# The State of *Framework*-aligned Assessment Tasks: Where are we?

Clarissa Deverel-Rico, Patricia Olson, Cari F. Herrmann Abell, and Chris D. Wilson
*BSCS Science Learning*

*Paper presented at the NARST 2024 Annual Conference in Denver, Colorado.*

**Abstract**: The emphasis on an equitable vision of science learning in current science education reform efforts sees students as contributing to knowledge-building through drawing on their rich cultural and linguistic backgrounds while engaging in the three dimensions to make sense of compelling, relevant phenomena. However, this vision will not be fully realized without coherence between curriculum, instruction, and assessment. As a majority of states have now adopted standards aligned to or adapted from the *Framework*, we see an urgent need for assessments that can support rather than conflict with equitable science learning. In this study, we seek to understand the current state of *Framework*-aligned assessment tasks. We have amassed 352 middle school tasks, originating from state-level assessment banks and assessment developers at universities or research organizations. Our preliminary findings from characterizing 104 tasks revealed that the majority of tasks target dimensions of the NGSS or *Framework*-based standards and include a phenomenon. However, there are challenges in framing phenomena that attend to students' interests and identities and engage students in three-dimensional sensemaking. Additionally, some phenomena are not based in real-world observations and are not authentic from students' perspectives, which makes it difficult for students to see connections of local or global relevance.

Corresponding Author: Clarissa Deverel-Rico, cdeverelrico@bscs.org

## Introduction

The science education reforms represented by the *Framework for K-12 Science Education* (NRC, 2012), along with other recent reports (NASEM, 2018; 2019), depict a shift away from previous ways of experiencing science in school as learning discrete facts through rote means towards students collaboratively figuring out real-world phenomena or solving design problems. This more equitable and student-centered vision of science education, which draws on sociocultural views of learning (Rogoff, 2003), sees students as contributing to knowledge-building through drawing on their rich cultural and linguistic backgrounds while engaging in the three dimensions that organize science learning in *Framework*-era reforms - the science and engineering practices, crosscutting concepts, and disciplinary core ideas - to make sense of compelling, relevant phenomena.

However, this vision will not be realized without coherence between curriculum, instruction, and assessment (Shepard et al., 2018). Curriculum and instruction have made significant progress beyond the traditional textbook and “cookbook lab” approach - through phenomenon-based, storyline curricular materials (Reiser et al., 2021) aligned with the *Framework's* vision that are designed to be educative (Davis & Krajcik, 2005) and are coupled with high-quality professional learning. Examples include BSCS Science Learning’s biology unit, Understanding for Life (BSCS Science Learning, 2023), OpenSciEd’s K-12 curricula (OpenSciEd, n.d.), and Learning in Places’ outdoor, place-based curricular resources and pedagogy (Learning in Places Collaborative, 2021).

While curriculum and instruction have made great strides, assessment has tended to continue to align with more traditional approaches through privileging canonical concepts above students’ experiences (Bang et al., 2013; Randall, 2021), and previous views on learning through focusing more on the knowledge contained in students’ heads (Penuel & Shepard, 2016; Shepard et al., 2018). With the majority of states having now adopted standards aligned to or adapted from the Next Generation Science Standards (NGSS; NRC, 2013), we see an urgent need for assessments that can support rather than conflict with the promotion of equitable science learning. Thus, we seek to understand the current state of purportedly *Framework*-aligned assessment tasks so that we may better understand where we are and note areas for improving assessment moving forward. We have gathered 352 middle school tasks from a wide range of state-level assessment banks and university or research groups as part of a larger project where we ultimately intend to validate a pool of *Framework*-aligned middle school science assessments for use with high-quality instructional materials. Through this first stage of our research, we ask, to what extent do currently available classroom assessments support the vision of *Framework*-era reforms which emphasizes more equitable learning opportunities for students?

## Conceptual Framing

Our conceptual framework is founded on the role of assessment as part of a coherent system that includes curriculum and instruction (Shepard et al., 2018) - one that is in line with sociocultural perspectives on learning (Rogoff, 2003) and supports a more equitable vision of how students experience science in school. There is broad consensus in the field of education that sociocultural perspectives offer the more encompassing theory of knowing and learning in accounting for the social nature of and cultural influences on learning. In building from cognitive perspectives, sociocultural theory represents a departure from focusing only on what is inside students' heads, and instead helps us see the larger ecology of the classroom and the many influences that contribute to learning (e.g., NASEM, 2018; Shepard et al., 2018). Previous notions of students learning through rote means or performing "cookbook-style" labs drew more on cognitive approaches to learning, while in the contemporary, reform-oriented views of science learning, students draw on their diverse backgrounds and experiences, pursue meaningful contexts, and engage with science learning in ways that foster seeing science as a valuable endeavor in their lives (NRC, 2012; NASEM, 2018; 2019). Further, sociocultural perspectives can be seen reflected in current reforms through the emphasis on sensemaking of phenomena, and through gaining access to and participating in disciplinary practice (Ford & Forman, 2006). Through such authentic activity (Brown et al., 1989), opportunities can be opened up for students to bring their interests and identities, which have historically been left at the classroom door, into science learning.

Next, we describe the qualities of assessments that better align with these perspectives and commitments to how students experience science learning.

## Approaches to Science Assessment that Cohere with the *Framework*'s Vision

In looking beyond the knowledge that students might accumulate, sociocultural perspectives see learning as entering into a community of practice, gaining knowledge and experience with the norms, values, and practices of that community. Thus, such perspectives compel us to rethink assessment to center this definition of what it means to learn, along with attending to the social and cultural contexts that shape student learning (Pellegrino et al., 2023). Further, in attending to equity in the *Framework*, the writers made more visible the existing inequities in science education and the need to remedy current practices so that all students have opportunities to participate in and succeed in science learning (NRC, 2012).

