Rhode Island and Vermont Multi-State Science Assessment

2018–2019

Volume 2: Test Development

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EXHIBITS

LIST OF APPENDICES

1. INTRODUCTION

In 2013, the Rhode Island Department of Education (RIDE) and Vermont Agency of Education (VT AOE) adopted the Next Generation Science Standards (NGSS). The RIDE and the VT AOE and their assessment vendor, the American Institutes for Research (AIR), developed and administered a new online assessment to measure the new standards. In 2017–2018, the Rhode Island Next Generation Science Assessment (RI NGSA) was administered as an independent field test in Rhode Island, and the Vermont Science Assessment (VTSA) was administered as an operational field test in Vermont. The RI NGSA and VTSA were administered operationally for the first time in 2018–2019. The RI NGSA and the VTSA measure the science knowledge and skills of Rhode Island and Vermont students in grades 5, 8, and 11 as an online assessment, constructed linearly on the fly, making use of several technology-enhanced item types.

In the remainder of this volume, the term *Multi-State Science Assessment* (MSSA) will refer to the RI NGSA and VTSA.

Additional details on the implementation of the assessments can be found in Volume 1 of this technical report.

The interpretation, usage, and validity of test scores rely heavily upon the process of developing the test itself. This volume provides details on the test development process of the MSSA, which contributes to the validity of the test scores. Specifically, this volume provides evidence to support the following:

- The test item specifications provided detailed guidance for item writers and reviewers to ensure that science items were aligned to the performance expectations they were intended to measure.
- The item development procedures employed for MSSA tests were consistent with industry standards.
- The development and maintenance of the item pool plan established an item bank, in which test items cover the range of measured performance expectations, grade-level difficulties and levels of cognitive engagement using both item clusters and stand-alone items.
- The test design summary/blueprint stipulated the range of operational items from each item type and content category that were required on each test administration. This document was implemented in the item selection algorithm for science.

Note that for the science assessments, as outlined in Volume 1, AIR works with a group of states that share common item development processes. In addition to developing items for each of those states, AIR develops and maintains the AIRCore item bank, which consists of items that are developed according to the same principles that are followed for the items owned by each of the states. Therefore, this volume focuses on the general test development activities.

For the MSSA test, items are drawn from an item bank that consists of AIRCore items, items owned by Rhode Island and Vermont, and items owned by several other states that share a Memorandum of Understanding (MOU) to share content, leadership, and new ideas and methods.

Specifically, all items developed under the MOU went through the same development process. In the remainder of this volume, the term *item bank* will refer to all items developed under the MOU unless stated explicitly otherwise.

1.1 CLAIM STRUCTURE

The goals, uses, and claims that the science item bank and subsequent tests would be designed to support were identified in a series of collaborative meetings held August 22–23, 2016, as an attempt to facilitate the transition from the NGSS to statewide summative assessments for science. AIR invited content and assessment leaders from ten states as well as four nationally recognized experts that helped author the NGSS. Two nationally recognized psychometricians also participated.

AIR staff and participating states collaborated to develop items and test specifications to measure the NGSS. The item specifications were generally accompanied by sample item clusters meeting those specifications. All specifications and sample item clusters were reviewed by state content experts and committees of educators in at least one of the states.

1.2 UNDERLYING PRINCIPLES GUIDING DEVELOPMENT

The item bank for science was established using a highly structured, evidence-centered design. The process began with detailed item specifications. The specifications, discussed in Section 2.2, described the interaction types that can be used, gave guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and provided sample items.

Items were written with the goal that virtually every item would be accessible to all students, either by itself or in conjunction with accessibility tools, such as text-to-speech, translations, or assistive technologies. This goal is supported by the delivery of the items on AIR's test delivery platform, which has received Web Content Accessibility Guidelines (WCAG) 2.0 AA certification. This platform offers a wide array of accessibility tools and is compatible with most assistive technologies.

Item development supported the goal of high-quality item clusters and stand-alone items through rigorous development processes managed and tracked by a content development platform. This system ensures that every item flows through the correct sequence of reviews and captures every comment and change to the item.

AIR sought to ensure that the items were measuring the performance expectations in a fair and meaningful way by engaging educators and other stakeholders at each step of the process. Educators evaluated the alignment of items to the performance expectations and offered guidance and suggestions for improvement. They participated in the review of items for fairness and sensitivity. Following item field testing, educators engaged in rubric validation, a process that refines rule-based rubrics upon review of student responses.

Combined, these principles and the processes that support them have been incorporated into an item bank that measures the performance expectations with fidelity and does so in a way that minimizes construct-irrelevant variance and barriers to access. The details of these processes follow.

1.3 ORGANIZATION OF THIS VOLUME

This volume is organized in three subsequent sections:

1. An overview of the science item development process that supports the validity of the claims that science tests are designed to support
2. An overview of the science item bank, the types of assessments the bank is designed to support, and methods for refreshing the bank
3. A description of test construction for the MSSA, including the blueprint, the test design, an evaluation of simulated test sessions, the operational blueprint match results, and the item exposure rates

2. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

2.1 OVERVIEW

AIR developed the science item bank in collaboration with the states that were part of the Memorandum of Understanding (MOU) using a rigorous, structured process that engaged stakeholders at critical junctures. This process was managed by AIR's Item Tracking System (ITS), which is an auditable content-development tool that enforces rigorous workflow and captures each item change and comment. Reviewers, including internal AIR reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, with all accessibility features and tools.

The process begins with the definition of item specifications, and continues with

- selection and training of item writers;
- writing and internal review of items;
- review by state personnel and stakeholder committees;
- markup for translation and accessibility features;
- field testing; and
- post-field-test reviews.

Each of these steps has a role in ensuring that the items can support the claims on which they will be based. Exhibit A on the following page describes how each step contributes to these goals. Each step in the process is discussed in more detail.

