# CRITERIA FOR PROCURING AND EVALUATING HIGH-QUALITY AND ALIGNED SUMMATIVE SCIENCE ASSESSMENTS

Version 1.0 – February 2018

## I. IN TRODUCTION

A growing number of states have demonstrated a commitment to ensuring better outcomes for all students by developing, adopting, and implementing rigorous science standards based on the National Research Council's *A Framework for K-12 Science Education*, such as the Next Generation Science Standards (NGSS). Fully meeting the vision set forth by the *Framework* and standards designed to implement it requires high-quality and aligned assessments that can provide actionable information to students, teachers, and families. Three-dimensional standards—those that integrate the Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs)—based on the *Framework* are comprehensive, and it is unlikely that most states will assess the full range of depth and breadth in a single summative assessment opportunity for each student. States have several decisions to make regarding how to translate the depth and breadth of their science standards into appropriate statewide summative science assessments. While those decisions will vary from state to state, there *is* a common vision underlying all three-dimensional assessment efforts—and this document describes the criteria that define those common features in a statewide summative assessment.

Achieve developed this document with extensive input from experts and practitioners in the science and assessment fields. It is grounded in our collective and evolving understanding of how best to assess multi-dimensional standards, in the research that defines what all students should know and be able to do in science, and in lessons learned from early state processes in developing three-dimensional assessments. Regardless of each state's approach, this document is intended to be a useful resource for anyone developing and/or evaluating statewide summative assessments aligned to their *Framework*-based three-dimensional science standards.

### THE PURPOSE AND AUDIENCE FOR THIS DOCUMENT

This document describes the features of a statewide summative science assessment that has been designed to embody standards based on the *Framework for K-12 Science Education*, such as the NGSS—to reflect its intent, grounded in the specific expectations of three-dimensional standards. Importantly, this document outlines the expectations for high-quality statewide summative science assessments that are designed and administered, in part, to meet federal requirements for science testing under Title I Part A of the Every Student Succeeds Act. As such, the criteria and evidence described here are grounded in the expectations outlined in the *Framework* and the NGSS as well as those described by federal peer review guidelines. In other words, while the priority for these criteria is to embody the intent of the NGSS and *Framework,* they are intentionally bounded by what would be needed and feasible to meet federal expectations for statewide summative assessments. They do not describe the expectations for other forms of science assessments that states and districts might use, such as interim or benchmark assessments or classroom-embedded summative and formative assessments. As such, expectations for a complete state system of science assessment is beyond the scope of this document. It is important to note that this is not because specifying the criteria for a full system of assessments is not important, but because this is a common component of the assessment system that all states are grappling with.

This document is intended to support state assessment directors, science supervisors, science assessment leads, test developers, and organizations that conduct independent evaluations of alignment of statewide summative assessments to state standards.

## TERMINOLOGY USED IN THIS DOCUMENT.

Throughout this document, the term 'assessment' is used to refer to the full suite of statewide summative science assessments being developed or selected by a state for a given grade level (inclusive of multiple forms, years of administration, etc.). Some of the evidence descriptors are specific to what an evaluator might examine on an operational test form (the tests that students might see, plus answer keys and associated alignment claims)—these are labeled as 'test forms' and are distinguished from 'documentation', which include supporting information that relates to the development and interpretation of the entire assessment suite.

The term 'tasks' is used instead of the more traditional 'items' to better reflect the nature of questions on assessments designed for Framework-based standards. A task includes all scenario/stimuli and prompts associated with a common activity; it can utilize multiple item formats, can have multiple parts, and can require students to respond to open-ended questions. The term 'prompt' is used to identify the specific questions associated with a task. Generally, one or more prompts combine to form a task. A 'scenario' is the phenomenon- or problem-based context used to engage students in the scientific thinking required by the task. A scenario is coherent, engaging, relevant, and provides students with the scientific information (descriptions, data, models, arguments, etc.) they need to successfully respond to the task using the SEPs, CCCs, and DCIs targeted by the task. Throughout the document, 'targeted standards' are referenced—these indicate the state standards a task is intended to assess, and includes both complete performance expectations as well as the specific SEPs, CCCs, and DCIs.

This document contains science-specific (e.g., scientific interpretations of the word 'evidence') and NGSS-specific (e.g., the use of the word 'element' to refer to the specific bulleted ideas described in the *Framework* and the NGSS appendices) uses of words and phrases to convey intentional ideas. A full glossary of specific language uses can be found in Appendix A.

This document is also built on some key principles underlying assessments for which these criteria are appropriate. These principles are detailed in Appendix B.

## EQUITY IN SCIENCE ASSESSMENTS

Ensuring that all students, including those from non-dominant groups, have access to a high-quality and rigorous science education that prepares them for college, career, and citizenship is at the heart of the *Framework* and the NGSS. This emphasis on student equity must extend to current efforts in assessments. Because statewide summative assessment data is used to evaluate and act on student science proficiency among student subgroups, it is imperative that *Framework*-based tests intentionally support students from non-dominant communities in demonstrating their scientific knowledge and abilities. It is difficult to make a validity argument for an assessment if students are incorrectly answering questions because of linguistic barriers or language mismatch, poor engagement, cultural insensitivities or bias, or inappropriately signaled scenarios that lead

students to answer the posed questions without using the targeted knowledge and skill. Because other resources provide extensive guidance about general accessibility and accommodations in assessments, this document focuses on the aspects of student equity and diversity that are most closely tied to content on science assessments, including the design of phenomena, problems, and tasks eliciting three-dimensional performances from students. This is embedded throughout the criteria, rather than posed as a separate expectation, to emphasize that a focus on equity cannot be separated from expectations for high-quality and aligned assessments—one cannot have a high-quality assessment that doesn't support all students. For more detail about how diversity and equity are included in each criterion, please see the FAQs.

## II. OVERVIEW OF SCIENCE ALIGNMENT CRITERIA

The criteria for science build on those described for mathematics, English language arts, and testing practice by the CCSSO Criteria for Procuring and Evaluating High Quality Assessments (CCSSO, 2014). Like the CCSSO Criteria for aligned mathematics and ELA assessments, the current document describes the features all science assessments should demonstrate to be considered aligned to *Framework*-based science standards, as well as the kinds of evidence test developers could provide to show how well a given assessment meets the criteria. These criteria and associated evidence descriptors describe the baseline of common features for assessments. As states articulate their goals and intended uses for their science assessment, they may add to the criteria as appropriate. Additionally, the criteria challenge states to envision three-dimensional items, which are accessible by all students and grounded in the vision of the *Framework for K-12 Science Education*.

To demonstrate it is aligned to the NGSS or similar Framework-based standards, statewide summative science assessments must meet the following criteria:

| Criterion | Description |
|---|---|
| 1. **Design.** Assessments are intentionally designed to assess state science standards in order to provide evidence to support, refute, or qualify state-specific claims about students' achievement in science. | Assessment tasks, and the precise determinations of how well they align to standards, are informed by the design of the assessment, including how tasks individually and collectively provide valid evidence to support an assessment's claims and reporting priorities, and under what conditions. |
| 2. **Three-dimensional performance.** Assessments require students to make sense of phenomena and solve problems by integrating the three dimensions. Assessment tasks elicit sense-making and problem solving by focusing strongly on reasoning using scientific and engineering evidence, models, and principles. | Assessments provide evidence of student knowledge and practice described by the targeted standards by requiring students to use the three dimensions (SEPs, CCCs, and DCIs) to identify and interpret evidence and engage in scientific reasoning as they make sense of phenomena and address problems. |
| **3. Phenomena.** Assessment scenarios focus on relevant, engaging, and rich phenomena and problems that elicit meaningful student performances. Assessment tasks are driven by meaningful and engaging scenarios. | Assessment tasks are situated in the context of meaningful scenarios, and are designed to elicit grade-appropriate, three-dimensional responses (i.e., responses in which students use multiple dimensions together). |