These perspectives and commitments extend to assessment to consider how practices in assessment design and use can increase access and achievement for the diverse students in today's classrooms (Furtak & Lee, 2023; Philip & Azevedo, 2017). Many contemporary reports and scholars articulate how classroom assessment can embody such commitments to sociocultural perspectives and equity (e.g. Harris et al, 2023; NASEM, 2017; Penuel et al, 2019; Fine & Furtak, 2020, Shepard, 2021). In this paper, we draw on Furtak and colleagues' (2020) emergent design heuristics which expands on the original design criteria put forth by the report, *Developing Assessments for the Next Generation Science Standards* (NRC, 2014). Six criteria for equity-driven assessments emerged from the work of five research teams: "Based on relevant phenomenon or scenario, Explicit attention to language, Includes scaffolds, Explicit attention to identities of learners, Engages and supports student sensemaking, Accompanied by tools and routines to support teacher implementation" (Furtak et al, 2020, p. 1492). Lastly, science assessments should also align with the science standards and be based on theorized models of learning in the discipline (NRC, 2014; Shepard et al., 2018; Harris et al., 2023); see Wertheim et al (2016) or Harris et al (2023) for a thorough discussion of how tasks can integrate the three dimensions of the NGSS into assessment in developmentally appropriate ways and tools for unpacking the three dimensions in learner-centered ways (Inquiry Hub & BSCS Science Learning, n.d.; State Performance Assessment Learning Community, n.d.).

The criteria we focus on are tied to the sociocultural underpinnings of the *Framework* which is seen through attention to the situated nature of learning by highlighting the importance of grounding learning in phenomena, through an emphasis on collaborative sensemaking of phenomena rather than a focus on learning through rote means, and in attending to students' interests and identities as key resources for learning science.

### ***Phenomena***

In the current era of science reforms, grounding learning in real-world, observable phenomena attends to the situated nature of learning, supporting students in seeing how science connects to their lives and a coherent view of learning from students' perspectives (Edelson et al., 2021; Penuel et al., 2019). The curricula mentioned earlier, for example, start with an anchor event which motivates the figuring out to be done throughout the instructional unit. In OpenSciEd's 7th grade unit focused on metabolic reactions, for example, the anchoring phenomenon is a middle school girl experiencing fatigue and digestion issues without a clear cause (OpenSciEd, 2022). This motivates investigations into various body systems to figure out the cause of the issue. There are different types of phenomena that can anchor such curricular units: generalized or contextualized (Kang et al, 2014), everyday, contemporary scientific, societally-relevant, culturally-relevant (Suarez & Bell, 2019).

Just as phenomena are considered crucial for day-to-day learning, they are also important for assessment. Further, as phenomena are based on observable events, students need data to notice and interact with; data can be presented or collected by students and be quantitative or qualitative. For instance, a task from the Kentucky Department of Education (2021) presents observations that vegetables grow bigger in Alaska than Kentucky, highlighting images of enormous vegetables from Alaska, and includes quantitative data about average rainfall, temperature, and hours of sunlight in both locations throughout the year.

### *Explicit Attention to Students' Interests and Identities*

A key assumption of the current reform efforts is based on research that finds connecting learning to students' experiences and backgrounds is critical for learning at school and may influence the choice to pursue a career in a STEM field (e.g. NRC, 2012). This is important for a vision of science education committed to broadening participation in science and providing opportunities for students to see how science connects to and can be useful for their own lives and communities (Tan & Calabrese Barton, 2012).

For assessments to be considered valid and reliable indicators of learning, students need to have had the opportunity to learn the ideas and practices that are being assessed, and to have occasion to demonstrate what they know and can do through meaningful connections with their interests and identities (Penuel et al., 2019). For example, in the Science And Integrated Language (SAIL) 5th grade curriculum centered on water quality, students argue in an assessment task whether bottled water or tap water is a better water source in their city (NYU SAIL, n.d.). Further, assessment tasks need to be accessible, and include scaffolds that support students' access to using disciplinary practice and knowledge in order to explain a phenomenon or solve a design problem. Tasks must also be attentive to language; in part, to reduce construct-irrelevant variance through an overreliance on the ability to read and write in standard English. Allowing students to respond with their full linguistic repertoires and providing multiple modalities for demonstrating learning are options for mitigating this (Achieve, 2019). Taking these steps can support students in giving them more equitable opportunities to show what they know and can do by disrupting assessment design that has historically reflected only the dominant culture of American schools or science (Randall, 2021; Randall et al, 2021).

### ***Sensemaking***

Sensemaking from a sociocultural perspective focuses on the figuring out that occurs collaboratively as students engage in the three dimensions, while drawing on previous experiences and backgrounds, working towards developing an explanation of a phenomenon or solving a design problem (Odden & Russ, 2019). In assessment, this means moving away from regurgitating memorized facts or representing what has already been learned to leveraging prior experiences and learning towards figuring out a novel phenomenon or problem. Further, sensemaking is cued through the problematization (Phillips et al., 2017) of phenomena, which makes clear to students what's at stake, and motivates the targeted sensemaking (Badrinarayan et al., 2023). This is a distinction from assessment opportunities that may 'cover' the three dimensions or include them in decontextualized ways. For instance, a task item could ask students to lay out the steps of an investigation but not include a phenomenon for students to decide how to design an investigation around. Or, students may be prompted to model the chemical reaction that occurs in cellular respiration, but without providing any observations or data to motivate students and to support sensemaking. An alternative to this latter example could be to present the phenomenon of how you might get out of breath at higher elevations and then ask students to model an explanation using cellular respiration, which results in demonstrating how oxygen is a limiting factor in this phenomenon.

## Towards Understanding the State of Assessments

An analysis of the extant science assessments available within two years of the publication of the NGSS made by Jill Wertheim and colleagues (2016) showed that the gaps to work on were to "align with each of the three dimensions of the NGSS, focus on big ideas in science, probe science and engineering practices in a way that engages students in reasoning with evidence; and give students a platform in which they can draw on their knowledge and skills as needed to investigate scientific questions and problems" (p. 42). While these are important criteria to continue to attend to in developing *Framework*-aligned assessments, we build on these and look towards what may now be the more urgent challenge in assessment - to what extent do currently available classroom assessments support the vision of *Framework*-era reforms which emphasize more equitable learning opportunities for students?

