From Classroom Tasks to Assessment Systems: Implications for Assessing Science Learning at Scale.

CONNECTION TO ASSESSMENT PURPOSE

The *Framework for K-12 Science Education* and standards based on it establish ambitious and exciting goals for student performance. There are a wide range of inferences we want to make about student learning to make sure all students are supported in achieving that vision, and we need to develop, select, and use assessments that will surface the right evidence at the right time for the right purpose so that we can help students meet those goals. By examining individual tasks designed for three-dimensiona standards, the Task Annotation Project in Science (TAPS) revealed the kinds of considerations that must be weighed when deciding whether assessments are designed for three-dimensional standards and can support students' science learning effectively. These lessons learned have implications for the design of all science assessments, from classroom-level formative processes and summative assessments to district- and state-wide assessment systems.

What does this mean for systems of assessment in science?

There are **two primary issues** that need to be navigated:

"Tasks designed for three-dimensional standards" is an umbrella for many different interpretations of student learning and performance goals.

One of the reasons that it is so difficult to define what alignment means for three-dimensional standards is the sheer range of priorities, vision, and definitions different stakeholders have when they use the phrase "three-dimensional task." The phrase has become short-hand for a wide range of very different assessments designed for different purposes. The reality is that despite our desire to learn everything about what a student knows and can do from a single task, no single assessment can provide evidence that is useful to serve the wide range of purposes for which we use assessments.

For each goal and purpose, there are a series of trade-offs to consider that change how tasks reveal student progress.

Given the range of goals we have for student learning, it is no surprise that no single science assessment task or instrument, or any one particular approach to science tasks, is going to surface everything we want to know about student learning and performance. What is important is that we are clear about what is and is not being assessed, why those choices were made, and what inferences we should and should not make about student learning based on their responses to that task.

Decision-makers can intentionally use the range of inferences and trade-offs that need to be considered to drive the design of high-quality science assessment systems in their school, district, and state.

Using this resource: a guide for science assessment system decision-makers and leaders

This guide describes the kinds of decisions and features leaders should consider when designing science assessments, from individual tasks and items to full-scale assessment systems. This guide can be useful for:

- **Designing specifications for a coherent system of science assessments** within a state or district. For more information about systems of assessment in science, see this suite of resources.
- **Communicating common expectations to vendors and developers,** so that the assessments designed meet the needs of the stakeholders using them.
- **Fostering collaboration with other schools, districts, and states** by developing a common understanding of what science assessments can and should look like, and agreeing on non-negotiable features and high priorities to meet the needs across multiple contributors and users.

How can decision-makers navigate these issues to create effective science assessment systems?

Step 1: Decide: What do we want to know about student learning?

Recommendation: Before designing or selecting assessments, be clear about which inferences of student learning are priorities for particular assessments. Be transparent about your focus, and plan your tasks, scoring, and feedback accordingly.

When making decisions about 1) how to design or modify an assessment, and 2) whether an assessment (task, or series of tasks) is the right assessment for a given purpose, it is important to be clear about what inferences or claims are going to be made about student performance. The table below highlights some common inferences that stakeholders frequently want to make about students' science learning—each inference requires different kinds of evidence from tasks, one task may provide evidence for more than one inference if designed well, and no single assessment task will be able to provide evidence to support all inferences. While this is not new or specific to science assessments, the three-dimensional nature of the standards, coupled with the student-centered learning goals described by the *Framework*, make this a more nuanced and particularly important conversation to have when designing and using high-quality three-dimensional assessments.

Common inferences stakeholders want to make about students' science learning	
DIMENSION-SPECIFIC	Dimension-specific inferences, such as the degree to which students: • understand components of specific DCIs, SEPs, CCCs. • can use multiple aspects of specific DCIs, SEPs, CCCs.
THREE-DIMENSIONAL	Three-dimensional inferences, such as the degree to which students: • can use specific, grade-appropriate elements of three dimensions together in service of sense-making. • can make sense of phenomena or solve problems, using any dimensions as appropriate. • can use multiple SEPs, DCIs, CCCs together to make sense of phenomena or solve problems.
SCIENCE COMPETENCIES	Inferences about other competencies in science connected to three-dimensional performances, such as the degree to which students: • can access and engage with highly ambiguous or uncertain scenarios. • can flexibly make sense of diverse kinds of phenomena and problems. • can transfer their understanding across contexts. • Exhibit creative and flexible thinking and problem solving. • Use evidence- and reasoning-based thinking/ demonstrate science literacy.

