned to the performance expectations of NGSS.

The 2014 NRC report, *Developing Assessments for the Next Generation Science Standards*, stresses that “assessments that are now in wide use were not designed to meet this vision of science proficiency and cannot readily be retrofitted to do so” (pg. 12). While it is true that most assessments do not demand the kind of reasoning with evidence or the range of science practices required by the NGSS, there are assessments that target closely related goals, at least in part. We argue that the challenge of creating assessments necessary to implement NGSS demands that we make use of existing tasks. In short, we must build on what exists already. Indeed, many of the assessment challenges posed by implementing the NGSS have been tackled previously in some form or another. Hence, we

¹ SNAP Team members: Jonathan Osborne (PI), Ray Pecheone, Helen Quinn, Susan Schultz, Nicole Holthuis, Jill Wertheim, Paolo Martin

have conducted a wide-ranging search of what exists with the goal of collecting examples of tasks that address each of the three dimensions of the NGSS, even if the tasks are not completely aligned to NGSS.

Several common characteristics of existing assessment tasks have emerged from our analysis of the landscape of existing assessments. Our observations of these trends illustrate the fundamental changes required to create tasks fully aligned to the new standards. This report describes those characteristics that are needed to support the vision of science learning described in the *Framework* and those of current assessment resources, and deconstructs existing tasks to illustrate (1) common ways that existing science assessments fall short of the new goals, and (2) promising approaches to meeting the goals of the NGSS.

Background

How the NGSS differ from previous standards and implications for assessment

Standards set common goals for K-12 learning across a state, and the way the standards present the learning goals lays a foundation upon which curriculum, instruction, and assessments are built (NRC, 2001). A significant shift in standards initiates changes in each of these three pillars of classroom learning. Here we make the nature of these shifts explicit and consider their implications for changes in assessment.

Many previous state standards followed the model established by the *Benchmarks for Science Literacy* (AAAS, 1993; 2008) and/or the *National Science Education Standards* (NRC, 1996) in which the focus is primarily on building students' core science content knowledge, and although engaging students in science inquiry was encouraged as a way to build science literacy and skills, they were treated as separate, and in effect, secondary to content. This focus on learning content separate from inquiry, in addition to a lack of coherence across grades and disciplines, is often linked with students experiencing science as lists of decontextualized facts (Gorin & Mislevy, 2013).

The design of the new standards signals a change from previous standards in several fundamental ways, including elevating science practices and common themes in science to the same level of importance as content, integrating engineering concepts and practices into science learning, and taking an approach to deeper learning about fewer topics with particular attention to coherence across grades (NRC, 2012, 2014; Achieve, Inc., 2013). Table 1 summarizes the goals for changing science education that underpin the shifts in the structure and substance of the standards.

Science Education Will Involve Less	Science Education Will Involve More
Rote memorization of facts and terminology	Facts and terminology learned as needed while developing explanations and designing solutions supported by evidence-based arguments and reasoning
Learning of ideas disconnected from questions about phenomena	Systems thinking and modeling to explain phenomena and to give a context for the ideas to be learned
Teachers providing information to the whole class	Students conducting investigations, solving problems, and engaging in discussions with teachers' guidance
Teachers posing questions with only one right answer	Students discussing open-ended questions that focus on the strength of the evidence used to generate claims
Students reading textbooks and answering questions at the end of the chapter	Students reading multiple sources, including science-related magazines, journal articles, and web-based resources Students developing summaries of information
Preplanned outcomes for "cookbook" laboratories or hands-on activities	Multiple investigations driven by students' questions with a range of possible outcomes that collectively lead to a deep understanding of established core scientific ideas
Worksheets	Students writing journals, reports, posters, media presentations that explain and argue
Oversimplification of activities for students who are perceived to be less able to do science and engineering	Providing supports so that all students can engage in sophisticated science and engineering practices

Table 1. Implications of the Vision of the Framework and the NGSS. (NRC, 2015: pg. 11).

The conceptual shifts that mark the main differences between NGSS and previous standards are described in Appendix A of the NGSS (Achieve, Inc., 2013), and the ways the system of science education as a whole must change to successfully implement the standards are explored at length in the NRC (2015) report on implementing the NGSS (e.g., making assessment part of instruction such that evidence of learning can be gathered from classroom work products). The NRC (2014) report on developing assessments for the NGSS provides a thorough summary of the ways that the standards demand changes in assessment development. When considering the landscape of existing assessments, three of the changes outlined in these documents highlight particularly large gaps between what is needed and what has been produced in the past: integration of the three dimensions, inclusion of engineering within science topics, and a focus on fewer and deeper ideas.

Three dimensions. The standards are presented as a blend of three dimensions: disciplinary core ideas, science and engineering practices, and cross-cutting concepts. Under the vision of science learning outlined in the *Framework*, proficiency with any given concept could not be demonstrated just by knowledge of the relevant facts, but would require students to demonstrate how they can apply their knowledge to engage in the practices of science and to reflect on common themes that cross science disciplines. For assessments to support this vision of 3-dimensional learning, tasks must probe each of the three dimensions in a way that exposes developing proficiency of each dimension for formative uses, and for summative uses tasks must probe how well students are able to apply their science and engineering knowledge to engage with phenomena using the practices.

Engineering: To date, few states have formal requirements for engineering or technology education and in the rare cases where these subjects are taught they are separated from other STEM (science, technology, engineering, and math) subjects (NRC, 2009). Based on the premise that reasoning about science and engineering ideas with the tools that all STEM subjects bring likely contributes to a more robust understanding and interest in all these subjects (NRC, 2009; 2014), NGSS supports an approach to teaching science that breaks down the barriers that typically separate each of the STEM sub-disciplines by making explicit connections between science topics and engineering and math concepts. Engineering is integrated into science such that it is addressed as a practice, core idea, and cross-cutting concept, deeply intertwining the learning expectations for the two fields. Milano (2013) combined K-2 ideas for weather and climate and engineering design to describe an example of how this integration could support a robust learning experience around this topic:

By bundling these three performances together, students would have the ability to observe the natural phenomenon of sunlight warming Earth's surface, then generate questions about what kinds of problems that might cause in their everyday life, and finally apply their acquired knowledge of the effects of sunlight to the design of structure that will solve their problem.
(p. 5)

Following Milano's example, NGSS provides a foundation for moving beyond acquisition of factual knowledge and even investigations of phenomena, to ask students to identify problems and design solutions. With few existing resources to support K-12 engineering and technology learning (NRC, 2009), assessments that operationalize this vision of increased integration will be critical to support implementation of this conceptual shift.

Fewer, deeper ideas: Many state standards overwhelm science instruction with far too many topics, obscuring what is most important for students to learn and the ways students need to build on those ideas repeatedly across the grades (Roseman and Koppal, 2008). The new standards were designed to focus learning around the most important and enduring concepts, and they revisit those big ideas in each grade band with expectations that increase in sophistication. Assessments that support this shift should move away from covering vast amounts of discrete facts, and instead evaluate the ways students' mental models are building in sophistication and complexity. Assessments that support the learning of fewer ideas and coherence across subjects and grades should take into account the expectations from prior and future grades, the ways that tasks can draw on multiple science and engineering subjects, and focus only on the ideas that are central (and what is not central) to science literacy as described in the K-12 Science Framework (NRC, 2012).

These priorities mark an important change from previous standards and underscore the need for assessments that are markedly different from those used for previous state science tests. Moreover, for some states accommodation of these different assessments may even require reconceptualization of the assessment system for science.

A new assessment system. Implementation of NGSS requires the development of entirely new banks of assessments, which presents both a need and an opportunity to create a “next generation system of assessments” (NRC, 2014). A next generation assessment system should provide teachers with a coherent set of resources designed to provide the critical feedback to support student learning, in addition to classroom-based and external monitoring and evaluating achievement of the goals of the NGSS (Black and Wiliam, 1998; NRC, 2001, 2006, 2014). In a separate paper (Osborne et al., 2015) the SNAP team has described an example of such a system. In that paper, it is suggested that NCLB-mandated assessment should be based on a combination of short performance tasks and computer-based constructed and selected response items. This is to be supplemented by a task bank of classroom assessments (including examples of both stand-alone performance tasks and longer curriculum embedded learning tasks with assessment elements). Together, these two elements could establish a system which could support and monitor science teaching and learning of the new standards.

Such an assessment system would require the development of a bank of short performance tasks and computer-based assessments for all performance expectations² in the NGSS, as well as the classroom performance assessments that would be aligned to the broader goals of each subject. Examples of such tasks are needed soon if they are to signal the new priorities for teaching and learning to teachers, administrators, and resource developers. Yet, this endeavor requires substantial changes for state assessment systems. And, once decisions are made at the state level, early exemplars of each task type will be needed to support the training of item writers to elicit a different kind of thinking while also maintaining the necessary attention to important psychometric properties, and other factors required for high-quality assessment such as accessibility by all learners.

A need for model assessments. Example assessments will be an important source of guidance for teachers and curriculum developers who are preparing new instructional materials. Assessments operationalize the standards and instantiate the expected performances in a way that helps to communicate what competency looks like as defined by the new standards, and what proficiency judgments will be based on. They 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. For example, an end-of-unit assessment task that asks students to draw on the principles they have learned about photosynthesis and energy flows to construct their own explanation of energy flows (Fig 1a) sends a very different message about the kinds of learning activities students would need to experience in the classroom compared to a simple task (Fig 1b) that asks students to select the name of the process that transfers heat energy from the sun from four possible answers.

² The goals for each topic are laid out in a set of *disciplinary core ideas* (DCI), *science and engineering practices*, and *cross-cutting concepts*; each topic also has several Performance Expectations (PEs) that blend elements from each dimension. The goals described for each of the dimensions, however, are much broader in scope than the goals defined by the PEs.

Brent and Emilio heard that growing plants on the roof could lower energy usage. Write an Energy Story to explain to them what happens to energy from the sun in the picture and how it can be used to lower energy use in the house.



A roof with plants growing on it

Remember to include:

- Where does energy come from?
- How does energy move/transfers from place to place?
- Where does energy go/stored?
- How does energy change/transform?

You received feedback on your story about lowering energy use. Where in your story did you change what you wrote based on this feedback?
Did the feedback help you write a better story? Explain how.

Fig 1a: A screenshot from one curriculum-embedded assessment task adapted from the Web-based Inquiry Science Environment (WISE). This task follows a series of activities using a computer model of an ecosystem to investigate how plants use and transform energy. Students write an initial essay, get automated guidance, revise, and also comment on their revision process (see Gerard, Ryoo, McElhaney, Liu, Rafferty, & Linn, 2015).
Reproduced with permission from wise.berkeley.edu.

Heat energy from the Sun is transferred to Earth primarily by which of the following processes?

