New Meridian

WHITE PAPER

The New Meridian Framework for Quality Review of NGSS Science Assessment Items

By New Meridian

December 15, 2020

New Meridian

Introduction and Purpose

Developing high quality science assessments based the science standards has presented significant challenges to educators and test developers. Processes have not evolved quickly enough to meet the challenges of modern science assessment development and the field has not formed a consensus view of best practice. Some of the major challenges include how to approach a design that models the richness of the standards, how to design equitable and fair science assessment, and how to connect assessment claims to the major features of a quality science assessment item or task.

While there has been extensive prior work to support the development of science curriculum, instruction, and classroom assessment based on new science standards, there has yet to be a framework specifically geared toward the needs of developers of large-scale science assessment. New Meridian has sponsored the development of this framework to address those needs and further advance the field of science assessment. The authors have synthesized an approach for thinking about, analyzing, and evaluating item quality. This document lays out the critical elements of a quality science assessment item or task. These are structured into a process that can be used to evaluate and ensure that science items and tasks exhibit those critical qualities.

The authors hope that this work can be used broadly by states as they develop new science assessments to reflect Next Generation Science Standards (NGSS) and similar standards based on *A Framework for K-12 Science Standards*[1] and will support their pursuit of developing high-quality items and tasks designed specifically for large-scale assessments. These assessments will need to measure three-dimensional (3D) expectations—those that integrate Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs)—in equitable and fair ways.

The criteria outlined in this framework identify the features of high-quality, three-dimensional science assessment design. States can apply these criteria to develop or review their large-scale assessments. The criteria apply to all assessments designed for multi-dimensional standards based on *A Framework for K-12 Science Education*.

[1] National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

This *Framework for Reviewing Three-Dimensional Science Assessment Items* consists of three parts:

Part 1: Critical foundations for developing high-quality Items and Tasks. This section identifies the metadata and features of task design and implementation that establish the necessary foundation for high-quality, multidimensional items and tasks prior to any content or quality review.

Part 2: Indicators of quality science assessment tasks: item- and task-level analysis. This section focuses on the indicators and processes for reviewing items and tasks and outlines a two-part process: a task-level prescreen and a deeper descriptive task and item review.

Part 3: Guidance for implementing reviews. This section describes the procedures and processes New Meridian recommends to implement reviews according to the indicators identified in the framework.

Terms and Definitions

This framework uses the following definitions for tasks, items, and scenarios:

Tasks refer to all scenario/stimuli and prompts/questions associated with a single coherent activity that is designed to monitor progress toward a specific target (e.g., performance expectation or bundle of performance expectations). Tasks can include single or multiple items/prompts, multiple parts or sections, and multiple formats.

Items refer to specific prompts or questions associated with a task—generally, the smallest unit that would be used to derive score points. One or more items usually combine with a scenario to form a task.

Scenarios refer to the phenomenon- or problem-based contexts used to engage students in the scientific thinking required by the task. This includes all stimuli, including text descriptions, data, models, arguments, etc. This contextual information may be presented at the beginning of a task as well as introduced at multiple times throughout the task.

Designing for equity and inclusion.

This framework reflects the commitment to equitable science education for all learners that is central to *A Framework for K-12 Science Education*, as well as NGSS and similar standards. While the focus of this framework is on content quality and alignment to multi-dimensional standards, features of equitable assessments cannot be disentangled from quality measures. High-quality science assessments are intentionally designed to support diverse learners in demonstrating their proficiency.

This framework guides authors and reviewers by outlining items and task features in each section that support equity and access. Content development and review processes should also include a diverse representation of stakeholders who review disaggregated student data for:

- Relevance
- Comprehensibility
- Coherence through the student lens
- Appropriate and supportive language

Emphasis should be placed on "sense-making" using the multiple dimensions, rather than assessing vocabulary, rote knowledge, and other isolated features exhibited in traditional science assessments, which have disadvantaged students in the past.

