Thinking About Claims for the Next Generation Science Standards

Mary Norris, Virginia Tech

Brian Gong, Ph.D., National Center for the Improvement of Educational Assessment

Paper presented at National Council for Measurement in Education Conference

Toronto, Canada

April 2019

During the past 50 years there has been a sustained call for improved science, technology, engineering, and mathematics (STEM) education in the US. Reports recommend that the federal government should support the development of strong math and science standards (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007; National Commission on Excellence in Education, 1983; Rutherford & Ahlgren, 1990). Work toward these goals has most recently resulted in the creation of the *Next Generation Science Standards: For States, By States* (*NGSS*) (NRC, 2013), a set of standards that has been adopted in some form by over 38 states and territories to date. States are just beginning to create assessments for these new standards. In this paper, we describe how the NGSS differ from previous standards and the challenges for assessments designed to measure them.

The Standards

The structure of the NGSS is different from the structure of earlier science standards. The standards consist of a set of *performance expectations* (*PEs*), which are assessable, three-dimensional statements about what students should know and be able to do. Other science standards are assessable statements, but they are not intentionally three dimensional (NRC, 2013). Each performance expectation of the NGSS is a combination of three dimensions: a disciplinary core idea (DCI), a science and engineering practice (SEP), and a cross-cutting concept (CCC). This explicit integration of content and process is very different from traditional science standards in which content and practices are usually listed separately. Next, we explain the three components of the NGSS in greater detail.

Disciplinary core ideas are key concepts of science and engineering that are necessary for a deep understanding of a wide variety of phenomena within a multitude of disciplines. They are, in essence, the "big ideas" of science and engineering. They are arranged by the three main scientific disciplines (Life Science, Earth and Space Science, and Physical Science) and Engineering Design. There are twelve very broad disciplinary core ideas (e.g. Ecosystems: Interactions, Energy, and Dynamics; Earth's Systems; Energy; and Engineering Design) each of which is defined across subcategories (e.g. Definitions of Energy; Conservation of Energy and Transfer of Energy; Relationship Between Energy and Forces; and Energy in Chemical Processes and Everyday Life) and across four grade spans. This results in a much larger collection of statements explaining how students are expected to apply these ideas to different phenomena and at different stages of academic development. *Science and engineering practices* are statements about what scientists and engineers do. They are eight activities that describe the process of science. The seven *cross-cutting concepts* are ideas and ways of thinking that allow students to link ideas across scientific disciplines. Figure 1 lists very general statements of the twelve broad DCIs, the eight main SEPs, and the seven CCCs (NRC, 2013). Similarly to DCIs, both SEPs and CCCs are defined in greater detail in terms of what students at different grade levels should be able to do.

Disciplinary Core Ideas	**Science and Engineering Practices**	**Cross-cutting Concepts**

LS1. From Molecules to Organisms: Structures and Processes	SEP1. Asking questions and defining problems	CCC1. Patterns
LS2. Ecosystems: Interactions, Energy, and Dynamics	SEP2. Developing and using models	CCC2. Cause and Effect
LS3. Heredity: Inheritance and Variation of Traits	SEP3. Planning and carrying out investigations	CCC3. Scale, Proportion, and Quantity
LS4. Biological Evolution: Unity and Diversity	SEP4. Analyzing and interpreting data	CCC4. Systems and System Models
ESS1. Earth's Place in the Universe	SEP5. Using mathematics and computational thinking	CCC5. Energy and Matter
ESS2. Earth's Systems	SEP6. Constructing explanations and designing solutions	CCC6. Structure and Function
ESS3. Earth and Human Activity	SEP7. Engaging in argument from evidence	CCC7. Stability and Change
PS1. Matter and Its Interactions	SEP8. Obtaining, evaluating, and communicating information	
PS2. Motion and Stability: Forces and Interactions		
PS3. Energy		
PS4. Waves and Their Applications in Technologies for Information Transfer		
ETS1. Engineering Design		
Figure 1. Components of the NGSS		

Each PE is a combination of a DCI, an SEP, and a CCC. An example of a PE for middle school physical science is MS-PS3-1: "Construct and interpret graphical displays of data to describe the relationships of kinetic energy to the mass of an object and to the speed of an object" (NRC, 2013, p.61). Figure 2 shows how this PE is related to its components. Where older science standards may have specified that students would be able to define kinetic energy and state how it is related to the mass and speed of an object (DCI PS3.A) and that students will be able to make sense of linear and nonlinear graphs (SEP 4), the PE explicitly states how students will use linear and nonlinear graphs and the concept of proportionality to describe these relationships. Older standards tended to assess content knowledge and skill competency separately. The NGSS always integrate the two.

The dimensions of the NGSS should not be separated for assessment purposes (NRC, 2013). To ask a student to state the law of conservation of energy (part of DCI PS3) or read a graph (one aspect of SEP4) is not in the spirit of the NGSS. To ask a student to find patterns in data using a graph and to interpret these patterns in terms of conservation of energy would be, however, because it would call upon the student to use aspects of DCI PS3, SEP5, and CCC1 concurrently. In fact, the eight SEPs are called "practices" rather than "skills" to emphasize this difference—that "engaging in scientific investigation requires not only skill, but also knowledge

that is specific to each practice" (NRC, 2012, p.30). All PEs are three dimensional by definition. This might lead one to conclude that the most logical domain for an NGSS assessment is all of the PEs. However, as we will show in the next section, this practice may prove problematic.

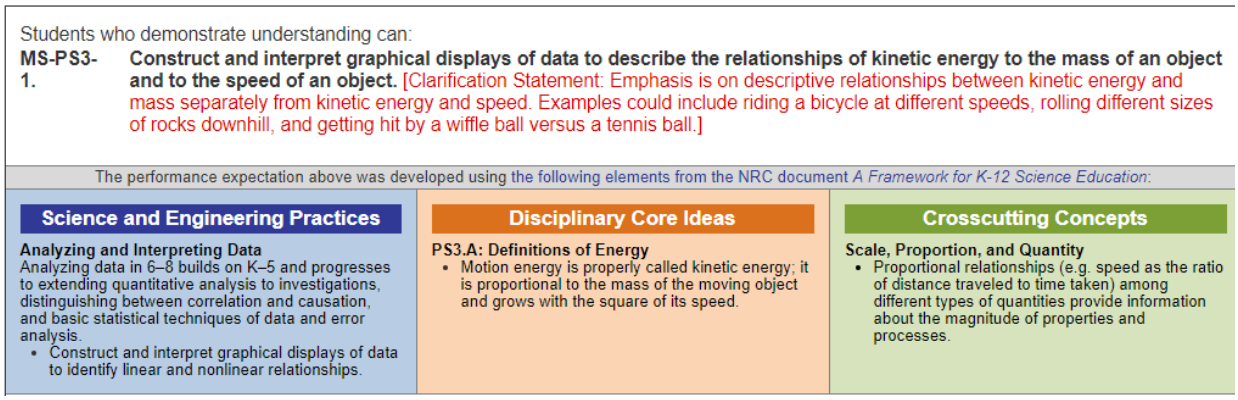

Figure 2. Example of NGSS Performance Expectation. Reprinted from Next Generation Science Standards, by National Research Council, 2013. Retrieved from https://www.nextgenscience.org/.

Defining the Domain of the NGSS for Assessment

There are 208 PEs in the NGSS—33 for grades K-2, 45 for grades 3-5, 59 for grades 6-8, and 71 for grades 9-12. If one were to list all possible combinations of SEPs and DCIs across the most generally stated DCIs listed in Fig. 1, the results would be 448 possible outcomes. The PEs, however, are composed from much finer grained aspects of the DCIs such as the one shown in fig. 2. There are 38 of these more-detailed statements for grades K-2, 60 for grades 3-5, 91 for grades 6-8, and 97 for grades 9-12. Combining these with SEPs and CCCs results in over 16,000 possible combinations. The PEs of the NGSS are slightly more than one percent of these possible combinations. While the PEs were intentionally selected to represent what students should be able to do by the end of each grade span, they are randomly and sparsely scattered among the matrix of possible combinations. The small number and random representation of the PEs as well as their three-dimensional nature mean that choices about assessing the NGSS may require different considerations than those for assessing traditional standards. Some consideration should be given to how the domain for assessment will be defined and how it will affect the types of claims that can be made about students.

There are multiple practical considerations that will affect how states define the test domain. First, the state will need to decide how many assessments the it will administer each year, at what grades they will be administered, and whether they will be cumulative across grades or grade spans. Currently, the federal requirement is that states administer a minimum of three tests per year, one in each grade span—elementary, middle, and high school. Each of these tests can cover only the PEs assigned to the year/course at which it is administered (e.g. Grade 3 or Middle School Biology), all PEs for the grade span (e.g. all Grade 3-5 PEs), or even all PEs which have been covered up to that year (e.g. all Grade K-5 PEs). Second, the state should decide whether the PEs will comprise the test domain or if some other combination of DCIs,

SEPs, and CCCs will be used to define it. Next we present three units of analysis which states might use to define the test domain along with how each affects the types of claims that can be made based on the test results.