- A. conduction
- B. convection
- C. evaporation
- D. radiation

Fig 1b) An item from the 2012 middle school (8th grade) MCAS earth science test for Massachusetts. Reproduced from doe.mass.edu

Synopsis of the landscape review: characteristics of promising existing assessments

This study has sought to identify model assessments that showed promise as an approach to assessing an important part of NGSS, even if they were not fully aligned as written. For example, perhaps a task is shown to be effective at probing students' ideas about developing a model, but it does not require knowledge of any Disciplinary Core Idea. The way such a task evaluates ideas about modeling might then become the foundation for the development of a new multidimensional task that is more closely aligned to NGSS.

The SNAP team conducted a wide-ranging search for suitable assessments drawing on their own knowledge and that of a set of experienced advisors. From this work they assembled promising assessment resources into a task bank, designed a set of evaluation criteria, and used the criteria to review a sample of these assessments (see Appendix I for a complete description of the methods). The task bank was supplemented and reviewed by the SNAP advisors³. It includes 203 assessment resources that cover the range of item formats that would be part of this system (multiple choice, constructed response, and short and extended performance tasks), assessments designed for different purposes (e.g., curriculum-embedded, external summative), and aligned to constructs relevant to NGSS (DCIs, science and engineering practices, and cross-cutting concepts). These evaluation criteria characterize existing assessments in terms of:

- 1) **Characteristics:** where they fit in the assessment system (grade, subject, format, timescale, etc.);
- 2) **Alignment:** how the tasks they contain are and are not aligned to the NGSS and other prioritized assessment practices (e.g., accessibility); and
- 3) **Evaluation:** whether they have tasks that hold promise as models for NGSS assessments.

The contents of the task bank were summarized in two ways. First, a sample of the task bank was studied in detail using the evaluation criteria to describe which aspects of NGSS are covered well by existing assessments, and which aspects require particular attention. The sample was selected to be broadly representative of the assessment formats represented in the task bank – selected response, performance tasks, and a range of technology-enhanced formats. Fifty-one assessment resources were evaluated in this sample, but there is substantial variability in the size of any one resource. For instance, an assessment resource might contain one multi-component performance assessment, or it could contain as many as 100 individual items; overall, roughly 400 tasks were analyzed.

The second summary of the task bank is a description of general trends observed across the entire bank showing the ways most existing assessments do, or do not match the goals of NGSS, and the changes in approach to item development needed for the new generation of assessments.

³ See snapgse.stanford.edu for the list of advisors.

A brief overview of the findings for both analyses follows in the next section. The study of the sample is described in greater detail in the Appendix. The summary of the general trends is elaborated and illustrated with numerous example items later in the Gap Analysis.

Summary of findings from the Landscape Review. 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 2). For example, the review identified 19 assessments (out of 51) that contain tasks aligned to elements of the practice *developing and using models*. These assessments use a variety of formats, from multiple-choice to performance tasks, different content areas, and platforms (e.g. computer-based), providing diverse examples of ways to probe this practice. When it comes to assessing students' ability to 'ask questions' only one form of question was commonly found. It is unclear if this a reflection of the lack of imagination and creativity of item writers, or alternatively, 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 3), although the assessment of *cause and effect: mechanism and explanation* can be found in these assessments about twice as frequently as the other concepts. One conclusion that can be drawn from the scarcity of existing assessments aligned to the cross-cutting concepts is that there is a substantial need for more models that show ways to probe these concepts across diverse formats, content areas, etc. Another is that the cross-cutting concepts bring us into uncharted territory that create a mismatch between the NGSS and nearly all existing assessments.

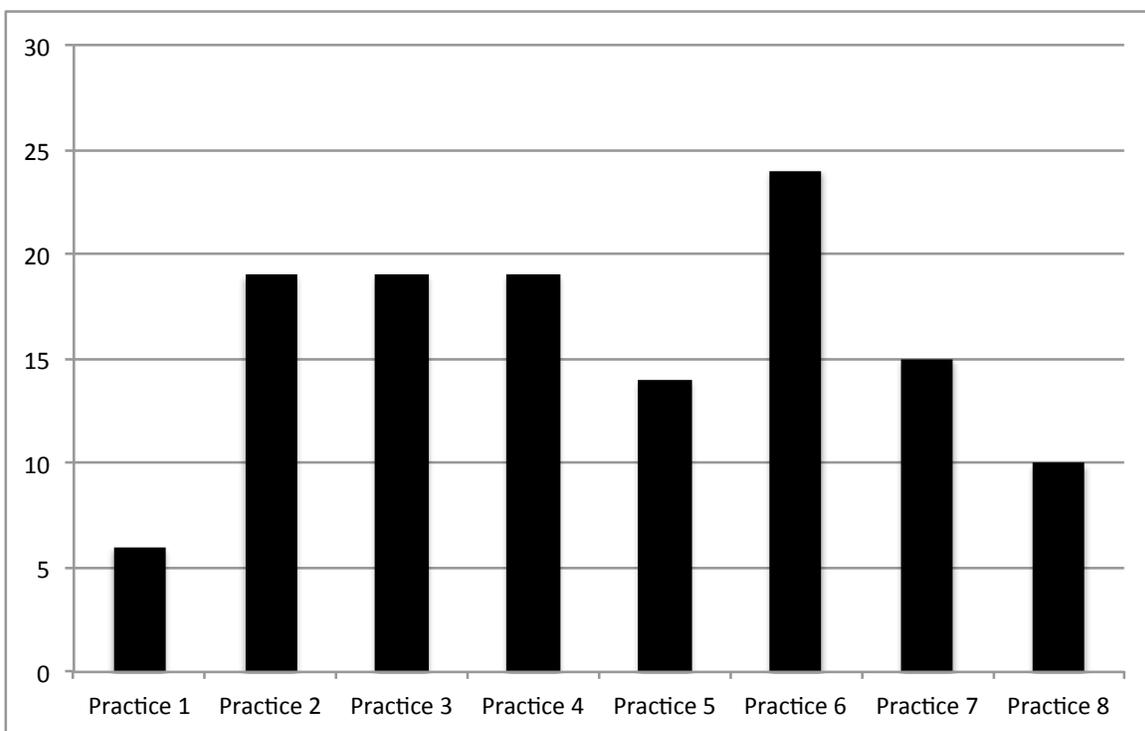


Figure 2. Bar graph showing the number of assessments in the task bank that have at least one task that targets a science and engineering practice. Practice 1: asking questions and defining problems; Practice 2: developing and using models; Practice 3: planning and carrying out investigations; Practice 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)

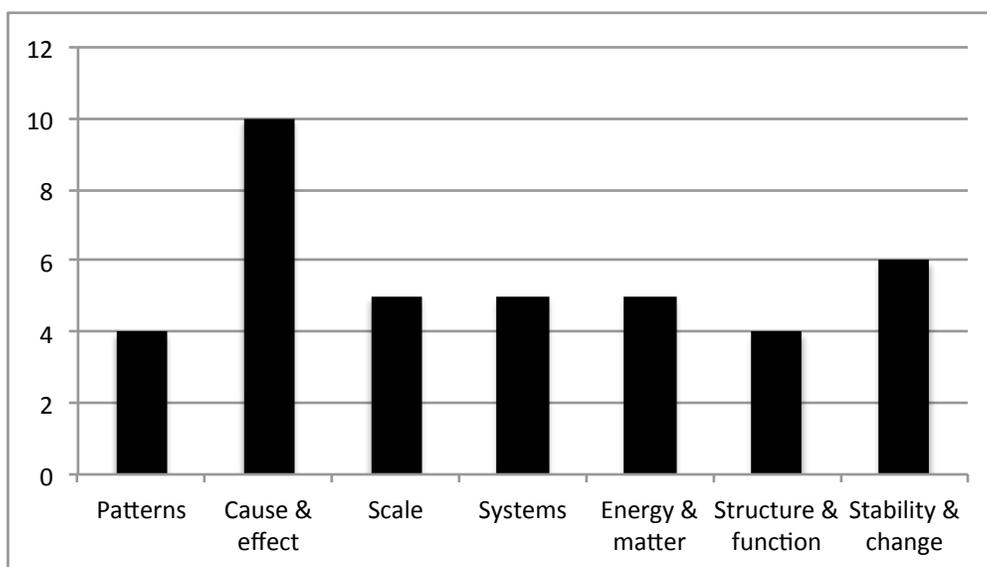


Figure 3. Bar graph showing the number of assessments in the SNAP task bank that have at least one task that aligns to a cross-cutting concept. (N=51)

As discussed above, the separate assessment of content and practices was typically a core design principle for the previous generation assessments, and this means that integration of the two elements requires a fundamental change to assessment design. Therefore, identifying assessments 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, indicating that these efforts proved successful. 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 4). Un-integrated tasks were composed of a cluster of items, each of which probed one dimension but no single item was aligned to more than one dimension.

Another area of interest is assessments that make connections to the Common Core. The NRC (2015) *Guide to Implementing the NGSS* encourages making connections with other curriculum subjects, and the NGSS draws explicit connections to Common Core standards for each topic. Twenty-seven percent of assessments reviewed had some connection, but almost two-thirds of those connections were to math. The practice of “*Obtaining, evaluating, and communicating information,*” which focuses on critical reading and summarizing of key ideas from science and engineering texts, and using words and visuals to communicate scientific concepts, and has clear potential to align to the Common Core Language Arts but few existing science assessments were found to evaluate this practice.

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 review 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 (Table 2). 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*.

State	Description of resources available	Sample tasks
DE	Task bank of performance assessments with anchors and several examples of student work Notes: these are really extended constructed response items	HS PAs: http://dedoe.schoolwires.net/Page/571
CT	Task bank with one curriculum-embedded performance assessment (CEPA) for each grade for classroom use and sample items from on-demand selected-response assessment	http://www.csde.state.ct.us/public/csde/cedar/assessment/capt/resources/released_items/2013/2013%20CAPT%20Released%20Items%20%28Science%29.pdf

IL	Task bank includes several classroom assessment tasks with scoring guides and numerous examples of student work Notes: these are not great items but this is a nice example of significant attention to the teaching and learning side of the assessment system.	http://www.isbe.state.il.us/ils/science/capd.htm http://www.isbe.state.il.us/ils/science/stage_B/assessment.htm
LA	Task bank with CEPAs is available online, designed for use as classroom activities or assessments Notes online formative assessment tool is not accessible to the public	e.g., for HS http://www.louisianabelieves.com/resources/library/teacher-support-toolbox-library/9-12-grade-science-teachers
OR	Task bank includes diverse resources: Online performance assessments developed for NGSS, sample on-demand tests, engineering design notebook templates, inquiry prompts	http://www.ode.state.or.us/search/results/?id=240 http://www.ode.state.or.us/wma/teachlearn/testing/scoring/guides/2011-12/science_engdesign_notebooktemplate_ms.pdf http://www.ode.state.or.us/wma/teachlearn/subjects/science/assessment/sample_sbperftask_genetic-engineering.pdf
UT	Task bank is an example of a combination of computer-based performances and interactive selected-response items as well as traditional multiple choice items	https://login1.cloud2.tds.airast.org/student/V106/Pages/LoginShell.aspx?c=Utah_PT

Table 2. Examples of assessment resources developed by States and released to the public that move beyond paper/pencil-based multiple-choice assessments.