| | |
|---|---|
| 4. **Scope.** Assessments are balanced across domains, and assess a range of knowledge and application within each dimension. | The summative assessments sample across conceptual understanding of core science ideas and crosscutting concepts, elements of scientific practices, and purposeful application of science as described by *Framework*-based standards. |
| **5. Cognitive complexity.** Assessments require a range of analytical thinking. | The assessments allow for robust information to be gathered for students with varied levels of achievement by providing opportunities that require all students to demonstrate varying levels of reasoning across life, physical, and Earth and space sciences as well as engineering, via SEPs and CCCs that range in grade-appropriate sophistication. Accommodations maintain the range of higher order analytical thinking skills as appropriate. |
| **6. Technical Quality.** Assessment tasks are of high technical quality and represent varied task types. | High-quality, fair, and unbiased tasks and a variety of types are strategically used to assess the standard(s). Tasks are designed with a focus on ensuring students from non-dominant communities are supported in demonstrating what they know and can do in science. |
| **7. Reports**. Assessments reports yield valuable information on student progress toward three-dimensional learning**.** | Assessment reports should be designed with specific uses in mind, transparently detail those uses, and illustrate student progress on the continuum toward the goals established by the standards at each grade band. Reports should reflect the intent of the standards by focusing on the integration and application of the knowledge and skills described by the standards, and how they are addressed by the assessment. |

**This document does not address every aspect of assessment design that would need to be considered as states develop and evaluate their assessments**; rather, it focuses on the features of content alignment (across all three dimensions) to the *Framework* and the NGSS. Many of the other important considerations states will have to contend with (e.g., accessibility) are addressed in the CCSSO Criteria.

The criteria, and evidence needed to meet the criteria, presented in this document represent a few notable shifts from traditional alignment expectations:

1) **The importance of an intentional design approach**. Traditional conceptualizations of alignment, that prioritize how well items "hit" targeted standards and "cover" the breadth of standards will not work for the NGSS given the breadth and depth of expectations both within a given standard and across the range of standards for a given grade level or band. To effectively assess the NGSS within common summative testing constraints, states will need to establish their priorities for the assessment. For example, states will need to determine:
    - Their purpose(s) and use(s) for the assessment;

- The claims they want to be able to make—about students, teachers, schools, districts, program evaluation, etc.;
- How items are designed to be accessible to all students;
- The evidence needed to support or refute those claims (and what other sources of information are available, such as via classroom embedded assessments, interim assessments, etc.); and
- How the factors above influence the aspects of NGSS that manifest on the assessment (e.g., assessment blueprint, task design), such as:
  - Which performance expectations, SEPs, CCCs, DCIs;
  - The types of scenarios or contexts students need to address;
  - Task formats;
  - The proportion of assessments devoted to different types or classes of performance;
  - Student sampling considerations (e.g., what evidence is coming from all students? From a subset?)
  - Use and possible consequences of the assessment.

Effectively, the claims, purpose, and design for the assessment should transparently prioritize the features of the NGSS that state leaders determine should be measured. It should be noted that the criteria themselves constitute a series of claims about what needs to be prioritized in an NGSS assessment.

2) **The need for evidence to support design decisions and rationales.** NGSS assessments involve many decisions, and the use of different test forms within and across test administrations may be key to many assessment designs. To document the approach and rationale underlying assessment decisions, it is important that test developers provide substantial documentation (described in detail below) of the assessment development process, and that independent alignment studies incorporate a review of this documentation in their reports. These processes can help ensure generalizability across test forms and administrations, as well as help make assessment decisions and rationale regarding NGSS translation to the assessment explicit and transparent.

3) **The need to redefine content centrality and complexity**. In traditional approaches to alignment, assessment items are generally designed to match the content presented by a standard, and evaluated for how well items match that standard. The NGSS and similar *Framework*-based standards reflect more comprehensive learning goals—any given task or task component may connect to substantial parts of one or more standards, but will likely not fully assess a given standard. The need to reconceptualize "alignment" to appropriately embody the NGSS and *Framework* is the major driving force behind the development of these criteria.

## A NOTE ABOUT INTERPRETING THIS DOCUMENT.

The *Framework,* NGSS, and similar standards that have been adopted since 2013 revolutionized science education by providing standards written as performance expectations that value the three dimensions of

science education equally to increase opportunity for student engagement and understanding. Implementing these standards at scale takes time, and the field is still in transition. This has implications for these criteria, including:

- These criteria represent current best thinking about how to approach NGSS assessments. Over time, as we learn more about assessing rigorous multi-dimensional performance expectations and as assessment practices (task and test form [booklet] designs, platforms, statistical models, etc.) becomes more sophisticated, it would be appropriate to revise these criteria to include new lessons learned and more specific targets.

- The first generation of new science assessments are unlikely to fully meet all of these criteria. Meeting the criteria will involve iterative assessment development processes, commitment to involving NGSS and *Framework* expertise in development and evaluation processes, rigorous construct validation, and careful professional learning for assessment developers and item writers. The importance of this cannot be overstated.

For further information, please see the Frequently Asked Questions in Appendix D.

## III. EVIDENCE TO MEET THE CRITERIA

For each criterion, this section includes:

- The criterion statement.
- A summary box that includes a high-level description of the what the criterion means.
- A paragraph rationale for why providing evidence for the criterion is an important feature of NGSS assessments.
- A description of the evidence to be collected from test tasks on test forms—an operational test that students might see, plus answer keys and associated metadata, and
- A description of documentation evidence, supporting information that relates to the development and interpretation of the entire assessment program--(e.g., test blueprints, explanatory materials, rationale, cognitive lab results, survey results, etc.)

The evidence detailed here describes what statewide summative science assessments will need to demonstrate in order to fully meet each criterion, and they walk the line between currently achievable and aspirational. Some of these descriptors refer to information some states/assessment programs may not yet collect; additionally, many states may want to use the criteria to support their work with developers on assessments that are yet-to-be designed (and therefore do not yet have this evidence). Levels of rigor of the evidence needed to demonstrate that an assessment meets the criteria and is aligned to *Framework*-based standards will vary depending on the stage of assessment development; additional detail in included in The CCSSO Criteria for Procuring and Evaluating High Quality Assessments.

**CRITERION 1: DESIGN**. ASSESSMENTS ARE INTENTIONALLY DESIGNED TO ASSESS STATE SCIENCE STANDARDS TO PROVIDE EVIDENCE TO SUPPORT, REFUTE, OR QUALIFY STATE-SPECIFIC CLAIMS ABOUT STUDENTS' ACHIEVEMENT IN SCIENCE.

| | |
|---|---|
| **Summary** | Assessment tasks, and the precise determinations of how well they align to standards, are informed by the design of the assessment, including how tasks individually and collectively provide valid evidence to support an assessment's claims and reporting priorities, and under what conditions. |

The depth and breadth of knowledge and practice expected by the NGSS and similar three-dimensional *Framework*-based standards will likely not be fully assessed on statewide summative assessments based on current and typical constraints states are facing (e.g., limited testing time and once-per-grade-band testing). As such, assessment tasks and design (including blueprints, task formats and specifications, etc.) must reflect intentional, state-specific decisions about the purpose, claims, and intended use of the assessment.

Evidence from assessment tasks found on test forms as well as assessment program documentation must do two things:

1) The evidence must show that the assessment meets the criteria described in this document, as the common baseline for all assessments claiming alignment for *Framework*-based standards; and
2) The evidence must demonstrate that the assessment provides the necessary and sufficient information to meet the state's claims and purpose. State's purposes and claims for their science assessment could manifest in a number of decisions about the assessment design, including how content across the three dimensions is sampled in blueprints and test forms (e.g., which standards; how much of each standard; the necessary item formats; the range of content included on different test forms; and the specific qualities of three-dimensional performances that are advantaged on the assessment, such as transfer tasks, emphasis on sense-making processes, and integrated vs. discipline specific performances).

