Reconceptualizing Alignment for NGSS Assessments

Aneesha Badrinarayan, Achieve, Inc

Jill Wertheim, Stanford/SCALE

With

Bill Penuel, CU Boulder
Joe Krajcik, Michigan State University
TJ Smolek, Michigan Department of Education

NARST Annual Meeting
Baltimore, MD
March 31, 2019

Reconceptualizing Alignment for NGSS Assessments

The Next Generation Science Standards (NGSS) are designed around a vision of science learning where students draw on the practices of science and engineering, disciplinary knowledge, and common themes to make sense of phenomena around them (NRC, 2012). The transformation of science standards from one dimension to three has set in motion rapid development of materials to support teaching and learning that can support both teachers and students in meeting these new expectations. Although development of curricula, assessment, and professional development are all in swift progress, there is wide variation in how people interpret what it means for students to meet the demands of these standards--even among those experts whose work define these goals for the field. Because assessments often most explicitly embody the goals of standards, this variation comes into sharp focus when examining assessments. Experts agree that new assessments for NGSS must look very different from previous assessments (NRC, 2014), but differences in approach to the new assessments can have dramatic and long-lasting impacts on the field for several reasons, including:

1. Different approaches to assessments feed back into differences in instruction and professional learning (Hannaway & Hamilton, 2008). Assessments that are aligned to the standards through a liberal interpretation (or misinterpretation) of the learning goals have the potential to drive learning toward a less ambitious vision of science learning than the one laid out in the *Framework*;

2. Differences in assessment of the standards, particularly in the treatment of non-routine reasoning and problem-solving skills, also have the potential to magnify inequities across schools, districts, and states (Darling-Hammond & Adamson, 2010); and

3. Fundamental differences the way standards are interpreted for assessment result in variations in the meaning of assessment outcomes, leading to the potential of reporting and feedback that inadvertently misleads educators, students, parents, and policymakers about the degree to which students are meeting the ambitious vision of the *Framework* (Fuhrman, 1994; Spillane, 1998).

Assessments for states, districts, and classrooms are being developed to support implementation of the standards with a sense of urgency, but with little consensus about the features necessary for an NGSS assessment, the results of those efforts vary widely. Preliminary evaluation of a collection of 68 assessment tasks developed for NGSS by curriculum developers, researchers, and states revealed tremendous range in approach to phenomena, integration of the three dimensions, focus on students' reasoning, and how of all of these features work together to provide information about students' learning. These variations lead to questions about the degree to which these differences represent a range of valid approaches to meeting the standards. Is one approach more effective than the others? Should several approaches

together be part of each students' opportunities to demonstrate their proficiency with the standards? Do any of the approaches misrepresent the goals of the standards?

To provide clarity, unity, and a platform for meaningful discussion in the field moving forward, a team of researchers and practitioners created set of task analysis tools to define criteria for evaluating the degree to which tasks meet the goals of high-quality assessment for the NGSS. The team used these tools to analyze the collection of assessments developed for NGSS to 1) surface, from the ground up, common and diverging perspectives about how these criteria *should* be reflected in assessments; 2) illustrate ways that assessments being developed for NGSS meet these criteria; and 3) illustrate ways that assessments do not meet the criteria for high-quality assessment of the NGSS.

Task Evaluation Instruments

To elicit the broad rage in approaches to evaluating the NGSS, the project team, consisting of members of National Research Council committees that developed the *Framework for K-12 Science Education* and related reports, experts in science assessment and policy, and state administrators developed a set of tools to evaluate the ways existing assessments are designed to evaluate students' progress with the standards. The tools were designed to elicit evidence of the ways in which the task supports (or does not support) students' progress with the core principles of science learning described in *A Framework for K-12 Science Education* (NRC, 2012)*,* as well as foundational principles of high-quality assessment summarized in *Developing Assessments for the Next Generation Science Standards* (NRC, 2014)*, Criteria for Procuring and Evaluating Science Assessments* (Achieve, Inc., 2018)*, Knowing what Students Know* (NRC, 2001)*, Criteria for High-quality Assessment* (Darling-Hammond et al., 2013), and the professional expertise of researchers and practitioners involved in assessment development and evaluation efforts.

The first tool used in this study was designed to structure a pre-screening analysis. The eight questions in the pre-screening tool (Appendix I) were developed around general, fundamental goals for three-dimensional science learning, including the presence of the three dimensions and the use of real phenomena. This tool is used to ensure that tasks that were included in the full analysis meet the most basic goals of the NGSS.

The second tool, the Task Screener (Appendix II), was designed for a more comprehensive evaluation of the way each task elicits evidence of students' progress with the NGSS. This tool consists of four categories, each of which captures information along as many as 17 indicators:

A. Tasks are driven by high-quality scenarios that focus on phenomena or problems.
B. Tasks require sense-making using the three dimensions.
C. Tasks are fair and equitable.
D. Tasks support their intended targets and purpose.

The indicators elicit detailed information about the task, including about the nature of the assessment scenarios, the how the three dimensions are used specifically to make sense of the scenario, the use of multiple modalities in the task and responses, and how the task design supports meaningful use of the student data.

Importantly, these tools were not used to assign scores or value-based judgments to any part of the tasks. Instead, the tools were used as a comprehensive guide to collecting evidence and citing reasoning about the presence or absence of the indicators. A secondary analysis of the task evaluations and a survey designed to probe ideas that emerged from the task evaluations were used to elicit experts' ideas about common features that should be present in three-dimensional science assessments.

Task Evaluation

Project leaders sought to examine tasks as part of this study that are (or will be) publicly available, including soliciting a diverse and expansive call for submissions of NGSS assessments from state departments of education, district offices, curriculum developers, and academic researchers. Seventy-one classroom tasks were included and twenty-eight statewide summative assessments representing life, earth and space, and physical sciences in each K-12 grade band (grades 3-5, 6-8, 9-12).

Project leaders selected a group of 40 reviewers to conduct a comprehensive task evaluation of the tasks that satisfied the Prescreen Tool criteria. Reviewers were recruited to represent five major stakeholder groups: 1) NGSS/*Framework* writers with experience with assessment, 2) research-based task developers, 3) state science and assessment leadership, 4) classroom educators, and 5) building and district-level administrators. Reviewers received six hours of training on using the evaluation tool to describe the degree to which the task met the criteria and to annotate the relevant parts of the task with reasoning the reviewer used to determine whether or not each criterion was met.