Step 2: Surface and Communicate: How do we want students to show us their learning?

Recommendation 2: Deliberate with stakeholders and make decisions about how different trade-offs will be made to surface evidence of student learning that aligns with both the intended use of the assessment as well as the goals, values, and perspectives at the table.

There are a variety of ways assessment tasks can surface evidence to support or refute each inference about student learning. Once decisions about the purpose of a task are made, it is important to think about how assessments will reveal student learning. These considerations, discussed below, interact with both the purpose of the assessment (what kinds of inferences we want to make about student learning, and what kind of feedback we want to share) as well as the needs and perspectives of stakeholders (how much does the group value features like student choice, authentic sense-making, etc.). Having honest conversations about what are the foregrounded priorities for particular assessments and specific purposes, and what will be foregrounded elsewhere in the system, can help make sure all values and concerns are addressed, and that all stakeholders have a better grasp of how to interpret student performance and use that information.

Some common considerations decision-makers and stakeholder groups may want to consider include:

Features	What are the tradeoffs and implications for a task?
DCIs and Phenomena	***The issue:*** in some tasks, the goal of a three-dimensional performance is to show whether students deeply understand a DCI or set of DCIs ; in other tasks, the goal of three-dimensional performance is to show whether students can make sense of phenomena, using the DCI as appropriate. In an ideal situation, both of these happen at the same time; in practice, many assessments prioritize one or the other for various reasons.
	When **surfacing DCI understanding** is the goal, we often see… • Students conceptual understanding of a DCI can be probed more deeply. • SEPs and CCCs are used in service of showing the DCI, rather than their use as sense-making tools. • The phenomenon- or problem-based scenario often feels more contrived/like "school science" because they have been chosen to surface the DCI more completely rather than to provide a context for sense-making. • Sense-making about a phenomenon is de-emphasized in service of describing facets of a core idea.
	When **making sense of phenomena** is the goal, we often see… • Sense-making using the three dimensions is foregrounded. • The phenomenon- or problem-driven scenario is often more authentic/meaningful to students. • SEPs and CCCs can be used in service of sense-making more easily. • The task surfaces evidence of student thinking using the parts of the three dimensions, including the DCI. Only parts of the dimensions may be assessed because they were most relevant to the phenomenon, while other parts may not be assessed.
Depth and Breadth	***The issue:*** in some tasks, student thinking with a limited set of SEPs, CCCs, and DCIs are assessed deeply; in other tasks, small grainsize samples of many SEPs, CCCs, and DCIs are assessed.
	When **depth** is foregrounded, we often see… • More complete and varied evidence of student thinking related to a small set of targets. • Multiple elements of the same targeted SEPs, CCCs, and DCIs assessed. • Many opportunities for students to make facets of their thinking visible. • Specific and in-depth feedback about a targeted set of learning goals.
	When **breadth** is foregrounded, we often see… • Wider range of evidence of student thinking across a wide range of learning targets, including multiple SEPs, CCCs, and DCIs. • Small samples of student performance used as a proxy for a wide range of wider learning targets (e.g., 1 or 2 elements used as an indicator of a whole dimension). • More general feedback about wider learning targets.
How much transfer is expected.	***The issue:*** Some tasks focus on making sure that students understand recent learning goals in contexts that are the same as or very similar to the learning context (embedded or near transfer). In other tasks, students are asked to apply their learning to contexts that might be quite removed and different from the learning context (far transfer).
	When **embedded or near transfer** is prioritized, we often see… • Tasks that are useful for informing instructional moves. • Tasks that give students the opportunity to practice using DCIs, SEPs, and CCCs in familiar ways. • Tasks that can easily be used both as part of teaching and learning as well as to assess student progress. • Performances that may not reveal students' capacity to generalize from phenomena or problems they are studying to new problems.
	When **far transfer** is prioritized, we often see… • Tasks that can be useful for informing instruction and that reveal if students can use what they have learned so far in diverse contexts. • Tasks that give students the opportunity to connect and use SEPs, CCCs, and DCIs developed through diverse learning experiences to make sense of unfamiliar phenomena and problems. • Tasks that require careful analysis to ensure comparability and that students have had adequate opportunity to learn.