The evidence descriptors below describe the necessary features of the design that need to be detailed in documentation and manifested on test forms for three-dimensional science assessments.

| **Table 1: Evidence Descriptors for Criterion 1.** |
|---|
| **To fully meet Criterion 1, test forms must demonstrate the following:**<br><br>**Providing evidence to support state assessment claims.** Each task contributes evidence for particular claims and subclaims. Tasks, taken together, provide the evidence needed to support the assessment purpose, claims and subclaims, assessment design, and reporting categories. |

**To meet Criterion 1, documentation must describe the relationship between an assessment's claims, reporting categories, blueprint, and task design, describing in what ways the assessment is designed to produce the necessary evidence for the assessment's target, including:**

- **Use.** The intended users and appropriate uses of the assessment results are clearly identified. (e.g., Is this assessment being used to make decisions about individual student placement? Program improvement for districts? Curriculum evaluation? Accountability at various levels?).

- **Domain**: The standards, elements, competencies, knowledge, and/or skills being assessed are defined specifically enough to allow differentiation from other likely interpretations by intended users, and specifically enough to guide test development.

- **Claims about student performance**: Specific statements about student capabilities that the assessment is designed to measure. These claims represent the priorities, depth, and breadth of the state's standards, and are specific enough that assessment tasks can be evaluated with regard to how well they provide evidence to support or refute the claims.
- **Task-level claims**, including:
    - The specific knowledge and practice targeted by the task (i.e., core components or substantial parts of SEP, CCC, DCI elements included in the grade-band that are intended to be assessed by each prompt within tasks, and the tasks as a whole)

    - Documentation that shows how the knowledge and practice targeted by a task connects to a substantial part of a standard/performance expectation at grade-level, and what evidence of proficiency looks like.

- **Opportunity to learn (OTL):** The kinds of student learning experiences that would prepare students to perform well on the assessment are specified. Given the progressive nature of the standards, OTL considerations should include both the tested year as well as science learning from previous years.

- **Attention to multiple dimensions of equity and diversity:** These can include, but are not limited to, culture, language, ethnicity, gender, and disability. Assessment documentation should clearly describe how these multiple dimensions were accounted for in (a) the blueprint development process, (b) task development and evaluation processes, including the development of task templates and evaluation rubrics, and (c) the content and format of contexts, phenomena, and problems used on assessments. This includes empirical evidence related to bias.

- **Evidence:** The type, quality, and amount of evidence that the assessment will provide about individual and group student performance.

- **Connecting evidence and use:** How the evidence provided by the assessment clearly matches the intended uses of the assessment (e.g., if the assessment is intended to be used by teachers, what information (and on what timescale) will be provided such that teachers can use the feedback to inform practice/instruction?) and the intended interpretations is described.

**CRITERION 2: THREE-DIMENSIONAL PERFORMANCE.** ASSESSMENTS REQUIRE STUDENTS TO MAKE SENSE OF PHENOMENA AND SOLVE PROBLEMS BY USING THE THREE DIMENSIONS TOGETHER. ASSESSMENT TASKS ELICIT SENSE-MAKING AND PROBLEM SOLVING BY FOCUSING STRONGLY ON REASONING USING SCIENTIFIC AND ENGINEERING EVIDENCE, MODELS, AND PRINCIPLES.

| | |
|---|---|
| **Summary** | Assessments provide evidence of student knowledge and practice described by the targeted standards by requiring students to use the three dimensions (science and engineering practices, disciplinary core ideas, crosscutting concepts) to identify and interpret evidence and engage in scientific reasoning as they make sense of phenomena and address problems. |

The NGSS and similar standards set the expectation that students demonstrate what they know and can do via purposeful application. The expectation, then, is for tasks that require students to use the three-dimensions to make sense of phenomena or to define and solve authentic problems. This contrasts with restating an idea, plugging information into a formula, analyzing a chart without needing to use any DCI understanding, or stating a step of a procedure or process.

Three-dimensional performances—those that demonstrate students' abilities to harness and use the SEPs, CCCs, and DCIs together to make sense of phenomena and solve problems—are a hallmark of the *Framework*, NGSS, and other similar standards. Assessments designed for *Framework*-based standards must engage students in using all three dimensions *together* to assess their capabilities to apply appropriate practices, crosscutting concepts, *and* disciplinary core ideas in their efforts to make sense of an engaging phenomenon or to solve an authentic problem. This involves three important, interrelated but distinct steps to determine whether individual prompts and tasks as a whole require students to: 1) demonstrate and use each targeted dimension appropriately; 2) use multiple dimensions together, and 3) use multidimensional performances to sense-make (defined here as reasoning with scientific and engineering evidence, models, and scientific principles).

The evidence descriptors below describe the necessary features on science assessments for assessing each dimension, integrating the dimensions together, and engaging students in meaningful sense-making.

| **Table 2: Evidence Descriptors for Criterion 2** |
|---|
| **To fully meet Criterion 2, test forms must demonstrate the following:**<br><br>**<u>Sense-making using the three-dimensions</u>** |

**Reasoning with evidence, models, and scientific principles.** All assessment tasks require students to connect evidence (provided or student-generated) to claims, ideas, or problems (e.g., explanations, models, arguments, scientific questions, definition of or solution to a problem) by using the grade appropriate SEPs, CCCs, and DCIs elements as the fundamental component of their reasoning. All prompts—including stand-alone prompts and those in multi-component tasks—require students to engage in one of the following activities:

- **Generating evidence.** Tasks require students to use SEPs, CCC, and/or DCIs to make sense of data, observations, and other kinds of information to generate evidence for scientific sense-making or solving a problem.
- **Applying evidence to claims with reasoning**. Tasks require students to use SEPs, CCC, and/or DCIs to interpret evidence and/or models to make, evaluate, support, and/or refute claims (e.g., ideas, predictions) about a problem or phenomenon.
- **Reasoning about the validity of claims.** Tasks require students to use SEPs, CCC, and/or DCIs to evaluate claims, ideas, and/or models based on the quality of evidence, additional or revised information, or the reasoning relating the evidence to the claim.

**Coherence and Supports.** Multi-component assessment tasks require students to progressively make sense of a phenomenon or address a problem; this includes that prompts within multi-component tasks build logically and support students' sense-making such that by the end of the task, students have figured something out. Supports included in the tasks (e.g., scaffolds, task templates) support sense-making and do not diminish students' ability to demonstrate the targeted knowledge and practice.

**<u>Assessing each dimension.</u>**

- All tasks elicit grade-appropriate thinking. Successful completion of all prompts (including stand-alone and part of multi-component tasks):
    - requires students to demonstrate understanding of and facility with the grade-appropriate[1] elements of the SEPs, CCCs, and DCIs[2] [cannot fully be answered using below grade-level understanding]
    - Does not require unrelated (not targeted) SEP, CCC, or DCI elements.
- The emphasis throughout the entire assessment is on the elements, parts of elements, and levels of sophistication that distinguish the performance at that grade band from those at a higher or lower grade band.
- **There are no tasks where rote understanding—associated with any dimension—is assessed in isolation**. In other words, prompts that ask students to 1) recall vocabulary, isolated factual statements, formulas or equations, 2) focus restating or identifying steps to a process, or 3) simply

[1] Note that 'grade-appropriate' is intended to distinguish from grade-band expectations from previous or future grade bands. In other words, if a state is assessing PEs from grades 3-5 on a 5th grade assessment, it is acceptable to assess DCIs included in 3rd and 4th grade. What would not meet the grade-appropriate expectation are MS or HS PEs assessed at a K-2 or 3-5 level

[2] Grade appropriate as defined by NGSS appendices E, F, G, and the foundation boxes associated with the NGSS PEs. It may be helpful to refer to the *Framework* for further elaboration of the three dimensions.

restate the language included in a DCI, SEP, CCC, or the prompt itself are not aligned to any dimension.