Exhibit A. Summary of How Each Step of Development Supports the Validity of Claims

	Supports alignment to the performance expectations	**Reduces construct-irrelevant variance through universal design**	**Expands access through linguistic and other supports**
Item specifications	Specifies item interactions, content limits, and guidelines for meeting task demands and levels of cognitive engagement requirements and adjusting difficulty.	Avoids the use of any item interactions with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selection and training of item writers	Ensures that item writers have the background to understand the performance expectations and specifications. Teaches item writers about selection of item interactions for measurement and accessibility.	Training in language accessibility, bias, and sensitivity helps item writers avoid unnecessary barriers.	
Writing and internal review of items	Checks content alignment and evaluates and improves overall quality.	Eliminates editorial issues, and flags and removes bias and accessibility issues.	
Markup for translation and accessibility features		Adds universal features, such as text-to-speech for science, that reduce barriers.	Adds text-to-speech, braille, American Sign Language (ASL), translations, and glossaries.
Review by state personnel and stakeholder committees	Checks content and cognitive complexity alignment; evaluates and improves overall quality.	Flags sensitivity issues.	
Field testing	Provides statistical check on quality and flags issues.	Flags items that appear to function differently for subsequent review for issues.	May reveal usability or implementation issues with markup.
Post-field-test reviews	Final, more focused check on flagged items. Rubric validation ensures that scoring reflects performance expectations.	Final, focused review on items flagged for differential item functioning.	

2.2 ITEM SPECIFICATIONS

AIR is working with a group of states, psychometricians, and science experts including the authors of the Next Generation Science Standards (NGSS) to develop powerful innovative solutions to the challenges of measuring the NGSS. Participating states include Connecticut, Hawaii, Idaho, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. New Hampshire participates in some activities. This collaboration has yielded item specifications for NGSS performance expectations, sample item clusters for each specification, and hundreds of NGSS item clusters and stand-alone items in various stages of development. Under this collaboration, states have developed item specifications jointly.

Test item specifications are documents that are designed to guide the work of item writers as they craft test questions and the reviews of those items by stakeholders. These specifications are intended to serve as a roadmap for writers to facilitate the creation of items that are properly aligned to the three-dimensions that comprise each Next Generation Science Standard and that together properly structure into coherent item clusters and stand-alone items. Exhibit B on the following page provides a sample of the item specifications developed by content experts for a middle school life sciences performance expectation (PE). Item specifications in science include the following:

- **Performance Expectation.** This identifies the PE being assessed.
- **Dimensions.** This identifies the Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs) that the PE assesses.
- **Clarifications and Content Limits.** This delineates the specific content that the PE measures and the parameters in which items must be developed to assess the PE accurately, including the lower and upper complexity limits of items. Specifically, content limits refine the intent of the PE and provide limits of what may be asked of test takers. For example, content limits may identify the specific formulae that students are expected to know or not know.
- **Science Vocabulary.** This section identifies the relevant technical words that students are expected to know, and related words that they are explicitly not expected to know. These categories should not be considered exhaustive, as the boundaries of relevance are ambiguous, and the list is limited by the imagination of the writers.
- **Content/Phenomena.** This section provides examples of the types of phenomena that would support the effective items related to the PE in question. In general, these are guideposts, and item writers seek comparable phenomena, rather than drawing on those within the documents.
- **Task Demands.** In this section, the PEs and associated evidence statements are broken down into specific task demands aligned to each PE. Task demands denote the specific ways in which students will provide evidence of their understanding of the concept or skill. Specifically, the task demands identify the types of interactions and activities that item writers should employ. Each item should be clearly linked to one or more of the

task demands, and the verbs guide the types of interactions writers might employ to elicit the student response.

Exhibit B. Sample Science Item Cluster Specifications for Middle School

<table>
<tr><td>Performance Expectation</td><td colspan="3">MS-LS1-1
Conduct an investigation to provide evidence that living things are made of cells; either one cell or many different numbers and types of cells.</td></tr>
<tr><td>Dimensions</td><td>Planning and Carrying Out Investigations
Conduct an investigation to produce data to serve as the basis for evidence that meets the goals of an investigation.</td><td>LS1.A: Structure and Function
All living things are made up of cells, which is the smallest unit that can be said to be alive. An organism may consist of one single cell (unicellular) or many different numbers and types of cells (multicellular).</td><td>Scale, Proportion, and Quantity
Phenomena that can be observed at one scale may not be observable at another scale.</td></tr>
<tr><td>Clarifications and Content Limits</td><td colspan="3">Clarification Statements
Emphasis is on developing evidence that living things are made of cells, distinguishing between living and non-living things, and understanding that living things may be made of one cell or many varying cells.
Content Limits
<u>Students do not need to know the following</u>:
The structures or functions of specific organelles or different proteins
Systems of specialized cells
The mechanisms by which cells are alive
Specifics of DNA and proteins or of cell growth and division
Endosymbiotic theory
Histological procedures</td></tr>
<tr><td>Science Vocabulary Students are Expected to Know</td><td colspan="3">Multicellular, unicellular, cell, tissue, organ, system, organism hierarchy, bacteria, colony, yeast, prokaryote, eukaryote, magnify, microscope, DNA, nucleus, cell wall, cell membrane, algae, chloroplast(s), chromosome, cork</td></tr>
<tr><td>Science Vocabulary Students are Not Expected to Know</td><td colspan="3">Differentiation, mitosis, meiosis, genetics, cellular respiration, energy transfer, RNA, protozoa, amoeba, histology, protista, archaea, nucleoid, plasmid, diatoms, cyanobacteria</td></tr>
<tr><td colspan="4">Phenomena</td></tr>
<tr><td>Context/ Phenomena</td><td colspan="3">Some example phenomena for MS-LS1-1 include:
Plant leaves and roots have tiny box-like structures that can be seen under a microscope.
Small creatures can be seen swimming in samples of pond water viewed through a microscope.
Different parts of a frog’s body (e.g., muscles, skin, tongue) are observed under a microscope, and are seen to be composed of cells.
One-celled organisms (e.g., bacteria, protists) perform the eight necessary functions of life, but nothing smaller has been seen to do this.</td></tr>
</table>

	• Swabs from the human cheek are observed under a microscope. Small cells can be seen.

