Frequently Asked Questions

The following questions may help clarify some of the specifics about the Criteria for Procuring and Evaluating High Quality Science Assessments.

QUESTION 1: WHO IS THE PRIMARY AUDIENCE FOR THIS DOCUMENT?

The primary audience for this document are state science and assessment leads involved in procuring and developing new science assessments; assessment vendors and developers; and those involved in conducting independent alignment studies to determine whether new science assessments align to three-dimensional standards based on the Framework.

QUESTION 2: WHICH ASSESSMENTS ARE THESE CRITERIA DESIGNED TO SUPPORT?

These criteria are designed to support statewide summative assessments in science that are used for federal accountability purposes. While the criteria do indeed reflect common goals for many 3D assessments, the exact evidence descriptors, as well as additional assessment-specific criteria, will vary depending on the scope and purpose of the assessment.

QUESTION 3: WHY IS REASONING A FOCUS?

A focus on interpretation of information and reasoning with evidence, models, and scientific principles emerged through criteria development as the writers sought to better explain the features of meaningful multidimensional performances. As the writers unpacked the SEPs, CCCs, and DCIs, as well as evaluated a wide range of sample items, it became clear that the common thread separating meaningful performances that embodied three-dimensionality from those that engaged the dimensions either superficially or in contrived ways was an emphasis on reasoning. It should be noted that the criteria emphasize reasoning with evidence, models, and scientific principles as a core principle of all multidimensional performances, not only those that engage practices traditionally associated with claims, evidence, and reasoning (explanation, argumentation).

QUESTION 4: MY STUDENTS KNOW THE INPUTS/OUTPUTS OF PHOTOSYNTHESIS--ISN'T THAT ENOUGH?

To make sure that we are asking students to respond to meaningful tasks that truly assess the goals for student learning, it is important to always reflect on why we care about specific knowledge and skills—why do we want students to know the inputs and outputs of photosynthesis? The *Framework* describes that we want students to be able to make sense of the natural world and solve problems—and doing this effectively in science requires that students have deep conceptual understanding of core ideas and crosscutting concepts, and knowledge and skills associated with the practices. Simply identifying if students know the inputs and outputs of photosynthesis—via definitions or filling in a diagram—doesn't actually provide evidence that students can sense-make using ideas about photosynthesis. The criteria address this directly by emphasizing *application* of conceptual understanding. Students need to know content—but demonstrating that knowledge by reciting facts is insufficient. Instead, the criteria require that assessments measure knowledge—like the inputs and outputs of photosynthesis—via application to specific phenomena, via grade-appropriate practices and crosscutting concepts.

QUESTION 5: HOW DO WE DEAL WITH THE DEPTH AND BREADTH OF THE STANDARDS, GIVEN THE LIMITATIONS OF STATEWIDE SUMMATIVE ASSESSMENTS?

The criteria provide a baseline for ensuring that assessments embody the depth and breadth of the standards by:

- Requiring that all domains and dimensions be represented (Criterion 4)
- Requiring an assessment design that clearly identifies the purpose and claims for the assessment, and how tasks are designed to meet standards while providing evidence to support/refute claims (Criterion 1)
- Emphasizing reasoning with the three dimensions as a mechanism to support depth of understanding (Criterion 2)

Beyond the criteria, states should also consider the role of a system of assessments to fully assess the depth and breadth of the standards.

QUESTION 6: WHY DOES COMPLEXITY NEED TO BE RE-THOUGHT FOR THREE-DIMENSIONAL ASSESSMENTS?

The NGSS pose a fundamentally different sort of standard than previous content standards for several reasons, but perhaps most prominently, because of the intentional integrations of three dimensions (SEPs, CCCs, and DCIs). This difference poses a series of additional considerations for interpreting the complexity within NGSS, including:

- **Intentional integration of science and engineering practices into performance expectations.** Previous science content standards focused on the facts students should know; as a result, complexity language systems like DOK that focused on categorizing the action or way in which students demonstrate their knowledge provided useful and discerning information. With three-dimensional standards and the intentional integration of the SEPs into the performance expectations themselves, DOK categories are too broad to provide useful and discerning information.
- **Three-dimensional learning and assessment**. The NGSS define grade-band specific PEs for all students, and these expectations are built by combining expectations from within each dimension. Dimensions scale in complexity both over grade-band progressions as well as within a grade band. Consequently, the complexity associated with any given performance needs to reflect the three-dimensional expectations as well as the implicit expectations associated with each dimension at that grade-band.