The task analysis process consisted of three phases. In phase 1, the leadership team used the Prescreen Tool to evaluate the full set of forty classroom tasks, and found that 31 (43%) of the assessments submitted met baseline requirements of NGSS assessments. All released sample items from NGSS states were included in the study.

In Phase 2, all reviewers, including five of the project leaders, were divided into teams of three to use the Task Screener to do a full analysis of the tasks that satisfied the criteria of the Prescreen Tool. The reviewers evaluated the 31 classroom tasks and 28 statewide summative assessment samples.

Each team was assigned 2-4 tasks to evaluate and annotate such that each task was analyzed by three independent reviewers. The composition of each group represented a range of expertise,

and no reviewer evaluated tasks that they had been involved in developing. Following their individual review, reviewers met with their teams to combine reviews into a form that captured their consensus about the qualities of the task. Reviewers also annotated each task to illustrate the evidence they used to justify their decisions about each of the categories.

In Phase 3, the task annotations were reviewed by an additional team of three reviewers, including NGSS and subject-area experts as well as task developers. The leadership group conducted a final review of the task evaluations and annotations to ensure that there was consistent interpretation of the criteria across the different types of tasks and groups of reviewers.

Analysis

The final task evaluations were analyzed for evidence of themes related to reviewers' (including those who participated in both initial and secondary reviews) perspectives about ways the tasks meet the criteria in the Task Screener and ways they fall short of meeting the criteria. A survey was created to probe more deeply reviewers' perspectives based on themes that emerged from the task reviews.

The survey consists of 20 questions that asked reviewers to select statements that described their beliefs about essential characteristics for NGSS assessments and to provide reasoning for these beliefs. The survey was completed by 95% of reviewers (N=62). Survey results were coded according to two themes related to perspectives about features of NGSS assessments: ideas shared by all reviewers, and ideas for which groups of reviewers' ideas differed.

The patterns that emerged from the task reviews and surveys reveal a set of features that the project participants agree are essential for high-quality NGSS assessments. The analysis also revealed a set of features that are considered "variable," in that the participants differ in their perspectives about how they should be implemented or whether or not they are essential. The sections that follow describe the major points of consensus and points of divergence about essential features for NGSS assessments among the expert reviewers.

Common features for tasks that elicit appropriate student performances. There was overwhelming consensus among both the project leadership as well as the reviewers about features that are essential to the fabric of science assessments designed for the NGSS--and when disregarded, substantially misrepresent the intent of the *Framework* and the NGSS. These included:

1. **Sense-making is the litmus test for NGSS assessments.** Throughout the task evaluations, sense-making emerged as one of the most highly valued features of NGSS tasks (Figure 1). While there were divergent perspectives on other aspects of three-dimensionality (e.g., level of sophistication, number of dimensions required for a task), tasks that did not require students to meaningfully engage in sense-making were not considered aligned to/designed

for the NGSS. Moreover, there was consensus about how sensemaking is elicited in an ideal NGSS task. Reviewers agreed that students should use the dimensions being assessed to "figure something out" about the phenomenon or problem--that sense-making in assessment is practically defined as the construction of an understanding of a phenomenon or problem using Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and/or Crosscutting Concepts (CCCs).

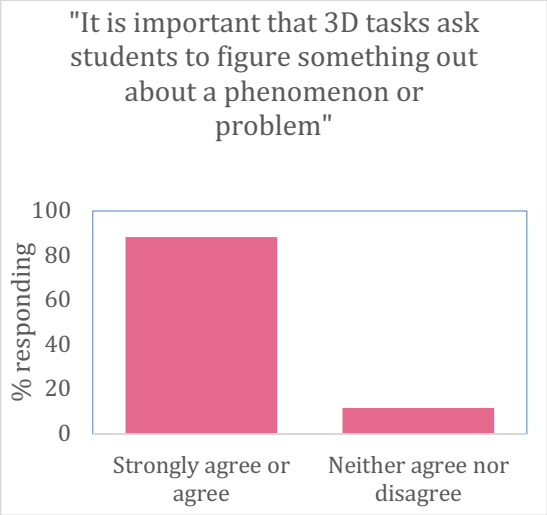

Figure 1: Sense-making. 90% of respondents agreed or strongly agreed that 3D tasks must ask students to figure something out about a phenomenon or problem.

2. **Accessible phenomena or problem-driven scenarios must motivate student responses.** Key features of phenomena- or problem-based scenarios emerged as a critical element of NGSS tasks (Figure 2). While initially viewed as "context," project participants routinely identified that tasks must require students to **address a phenomenon or problem**, with a phenomenon or problem defined as a specific instance (not topic or statement) grounded in real-world observations. Tasks that asked students to address general principles or topics (e.g., evolution; tornadoes) rather than specific instances (e.g., the observation that the number of deaths of swallows in a specific area decreased) were routinely considered poorer-quality tasks. Importantly, these judgments were tied to the idea that without a specific phenomenon or problem to address, tasks did not appropriately elicit sense-making using the targeted three dimensions--instead, student responses reflected rote knowledge or procedural use of the practices. Specific observations cited by project participants included that information provided by the scenario must be required to respond to the question, the targeted dimensions must be required to address the scenario, and students must be required to figure something out about the

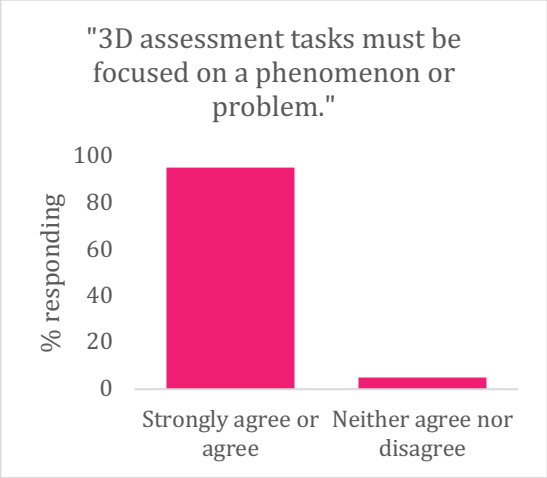

Figure 2. Phenomenon or problem-focused. 96% of respondents agreed or strongly agreed that 3D tasks must be focused on a phenomenon or problem.

phenomenon or problem presented in the scenario. In contrast, tasks that present a phenomenon as a hook but require that students merely recall (or represent) general principles, are not aligned sufficiently to NGSS. Relatedly, several features of scenarios were identified as critical to being able to elicit meaningful 3D performances. These included specificity of the observations, the quality of data and information provided, problematization of the scenario (i.e., making clear the uncertainties students have to make sense of), comprehensibility for a wide range of students, and being actually explainable using grade-appropriate SEPs, CCCs, and/or DCIs.