This Performance Expectation and associated Evidence Statements support the following Task Demands.
Task Demands
1. Identify from a list, including distractors, the materials/tools needed for an investigation to find the smallest unit of life (cell).
2. Identify the outcome data that should be collected in an investigation of the smallest unit of living things.
3. Evaluate the sufficiency and limitations of data collected to explain that the smallest unit of living things is the cell.
4. Make and/or record observations about whether the sample contains cells.[a]
5. Interpret and/or communicate data from the investigation to determine if a specimen is alive.
6. Construct a statement to describe the overall trend suggested by the observed data.

Note. [a]Denotes those task demands that are deemed appropriate for use in stand-alone item development.

The specifications help test developers create item clusters and stand-alone items that will support a range of difficulty, furthering the goal of measuring the full range of performance found in the population, but remaining on grade-level.

2.3 SELECTION AND TRAINING OF ITEM WRITERS

All item writers developing science items at AIR have at least a bachelor's degree, and many bring teaching experience. All item writers are trained in

- the principles of universal design;
- the appropriate use of item interactions; and
- the NGSS specifications.

Key materials are shown in Appendix A. These include

- AIR's Language Accessibility, Bias, and Sensitivity Guidelines; and
- a training (presented using Microsoft PowerPoint) for the appropriate use of item interactions.

2.4 INTERNAL REVIEW

AIR's test development structure utilizes highly effective units organized around each content area. Unit directors oversee team leaders who work with team members to ensure item quality and adherence to best practices. All team members, including item writers, are content-area experts. Teams include senior content specialists who review items prior to client review and provide training and feedback for all content-area team members.

AIRCore and MOU science items go through a rigorous, multiple-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the entire process. A sample item review checklist that our test developers use is included in Appendix B. The AIRCore and MOU science internal review cycle includes the following phases:

- Preliminary Review
- Scoring Entry and Review
- Content Review One
- Edit Review
- Senior Review

2.4.1 Preliminary Review

Preliminary Review is conducted by team leads or senior content staff. Sometimes Preliminary Review is conducted in a group setting, led by a senior test developer. During the process, team leads or senior content staff analyze items to ensure the following:

- The item aligns with the performance expectation.
- The item matches the item specification for the skills being assessed.
- The item is based on a quality scientific phenomenon (i.e., it assesses something worthwhile in a reasonable way/it is a discrete observation that grounds a scenario that allows for the assessment of something worthwhile in a meaningful way).
- The item is properly aligned to the task demands.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question.
- The stimulus is clear, concise, and succinct (i.e., it contains enough information to know what is being asked, it is stated positively, and it does not rely on negatives—such as *no, not, none, never*—unless absolutely necessary).

For selected-response item interactions, test developers also check to ensure that the set of response options are

- as succinct and short as possible (without repeating text);
- parallel in structure, grammar, length, and content;
- sufficiently distinct from one another;

- all plausible (but with only correct option); and
- free of obvious or subtle cuing.

2.4.2 Scoring Entry and Review

At Scoring Entry level, the item writer inputs the machine scoring so that it can be reviewed by the team lead or senior staff that is reviewing the item prior to Content Review One. This step is kept separate from Preliminary Review so that the senior staff can suggest changes to the interaction at Preliminary Review without requiring the writer to overhaul scoring that they have already created. It also allows the senior staff to ensure that the scoring suggested by the writer at Preliminary Review is appropriate. This ensures the scoring is entered once, streamlining the process. At this level, the scoring is analyzed to ensure the following:

- The scoring works as it is intended (i.e., the student gets a point for ALL correct responses and no points for ALL incorrect responses).
- The student receives a point for every unique piece of information they reveal about their understanding through their responses.
- Dependent scoring between and within interactions is captured.
- The way in which the scoring is set up is unambiguous and matches the questions asked (i.e., if we tell them they must round to a certain decimal place, we score them as such).

The senior staff approves the intent of the scoring at Preliminary Review. At scoring entry, the writer inputs this approved scoring, after which the senior staff checks the functionality of the scoring. Once the scoring is determined to be working correctly, the senior staff signs off on it and moves it to Content Review One.

2.4.3 Content Review One

Content Review One is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on all the criteria identified for Preliminary Review. He or she also ensures that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item both from the perspective of potential clients as well as his or her own experience in test development.

2.4.4 Edit Review

During Edit Review, editors have four primary tasks:

1. Editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.
2. Editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to make sure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. They ensure that the keys are correct and that all

information in the item is correct. For items with mathematical tasks, editors perform all calculations to ensure accuracy.

3. Editors review all material for fairness and language accessibility issues.
4. Editors confirm that items reflect the accepted guidelines for good item construction. In all items, they look for language that is simple, direct, and free of ambiguity with minimal verbal difficulty. Editors confirm that a problem or task and its stem are clearly defined and concisely worded with no unnecessary information. For multiple-choice interactions, editors check that options are parallel in structure and fit logically and grammatically with the stem and that the key accurately and correctly answers the question as posed, is not inappropriately obvious, and is the only correct answer to an item among the distractors. For constructed-response interactions, editors review the rubrics for appropriate style and grammar.

2.4.5 Senior Review

By the time a science item arrives at Senior Review, it has been thoroughly vetted by both content reviewers and editors. Senior reviewers (in particular, senior content specialists) look back at the item's entire review history, making sure that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy, alignment to the performance expectation, and consistency with the expectations for the highest quality. They check whether the scoring is working as intended and that the scoring assertions adequately address the evidence the student provides with each type of response.