Summary of the trends from the Task Bank. Emerging from this review are several critical gaps between features of assessment tasks required for NGSS and those of existing tasks. Examination of the entire task bank confirmed this conclusion, as it identified few assessments with tasks that were better fit for NGSS than those sampled for the review. Nevertheless, for most of the features, assessments that model effective approaches to incorporating those features were identified. In some cases, this required looking beyond the task bank and doing targeted searches of the research literature, but the important message is that for each dimension of NGSS, there are existing tasks that can be immensely useful resources for guiding the design of new tasks.

Four of the “gaps” and their implications for the design of a new generation of assessments are described in brief below and later in detail in the Gap Analysis.

1. **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.

2. Focus on the big ideas in science: the NGSS emphasize the big ideas and themes in science, and fine details are included only as they are central to making sense of the big ideas. In fact, the writers of NGSS deliberately excluded some topics that have long been part of science classes because they were considered non-essential for contributing to students' understanding of the big ideas. But fine details, often discrete facts, are generally much easier to assess than big ideas. So it is unsurprising that many existing tasks evaluate these facts, and that many evaluate knowledge and skills that are not directly targeted in the new standards.
3. Evaluate 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.
4. Use a variety of task formats: Most existing tasks probe a very small piece of content or practice, and therefore provide limited insight into what students think and what they can do. Well-designed performance tasks are able to probe much more deeply into students' reasoning and their ability to draw on their 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.

The following section explores each of these four gaps using numerous example tasks to illustrate both the ways most existing tasks do not meet the needs of NGSS and the ways some tasks model approaches that better fit those needs.

Gap Analysis: implications for the development of assessments for the NGSS

The landscape review 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 review did identify some tasks that model ways of addressing the gaps. In this section examples of tasks that fall into these four critical areas are contrasted with those that do not to make explicit the characteristics that new assessments should avoid, and to provide models of the ways that some assessments might better align with the goals of the NGSS.

1. Alignment and integration of assessments to the three dimensions.

The *K-12 Science Framework* (NRC, 2012) describes the importance of learning science in three dimensions: Disciplinary Core Ideas, Science and Engineering Practices, and Cross-cutting Concepts.

“...in order to facilitate students' learning, the dimensions must be woven together in standards, curricula, instruction, and assessments. When they explore particular disciplinary ideas from Dimension 3, students will do so by engaging in practices

articulated in Dimension 1 and should be helped to make connections to the crosscutting concepts in Dimension 2.” (pg. 29)

As discussed above, certain contexts might call for assessing the dimensions separately, particularly in the case of formative assessments. But in general, to evaluate the kind of engagement with scientific concepts in the manner described in the *Framework*, assessment tasks should be aligned to the three dimensions so that the targeted knowledge is integrated with a practice and a cross-cutting concept. A task that consists of multiple interrelated items might probe the three dimensions in its entirety, though each of the component items might probe just two or even one dimension (NRC, 2014).

There are several common ways that existing tasks appear to be multidimensional but probe only one dimension. Some tasks use a context relevant to a DCI, but they engage students in a practice in a way that they do not have to draw on their content knowledge. Conversely, some tasks present a problem or question as the context of a task, but responding to the task only requires declarative (factual) knowledge without use of a practice. In other cases, tasks present a problem but provide insufficient information to solve it, so the only way students can get to the correct answer is if they already were familiar with the problem. A common example involves asking students to analyze the features of the Grand Canyon to explain how it formed; many students know that the Grand Canyon is formed by erosion so their response requires only the recall of that fact, not analysis of evidence presented in the task.

The next two figures show examples of tasks that probe one dimension at a time. The first example (Figure 4) presents the problem of oil spills and their economic and ecological impacts, and shows some data relevant to the problem. But item (a) asks students to interpret the graph without using any content knowledge. Moreover, this basic level of interpretation, reading a bar graph, would not be considered aligned to a high school practice. Items (b)-(f) demand only content knowledge with no data analysis or other science practices. Multidimensionality can be achieved using tasks comprised of several components, but at least some of these components would need to blend more than one dimension, such as having students cite data from a graph to support an explanation of how damages to an ecosystem can impact an economic system.

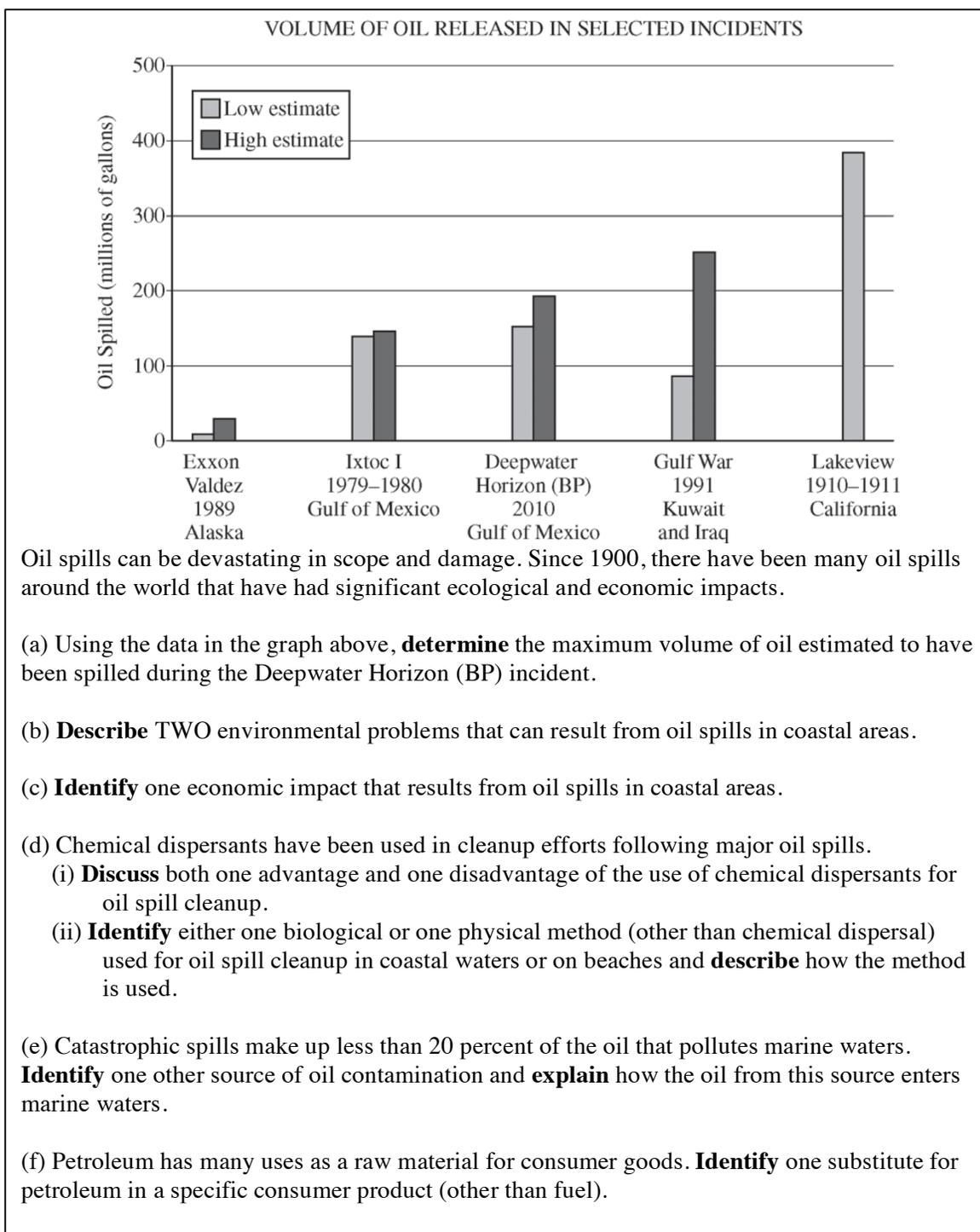


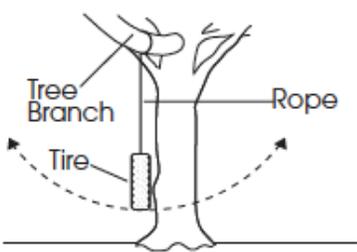
Figure 4. An Advanced Placement Environmental Science (2015) task from the College Board that probes graph reading and content knowledge about oil and environmental impacts of oil. Reproduced from collegeboard.org.

Although the task in Figure 5 (Ohio, Grade 8) presents physical science content about the motion of a pendulum in the context of an investigation, the task is not multidimensional. Indeed, a student could identify a pattern on the data table without knowing about the

phenomenon being investigated. Furthermore, there is a second single-dimensional route to the answer. 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. Although data analysis is one of the science and engineering practices, alignment to that practice should require more than selecting the correct answer based on a pattern in a data table; for instance, if this task asked students to analyze the data and to cite these data to describe a pattern, it would become a two-dimensional item (*identify a pattern* is a cross-cutting concept).

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

Tasks that require students to engage with science data but do not require students to use any content knowledge to analyze, interpret, or perform any other science practice are also common. In Figure 6, students must look for a pattern in the data and extend the pattern, but there is no explicit connection to the underlying phenomenon. Some students might observe a pattern of numbers increasing by one or two over the day. And some students might know that the numbers will eventually decrease later in the day, but this task cannot be used to tease apart the students who are just evaluating a number pattern and those who are thinking about how air temperature tends to rise and fall over a day and to make a predication about the pattern while drawing on their content knowledge.

15. A student recorded air temperatures on Monday and Tuesday.

Air Temperature Information

Time	Monday	Tuesday
8:00 AM	45°F	50°F
9:00 AM	47°F	52°F
10:00 AM	48°F	53°F
11:00 AM	50°F	55°F
12:00 PM	52°F	57°F
1:00 PM	?	?

A. Describe a pattern in the student's recorded data.

B. Based on the pattern, predict the **most likely** air temperature at 1:00 PM for each day.

Monday: _____°F

Tuesday: _____°F

Figure 6. A data analysis task for 4th grade from Alaska. Reproduced from education.alaska.gov.

Models of multi-dimensional assessment tasks

Figure 7 from the International Baccalaureate® Environmental Science exam (2010) shows a farming system annotated with information about key elements of the system. To answer the first part of the question students must analyze the information provided, compare this information to their knowledge of general characteristics of different types of agricultural systems, and decide which type this case fits best. Students present an argument to support their claim, draw evidence from the diagram and table, and justify it

with their knowledge of agricultural systems. Part b of the question has students use the diagram and table as sources of evidence to demonstrate how this farm is an example of a system.