# Method

We began by gathering extant assessment tasks that were said to align with the NGSS or standards based on the *Framework*, and were based around a phenomenon, sometimes called phenomenon-based item clusters. We have amassed 352 middle school tasks, originating from across 22 state-level assessment banks and university or research groups (see Table 1). To name a few, tasks that originated from state-level assessment banks include Utah (Utah State Board of Education, n.d.), Kentucky (Kentucky Department of Education, 2021), and Louisiana (Louisiana Department of Education, 2021). Some were sample tasks intended to support instructional alignment with the NGSS and some were released items from previous years' standardized tests. Examples of university- or research group-developed tasks include the

Curriculum Independent Next Generation Assessments (Lawrence Hall of Science, n.d.) and Stanford NGSS Assessment Project (SNAP; Stanford Graduate School of Education, n.d.).

Table 1. Middle School Classroom Assessment Task Sources

| **Source** | **Number of Tasks** |
|---|---|
| 3D Middle School Science (3DMSS) | 1 |
| Achieve NGSS/CCSS-M Sample Tasks | 4 |
| ASPECt-3D Project (BSCS Science Learning) | 14 |
| Curriculum Independent Next Generation Assessments (CINGA) | 125 |
| Connecticut State Department of Education | 7 |
| Delaware State Department of Education | 1 |
| Kentucky Department of Education | 18 |
| Louisiana Department of Education | 27 |
| Los Angeles County Office of Education | 14 |
| Massachusetts Innovative Science Pilot | 2 |
| Massachusetts Consortium for Innovative Education Assessment (MCIEA) Task Bank | 12 |
| Michigan State Dept of Education | 1 |
| Michigan State University Create for STEM Institute – M-PLANS project | 6 |
| New Meridian Corporation Resource Center | 3 |
| Next Generation Science Assessment (NGSA) | 49 |
| Envision Learning Partners' Performance Assessment Resource Bank | 1 |
| Washington State's SAGE Project | 1 |
| Stanford NGSS Assessment Project (SNAP) | 6 |
| Tennessee District Science Network | 5 |
| Utah State Board of Education | 52 |
| Wisconsin State Department of Ed | 3 |
| Total | 352 |

See Appendix for links to publicly available tasks.

This paper focuses on an analysis of a subset of 104 tasks. This subset was identified while working in tandem with our next phase of research, which required identifying assessment tasks targeting Performance Expectations aligned with the OpenSciEd middle school units selected for our main validity study.

## Task Screening and Characterization

For each task, we first checked and updated alignment with the NGSS Performance Expectations. While many tasks had supplementary materials that included information about alignment with one or more Performance Expectations, not all did so, and some were aligned with standards using a different classification than the NGSS. For those tasks, we crosswalked those with the appropriate NGSS Performance Expectation. Further, not all tasks were clearly stated to align with each dimension and element in the NGSS Performance Expectation (e.g. with specific elements of the SEP or CCC), so identifying and cataloging that alignment was an additional step our team took. Most tasks purported to align with one Performance Expectation, however about ten percent (n=38) assessed more than one. Though no emphasis was placed on gathering tasks with a particular domain focus, a little more than half were aligned with physical science (n=192), followed by about a third of tasks aligned to life science (n=115), about a quarter earth and space science (n=89), and two percent aligned with engineering (n=7).

Each task was evaluated on the criteria outlined in our conceptual framing; we used the Achieve (2018a) science task screener as a preexisting tool which is organized into four criteria: Tasks are: (A) driven by high-quality scenarios that focus on phenomena or problems, (B) require sense-making using the three dimensions, (C) are fair and equitable, and (D) support their intended targets and purpose. To ensure reliability, our team initially screened tasks as a group and discussed areas of disagreement, at the same time unpacking the components of each criterion to reach consensus.

## Analytic Approach

Our data sources include the middle school assessment tasks, the screener response data, and additional coding that we included to surface variation among our criteria of interest. We used the Achieve task screener as a starting point for characterizing the tasks in our pool. In engaging in an iterative process of noticing what the Achieve task screener was providing us and looking to surface further variability of characteristics, we refined our coding approach, resulting in a blend of screening items originating from the Achieve task screener and screening items that our team added. For example, Criterion A of the Achieve task screener includes prompts to determine if a phenomenon or problem is present with additional prompts that suss out how students are able to engage with the phenomenon or problem (e.g. if real-world observations are included; if they are presented as puzzling). We were interested in understanding the variability in the *kinds* of phenomena used, so our coding approach accounts for that - to characterize phenomena or problems as everyday or lab-based, for example. Table 2 presents how we organized our approach to analyzing the tasks along the three main criteria. Two of the authors independently coded and then met to adjudicate all disagreements.

Table 2. Task Characterizing Approach

| **Category** | **Code** | **Definition** | **Examples** |
|---|---|---|---|
| **Phenomenon*** | Specific and/or localized, scientific phenomena | Based on a specific and/or localized real-world phenomenon; includes images, real-world data that ground the scenario in its specific instance | Intense algae blooms on the Gulf of Mexico coastline in 2018 (Utah task) |
| | Everyday phenomenon | Phenomenon that is likely encountered by students in many places and circumstances; likely fictitious or nonspecific components to scenario | A person begins to cry while cutting onions (BEAR task) |
| | Engineering/Problem-based Phenomenon | Engineering or solving a design problem/challenge is primary context and objective of task | Designing a smartphone case to protect the phone from fall damage (CINGA task; Utah task) |
| | Generalized real-world phenomenon | Based on a real-world phenomenon but does not draw on a specific instance. Parts of task may be fictitious | A fictitious person practices throwing a ball to win a game |
| | Lab-based scenario | A scenario most likely to occur in lab- or school-based settings | A student attempts to identify unlabeled chemical substances |
| | Scenario only | Not based in observation of a phenomenon; the scenario relates to the task, but does not drive the prompts | Two characters witness a chemical reaction, then compare particle models representing the reaction |
| **Attending to Learners' Interests and Identities** | Opportunity for students to find relevance; to their own lives, their community, or to more universal connections or global issues. | Opportunity to find relevance is not dependent on whether students find the phenomenon interesting, but how it *could* relate to students' lives or the communities that they are embedded in | Own lives: A house gets cold during the winter, and its owners want to make improvements (Utah task)<br>Community: Local highway engineers investigating solutions to reduce the severity of car crashes (Kentucky task)<br>Global: Chemical engineers determining if a fuel production process releases greenhouse gasses (M-PLANS task) |
| | Opportunity for students to contribute their own ideas | Students are explicitly given the opportunity to contribute their own ideas and solutions, drawing on prior experiences, interests, and their identities | Students are asked to independently come up with a way to cool soup without diluting it (SNAP task) |