**Please see below for guidance on the important features associated with each dimension.**

**Integrating multiple dimensions.**

- Multi-component tasks assess three-dimensional performances.
- All tasks are science tasks.
- All multi-component tasks require students to explicitly apply at least two dimensions at appropriate levels of sophistication to successfully complete the task. (This contrasts with tasks that may connect to a dimension, but not require grade-appropriate use for successful completion.)
- The vast majority of assessment prompts (individual questions; these can be stand-alone tasks or parts of multicomponent tasks) explicitly require students to explicitly apply at least two dimensions at grade-appropriate levels of sophistication to successfully complete.
- Tasks targeting a specific standard or set of standards individually reveal a key component of the scientific understanding associated with those PE targets (i.e., individual tasks provide a piece of evidence to support a claim about student proficiency with that standard; tasks should assess what is most important about targeted DCIs, SEPs, CCCs, and/or PEs).
- Collectively, the successful completion of the set of tasks and tasks targeting a PE or bundle of PEs reveals sufficient (but not necessarily comprehensive) evidence of student proficiency for a given PE or bundle of PEs including all three dimensions[3].
- The knowledge and skills required of students should not exceed assessment boundaries specified in their state's standards.

**To meet criterion 2, test documentation should provide:**

- Test blueprints and other specifications as well as exemplar test tasks for each grade level/band assessed, demonstrating the expectations above are met.
- A rationale for the selection of DCI, CCC, and SEP elements for each item, including the relationship between the assessment design and goals, the elements selected, and how the task assesses those elements.
- A rationale for how parts of DCI, CCC, and SEP elements are selected (e.g., how were the most important components of these elements chosen? What were the criteria used for unpacking?)

[3] Determinations of what constitutes "sufficient" will depend on expert evaluation and a state's purpose and claims for their assessments; truly sufficient and comprehensive evidence will likely require a much broader range of evidence than what can realistically be provided on a statewide summative assessment.

- Evidence for whether all groups of students—including those from non-dominant groups—are actually using the knowledge, abilities, and processes described by grade appropriate elements of the dimensions to respond to assessment tasks (e.g., findings from cognitive labs that intentionally sample students from a wide range of ability, economic, racial, ethnic, and linguistic backgrounds).

## GUIDANCE TO SUPPORT CRITERION 2:

Because three-dimensional learning as a construct is relatively new to the field, this section provides some additional guidance regarding what assessing the dimensions should "look like". This section includes possible examples for assessing the three-dimensions, but these examples are not intended to be comprehensive, prescriptive, or exclusive—rather, they are intended to support developers and evaluators as they pursue three-dimensional assessments. It should be noted that different types of tasks—those that are designed to foreground and prioritize different capabilities and competencies—will likely be needed across an assessment to represent the student performance associated with each dimension and their use together.

It should be noted that guidance across all three dimensions assumes and emphasizes the importance of the foundation boxes/elements/language from the framework used to detail the SEPs, CCCs, and DCIs as part of the process for determining alignment.

## SCIENCE AND ENGINEERING PRACTICES (SEPS).

**Application of SEPs in both phenomenon- and problem-based scenarios.** All eight practices can be applied in the context of making sense of phenomena (science) and solving problems (engineering). Assessments should engage students in demonstrating their ability to use SEPs in a variety of different contexts. For example, middle school students could be asked the following, by practice:

- **Asking Questions and Defining Problems.** Given a challenging situation, students formulate a problem to be solved with criteria and constraints for a successful solution. Similarly, given an intriguing observation, students use their knowledge to formulate a question to clarify relationships among variables in the system that could account for the phenomenon.
- **Developing and Using Models.** Develop a visual representation to propose a mechanism for a phenomenon being examined, based on presented data and student understanding, that is used to predict a future observation under different conditions (in contrast to simply diagramming a representation). Critique and revise a diagram that represents a conceptual model of a natural or human-made system to support solving a problem related to that system.
- **Planning and Carrying Out Investigations.** Given three different solutions to a problem, design [and conduct, or describe expected and unexpected outcomes for] an investigation to determine which best meets the criteria and constraints of the problem.
- **Analyzing and Interpreting Data.** Given a data set and research question, analyze and display the data to answer the question. This contrasts with simply reading a graph or chart.
- **Using Mathematics and Computational Thinking.** Formulate an equation based on data, and use that equation to interpolate or extrapolate possible future outcomes to answer a question or propose a solution to a problem.

- **Constructing Explanations and Designing Solutions.** Explain the likely reason for an experimental result, or design and compare solutions to see which best solves a problem.
- **Engaging in Argument from Evidence.** Describe how given evidence supports or refutes a claim.
- **Obtaining, Evaluating and Communicating Evidence.** Compare the credibility of information from two different sources, and summarize findings to give proper weight and citations to alternative arguments.

**Assessing SEPs, not simply skills.** Across the assessment, tasks should provide evidence of students' facility using the practices for sense-making by requiring the use of practices to make sense of phenomena or solve problems, not simply skills used to carry out a procedure. While skills—purely procedural aspects of scientific endeavors—are important to science, they do not represent a connection to sense-making and therefore are not targeted specifically by the *Framework.* SEPs are assessed when students use them as meaningful tools to deepen their exploration or sense-making of the phenomena/problems at hand, from the student perspective. This is in contrast to assessing skills in isolation, without a connection to the phenomenon or problem being addressed by student sense-making.

Some examples of skills vs. SEPs include:

| Assessing Skills | Assessing SEPs |
|---|---|
| Describing a simple observational pattern from a graph (e.g., there is an increase) | Analyzing patterns in a graph to provide evidence to answer a question or support/refute an idea. |
| Taking an accurate reading from a graduated cylinder | Defining the variables and measurements needed to be part of an investigation in order to answer a question. |
| Labeling the consumers, producers, and decomposers in a food web. | Using a given food web to make a prediction about what happens when one of component of the food web is eliminated, or making a recommendation about how to alter an ecosystem to get a desired outcome. |

**Assessing appropriate SEPs**. Assessment tasks should include those grade-appropriate SEP elements that are most appropriate to the student performance being targeted, the assessment context and design features, and the scenario at hand. The interconnectedness of the SEPs make three things possible, and indeed perhaps ideal:

1) Multiple SEPs (or parts of SEPs) can be used to assess a standard, bundle of standards, or bundle of parts of standards, as demonstrated by a complete student performance.
2) SEPs can take a variety of forms even within a particular practice (e.g., part of "developing a model" includes evaluating and critiquing models; refining models based on new information; and using developed models to predict future outcomes).
3) Enhance students' abilities to access the assessment task by providing ways to make their thinking visible, rather than a focus on stating the right answer.

As an example, suppose a cluster is being developed to address a student performance that involves the SEP element "conduct an investigation to produce data to serve as the basis for evidence to answer scientific questions." Given the constraints of summative assessment, it may be appropriate to have a student focus on planning and evaluating the investigation plan; evaluating the resulting data and methodology to reflect on investigation plan; and/or refining and investigation plan to produce more appropriate data for the question at hand. These activities assess students' knowledge and abilities of the targeted practice, but are more appropriate to the testing context. As cluster development continues, it may become obvious that to reveal student understanding and to fully address the scenario, it is necessary to ask students to analyze and interpret data, use the data to create a model to enable predicting outcomes, or use the data as evidence in and argument. These modifications and additions of SEPs, assuming they remain grade-appropriate, clearly connect back to the PEs, and are often necessary to evaluate student performance and to provide evidence for a state's assessment purpose and claims.

### CROSSCUTTING CONCEPTS (CCCS)

**Breadth of CCC Applications in Assessment.** CCCs are an integral component of the *Framework* and the NGSS, and represent ways that scientists and engineers advance their thinking. The CCCs should be used by students to deepen their understanding of a scenario at hand through a range of applications, including:

- Making connections across multiple science experiences, phenomena, and problems
- Probing a novel phenomenon or problem to support new questions, predictions, explanations, and solutions
- Using different CCCs as lenses to reveal further information about a scenario.