Although the content targeted by this item is not aligned with any DCI in the NGSS, it provides an example of a task composed of three two-dimensional components, which all together comprise a three-dimensional task.

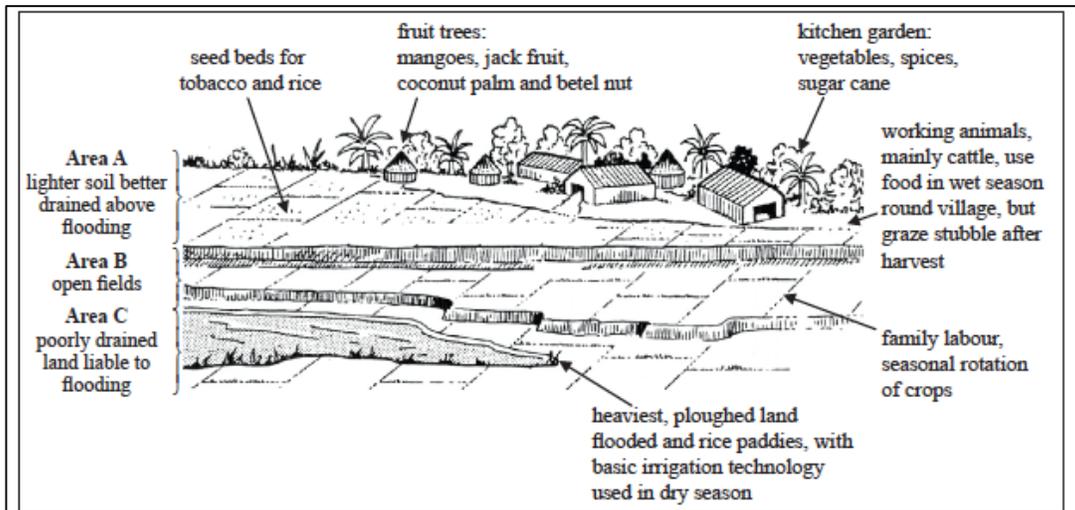


Figure 2(b)

Month	March	April	May	September	March
Season	Pre-monsoon			Wet season	
Area A	cattle in yard, mangoes, vegetables			repairing and thatching, green coconuts, betel nuts	
Area B	jute			wheat, tobacco, mustard	
Area C	grazing, rice (flooding)			grazing	

[Source: Adapted from M Carr, *Patterns, Process and Change in Human Geography*, Macmillan, (1987), page 142]

- (a) State, giving two reasons, whether this system is more typical of farming in a more economically developed country (MEDC) or a less economically developed country (LEDC).

[2]

(b) Complete the systems diagram below to show **three inputs, processes and outputs** for the farming system shown in Figure 2(a) and Figure 2(b). [3]

Inputs

1.

2.

3.

Processes

1.

2.

3.

FARM

Outputs

1.

2.

3.

(c) With reference to Figure 2(a) and Figure 2(b), describe **two ways** in which the farming system has been developed in response to variations in the local environment. [2]

Figure 7. A three-part task (a-c) from the International Baccalaureate® Environmental Science exam (2010). The first image shows a farming system and the table below shows the farming activities in that system in Areas A, B, and C. The questions ask students to analyze and interpret data, construct a system model, and make an evidence-based claim.

Reproduced with permission from the International Baccalaureate®.

Another example of a multi-dimensional task (Figure 8) is from the Next Generation Science Assessment (NGSA)⁴ project for the topic: matter and its interactions (Grades 6-8). In this task, students watch a short video of a phenomenon, dye-coated candies put into water at different temperatures. Students draw models and write an explanation to show why the dye on the candies spread differently at the different temperatures. This two-part task requires that students use their physical science knowledge to develop a model that shows the cause of a phenomenon (DCI and science practice) and to construct a written explanation for the phenomenon (DCI, science practice). Also embedded in this task is a cross-cutting concept, *cause and effect: mechanism and explanation*. In a fully three-dimensional task this element of the task would not be embedded, but would ask students explicitly to include the mechanism that caused the phenomenon they observed.

⁴ NGSA is a multi-institutional research and development collaboration among Michigan State University, SRI International, University of Illinois Chicago, and Concord Consortium.

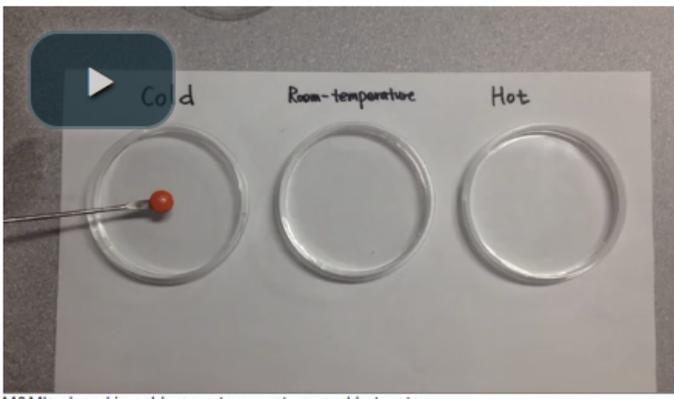
Question #1

Watch the video clip. Construct a model to explain why the M&M behaved differently in cold, room temperature, and hot water. Your model should include both pictures and words to explain the behavior of M&M particles in the water at different temperatures.

Cold Water (5°C)	Room Temp. Water (20°C)
Hot Water (80°C)	

Describe how your model explains the observed behavior of the M&Ms.

[Make drawing](#)



M&M's placed in cold, room-temperature, and hot water.

Figure 8. a two-part task for grades 6-8 on energy and states of matter that incorporates a short video. Reproduced with permission from the Next Generation Science Assessment project (NGSA).

Not all tasks that target reasoning in three dimensions have to require a lot of time. For example, the short-response items in Figure 9 are a series of questions designed to provide insight into students' ideas about a physical science concept. The items require students to analyze a phenomenon, decide if the statement is a correct description and mechanism to explain the predicted motion using their knowledge of the earth's gravitational force. These types of item sets provide a quick way for teachers to check whether their students are on track because each question, as in the examples shown in Figure 9, is grounded in a common student misconception or a critical concept central to the topic.

2. Focus on big ideas in science

Another trend that emerged from the landscape review is that many existing tasks focus on the fine details of a scientific concept at the expense of the big ideas. In other words, assessments often prioritize what is easily assessable over what is most important for students to know and be able to do. The disciplinary core ideas and the cross-cutting

APPLES & EARTH



Look at the picture and decide if each statement is true or false. Explain your answer in the space below each statement.

True False

1 Earth pulls with an equal force on the big apple and the small apple.

True False

2 Earth pulls on the big apple, and the big apple pulls equally hard on Earth.

True False

3 When an apple drops, it will fall at a constant speed because the gravitational pull from Earth is a constant.

Figure 9. A diagnostic task from the WestEd Making Sense of SCIENCE: Force & Motion formative assessments. Reproduced with permission from WestEd.

concepts in NGSS both reinforce the commitment of the design of the *Framework* to focus on the most important and broadly explanatory ideas of science and to de-emphasize the details that are not essential to understanding those ideas:

Specify big ideas, not lists of facts: Core ideas in the framework are powerful explanatory ideas, not a simple list of facts, that help learners explain important aspects of the natural world.” (NRC, 2012: Pg. 254)

There are several common ways that existing assessments fall short of this goal, particularly by testing isolated facts that are not explicitly linked to larger themes, and by probing content or vocabulary that was deliberately excluded from the NGSS because it is not essential for understanding the big ideas.

Many existing tasks ask students about content knowledge that is so narrow in scope that it would be difficult or impossible to infer what the student thought about the core principles. For example, a task from TIMSS (2011) (Figure 10) 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 10. 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:

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)

The framing of the DCI statement for the Life Science subtopic “structure and function” around systems and subsystems demands development of assessments that do not simply focus on recall of the function of one element of a system, but push students to consider either how that element is composed of systems of cells with specific functions or how it functions as part of the reproductive system. Although Figure 10 poses a question that is related to this DCI, and could be a useful item as part of a set, an assessment that probes the kind of reasoning that is the goal of the *Framework* and *Standards* would need to get at a broader concept than the function of an individual organ. Moreover, an assessment aligned to the NGSS Performance Expectation for this idea (MS-LS1-3) would have to look very different from the one shown in Figure 10.

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

Performance expectation (PE) MS-LS1-3 would require an assessment in which students made a claim about what makes a body a system composed of subsystems down to the cellular level. Instead of stating the function of the uterus, students would have to use their knowledge of its function in relation to other systems to cite evidence that supports an argument about how the uterus is both part of a system and how it has a system within it. For instance, students could be given a diagram that shows an early stage embryo in the uterus and the surrounding circulatory system, and they could be asked to use the evidence from this diagram plus their general knowledge to describe processes taking place in the embryo and ways those processes are supported by the body systems of the mother.

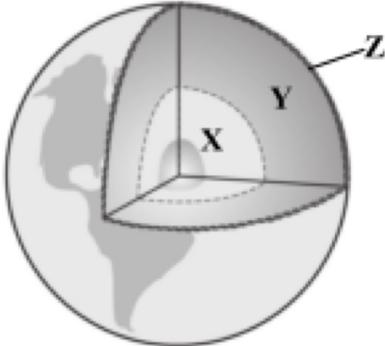
This comparison between an existing task and the new standards illustrates the significant changes between content expectations in NGSS compared with previous standards. Decisions about the content to include in the *Standards* primarily focused on a set of core

principles and the concepts necessary to understand each principle. As a result, some content that has been a part of most previous standards was eliminated from the NGSS in an effort to allow for this deeper learning.

“The Framework identified a smaller set of Disciplinary Core Ideas that students should know by that they graduate from high school, and the NGSS are written to focus on the same. It is important that teachers and curriculum/assessment developers understand that the focus is on the core ideas-not necessarily the facts that are associated with them.” (NGSS, pg. 2)

Assessments can encourage and help communicate this focus on deep learning of a smaller set of ideas if they closely adhere to the boundaries around each core idea. However, many existing items fall outside of these boundaries. The item shown in Figure 11 asks students to label three layers of the earth, recall a characteristic of each layer, and name a way that two of those layers interact. The closest DCI, ESS2.A (below), is about the flow of energy and matter as a driver of large earth system processes. Knowledge of the names of the layers of the earth might be useful for discussing these system processes, but focusing on the detailed factual knowledge required for this question distracts from the more significant objectives of performance expectation MS-ESS1-4 shown beneath.

The diagram below shows three main layers that compose Earth. The layers are labeled X, Y, and Z.



1. Identify each of the three layers of Earth (X, Y, and Z) labeled in the diagram.
2. Describe one characteristic of the layer labeled X.
3. Describe one characteristic of the layer labeled Y.
4. Describe one way that the layer labeled Y interacts with the layer labeled Z.