3. **Tasks must <u>require</u> students to use both practices and disciplinary ideas together throughout the course of the task.** There was consensus among the project participants that tasks that do not evaluate students' use of disciplinary content while engaging with science and engineering practices are inappropriate for NGSS. For example, tasks that only require students to use logical reasoning to interpret data (i.e., no disciplinary understanding needed) were not considered aligned to the NGSS. Notably, this requirement was at the level of the task as a whole--expectations for multidimensionality did not apply to every individual question or prompt, but rather that the 1) majority of the task questions required students to use at least 2 dimensions in service of sense-making, and 2) that single-dimensional questions still required some reasoning and were used primarily as scaffolds for students and/or supports for interpreting student responses.

4. **Scoring guidance must provide sufficient support for interpreting student responses.** Many tasks submitted for review were rejected even prior to prescreening because they had no support for interpreting student responses. Moreover, even in the tasks that were reviewed, support for interpreting student responses was routinely identified as a weakness. Inclusion of the Performance Expectation(s) (PE) being assessed and sample proficient responses were considered sufficient support to warrant inclusion, yet many assessments did not provide even this minimal guidance as to how to use the assessment to evaluate students' progress with NGSS. Reviewers also noted a wide variation in approaches to scoring guides, including:

    - Some tasks claimed to assess all three dimensions in a PE, but only describe how two of the three dimensions (e.g., the DCI and SEP) are to be evaluated.

    - Some scoring guidance identified a large number of PEs that could be involved in student responses, but did not tie interpretation of specific student responses to particular PEs or parts of PEs.

    - Some scoring guidance made claims about the competencies being assessed when those competencies were not being elicited by the task.

    - Many scoring guides did not include appropriate support for interpretation of student responses relative to the purpose of the assessment (e.g., different kinds of support for formative assessments vs. summative assessments).

5. **Assessment purpose matters when defining specific manifestations of these common features.** Project participants emphasized the influence of the purpose of the assessment on the relative expectations for these common features in tasks. For example, short, targeted assessments that required relatively straightforward student responses could be well supported by more narrow phenomena and problem-based scenarios while more expansive tasks--such as those targeting entire PEs or parts of multiple PEs--required more expansive scenarios that could sustain deeper investigation. Similarly, expectations for multidimensionality and grade-appropriateness of the dimensions were higher for summative assessments that occurred at the end of units or grades than for more embedded assessments. This was related to divergences regarding appropriate priorities and purposes for NGSS assessments--for example, while project participants might agree that a task deeply assesses DCI understanding while backgrounding the SEP, they routinely disagreed on whether a DCI focus is an appropriate target for different kinds of assessments designed for the NGSS. These divergences will be described in further detail in the following presentation.

Divergent Perspectives. Despite broad consensus about the *presence* of most indicators in the evaluation tool, divergent perspectives emerged regarding the *importance* of some indicators, especially with regard to relative importance. In other words, across reviewers, there was largely consensus about whether indicators were present in tasks (with some exceptions noted below), but there were divergences with regard to how critical these features were to the overall quality of tasks. Major divergences included:

1. **The nature of crosscutting concept representation in student assessment.** Reviews revealed widely varying ideas about what it means to assess the crosscutting concepts (CCCs) (Figure 3; Table 1)). Reviews and surveys revealed four different conceptualizations of assessing crosscutting concepts:

    - Ideas associated with CCCs are implicitly assessed. Student responses are themselves an example of a crosscutting concept element; understanding the CCCs is <u>necessary</u> to respond to the task and evidence of this dimension is underpinning their response.

    - Ideas associated with CCCs are explicitly assessed. These tasks directly evaluate student understanding of ideas associated with CCCs, such as “can you identify a pattern?”--while these still focus on knowledge in use, the questions themselves provide significant scaffolds for students to use the CCCs as part of the task.

    - Tasks are developed to require considerable reasoning, such that while DCIs and SEPs are necessary, they are not sufficient to respond to the task, requiring that students use ideas embedded in the CCCs to fully respond to these questions. These tasks tend to include more uncertainty and multiple possible correct approaches/solutions/interpretations.

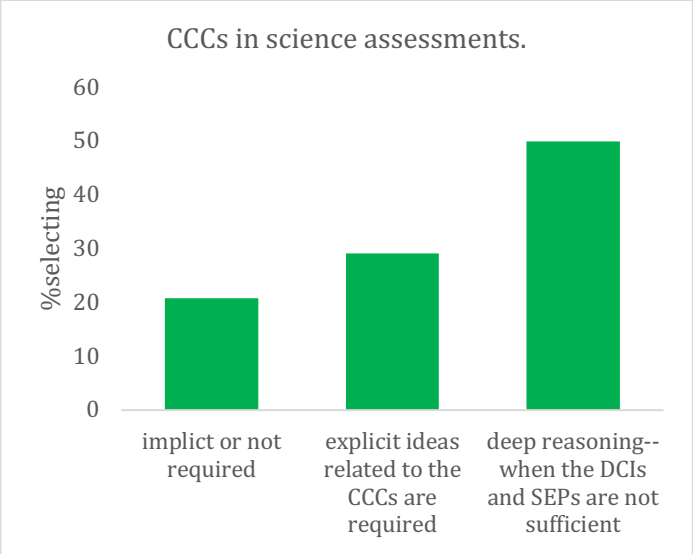

Figure 3. Representation of CCCs in assessments. Reviewers held divergent perspectives on what was considered sufficient to assess the CCCs. 21% of reviewers felt that explicit assessment of the CCCs was not necessary, while 79% expressed that some explicit understanding of CCCs—either specific ideas or CCCs as part of deeper reasoning, when DCIs and SEPs are not sufficient—is an important requirement of 3D tasks.

It is notable that there was one common idea about the role of CCCs that was NOT implemented in any of the assessments that were reviewed. That is, no assessments used CCCs as a tool for helping students recognize connections and common themes across different science disciplines. Nor was it identified as an important goal by reviewers.

Table 1: Examples of divergent reviewer perspectives on the CCCs.