The conversation among educators over how to ensure equitable science assessments, particularly those designed for large-scale use, evolves every day. We expect the features described here to establish a minimum threshold: the floor, not the ceiling. We look forward to updating and enhancing criteria for equitable assessments as design processes and expectations progress in the field.

New Meridian

Part 1: Critical Foundations for Developing High-Quality Items and Tasks

Development of high-quality items and tasks builds upon the foundations of strong test design:

1. **Purposeful design**. Test developers must clearly articulate how an assessment supports claims about student mastery of the domain based on evidence of mastery generated through student engagement with the tasks. Blueprints should thoughtfully outline how the assessment samples the domain and multiple dimensions. Reporting categories should reflect how the domain is organized and coherently organize the claims to support interpretation. At a minimum, this includes the following:
    a. **Domain:** An overview of the standards, elements, competencies, knowledge, and/or skills being assessed, defined specifically enough to 1) allow differentiation from other likely interpretations by intended users, and 2) guide test development. While the exact documentation will vary from state to state, this might include contextualized item specifications, state-created development frameworks, and blueprints, as well as other documentation.
    b. **Task-level claims**, including:
        i. The specific knowledge and practice targeted by each task (i.e., core components or substantial parts of the *Framework* SEP, CCC, DCI elements included in the grade band that are intended to be assessed by each prompt within tasks, and the tasks as a whole)
        ii. Documentation that shows how the knowledge and practice targeted by each task connects to a substantial part of a standard/performance expectation at grade-level, and what evidence of proficiency looks like
    c. **Attention to multiple dimensions of equity and diversity**: Test developers should consider dimensions of equity and diversity throughout the test development process, including diverse representations of culture, language, ethnicity, gender, and disability. Test developers should attend to these dimensions throughout the test development process, including (a) the blueprint development process; (b) the task development and evaluation processes, including the development of task templates and evaluation rubrics; and (c) the content and format of contexts, phenomena, and problems used on assessments. Test developers should consider empirical evidence related to bias and sensitivity as they become available through field testing.

d. **Stakeholder involvement and engagement:** Test developers should engage diverse stakeholders throughout the development process, including recruiting teacher involvement and diverse representation within the item writing and review processes.

e. **Technology specifications:** Test developers should consider and document all technology required to use the items/tasks (e.g., Technology-Enhanced Item types; QTI format; use of simulations, videos, and photographic images; and technology needed for intended accessibility supports).

f. **Pretesting**. Pretesting items and tasks with students generates critical data to support evaluation of quality, difficulty, accessibility, and fairness. States should collect and review pilot, field-test, and operational data on how items and tasks perform. This may include descriptive data from cognitive labs capturing students' reflections on what the item and task is measuring, and/or quantitative item statistics disaggregated by demographic categories to evaluate item and task performance.

Part 2: Indicators of Quality Science Assessment Tasks: Item- and Task-level Analysis

At the heart of high-quality assessments are the items and tasks that comprise those assessments. The indicators described here were developed based on expert understanding of how to design assessments for the NGSS, a review of state summative assessment items, and previously developed and widely used documents intended to support the design and vetting of high-quality NGSS tasks[1].

[1] Foundational documents that provided a basis for the indicators here included Science Alignment Criteria, the Science Task Prescreen, and the Science Task Screener indicators and processes, developed by Achieve in collaboration with states and experts to exemplify the cor4 features of NGSS assessments from large-scale models to instructionally relevant tasks. While the criteria and guidance from these documents provides a basis for the criteria described in this current framework, all indicators included here were tested and modified for current large-scale assessment item and task review as appropriate.

Table 1 outlines the core features of high-quality items and tasks aligned to the *Framework*.

Table 1: Core features of item and task review.