- **Relationships within and among dimensions and standards**. NGSS performance expectations are not intended to be performed in isolation—they are intended to be bundled with other performance expectations (parts or whole), and applied appropriately in contexts. There are many ways the standards can be bundled, which means that any given instructional or assessment task can involve several SEPs, CCCs, and DCIs, and they can be foregrounded/backgrounded in different ways. This influences how we conceptualize and talk about complexity, as any given PE can be demonstrated in a variety of ways, depending on what additional concepts, skills, and connections are emphasized, and how much support students are provided to pursue the task.
- **Novel phenomena and real-world problem focus**. At its core, the NGSS performance expectations and tasks designed for the NGSS require students to develop deep enough conceptual and practical understanding that they can transfer that understanding to a range of real-world scenarios. These scenarios themselves often drive how student performance is emphasized. The familiarity / novelty of the context, the availability of observations and data, the level of collaboration expected, and student prior experience play major roles in determining the complexity potential of the performance.

QUESTION 7: WHY DO THE CRITERIA SUGGEST THAT SUBSCORES NOT BE BROKEN DOWN BY DIMENSION?

The criteria emphasize that the dimensions not be disaggregated for individual students (i.e., giving students an SEP score, a CCC score, and a DCI score) for a few important reasons:

1) Disaggregating by dimension is inconsistent with signaling the importance of three-dimensional learning and performances, and may provide feedback that is inconsistent with how we expect teaching and learning to build toward three-dimensional learning and performances.
2) Assessment tasks are intended to elicit authentic scientific performances—that is, they are intended to be multidimensional. In truly multidimensional tasks and prompts, student performance cannot easily be attributed to specific dimensions. Disaggregation is not supported both because a) the dimensions interact and overlap with one another quite heavily (e.g., CCCs and SEPs; CCCs and DCIs), and b) because student performance in response to any given multidimensional prompt or task is driven by knowledge and skill associated with multiple dimensions, the scenario presented, scaffolding provided, etc.

Importantly, the criteria emphasize meaningful scores over simply providing more data. Score reports should be directly connected to the purpose, claims, and evidence the assessment is designed to provide, AND the evidence that can be validly and reliably acquired by the assessment. These guidelines are for scores provided to individual students—because information provided at the classroom, school, or district level involves more data across students, it's possible that scores at higher levels of aggregation can report smaller grainsize information with more integrity.

QUESTION 8: I THOUGHT A SYSTEM OF ASSESSMENTS WAS NEEDED TO ASSESS THE FULL SCOPE OF FRAMEWORK-BASED STANDARDS—WHY DON'T THE CRITERIA ADDRESS THE SYSTEM?

The NRC report *Developing Assessments for the Next Generation Science Standards* emphasizes that, given the depth and breadth of the NGSS performance expectations, students will need multiple opportunities using a variety of assessment strategies and formats to demonstrate proficiency on the full scope of the standards. This document tackles on piece of that assessment system—the statewide summative piece—and answers the question "What should a statewide summative science assessment comprise to meet federal requirements for valid and reliable assessments aligned to the depth and breadth of state standards, given usual resource and capacity constraints?" The criteria were designed to guide the statewide summative assessment in particular because this is a pressing need for ***all*** states as they work to establish their assessment systems. The focus of the document is driven by meeting states where they are, and over time there will be additional support for system-wide considerations.

QUESTION 9: WHAT DO YOU MEAN BY "STANDARDS BASED ON *A FRAMEWORK FOR K-12 SCIENCE EDUCATION*"?

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.*

QUESTION 10: WHERE ARE ALL THE ADDITIONAL DETAILS, LIKE EVALUATION GUIDANCE AND SCORING CUTOFFS?

Because the criteria are new, an alignment evaluation methodology—that includes information like task-specific three-dimensional expectations, scoring guidance for each criterion, and cutoffs for different levels of alignment—has not been developed yet. Subsequent versions of the criteria will include additional information.

QUESTION 11: HOW ARE EQUITY AND ACCESS ADDRESSED IN THE ALIGNMENT CRITERIA?

Ensuring that science is accessible to and supportive of all learners is central to all efforts connected to the *Framework*, including assessments. As a result, the Criteria integrates equity and accessibility considerations throughout the document. Specifically:

- **Criterion 1** requires that assessment documentation explicitly describe how the design of the assessment attends to multiple aspects of diversity and equity.
- **Criteria 2 and 4** require that assessments meaningfully assess student learning by focusing on conceptual learning and demonstration of that understanding via a range of SEPs.
- **Criterion 3** requires that assessment items be driven by phenomena and problems that are engaging to a wide range of students, including those from non-dominant groups of students; that scenarios be grounded in current best practices for phenomena and problems that are accessible to a wide range of students and are consistent with classroom practices that engender scientific identify and agency in all students; that documentation clearly describe how accessibility to non-dominant student groups was ensured; and that language and math requirements not only do not impede student demonstration of science, but also that research and evidence regarding how students from non-dominant groups may interpret scenarios and questions be included in the design of equitable and accessible assessment opportunities.
- **Criterion 5** requires that all students, including those that identify as part of non-dominant groups be given the opportunity to demonstrate a range of higher order thinking. This includes ensuring that any included accommodations maintain high levels of higher order analytical thinking as appropriate.
- **Criterion 6** requires that items be fair, unbiased, and accessible to all students; that items be clear to all students; that mathematics and language requirements be supportive of students from non-dominant groups in their interpretation of assessment tasks; and that assessment tasks be designed to elicit the targeted knowledge and skill from all students, including taking into account research and data from cognitive labs about reasoning and processes used by students from non-dominant groups to answer particular questions.
- **Criterion 7** requires that assessment design programs provide evidence that reports are valid for all students, including those from non-dominant groups.

Additionally, good testing practice across all content areas is predicated upon fair, unbiased assessments that accurately measure student proficiency relative to standards for all students, including those from non-dominant communities. These content agnostic expectations are not included in the criteria document because they have been extensively discussed elsewhere; we expect that states and assessment developers will follow this guidance, including the Standards for Educational and Psychological Testing (AERA 2014), Criteria A.1-5 of the CCSSO Criteria (CCSSO 2013), and other research- and practice-based guidance for creating equitable assessments.

Of note are two exclusions: the NRC's *Developing Assessments for the Next Generation Science Standards* recommend one feature of assessments that would support wide ranges of students in demonstrating three-dimensional learning: the use of technology enhanced items (TEIs). For these alignment criteria, we chose to not require TEIs for a different equity purpose: several states will be required to administer paper-and-pencil tests to some subset of students, and we felt it critical that the criteria also support these students. We stand by the NRC recommendation to the extent possible, and hope that states will continue pushing for innovation in their assessments to better support accessible three-dimensional student performance.

One aspect of student equity considerations that merits a discussion is the tension between standardized science tests and current thinking about the best ways to support diverse students in demonstrating science proficiency. The *Framework* and standards like the NGSS are built on decades of research that suggest that active, coherent, engaging learning experiences—those that allow students to develop an identity as scientists and engineers as well as scientific agency—enable all students to achieve proficiency in science. Good three-dimensional teaching and learning is predicated on these principles, and values student-driven discourse and active sense-making and problem-solving via the SEPs, CCCs, and DCIs. It can be challenging to mirror these experiences in statewide summative assessments, despite them being more valid representations of student performance relative to new science standards. This is in part due to constraints such as time and resources, and in part a tradeoff associated with a different equity argument: the need for valid, reliable instruments that can be administered to large groups of students and used to monitor student progress and support programmatic interventions for all students, including those from non-dominant groups and communities. Indeed, the statewide summative assessment is often the single opportunity to pose a common vision and metrics for success, providing the data to support local and state policy makers as they make decisions to better support students—provided, of course, that the assessments are actually eliciting the right kind of information from all students. This tension is important to consider—in this document, we suggest criteria that reflect to the extent possible best practices for science, but encourage states to constantly push toward bridging these competing needs by adopting assessment systems, technology enhanced assessment items, and innovative assessment designs.

QUESTION 12: HOW DO THESE CRITERIA RELATE TO THE CCSSO CRITERIA FOR PROCURING AND EVALUATING HIGH-QUALITY ASSESSMENTS?

These criteria were developed to meet a similar and complementary need for guiding summative assessment design in science. As a result, these criteria are designed to complement the CCSSO Criteria by addressing science-specific alignment and design considerations. The content agnostic components of the CCSSO criteria (e.g., Criteria A, D, E, and F) apply to science as well; where three-dimensional standards pose particular expectations regarding these components, they are addressed here.

QUESTION 13: THESE SEEM PRETTY HARD TO MEET. WHAT IF WE CAN'T ACCOMPLISH EVERYTHING YET?

These criteria were designed to balance aspirational with practical, and as a result, they represent what we want state assessments to strive for. Right now, many assessments may not meet all of the alignment criteria, but the articulation of the criteria allow for specific feedback to be provided, and specific improvement targets to be determined. We're confident that state assessment programs can achieve much of what is included here, and over time, their development efforts to meet the criteria will result in better assessments for students.

QUESTION 14: HOW DO I KNOW WHAT, SPECIFICALLY, NEEDS TO BE IN EVERY TASK?

The devil, as they say, is in the details, and there are a lot of questions about task-specific analysis that are out of scope for this particular document. Here, the focus is on tasks, across the assessment, providing rigorous evidence for both grade-specified expectations and test- and task-level claims. For more specific guidance about tasks, we plan to release task specific tools, examples, and connections between those tools and these criteria in summer and fall of 2018.