Crosscutting Concepts in assessments	
CCCs must be assessed	"CCCs are the single most important innovation of the NGSS--they are the connection to higher order thinking for ALL students, and not assessing them prevents us from signaling and supporting all students in developing the thinking skills they are capable of." "CCCs are likely the most transferrable ideas in the NGSS--they can help students approach situations and problems outside of science too. So we have to make sure they are developing them!"
CCCs do not need to be assessed	"The CCCs' power is in how educators use them to connect to prior knowledge gained by the student from other classes or subjects. Them being assessed explicitly is not necessary." "One or more CCCs "fall out" of SEP use in the context of a DCI--it is not necessary for students to demonstrate a separate grasp of the CCCs. Such contextualized knowledge use has the potential to provide strong evidence that students have robust and flexible command of a discipline, but is not valuable in a vacuum."

2. **The relative emphasis of disciplinary ideas vs. science and engineering practices.** Assessments for NGSS tend to emphasize evaluation of students' proficiency with the DCI or the science and engineering practice (SEP)--while there are some examples of more balanced assessments, many tended to emphasize one dimension over the other, and individual developers' tasks tended to display a similar priority across tasks. Similarly, project participants were divided and diametrically opposed in their ideas about which dimension *should* be prioritized and the degree to which it is acceptable to foreground one dimension. Survey results revealed that these differences are associated with divergent philosophies about important features of science learning--for example, 43% of project participants indicated that the primary purpose of science teaching, learning, and assessment should be focused on understanding science principles, with the SEPs as a mechanism for students to show their understanding of those ideas rather than as a critically important target in their own right. In contrast, 52% of participants indicated that reasoning using the three dimensions is more important than a focus on assessing knowledge of specific science concepts. Importantly, both "camps" felt that routine and significant backgrounding of 1 or 2 dimensions could still provide sufficient evidence for three-dimensional performances.

3. **The nature of the phenomenon and problem-based scenarios.** The nature of the phenomenon or problem that the task scenario is based on also elicited vastly different ideas. While participants agreed on the importance of specific phenomena or problem-based scenarios that required students to use multiple dimensions, reviewers had very different perspectives on the relative importance of relevance, authenticity/real-world observations, and degree of problematization. These views tended to fall into two contrasting categories summarized in Table 2.

Table 2: Examples of divergent perspectives regarding scenarios

Perspective 1	Perspective 2
Scenarios should present a real, specific instance of a scientific phenomenon and the relevance or importance of the phenomenon should be made clear to students.	Scenarios must be able to elicit the targeted performance, but relevance is not needed to motivate student responses and is not a critical feature of NGSS assessments Students should be able to apply their understanding even in scenarios that are not connected to their experiences.
Scenarios and task prompts should support	Scenarios are tools to elicit the targeted

students in coherently and progressively making sense of a targeted phenomenon/problem.	dimensions and show how the dimensions can be used to make sense of phenomena. Coherence from the student perspective is not necessary.
Phenomena and problems should be specific such that they require students to address that specific instance.	Phenomena and problems must be based on real science but they can be general as long as students must use the targeted science principles to engage with them.

4. **How to define appropriate levels of sophistication of student performance (elements, grade-appropriateness).** There were clear differences regarding how reviewers viewed appropriate levels of sophistication for NGSS tasks, particularly with regard to the CCCs and SEPs. For example, some reviewers use the full spectrum of the grade-specified targets described in the NGSS (including the foundation boxes, and appendices E, F, and G) as appropriate for summative student assessments, particularly in Middle School and High School (e.g., a Middle School task could focus on any grade 3-5 elements of the SEP or CCC), while others suggested that to support claims about grade-appropriate proficient performances, tasks should focus on eliciting only those competencies associated with the aspects of the grade-band targets that are distinct from the targets of the grade bands above and below. Similarly, reviewers diverged with regard to how much guidance/cueing should be provided to students to elicit the appropriate levels of multidimensional performances.

5. **Manifestation of equity considerations in tasks.** While all project participants agreed that tasks should be equitable and fair, there was disagreement about 1) what this means for assessments, and how this connects to equity considerations gaining more traction in NGSS instruction, and 2) the relative importance of equity considerations over eliciting targeted student responses. These differences highlight a tension between some lines of thinking regarding assessment and current moves in the field regarding equitable and accessible science teaching and learning. For example, some reviewers felt strongly that students must be given opportunities to meaningfully connect scenarios/tasks to their own experiences (particularly in classroom assessments), while others suggested that as long as the task is comprehensible, specific considerations for students' motivation or connection to the task were unnecessary. Similarly, some reviewers suggested that an important aspect of student assessment is the support for student identity and agency as scientists, while others suggested that this is an inappropriate target for assessment (across various types of assessment).

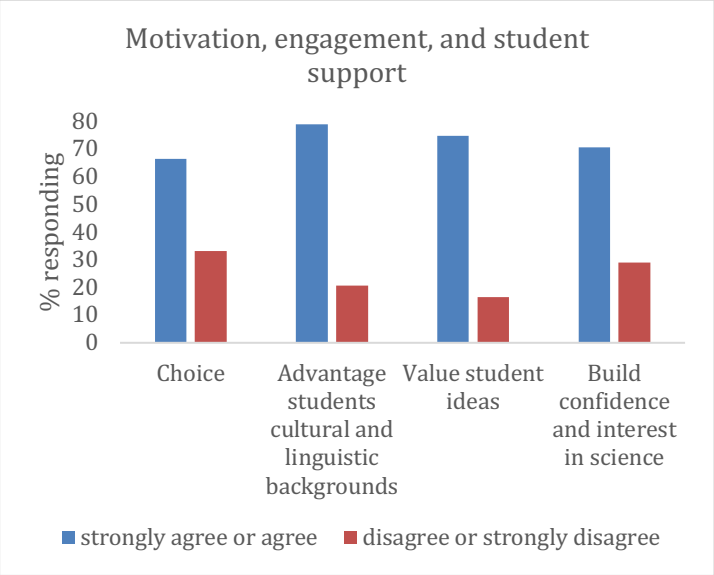

Figure 4. The role of motivation and student engagement in 3D tasks. While a large percentage of reviewers agreed or strongly agreed that features of tasks like student choice, attention to students cultural and linguistic backgrounds, valuing student ideas as an important aspect of task performance, and building student confidence in science were important components of 3D tasks, a non-negligible percentage of expert reviewers did not view these as important elements of 3D tasks.