Feature	Rationale for inclusion based on the NGSS and similar standards based on *A Framework for K-12 Science Education.*
The quality of the phenomenon- or problem-based scenarios.	Meeting the expectations of the NGSS and similar multidimensional standards based on the *Framework* requires that students demonstrate the degree to which they can use the three dimensions to make sense of phenomena and problems. In assessment, scenarios grounded in specific phenomena and problems provide the structure for students to make their facility with three-dimensional targets visible. The quality of the scenario plays a large role in determining how well tasks and items can elicit meaningful multi-dimensional thinking; as a result, this framework outlines a review process that attends substantially to the quality of scenarios grounded in phenomena and problems.
The degree to which multi-dimensional targets are assessed.	The NGSS and similar multi-dimensional standards require students to demonstrate the degree to which they understand and can use the three dimensions of science education—Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs)—together to make sense of phenomena and problems. In assessment, meeting these standards requires that items and tasks elicit student understanding and performance relative to specific dimensions as well as their integrated use. This framework outlines a review process that attends to the three dimensions, separately and together.
The degree to which sensemaking is required to respond to the task.	In the NGSS and similar standards, sense-making[2] distinguishes meaningful, multi-dimensional performances from more isolated and superficial demonstrations of the three dimensions. Demonstrating the three dimensions as expected by the standards requires that they be used in service of sense-making; in other words, it is not sufficient to define scientific words or skills. Rather, science ideas and practices must be demonstrated as students are applying them to "figure out" aspects of phenomena and problems. This framework addresses sense-making, both in terms of how scenarios are set up to enable and require it, as well as whether the dimensions are engaged in service of it.

This framework describes a two-part process to conduct an efficient and comprehensive item and task review:

- **Part 1: Prescreen.** Conduct an initial **task-level prescreen** to evaluate for a minimum quality threshold.
- **Part 2: Descriptive Review.** For tasks that satisfy prescreen requirements, conduct an in-depth **item- and task-level descriptive review**.

[2] For a practical guide to sense-making in assessment tasks, please see this resource, developed by Achieve as part of a collaborative project to support understanding high-quality science assessments

Task-level Prescreen

Conduct an initial prescreen for basic criteria that indicate high-quality science tasks designed for multi-dimensional standards.

Table 2 presents the quality measures and specific indicators for the task-level prescreen process.

Table 2: Task-level Prescreen Quality Measures and Indicators

Quality Measure	Specific Indicators
A phenomenon or problem drives the task.	a. A phenomenon or problem is present. b. The scenario, grounded in the phenomenon or problem, establishes a meaningful context for successfully responding to all items in the task. c. The scenario, grounded in the phenomenon or problem, is necessary to respond to the majority of items posed in the task successfully.
As a whole, the task requires sense-making.	a. Rote knowledge cannot be used to successfully respond to most of the questions in the task. b. The majority of the questions require some kind of reasoning to respond successfully.
Appropriate disciplinary core ideas (DCIs) are required to respond successfully to the task.	a. DCIs required are grade appropriate. b. The targeted DCIs are required (i.e., what is claimed is what is assessed).
Appropriate science and engineering practices (SEPs) are required to successfully respond to the task.	a. SEPs required are grade appropriate. b. The targeted SEPs are required (i.e., what is claimed is what is assessed).
Multiple dimensions must be used together to successfully respond to the task.	a. Dimensions are not assessed in isolation within individual items or tasks b. Over the course of the task, multiple dimensions are used together.
The task is comprehensible and coherent.	a. The task is clear and makes sense to the students intended to respond to the task.

Item- and Task-level Descriptive Review

Tasks that meet the requirements of the prescreen should be analyzed more deeply using a descriptive review at the item and task level. Quality and alignment indicators for the descriptive review fall into four categories:

1. **Scenario quality**
2. **Three-dimensional performance**
3. **Technical quality**
4. **Cognitive complexity**

Table 3 presents the quality measures and specific indicators used in the descriptive item- and task-level review.