|  |  |  |  |
|---|---|---|---|
|  | Opportunity for students to make decisions about how to approach the task | Students can decide to respond to a prompt in more than one way. Limited: more than one way to approach a prompt within the task, but response space is heavily constrained<br>Open-ended: students are still under some constraints, but have more freedom to respond in different ways | Limited: students choose two climate differences (out of the three presented to them) to use in an argument<br>Open-ended: students use given data to create a display to compare the climates of two cities, and are told to do so in whatever way makes the most sense to them |
| **Sensemaking** | Opportunity to use NGSS dimensions in an integrated manner | The task includes at least one opportunity for students to use 2 or 3 dimensions together | Students analyze data for patterns to identify the cause of a fan spinning (ASPECt task) |
|  | Multidimensional sensemaking of phenomenon involving authentic uncertainty | Students use NGSS dimensions to develop a fresh design solution or explanation of a phenomenon, rather than solely providing information that they have already learned or experienced. | Students identify a way to increase crop yields through modeling the functions of plant cell organelles (Utah task) |
|  | Phenomenon is problematized | Task identifies a gap in knowledge about the phenomenon for students to resolve, and makes clear what's at stake | Deer populations have declined and scientists are unsure why (SNAP task) |
|  | Sensemaking is minimized due to overreliance on rote means | Task focuses on prompting students to show generalized content/process knowledge and memorized information | Students calculate the kinetic energy of a given object at several different speeds |
|  | Sensemaking is minimized due to scaffolding | Too much scaffolding may get in the way of students' sensemaking | Students clarify a design problem by matching pre-written answers to given steps in the engineering design process |
|  | Sensemaking is minimized due to the task "giving away the figuring out" | Task states information or ideas that students would have otherwise figured out themselves as part of the sensemaking process | Task states that a given experiment design is unfair due to failure to control variables. |

* Coded mutually exclusively. See Appendix for links to publicly available tasks.

## Analysis and Findings

We report on the findings from the subset of 104 tasks, focusing on three criteria: (1) the type of the scenario or phenomenon, (2) attention to students' interests and identities, and (3) opportunity to engage in three-dimensional sensemaking of the phenomenon.

### Phenomena

Our analyses of the responses to the prompts included in Criterion A of the screener that help determine the type and characteristics of the phenomenon or problem framing the task revealed a range of phenomena. The vast majority of tasks, about 89%, had scenarios, phenomena, or design problems that were rooted in real-life and/or in observable, scientific phenomena. We also looked to further specify the type of phenomenon or problem framing each of these tasks, which breaks down into 20% specific and/or localized, contemporary scientific phenomena, 14% everyday phenomena, 9% engineering or problem-based phenomena. 36% of tasks included phenomena that were still rooted in real-life though they were characterized as more of a general occurrence (rather than a specific one) or fictionalized phenomena (as opposed to cited, specific instances). Lastly, 11% were not based on real-world phenomena and instead were primarily decontextualized scenarios that only occur in school-based, laboratory settings, like a task that asked students to use observable characteristics to distinguish between unlabeled chemical substances.

Table 3. Distribution Of Type of Scenario, Phenomenon, or Problem Across Tasks (n=104); Mutually Exclusive Coding

| **Type of Scenario, Phenomenon, Problem** | **% of tasks** |
|---|---|
| Specific and/or localized, contemporary scientific phenomenon | 20% |
| Everyday phenomenon | 14% |
| Engineering/problem-based phenomenon | 9% |
| Generalized real-world phenomenon | 36% |
| Lab-based scenario; scenario only | 10% |
| **Total percent of tasks including any scenario, phenomenon, or design problem** | **89%** |

## Attending to Learners' Interests and Identities

Next, we focused on the parts of the screener that attended to learners' identities and interests. These prompts were primarily concentrated in Criterion C, and included assessing whether tasks provided "ways for students to make connections of meaningful local, global, or universal relevance," and how tasks cultivated "students interest in and confidence with science and engineering" (Achieve, 2018a) through occasions for reflecting students' own ideas or making decisions on how to approach completing a task.

We looked at the opportunity students had to connect with the scenario on an individual, community, and more global level and found that 57% of tasks could reasonably connect with students' own lives, 37% on more of a global scale, and 5% of tasks connect to students' communities. One task with an individual-level relevance concerned a student whose family's house got cold in the winter, and who wanted to make improvements around the house to keep it at a more consistent temperature (Utah State Board of Education, n.d.). Another task presented a case with scientists attempting to determine whether greenhouse gasses were released during the production of a certain fuel, which had global relevance since it gave students the opportunity to think about ideas related to climate change and sustainable energy (Michigan State University Create for STEM Institute, n.d.). Lastly, a task that centered around local highway engineers investigating solutions to reduce the severity of car crashes had community-level relevance, as students could think about similar measures in their own communities (Utah State Board of Education, n.d.). About 13% of tasks did have overlapping opportunities, meaning they were coded as providing more than one opportunity to find connection - the more likely combination being on the individual and global scale. One example of this combination was centered around disposable plates: in evaluating styrofoam and sugarcane plates based on characteristics like cost and ability to decompose, students could think about how they use disposable plates in their own lives as well as wider issues of sustainability (Kentucky Department of Education, 2021).