One tension that arose regularly were formats for task presentation and student responses (Figure 4; Table 3). While some reviewers strongly supported the idea that tasks should 1) present information in multiple modalities effectively and 2) use as many words as needed but no more (allowing for diagrams, simulations, etc. to take the place of some text), others felt that this was unnecessary as long as reading load was limited. Similarly, some reviewers opposed tasks that emphasized written responses as the primary mechanism for demonstrating understanding while others suggested that written responses might be necessary as the most explicit representation of students' ideas.

Table 3: Examples of divergent reviewer perspectives on motivation and engagement in science tasks.

Motivation and engagement	
Motivation and engagement are critical components of 3D tasks	"If we want to support all students, it is critical that assessments actually provide meaningful feedback about student learning. To do so, they must ask all students to show what they do know and can do, and value a wide range of ways of knowing." "If we don't design assessments that support all students, what's the point?"
Motivation and engagement are not necessary for 3D tasks	"In the end, tasks measure what a student knows and can do. Having rich task that keeps the student engaged and motivated and empowered is nice, but accurately assessing the student is the main goal of the assessment. DCIs and SEPs should be the focus. DCIs are obviously very important because they are reflective of the science content knowledge we expect students to have." "It is not the responsibility of assessments to play a social justice role. Science facts are inherently unbiased--focusing on empirical science ideas is the best way to support students."

	"All students won't be interested in all subjects in school, including science--it is unfair to expect assessment tasks to build confidence or attend to students' cultural backgrounds."

6. **Appropriate priorities for NGSS assessments.** Reviewers had many different perspectives about appropriate targets for NGSS assessments, and how these contribute to overall judgments about student proficiency. These included degree of sense-making; degree of student inclination; degree of transfer, requirements for multiple aspects/elements of individual SEPs and CCCs; use of multiple SEPs, CCCs and DCIs together; integration across multiple domains; foregrounding of SEPs; Foregrounding of DCIs; degree of sophistication expected for each dimension. While reviewers acknowledge that it would be unlikely that all priorities could be addressed in individual tasks, there were divergent perspectives about 1) which features were most important for monitoring student progress toward the vision of the *Framework* and NGSS, and 2) whether all of these were appropriate targets.

Implications for shifts in assessment development

The task evaluations revealed broad consensus that assessments that are to support NGSS learning must do more than just assessing the three dimensions of a Performance Expectation. Indeed, design and development procedures used prior to NGSS need to be modified to address a new set of requirements for high-quality assessments. The experts that participated in this study largely agree that many of these changes fall into the following four categories, though there are differing opinions as to some of the specific features required of new assessments:

1. **Context**. NGSS assessments need to present scenarios based on a context that is grounded in a real problem and phenomenon, and students should need to use the three dimensions to *figure something out* about the scenario. This means that students do not just answer questions that are related to a real scenario, but they answer questions to reveal something they didn't already know about it using grade-appropriate representations of evidence. There are different ideas, however, about how relevant and coherent the task scenario needs to be to students and whether or not it can be a general phenomenon (e.g., a volcano) or if it needs to be based on a specific instance (e.g., the 1980 eruption of Mt St Helens).

2. **Three dimensions.** NGSS assessments must require students to use multiple dimensions to engage in sensemaking. In other words, students must be asked to demonstrate their understanding of a phenomenon or problem using a combination of Disciplinary Core Ideas (DCI), Crosscutting Concepts (CCC), and Science and Engineering Practices (SEP). The DCI and SEP, in particular, must be used together to respond to a task. There is some debate as to the role of the crosscutting concept in assessments, particularly the degree to which the assessments must elicit explicit evidence of the CCC.

3. **Scoring**. NGSS assessments must include guidance for interpreting students' responses across all of the dimensions that the task is designed to assess. If an assessment is meant to support 3-dimensional classroom instruction, for example, the scoring guide must give teachers the information they need to evaluate students' proficiency with all three dimensions. Moreover, if a task just assesses parts of each dimension, the scoring guide should indicate which parts of the dimension are being measured and which parts are not.

4. **Equity.** Fairness and equitability is a priority for all assessments. The prominence of phenomena and problems in NGSS assessments in particular demands special attention to the accessibility of the task context. In addition, task prompts that elicit evidence of multiple dimensions tend to require substantial written responses. These requirements introduce new considerations for determinations of equity for assessments, but there are many different perspectives on what assessments must do to address them. For example, some experts believe that tasks should prioritize multimodality to provide multiple avenues for students to access the context and provide evidence of their proficiency with the Performance Expectations. Others believe that limiting the reading and writing load is sufficient for attending to the range in English reading and writing abilities. Some experts believe that assessments must initiate students' motivation or interest in the task by helping students connect the scenario to their own experiences or by putting students in the position of a scientist who is tasked with solving a problem. These reviewers believe motivation and engagement is critical to ensuring that all students engage with the task sufficiently to provide their optimal performance, while others believe that providing motivation should not a requirement for an assessment.

Conclusions

Assessments are often the clearest means of communication of what is expected of students, making them critical to the successful implementation of new standards. But they can only be effective in this role if the assessments themselves accurately represent the new goals. Through the evaluation of a comprehensive sample of the assessments being developed for NGSS, this study reveals that 1) many development efforts fail even to include the key features required to assess the NGSS, and 2) among those that are making useful and informative progress in defining how these features can be woven into assessments, there are dramatically different perspectives about how these features can -- and should -- be implemented. Professional learning for new developers of assessments for NGSS is essential to incorporate the critical features for high-quality assessment and to make the appropriate decisions about the best practices for implementing those features in their specific context.

References

Achieve, Inc. (2018). Criteria for Procuring and Evaluating Science Assessments. Washington: Achieve, Inc.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Palo Alto, CA: Stanford

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., … Hakuta, K. (2013). Criteria for high-quality assessment. *Stanford Center for Opportunity Policy in Education (Online).* Retrieved Jan, 2017 from *https://Edpolicy. Stanford. Edu/Sites/Default/Files/Publications/Criteria-Higher-Quality-Assessment_2. Pdf.*

Fuhrman, S. (1994). Politics and Systemic Education Reform. CPRE Policy Briefs.

Hannaway, J., & Hamilton, L. (2008). *Performance-based accountability policies: Implications for school and classroom practices.* Washington: Urban Institute and RAND Corporation.

National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academy Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. (J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty, Eds.). Washington, DC: National Academies Press.

National Research Council. (2015). *Guide to Implementing the Next Generation Science Standards*. Washington, DC.