2.5 Review by State Personnel and Stakeholder Committees

All science items have been through an exhaustive external review process. Items in the science bank were reviewed by content experts in one or several states and reviewed and approved by multiple stakeholder committees to evaluate both content and bias/sensitivity.

2.5.1 State Review

After items have been developed for a state participating in the MOU, content experts from the state that owns the item review any eligible items prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, alignment changes, or task demand updates. An AIR director for science reviews all client-requested edits while considering the science item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients have the option to present these items to the committee (based on the edits made) or withhold them from committee review.

AIRCore items are reviewed by at least one or two states. The states provide feedback on the AIRCore items, and the AIR science leadership gathers suggestions and makes edits that improve the AIRCore item. Not all suggestions are implemented, as these items are owned by AIR. Further, most MOU states accept or reject AIRCore and MOU items (as they appear at the time), to be presented to their committees. Some clients skip this step and allow AIR to review all items with their committees before reviewing them. These items can be either set for field testing in a future administration or already at locked operational pool.

2.5.2 Content Advisory Committee Reviews

During the Content Advisory Committee (CAC) reviews, items are reviewed for content validity, grade-level appropriateness, and alignment to the performance expectation. CAC members are typically grade-level and subject-matter experts. During this review, educators also ensure that the scoring assertions make clear what is being scored as correct and give credit where they should (see more information in the Rubric Validation section that follows).

Items developed for each state under the MOU are reviewed by the state that owns the items. AIRCore items are reviewed by the CAC of one or more states. In most cases, items are seen by multiple state committees prior to their field-test or operational use.

A summary of the committee meetings appears in Exhibit C, with further details about the participants in Appendix C.

Exhibit C. Summary of Content Advisory Committee Meetings

Project	Meeting	Number of Committee Members	Number of Items Reviewed
AIRCore	March 2018	26	152
Connecticut	February 2017	41	45
	May 2017	42	40
	October 2017	41	75
	November 2017	35	41
	January 2018	33	42
	October 2018	45	84
	November 2018	49	235
	December 2018	32	56
	January 2019	44	65
	September 2019	50	60
Hawaii	July 2017	22	25
	September 2017	20	65
	October 2018	29	85
	February 2019	21	44
Idaho	December 2018	21	111
MSSA[a]	January 2018	42	73
	March 2018	28	100
	January 2019	21	116
Oregon	August 2017	10	110
	August 2018	20	257
	December 2018	16	62

Project	Meeting	Number of Committee Members	Number of Items Reviewed
Utah	July 2017	23	55
	December 2017	36	48
West Virginia	January 2017	28[b]	39
	October 2018	10	191
	July 2019	12	50
Wyoming	December 2017	17	51
	October 2018	14	37

Note. [a]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[b]Number of Committee Members includes total committee members for English language arts (ELA), mathematics, and science. The number for science-only committee members is not available.

2.5.3 Language Accessibility, Bias, and Sensitivity Committee Reviews

During the bias and sensitivity reviews, stakeholders review items to check for issues that might unfairly impact students based on their background. For example, some states include representatives from student populations such as Special Education, low vision, and the hearing impaired. Further, diverse members of this committee represent students of various ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns.

A summary of the committee meetings appears in Exhibit D, with additional details about the participants in Appendix D.

Exhibit D. Summary of Fairness Committee Meetings

Project	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
AIRCore	March 2018	13	152	N/A
Connecticut	February 2017	6	45	1
	December 2017	9	75	N/A
	December 2017	10	41	N/A
	February 2018	3	42	N/A
	November 2018	11	319	38
	December 2018	10	56	N/A
	January 2019	9	65	N/A
	September 2019	9	48	N/A[a]
Hawaii	July 2017	22	25	2
	September 2017	20	65	13

Project	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
	October 2018	29	85	6
	February 2019	21	44	0
Idaho	December 2018	15	111	1
MSSA[b]	January 2018	21	73	14
	March 2018	11	100	24
	January 2019	14	116	18
Oregon	August 2017	5	110	5
	August 2018	9	256	56
	December 2018	11	62	13
Utah	August 2017	6	44	2
	December 2017	6	48	1
West Virginia	January 2017	28[c]	34	N/A
	January 2019	10	191	N/A
Wyoming	December 2017	5	51	3
	October 2018	5	37	N/A

Note. [a]Number of rejected items has not been finalized through client resolution at the time of writing this report.
[b]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[c]Number of Committee Members includes total committee members for ELA, mathematics, and science. The number for science-only committee members is not available.

2.5.4 Markup for Translation and Accessibility Features

After all approved state- and committee-recommended edits have been applied, the items are considered *locked* and ready for a portion of the accessibility tagging. Text-to-speech tagging is applied prior to field testing while Spanish translations and braille are applied post-field test. Accessibility markup is embedded into each item as part of the item development process rather than as a *post-hoc* process applied to completed tests.

Accessibility markup, whether translations or for text-to-speech, follow similar processes. One trained expert enters the markup, then a second expert reviews the work and recommends changes if necessary. If there is disagreement, a third expert is engaged to resolve the conflict.

Currently, science items are tagged with text-to-speech. Spanish translations, including Spanish text-to-speech and braille, are available for a subset of items.

2.6 FIELD TESTING

A large pool of items was field tested in nine states in spring 2018 for science. For Hawaii, Oregon, and Wyoming, items were embedded as field-test items in the legacy science test. Connecticut and Rhode Island conducted an independent field test in which all students participated, but no scores

were reported. In New Hampshire, Utah, Vermont, and West Virginia, an operational field test was administered.