The first (and most obvious) approach to defining the test domain is to use PEs as the unit of analysis. California is an example of a state which is approaching NGSS assessment this way. States which use PEs as the unit of analysis will need to decide how the PEs relate to the domain of science and how many and which PEs will be used to define this. For instance, a state could define the domain of the assessment as all of the PEs in the standards or some subset of all PEs. If a state decides to assess a cumulative set of PEs for a grade span, then it will be challenging for the state to assess all of the PEs in a single assessment. For example, the smallest number of PEs in a commonly assessed grade span is 45 PEs in grades 3-5. Allocating only two items per PE (every state's design that we have seen allocates more than two items to assess a PE) would result in an assessment with 90 items. This is more than almost any science assessment we are aware of. A solution to this is to define the domain of the assessment across a single grade/course or to define it as a subset of the PEs for the grade span, grade, or course. Due to the sparse and random representation of the PEs across the entire matrix of possible combinations of DCIs, SPEs, and CCCs, even the entire set of PEs for a grade/course may fail to provide the basis for collecting evidence to support a cohesive claim about what students know and can do. Choosing a subset of PEs to assess may narrow specific claims or require general claims to be even more general. For state which choose PEs as the unit of analysis, claims about the domain of science to be assessed will be all or some of the PEs and claims about students will be about their performance within this defined domain of science.

A second possible approach to defining the test domain is to use PE clusters. A PE cluster is two PEs which are assessed together and in which the two DCIs, SEPs, and CCCs may be mixed together to create combinations that are not included in the NGSS (see Fig. 3).

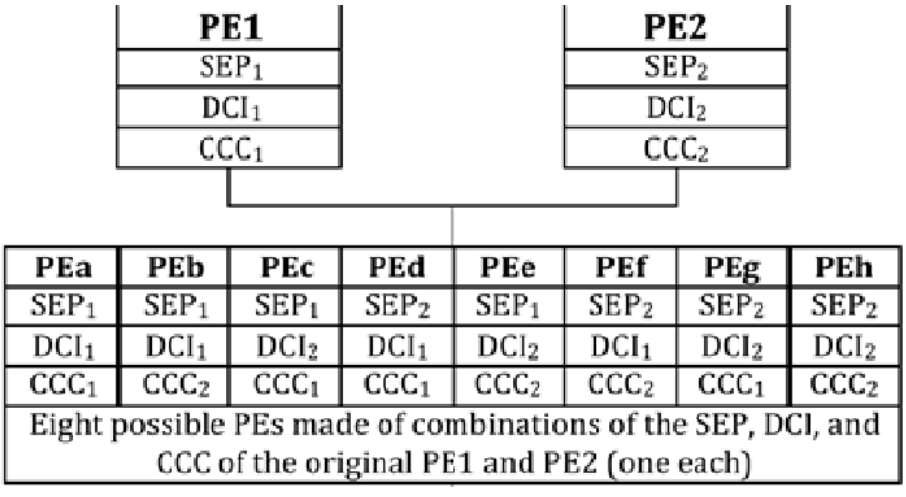

Figure 3. A PE cluster created by combining PE1 and PE2.

In this approach, the assessment domain may include both PEs that were defined in the NGSS (PE1 and PE2) and a number of additional PEs made up of the parts of the original two PEs. Washington is an example of a state that has defined the domain through PE clusters. The domain—as defined by the PEs' possible cross-combinations—is much larger in the possible combinations of two PEs than the domain defined by two individual PEs. States which use this approach, then, will likely be unable to assess all PEs. However, this approach may increase the cohesiveness of the domain of science for assessment because the same DCIs are approached with different SEPs and CCCs or the same SEPs and/or CCCs are applied to multiple DCIs. Some possible claims associated with this definition of the NGSS domain are:

- "The domain of science to be learned and assessed is not just the two individual PEs drawn from the NGSS, but also all the other possible combinations of SEPs, DCIs, and CCCs taken together in combinations of at least one of each."
- "The student has been assessed on all possible three-dimensional combinations associated with the domain definition."

The second claim is of special interest. It may be that students will not have been instructed on combinations that are not in the NGSS. If this is the case, then states might claim that students are able to generalize by applying the SEPs and/or CCCs to new DCIs.

The final approach we discuss for defining the test domain is to use the SEPs as the unit of analysis. This approach might be useful for states which want to assess a student's ability to plan, conduct, and communicate about an investigation across a single DCI. This approach is illustrated in Fig. 4. Kentucky is an example of a state that is considering such an assessment design. Because the NGSS include no such combinations, it is likely that states that choose this method to define the domain of assessment will be testing students on combinations that have not been directly taught. A possible claim associated with this definition of the NGSS domain is:

- "The student has demonstrated ability to apply [all eight of] the Scientific and Engineering Practices with important DCI and CCCs to scientifically investigate and solve important problems."

SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8
DCI1	DCI1	DCI1	DCI1	DCI1	DCI1	DCI1	DCI1
CCC1	CCC1	CCC1	CCC1	CCC1	CCC1	CCC1	CCC1
An example set of PE associated with an investigation involving all eight SEP for the same DCI and CCC.							

Figure 4. Three-dimensional assessment targets created by applying all SEPs across the same DCI and CCC. This method creates a more cohesive set of targets than choosing from existing PEs.

Defining the domain of an assessment is important to ensure that it will elicit the evidence needed to support the claims that will be made. Modern test validity theory views

assessments as providing evidence to support claims or interpretations about students in relation to a domain. Validation may be thought of as making a set of claims based on assessments, being clear what the claims are about, and why those claims might be reasonable and then the process of assembling and evaluating evidence about the degree to which those claims are supported (Kane, 2006). Must of the evidence for validation requires data from operational use, but much of the evidence can also be gathered during test design and development. "Evidence-centered design" is a perspective and test-development approach that states that tests should be designed intentionally to provide sufficient evidence to support the intended interpretations/claims and uses (Mislevy, Steinberg, & Almond, 1999). This includes intentional alignment of the test domain and the intended claims that will be supported. The NGSS lend themselves to multiple interpretations of how to use the PEs and underlying three-dimensional components to define the domain of science to be learned and assessed. The claims about the domain definition will shape the intended interpretations, test design, and validation focus. States should carefully formulate their claims regarding the NGSS domain definition.

Describing Test Performance

In both validation and evidence-centered design, it is essential that the test sponsor and test developer have claims clearly stated in enough detail to guide test development, interpretation, and use. Claims in state summative assessments are typically found in at least three places: the definition of the domain; the definition of the construct that is being measured; and the choice of aspects of performance or attribute quality that will be used to describe different levels of achievement. We have discussed issues related to domain definition above and will now discuss aspects of quality.

A common interpretation of test scores is how well students perform within the domain or construct. One of the key choices that must be made in test construction is what aspect or aspects of quality will be used to order the performances. There are many possibilities to choose from. Each option emphasizes different aspects of student performance and will require different evidence to support it. We list some of these aspects of performance in Fig. 5 along with how they might be interpreted as applied to the NGSS.

Aspect of Performance	**Description**	**Relating to NGSS**
Content breadth	Amount of key content or key content standards	Possible implementations include performing the same SEPs or CCCs across more DCIs/SEPs/CCCs
Cognitive complexity	Degree of difficulty of thinking about the content, applying the content's knowledge and skills	This may be defined in terms of the "depth" of the CCCs, the extent of the SEPs, the demand of the DCI, or some combination and may be informed by a novice-to-expert approach.

Degree of Correctness	Number of errors in performance; regularity with which performance can be repeated	The way this is applied depends upon the unit of analysis—SEP, DCI, CCC, or combinations
Degree of Challenge	Difficulty of assessment items which address the same content	This may be informed by within-grade and across-grade learning progressions.
Fluidity	How quickly/automatically person can perform assessment tasks	May imply that either assessment or items are timed
Degree of Independence	Extent to which directions, background information, or other scaffolding are provided	Background information may be provided during assessment
Sophistication of Solution	How expertly does student formulate and solve the problem? To what extent does the student think about general problem-solving techniques (metacognition)?	For the NGSS, this may be reflected in more or less expert versions of CCCs and DCIs, or more or less expert applications of CCCs, DCI, and SEP across problems and disciplines
Figure 5. Common aspects used to differentiate quality of performance		

Two of these aspects of quality—cognitive complexity and sophistication of solution—may be informed by research about the differences between novice and expert thinking in science. We provide a summary of some of these ideas in Figure 6. Science research indicates that novices tend to explain phenomena within a personal framework of theories/ideas which do not agree with and/or are not organized in the same way as the guiding theories/ideas of the discipline while experts are able to explain diverse (i.e. cross discipline and/or cross scale) phenomena correctly within the guiding theories/ideas of a discipline even for problems that are novel (Vosniadou, 2014; Harrison & Treagust, 2001). Another novice-expert difference comes from Kuhn, Amsel, and O'Loughlins' (1988) work on scientific thinking skills which found that part of what differentiates experts from novices is the ability to recognize that they are using theories (perhaps naïve theories) to make sense of a phenomenon. At the most basic level, novices fail to differentiate theories from evidence (Kuhn et al., 1988). As they progress toward mastery, they begin to use theories and models to evaluate evidence and evidence to modify theories and models (first subconsciously, then consciously) (Kuhn et al., 1988). Experts are also able to think about how they apply theory (Kuhn et al., 1988).