Across all tasks in an assessment, the range of applications associated with crosscutting concepts should be addressed. This could look like using multiple CCCs to probe a given scenario to provide different components of an explanation, argument, question, or hypothesis; asking questions about or proposing an experiment to address a phenomenon for which students are unlikely to have sufficient DCI understanding to fully explain; relating a specific phenomenon/data/model to a different phenomenon, possibly at a different scale, to support near or far transfer of knowledge. Some examples of middle school student performance that could be linked to these applications include:

- **Patterns.** Use identified patterns in data to predict future outcomes in specific scenarios that students are unlikely to fully be able to explain with the grade-band DCIs (to distinguish from DCI application), or anticipate additional data to better understand a phenomenon or solve a problem.
- **Cause and Effect.** Critique the conclusion of an experiment by distinguishing between situations that provide correlational rather than causal relationships between variables.
- **Scale, Proportion, and Quantity.** Use observations and mechanisms at a microscopic scale to predict macroscopic events or solve macroscopic problems.
- **Systems and System Models**. Given an observation, propose a mechanism for how a series of events in a different subsystem may account for the observed phenomenon or problem.

- **Energy and Matter**. Analyze the flow of energy through a system to predict what may occur if the system changes. (This example combines two CCCs, if engaged appropriately: energy and matter, and systems and system models.)
- **Structure and Function**. Evaluate the potential uses of a new material based on its molecular structure.
- **Stability and Change**. Given a system in dynamic equilibrium (stable due to a balance between continuing processes) that has become destabilized due to a change, determine which feedback loops can be used to re-stabilize the system.

**Crafting tasks that are most likely to elicit students' understanding and use of CCCs.** Students' facility with the CCCs often comes to the foreground when their understanding of DCIs is insufficient to explain a phenomenon or solve a problem—in these situations, they must apply crosscutting concepts to learn more about a phenomenon or solve a problem. Assessment developers can use this idea to create situations that make it more likely that students' will engage and use crosscutting concepts.

**Note: Because the CCCs 1) often overlap extensively with DCIs and SEPs, and 2) may be used in different ways by students as they are sense-making, claims about student performance on CCCs should be made extremely carefully. Claims/reports that call out student performance relative to the CCCs should be very carefully evaluated.**

## DISCIPLINARY CORE IDEAS (DCIS)

**Application of DCIs in Meaningful Contexts.** Tasks assessing DCIs cannot be answered successfully by restating a DCI or part of a DCI; they require students to apply the understanding associated with the DCI (i.e., to reason about or with the targeted DCI) in a meaningful context, such as interpreting evidence or defining or solving a problem. Tasks and prompts that assess factual knowledge in isolation are not acceptable.

**Focus on Essential Aspects of DCIs.** In cases where it is not feasible or reasonable to assess a DCI fully, tasks and prompts should target those parts of DCIs that have the most explanatory value—those that are most central to the grade-level understanding or that students will need for future work. This should be determined through careful (and documented) unpacking of the DCIs, informed by expert judgment, and consideration of the standards and *Framework*, NGSS appendix E, and research about learning.

## SCAFFOLDING AND SUPPORTS

For all three dimensions, and their use together, the scaffolding and supports included should enhance students' ability to deeply reason and engage with the targeted dimensions, phenomena, and problems; this is in contrast to scripts, guides, or supports that inhibit students' ability to demonstrate the range of their thinking and abilities.

**CRITERION 3: PHENOMENA.** ASSESSMENTS FOCUS ON RELEVANT, ENGAGING, AND RICH PHENOMENA AND PROBLEMS THAT ELICIT MEANINGFUL STUDENT PERFORMANCES. ASSESSMENT TASKS ARE DRIVEN BY MEANINGFUL AND ENGAGING PHENOMENA AND PROBLEMS.

| | |
|---|---|
| **Summary** | Assessment tasks are situated in the context of meaningful[4] scenarios and are designed to elicit grade-appropriate, three-dimensional responses. |

An important feature of the new standards is that students are expected to demonstrate their knowledge and abilities purposefully, as part of making sense of natural phenomena and solving authentic problems. To measure students' abilities to accomplish such complex tasks, assessments will need to use detailed scenarios involving phenomena and problems, accompanied by one or more prompts, to provide a rich context both to engage students' interest, and to enable them to demonstrate their capabilities. Assessment tasks should be situated in contexts such that they elicit student responses that demonstrate understanding, application and integration of core ideas, practices, and crosscutting concepts that were developed through appropriate three-dimensional classroom learning experiences that intentionally advantage student funds of knowledge.

The evidence descriptors below describe the necessary features of the scenarios included on science assessments.

| **Table 3: Evidence Descriptors for Criterion 3** |
|---|
| **To fully meet Criterion 3, test forms must demonstrate that:**<br><br>● **Phenomena or authentic problems drive all student responses.** All assessment tasks (multi-component and stand-alone) posed to students involve phenomena and/or problems, and both phenomena and problems must be present on each assessment form. Information related to the phenomenon provided by the scenario (e.g., graphs, date tables) is necessary to successfully answer the prompts posed by the task. |

[4] Note that 'meaningful' here refers to making sense of the natural world and solving problems. It is meant to distinguish between tasks that require sense-making using both provided information and learned information and abilities, versus those that might require superficial information, such as numbers in a chart that need to be plugged into a calculation.

- **Relevant and engaging scenarios.** Contexts used on an assessment must:
    - Be puzzling and/or intriguing
    - Be explainable using scientifically accurate knowledge and practice
    - Be observable and accessible to students (firsthand or through media, including through tools and devices to see things at large and small spatial and temporal scales or to surface patterns in data). Specifically, contexts:
        - Use real or well-crafted data that are grade-appropriate and accurate;
        - Use real pictures, videos, scenarios; and
        - Are locally relevant, globally relevant, and/or exhibit features of universality.
    - Be comprehensible at the grade-level and for a range of student groups. This includes ensuring phenomena are unbiased and accessible to all students, including female students, economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, students with limited English language proficiency, and students in alternative education.
    - Be observations associated with a specific case or instance(s), not a statement or topic (e.g., a problem arising from a specific hurricane, rather than explaining the topic of hurricanes).
    - Include diverse representations of scientists, engineers, and phenomena, and problems to be solved.
    - Use as many words as needed, but no more.
    - Be supported by multimodal representations (e.g., words, images, diagrams)
    - Emphasize relevant mathematical thinking (e.g., analysis, interpretation, or numerical reasoning) rather than only formulas and mathematical equations to be applied rotely[5].
    - Build logically and coherently, when multiple phenomena/parts of a scenario are used.
    - Support student engagement throughout the entirety of the task.
- **Grade level-appropriate.** Task contexts are designed to elicit appropriate grade level of disciplinary core ideas, practices, and crosscutting concepts specified in NGSS appendices E, F, and G (as described in criterion 2). Therefore, contexts:

[5] Some standards may include expectations for specific mathematical procedures, formulas, and equations. When specified (explicitly or implicitly), students should be provided with sufficient supports to perform the necessary math in service of demonstrating their understanding of the science ideas and practices.

- Require grade-appropriate SEPs, CCCs, and DCIs (cannot be fully answered by using below grade-level understanding).
- Do not require unrelated[6] (not-targeted) SEP, CCC, or DCI elements.
- Use grade-appropriate vocabulary and syntax.

[6] Unrelated here refers to the assessment target, not a "PE": all SEPs, CCCs, and DCIs required to successfully complete the task should be part of the assessment target, and those not part of the assessment task target should not be required to complete the task.

**To meet Criterion 3, test documentation must describe how the above criteria are consistently met, including:**

- Test blueprints and other specifications as well as exemplar test tasks are provided, demonstrating the expectations above are met consistently across the assessment, including multiple tests forms, grades, and administrations as appropriate.

- Documentation includes qualitative and quantitative information that describe how well a wide range of student populations can engage with the set of phenomena- and problem-driven assessments scenarios. These could include disaggregating data to look at fairness by evaluating:

    - Text complexity information related to the assessment scenarios.

    - Student feedback/survey data to provide evidence for engaging a wide range of students, including those from non-dominant groups, at the targeted grade-level (e.g., student interest and engagement survey data, data from cognitive labs and talk-a-louds).

    - Justification for features of universality of the phenomena and problems used in assessment scenarios, including principles of universal design, accessible language, and broadly accessible contexts that do not depend on idiosyncratic symbols or the cultural experiences of a particular group of students.