In 2019, a second wave of items was field tested in nine states. For Hawaii, Idaho elementary school, and Wyoming, unscored field-test items were added as a separate segment to the operational (scored) legacy science test. For a sample of Idaho middle schools, an independent field test in which students were administered a full set of items was conducted. In Connecticut, New Hampshire, Oregon, Rhode Island, Vermont and West Virginia, field-test items were administered as unscored items embedded within the operational items. AIR's field-test process is described in detail in Volume 1, Section 3.2.1, of this technical report.

2.7 POST-FIELD-TEST REVIEW

Following the field test, items were subject to a substantial validation process. This included rubric validation and data review. These processes are described in the following sections.

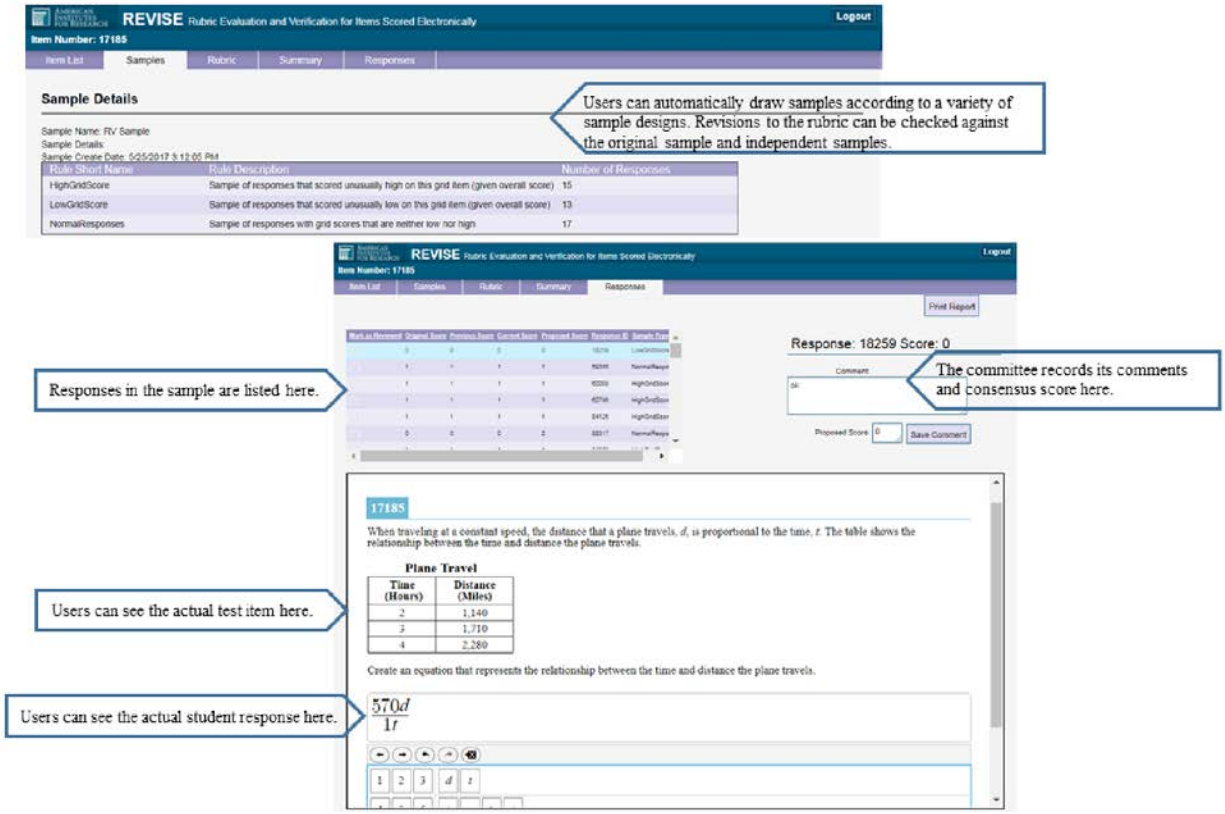
2.7.1 Rubric Validation

The validation process of field-test items begins with rubric validation to verify and make any necessary revisions to the scoring rubrics. The rubric validation process occurs in two phases. During the first phase, AIR content experts work with the analysis team to prepare for the rubric validation meetings. The AIR content experts use the Rubric Evaluation and Verification for Items Scored Electronically (REVISE) system to generate student responses that are scientifically sampled to overrepresent responses most likely to have been mis-scored. Specifically, the sample overrepresents: (1) low-scored responses from otherwise high-scoring students, and (2) high-scored responses from otherwise low-scoring students. This process allows AIR to identify any potential scoring concerns before the rubric validation meeting, such as unanticipated (but accurate) responses, equivalent responses that were not originally considered, and responses that are getting credit but should not (based on the content and the item rubric). At this point, the rubrics may be adjusted and responses rescored.

The second phase of rubric validation involves committees of educators in each state. The committees review the response samples generated by AIR to make recommendations to change or to confirm the rubrics of each item. The committee recommendations are then discussed with the owning state to resolve any inconsistencies. The rubric is then edited or confirmed based on this resolution.

Exhibit E on the following page shows the features of REVISE.

Exhibit E. Features of the REVISE Software



After the rubric validation meetings, AIR staff apply the approved revisions to the rubrics, and any items rejected as part of the process are rejected in ITS. ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the senior content expert once the rubric has been changed, rescoring the entire sample has been completed, and it has been verified that the scoring used the final rubric functioned as intended.

Following rubric validation, all items are subject to statistical checks, and flagged items are presented in data review committees.

2.7.2 Data Review

Following rubric validation, all items are rescored, and classical item statistics are computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning (DIF) statistics. The states established standards for the statistics, and any items violating these standards are flagged for a second educator review. Even though the scoring assertions were the basic units of analysis to compute classical item statistics, the business rules to flag items for additional educator review were established at the item level, because assertions cannot be reviewed in isolation. A common set of business rules was defined for all the states participating in the field test. The classical item statistics were computed on the data of the students testing in the state that owned the item. For Rhode Island and Vermont, which share their item development, statistics were computed on the combined data of students testing in both states.