Novice	**Expert**
Refer to personal/naive theories when explaining scientific phenomena[1,2]	Explain diverse (i.e. cross discipline and/or scale) phenomena within guiding

	theories/concepts of discipline even for novel problems[1,2]
Frame problem in terms of irrelevant features; approach problem with little direction and do not self monitor progress; unwilling to switch strategies despite inconsistencies/unreasonable results[3,4]	Use underlying theories and relevant concepts to model problem; anticipate multiple outcomes; self monitor for progress, consider what can be learned from errors, and change strategies if needed[3,4]
Fail to distinguish evidence from theories (may be naïve theories) or use evidence and theories/concepts in explanations; may be selective in considering evidence[5,6]	Recognize and use scientific theories and concepts to evaluate evidence AND evidence to evaluate theories[5,6]
Likely to consider small number of random confirming cases sufficient evidence for cause and effect relationship between variables and a small number of random disconfirming cases sufficient evidence for no cause and effect relationship[5]	Recognize the nature of random variability in data and evaluate cause and effect relationships accurately even when data include random variability[5]
No description provided.	Recognize domain-specific paradigms that guide thinking and learning in discipline; work on developing habits of mind which transcend domains[7]

Figure 6. Differences in novice to expert performance of science.

Notes: 1. Vosniadou, 2014

2. Harrison & Treagust, 2001

3. Baxter & Glaser, 1997

4. Schoenfeld, A. H., 1985

5. Kuhn, D., Amsel, & O'Loughlin, 1988

6. Gotwals, A. W., Songer, A. B., and Bullard, L., 2012

7. Halloun, I., 2011

In the prior two sections, we have discussed considerations in defining the test domain and possibilities for differentiating the quality of performance for assessments of the NGSS. Both the test domain and aspects of quality used to differentiate performance must reflect the three-dimensional structure of the NGSS. In state summative assessments, the test domain and aspects of quality are used to create performance level descriptors (PLDs)—claims of what

students know and can do based on test scores. In the next section we describe types of PLDs and provide three examples of PLDs for NGSS to show different ways that these factors may be combined.

Performance Level Descriptors

In addition to the general PLDs associated with score reports, states have begun to develop more-detailed PLDs which they use to guide test development and interpretation (Schneider and Egan, 2014). Without the guidance these PLDs provide, there is a danger that validation will occur after assessments have been designed and administered. Such practice fails to recognize that useful evidence can be gathered during test design and development. Tests should be designed intentionally to provide sufficient evidence to support the intended interpretations and uses (Mislevy, et al., 1999). Next, we review four types of PLDs identified by Schneider and Egan (2014) and then build on this model by expanding the notion of cognitive complexity to aspects of quality and showing how these can encompass the three dimensions of the *Standards* using a novice-to-expert perspective.

Schneider and Egan (2014) describe four types of PLDs which are developed and refined during assessment development: 1) Policy PLDs, 2) Range PLDs, 3) Target PLDs, and 4) Reporting PLDs. These can be considered as stages in PLD development. Range PLDs are developed from Policy PLDs, Target PLDs are developed from Range PLDs, and Reporting PLDs are developed from Range and Target PLDs. When PLDs are created near the beginning and revised throughout the assessment development process, they act as a guide to create a system in which standards, curriculum, and assessment are aligned both with each other and with the claims that a state wishes to make about student learning. This section of our paper describes the four types of PLDs along with the purpose of each.

Policy PLDs are created early in the process of assessment development. A Policy PLD is an initial statement of the claims and/or uses a state wishes to make about student performance at different levels based on assessment scores. For instance, a state may wish to claim that students have mastered a set of science skills, general science content, content within a specific science domain (i.e. life science), ways of thinking, or some combination of these. Establishing what these claims will be at the outset helps to ensure that scores on an eventual assessment will support the inferences the state makes about student performance.

An important purpose of Policy PLDs is to provide general guidance for the test development process. If an assessment is to be intentionally designed to provide evidence for claims and uses, then stating claims is an essential initial step. Policy PLDs may be quite general--much more general than later PLDs. They serve as a starting point for the development of a set of Range PLDs which will be detailed enough to guide curriculum and item development. It is possible that the focus of assessment will evolve as curricula and assessments are developed. In this case, the Policy PLDs may need to be revised. Assessment and PLD development are iterative and interwoven processes.

Range PLDs are more detailed versions of the Policy PLDs. While policy PLDs may only describe performance within general reporting categories (i.e. overall or domain-specific

science knowledge and/or skills), Range PLDs describe observable performances for each level of each standard being assessed. Therefore, a set of Range PLDs is much longer than a set of Policy PLDs. Although student performance exists along a continuum, the assignment of students to PLDs collapses performances into discrete categories. The range PLDs generally describe the performance of a student at the middle of the range for each level.

The purposes of Range PLDs are to guide curriculum and item development, and to allow evaluation of claims about the construct. Range PLDs should provide a clear distinction between the quality of the performance at each level and reflect higher quality performance for higher levels. We describe aspects of performance that are commonly used to distinguish between performance levels in Figure 5. In writing Range PLDs, states choose one or more of these or similar aspects with which to describe performance quality. The choice will help to define what types of evidence are needed to support the claim and thus guide the item specification process.

Target PLDs (which support interpretation of performance at the transitions from one performance level to another) are used to set cut scores. They are based on the Range PLDs. While range PLDs typically describe the performance at the midpoint of a performance level, target PLD's describe the performance at the edges. They answer the question: "What is the minimum performance that a student could have and still be placed at this level of performance?" For example, the Range PLD used to set the cut score between "not proficient" and "proficient" would describe the performance of the lowest scoring "proficient" student or the highest scoring "below proficient" student. Target PLDs will likely be shorter than Range PLDs because they only need to describe the aspects of performance that are necessary to distinguish between students who perform very near the boundary (Schneider and Egan, 2014).

Reporting PLDs are the claims about student performance that are shared with stakeholders. Ideally, these are developed from the Range and/or Target PLDs after cut scores are set (Schneider and Egan, 2014). Note that if the Reporting PLDs differ from the Range or Target PLDs, it should be evaluated whether the construct and intended claims for interpretation and use of assessment results in relation to the construct have changed.

Next, we give three examples of Range PLDs and discuss the implications of the choice of quality indicator for test development, interpretation and use. The first example differentiates PLDs by breadth of knowledge, the second example differentiates PLDs by degree of student independence, and the third PLD differentiates by sophistication of solution. The first two examples are for performance within the DCI PS3.B which deals with conservation of energy and the third example describes performance within the Science and Engineering Practice: Asking Questions and Defining Problems. All examples are written for the grades 9-12 grade band.

<u>PLD Differentiation Example 1: Breadth of knowledge</u>

Our first example, shown in Figure 7, differentiates between quality of performance along a single scale—breadth of knowledge. Note that descriptions of performance refer to what a student can do consistently. None of the descriptions refer to what a student cannot do. Evidence to support this claim must be strong enough to support the interpretation of a consistent

performance. The descriptions are additive—students at the Developing level can do what students at the Novice level can do plus more—and refer to breadth of multiple subdomains of knowledge. The first subdomain is types of energy. Students at higher performance levels are able to identify and represent more types of energy when using the law of conservation of energy to understand a system than students at the level below. The second subdomain is ways to model the conservation of energy in a system. Students at higher performance levels are able to model the conservation of energy conceptually, graphically, *and* quantitatively while those who perform at lower performance levels can only model systems conceptually and graphically.

Novice	Developing	Proficient	Expert
Student is consistently able to:	Student is consistently able to:	Student is consistently able to:	Student is consistently able to:
Identify and represent flow and conservation of matter and mechanical energy within a system conceptually and graphically.	Identify and represent flow and conservation of matter and mechanical and thermal energy within and between non-living systems conceptually and graphically.	Identify and represent the flow and conservation of matter and multiple types of energy (i.e. mechanical, thermal, electric, magnetic) within and between non-living systems conceptually, graphically, and quantitatively.	Identify and represent the flow and conservation of matter and all types of energy (i.e. mechanical, thermal, electric, magnetic, nuclear and chemical) within and between both living and non-living systems conceptually, graphically, and quantitatively.

Figure 7. Example 1--PLD differentiation using breadth of knowledge

This type of differentiation may be most useful when desired claims foreground content, or some other aspect of the construct that can be expressed in discrete chunks, especially if they follow some sort of sequence. For the NGSS, content will usually be expressed as some aspect of the DCIs or the CCCs. Although claims for the NGSS may foreground a single dimension of the NGSS, they should include all three dimensions. Our example is written to describe performance on Crosscutting Concept 5, Matter and Energy at the High School level. Deciding what will be foregrounded in the claims (CCC 5 in this case) at the beginning of the test development process focuses blueprint creation and item development to create close alignment between scores and desired claims. Although our example foregrounds a single dimension of the NGSS, this does not mean that we have ignored the other two. The performances we describe are three dimensional. They require students to use the Crosscutting Concept across multiple DCIs and two SEPs. We unpack the contents of the PLD in more detail in the next paragraph.