    - Evidence that the scenarios presented on an assessment are sensitive to the cultural and linguistic backgrounds of students from non-dominant groups. This includes ensuring that students from non-dominant groups are interpreting language and other representations as intended, and that those scenarios are eliciting the intended knowledge and reasoning.

- Documentation describes any assumptions made about phenomena and problems students may commonly experience in the classroom, and how they relate to (but are unlikely to be exactly the same as) those present on the assessments. This could include:

    - Information about commonly used curriculum materials;

    - Survey data from teachers, schools, and/or districts;

    - State/region-specific considerations;

    - Depth of exposure to related phenomena and problems; and

    - The range of grade-appropriate phenomena that may be used for a given standard/performance expectation.

## **CRITERION 4: SCOPE**. ASSESSMENTS ARE BALANCED ACROSS DOMAINS, AND ASSESS A RANGE OF KNOWLEDGE AND APPLICATION WITHIN EACH DIMENSION.

| | |
|---|---|
| **Summary** | The assessments sample across conceptual understanding of core science ideas and crosscutting concepts, scientific and engineering practices, and purposeful application of science as described by *Framework*-based standards. |

While it is not feasible for individual statewide summative assessment test forms to fully cover all standards, assessments should represent the breadth and depth of the standards by sampling the most important aspects of PEs, across all domains at the relevant grade levels or grade band. Individual test forms should reflect the standards, and there should be a plan for assessing the full range of standards through multiple forms within or across tested years. All decisions about sampling should be transparent and clear to teachers, students, and parents.

**Table 4: Evidence Descriptors for Criterion 4**

**To meet Criterion 4, test forms must demonstrate the following:**

**Balance Across Domains.** Tasks are balanced across the disciplines (physical science, life science, earth science, engineering) as appropriate, roughly mirroring the disciplinary distribution in the grade-level or grade-band standards that students should have covered in the targeted instructional period (e.g., an assessment designed for the NGSS in 5$^{th}$ grade, the distribution of DCIs should be: 35-50% physical science, 10-20% life science, 25-40% Earth and space science, and 10-20% engineering core ideas to roughly reflect the distribution of the PEs).

**Range of Knowledge within SEPs and CCCs.** Tasks are balanced to assess students' capabilities across a range of the SEPs and CCCs that are represented in the state's standards at the target grade-band—in other words, multiple SEPs and CCCs should be represented on the assessment such that the range of student performance described by the grade level/band of standards is addressed. A variety of elements are assessed for each targeted SEP and CCC.

**Valuing All Three Dimensions.** The majority of the assessment is comprised of multi-component assessment tasks. Across the tasks that comprise an assessment, all three dimensions are regularly assessed.

**To meet Criterion 4, test documentation must include:**

- Test blueprints and other specifications as well as exemplar test tasks for each grade level are provided, demonstrating the expectations above are met.

- State-specific design considerations that influence how assessed content is represented on the blueprint and varies across test forms are provided, if appropriate
- Across multiple administrations, the full range of SEPs, CCCs, and DCIs represented in a state's standards are assessed.
- A sampling plan that includes which PEs, SEPs, CCCs, DCIs will be assessed on the assessment is made publicly available.

## **CRITERION 5: COGNITIVE COMPLEXITY**. ASSESSMENTS REQUIRE A RANGE OF ANALYTICAL THINKING.

| | |
|---|---|
| **Summary** | The assessments allow for robust information to be gathered for students with varied levels of achievement by providing opportunities that require all students to demonstrate varying levels of reasoning across life, physical, and Earth and space sciences as well as engineering, via SEPs and CCCs that range in grade-appropriate sophistication. Accommodations maintain the range of higher order analytical thinking skills as appropriate. |

Three-dimensional tasks require higher order analytical thinking skills. Well-designed summative assessments require a range of higher order thinking so that all students—including those at the lower and higher ends of the achievement spectrum—can demonstrate their knowledge and abilities in ways that reflect three-dimensional teaching and learning.

The evidence descriptors below describe the necessary features for assessing a range of higher order analytical thinking.

| **Table 5: Evidence Descriptors for Criterion 5** |
|---|
| **To meet Criterion 5, test forms must include:**<br><br>**Tasks that assess a range of higher-order, analytical thinking.** Tasks on a test form demonstrate a distribution of higher order analytic thinking skills[7] for each grade-level and science domain that is sufficient |

[7] Traditional measures of complexity (e.g., DOK) might not fully reflect the range of higher-order thinking as students demonstrate three-dimensional standards. However, the field is still determining alternative approaches to cognitive complexity in science; if traditional DOK definitions and methodology are used as the cognitive complexity methodology framework, assessments should include:

- 0% DOK1 tasks
- Emphasis on incorporating a higher proportion of DOK 3 tasks

to assess the depth and breadth of the state standards, preferably as evidenced using classifications specific to science and drawn from the requirements of the standards themselves. These could include:

- The extent of scaffolding or support provided to use the dimensions to make sense of phenomena and solve problems
- The level of inference and reasoning required by the analysis of data, models, and evidence, as determined by the scenarios, information provided in the task, and prompts
- The number of processes or kinds of information that need to be connected by the student to successfully answer the question.
- The sophistication of targeted elements (e.g., SEP elements) and their use together[8].
  - The type of transfer of knowledge required by a task (e.g., across closely related phenomena; integrative phenomena that support sense-making using multidisciplinary ideas; transfer of skills and CCCs across disciplinary contexts). It should be noted that this requires knowledge about what students were previously exposed to during instruction, and might not be an appropriate complexity factor for all statewide summative assessment programs.

**To meet Criterion 5, documentation must include:**

- Test blueprints, and other specifications demonstrate that the distribution of cognitive demand for each grade-level and content area is sufficient to assess the depth and breadth of the state's standards, such as:
  - Specifications and descriptions of an appropriate methodology used to determine cognitive complexity; if Webb's Depth-of -Knowledge (DOK) is used, specifications include a rationale and justification for how DOK was interpreted for three-dimensional standards. If alternative models or additional factors of complexity (e.g., degrees of transfer) are used, justification for how this was applied is provided (e.g., for transfer, opportunity to learn information is needed).
  - A rationale to justify the distribution of cognitive complexity for each grade level and science domain
  - Evidence that the full range of higher order thinking and analytical skills are required of all students, including those using any available testing accommodations, gifted and talented students, economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, students with limited English language proficiency, students in alternative education programs, and female students.
- Exemplar test tasks for each grade level illustrate each level of cognitive complexity, and accompanied by a description of the process used to determine a task's cognitive level.

- To the extent possible, a small subset of tasks should reflect the complexity (e.g., complex reasoning, making several connections within/among content areas, select or devise one approach among many alternatives) of DOK 4 without the time factor.

[8] For examples of SEP expectations that range in sophistication, please see NGSS Appendix F and the Sample Task Formats by the Research and Practice Collaboratory.

## **CRITERION 6: TECHNICAL QUALITY.** ASSESSMENT TASKS ARE OF HIGH TECHNICAL QUALITY AND REPRESENT VARIED TASK TYPES.

| | |
|---|---|
| **Summary** | High-quality, fair, and unbiased tasks designed with a student equity focus and a variety of types are strategically used to assess the standard(s). |

Assessments will not be able to measure students' achievement of their state's standards if the tasks they comprise are not of high quality. The evidence descriptors below describe the features of technical quality that need to be present in statewide summative assessments in science.

**Table 6: Evidence Descriptors for Criterion 6**

**To meet Criterion 6, test forms must demonstrate the following:**

- **Multiple task types.** Test forms include multiple task types and the range of task types used supports the goals of the assessment.
- **Accuracy.**
  - All tasks are scientifically accurate and reflect high-quality scientific endeavors (e.g., well-crafted and grade-appropriate data; accurate and grade-appropriate representation of experimental design).
  - All tasks are free from technical errors.
- **Clarity.** Tasks are written and illustrated clearly so that they are easily understood by students.
- **Equitable and Free from Bias.** Tasks are accessible to all student groups, including economically disadvantaged students, students with limited English language proficiency, students with disabilities, students from all major racial and ethnic groups, female students, students in alternative education programs, and gifted and talented students.
- **Appropriate level of mathematics and English language arts/literacy skills.** Tasks do not require reading or mathematics beyond what is required by the SEP, CCC, and DCI as specified by the targeted elements, by the assessment boundaries described in the standards, or a state's grade-level mathematics and ELA standards.