Table 3: Item- and Task-level Descriptive Review Quality Measures and Indicators

Quality Measure	Specific Indicators
1. Scenario quality. Indicators in this category describe the features of the scenario provided to students.	**1. Task scenario is sufficient, engaging, relevant, and accessible to a wide range of students. The scenario must:** a. be observable and accessible to a wide range of students: i. Uses real-world observations. ii. Uses at least two modalities (e.g., text, images, video, data tables). iii. Employs real or well-crafted data. b. present a puzzling/intriguing problem. c. use grade-appropriate SEPs, CCCs, DCIs. d. use grade-appropriate data. e. present a local, global, or universal context that is relevant and clear to students. f. be comprehensible to a wide range of students at grade-level. g. use as many words as needed, no more. h. include sufficiently rich content to drive and sustain performance through the task. i. use diverse representations of scientists and engineers, as appropriate. j. be built logically and coherently (when multiple components of a scenario are introduced throughout a task). **2. Task scenarios must be grade-appropriate and:** a. require grade-appropriate SEPs, DCIs, and CCCs to respond. b. not require information that is outside the bounds of the targeted dimensions outlined in the standards. c. use grade-appropriate vocabulary and syntax, based on accepted standards in science and English language arts.

2. Multi-dimensional performance. Indicators in this category determine the degree to which tasks and items require students to use the Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs) and Crosscutting Concepts (CCCs) in service of sense-making.	**1. Reasoning with evidence, models, and scientific principles (i.e., sense-making). A fundamental difference between multi-dimensional items and tasks and more traditional science assessments is that these new tasks and items require sense-making from the student to answer the questions being asked.** a. Item level: individual items require students to engage in generating evidence, reason with evidence, or reason about the validity of claims related to a phenomenon or problem. b. Task-level: Assessment tasks require students to connect evidence (provided or student generated) to claims, ideas, or problems (e.g., explanations, models, arguments, scientific questions, definition of/solution to a problem) by using the SEPs, CCCs, and DCIs as a fundamental component of their reasoning. **2. Assessing each dimension, and multiple dimensions together. For each dimension (DCIs, SEPs, CCCs), alignment indicators include the following:** a. Which element of the dimension is required to respond to the item/task b. The grade-band at which the dimension is engaged c. Whether the dimension is engaged in service of sense making (in contrast to rote information) *It should be noted that more weight/emphasis should be placed on DCIs and SEPs, as CCCs prove challenging to assess in most large-scale contexts.* i. **Item level:** Individual items require students to use each dimension at grade level in service of sense making; this can be evaluated for each dimension across the indicators described above. ii. **Task level:** Across a task, students are required to use at least two dimensions together to make sense of phenomena and/or problems.
3. Technical quality. These indicators describe the technical quality of items. These indicators should <u>all be met for all items and tasks.</u>	**1. Accuracy** a. Scientific accuracy b. Free from technical errors **2. Clarity:** Items and tasks are written and illustrated clearly so that they are easily understood by students. **3. Equitable and free from bias and sensitivity concerns:** Items and tasks are accessible to all student groups, including economically disadvantaged students, students with limited English language proficiency, students with disabilities, students from all major racial and ethnic groups, female students, students in alternative education programs, and gifted/talented students. 4. **Appropriate level of mathematics and ELA/literacy:** Items and tasks do not require reading or mathematics beyond what is required by the SEP, CCC, and DCI as specified by the targeted elements, by the assessment boundaries described in the standards, or by a state's grade-level mathematics and ELA standards.

4. Cognitive complexity. These indicators describe the level of sensemaking required to respond to tasks.	Tasks are evaluated according to a framework designed specifically for NGSS assessments[3], which focuses on determining the level of thinking required by large-scale assessments and builds on the multi-dimensional and progressive nature of NGSS tasks. This work, developed by Achieve Inc., is based on the *Task Analysis Guide in Science*[3].

[3] Achieve developed *A Framework to Evaluate Cognitive Complexity in Science Assessments* to support monitoring cognitive complexity measures in three-dimensional assessments. The Achieve framework draws from research on cognitive complexity and examples of student performance; task complexity in classroom assessment tasks; and the specific design and approach of large-scale assessments designed for NGSS and similar standards. The Achieve framework uses that measure, rather than other complexity frameworks, because it is designed to reflect the nuances and distinguishing features of 3D assessments.