This example foregrounds CCC 5, Matter and Energy, within the context of multiple DCIs and SEPs. Figure 8 lists the CCC, DCIs, and SEPs which we intend our example to refer to. The entire text of the CCC and DCIs is included. For SEP 2, Developing and Using Models,

we have included the initial paragraph, but not the detailed bullets and for SEP 5, Using Mathematical and Computational Thinking, we have included the bullets without the initial paragraph. We feel that these excerpts provide enough information to understand Example 1. The full versions can be found in Volume 2 of the NGSS (NRC, 2013). The example does not cover all sections of each dimension. Those sections of the CCC, DCIs, and SEPs which are not included are italicized.

<table>
<tr><td colspan="2">CCC 5—Matter and Energy
The total amount of energy and matter in closed systems is conserved.
Describe changes in energy and matter in a system in terms of energy and matter flows into, out of, and within that system.
Energy cannot be created or destroyed—it only moves between one place and another place, between objects and/or fields, or between systems.
Energy drives the cycling of matter within and between systems.
In nuclear processes, atoms are not conserved, but the total number of protons plus neutrons is conserved.</td></tr>
<tr><td>DCIs</td><td>SEPs</td></tr>
<tr><td>PS1.B Chemical processes are understood in terms of collisions of molecules, rearrangements of atoms, and changes in energy as determined by properties of the elements involved.</td><td rowspan="2">SEP2—Developing and Using Models
Use, synthesize, and develop models to predict and show relationships among variables between systems and their components in the natural and designed world(s).</td></tr>
<tr><td>PS2.B Forces at a distance are explained by fields that can transfer energy and that can be described in terms of the arrangement and properties of the interacting objects and the distance between them. These forces can used to describe the relationship between electrical and magnetic fields.</td></tr>
<tr><td>PS3.A-B The total energy within a system is conserved. Energy transfer within and between systems can be described and predicted in terms of the energy associated with the motion or configuration of particles (objects). Systems move toward stable states.</td><td rowspan="5">SEP5—Mathematics and Computational Thinking
--Create and/or revise a computational model or simulation of a phenomenon, designed device, process or system.
--Use mathematical, computational, and/or algorithmic representations of phenomena or design solutions to describe and/or support claims and/or explanations.
--Apply techniques of algebra and functions to represent and solve scientific and engineering problems.
--Use simple limit cases to test mathematical expressions, computer programs, algorithms, or simulations of a process or system to see if a model “makes sense” by comparing the outcomes with what is known about the real world.
--Apply ratios, rates, percentages, and unit conversions in the context of complicated measurement problems involving quantities with derived or compound units.</td></tr>
<tr><td>PS3.C A field contains energy that depends on the arrangement of the objects in the field.</td></tr>
<tr><td>PS3.D Photosynthesis is the primary biological means of capturing radiation from the sun. Energy cannot be destroyed; it can be converted to less useful forms.</td></tr>
<tr><td>LS1.C The hydrocarbon backbones of sugars produced through photosynthesis are used to make amino acids and other molecules that can be assembled into proteins or DNA. Through cellular respiration, matter and energy flow through different organizational levels of an organism as elements are recombined to form different products and transfer energy.</td></tr>
<tr><td>LS2.B Photosynthesis and cellular respiration provide most of the energy for life processes. Only a fraction of the matter consumed at the lower level of the food web is transferred up, resulting in fewer organisms at higher levels. At each link in an ecosystem, elements are combined in different ways and matter and energy are conserved. Photosynthesis and cellular respiration are key components of the global carbon cycle.</td></tr>
<tr><td colspan="2">Figure 8. Components of NGSS used to create Example 1</td></tr>
</table>

We did not consider individual PEs when developing this PLD. A posteriori comparison of Fig. 8 and the NGSS, however, reveals that Fig. 8 includes the components (as listed in the *Standards*) of the following PEs: HS-PS1-4, HS-PS3-2, HS-LS1-5, HS-LS1-7, and HS-LS2-4. Item specifications created from Fig. 8 might or might not include these PEs and would probably include new combinations of SEPs and DCIs with CCC 5. If one were to create PLDs from a list of performance expectations, it is likely that they would either be much more specific (concern a much smaller domain) or much less cohesive. The combinations of SEPs, DCIs, and CCCs included in the NGSS are sparse and randomly distributed within the matrix of all possible combinations. This distribution makes it difficult to write cohesive PLDs for high level claims by considering only PEs.

In this example, the descriptions of performance at each level are based on novice/expert differences which come from the literature on conceptual change. Novices do not refer to the organizing theories and ideas of the discipline (in this case, conservation of energy) when making sense of phenomena, while experts tend to frame questions and problems using accepted theories and concepts (Vosniadou, 2014; Harrison & Treagust, 2001). Here, we differentiate performance based on the extent to which students can use the organizing idea of energy conservation as described in CCC 5. As students move toward expertise, they are able to consider energy conservation across a greater range of energy types, phenomena, and scales.

The choices of which types of energy to include at each level of performance are informed by traditional high school science curricula in which courses are arranged by discipline. At the novice level, students are only expected to know two highly related forms of energy (gravitational potential and kinetic) which are usually taught together in a physical science course. Developing students are expected to add thermal energy when considering energy flow between and within systems. While this type of energy is not typically taught at the same time as mechanical energy, it is usually explained in terms of potential and kinetic energy of particles and referred to when discussing conservation of energy. For instance, teachers usually explain that some of a falling object's kinetic energy transforms to heat when it hits the ground. At the Proficient level, students must be able to include more diverse types of energy within their models. Electric and magnetic potential energies are usually explained in terms of the position of charges or moving charges within a field. Including these energies (which are usually modeled on a microscopic scale) together with the previous types (usually modeled on a macroscopic scale) in the same model and both conceptually and quantitatively is a step toward expertise. Finally, an expert is able to move between disciplines and across scales to include all types of energy within the same model of energy conservation.

PLD Differentiation Example 2: Degree of independence

A second example, shown in Fig. 9, also differentiates performance quality along a single dimension, that of degree of student independence. This method of differentiation can be helpful when claims concern what students can do with knowledge. Students may be able to perform skills with assistance or scaffolding that they cannot perform otherwise. Progressive degrees of

independence signify a move toward greater mastery of the skill. For the NGSS, what students do with scientific knowledge is expressed in the SEPs. The example foregrounds a set of SEPs as applied within the context of the conservation of energy.

In this example, the PLD claims are aligned to SEPs 2-6. Whereas our first example foregrounded a single CCC, this example foregrounds a group of SEPs. As for our first example, these descriptors are also three dimensional. Each includes multiple DCIs and CCCs. Fig. 10 lists the names of the relevant components from the *Standards* which are addressed in Example 2. We have not included the text as we did for our first example but feel that the given information is sufficient to illustrate the relationship between the PLD and the components. The full text can be found in Volume Two of the NGSS (NRC, 2013).

Novice	**Developing**	**Proficient**	**Expert**
Student is consistently able to:	Student is consistently able to:	Student is consistently able to:	Student is consistently able to:
Follow step-by-step instructions to carry out an investigation using a mathematical model of energy flow and conservation in Earth's atmosphere, ocean, and land including collecting, analyzing, and interpreting data and constructing an explanation for observations.	Carry out an investigation (including collecting, analyzing, and interpreting data and constructing an explanation for observations) about energy flow and conservation in Earth's atmosphere, ocean, and land using a mathematical model when instructions define relevant variables, how to control and measure (if appropriate) them, and how to analyze the data.	Design and carry out an investigation (including defining and controlling relevant variables, collecting, analyzing, and interpreting data and constructing an explanation for observations) about energy flow and conservation in Earth's atmosphere, ocean, and land using a mathematical model in which relevant variables are defined and controlled given relevant equipment and instruction in how to use it and/or access to a relevant data set (e.g. Solar and climate data).	Design and carry out an investigation (including designing experimental setup, choosing equipment and/or finding and choosing relevant data, defining and controlling relevant variables, collecting, analyzing, and interpreting data and constructing an explanation for observations) about energy flow and conservation in Earth's atmosphere, ocean, and land using a mathematical model.