The possibility of students making connections of personal or global importance largely stems from the phenomenon or problem framing the task and in connecting with the different types of phenomena mentioned earlier. Some of the opportunities for making these connections come from phenomena that would be considered societally-relevant through attention to environmental impact or through phenomena that could have more global implications, like how fertilizer runoff can cause red algae blooms which can worsen symptoms of asthma in some populations (Utah State Board of Education, n.d.). Some of the tasks were culturally-relevant through phenomena that are based in students' interests and identities, like playing video games, dropping smartphones, playing sports, or cooking, to name a few.

Beyond the phenomenon framing the task, some tasks also fostered students' confidence in engaging with disciplinary practice. About 15% of tasks included opportunities for students to reflect their own ideas and prior experiences; for example, in a task geared towards explaining thermal energy transfer of cooling soup with ice, students were asked for their ideas on how to cool soup without the side effect of dilution (Stanford Graduate School of Education, n.d.). Further, tasks also provided limited (51%) or more open-ended (8%) opportunities for making decisions about how to approach responding to the prompts, for example, in deciding how to approach an engineering problem given a range of criteria and constraints. In one such task with an open-ended opportunity to make decisions, students were shown a variety of smartphone cases with different characteristics and asked to design a sturdy, slim smartphone case with a low cost. There were a variety of possible answers and approaches to this prompt, and any response which fully and correctly justified its choices was considered to be "correct" according to the teacher-facing materials that accompanied this task (Utah State Board of Education, n.d.).

Lastly, in considering the accessibility of the task, a preliminary takeaway we found was that roughly 36% of tasks appeared to rely heavily on reading and/or writing in standard English - an important design consideration for minimizing construct-irrelevant variance. A tension we have noticed, as designers and adaptors of assessments, is in providing sufficient context for students to engage with, to understand what's at stake, and sensemake with, without introducing a reading burden. This assessment design criterion is a challenge that needs continued attention and approaches for finding a balance between accessibility and sufficient problematization.

Table 4. Distribution of Tasks’ Attendance to Learners’ Interests and Identities (n=104)

| **Category** | **% of Tasks** |
|---|---|
| **Relevance to…** | |
| Students’ own lives | 57% |
| Students’ communities | 5% |
| Global issues | 37% |
| **Opportunities for students to contribute their own ideas** | 15% |
| **Opportunities for students to make decisions about how to approach task*** | |
| Open-ended | 8% |
| Limited | 51% |
| **Heavy reliance on reading/writing in standard English** | 36% |

* Coded mutually exclusively

**Three-dimensional sensemaking of phenomenon**

Criterion B of the task screener focuses on assessing students’ opportunity to engage in sensemaking. Though the majority of tasks, about 96%, provided some opportunity for reasoning with at least one dimension, 42% of tasks were found to engage students in multi-dimensional sensemaking of a phenomenon or problem - where authentic uncertainty of a problematized phenomenon was central and clear from students’ perspectives, as opposed to the teacher’s or developer’s perspective.

An important component of motivating and cuing sensemaking is through problematizing, as described earlier. We found that about 44% of tasks framed phenomena as problematized. One such task stated in its introduction that the deer population in a certain state had declined over time and that scientists were attempting to figure out why, thus pointing out a gap in what is known about this phenomenon (Stanford Graduate School of Education, n.d.). It then defined the stakes of the situation by stating that changes to one population affect the whole ecosystem. This problematized framing of this phenomenon motivated sensemaking in future prompts, which themselves further problematized the phenomenon by continuing to point out knowledge gaps while asking for explanations, rather than implying that responses would be restatements of information that was already known. For example, after reviewing data on various possible causes of the population decline, students are asked to "analyze patterns to help [them] decide" why the deer population changed - emphasizing that the cause of the population decline was still unknown and needed to be determined by the student.

While some tasks had promising phenomena in that they would likely be seen as compelling or relevant to students' lives, they were not all framed as problematized and might even "give away" the figuring out, resulting in a task where students demonstrated knowledge on a foregone explanation but did not sensemake. Similarly, some tasks would present the phenomenon as puzzling, but then opportunities for three-dimensional sensemaking would be minimized; we saw this happen through items that called on rote knowledge (31%), included too much scaffolding (27%), and gave away the figuring out (19%). A task that minimized sensemaking through calling on rote knowledge might present a phenomenon, then turn its focus toward prompting students to show generalized content and process knowledge, e.g. asking for statements of related science ideas and definitions or performing calculations from a memorized formula. A task with too much scaffolding would not necessarily turn its focus away from the phenomenon, but could nonetheless reduce opportunities for sensemaking through, for example, prompting students to select the correct explanation of a phenomenon from several answer choices rather than allowing them to construct their own explanations. Tasks that gave away the figuring out often stated components of the phenomenon's explanation or of the design solution, thus closing or narrowing the knowledge gap identified by the problematized phenomenon and reducing students' opportunity to sensemake around the phenomenon.

Table 5. Distribution of Tasks' Attributes Related to Sensemaking (n=104)

| Category | % of Tasks |
|---|---|
| **Problematized framing of phenomenon** | 44% |
| **Opportunity to reason using at least one NGSS dimension** | 97% |
| **3D sensemaking involving authentic uncertainty** | 42% |
| **Sensemaking minimized by…** | |
| Over reliance on rote knowledge | 31% |
| Too much scaffolding | 27% |
| Giving away the figuring out | 13% |

## Discussion

This study aims to describe the current state of science assessment through identifying characteristics of tasks that suggest strengths and ongoing challenges in assessment development in order to understand where we are as a field and contribute to an ongoing conversation about assessment design and use that supports a more equitable vision of science education. Thus far, we have found that the majority of middle school science assessments target dimensions of the NGSS and include a scenario, phenomenon, or design problem. However, there were challenges in framing phenomena that attend to learners' interests and identities and engage students in three-dimensional sensemaking.