Figure 11. A multiple choice item from the 8th grade Massachusetts state science test, MCAS (2011). Reproduced from doe.mass.edu.

From NGSS:

Disciplinary Core Idea MS-ESS1.C

The geologic time scale interpreted from rock strata provides a way to organize Earth’s history. Analyses of rock strata and the fossil record provide only relative dates, not an absolute scale.

Sample Performance Expectation MS-ESS1-4

Construct a scientific explanation based on evidence from rock strata for how the geologic time scale is used to organize Earth's 4.6-billion-year-old history.

Models of assessments that probe big ideas

An example of a task that deeply explores a foundational concept is shown in Figure 12. 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. This task adheres closely to probing only content required to evaluate students' proficiency with the big idea central to the DCI PS1.A: Structure and Properties of Matter.

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?

- Fish are able to breathe the oxygen in the water.
- Bubbles are made of air.
- The pan has been boiling for 10 minutes and it is still bubbling.
- 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 12. 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).

Although tasks that are aligned to the content goals described in the NGSS do exist, the overwhelming lack of alignment between the DCIs and content targeted by promising assessments in the task bank was one of the most alarming findings of the study.

3. Probe the full breadth of science and engineering practices

The central role that science and engineering practices play in the *K-12 Framework and Standards* has led to the practices being described in far more detail than in previous standards. These new goals are novel in several ways: a) they include some practices that were not explicitly defined in previous standards; b) they combine engineering practices with those for science, and c) they break the practices down into different component practices from previous standards. Each of these changes produces a gap between existing assessments that were developed for a different set of goals and what is needed for assessing the NGSS.

Many of the science and engineering practices overlap with the practices or inquiry skills that were used in the National Science Education Standards (NSES) and other previous standards, such as *planning and carrying out investigations* and *analyzing and interpreting data*. In many cases, though, specific components of the Science and Engineering Practices are different from those defined elsewhere. Figure 13 shows the components of one practice in NGSS and a short excerpt from the description of the same practice in NSES. The differences in the components of the practices naturally lead to differences in the way the practices are assessed; tasks aligned to NSES might focus on knowing about types of investigations and the kinds of questions they can address, whereas those for NGSS might focus on conducting an investigation and collecting data in addition to planning and evaluating the design of investigations.

Planning and Carrying out Investigations for middle school in NGSS

- **Plan an investigation** individually and collaboratively, and in the design:
 - identify independent and dependent variables and controls,
 - what tools are needed to do the gathering,
 - how measurements will be recorded, and how many data are needed to support a claim.
- **Conduct an investigation** and/or evaluate and/or revise the experimental design to produce data to serve as the basis for evidence that meet the goals of the investigation.
- **Evaluate the accuracy** of various methods for collecting data.
- **Collect data** to produce data to serve as the basis for evidence to answer scientific questions or test
- Collect data about the performance of a proposed object, tool, process or system under a range of conditions

An excerpt from the middle school fundamental understandings of inquiry in the National Science Education Standards (NSES)

- Scientific investigations involve asking and answering a question and comparing the answer with what scientists already know about the world.
- Scientists use different kinds of investigations depending on the questions they are trying to answer.
 - Types of investigations include describing objects, events, and organisms; classifying them; and doing a fair test (experimenting)
- Simple instruments, such as magnifiers, thermometers, and rulers, provide more information than scientists obtain using only their senses

Figure 13. The components of the science and engineering practice *planning and carrying out investigations* as they are described in NGSS and an excerpt of the components for this practice as described in NSES (NRC, 1996).

Indeed, there are numerous existing assessments that evaluate specific details of planning an investigation, such as identifying independent and dependent variables (Figure 14), and there are also many tasks that ask students to evaluate the methods or design of an investigation (Figure 15). But very few existing assessments engage students in planning their own investigation, including identifying the data they would need to collect and the appropriate methods to collect them, going through the process of conducting the investigation to collect data, or evaluating the limitations of the methods.

Dan and Dawn want to know if there is any difference between the mileage expected from bicycle tires from two different manufacturers. Dan will put one brand on his bike and Dawn will put the other brand on her bicycle. Which of the following variables would be MOST important to control in this experiment?

- a) The time of day the test is made.
- b) The number of miles traveled by each type of tire.
- c) The physical condition of the cyclist.
- d) The weather condition.
- e) The weight of the bicycle used.

Figure 14. A task that probes students' ability to identify the variable being controlled in an experiment from the Iowa Assessment Handbook (Enger & Yager, 1998).

Scientists are interested in knowing how much mass is contained in the Universe. The total mass of the Universe will affect the way in which the universe will end.

Astronomers are able to calculate the mass of individual galaxies in two ways. The first is by looking at how fast they spin. The faster a galaxy spins the more mass it must contain. The second method is to add up the masses of all the visible objects in the galaxy.

In the 1930s an astronomer, Fritz Zwicky, used both of these methods to calculate the mass of galaxies in the Coma cluster. The two methods led to very different results. There was far too little visible matter to account for the rate at which the galaxies were spinning. Zwicky concluded that there was something there which was exerting gravitational attraction but which could not be seen.

Zwicky's findings were mostly ignored for the next 40 years. However, other astronomers also found that there were other measurements that could not be explained by the mass of the visible universe. The 'missing' mass was named dark matter because it does not emit electromagnetic radiation and could not be detected by telescopes. It is now calculated that only about 5% of the mass of the universe comes from visible matter such as stars and planets

Scientists using the Large Hadron Collider (LHC) will recreate the conditions in the universe just after the 'Big Bang'. They are hoping that they will be able to create and study dark matter particles. This will allow them to develop a better description of the way in which the Universe began, and what it is all made from.

Suggest why Zwicky thought that it was worthwhile to use two different methods to work out the mass of the galaxies?

Figure 15. A constructed response item that probes students' ability to evaluate the reasoning behind the methods for an analysis. From the Science in Society A-level exam. Reproduced with permission from AQA <http://www.aqa.org.uk/subjects/science/as-and-a-level/science-in-society-2400>.

Figure 16 shows one of these rare examples of a multiple-choice task that asks students to identify the data needed to investigate a question. Figure 17 shows another example: a constructed response task that asks students to design an investigation that would help to determine which of two arguments can be used to correctly explain a phenomenon. A task such as this one might be more effective at eliciting correct responses if it were explicit about the elements of an investigation that must be included (e.g., data to collect, variables to control, question to investigate, etc.), but the simple, well-constrained task provides an example of a promising foundation for assessing this practice.

QUESTION 18.3

At the end of the article Ferwerda refers to scientists who say that carbon dioxide is not the main cause of the Greenhouse effect.

Karin finds the following table showing the relative Greenhouse effect caused by four gases:

Relative Greenhouse effect per molecule of gas			
Carbon dioxide	Methane	Nitrous oxide	Chlorofluorocarbons
1	30	160	17 000

From this table Karin cannot conclude which gas is the main cause of the increase of the Greenhouse effect. The data in the table need to be combined with other data for Karin to conclude which gas is the main cause of the increase of the Greenhouse effect.

Which other data does Karin need to collect?

Data about the origin of the four gases.

- A. Data about the absorption of the four gases by plants.
- B. Data about the size of each of the four types of molecules.
- C. Data about the amounts of each of the four gases in the atmosphere.

Figure 16. A question about identifying the data needed to investigate a question from a multi-part item in PISA (released items from 2000-2006). Reproduced from nces.ed.gov/pisa.

Kayra and Emre are studying plants. They have learned that characteristics such as the height of plants and the color of fruit are inherited.

They are looking at some green and red peppers.



Kayra thinks they are different kinds of peppers, because they are different colors.

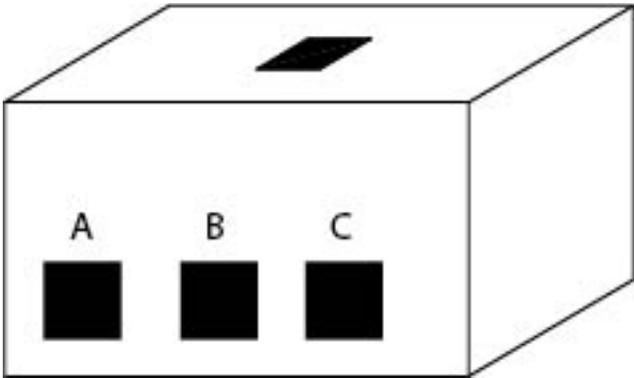
Emre thinks that they are the same type of pepper, and red peppers are red because they have been left on the plant longer and have ripened.

Describe how you could set up an investigation to decide whether Kayra or Emre is correct.

Figure 17. A constructed-response task about designing an investigation to test two ideas from TIMSS (2011). Reproduced from nces.ed.gov/timss.

As described above, 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. Strong examples can be found in research journals (Figure 14), but the challenges of getting access, collecting, and adapting them for classroom or large scale use often precludes their utility as models.

The person demonstrating the operation of the box below has two containers of balls, one with black balls and one with white. First, a black ball is dropped into the opening on top of the box. After about 3-5 seconds a white ball comes out of the A opening on the front. Next, a white ball is dropped in the top. A black ball comes out of opening B. Throughout the demonstration, whenever a ball is put in the top opening, the other color ball comes out the front. All the balls, regardless of color, always come out first from opening A, then B, Then C and back to A.



- Procedure
- Students collect ideas from group members
- Brainstorm
- Rely on previous knowledge about black boxes and everyday things
- Create a representation
- Use an analogy
- Test for explanatory adequacy
- Test for predictive adequacy

Students were told that they would revisit this modeling strategy list and add to it as the class proceeded. The discussion of models was then put aside, and two weeks of instruction on Mendel’s model of simple dominance, meiosis, and use of computer models ensued. Strategies for production and evaluation of models were discussed again only after the students had more experience with specific biological modeling problems.

Figure 18. A description of an assessment task that targets students’ development and use of a model (modified from Cartier, 2000).

Using mathematical and computational thinking also has not been as prominent a goal of previous standards, but the direct connections that can be made to the math standards in the Common Core State Standards bring extra attention to this practice. However, the landscape study found that overall it was probed infrequently (14 of the 51 assessments reviewed include a task that probes this practice). The low number of assessments that have items aligned to this practice may be explained by the nearly exclusive focus on the assessment of content knowledge that has dominated science assessment for the last decade. Some organizations, such as the College Board, have already changed their frameworks to prioritize computational thinking in their item development process, and numerous examples of ways to incorporate this skill can be found in their released Advanced Placement tests (see apcentral.collegeboard.com). The emphasis in the *Framework* is on using mathematical computational thinking to reason with data rather than algebraic representation. It focuses on “the abilities to view data from different perspectives and with different graphical representations, to test relationships between variables, and to explore the interplay of diverse external conditions” (NRC, 2012, p. 65). Yet, many of the existing assessments that target this skill have expectations that far exceed the competencies outlined in the framework.