Figure 9. Example 2--PLD differentiation using degree of independence

SEPs	
SEP2—Developing and Using Models **SEP3—**Planning and Carrying Out Investigations **SEP4—**Analyzing and Interpreting Data **SEP5—**Using Mathematics and Computational Thinking **SEP6—**Constructing Explanations and Designing Solutions	
DCIs	**CCCs**
HS--ESS2.C—The roles of water in Earth's surface processes	**CCC1—**Patterns
	CCC2—Cause and Effect
HS-ESS2.D—Weather and climate	**CCC3—**Scale, Proportion, and Quantity
HS-ESS3.C—Human impacts on Earth systems	**CCC4—**Systems and System Models
	CCC5—Energy and Matter
HS-ESS3.D—Global climate change	**CCC7—**Stability and Change
Figure 10. Components of NGSS used to create Example 2	

Again, we did not consider individual PEs when developing the PLD. We can, however, map the components that are included in the PLD (those in Fig. 10) onto existing PEs and find where they intersect. This process shows that we have included the components of the following PEs in our domain: HS-ESS2-2, HS-ESS2-4, HS-ESS2-6, HS-ESS3-3, HS-ESS3-4, HS-ESS3-5, and HS-ESS3-6. Using these tables to create item specifications would not insure that all (or any) of the individual PEs were represented by the items and it is possible that items would be aligned to combinations of the components that are not represented by the PEs. In fact, such an alignment may be necessary to create a set of tasks for which student responses could provide sufficient evidence for the claims.

Note that, as in our first example, each descriptor in this set describes what students can do. Nevertheless, this set of claims requires evidence that is very different. Items on a large-scale test to provide evidence for either PLD could be created around a phenomenon. However, items to measure the performance of the CCC would be designed to elicit responses about the energy flow involved in the phenomena, while items to measure performance of the SEPs will likely involve an investigation scenario about the phenomenon. Deciding on this focus in the early stages of test development can help to focus the development of tasks for which responses will support the claim.

In this example, levels of performance are distinguished by the level of independence. According to the PLDs, a student who performs at the Novice level can perform all of these skills consistently only when given help in the form of detailed instructions. This would be akin to a student performing a "cookbook" experiment in which s/he follows instructions to collect and analyze data and answer guided questions to confirm the law of conservation of energy. A student at the Developing level can interpret data and explain the results independently (SEPs 4 & 6), but still needs help to perform the other parts of the investigation. A student at the

Proficient level can perform even more SEPs independently, and a student at the Expert level can perform all SEPs independently in the context of this DCI.

It is easy to see that the PLDs in example 2 could be written for multiple DCIs in the form of an "and" statement or a choice of DCIs with an "or" statement. Each of these choices has implications for test development. As written, item clusters designed around phenomena concerning conservation of energy could be aligned to both the appropriate SEPs and the DCI. If the PLDs were written with an "and" component, the test would need to present multiple phenomena concerning multiple DCIs. Creating the PLDs at the beginning of test development helps to guide the item writers so that items will align with the desired SEPs and DCIs. Alignment with appropriate criteria is a necessary condition for scores to support the claims made in the PLDs.

PLD Differentiation Example 3: Degree of Sophistication

Our third example (shown in Fig. 11) differentiates performance by considering the degree of sophistication of the solution. This type of claim is likely to be useful when claims are about what students can do with knowledge. Our sample claim is about student performance of the Science and Engineering Practice of "Asking Questions". Careful analysis will show that aspects of other SEPs such as "Planning Investigations" are also represented, but they are not the primary focus of the PLD. The performances are three dimensional. They require students to incorporate their core knowledge and crosscutting concepts as they formulate questions. The DCIs and CCCs that are represented will vary by the phenomenon that is chosen (i.e. forest fires, tick-borne diseases). For this reason, we do not provide a table listing the components that are represented in the PLD.

We have described "degree of sophistication" as how expertly the student formulates and solves a problem and the extent to which they think about their problem-solving technique. For our example, the problem is to formulate a scientific question about some natural phenomenon. We describe the progression toward expertise by the extent to which students are able to use scientific theories as tools to shape scientific questions. This is similar to Kuhn et al.'s (1988) distinction of novices and experts based on how they use theories to evaluate evidence and evidence to revise theories. Because scientific questions are those which are answerable through data collection and analysis, they will define relevant and measurable variables and account for variables which can and should be controlled. We include distinctions of this in our descriptors as well.

Novice	**Developing**	**Proficient**	**Expert**
Student asks questions which do not consider relevant core knowledge about complex natural phenomena (i.e. increased	Student is able to ask questions which consider some relevant core knowledge about complex natural phenomena (i.e. increased frequency	Student is able to ask questions which consider all relevant core knowledge about complex natural phenomena (i.e. increased	Student is able to ask questions which consider all relevant core knowledge about complex natural phenomena (i.e. increased frequency of

frequency of earthquakes in Oklahoma, changes in annual acreage burned by wildfires, or variations in the prevalence of deer-borne diseases) and which are posed in a form that either negates the need for investigation or prevents valid, evidence-based answers from being developed. For instance, the question may be about producing a perpetual motion machine or it may have an answer based on opinion.	of earthquakes in Oklahoma, changes in annual acreage burned by wildfires, or variations in the prevalence of deer-borne diseases) and which are posed in a form that prevents valid, evidence-based answers from being developed. For instance, the question considers at least one relevant core idea but fails to consider another core idea that is crucial to producing a valid answer, or question fails to control important factors such that answers will be inconclusive.	frequency of earthquakes in Oklahoma, changes in annual acreage burned by wildfires, or variations in the prevalence of deer-borne diseases) and which are posed in a form that allows valid, evidence-based answers to be developed. For instance, the question considers what data will be needed, if they are available/able to be measured, what factors should be controlled, and how data might be manipulated to provide evidence for an answer.	earthquakes in Oklahoma, changes in annual acreage burned by wildfires, or variations in the prevalence of deer-borne diseases) and which are posed in a form that allows valid, evidence-based answers to be developed; and to evaluate the contributions of core knowledge to formulating the questions. For instance, the question considers what data will be needed, if they are available/able to be measured, what factors should be controlled, and how data might be manipulated to provide evidence for an answer.
Figure 11. Example 3—PLD differentiation using sophistication of solution			

All levels of performance require students to ask questions about complex natural phenomena. What differs is the extent to which their questions are scientific (able to be answered through data collection) and the extent to which they consider core scientific knowledge (the DCIs) in their formulation. We have not listed the specific DCIs, CCCs, or PEs represented in these descriptors because they are likely to be phenomenon specific and because our primary purpose is to illustrate a way to apply "degree of sophistication" to distinguish performance. As students move from novice to expert performance, they are able to incorporate DCIs into their questions more correctly and completely. At the highest level, they are able to explain how DCIs affect question formulation.

When the degree of sophistication is used to distinguish performance quality, it is unlikely that claims will focus on content knowledge (DCIs). Solutions are created through use of SEPs and/or CCCs, so this type of distinction is most useful when these dimensions of the NGSS are the primary focus of performance. Some states are collapsing the SEPs into three

larger practices. For instance, a state might collapse the first three SEPs (Asking Questions and Defining Problems, Developing and Using Models, and Planning and Conducting Investigations) into a single practice called "Investigating a Phenomenon". In this case, our descriptors could be modified to more explicitly include "Developing and Using Models" and "Planning and Conducting Investigations". Of course, the claims above could also be modified to refer to specific phenomena and specific DCIs. In fact, such detail would be necessary when developing Range PLDs for item specification.

As written, gathering evidence for the claims will require multiple items which align to the same DCI and which differ in the extent to which they require students to incorporate relevant DCIs into their answers. A potential problem is that for students who answer incorrectly, we might not know whether they do not know the DCIs or they do not think to incorporate them into their questions. Finally, note that students who perform at the novice level "ask questions" while students who perform at higher levels are "able to ask questions". Remember that claims must be able to be supported with evidence. Novice responses are those that show no evidence of relevant considering core knowledge or crosscutting concepts. These are the types of questions we have evidence for. Other responses will show varying degrees of using DCIs and CCCs in forming questions. These provide evidence that students "can" ask these types of questions, not that they do it consistently.

We have presented several types of claims that states may make about the domain of and student performance on assessments of the NGSS. It is always important to have coherence between claims and test blueprints, but this will be more critical when considering science assessments of the NGSS because the claims are more complex. Ideally, states will have a more detailed claim such as a PLD to *guide* the development of the more detailed test blueprint. It should not be the case that the more detailed test blueprint is developed prior to the claim or is developed without a more detailed claim or PLD. The reason the claim precedes the test blueprint is that in the spirit of evidence-centered design, the assessment is *designed* to provide sufficient evidence to support the claim. Without a claim (including intended use) it is not possible to decide what the evidence should be. If a state does not have aligned PLDs and test blueprints, it can always work to develop them and make them more coherent. In the next section, we present examples of PLDs from three states who have adopted the NGSS or their own NGSS-like standards.

Examples from States

In this section of the paper, we illustrate some different approaches that states are taking toward defining NGSS claims in the form of performance level descriptors (PLDs) while trying to retain fidelity to the multidimensional character of the NGSS. We discuss three PLD examples in terms of some implications of each approach for construct/domain definition, test development, and validation. The first example is a very general claim from California; the second example is one level of a Grade 11 Science PLD from Washington; and the third example is an excerpt from the Kansas Grade 11 Science PLD.