Features	What are the tradeoffs and implications for a task?

Choice and Specificity

The issue: iIn some tasks, student choice is prioritized through ways of approaching tasks, the content students engage with, and how they communicate their thinking—this inherently requires some flexibility in the exact nature of the sense-making with the SEPs, CCCs, and DCIs that are assessed. In other tasks, the goal is to determine whether students know and can use a specific set of SEPs, CCCs, and DCIs, limiting the choices students make.

When **student choice** is prioritized, we often see…

- A variety of SEPs, CCCs, and DCIs engaged in sense-making that may differ across students (although not necessarily, depending on the nature of the choices).
- High levels of student ownership, confidence, and interest.
- High levels student engagement.
- A window into student inclination, competencies, and facets of student learning that students choose to make visible (e.g., opportunities for creative thinking, creative communication styles, etc.) with an emphasis on "are students able to address this kind of phenomenon or problem, and how do they go about doing so?"

When **specificity** is prioritized, we often see…

- Students engaging in a specific set of SEPs, CCCs, and DCIs that are common across students and tied to a specific learning goal.
- Student thinking relative to a specific set of learning goals, with an emphasis on "do students understand and are able to use this targeted learning?"
- Higher degree of comparability across students' performances.
- Opportunities to compare student performance and feedback.

How much student thinking can be made visible.

The issue: Making student thinking visible is critical. In some tasks, students are asked to make each step of their thinking visible, through written descriptions, models, discourse, etc. In other tasks, a simpler student performance, such as selecting a response or making a prediction, is used as evidence of student thinking.

When **each step of student thinking** is prioritized, we often see…

- More complete and varied evidence of student thinking related to a small set of targets.
- Multiple elements of the same targeted SEPs, CCCs, and DCIs assessed.
- Many opportunities for students to make facets of their thinking visible.
- Specific and in-depth feedback about a targeted set of learning goals.

When **simpler student performances** are used, we often see…

- Shorter tasks for teachers and students that are faster to engage with and provide feedback.
- Tasks that are more likely to be a quick check on learning/representation of ideas.
- Less writing/language expected from students, among fewer ways to make student thinking visible.
- Fewer facets of students' current understanding are revealed.

It should be noted that these are just a sample of the kinds of considerations and trade-offs that should be part of assessment conversations, and these are generally not "either-or" decisions, but rather should be viewed as sliding scales—it is certainly possible to have assessments that live in the middle of the kinds of considerations cited above, but similar trade-offs are still being made in those instances. Some other features stakeholders may consider include:

- Degree to which factors that promote student agency and identity in science are emphasized.
- Degree of sophistication of each dimension.
- Time for feedback.
- Cultural relevance.
- Degree of integration of science with other subject areas.

Bottom line: If we want to support all students in science, we need to monitor progress across the full range of inferences and trade-offs, distributed across a system of assessments appropriately.

Connecting inferences and trade-offs: examples

How do inferences and trade-offs manifest in an assessment task?

Sample Middle School Task:

HEATED CUP OF WATER

Thermal energy is slowly transferred to a cup of water by heating it in a microwave, but there is no change in the state of water. Imagine that you had a very powerful tool that allowed you to see how the water molecules are moving after thermal energy is transferred to the water by heating.

Construct a model to explain the motion of the water molecules before and after water is heated.

Be sure your model includes pictures and a key.

Explore the full annotation of this middle school physical science task.

Once we start thinking about assessments as a series of tradeoffs, we can move away from thinking of assessments as "good" and "bad", and toward thinking about what choices were made, and how this influences how the assessment can be used. As an example, consider this middle school physical science task that focuses on what happens to water molecules when they are heated. Some trade-offs that were made include:

- Focusing on demonstrating their understanding of a DCI using an SEP in the context of phenomenon, rather than focusing on making sense of a phenomenon.
- Emphasizing making their thinking visible, requiring more time from students and teachers for the scope of the assessment.
- Focusing on depth of student understanding related to a particular DCI and SEP target, rather than breadth across multiple targets.
- Emphasizing specificity of the assessment target.

Based on these tradeoffs...