[3] Tekkumru-Kisa, Miray & Stein, Mary & Schunn, Christian. (2015). A framework for analyzing cognitive demand and content-practices integration: Task analysis guide in science: TASK ANALYSIS GUIDE IN SCIENCE. Journal of Research in Science Teaching. 52. 10.1002/tea.21208.

Part 3: Implementing Reviews – How New Meridian Operationalizes this Framework

Reviewers: Recruitment and Panel Composition

States should conduct these reviews with a small panel of expert reviewers who are knowledgeable in how to apply the indicators described in Part 2 (Indicators of Quality Science Assessment Tasks: Item- and Task-level Analysis) to large-scale assessments. Reviewers should have grade-band specific domain expertise, deep familiarity with the NGSS and similar standards, familiarity with classroom implementation of the NGSS, and familiarity with large-scale summative assessment. The review panel should reflect appropriate diversity, including at a minimum racial, ethnic, gender, and geographical diversity. We recommend panels large enough to allow for three reviewers per item review block, ensuring that all items include individual and expert consensus review. The exact size of the review panel will depend on the number of states and tasks to be reviewed.

Reviewers: Training and Calibration

Prior to engaging in any review processes, reviewers should undergo an intensive training and calibration process, spanning many different task development approaches. Reviewer training should include understanding the features of high-quality scenarios; the strategies for assessing each dimension (and the dimensions together) in service of sense making;, and how to review scenarios and tasks for equity and fairness. Training should also include how to review the indicators described in this framework. Following the training, reviewers should review sets of diverse items for calibration purposes, particularly those with design features similar to items they may be reviewing in the upcoming cycles. Reviewers should meet at least twice a year to re-calibrate and extend their understanding of item development and implementation, as these processes are expected to evolve.

Review Process

Once reviewers are recruited, trained, and calibrated, New Meridian recommends the following review process:

1. **Internal Screen.** Prior to content review, New Meridian staff screen the submitted information for the indicators described in Part 1 (Must-Have Features for Item and Task Submissions) and organize the information within a system to enable efficient review.
2. **Task Assignment.** Tasks are then assigned to a panel of at least three reviewers for individual and consensus review. Tasks should be assigned based on the expertise and diversity features noted above. New Meridian will assign a lead reviewer and/or separate facilitator who collates reviews and leads the writing process for the final report as needed.

3. **Individual to Consensus Reviews.** For both the prescreen and descriptive reviews, reviewers should follow an individual-to-collective review process: each reviewer should review the tasks independently and record their evidence, reasoning, and final judgements prior to any group discussions. During the group discussion, the facilitator/lead reviewer should conduct a discussion to ensure consensus on each indicator for each item or task.
    a. **Prescreen.** Reviewers should first prescreen all assigned tasks to determine which will undergo the more in-depth descriptive review. This might involve New Meridian staff or the assigned review panel; best practices would suggest at least two reviewers connect on the prescreen and make decisions about tasks moving forward in the review process. Prescreen review should include documentation of how each task performed relative to each indicator, the evidence and reasoning used to make the judgement, and any overall holistic comments (particularly for tasks that are NOT moving onto the full review).
    b. **Full descriptive review.** For tasks that move on to the in-depth review, reviewers should once again individually review each scenario, item, and task relative to the appropriate indicators prior to consensus conversations.
4. **Final report**. New Meridian will share the results of each descriptive task review to contributing states. New Meridian believes these reviews will inform and guide states in their future science item development and thus help elevate overall quality of largescale science assessments nationally. Science assessment content in the New Meridian Science Exchange Item Bank will be tagged with the item- and scenario-level review data to support subscribing states in their selection of tasks to meet their assessment needs.

Additional Notes

Some states may require additional reviews prior to including items on their assessment—internal reviews, reviews by state teacher panels, etc. While these reviews are the state's responsibility, New Meridian will make all review and training materials publicly available and will support states in training reviewers if states wish to use this process.