As previously stated, the PLD and test development processes are iterative. The PLDs we present are snapshots of work from ongoing and complex projects. We have chosen these three examples because each one illustrates a different approach to assessment of the NGSS. We will examine the structure of each PLD and explain how the structure of the PLD determines what types of evidence are needed to support the claims. This, in turn, has implications for test design. These examples will show how PLDs and claims can guide test development in an Evidence-centered Design approach.

Example 1: California NGSS Policy PLDs

California field-tested its NGSS-based assessment last spring, 2018. Test items are aligned to individual PEs. For school-level reporting, the full set of PEs at each grade span is sampled each year between all of the test forms through a combination of common/matrix sampling. For student-level reporting, the test will include a sample of the full set of PEs such that all PEs will be represented over three years. While individual students will not be assessed on all PEs, this design reduces the danger of limiting the curriculum. Because teachers do not know which PEs their students will be assessed on each year, they are encouraged to teach in a way that allows students to master all of the PEs in the Standards. Although the PLDs for California were not public at the time of this writing, the general claims listed in Fig. 12 are very similar to Policy PLDs. We will illustrate how these claims could be used to develop different sets of Range PLDs and how the different Range PLDs have different implications for test development.

Because California has stated that it intends to align items to individual PEs, a Range PLD would need to be developed for each PE. A useful next step in the test development process would be to determine what aspect(s) of performance quality will be used to distinguish performance levels and if it will be consistent across all PEs. For instance, it may be that as students become more proficient, they are expected to master a greater number of PEs (breadth of content) or it may that that they are expected to perform the PE with less support (degree of independence). Such decisions may be made intentionally, or they may emerge naturally during discussion about what constitutes different levels of performance. In either case, explicitly recognizing these aspects of quality before developing items would allow them to be used in the item specifications. This intentional alignment of claims, PLDs, and item specifications strengthens the relationship between test scores and claims.

CAST Claims

The CAST has four claims—one overall claim for the entire assessment, and three separate science domain claims. Table 1 shows the claim statements for CAST.

Table 1. CAST Claims

Domains	Description
3D Overall	Students can demonstrate performances associated with the expectations of the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts across the domains of Physical Sciences, Life Sciences, Earth and Space Sciences, and Engineering, Technology, and Application of Science.
3D Physical Sciences	Students can demonstrate performances associated with the expectations in the disciplinary area of Physical Sciences within the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts.
3D Life Sciences	Students can demonstrate performances associated with the expectations in the disciplinary area of Life Sciences within the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts.
3D Earth and Space Sciences	Students can demonstrate performances associated with the expectations in the disciplinary area of Earth and Space Sciences within the California Next Generation Science Standards, through the integration of Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts.

From: https://www.cde.ca.gov/ta/tg/ca/documents/castblueprint.pdf

Figure 12. Example: California Policy claims for NGSS assessment

Example 2: Washington NGSS PLDs

Washington administered a new NGSS-based science assessment for the first time this spring. The state assesses students in grades 5, 8, and 11 on standards from the current grade span (i.e. 3-5, 6-8, or 9-11) using a fixed form test. The sample PLDs come from a draft dated January 2018. At the high school grade span, the test is projected to take about 150 minutes with 120 minutes devoted to item responses. There are three levels for each PLD. Level 2 describes the lowest level of performance and level 4 describes the highest level of performance. Figure 13 shows the mid-level descriptor for Grade 11, level 3. Note that the PLD consists of a general claim at the top, followed by a set of statements that include more specific knowledge and skills. This is a very common form for PLDs. We have coded the descriptor to show how it aligns with SEPs, DCIs, and CCCs. The two columns with coding for CCCs and Domain/DCI are our additions, as is the color-coding for SEP; these are not included in Washington's PLD.

As we examine the structure of this example, we will consider each set of claims separately--the general claim first, and then the more specific statements. The general claim states the students will be able to effectively apply both the SEPs and the CCCs to "explain phenomena and design solutions," but does not specifically mention the DCIs. It then goes on to list six of the eight SEPs individually (as shown by our color coding). Evidence about performance on each SEP and CCC will be needed to support this claim. Evidence will also be needed about some set of DCIs in order to maintain the 3-D nature of the NGSS, but it will not be needed about every DCI. A test blueprint, then, will need to consider how many items

address each SEP and CCC, but not how many items address each DCI. Next, we examine the more specific claims of this PLD.

Grade 11 Level 3		
An 11th grade student performing at **level 3** effectively applies science and engineering practices and crosscutting concepts to explain phenomena and design solutions to problems in the natural and designed world. The student develops models and uses information and patterns in data to support scientific arguments, describe relationships among variables, and predict how the variables will change over time. The student analyzes patterns in data to evaluate how well a solution meets the criteria and constraints of the problem. The student uses data, mathematical and computational thinking, and scientific principles to construct explanations of scientific processes and arguments about how systems and system parts will change over time.		
CCCs	**In addition to the skills and knowledge demonstrated at Level 2, a student performing at Level 3 can do things like:**	DCI Topics
1, 2, 5, 7	1. Develop and use a model of atomic structure and patterns in data to predict properties of matter and to make and support arguments about the effect of temperature on reaction rates.	HS-PS1
2	2. Plan an investigation to collect data that can, with mathematical and computational thinking, support a quantitative argument about the effect of net force and mass on the acceleration of an object.	HS-PS2
2, 4, 5,	3. Design a device that converts energy from one form to another, and develop and use a model to quantitatively describe how energy changes in one part of a system affect other parts of the system.	HS-PS3 HS-ETS1
2, 4, 7	4. Develop and use a model to quantitatively predict how a change in medium will affect amplitude, frequency and wave speed.	HS-PS4
4, 6,	5. Use data to develop a model and construct an explanation of how DNA determines protein structure and how multicellular organisms are organized into interacting systems with specialized functions.	HS-LS1
2, 3, 4, 5, 7	6. Use mathematical and computational thinking to construct a quantitative argument about the cycling of matter and flow of energy among organisms in an ecosystem.	HS-LS2
2	7. Ask questions to describe relationships among DNA, chromosomes, and traits, and use evidence to construct arguments about causes of inheritable genetic variation.	HS-LS3
1, 2	8. Use data to construct an explanation of how given factors result in evolution and to construct an argument about how environmental conditions affect genetic variation within populations.	HS-LS4
1, 3, 5, 7	9. Use mathematical and computational thinking to qualitatively predict the motion of objects in the solar system, and use information to describe that the processes and elements produced within stars depend on the mass and age of the star.	HS-ESS1
2, 4, 5, 7	10. Develop a model that describes how changes in climate are caused by variations in energy flow into and out of Earth's systems.	HS-ESS2
2, 4, 7	11. Use data from climate models to predict the rate of change in climate and whether impacts on Earth's systems are reversible.	HS-ESS3
4	12. Define qualitative and quantitative criteria for a successful solution to a major global problem that takes into account what people need and want.	HS-ETS1

Figure 13. Draft of Washington PLD for NGSS Science, Grade 11, Level 3

Key: SEP1 SEP2 SEP3 SEP4 SEP5 SEP6 SEP7 SEP8

Our color coding of the 12 specific statements shows that they include all eight SEPs. In addition, we have added two columns to the PLD to show alignment with CCCs and DCI Topics. Our coding indicates that the set of statements includes all CCCs and all DCI topics. A comparison of the statements to the specific PEs of each DCI topic shows that each one corresponds to one or more PEs listed under the topic. Although, the correspondence is not perfect, it is close.

What does this mean for assessment? Some states have approached NGSS test design by developing item clusters that map onto PE bundles (sets of PEs). For a state that wishes to make claims about individual PEs (such as our first example), it could be useful to tighten the language of the statements so that they more closely mirror the PEs. However, Washington's general claim is that students will master all SEPs and CCCs. In this case, direct alignment with individual PEs is not necessary. While items may map directly onto PEs, items can also be created which mix and match the SEPs, DCIs, and CCCs from the original PEs to create new combinations—some of the many possible PEs that do not occur in the *Standards*. What is important is that the specific statements reflect these new PEs. If they do not, then one or the other should be revised.

Finally, note that the statement preceding the general claim includes the phrase "students can do things like." The inclusion of "like" allows for a test blueprint that includes items to measure student performance on many, but not necessarily all the statements. Next, we illustrate the dimensions of quality by which Washington distinguishes quality of performance and what implications this has for test design.

Washington's Level 2 and Level 4 descriptors are very similar to Level 3. Instead of including the Level 2 or Level 4 descriptors, we will describe how they are similar to and different from the Level 3 descriptor. First, all students are expected to learn all SEPs, all CCCs and the same set of DCIs (same "breadth of content"). What differs between levels is:

- Degree of independence—A student at Level 2 needs support to apply the SEPs
- Degree of correctness—Students at Level 3 apply the SEPs "effectively" while those at Level 4 apply them "effectively, consistently, and appropriately".
- Degree of challenge and cognitive complexity—For specific statement seven (about DNA and genetic variation), a student at Level 2 can ask questions "to identify relationships" while a student at Level 3 can ask questions "to describe relationships", and a student at Level 4 can ask questions and use scientific reasoning to "evaluate relationships".