Spillane, J. P. (1998). State policy and the non-monolithic nature of the local school district: Organizational and professional considerations. *American Educational Research Journal*, *35*(1), 33–63.

Science Task Prescreen

Introduction

The purpose of the Science Task Prescreen is to conduct a quick review of tasks to determine whether tasks might be designed for standards based on the *Framework for K-12 Science Education*, like the Next Generation Science Standards (NGSS). Because it is currently difficult to find or design three-dimensional assessment tasks, the prescreen is intended to reveal whether tasks include challenges—identified in this document as "red flags"—commonly found in science assessments.

Those interested in pursuing a more rigorous evaluation of tasks should use the Science Task Screener; however, the Task Screener assumes a deeper understanding of *A Framework for K–12 Science Education* and the NGSS. Those that are familiar with the development of tasks, but not very familiar with the Framework should start with the prescreen as a bridge for educators and developers to the differences in how to develop tasks for the NGSS and Framework. For those less familiar with the Framework, it will be particularly helpful to use the prescreen as part of a collaborative professional learning process, to help build a common understanding of the questions and what constitutes as evidence to address them.

Because the prescreen is a quick screening tool as opposed to a comprehensive evaluation tool, the questions in the prescreen focus on features that are non-negotiable, easily identified, and reflect the most serious "fatal flaws" seen in attempts to develop science tasks. While there are indeed many other critically important features of science assessments, they are excluded here for the purposes of screening, and are addressed in the Task Screener.

Using the Task Prescreen to evaluate science assessment tasks

The prescreen is organized around a short series of yes-or-no questions. In applying the prescreen to a task, follow these simple steps:

1. Read through the task and complete the task as though you were a student.
2. Read through any additional support materials for the task.
3. Answer the questions in the prescreen regarding the task and note any red flags.
4. Discuss the answers to the questions and evidence to support those answers with other reviewers.
5. Use your analysis to determine next steps for the task.

Because the prescreen is applied at the level of the task rather than individual questions, reviewers will need to answer the questions based on evidence from the task as a whole. Because the prescreen can be used on tasks designed for a number of purposes and intended uses, there is no associated scoring guide—instead, reviewers should consider the red flags they have identified and determine, based on their needs, whether the assessment:

A. **Should be used as-is, without further evaluation.** This is most appropriate for lower-stakes assessments, such as minor assessments an individual teacher might use in the classroom. It is important to remember that even if no red flags are identified, the task may have major flaws.

B. **Warrants further review.** This might be particularly relevant for assessments that are used as major components of a lesson or unit; used across multiple classrooms or schools; or used in other high-impact, higher-stakes scenarios, such as tasks used as part of district- or state-wide assessment efforts. Red flags can be used to determine if the assessment has potential and to focus the major areas of improvement that might be needed.

C. **Should not be used.** Reviewers can use the red flags to determine that, for their current purposes, the task should simply not be used.

While it is possible for the prescreen to be applied by an individual, it is more powerful when used as part of a collaborative review process. While the questions are high level, they can drive very meaningful conversations and help reviewers come to a common understanding of features of NGSS tasks. Reviewers should carefully discuss their answers to the questions and the evidence in the task that led them to those answers to come to a common understanding of language and expectations.

Science Task Prescreen

Task Title_____ **Grade** _____ **Date** _____ **Rating:**_____

Targeted SEP:_____ **DCI:**_____ **CCC:** _____

Intended Task Purpose:_____

Before you begin: Complete the task as a student would. Then, consider any support materials, such as information about the task provided to teachers and answer keys/rubrics.

Prescreen: Answer the following high-level questions to identify any major red flags (🚩) in your task. If you find one or more red flags, consider the purpose of the task and the evidence gathered to determine whether the task warrants a deeper dive.

Question	Yes	No
1. Is there a phenomenon or problem driving the task?	☐ 🚩	☐
2. Can the majority of the task be answered without using information provided by the task scenario?	☐	☐ 🚩
3. Can significant portions of the task be answered successfully by using rote knowledge (e.g., definitions, prescriptive or memorized procedure)?	☐ 🚩	☐
4. Does the majority of the task require students to use reasoning to successfully complete the task?	☐	☐ 🚩
5. Does the task require students to use some understanding of disciplinary ideas to successfully complete the task?	☐	☐ 🚩
6. Do students have to use at least one science and engineering practice to successfully complete the task?	☐	☐ 🚩
7. Are the dimensions assessed separately in the majority of the task?	☐ 🚩	☐
8. Is the task coherent and comprehensible from the student perspective?	☐	☐ 🚩

Based on your assessment needs, make a recommendation about this task moving forward (choose 1) :

☐ Should be used as-is, without further evaluation.

☐ Warrants further review.

☐ Should not be used.

Summarize your evidence and reasoning:

Science Task Screener

Introduction

The purpose of the Next Generation Science Standards (NGSS) Task Screener is **1)** to determine whether classroom assessment tasks are high quality, designed to elicit evidence of three-dimensional performances, and designed to support the purpose for which they will be used, and **2)** to provide a group of reviewers with a common set of features to ground conversations about what it "looks like" for students to demonstrate the kinds of performances expected by three-dimensional standards. This Screener builds off the criteria in Category III of the EQuIP Rubric for Science by more clearly specifying features for the assessment tasks embedded in lessons and units.

The directions for using the Task Screener assume an understanding of *A Framework for K–12 Science Education* and the NGSS, including how the NGSS are different from past standards as outlined in Appendix A of the NGSS and the Innovations of the NGSS. The Task Screener focuses on determining whether what is new and different about three-dimensional expectations are accurately represented in the tasks being evaluated.

Task Screener Overview

The Task Screener is organized around four criteria:

A. Tasks are driven by high-quality scenarios that focus on phenomena or problems.

B. Tasks require sense-making using the three dimensions.

C. Tasks are fair and equitable.

D. Tasks support their intended targets and purpose.

Each criterion includes:

1. A set of indicators to help reviewers determine whether the criterion is met
2. A set of response forms for gathering and analyzing evidence, providing suggestions for improvement, and rating the task

Using the Task Screener properly demands the collection of specific and detailed evidence to support claims about how well each criterion is addressed within a task.