**To meet Criterion 6, documentation must include:**

- Specifications demonstrate that the distribution of task types for each grade level and content area is sufficient to strategically assess the depth and complexity of the standards being addressed. Task types may include, for example, selected-response, two-part evidence-based selected-response, short and extended constructed-response, technology-enhanced tasks, and performance tasks.
- To support claims of quality, the documentation provides:

- Exemplar tasks for each task type used in each grade band, including training tests, task patterns/templates;
- Rationales for the use of the specific task types to assess particular performances;
- Specifications showing the proportion of task types on a form;
- The kinds of responses/work students will produce to demonstrate a range of proficiency.
- A scoring plan (e.g., machine-scored, hand-scored, by whom, how trained), scoring guides and sample student work to confirm the validity of the scoring process for constructed response and hands-on performance tasks; and
- A description of the process used for ensuring the technical quality, alignment to standards, and editorial accuracy of the tasks used on operational test forms.
- A description of the process used to ensure that tasks on operational test forms relate to a range of experiences, do not depend on familiarity with terms or cultural practices that are limited to particular groups.
- Evidence addressing differential functionality of tasks or task prompts across different student subgroups.
- Evidence that all students, including English learners and students from a range of non-dominant communities, are interpreting the scenarios appropriately and are using the targeted knowledge and practice to answer the questions posed.
- An explanation for how research about relevant common student misconceptions or problematic ideas was used to support task and/or task development.
- Discipline-specific features and their inclusion in task design and evaluation

## **CRITERION 7: REPORTS.** ASSESSMENT REPORTS YIELD VALUABLE INFORMATION ON STUDENT PROGRESS[9] TOWARDS THREE-DIMENSIONAL LEARNING.

| | |
|---|---|
| **Summary** | Assessment reports should illustrate student progress on the continuum toward the goals established by the standards at each grade level and/or band and support the types of decisions the assessment is designed to support. Reports should focus on connecting the assessment purpose and appropriate uses of the assessment information, and on the integration and application of the knowledge and abilities described by the standards, and how they are addressed by the assessment. |

[9] The general expectations for this criterion are described in the *CCSSO Criteria for Procuring and Developing High Quality and Aligned Assessments [Criteria D.1 and D.2];* here, we address science-specific implications and restate some portions of the CCSSO criteria for clarity.

Most users of assessment information do not see test forms or the range of documentation—their interpretations and decisions are based information provided by assessment reports, which are presumed to accurately represent student performance. Therefore, reports must reflect what is truly assessed, maintain the integrity of the standards, and support the types of decisions the assessment is designed to support. Information should be provided to guide accurate and effective interpretation of results, including estimate error bars, how cut scores were determined, and how to effectively use results consistent with the intention of the assessment.

| **Table 7: Evidence descriptors for Criterion 7** |
|---|
| **To fully meet criterion 7, test forms must demonstrate the following:**<br><br>**Meaningful Scores**. Assessment tasks (including individual prompts when appropriate) associated with reported scores and subscores meaningfully contribute relevant information about student performance (e.g., if a state is reporting subscores for competencies, such as "communicating information about patterns in life sciences", tasks that contribute score points to that subscore clearly require students to demonstrate grade-appropriate facility with all aspects of that competency, without being confounded by other factors). |
| **To meet criterion 7, documentation must demonstrate [adapted from the CCSSO Criteria for Procuring and Evaluating High Quality Assessments Criteria D.1 and D2):**<br><br>Reporting categories maintain the intent of the standards--they are three-dimensional, and not disaggregated by dimension (i.e., SEPs, CCCs, and DCIs are not given separate subscores) unless there is extensive evidence that those subscores are valid and reliable (e.g., because the CCCs heavily overlap with the SEPs and DCIs, and multiple CCCs may be used by students in a given scenario, it is difficult to make a case for reporting CCC subscores).<br><br>The scores reported can be supported by the assessment design, as demonstrated by evidence. This includes:<br>○ Data confirming that test blueprints include a sufficient number of tasks for each reporting category, so that scores and subscores reliably lead to the intended interpretations and minimize the possibility of misinterpretation[10] and misuse.<br>○ Information confirming that test tasks provide the appropriate evidence for reporting categories, including demonstration of knowledge, application, and skill at the appropriate grade-level[11]. Scores |

[10] This includes ensuring that reporting categories are not misleading. For example, suppose a state intends to report a Life Science subscore, based on cluster-based assessment design that has one cluster (~10 questions) about a life science PE bundle that is focused on 3 LS1 PEs. However, life sciences in any given grade band comprise several DCIs and topics; it would be misleading for a student to be given a life sciences score that only reflected tasks from a small subset of those DCIs and topics

[11] Reporting structures need to be supported by the assessment blueprint and evidence supporting the validity of those scores. Some score reporting options that will likely be supported by sufficient numbers of tasks are:

- Specific three-dimensional competencies within a discipline (e.g., Students can develop and use models to describe cause and effect relationships in the life sciences) or a specific core idea ((e.g., Students can develop and use models to describe cause and effect relationships that drive interdependence in ecosystems)

are meaningful—there is sufficient evidence to indicate that the generalization they are making is valid and reliable.

- If subscores are reported, their value for students, parents, teachers and districts should be clearly described. This could be demonstrated by:
  - Evidence from stakeholder surveys and focus groups
  - Clear explanations detailing how the information should be used by teachers, schools, and districts
  - Clear explanations about how the scores relate to opportunity to learn.

Evidence indicates that assessments yield valid and reliable scores for all students, including those from non-dominant communities.

Reporting should recognize the challenges with transitioning to new standards, and provide clear support for how to interpret results for specific intended users.

## III.REFERENCES


Council of Chief State School Officers (2014). *Criteria for Procuring and Evaluating High Quality Assessments.*

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: National Academy of Sciences.

National Research Council (2012). *A Framework For K-12 Science Education: Practices, Crosscutting Concepts, And Core Ideas.* Committee on a Conceptual framework for New K-12 Science Education Standards. Board on Science Education. Washington, DC: National Academy Press.

NGSS Lead States (2013) *Next Generation Science Standards: For States By States*. Washington, DC: National Academy Press.


## IV. ACKNOWLEDGEMENTS


These criteria were developed as part of a collaborative process coordinated by Achieve involving over 50 individuals from states, science education researchers and experts, assessment experts, and writers of the Next Generation Science Standards, *A Framework for K-12 Science Education,* and *Developing Assessments for the Next Generation Science Standards.* We are immensely grateful to everyone who contributed to this draft document through drafting, review, in-person meetings, and targeted conversations, and we invite public feedback on this document to improve it. The following individuals were major contributors throughout the process, and this document would have been possible without their work and insight:


- Reporting on the classes of phenomena students are asked to explain.

All options have tradeoffs associated with them, and states should carefully consider these options, if they choose to provide subscores.

Danine Ezell, *San Diego Unified School District and San Diego County Office of Education (Retired)*

Cary Sneider, *Associate Research Professor, Portland State University*

James Pellegrino, *Professor, University of Illinois, Chicago*

William Penuel, *Professor, University of Colorado, Boulder*

Melanie Cooper, *Professor, Michigan State University*

Tamara Smolek, *Michigan Department of Education*

Substantive feedback was also provided by our colleagues at the following organizations: American Association for the Advancement of Science, Council of State Science Supervisors, National Center for the Improvement of Educational Assessment, the Stanford NGSS Assessment Project, the National Science Teachers Association, and the WebbAlign Institute.