Because different claims distinguish between performance levels using different aspects of performance, different types of evidence are needed to support each type of claim.

The implications of these PLDs for test development are numerous. Because the claims center around PE bundles, the test blueprint will align item clusters with PE bundles. Because claims are not made across DCIs, the PEs of each bundle will come from the same DCI.

Because descriptions of differences in level of performance include multiple aspects of performance, items which measure the same material across these aspects must be created.

Example 3: Kansas NGSS PLDs

Kansas administered the operational form of its NGSS-based state assessment in 2017. The state assesses students at grades 5, 8, and 11. The fifth-grade assessment covers the fifth-grade standards (mastery of the K-4 Standards is considered foundational), the eighth grade assessment covers all middle school standards and the eleventh grade assessment covers all high school standards. The test is designed to be completed in two 45-60 minute sessions and the test domain includes all of the standards in order to avoid unintentional narrowing of the curriculum. A statement on the Kansas Department of Education website reads:

> One significant change is that the days of the tested indicator are gone. Tested indicators were intended to give teachers a better handle on what the assessment was going to be addressing, but were misused and abused to become either the only things that students were expected to learn, or were used for drill-and-kill rote memorization activities. At each of the tested levels, the full scope of the standards will be addressed on the assessment. (https://community.ksde.org/Default.aspx?tabid=5989)

Fig. 14 shows an abbreviated form of the Kansas PLDs for Grade 11. We have colored and emboldened key parts of the text to clarify our analysis. PLDs are organized around Domains (Physical Science, Life Science, Earth and Space Science) and DCI topics. A general claim ("Claim 1," "Claim 2,"…) is made about each Domain followed by more specific claims about the DCIs within the domain ("Target A," "Target B," …) plus an additional target for Engineering Design for each domain. The complete PLD has a total of three "Claims" and 18 "Targets" corresponding to the three NGSS content domains of Physical, Life, and Earth/Space Sciences, and the 18 DCI topics. We have chosen to include only a few key parts of the PLD to clearly illustrate claims that are different from the previous examples and to consider what types of evidence are needed to support these different claims. We do not claim that all Kansas PLDs follow this pattern.

The Kansas PLD provides the types of information and level of detail that can be used to guide item specification. It is an example of a Range PLD. While Kansas uses a much briefer PLD on individual score reports, it provides the example shown below to parents and students online to help them understand more precisely what their scores indicate about their skills and knowledge.

Claim/Target	**Level 2**	**Level 3**	**Level 4**
Claim 1: Physical Science	Students in this range typically comprehend and **describe** scientific ideas, connecting concepts, and procedures or practices (targets A-E), and they apply scientific and engineering knowledge	Students in this range typically comprehend and **explain** scientific ideas, connecting concepts, and procedures or practices (targets A-E), and they apply scientific and engineering knowledge	Students in this range typically comprehend and **analyze** scientific ideas, connecting concepts, and procedures or practices (targets A-E), and they apply scientific and engineering knowledge

	consistently to problems of **low complexity** and **inconsistently** to problems of **moderate complexity** in the physical sciences (targets A-F).	**consistently** to problems of **moderate complexity** and **inconsistently** to problems of **high complexity** in the physical sciences (targets A-F).	**consistently** to problems of **high complexity** in the physical sciences (targets A-F).
Target C*: (We have **emboldened** the ways that students interact with content to clarify our narrative.) **Targets A, B, and D-F are omitted from our example.*	Students can **use** Newton's second law to describe force and motion relationships, **explain** the concept of conservation of momentum, and **describe** and **predict** forces that act at a distance. **No SEPs**	Students can **compare** the effects of forces on an object's motion, **use a mathematical representation** to support the claim there is conservation of momentum in a system, and **use mathematical representations** to describe and predict forces that act at a distance. **SEP5**	Students can **analyze evidence that supports** Newton's second law of motion, **use mathematical representations** to explain the conservation of momentum, and **use models** and **mathematical representations** to describe and predict forces that act at a distance. **SEP2, SEP4, & SEP5**
Claim 2*: Life Science **Targets A-F are omitted from our example.*	Students in this range typically comprehend and **describe** scientific ideas, connecting concepts, and procedures or practices (targets A-E), and they apply scientific and engineering knowledge **consistently** to problems of **low complexity** and **inconsistently** to problems of **moderate complexity** in the life sciences (targets A-F).	Students in this range typically comprehend and **explain** scientific ideas, connecting concepts, and procedures or practices (targets A-E), and they apply scientific and engineering knowledge **consistently** to problems of **moderate complexity** and **inconsistently** to problems of **high complexity** in the life sciences (targets A-F).	Students in this range typically comprehend and **analyze** scientific ideas, connecting concepts, and procedures or practices (targets A-E), and they apply scientific and engineering knowledge **consistently** to problems of **high complexity** in the life sciences (targets A-F).
Claim 3: Earth and Space Science	*Omitted from our example*		

Figure 14. Example: Kansas PLDs for NGSS Science, Grade 11

Claims for each domain are the same—that the student has some degree of skill or scientific knowledge, can connect "content" and "procedures or practices", and "can apply scientific and engineering knowledge". These claims mirror two of the three NGSS dimensions—SEPs and DCIs—but use different language to describe them. There is no direct

mention of the CCCs in the claims. The PLD language gives content knowledge prominence over practices. This is in stark contrast to the Washington general claim that promotes SEPs and CCCs with no mention of DCIs (content). Because the claims do not mention specific "procedures," "practices," or content, they do not narrow the test domain beyond scientific ideas, concepts, procedures and practices and engineering knowledge within physical, life, and earth and space science.

For the claims, levels of performance are differentiated in three ways:

- Cognitive complexity—At level 2, students can "describe", at Level 3 they can "explain", and at Level 4 they can "analyze".
- Degree of challenge—Students can consistently apply knowledge to problems of "low complexity" at Level 2, of "moderate complexity" at Level 3 and of "high complexity" at Level 4.
- Degree of correctness—Students at Level 2 can apply knowledge to problems of moderate complexity "inconsistently" while those at Level 3 can apply knowledge to the same types of problems "consistently".

None of these aspects of performance are explicitly defined in the Claims. However, the Target descriptors provide enough detail to consider how they are operationalized. Next, we analyze the Target descriptors.

The Target descriptors are arranged by DCI topic. Each descriptor includes the content of the DCI--Newton's second law, conservation of momentum, and forces at a distance for the topic shown--but they do not directly mirror the language of the PEs. Looking across the descriptors, it is clear that all students are expected to interact with all of the content, but at different levels of complexity. To more clearly point this out, we have emboldened the action terms in the descriptors. The emboldened terms show that students at level two can "use" Newton's second law, students at level three can "compare the effects of forces" on motion, and students at level four can "analyze evidence that supports" the law. This language matches the language in the general claims which will facilitate the creation of items that align with both Claims and Targets. However, it does not match directly with the NGSS. In order to relate the descriptors more closely to the language of the NGSS, we have interpreted the ways that students interact with content in terms of the SEPs and color coded the Target descriptors.

The color coding reveals a pattern. A student at Level 4 can analyze evidence, use mathematical representations, and use models within given contexts (DCIs). A student at Level 3 can use mathematical representations within given contexts, and a student at Level 2 cannot perform any of these practices, although he is familiar with the material of the DCIs. Students at all levels are expected to engage with all the content—Newton's second law, the conservation of momentum, and forces at a distance—but students at higher levels are expected to perform more of the SEPs. We do not know whether this pattern was an intentional part of the PLD design, or whether it emerged during the development process, but it could be used to translate the claims and descriptors into more "NGSS-like" language and thereby strengthen the alignment between standards and claims.

According to our coding, the level four performance in this topic requires students have knowledge of all of the DCIs and to perform all but two of the SEPs listed in the PEs. The two missing SEPs are ones that may be especially difficult to measure on a large-scale assessment—planning and carrying out an investigation and designing a device. Because we did not analyze all the Kansas Targets in this way, we cannot say whether they follow the same pattern. This is an example, however, so we will discuss the implications for test development as if they did.

The first implication is for test length. Although the exclusion of selected SEPs narrows the test domain, what is left is still quite large. As written an assessment would need to provide evidence for each claim within each target. Items which test content devoid of SEPs would be needed as well as items which measure performance on the same content through different SEPs. The second implication is for score reporting. The relationship between Target performances and Claim performances would need to be defined. For instance, if the Claim performance is seen as the mean Target performance, then Target performance levels could simply be averaged. However, if the Claim performance means that students perform close to that level in all areas, then the scoring rule would need to recognize the presence of a very high and a very low score.

Summary

As shown by our example PLDs, there are many types of claims that can be made about student performance on the NGSS. Each of our examples illustrates a different type of claim about student performance, and each will need different types of evidence to support it.

California's claims about performance of specific PEs will require assessments with multiple items aligned to each PE. In creating Range PLDs, choices will be made about how to distinguish between qualities of PE performance. These choices will have implications for item specification. For instance, if performance levels are to be distinguished by level of independence, then items will need to measure performance of the same PEs with different levels of support.