Some inferences this assessment **can** support:

1. How well students understand a specific DCI element.
2. Whether students can use their understanding of diagrammatic models to show their DCI understanding effectively.

Some inferences this assessment **should not be used to** support:

1. Whether students can use the three dimensions together to make sense of phenomena.
2. Whether students can transfer their understanding of properties of matter flexibly across contexts.

This assessment is appropriate to meet a specific need, but if every assessment task students complete made the same tradeoffs, it would not support the full range of inferences we need to be able to make to support students' science learning.

How can inferences and trade-offs support system design decisions? A state-level example.

Two states decided they would like to collaborate on the design of interim assessment tasks that teachers could use in the classroom to monitor student progress toward end-of-year learning goals. During the process, state leaders and task developers realized that although they all had the same goal in mind of developing three-dimensional, curriculum agnostic tasks, they had very different ideas about what the final products should look like. State leaders decided to lead the development teams through a conversation that identified 1) which claims about student learning these tasks should be able to make, 2) what were the constant features that needed to be in each task, regardless of approach, and 3) what trade-offs each group needed to consider and surface.

This process surfaced several different priorities and trade-offs task developers were implicitly making, which resulted in tensions during the development process (e.g., developers who prioritized core ideas vs. those more interested in flexible sense-making; those who wanted interim tasks to mirror the design decisions of the statewide summative vs. those who wanted the interim assessments to provide complementary information).

When these differences surfaced, state leaders were able to diagnose the issues; transparently consider trade-offs, and decide, as a community, which features needed to be foregrounded consistently, and which features could vary across different tasks; look at all of the tasks—across groups—and see where and how often different trade-offs were being made; develop clear instructions for using the assessments, including how student responses should be interpreted, what was and was not being assessed, and providing clear pointers about additional evidence of student learning needed; and ensure that each state's internal system to ensure coherence within individual systems and identify other opportunities for collaboration.

Potential Pitfalls

As these conversations and decisions proceed, be aware of some "red flags" that might emerge that would not be appropriate for effectively monitoring student learning toward three-dimensional science goals:

Red Flag	What is the issue?	Why is this a problem?
Equity and fairness are sacrificed as a trade-off.	Decisions that negatively impact whether tasks are fair and equitable are justified as an appropriate trade-off.	Tasks that do not support diverse learners do not provide valid information for making any inferences about student performance.
The same trade-offs are made every time.	All the assessments students see across a year or grade-band are the same kinds of assessments, making the same trade-offs —and therefore supporting the same inferences—each time.	Supporting students' science learning requires that we build evidence to support a wide range of inferences—if the same choices about assessments are made every time, we are only surfacing a small piece of student learning.
There is a mismatch between claims and evidence.	Scores, grades, or other reports/feedback of student progress and performance make claims about aspects of three-dimensional science performances they do not support.	Feedback is not valid unless the tasks surface evidence that supports those assertions.
Trade-offs are used to justify tasks that don't meet the non-negotiable features of 3D assessments.	A combination of the inferences and trade-offs discussed above is used to justify tasks that are rote; one-dimensional; not connected to a phenomenon or problem, etc.	Three-dimensional tasks must meet certain baseline features to support the goals of standards based on *A Framework for K-12 Science Education*. For more information, see the Science Task Prescreen and Task Screener.

Implications for Decision-Makers, Educators, and Developers.

Decision-makers should:

- Plan assessment systems intentionally, so that assessments are coherent and complementary from the classroom to external assessments, and across K-12. For more support in how to plan your system, see these resources and this guide.
- Have open conversations with developers/vendors such that everyone is clear about priorities and non-negotiables, what trad-eoffs are appropriate, what evidence is being surfaced, and how that information should be used.
- Be clear about how this information should be used by educators, students, families, and policy makers.

Educators should:

- Be clear about which inferences are important to make at different times, and select assessments accordingly.
- Be discerning and advocate for what you, your students, and families need to help all your students succeed.

Developers should:

- Decide which inferences are important to support in the assessments they are developing. Those developing assessment banks or a single assessment instrument will be making different decisions than those developing classroom assessments as part of instructional materials.
- Be clear and transparent about which inferences you are supporting, which trade-offs you are making, and how this influences who uses the resulting information, and for what purpose.
- Make sure to include the non-negotiable features of 3D assessments, regardless of the other trade-offs being made.