For example, Figure 19, from Achieve, Inc., shows an excerpt from a performance task designed to target mathematical and computational thinking for NGSS. In steps A-E of this task, students draw models of solar ovens and show how energy is transferred in each design, test and implement the most efficient design, explain the reasoning behind key decisions, and construct a computational model to simulate use of the oven and discuss tradeoffs in optimizing the design. Step F, shown below, has students use their calculations from the computational model to guide decision making about optimizing their solar oven design, and they articulate how their calculations were used to make these decisions. Students also consider the ways in which the use of a mathematical equation was and was not useful for guiding their re-design and make observations about uncertainty associated with their equation. The development of a computational model from scratch exceeds the goals for this practice in middle school, but the use of a mathematical model to inform design decisions fits well within the middle school goals, and discussing the role of uncertainty in the application of this model aligns to middle school goals for *analyzing and interpreting data*⁵.

⁵ It should be noted that Achieve developed a set of nine model classroom performance assessments for middle school and high school (<http://www.nextgenscience.org/classroom-sample-assessment-tasks>). To our knowledge these tasks have not been pilot tested yet, and we expect that the substantial reading load, in addition to the inclusion of activities that we consider better fit for high school, will make them particularly challenging to students.

F. Using what you have learned from working with your spreadsheet calculations and the temperature data you collected during your first test, refine the design of your solar oven and update your design plan to further maximize the change in temperature within the oven and to meet the constraints you identified. Discuss the reasoning behind the components of your design as they relate to the components of the equation and your spreadsheet calculations. Test your redesigned oven. Using the same criteria as stated in part B, collect peak temperature data for your oven and calculate the change in temperature between the inside of your oven and the temperature outside your oven. Compare the two temperature values and discuss why you think the changes you made lead to the change in the temperature values. In your discussion, consider the spreadsheet calculations you made in task Component E. How were these calculations useful for predicting the change in temperature values you observe here? Were there limitations to the usefulness of your equations? Specifically describe which variables were the most uncertain or most difficult to choose values for, given your design.

Fig 19. An excerpt from a middle school physical science performance task on designing a solar cooker. Adapted from Achieve, Inc. (nextgenscience.org).

Figure 19 also serves as an example of one of the few assessments in our task bank that targets engineering concepts and practices. Despite the absence of national engineering learning standards, 36 states have a strong presence of engineering in their standards, and 12 of those states have engineering in the science standards (Carr et al., 2012). But the slow increase in adoption of this subject into K-12 classrooms has not yet led to development of a substantial collection of model assessments. The few places where tasks for K-12 engineering are publicly available will be critical resources for guiding development of model assessments. Some places where engineering tasks have been made publicly available include:

- Massachusetts Department of Education (<http://www.doe.mass.edu/mcas/search/>), which has been a leader in introducing engineering into K-12 classrooms
- NAEP Technology and Engineering Literacy (<https://nces.ed.gov/nationsreportcard/tel/>)
- select research and development groups that have created resources that could be modified for use as assessments (e.g., <http://concord.org/stem-resources/subject/engineering>)

4. Assessments for the NGSS require the use of a variety of formats.

The *Framework* and the *NGSS* are intended to advance a vision of scientific literacy in which students gain the intellectual tools needed to make sense of scientific phenomena in the world around them.

By the end of the 12th grade, students should have gained sufficient knowledge of the practices, crosscutting concepts, and core ideas of science and engineering to engage in public discussions on science-related issues, to be critical consumers of scientific information related to their everyday lives, and to continue to learn about science throughout their lives. They should come to appreciate that science and the current scientific understanding of

the world are the result of many hundreds of years of creative human endeavor (NRC, 2012: pg. 9)

Assessments can make explicit the kind of scientific reasoning that students should be prepared to do and can provide insight into students' progress toward that goal. Assessments, therefore, should present the opportunity for students to perform activities relevant to investigating phenomena and solving problems such as devising methods to collect and analyze data, using models to evaluate their analyses, and make claims and justify their responses. Clusters of short response tasks (Figure 20) can provide brief glimpses of how students conduct these activities, and in some settings these are the closest approximation of students' reasoning about phenomena and problems that is feasible.

A scientist performs two investigations. Before the investigations, she determines the characteristic properties and molecular formula of each of the starting substances. In the first investigation, she mixes liquid nitrogen with liquid water. She observes a gas form and collects samples of all the ending substances. In the second investigation, she places a piece of solid lithium in liquid water. She observes a gas form and collects samples of all the ending substances. She determines the characteristic properties and molecular formula of the ending substances. The table below summarizes her findings.

Investigation #1				
	Substance	Boiling Point	Flammable	Molecular formula
Starting Substances	Liquid Nitrogen	-196°C	No	N ₂
	Liquid Water	100°C	No	H ₂ O
Ending substances	Gas 1	-196°C	No	N ₂
	Liquid 1	100°C	No	H ₂ O
Investigation #2				
	Substance	Boiling Point	Flammable	Molecular formula
Starting Substances	Solid Lithium	1342°C	Yes	Li
	Liquid Water	100°C	No	H ₂ O
Ending substances	Solid 2	924°C	No	LiOH
	Gas 2	-253°C	Yes	H ₂

- 1) Did a chemical reaction occur during either of the investigations?
 - A. A chemical reaction occurred in both investigations.
 - B. A chemical reaction occurred only in the first investigation (nitrogen and water).
 - C. A chemical reaction occurred only in the second investigation (lithium and water).
 - D. A chemical reaction did not occur in either investigation.
- 2) Explain what a chemical reaction is and describe what indicators can be used to determine whether or not a chemical reaction occurred during these investigations.

Figure 20. A two-part task that asks students to analyze data from two investigations and to write an explanation for their reasoning. Reproduced with permission from the AAAS Science Assessment project.

But the design of the system of assessments proposed by SNAP (see Osborne et al., 2015) is based on the premise that measurement of the learning goals embedded in the NGSS requires the inclusion of some opportunities for deep, extended reasoning about phenomena. Investigating phenomena is not always straightforward with one logical path to a correct answer, but often encounters problems that must be solved or worked around. Performance tasks can elicit this extended reasoning across the three dimensions, and can provide opportunities to observe students drawing on multiple elements of each dimension as they are needed to solve problems and answer questions. Not all performance tasks fill this role, however, and there are some trends in ways existing performance tasks fail to elicit this kind of reasoning from students.

Students often need guidance when they investigate and solve problems, but many existing performance tasks set such a prescribed path for them that they essentially measure students' ability to follow the directions without the reasoning and decision-making essential to these activities (Duschl and Bybee, 2014). In the example shown in Figure 21, students follow a five-step procedure for collecting data to fill in a table. They make a prediction in the middle of the experiment and describe their observations at the end. Students are not asked to make decisions about how to collect these data, but they are also not given enough context to consider what scientific phenomenon is being investigated and what question about that phenomenon they might seek to answer. The goal of *planning and conducting investigations* is not to train all students to become lab scientists, but it is to “help students become more critical consumers of scientific information” (NRC, 2012: pg. 41). A performance task aligned to the NGSS might involve students in considering what a valid and reliable test of a hypothesis might look like and identifying sources of uncertainty that they notice as they conduct an investigation.

You are planning to spend the day at the park with your friends. Since it is a hot day, you want to serve drinks that will stay cool. You have three types of containers you can bring. You have paper cups, Styrofoam cups, and metal cans. You aren't sure which one to choose, so you decide to test which type of container will keep liquid the coldest for 15 minutes.

Work in groups of four.

- Step 1. Set out the three types of containers and put a thermometer in each one.
- Step 2. Using the measuring cup, put equal amounts of ice water in each container. This amount should be enough to fill each container at least three-quarter full.
- Step 3. In the second column of the data sheet below, record the temperature of the water in each container after 30 seconds. Use the Celsius ($^{\circ}\text{C}$) scale for all of your temperature measurements.

	Temperature after 30 seconds (Step 3)	Temperature after 15 minutes (Step 4)	Total difference in temperature (Step 5)
Styrofoam cup			
Paper cup			
Metal can			

Now go back to your desk and answer Question 1, on the next page. After you answer the question, go back to the experiment and finish it. Look to see what time it is so you can return to your experiment when 15 minutes are up.

- Step 4. After fifteen minutes have passed, record the temperature of the water in each container.
- Step 5. Subtract the beginning temperature from the temperature that was taken after fifteen minutes. Record the difference. This is the amount that the temperature changed.

Figure 21. A performance task that provides step-by-step directions for students to follow to perform an investigation from the PALS project. Reproduced from pals.sri.com: permission pending.

Computer-based assessments offer promise as an avenue for engaging students in science and engineering practices without the burden of time and equipment required by most hands-on investigations. Figure 22 shows a typical computer-based performance task, which poses a question and asks students to run an experiment to collect data. But, the decisions students can make are constrained to drop-down menus and students can quickly run all possible variations of the experiment so that they do not have to make any decisions about which combinations of variables are relevant. Similarly, drawing the graph from the data table is constrained so that the axes are labeled and students click on the area that approximates the amount shown on the data table. Although the task touches on several science practices, the degree of engagement is often comparable to a multi-step multiple-choice task.

The NAEP computer-based task shown in Figure 23 has a similarly structured process for conducting trials to collect data for several variables related to the height to which a helium balloon can rise. In contrast to the previous task, however, students have to make predictions, analyze patterns in data, and select an appropriate explanation in a series of multiple choice items that require that students make sense of the data they collected.

Questions: 3 - 5 Science Grades 7-8 (2 out of 18) GUEST, GUEST (SSID: GUEST)

Back Next Save Pause

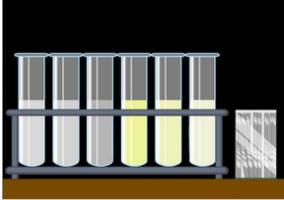
Masking Dictionary Calculator Notes Zoom Out Zoom In

3

Conduct experiments in which a metal is added to a solution and a chemical reaction occurs. You can change the variables in your experiments.

Choose a solution and its concentration to conduct an experiment at room temperature. Then choose a length of time to observe the reaction. The reaction produces a gas that will be collected in the syringe.

Click Start to begin the experiment. Data will appear in the table.



Solution **A**

Concentration **Weak**

Time (sec) **10**

Start

Clear All Rows

Solution	Concentration	Time (sec)	Gas Volume (mL)

4

Create a bar graph to show how the concentration of solution A affects the amount of gas produced.

Click on lines on the graph to indicate where the top of each bar should be.

- Round your answer to the nearest fifth.
- To remove your selection click on the graph again.