Claims about performance on SEPs and CCCs, such as Washington's, will require the creation of multiple items which align to each SEP and CCC, but not to each DCI. (Of course, items will need to align to DCIs to maintain fidelity to the 3-D nature of the NGSS, but they do not need to align to all of the DCIs).

Our third example illustrates the need to consider the relationship between a general claim and more specific subclaims. When subclaims are included without a qualifier (i.e. "Students can do things *like*"), then one must decide how to collapse varying levels of performance on subclaims into a single level. In our example from Kansas, for instance, we might ask what patterns of performance in the nine Targets constitute an overall performance at level 3. Is the overall performance based on the mean of the Targets, the minimum Target score, the overall scale score, or something else? In addition, subclaims supported by subscores may require longer tests and more complex measurement models. Understanding the relationship between claims (as stated in PLDs) and evidence allows for a more tailored process of assessment design by clarifying the types of evidence that should be collected throughout the assessment design process.

Conclusions

Of the many U.S. states, districts, and territories which have adopted the NGSS, most are still in the process of developing operational summative assessments. Our work analyzes the types of decisions that must be made in developing assessments of the NGSS to ensure that assessments of and claims about student performance will be aligned. Strong summative assessments of the NGSS should have the same characteristics as strong classroom assessments of the NGSS--they should elicit student thinking about DCIs and CCCs through engagement in SEPs applied to important phenomena—while sufficiently covering the breadth of the NGSS in a cost-effective manner (NRC, 2014). This is a challenging task. Most current large-scale assessments do not require students to integrate scientific practices and essential knowledge (NRC, 2014). We have shown some of the varied approaches states can use/are using to design assessment tasks and align them with the structure of the Standards in order to support their claims about student performance of science.

The PEs themselves are assessable three-dimensional statements, but, for multiple reasons, they still require states to make choices for assessment and instruction. First, they are not intended to be the curriculum (NGSS Lead States, 2013). Instead, they are meant to "clarify what students will know and be able to do by the end of the grade or grade band (NGSS Lead States, 2013, p. 1)." Second, even though the PEs are a subset of all possible combinations of SEPs, DCIs, and CCCs, there are still approximately 50 per grade span—too many to easily assess in a typical summative assessment. How many and which types of PEs must be assessed to provide sufficient coverage and how does this affect the claims that can be made about student skills and knowledge? The PEs of the Standards are a small subset of all the possible ways that SEPs, DCIs, and CCCs can be combined. States could decide to teach and/or assess some of these different combinations of SEPs, DCIs, and CCCs to support a different type of claim. Third, each PE requires a student to do science in ways that have not typically been measured by large-scale assessments. Assessments that can measure the interplay of complex knowledge and understanding through science practice required by the NGSS will not look or act like prior assessments. This means that, compared to earlier standards, assessing the NGSS requires states to make more complex decisions about claims, the test domain, reporting categories, and what aspect(s) of performance will be used to differentiate levels of performance. We have provided some ideas for how to approach these decisions intentionally.

Specific choices that are made will have different implications for validation, alignment, and test development. For instance, while it may be possible for a large state or group of states to develop and operationalize enough items to support a school level claim about the entire set of PEs, a test to do this at the individual level would likely be much too long and costly to administer. For tests of this type, the individual-level claims that could be supported would depend on how items map onto SEPs, DCIs, and CCCs. Careful and intentional creation of items that align with specific SEPs or DCIs might support more specific claims about performance in these areas. Scores on tests that lack this alignment would only support very general claims about individual student learning. States may choose to address all SEPs and/or CCCs across a subset of DCIs. They might test all SEPs across the same DCI by creating a

series of assessment tasks across the same phenomenon, or they could use sets of SEPs across different DCIs and phenomena. Scores from each of these test designs would support different types of claims about what students know and can do. Finally, choices about what distinguishes the quality of performance will impact test design and claims.

Perhaps the most damaging potential negative consequence of defining the test domain is an unintentional narrowing of the teaching curriculum. If a limited number of PES or DCIs are specified, is that all that will be taught? We have shown how some states are lessening this danger. For students to develop the skills and dispositions that are the goals of the Standards, they will need to engage with multiple SEPS (often together) to understand the same DCI and to use the same SEPs across different DCIs from multiple disciplines (NRC, 2014). They also need the opportunity to reflect on and discuss their work with each other (NRC, 2014). Science is an essentially social activity. Simply teaching the PEs, even though they are three dimensional, is unlikely to result in the desired types of learning.

Performance level descriptors (PLDs) combine these decisions and make them visible. Whether intentionally or unintentionally, PLDs define the scope of the test domain, the reporting categories, and the aspects of performance that are used to distinguish performance levels. This determines the types and numbers of items that will be needed to produce needed scores and subscores. These important decisions may be more difficult to coordinate when PLDs are developed after the test has been administered. It is important to remember that the purpose of the test is to gather evidence which will support desired claims about student learning, not to limit what claims can be made. The earlier in test development they are established, the more useful PLDs will be for the process. As we have repeatedly stated, PLDs, claims, and tests are related and should be developed in an iterative and interwoven process. Changes in one facet will likely require changes in the other two. Making these changes will help to guide the complex process of NGSS assessment.

According to the Framework, the intentions of the Standards include for all students to be prepared to engage in public discussions and debate about science and engineering issues; to continue their science and engineering education if desired; to be informed consumers of science and engineering in their lives; and to appreciate the "beauty and wonder of science" by the end of the twelfth grade (NRC, 2012, p. 2). As states decide how to assess student performance of the NGSS, it may be helpful to keep these intentions in mind. The NGSS and supporting documents leave room for different interpretations. States should consider how each choice in assessment development relates to claims about student knowledge in terms of interpreting the NGSS. Our work can provide guidance as they make these important decisions.

.

References

Baxter, G. P., & Glaser, R. (1997). *An approach to analyzing the cognitive complexity of science performance assessments* (CSE Report 452). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Gotwals, A. W., Songer, A. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (183-210), Rotterdam, The Netherlands: Sense Publishers.

Halloun, I. A. (2011). From modeling schemata to the profiling schema: Modeling across the curricula for profile shaping education. In M. S. Khine & I. M. Saleh (Eds.), *Models and modeling: Scientific tools for scientific inquiry* (77-96), London: Springer-Dordrecht.

Harrison, A. G., & Treagust, D. F. (2001). Conceptual change using multiple interpretive perspectives: Two case studies in secondary school chemistry. *Instructional Science, 29*, 45-85.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*, Academic Press, Inc: San Diego, CA.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design.* Princeton, NJ: Educational Testing Service.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising above the gathering storm:Energizing and employing America for a brighter economic future.* Washington, DC: The National Academies Press. https://doi.org/10.17226/11463

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for education reform*. Retrieved from https://ww2.ed.gov/pubs/NatAtRisk/intro.html

National Research Council (NRC). (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2013). *Next Generation Science Standards: For states, by state*s. Washington, DC: The National Academies Press. https://doi.org/10.17226/18290

Rutherford, F. J., & Ahlgren, A. (1990). *Science for all Americans*. New York, NY: Oxford University Press.

Schneider, C. & Egan, K. (2014). A handbook for creating range and target performance level descriptors. Downloaded from: https://www.nciea.org/sites/default/files/publications/Handbook_091914.pdf

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, Florida: Academic Press, Inc.

Vosniadou, S. (2014). Examining cognitive development from a conceptual change point of view: The framework theory approach. *European Journal of Developmental Psychology, 11*(6), 645-661.

Our work encourages states to approach each aspect of NGSS test development intentionally and outlines the types of choices that must be made and their implications. First, we review the three-dimensional structure and complex nature of the NGSS, show how they are different from traditional standards, and explain why the three dimensions should not be separated for assessment. Second, we illustrate multiple ways to define the domain of an NGSS assessment and what types of claims can be made for each. Although it is important for the test domain to remain three dimensional, the three dimensions can be arranged in numerous configurations. States must decide if they will include just the PEs, all SEPs or CCCs across a set of DCIs, or something else. We suggest that states and test developers define this intentionally and at the start of test development to ensure that the structure they have proposed aligns to the claims they wish to make about student performance. The analysis we provide is designed to inform these decisions. Third, we propose seven general aspects of quality that can be used to differentiate assessment task performance (see Figure 1). We incorporate descriptions of novice/expert differences from research on conceptual change in science and science expertise (shown in Figure 2) to apply the general aspects of quality to describe NGSS performance. Fourth, we introduce PLDs as a type of claim, show four types of PLDs from Schneider and Egan (2014), and describe how they are used to guide test development. We create three sample NGSS-based three-dimensional PLDs (one example is shown in Figure 4) to show the relationship between the domain for assessment (see example in Figure 3), the aspect of quality used to describe performance, and the structure of the PLD. We discuss the implications of each for test development. Fifth, we examine the structure of three states' developing PLDs and discuss the how the structure might impact test design. Finally, we discuss how the rich information inherent in detailed PLDs might be used to provide evidence for claims and better inform stakeholders about the expectations of the NGSS for what students know and can do