For AIRCore items, the data from students testing in Connecticut, Idaho grade 8, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia were combined (states that administered AIRCore items and utilized either an independent or operational test).

Volume 1, Section 4, describes in detail the statistical flags that send items to data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members are taught to interpret these flags and are given guidelines for examining the items for content or fairness issues.

For each of the states participating in the MOU, flagged items owned by the state were reviewed by a data review committee. The composition of the data review committees generally consisted of content experts from the state's department of education or state educators (in this case, the state educators were science teachers) and were supported by AIR content experts. AIRCore items were distributed over the data review committees of states participating in the MOU. In summer 2018, AIRCore field-test items were reviewed in webinars with committee members from several states in each session. Outcomes were decided by AIR science content leadership. In summer 2019, AIRCore field-test items were taken to Connecticut, Hawaii, and Idaho for committee review. Outcomes were decided by AIR science content leadership, taking the committees' recommendations into consideration.

At the start of each state-owned item data review meeting, AIR staff leads participants in a training session to familiarize them with the item development process, the purpose of data review, the meaning of the various flags, and the purpose of the data review committee. Committee members are taught to interpret the various flags and are given guidelines for examining the items for content or fairness issues. The training includes a group review of item cards, which detail specific item attributes (including grade level and alignment to the science performance expectations, the content and rubric of the item, and the various item statistics). A sample of the training materials used for these data review meetings appears in Appendix E. Participants use an online environment via laptop computers to review the items in order to interact with them in a manner similar to that of students, and also to view all statistics associated with each item.

Items are then reviewed by participants who are most familiar with the particular grade (band) level and content domain of these items. AIR content specialists, who are also well versed in item statistics, facilitate the discussion in each room with AIR psychometricians available to answer questions as they arise. At the end of each meeting day, AIR content specialists meet with the state content specialists to review the committee recommendations and decide whether to accept the item for inclusion in the operational pool or reject the item from the operational pool. Items that were rejected are potentially eligible for changes to the item and an additional field test.

Exhibit F on the following page summarizes the data review committee meetings. Details, including the composition of each committee, appear in Appendix F.

Exhibit F. Summary of Data Review Committee Meetings

Owner and Item Type	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
AIRCore	July 2018	18	84	8

Owner and Item Type	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Cluster			33	2
Stand-Alone			51	6
AIRCore	August 2019	N/A[a]	43	3
Cluster			0	1
Stand-Alone			43	2
Connecticut	August 2018	29	18	11
Cluster			7	5
Stand-Alone			11	6
Connecticut	August 2019	29	53	20
Cluster			14	6
Stand-Alone			39	14
Hawaii	August 2018	18	32	3
Cluster			7	1
Stand-Alone			25	2
Hawaii	August 2019	18	37	13
Cluster			17	5
Stand-Alone			20	8
Idaho	August 2019	10	12	6
Cluster			4	3
Stand-Alone			8	3
MSSA[b]	August 2018	2[c]	9	6
Cluster			2	0
Stand-Alone			7	6
MSSA[b]	August 2019	2[c]	14	4
Cluster			2	1
Stand-Alone			12	3
Oregon	September 2018	11	44	6
Cluster			28	5
Stand-Alone			16	1
Oregon	August 2019	4	8	7
Cluster			1	1
Stand-Alone			7	6
Utah	August 2018	16	40	6
Cluster			40	6
Stand-Alone			0	0
West Virginia	July 2018	4	3	1

Owner and Item Type	Meeting	Number of Committee Members	Number of Items Reviewed	Number of Items Rejected
Cluster			3	1
Stand-Alone			0	0
West Virginia	September 2019	4	7	6
Cluster			1	1
Stand-Alone			6	5
Wyoming	October 2018	19	16	6
Cluster			6	1
Stand-Alone			10	5
Wyoming	August 2019	10	16	5
Cluster			4	3
Stand-Alone			12	2

Note. [a]In summer 2019, AIRCore field-test items were taken to Connecticut, Hawaii, and Idaho for committee review.
[b]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[c]Conducted by the Rhode Island Department of Education and the Vermont Agency of Education science content experts.

3. SCIENCE ITEM BANK SUMMARY

Tests based on or inspired by the Next Generation Science Standards (NGSS) framework, such as the Multi-State Science Assessment (MSSA), adopt a three-dimensional conceptualization of science understanding, including Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs). Accordingly, the new science assessments are composed mostly of item clusters representing a series of interrelated student interactions directed towards describing, explaining, and predicting scientific phenomenon. Some stand-alone items are added to increase the coverage of the test without also increasing the testing time or testing burden.

AIR Assessment has built the science item bank in partnership with multiple states. The science item bank is robust and has been constructed to support multiple statewide science assessments. As described earlier, science items were written to the NGSS. The science item bank comprises AIR-owned items, which are shared with partner states. These items follow the same specifications, test development processes, and review processes. In 2018, AIR field tested more than 540 item clusters and stand-alone items, of which 451 (including items from all sources) were accepted and made available as operational items in 2019. In 2019, 347 item clusters and stand-alone items were field tested, of which 265 have passed rubric validation and item data review.

Each state using the science item bank selects items that are appropriately aligned and have passed required reviews (as described in Section 2, Item Development Process That Supports Validity of Claims) for use on its statewide assessment. The science item bank continues to grow as participating states continue to field test new items. Participating states collectively share the items and agree to field test new items each year.

3.1 CURRENT COMPOSITION OF THE SCIENCE ITEM BANK

The MSSA tests are composed of item clusters and stand-alone items. Item clusters represent a series of interrelated student interactions directed toward describing, explaining, and predicting scientific phenomena. Item clusters can consist of several item parts requiring the student to interact with the item in various ways. In addition, shorter items (stand-alone items) are included to increase the coverage of the assessments without also increasing testing time or testing burden.