While it is possible for the Screener to be applied by an individual, the real power of the Task Screener lies in the meaningful conversations it can drive among a team of reviewers as part of a collaborative process. Just as when using other resources in the EQuIP suite of tools, collaborative teams of users should:

1. Individually record criterion-based evidence using the provided response forms;
2. Individually make suggestions for improvement; and then
3. Collaboratively discuss findings with team members before checking one of the boxes under the "Evidence of Quality?" section included at the end of the screening process. As part of these discussions, reviewers should address any differences in how they interpreted the criteria and indicator language, as well as the evidence they found, to support a common understanding of the task, the expectations outlined in the screener, and how well the task met those expectations. A rating of "Adequate" means that the task meets the criterion. If the collaborative feedback is being used to improve the task or make decisions about how it should be used, use a blank set of response sheets to capture the consensus feedback.

Science Task Screener

Using the Task Screener. Use this screener tool to evaluate tasks designed for three dimensional standards. For each criterion, record your evidence for the presence or absence of the associated indicators. After you have decided to what degree the indicators are present within the task, revisit the purpose of your task and decide whether the evidence supports using the task.

Before you begin: Complete the task as a student would. Then, consider any support materials, such as information about the task provided to teachers and answer keys/scoring guidance.

A. Tasks are driven by high-quality scenarios that focus on phenomena or problems.	B. Tasks require sense-making using the three dimensions.
i. Making sense of a phenomenon or addressing a problem is necessary to accomplish the task. ii. The task scenario—grounded in the phenomena and problems being addressed—is sufficient, engaging, relevant, and accessible to a wide range of students.	i. Completing the task requires students to use reasoning to sense-make about phenomena or problems. ii. The task requires students to demonstrate grade-appropriate: a. SEP elements b. CCC elements c. DCI elements iii. The task requires students to integrate multiple dimensions in service of sense-making and problem-solving. iv. The task requires students to make their thinking visible.

C. Tasks are fair and equitable.	D. Tasks support their intended targets and purpose.
i. The task provides ways for students to make connections to meaningful local, global, or universal relevance. ii. Tasks include multiple modes for students to respond to the task. iii. The task is accessible, appropriate, and cognitively demanding for all learners, including students who are English learners or are working below or above grade level. iv. The task cultivates students' interest in and confidence with science and engineering. v. Tasks focus on performances for which students' learning experiences have prepared them (opportunity to learn considerations). vi. The task uses information that is scientifically accurate.	i. The task assesses what it is intended to assess, and supports the purpose for which it is intended. ii. The task elicits student artifacts that provide evidence of how well students can use the targeted dimensions together to make sense of phenomena and design solutions to problems. iii. Supporting materials include clear answer keys, rubrics, and/or scoring guidelines that are connected to the targeted three-dimensional standards and provide the necessary and sufficient guidance for interpreting student responses relative to all three dimensions and the target as a whole. iv. The task's prompts and directions provide sufficient guidance for the teacher to administer it effectively and for the students to complete it successfully while maintaining high levels of students' analytical thinking as appropriate.

Criterion A.

Tasks are driven by high-quality scenarios that focus on phenomena or problems.

<table>
<tr><th>Tasks designed for the NGSS include clear and compelling evidence that:</th><th>What was in the task, where was it, and why is this evidence?</th></tr>
<tr><td>i. Making sense of a phenomenon or addressing a problem is necessary to accomplish the task.</td><td>1) Is a phenomenon and/or problem present?

2) Is information from the scenario necessary to respond successfully to the task?</td></tr>
<tr><td>ii. The task scenario is engaging, relevant, and accessible to a wide range of students*.</td><td>Features of engaging, relevant, and accessible tasks (Check the appropriate box, then describe rationale with evidence)
<table>
<tr><th>Features of scenarios</th><th>Yes</th><th>Somewhat</th><th>No</th><th>Rationale</th></tr>
<tr><td>Scenario presents real-world observations</td><td>☐</td><td>☐</td><td>☐</td><td></td></tr>
<tr><td>Scenarios are based around at least one specific instance, not a topic, statement</td><td>☐</td><td>☐</td><td>☐</td><td></td></tr>
<tr><td>Scenarios are presented as puzzling/intriguing</td><td>☐</td><td>☐</td><td>☐</td><td></td></tr>
<tr><td>Scenarios create a “need to know”[1]</td><td>☐</td><td>☐</td><td>☐</td><td></td></tr>
<tr><td>Scenarios are explainable using grade-appropriate SEPs, CCCs, DCIs</td><td>☐</td><td>☐</td><td>☐</td><td></td></tr>
<tr><td>Scenarios effectively use at least 2 modalities</td><td>☐</td><td>☐</td><td>☐</td><td></td></tr>
</table></td></tr>
</table>

	Features of scenarios	Yes	Somewhat	No	Rationale
	If data are used, scenarios present real/well-crafted data	☐	☐	☐	
	The local, global, or universal relevance of the scenario is made clear to students[2]	☐	☐	☐	
	Scenarios are comprehensible to a wide range of students at grade-level	☐	☐	☐	
	Scenarios use as many words as needed, no more	☐	☐	☐	
	Scenarios are sufficiently rich to drive the task	☐	☐	☐	

Across all indicators, there is (choose one) **evidence of quality of this criterion.**

Suggestions for improvement of the task for Criterion A:

1. When considering whether the scenario creates a need to know for students, consider whether the scenario makes the uncertainty associated with explaining a phenomenon or solving a problem central, in ways that are likely to 1) connect with students' own experiences or knowledge, and 2) connect to disciplinary core ideas (regardless of whether those ideas are explicitly named or required by the task).

2. Consider whether an authentic stakeholder group is interested in the outcome of the scenario, and/or whether students are given enough information to answer the question "why should I care?".

Criterion B.
Tasks require sense-making using the three dimensions.

Tasks designed for the NGSS include clear and compelling evidence that:	What was in the task, where was it, and why is this evidence?	
i. Completing the task requires students to use reasoning to sense-make about phenomena or problems.	*Consider in what ways the task requires students to use reasoning to engage in sense-making and/or problem solving.*	
ii. The task requires students to demonstrate grade-appropriate: • SEP elements • CCC elements • DCI elements	Evidence of **SEPs** (which element, and how does the task require students to demonstrate this element in use?)	
	Evidence of **CCCs** (which element, and how does the task require students to demonstrate this element in use?)	
	Evidence of **DCIs** (which element, and how does the task require students to demonstrate this element in use?)	
iii. The task requires students to integrate multiple dimensions in service of sense-making and/or problem-solving.	*Consider in what ways the task requires students to use multiple dimensions together to sense-make and/or problem-solve.*	
iv. The task requires students to make their thinking visible.	*Consider in what ways the task visibly surfaces student thinking. Look for evidence of how the task surfaces current understanding, abilities, gaps, and misconceptions.*	

Across all indicators, there is (choose one) **evidence of quality of this criterion.**

Suggestions for improvement of the task for Criterion B:

Criterion C.
Tasks are fair and equitable.