There were some promising examples of real-world, authentic phenomena or design problems; at the same time, we saw tasks that presented phenomena that were not based in real-world observations and did not appear authentic from students' perspectives, likely making it a challenge for students to make connections of local, global, or universal relevance (Evans, 2023; Edelson et al., 2021; Furtak & Lee, 2023; Penuel et al., 2019). Further, if task scenarios are more like laboratory experiments or represent circumstances only encountered in school, there is a risk of reifying traditional approaches that we are attempting to shift away from, rather than moving towards students having opportunities to find the relevance of science beyond the classroom and connect with their lives (Lee & Grapin, 2022).

In progressing towards the development of assessments that support a sociocultural view on science learning, it is important that assessments more closely align with students' experiences in their day-to-day science learning, which are hopefully more student-centered in collaboratively figuring out phenomena. In our subset of tasks, we saw some opportunities for students to incorporate their own ideas and prior experiences, as well as have opportunities to make decisions about how to approach the task. When such classroom assessments are administered as more formal and individual, these could be seen as opportunities to collaboratively sensemake with the ideas presented through the task and students' own ideas. In figuring out how to design assessments that better support students' instructional experiences, we can continue to think about ways that perhaps disrupt the norm of seeing assessments as mainly formal and individual, and instead think creatively about how assessments could perhaps be completed collaboratively so that students may benefit from hearing and leveraging the perspectives of fellow students. For instance, on a mid-unit assessment in the OpenSciEd High

School biology unit, B.1 Ecosystem Interactions and Dynamics, students collaborate on the items that lead up to the final item, which is then completed individually (OpenSciEd, 2023).

Further, emphasizing three-dimensional sensemaking of authentic, problematized phenomena or problems, where students have observations or data to make sense with, and where they are figuring out a gap in the explanation or applying knowledge and skills to solve a design problem, will also support such instructional experiences. Though we saw that most assessments provided an opportunity to reason using at least one dimension (e.g. interpreting data to make a claim), the opportunity to sensemake around authentic uncertainty using the three dimensions seemed to be a higher bar to clear. We offer a couple suggestions moving forward, given our observations from characterizing the tasks: to problematize the phenomenon or problem and reduce the ways that sensemaking can be minimized.

Further, these three criteria - phenomena, attention to students' interests and identities, and three-dimensional sensemaking - appear to be mutually supportive. We have seen, for example, how a real-world, compelling phenomenon alone may not be enough; when phenomena are framed to be problematized and make clear what's at stake, followed with a coherent flow of prompts that support sensemaking, such tasks appear to provide an assessment experience that is more in line with students' opportunities to learn given *Framework*-aligned instruction. To attend to students' interests and identities, there could be more opportunities embedded for students to connect with the phenomena or problems presented in tasks so that students can see themselves as worthy contributors to knowledge-building and see how the science learned in school connects to their lives and interests. So, rather than presenting a phenomenon that is assumed to be relevant to students, items can explicitly solicit students' ideas and experiences, offer multiple ways for students to show what they know and can do, and ask how students see such tasks as relevant to them. We intend to systematically explore these interactions - between phenomena, attention to interest and identity, and sensemaking - in finalizing our analyses of the full set of tasks.

We want to acknowledge that assessments that exemplify these criteria well are challenging to write; we know this all too well as assessment writers and collaborators with teachers on other assessment-focused projects (Lo et al., 2022; Deverel-Rico, 2023). Our hope is that by compiling the bank of middle school assessments and highlighting examples that help us see what assessments can look like, we can work towards assessments that better reflect the vision set forth by the *Framework*.

These findings build on Wertheim and colleagues' (2016) analysis of the extant tasks available within a few years of the release of the NGSS, and provide a landscape of the currently available tasks to see how they align with the goals of the *Framework* as founded on sociocultural views of learning and equity. We acknowledge that this paper represents a subset of the pool of middle school assessment tasks that we have gathered and expect that the distribution of task characteristics will likely shift. However, we think that reporting on the subset of 104 is worthwhile to the field to have an understanding of the current state of *Framework*-aligned assessments that are widely available. We also do not have a broad sense of what kinds of assessments students are encountering in their K-12 science education nor how they are experiencing them. This would be helpful to know, as even though this paper reports on widely available tasks, we do not actually know the types of assessments that most students are encountering. Lastly, we also have not explored the accompanying materials to understand what kinds of supports teachers receive with these tasks and how the enactment guidance does or does not center student experience.

In the next phase of our research project, we look forward to further exploring the nuances in and interactions among task characteristics and finalizing the analyses of all tasks in our bank to identify assessments for use in our forthcoming pilot in middle school classrooms using *Framework*-aligned instructional materials. The validity evidence we gather about how tasks perform, combined with student experience and opportunity to learn data, can support making informed decisions to further ongoing conversations around developing assessment tasks that better cohere with *Framework*-era reforms and calls for assessment to be more equitable and meaningful for students (Penuel et al, 2019).

In closing, there is no doubt of the direction that science assessments must move towards aligning with, rather than holding back, an equitable vision of science education. Applying critical and sociocultural perspectives into the research and development of assessment is "much-needed and long overdue" (Pellegrino et al., 2023, p. 5). Such lenses help us scrutinize educational practices writ large to examine how such practices interact with the experiences of students that do not identify with the dominant culture (Ladson-Billings & Tate, 2005). Further, when we acknowledge how learning is co-constructed across many institutions (e.g. family, work, school, community) and transformed via participation in practices that themselves evolve (e.g. Nasir & Hand, 2006; Rogoff et al., 2007), we can work towards a more equitable, contemporary, and evolving definition of science learning and assessment that better reflects the pluralistic society in which we live.

## Acknowledgements

Thank you to Nathan Dadey who has painstakingly organized the majority of the middle school tasks we started with. Thank you to the university- and research-based developers who provided us with access to their assessments. Lastly, thank you to all of the dedicated people and teams developing assessments aligned to the NGSS and *Framework* - your hard work is appreciated! This material is based upon work supported by the National Science Foundation under Grant No. 2201438. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

Achieve. (2018a). Science Task Screener, Version 1.0. https://www.nextgenscience.org/sites/default/files/resource/files/Achieve%20Task%20Screener_Final_9.21.18.pdf

Achieve. (2018b). Science Task Prescreen. https://www.nextgenscience.org/sites/default/files/resource/files/Achieve%20Task%20PreScreener_Final_9.21.18.pdf

Achieve. (2019). Task Annotation Project in Science: Equity. https://www.achieve.org/publications/task-annotation-project-science-equity

Badrinarayan, A., Van Horne, K., Cooper, S. (2023). *Developing meaningful science assessments: Problematizing phenomena*. Retrieved from contextus.science on August 11, 2023.