Within each item (item cluster and stand-alone item), a series of explicit assertions is made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses across multiple interactions. For example, a student may correctly graph data points indicating that they can construct a graph showing the relationship between two variables, but they may make an incorrect inference about the relationship between the two variables, therefore not supporting the assertion that the student can interpret relationships expressed graphically. Table 1 below lists the science interaction types. Examples of various interaction types can be found in Appendix G.

Table 1. Science Interaction Types and Descriptions

Interaction Type	Associated Sub-Types	Description
Choice	Multiple-Choice	Traditional multiple-choice interaction allows the student to select a single option from a list of possible answer options.
	Multi-Select	Traditional multi-select interaction (checkboxes) allows students to select one or more options from a list of possible answer choices.
Text Entry	Simple Text Entry	Students type a response in a text box.
	Embedded Text Entry	Students type their response in one or more text boxes that are embedded in a section of read-only text.
	Natural Language	Students are directed to provide a short, written response.
	Extended Response	Students are directed to provide a longer, written response in the form of an essay.
Table	Table Match	Interaction allows students to check a box to indicate if the information from a column header matches information from a row header.
	Table Input	Interaction solicits a student to complete tabular data.
Edit Task	Edit Task	A student clicks a word and replaces it with another word that they type to revise a sentence.

Interaction Type	Associated Sub-Types	Description
	Edit Task with Choice	A student clicks a word or phrase and chooses the replacement from several options.
	Edit Task Inline Choice	Drop-down menus are placed through the text, and a student chooses the right option to complete the text.
Hot Text	Selectable	Selectable hot text interactions require students to select one or more text elements in the response area.
	Re-orderable	Re-orderable hot text interactions require students to click and drag hot text elements into a different order.
	Drag-from-Palette	Drag-from-palette hot text interactions require students to drag elements from a palette into the available blank table cells or *gaps* (text boxes) in the response area.
	Custom	Custom hot text interactions combine the functionality of the other hot text interaction sub-types. Students responding to a custom hot text interaction may need to select text elements, rearrange text elements, and/or drag text elements from a palette to blank table cells or drop targets in the response area.
Equation	N/A	Equation interactions require students to enter a response into input boxes. These boxes may stand alone, or they may be in line with text or embedded in a table. The equation interaction may have an on-screen keypad that may consist of special mathematics characters. Students may also enter their response via a physical keyboard.
Grid	Grid	Grid interactions require students to enter a response by interacting with a grid area in the answer space. The student may be required to draw a line or shape, plot a point, or create a graph. The student may also drag and drop or click on selectable hot spots.
	Hot Spot	Hot spot interaction sub-types allow you to create grid interactions with specific hot spot functionality. These interactions require students to select hot spot regions in the grid area.
	Graphic Gap Match	Graphic gap match interactions allow you to create grid interactions with specific drag-and-drop functionality. These interactions require students to drag image objects from a palette to specified regions (gaps) in the grid area.
Simulation	N/A	Simulation interactions allow the student to investigate a phenomenon by selecting variables to get output data. Some simulations are accompanied by animations.

Table 2 through Table 6 on the following pages provide the number of items in the across-state science item bank available for use in the spring 2019 statewide assessments. Appendix H provides the science item bank available by grade band, performance expectation (PE), and origin.

Table 2. Across-State Science Spring 2019 Operational and Field-Test Item Bank

Grade Band	Item Type	Total Sp19 AIRCore Items	Total Sp19 MSSA Items	Total Sp19 Other MOU State Items[a]	Total Sp19 Items
Elementary School	Cluster	32	9	85	126
	Stand-Alone	47	8	71	126
Middle School	Cluster	29	6	140	175
	Stand-Alone	50	7	89	146
High School	Cluster	30	5	66	101
	Stand-Alone	54	7	63	124
Total		**242**	**42**	**514**	**798**

Note. [a]Other Memorandum of Understanding (MOU) states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 3. Across-State Science Spring 2019 Operational Item Bank

Grade Band	Item Type	Sp19 AIRCore OP Items	Sp19 MSSA OP Items	Sp19 Other MOU State OP Items[a]	Total Sp19 OP Items
Elementary School	Cluster	32	5	39	76
	Stand-Alone	29	2	28	59
Middle School	Cluster	25	3	109	137
	Stand-Alone	26	2	29	57
High School	Cluster	28	2	36	66
	Stand-Alone	27	4	25	56
Total		**167**	**18**	**266**	**451**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 4. Across-State Science Spring 2019 Field-Test Item Bank

Grade Band	Item Type	Sp19 AIRCore FT Items	Sp19 MSSA FT Items	Sp19 Other MOU State FT Items[a]	Total Sp19 FT Items
Elementary School	Cluster	0	4	46	50
	Stand-Alone	18	6	43	67
Middle School	Cluster	4	3	31	38
	Stand-Alone	24	5	60	89
High School	Cluster	2	3	30	35

Grade Band	Item Type	Sp19 AIRCore FT Items	Sp19 MSSA FT Items	Sp19 Other MOU State FT Items[a]	Total Sp19 FT Items
	Stand-Alone	27	3	38	68
Total		**75**	**24**	**248**	**347**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 5. Across-State Science Spring 2019 Operational and Field-Test Item Bank by Science Discipline