Tasks designed for the NGSS include clear and compelling evidence of the following:	What was in the task, where was it, and why is this evidence?
i. The task provides ways for students to make connections to local, global, or universal relevance.	*Consider specific features of the task that enable students to make local, global, or universal connections to the phenomenon/problem and task at hand. Note: This criterion emphasizes ways for students to find meaning in the task; this does not mean "interest." Consider whether the task is a meaningful, valuable endeavor that some stakeholder group locally, globally, or universally would be invested in.*
ii. Tasks include multiple modes for students to respond to the task.	*Describe what modes (written, oral, video, simulation, direct observation, peer discussion, etc.) are expected/possible for student responses.*
iii. The task is accessible, appropriate, and cognitively demanding for all learners, including students who are English language learners or are working below or above grade level.	*Consider how the task supports all learners, including:*

	Yes	Somewhat	No	Rationale
Task includes appropriate scaffolds	☐	☐	☐	
Tasks are coherent from a student perspective	☐	☐	☐	
Tasks respect and advantage students' cultural and linguistic backgrounds	☐	☐	☐	

3. For more information about culturally and linguistically responsive classroom assessments, please see this resource.

Tasks designed for the NGSS include clear and compelling evidence of the following:	**What was in the task, where was it, and why is this evidence?**				
iii. *(continued)*		Yes	Somewhat	No	Rationale
	Tasks provide both low- and high-achieving students with an opportunity to show what they know	☐	☐	☐	
	Tasks use accessible language	☐	☐	☐	
iv. The task cultivates students' interest in and confidence with science and engineering.	*Consider how the task cultivates students interest in and confidence with science and engineering, including opportunities for students to reflect their own ideas as a meaningful part of the task; make decisions about how to approach a task; engage in peer/self-reflection; and engage with tasks that matter to students.*				
v. Tasks focus on performances for which students' learning experiences have prepared them (opportunity to learn considerations).	*Consider the ways in which provided information about students' prior learning (e.g., instructional materials, storylines, assumed instructional experiences) enables or prevents students' engagement with the task and educator interpretation of student responses.*				

Tasks designed for the NGSS include clear and compelling evidence of the following:	What was in the task, where was it, and why is this evidence?
vi. The task presents information that is scientifically accurate.	*Describe evidence for scientific inaccuracies promoted by the task.*

Criterion C. *continued*

Across all indicators, there is (choose one) **evidence of quality of this criterion.**

Suggestions for improvement of the task for Criterion C:

Criterion D.
Tasks support their intended targets and purpose.

Before you begin:

1. Describe what is being assessed. Include any targets provided, such as dimensions, elements, or PEs. :

2. What is the purpose of the assessment? (check all that apply)

- ☐ Formative (including peer and self-reflection)
- ☐ Summative
- ☐ Determining whether students learned what they just experienced
- ☐ Determining whether students can apply what they have learned to a similar but new context
- ☐ Determining whether students can generalize their learning to a very different context
- ☐ Other (please specify) _____

Tasks designed for the NGSS include clear and compelling evidence that:	What was in the task, where was it, and why is this evidence?
i. The task assesses what it is intended to assess and supports the purpose for which it is intended.	*Consider in what ways:* *1) Understanding and using all aspects of the assessment target is necessary to respond to the task?* *2) Any ideas, practices, or experiences* ***not*** *targeted by the assessment are necessary to respond to the task? Consider the impact this has on students' ability to complete the task and interpretation of student responses.* *3) The student responses elicited support the purpose of the task? (e.g., if a task is intended to help teachers determine if students understand the distinction between cause and correlation, does the task support this inference?)*

Tasks designed for the NGSS include clear and compelling evidence that:	What was in the task, where was it, and why is this evidence?
ii. The task elicits artifacts from students as direct, observable evidence of how well students can use the targeted dimensions together to make sense of phenomena and design solutions to problems.	*Consider what student artifacts are produced and how these provide students the opportunity to make **1)** sense-making processes, **2)** thinking across all three dimensions, and **3)** ability to use multiple dimensions together visible [note: these artifacts should connect back to the evidence described for Criterion B.]*
iii. Supporting materials include clear answer keys, rubrics, and/or scoring guidelines that are connected to the three-dimensional target. They provide the necessary and sufficient guidance for interpreting student responses relative to the purpose of the assessment, all targeted dimensions, and the three-dimensional target.	*Consider how well the materials support teachers and students in making sense of student responses and planning for follow up (grading, instructional moves), consistent with the purpose of and targets for the assessment. Consider in what ways rubrics include:* *1) Guidance for interpreting student thinking using in integrated approach, considering all three dimensions together as well as calling out specific supports for individual dimensions if appropriate:* *2) Support for interpreting a range of student responses, including those that might reflect partial scientific understanding or mask/misrepresent students' actual science understanding [e.g., because of language barriers, lack of prompting or disconnect between the intent and student interpretation of the task, variety in communication approaches]:* *3) Ways to connect student responses to prior experiences and future planned instruction by teachers and participation by students:*

Tasks designed for the NGSS include clear and compelling evidence that:	What was in the task, where was it, and why is this evidence?
iv. The task's prompts and directions provide sufficient guidance for the teacher to administer it effectively and for the students to complete it successfully while maintaining high levels of students' analytical thinking as appropriate.	*Consider any confusing prompts or directions, and evidence for too much or too little scaffolding/supports for students (relative to the target of the assessment—e.g., a task is intended to elicit student understanding of a DCI, but their response is so heavily scripted that it prevents students from actually showing their ability to apply the DCI).*

Across all indicators, there is (choose one) **evidence of quality of this criterion.**

Suggestions for improvement of the task for Criterion D:

Overall Summary

Consider the task purpose and the evidence you gathered for each criterion. Carefully consider the assessment purpose and intended use, your evidence, reasoning, and ratings to make a summary recommendation about using this task. While general guidance is provided below, it is important to remember that the intended use of the task plays a big role in determining whether the task is worth students' and teachers' time.

Final recommendation

☐ Use this task (all criteria had at least an "adequate" rating)

☐ Modify and use this task

☐ Do not use this task