Bang, M., Warren, B., Rosebery, A. S., & Medin, D. (2013). Desettling expectations in science education. *Human Development, 55*(5–6), 302–318.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-2.

BSCS Science Learning. (2023). *BSCS Biology: Understanding for Life*. Kendall Hunt Publishing. https://bscs.org/educator_resource/bscs-biology-understanding-for-life/

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational researcher, 34*(3), 3-14.

Deverel-Rico, C. (2023). Washington's Science Assessment Grounded in Equity (SAGE), Pilot Report. Boulder, Colorado: inquiryHub.

Edelson, D. C., Reiser, B. J., McNeill, K. L., Mohan, A., Novak, M., Mohan, L., Affolter, R., McGill, T. A. W., Buck Bracey, Z. E., Deutch Noll, J., Kowalski, S. M., Novak, D., Lo, A. S., Landel, C., Krumm, A., Penuel, W. R., Van Horne, K., González-Howard, M., & Suárez, E. (2021). Developing research-based instructional materials to support large-scale transformation of science teaching and learning: The approach of the OpenSciEd Middle School Program. *Journal of Science Teacher Education*, 32(7), 780–804.

Evans, C. M. (2023). Applying a Culturally Responsive Pedagogical Framework to Design and Evaluate Classroom Performance-Based Assessments in Hawaiʻi. *Applied Measurement in Education*, 36(3), 269–285.

Fine, C. G. M., & Furtak, E. M. (2020). A framework for science classroom assessment task design for emergent bilingual learners. *Science Education*, 104(3), 393-420.

Ford, M. J., & Forman, E. A. (2006). Redefining disciplinary learning in classroom contexts. *Review of Research in Education*, 30, 1–32.

Furtak, E., Kang, H., Pellegrino, J., Harris, C., Krajcik, J., Morrison, D., Bell, P., Lakhani, H., Suárez, E., Buell, J., Nation, J., Henson, K., Fine, C., Tschida, P., Fay, L., Biddy, Q., Penuel, W. R., & Wingert, K. (2020). Emergent Design Heuristics for Three-Dimensional Classroom Assessments that Promote Equity. In Gresalfi, M. and Horn, I. S. (Eds.), The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020, Volume 3 (pp. 1487-1494). Nashville, Tennessee: International Society of the Learning Sciences. https://repository.isls.org//handle/1/6354

Furtak, E. M. & Lee, O. (2023). Equity and Justice in Classroom Assessment of STEM Learning. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

Harris, C.J. Wiebe, E., Grover, S., & Pellegrino, J.W. (Eds.) (2023). *Classroom-Based STEM assessment: Contemporary issues and perspectives*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

Inquiry Hub & BSCS Science Learning. (n.d.) *Phase 2: Get to Know the Standards*. 5D Assessment. 5dassessment.org/tools/phase-2

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Educatio*n, 98(4), 674-704.

Kentucky Department of Education. (2021, May 17). Through Course Task Bank. https://education.ky.gov/curriculum/conpro/science/Pages/tct.aspx

Ladson-Billings, G., & Tate, W. F. I. V. (2005). Toward a critical theory of education. *Teachers College Record*, 97, 47–68.

Lawrence Hall of Science. (n.d.). *Curriculum Independent Next Generation Assessments (CINGA)*. https://lawrencehallofscience.org/educators/cinga/

Learning in Places Collaborative. (2021). Learning in Places website. learninginplaces.org

Lee, O., & Grapin, S. E. (2022). The role of phenomena and problems in science and STEM education: Traditional, contemporary, and future approaches. *Journal of Research in Science Teaching*, 59(7), 1301–1309.

Lo, A. S., Glidewell, L., O'Connor, K., Allen, A., Hermann-Abell, C. F., Penuel, W. R., Winger, K., & Lindsay, W. (2022). Promoting shifts in teachers' understanding and use of phenomena in instruction and assessment. Proceedings of the annual conference of the International Society of the Learning Sciences.

Louisiana Department of Education. (2021). *Practice tests.* Louisiana Believes. https://www.louisianabelieves.com/resources/library/practice-tests

Michigan State University Create for STEM Institute. (n.d.). *Story 6 - Making Hydrogen Gas*. Motivation - Planning Lessons to Activate eNgagement in Science.

Nasir, N. S., & Hand, V. M. (2006). Exploring Sociocultural Perspectives on Race, Culture, and Learning. *Review of Educational Research*, 76(4), 449–475.

National Academies of Sciences, Engineering, and Medicine. (2017). *Seeing Students Learn Science: Integrating Assessment and Instruction in the Classroom*. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2018). *English learners in STEM subjects: Transforming classrooms, schools, and lives*. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2019). *Science and Engineering for Grades 6–12: Investigation and Design at the Center*. 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 Academies Press.

National Research Council. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

NYU SAIL. (n.d.) *Unit 3: Why does it matter if I drink tap or bottled water?* https://www.nyusail.org/curriculum

Odden, T. O. B., & Russ, R. S. (2019). Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science Education*, 103(1), 187-205.