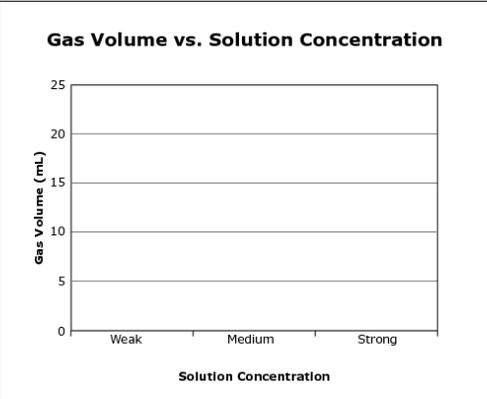


Figure 22. An example of an 8th grade online interactive task from Utah Dept. of Education in which students run trials using drop-down menus and create a bar graph based on the outcomes. Reproduced from schools.utah.gov.

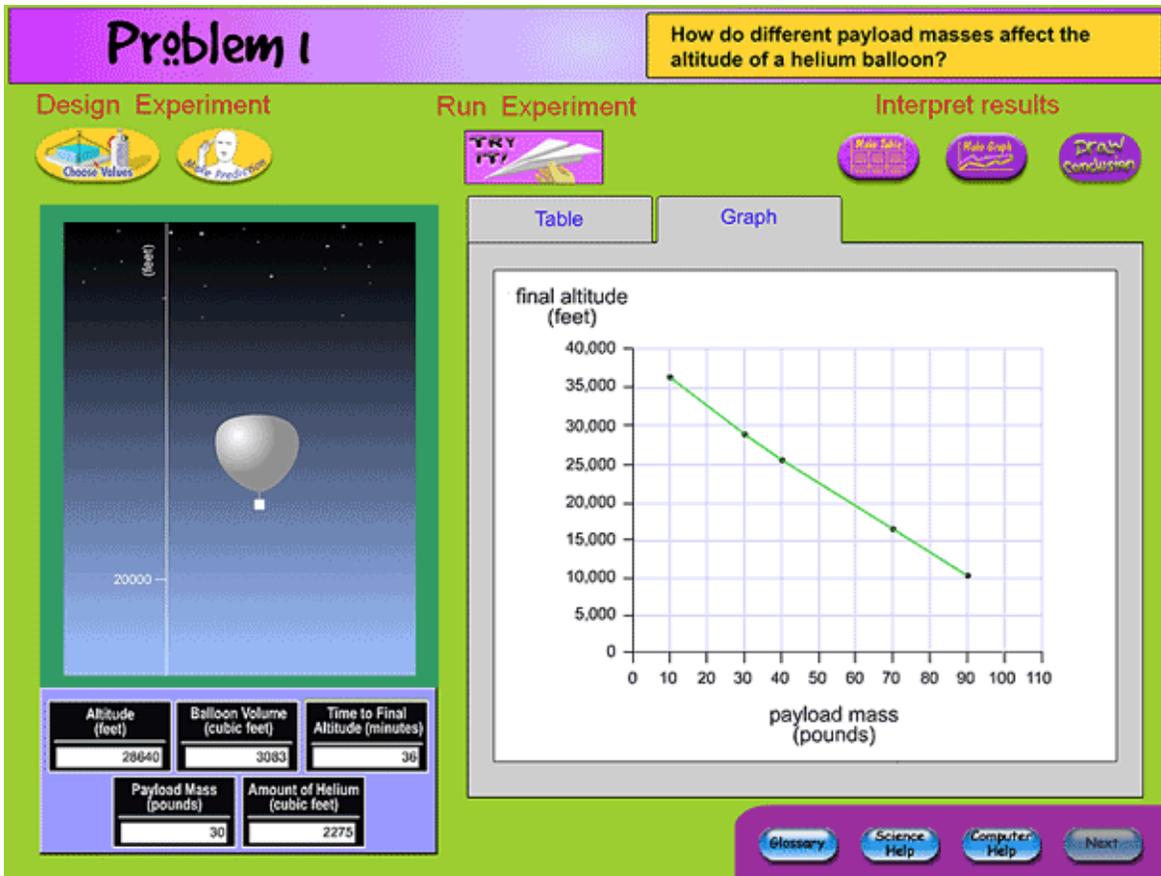


Figure 23. One screen from a NAEP simulation-based task in which students run a simulation to collect data for three closely related investigations about a helium balloon. Students select variables to test, run trials, and analyze a data table and graph to draw conclusions about observed patterns. Reproduced from nces.ed.gov.

One approach to balancing the need between giving students some structure to guide their investigation and giving student the opportunity to use science practices to reason with evidence is to separate the design of the investigation and the analysis of the results. For example, Figures 24a-d shows an item in which students are given a constrained scenario and are asked to describe a problem and an investigation to address the problem. After students describe and conduct their own investigations, they are presented with a data table from another “student’s” investigation and must analyze the results (Figure 24c), critique the design of the investigation, and draw conclusions based on the analysis (Figure 24d).

This performance task from Connecticut draws on five science practices as they are needed to reason through observations made from simple investigations around physical science content. An open-ended investigation in the beginning of the task enables students to become familiar with the problem space, but the rest of the task was standardized with a common investigation and data set, making sure that students’ performance on the rest of the practices being probed is not dependent on their investigation.

Fire Extinguisher

Some fires can be extinguished by smothering them with carbon dioxide gas (CO_2). A company is designing a fire extinguisher that uses the chemical reaction between vinegar and baking soda to produce carbon dioxide. Since the fire extinguisher must produce the gas quickly in order to put out a fire, the designers need your help in studying variables that affect how much carbon dioxide this reaction produces in a certain amount of time.

There are several variables that may affect the rate of carbon dioxide production in the fire extinguisher, such as the amount of baking soda, the concentration of vinegar solution, and the temperature of the vinegar solution. You will investigate two of these variables using a plastic bottle as a model fire extinguisher.

Your model fire extinguisher should only hold a maximum of 10 cc (cubic centimeters) of vinegar solution. Note: 1 cc = 1 mL.

Your task:

Part I: You and your partner will design and conduct an experiment to determine how the *amount of baking soda* affects how much carbon dioxide is produced in a *certain amount of time*.

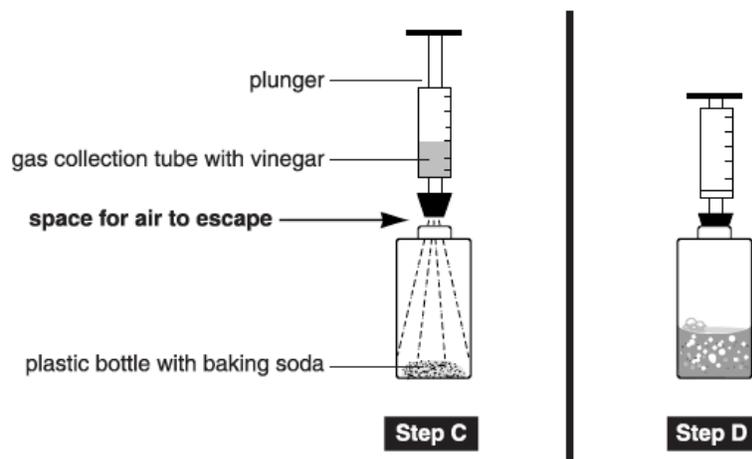
Part II: You and your partner will design and conduct an experiment to determine how *another variable* that you choose affects how much carbon dioxide is produced in a *certain amount of time*.

24a)

PART I

1. **State** the problem you are going to investigate. Clearly identify the *independent* and *dependent* variables that will be studied. Write your problem statement on page 5.
2. **Design** an experiment to solve the problem. Your experimental design should match the statement of the problem, should control for variables and should be clearly described so that someone else could replicate your experiment. Use a control and perform multiple trials, as appropriate. Write your experimental design on page 5.

Use the diagram below to help you set up your experiment. **Remember, your model fire extinguisher should only hold a maximum of 10 cc of solution. Note: 1 cc = 1 mL.**



24b)

Group B did not include a control in their experiment. What would be an appropriate control? Explain your answer fully including how the control would improve their experiment.

Write your answer in your answer booklet.

What conclusion can be drawn from Group B's experiment and results? Explain how valid you think this conclusion is.

Write your answer in your answer booklet.

24c)

Group B carried out the following experiment.

1. Make up solutions of 100%, 75%, 50% and 25% vinegar.
2. Place baking soda in a plastic bottle.
3. Add different concentrations of vinegar to the bottle.
4. Measure how much carbon dioxide gas is collected in 20 seconds.

Our results:

Concentration of Vinegar	Amount of Baking Soda	Amount of Carbon Dioxide Collected in 20 Seconds
100%	2 scoops	42 mL
75%	2 scoops	28 mL
50%	2 scoops	16 mL
25%	2 scoops	10 mL

24d)

Figure 24a-d. Excerpts from a grade 9-10 curriculum-embedded performance task from the Connecticut Department of Education that incorporates opportunities for students to follow directions to conduct an investigation, then use that procedure to design a new investigation. Students also answer several open-ended (24d) and multiple-choice questions in which they graph and analyze data, and evaluate experimental designs and analyses. Reproduced from sde.ct.gov.

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 [Next Generation Molecular Workbench](#) a collection of simulations that have been used in assessments such as Figure 25 from the [Interactions](#) project (Figure 25). Interactions⁶ embeds these simulations into activities that blend learning about the behavior of atoms and molecules with assessment questions that ask students to demonstrate that they can

⁶ Interactions is a collaboration between the Concord Consortium, the CREATE for STEM Institute at Michigan State University, and the University of Michigan.

use evidence from the simulations to perform scientific practices and to use microscopic molecular behavior to explain macroscopic phenomena.

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

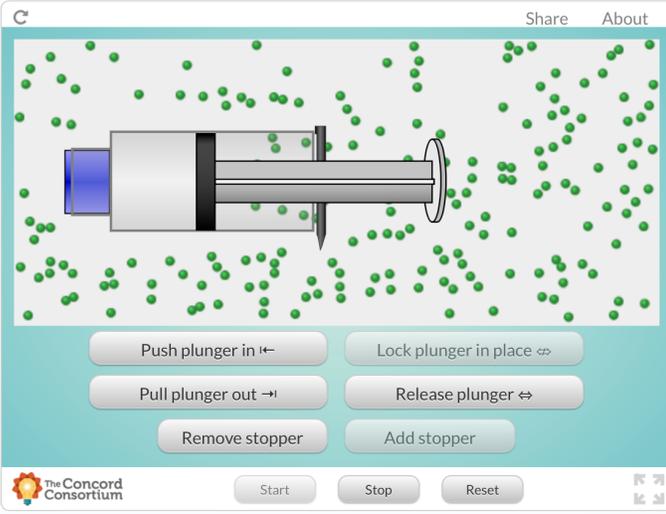
Revisit your initial model of a gas (the first question of this activity). Do the components of your initial model explain your observations of gas being compressed in the syringe? If not, what revisions would you make to your model?