This document was shaped by important discussions and invaluable feedback provided through a series of virtual and in-person meetings about alignment in general and the criteria, specifically. We'd like to extend a special thanks to these contributors, including individuals from across our 50 State Science Network, NGSS writers, and our research and practice partners.

## V. APPENDICES

### APPENDIX A: GLOSSARY

Given the unique features of the new science standards, it is especially important for a document of this sort to be clear about the intended meanings of essential vocabulary. The following definitions explain word usage within this document.

**Assess:** Successful completion of a task or prompt requires students to demonstrate the targeted understanding; the task cannot be successfully completed without the targeted understanding. This contrasts with items that may connect to or engage a dimension (e.g., incorporate it as part of the context or superficially touch on the dimension) but do not require that students understand the targeted DCI, SEP, and/or CCC to answer the question.

**Assessment:** The collection of tasks intended to represent the state standards for a particular purpose (e.g., a statewide summative assessment), inclusive of multiple test forms.

**Assessment program:** All of the assessments in a coherent suite that are, as a whole, designed to monitor student progress. This can include multiple grade levels or multiple types of assessments (e.g., interim and summative) designed by the same vendor and based on the same underlying assumptions and documentation.

**Blueprint:** A detailed plan that describes the core ideas, crosscutting concepts, practices, task types, levels of challenge, and other categories that will be assessed.

**Claims:** Statements about student performance that the assessment is designed to provide evidence to support or refute.

**Dimension:** One of the three components of science education identified in the *Framework* that comprise each standard in three-dimensional standards like the NGSS: the Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs).

**Element:** Bulleted statement derived from the language of the *Framework* that describe specific components of the disciplinary core ideas, science and engineering practices, and crosscutting concepts (e.g., the bulleted statements in NGSS Appendix F, G).

**Equity:** Providing multiple diverse and differently structured opportunities that allow all students to develop proficiency, agency, and identity as described by *the Framework* and the NGSS.

**Key component:** An essential part (determined through careful unpacking) of a performance expectation, DCI, SEP, or CCC. These should be, determined through careful unpacking and expert judgements, the "most important parts" of the targeted goals or parts of elements.

**Opportunity to learn (OTL):** Overlap between what students are taught and what is assessed.

**Performance expectation:** Statement of what students should know and be able to do after instruction.

**Phenomena:** Specific, observable events in the natural or designed world that students can study.

**Problem:** A situation that people wish to change, that might or might or might not be solved with engineering.

**Prompt:** A question (or item) posed to students to elicit targeted knowledge and skills. Prompts may comprise multiple parts, but they represent the lowest common unit of analysis—score points are assigned for subscores at the prompt level.

**Reasoning:** Rationale, grounded in logic and scientific understanding, explaining how evidence supports a claim.

**Reporting categories:** The categories that group assessed knowledge and skills into broad content areas or competencies to support users in interpreting results (e.g., an overall science score; domain-specific sub-scores).

**Scenario:** Phenomenon or problem situation used to elicit students' capabilities in a meaningful setting. This could include language intended to describe the phenomenon, datasets and observations, interactive stimuli, etc.

**Scientific Evidence:** Interpretation of factual data and information, used to support a claim, idea, or model.

**Scientific Principles:** Scientific ideas that can be leveraged as part of sense-making and problem solving, such as those described by the DCIs, CCCs, nature of science, interdependence of science, engineering, and technology, and natural laws and theories, as well as the knowledge associated with SEPs.

**Standard:** Guidelines for curriculum, instruction, and assessment.

**Targeted standard:** The performance expectation, SEP, DCI, and/or CCC an assessment task or prompt is intended to assess.

**Task:** Activity students are asked to complete to demonstrate competence. A task comprises a scenario and one or more prompts. In some states, tasks may be called item clusters, performance tasks, or testlets—a group of items connected to a common context. Multi-component tasks refer to tasks that have multiple, related prompts associated with a single complete performance, while stand-alone tasks include a single prompt.

**Test form.** The student-facing test, including the specific combination of tasks an individual student would be asked to answer. Tests may have multiple forms.

## APPENDIX B: PRINCIPLES UNDERLYING CRITERIA DEVELOPMENT

As this document as developed, the criteria were developed with a few assumptions and principles driving the work, including:

1. These criteria are intended for high-quality science summative assessments; that is, assessments designed for the NGSS or closely related *Framework*-based standards. This includes the following features:
    - Standards (e.g., NGSS PEs) are three-dimensional; that is, to meet a given standard, students need to demonstrate a performance at the nexus of the grade-appropriate science and engineering practices, disciplinary core ideas, and crosscutting concepts.
    - Three-dimensional performances are intended to be demonstrated through application—making sense of phenomena and/or designing solutions to problems as the mechanism to demonstrate facility with the standards.
    - The standards reflect progressive knowledge and skill—as such, assessments that are designed to measure achievement should focus on grade-appropriate performances. Grade-appropriateness is defined by the NGSS Appendices and the *Framework.*
    - All students should have the opportunity to learn the science needed to be proficient in all standards. While this does not mean that every student will be tested on each standard, the assessment should not promote power standards or the prioritization of some dimensions over others. Assessments should reflect instructional practices that are designed to support all learners, and should attend to opportunity to learn.

2. Most statewide summative science assessments will have to be created within the following constraints:
    - Administered once per grade-band
    - 1-2.5 hour administration; with a phenomenon-based approach, this will likely result in 30-60 questions.
    - Generate comparable individual student scores; additional reports can be generated at higher levels of aggregation.
    - Be administered in an on-demand format that can be taken by large groups of students, and scored relatively quickly.
    - Assessments must cover the depth and breadth of standards.
3. Three -dimensional learning cannot be assessed by assessing each dimension in isolation, or by focusing on rote knowledge (in any dimension).
4. The BOTA report (2014) states that assessing the NGSS requires the use of multi-component tasks; the assumption underlying the criteria is that states will likely use a combination of single- and multi-prompt tasks.
5. Multiple approaches to assessment and task designs are appropriate, and any assessment designed for the NGSS should use multiple task formats.
6. Statewide summative assessments are designed for specific purposes that might include program and/or curricular evaluation, etc.; student scores reflect achievement, rather than intended to provide formative feedback for teaching and learning.
7. Different assessments will have a range of claims and reporting categories for which they are designed. Designing an assessment that provides evidence for these claims and reporting categories may result in different approaches to task design, different units of analysis for meeting or evaluating the criteria (e.g., an item cluster, a segment of the assessment, all tasks associated with a given standard). To support the idea that assessments should be aligned to standards AND provide sufficient evidence to support the claims and reporting categories, this document will not prescribe exactly how criteria should be operationalized or what percentages of tasks are necessary to "meet" the criterion. These specifics would be determined in appropriate alignment methodologies developed to evaluate new assessments.
8. These criteria assume NGSS/Framework expertise and judgement to be operationalized for development or evaluation.
9. Assessments should be designed to reflect research about how students learn science.

## APPENDIX C: CRITERIA DEVELOPMENT

The criteria were developed by a *Framework* and NGSS writers to describe the core features that define assessments aligned to the NGSS. The writers developed the criteria through an iterative process that involved several rounds of external review from states, science education experts, assessment and measurement experts, and practitioners. During the development process, the criteria writers consulted many documents and ideas, including:

- *A Framework for K-12 Science Education*
- *The Next Generation Science Standards*

- *Developing Assessments for the Next Generation Science Standards*
- Resources developed to support three dimensional standards from a range of sources
- The NGSS Innovations, as described by PEEC
- Analysis of a wide range of sample tasks, including those in development that were shared by states and developers, released items and tasks, and items from pre-existing assessments, including state assessments, PISA, and NAEP
- Assessment best practices, including lessons learned from mathematics and English language arts, current efforts in science assessment design, and principles of evidence-centered design.
- Challenges states are currently encountering in negotiating for meaningful assessments

In particular, the criteria emphasize addressing gaps identified in current approaches to three-dimensional science assessments, in an effort to support states as they iteratively develop assessments.

## APPENDIX D: EXAMPLES

***Examples will be added in the next version of the criteria.***