OpenSciEd. (n.d.) OpenSciEd Middle School Science. openscied.org/middleschool

OpenSciEd. (2022). *7.3 Metabolic Reactions: How do things inside our bodies work together to make us feel the way we do?*. https://www.openscied.org/instructional-materials/7-3-metabolic-reactions/

OpenSciEd. (2023). *B.1 Ecosystem Interactions & Dynamics.* https://www.openscied.org/instructional-materials/b-1-ecosystem-interactions-dynamics/

Pellegrino, J. W., Grover, S., Harris, C. J., & Wiebe, E. (2023). Classroom Assessment in STEM Education: An Introduction to the Report. In C. J. Harris, E. Wiebe, S. Grover, & J. W. Pellegrino (Eds.), *Classroom-Based STEM assessment: Contemporary issues and perspectives*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

Penuel, W. R., & Shepard, L. A. (2016). Social Models of Learning and Assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment* (pp. 146–173). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118956588.ch7

Penuel, W. R., Turner, M. L., Jacobs, J. K., Horne, K., & Sumner, T. (2019). Developing tasks to assess phenomenon‐based science learning: Challenges and lessons learned from building proximal transfer tasks. *Science Education*, 103(6), 1367–1395.

Philip, T. M., & Azevedo, F. S. (2017). Everyday science learning and equity: Mapping the contested terrain. *Science Education*, 101(4), 526-532.

Phillips, A. M., Watkins, J., & Hammer, D. (2017). Problematizing as a scientific endeavor. *Physical Review Physics Education Research*, 13(2), 020107.

Randall, J. (2021). “Color‐Neutral” Is Not a Thing: Redefining Construct Definition and Representation through a Justice‐Oriented Critical Antiracist Lens. *Educational Measurement: Issues and Practice*, 40(4), 82-90.

Randall, J., Poe, M., & Slomp, D. (2021). Ain’t oughta be in the dictionary: Getting to justice by dismantling anti‐black literacy assessment practices. *Journal of Adolescent & Adult Literacy*, 64(5), 594-599.

Reiser, B. J., Novak, M., McGill, T. A., & Penuel, W. R. (2021). Storyline units: An instructional model to support coherence from the students’ perspective. *Journal of Science Teacher Education*, 32(7), 805-829.

Rogoff, B. (2003). *The cultural nature of human development*. Oxford, UK: Oxford University Press.

Rogoff, B., Moore, L., Najafi, B., Dexter, A., Correa-Chavez, M., & Solis, J. (2007). Children’s development of cultural repertoires through participation in everyday routines and practices. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of socialization: Theory and research* (pp. 490–515). New York, NY: Guilford Press.

Shepard, L. A. (2021). Ambitious Teaching and Equitable Assessment: A Vision for Prioritizing Learning, Not Testing. *American Educator*, 45(3), 28.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large‐scale assessment. *Educational measurement: issues and practice*, 37(1), 21-34.

Suarez, E., & Bell, P. (2019, April). Supporting expansive science learning through different classes of phenomena. Paper presented at NARST Annual Conference, Baltimore, MD.

Stanford Graduate School of Education. (n.d.) *Short Performance Assessments*. Stanford NGSS Assessment Project. https://scienceeducation.stanford.edu/assessments/short-performance-assessments

Stanford Graduate School of Education. (n.d.) Cooling Soup Task. *Short Performance Assessments*. Stanford NGSS Assessment Project. https://scienceeducation.stanford.edu/assessments/short-performance-assessments

State Performance Assessment Learning Community. (n.d.) *Learner-Centered Standards Unpacking Template*. Contextus.science. https://contextus.science/resources/resource-title-1-zdaw5-mntr6-k6935

Tan, E., & Calabrese Barton, A. (2012). *Empowering science and mathematics education in urban schools*. University of Chicago Press.

Utah State Board of Education. (n.d.). 8th Core Guide, SEEd 8.3.3 Formative Assessment. https://schools.utah.gov/curr/science?mid=1128&tid=1

Vygotsky, L. S. (1978). Mind in society: *The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). *An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS*. Palo Alto, CA: Stanford NGSS Assessment Project Team (SNAP).

## Appendix: Publicly-Available Science Assessment Task Sources

| Source | Link |
|---|---|
| 3D Middle School Science (3DMSS) | https://3dmss.bscs.org/ |
| Achieve NGSS/CCSS-M Sample Tasks | https://www.nextgenscience.org/resources/classroom-sample-tasks |
| ASPECt-3D Project (BSCS Science Learning) | https://assessment.bscs.org/projects |
| Curriculum Independent Next Generation Assessments (CINGA) | https://lawrencehallofscience.org/educators/cinga/ |
| Connecticut State Department of Education | https://ct.portal.cambiumast.com/resources/ngss-assessment/ngss-practice-test-answer-keys |
| Delaware State Department of Education | https://www.doe.k12.de.us/Page/3783 |
| Envision Learning Partners' Performance Assessment Resource Bank | https://www.performanceassessmentresourcebank.org/ |
| Kentucky Department of Education | https://www.education.ky.gov/curriculum/conpro/science/Pages/tct.aspx |
| Louisiana Department of Education | https://www.louisianabelieves.com/resources/library/practice-tests |
| Los Angeles County Office of Education | https://lacoepd.instructure.com/courses/327 |
| Massachusetts Innovative Science Pilot | https://ma-innov-sci.mypearsonsupport.com/practice-tests/ |
| Massachusetts Consortium for Innovative Education Assessment (MCIEA) Task Bank | https://mcieaclassroom.oscarscore.com/#/public/tasks/CCE |
| Michigan State Dept of Education | https://www.michigan.gov/mde/services/student-assessment/m-step/content-specific-information/online-practice-for-m-step-ela-math-science-and-social-studies |
| New Meridian Corporation Resource Center | https://resources.newmeridiancorp.org/released-items/ |
| Next Generation Science Assessment (NGSA) | https://ngss-assessment.portal.concord.org/middle-school |
| Washington State's SAGE Project | https://contextus.science/check-and-connect-assessments/#mstasks |
| Stanford NGSS Assessment Project (SNAP) | https://scienceeducation.stanford.edu/assessments/short-performance-assessments |
| Tennessee District Science Network | https://ngs.wested.org/tennessee-district-science-network |
| Utah State Board of Education | https://schools.utah.gov/curr/science?mid=1128&tid=1<br>*Assessments embedded within each Core Guide* |
| Wisconsin State Department of Ed | https://dpi.wi.gov/science/assessment/examples |