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



The simulation interface features a central window with a syringe and green particles. The syringe has a blue stopper on the left and a plunger in the middle. The particles are scattered throughout the cylinder. Below the window are several control buttons: 'Push plunger in ←', 'Pull plunger out →', 'Lock plunger in place ↔', 'Release plunger ↔', 'Remove stopper', and 'Add stopper'. At the bottom, there are 'Start', 'Stop', and 'Reset' buttons, along with a logo for 'The Concord Consortium' and a 'Share About' link.

Figure 25. 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/>.

WISE (Web-based Inquiry Science Environment) takes a different approach to computer-based curriculum-embedded assessment (Figure 26). WISE shows how entire computer-based curricula can incorporate simulations, videos, assessment, and a broad range of assessment formats from multiple-choice to drafting a persuasive letter to a congressional representative. In WISE, simulations are seamlessly embedded in learning experiences and formative and summative assessment; students make predictions or describe what they know about the phenomena being represented before they engage with the simulation, answer questions and complete short digital performance assessments interspersed throughout the simulation, and demonstrate their learning after using the simulation by completing questions, writing assignments, oral presentations, and other assessment products.

What would your general strategy be if you needed to design critters that would outcompete all other critters in a specific environment?

You will test multiple designs to see if you can create a population that out-competes all your classmates' designs. Every student will be given up to 10 tries to test a new critter design (and remove an old one) within a 5 minute competition.

In this competition, environmental conditions for growing grass will not remain constant. They will be changed randomly throughout the competition.

The screenshot shows a simulation interface for designing critters. On the left is a control panel with various settings and buttons. At the top, there are three toggle switches: 'gray-out-others?' (Off), 'follow-a-critter?' (Off), and 'show-energy?' (Off). Below these are input fields for 'Your critter shape: cyan heart' and 'behavior-dna: FFFFL'. There are sliders for 'speed' (1.00) and 'birthing-level' (25), and a checkbox for 'carnivore?' which is checked. A 'Place New Species' button is present, along with a warning message: 'WARNING: placing a new species will remove all your current critters.' Below the warning are fields for 'Countdown until you can place again: 0' and '# of species left you can introduce: 10'. At the bottom, there are several data boxes: 'Your Current Species Longevity: 0', 'Your Max. Species Longevity: 0', '# Alive Species: 0', '# Extinct Species: 0', 'Your Current Species Size: 0', 'Species Longevity Leader: -----', 'Longevity Record: -----', and 'time: 216'. The central area is a green field representing the simulation environment.

1. Make a Prediction: Do you think you will be able to design a population that will outcompete all the other populations in the ecosystem for the entire time?
2. What evidence do you find in the model to determine whether one of your designs out-competed all others?
3. Why wasn't every population you designed and tested equally successful at competing against the other populations that were randomly created by the computer?
4. What is the one big idea that you have discovered in this activity?

Figure 26. An excerpt from a high school task from the Web-based Inquiry Science Environment (WISE) embedded in a population biology unit which students set the parameters on a simulation of population that will compete with other students' populations and use evidence from the simulation to answer questions about the dynamic interactions in this ecosystem. Reproduced from wise.berkeley.edu.

In the task shown in Figure 26, students who are at the end of a unit on population biology, in which they have learned about ecosystem modeling, competition between individuals and between populations, and principles of fluctuation and stability, use what they have learned to engage in several science practices. Students design an organism that with the goal of outcompeting other students' organisms for a resource (grass). After students run the model, they use it to analyze their data and construct explanations and arguments about the observed phenomena.

Well-designed performance tasks can probe much more deeply into students' reasoning by requiring students 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).

Summary and conclusions

Numerous existing assessments were found to align at least in part to the goals laid out in the NGSS. Even though they are incompletely aligned to the new standards, they provide diverse examples of ways to probe some of the parts of the NGSS that are less familiar to most assessment developers, such as many of the science and engineering practices and cross-cutting concepts. This review has identified examples of assessment tasks aligned to each of the three dimensions, although some parts of the standards have fewer examples than others such as the practice *asking questions and defining problems* and all but one of the cross-cutting concepts, *cause and effect: mechanism and explanation* occurred.

Several themes emerged from the landscape review that point to common characteristics of existing assessments that are incompatible with the needs of a new assessment system. Tasks designed for a new assessment system must: 1) align with each of the three dimensions of the NGSS; 2) focus on big ideas in science; 3) probe science and engineering practices in a way that engages students in reasoning with evidence; and 4) give students a platform in which they can draw on their knowledge and skills as needed to investigate scientific questions and problems.

The landscape review of promising assessments revealed that most existing assessments are either not well aligned to the learning goals outlined in the NGSS, or do not advance the vision of the framework in one of the ways outlined above. At the same time, it also revealed that there are many assessments that fulfill at least one major goal of the NGSS and can provide a basis to guide the development of new, fully-aligned assessments.

There are opportunities and challenges ahead as assessment designers struggle to develop tasks that test the performance expectations in NGSS. But, it is critical to undertake this struggle, because in the end, the intent of the standards is read by teachers not from the framework or the standards themselves, nor even from curriculum materials developed to present these ideas, but from the assessments. Moreover, the assessments that teachers use as their guides are chiefly those used by or supported by the states. Thus working to make

assessments that truly reflect the vision of reform outlined in the *Framework* and implemented in NGSS, while challenging, is an essential task in realizing this vision.

Please note that the sources that provided permission to reproduce items in this report do not necessarily share the views expressed in the report.

Item sources

Figure 1a. “Photosynthesis.” *Web-based Inquiry Science Environment (WISE)*, Graduate School of Education, University of California, Berkeley. Retrieved August, 2015, from <http://wise.berkeley.edu/previewproject.html?projectId=7709>.

Figure 1b: Massachusetts Department of Elementary and Secondary Education (2012) Massachusetts Comprehensive Assessment System (*MCAS*). Retrieved August, 2015 from doe.mass.edu/.

Figure 4. The College Board (2015). “Environmental Science.” The College Board, New York.

Figure 5. Ohio Department of Education (2006). Ohio Achievement Tests: Science. Retrieved April, 2015.

Figure 6. Alaska Department of Education (2012). Alaska Standards Based Assessment for Science. Retrieved April, 2015 from www.eed.state.ak.us.

Figure 7. “Environmental Systems and Societies.” International Baccalaureate, 2010.

Figure 8. Next Generation Science Assessment (n.d.). Retrieved 2015, August 24 from <http://concord.org/projects/ngss-assessments>.

Figure 9. Daehler, K., and Folsom, J. (2014). Making Sense of SCIENCE Force & Motion: Formative Assessment Task Bank. San Francisco: WestEd. Retrieved from: <http://www.WestEd.org/mss>

Figure 10. International Association for the Evaluation of Educational Achievement (IEA)(2013). TIMSS 2015 Assessment Frameworks. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Figure 11. Massachusetts Department of Elementary and Secondary Education (2011) Massachusetts Comprehensive Assessment System (*MCAS*). Retrieved August 24, from <http://www.doe.mass.edu/mcas/search/>.

Figure 12. Assessments of Argumentation in Science: bubbles in water. Retrieved August 24, 2015 from Scientificargumentation.edu.

Figure 14. Enger, S. K., & Yager, R. E. (1998). The Iowa Assessment Handbook. Iowa City: Science Education Center, University of Iowa.

Figure 15. AQA (2011). Science in Society A-level exam. AQA, U.K.

Figure 16. The Organisation for Economic Co-operation and Development (2009). Take the test: sample questions from OECD’s PISA assessments. Retrieved August 24, 2005 from <http://www.oecd.org/pisa/pisaproducts/pisatakethetestsamplequestionsfromoecdspisaassessments.htm>.

Figure 17. International Association for the Evaluation of Educational Achievement (IEA)(2013). TIMSS 2015 Assessment Frameworks. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Figure 18. from Cartier, 2000.

Figure 19. Achieve (n.d.). Classroom Sample Tasks: Solar Cooker. Retrieved August 24 from http://www.nextgenscience.org/sites/ngss/files/HS-PS-Physics-SolarCooker_version2.pdf.

Figure 20. AAAS. (2015). Chemical reactions assessment item. Unpublished, Project 2061, Washington, DC.

Figure 21. SRI (n.d.). Performance Assessment Links in Science. Retrieved August 24 from <http://pals.sri.com/tasks/k-4/KeepCool/directs.html>. (contacted 12/1)

- Figure 22. Utah Dept. of Education (n.d.). Student Assessment of Growth and Excellence. Retrieved August 24 from <http://sageportal.org/training-tests/>.
- Figure 23. NCEES (n.d.). National Assessment of Educational Progress. Retrieved August 24 from <https://nces.ed.gov/nationsreportcard/studies/tba/tre/sim-description.aspx>
- Figure 24. Connecticut Department of Education (2004). Connecticut Aptitude Performance Test.
- Figure 25. The Concord Consortium (n.d.) Next Generation Molecular Workbench. Retrieved 2015, August 24 from <http://mw.concord.org/nextgen/#activities>.
- Figure 26. “How do populations change?” *Web-based Inquiry Science Environment (WISE)*, Graduate School of Education, University of California, Berkeley. Retrieved from <http://wise.berkeley.edu/previewproject.html?projectId=11346>, August 2015

References

- AAAS Project 2061. (1993, 2008). *Benchmarks for Science Literacy*. Washington, DC: American Association for the Advancement of Science.
- Achieve, Inc. (2002). *Staying on course: Standards-based reform in America's schools: Progress and prospects* [On-line]. Available: <http://www.achieve.org/files/5YearReportfinal.pdf>,
- Achieve, Inc. (2013). *Next Generation Science Standards*. Achieve, Inc.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Cartier, J. (2000). Assessment of explanatory models in genetics: Insights into students’ conceptions of scientific models. *University of Wisconsin-Madison*, 98(1).
- 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>.
- Enger, S. K., & Yager, R. E. (1998). *The Iowa Assessment Handbook*. Iowa City: Science Education Center, University of Iowa.
- Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., & Linn, M. C. (2015). Automated Guidance for Student Inquiry. *Journal of Educational Psychology*. Advance online publication. <http://dx.doi.org/10.1037/edu0000052>
- 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.
- Milano, M. (2013). The Next Generation Science Standards and Engineering for Young Learners: Beyond Bridges and Egg Drops. *Science and Children*, 51(2), 10.
- National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.
- National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academies Press.
- National Research Council. *Engineering in K-12 Education: Understanding the Status and Improving the Prospects*. Washington, DC: The National Academies Press, 2009.
- National Research Council. (2009). *Engineering in K-12 Education: Understanding the Status and Improving the Prospects*. Washington, DC: National Academies Press.
- National Research Council. (2011). *Assessing 21st Century Skills*. Washington, DC.
- 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.; 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.
- Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, 340(6130), 320–323.
- Roseman, B. J. E., & Koppal, M. (2008). Using National Standards to Improve K–8 Science Curriculum Materials. *The Elementary School Journal*, 109(2), 104–122.
- 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.