Grade Band	Science Discipline	Item Type	Total Sp19 AIRCore Items	Total Sp19 MSSA Items	Total Sp19 Other MOU State Items[a]	Total Sp19 Items
Elementary School	Earth and Space Sciences	Cluster	11	3	23	37
		Stand-Alone	12	4	23	39
	Life Sciences	Cluster	11	3	31	45
		Stand-Alone	17	3	22	42
	Physical Sciences	Cluster	10	3	31	44
		Stand-Alone	18	1	26	45
Middle School	Earth and Space Sciences	Cluster	9	2	37	48
		Stand-Alone	17	1	29	47
	Life Sciences	Cluster	8	2	57	67
		Stand-Alone	23	3	27	53
	Physical Sciences	Cluster	12	2	45	59
		Stand-Alone	10	3	33	46
	Engineering, Technology, and Applications of Science	Cluster	0	0	1	1
		Stand-Alone	0	0	0	0
High School	Earth and Space Sciences	Cluster	6	3	12	21
		Stand-Alone	11	3	11	25
	Life Sciences	Cluster	16	1	36	53
		Stand-Alone	35	2	29	66
	Physical Sciences	Cluster	8	1	18	27
		Stand-Alone	8	2	23	33
Total			**242**	**42**	**514**	**798**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 6. Across-State Science Spring 2019 Operational and Field-Test Item Bank by Disciplinary Core Idea

Grade Band	Science Discipline	Disciplinary Core Idea	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
Elementary School	Earth and Space Sciences	ESS1: Earth's Place in the Universe	7	2	13	22
		ESS2: Earth's Systems	10	3	25	38
		ESS3: Earth and Human Activity	6	2	8	16
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	10	3	20	33
		LS2: Ecosystems: Interactions, Energy, and Dynamics	4	0	11	15
		LS3: Heredity: Inheritance and Variation of Traits	2	3	6	11
		LS4: Biological Evolution: Unity and Diversity	12	0	16	28
	Physical Sciences	PS1: Matter and Its Interactions	6	2	13	21
		PS2: Motion and Stability: Forces and Interactions	7	1	18	26
		PS3: Energy	13	1	16	30
		PS4: Waves and Their Applications in Technologies for Information Transfer	2	0	10	12
Middle School	Earth and Space Sciences	ESS1: Earth's Place in the Universe	12	2	15	29
		ESS2: Earth's Systems	5	0	28	33
		ESS3: Earth and Human Activity	9	1	23	33
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	5	2	31	38
		LS2: Ecosystems: Interactions, Energy, and Dynamics	15	2	22	39
		LS3: Heredity: Inheritance and Variation of Traits	2	0	10	12
		LS4: Biological Evolution: Unity and Diversity	9	1	21	31
	Physical Sciences	PS1: Matter and Its Interactions	6	3	30	39
		PS2: Motion and Stability: Forces and Interactions	3	1	21	25

Grade Band	Science Discipline	Disciplinary Core Idea	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
		PS3: Energy	8	1	16	25
		PS4: Waves and Their Applications in Technologies for Information Transfer	5	0	11	16
	Engineering, Technology, and Applications of Science	ETS1: Engineering Design	0	0	1	1
High School	Earth and Space Sciences	ESS1: Earth's Place in the Universe	6	2	10	18
		ESS2: Earth's Systems	5	3	8	16
		ESS3: Earth and Human Activity	6	1	5	12
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	11	1	23	35
		LS2: Ecosystems: Interactions, Energy, and Dynamics	15	1	21	37
		LS3: Heredity: Inheritance and Variation of Traits	8	0	4	12
		LS4: Biological Evolution: Unity and Diversity	17	1	17	35
	Physical Sciences	PS1: Matter and Its Interactions	8	2	15	25
		PS2: Motion and Stability: Forces and Interactions	4	1	10	15
		PS3: Energy	4	0	11	15
		PS4: Waves and Their Applications in Technologies for Information Transfer	0	0	5	5
Total			**242**	**42**	**514**	**798**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

3.2 STRATEGY FOR POOL EVALUATION AND REPLENISHMENT

AIR and MOU states continue to develop items to replenish and grow the science item pool. The general strategy for targeting item development gathers information from three sources:

1. Characteristics of released items to be replaced
2. Characteristics of items that are overused
3. Tabulations of content coverage and ranges of difficulty to identify gaps in the pool

Before a test goes live, simulations are used to fine-tune the parameters of the algorithm that governs the item selection in a linear-on-the-fly (LOFT) test design. Among the many reports from the simulator are items that are seen by more than 20% of students. The characteristics of these items are the primary targets for development. Overused items become candidates for release two years from this time, once replacements have been introduced into the operational pool.

4. MULTI-STATE SCIENCE ASSESSMENT TEST CONSTRUCTION

4.1 TEST DESIGN

The Multi-State Science Assessment (MSSA) was administered online to students in grades 5, 8, and 11 using a linear-on-the-fly (LOFT) test design. Contrary to a fixed form, every student potentially sees a different set of items. Items are selected by an item selection algorithm so that the blueprint is met whenever possible. The algorithm that was used is the same algorithm that AIR uses for the administration of adaptive tests. The adaptive item selection algorithm selects items based on their content value and information value. At any given point during the test, the content value of an item is determined by its contribution to meeting the blueprint, given the content characteristics of the items that have already been administered.

During the test, the content value increases for items that exhibit features that have not met their designated minimum as the end of the test approaches. Vice versa, the content value decreases for items with content features for which the minimum has been met. The information value of an item is based on the item information function evaluated at the estimated proficiency. The proficiency estimate is updated throughout the test. By assigning a weight of zero to the information value of an item with respect to the underlying proficiency, the items are selected solely based on their contributions to meeting the blueprint. By assigning weights of zero to the information value of an item with respect to the underlying latent variable, the items are selected based solely on their contribution to meeting the blueprint. Details for AIR's adaptive testing algorithm are described in Appendix J. Adaptive Algorithm Design.

For the 2018 independent field test, a segmented design was used; items were administered grouped in four segments. The segments correspond to each of the three science disciplines and a (additional) field-test segment that could contain items from all three science disciplines.

In 2018, the order of the segments corresponding to the science disciplines was randomized over students. The additional field-test segment consisted of one item cluster and was always presented

at the end of the test (segment four). The primary purpose was to collect additional student responses for the item clusters that had low exposure in the first three segments.

In 2019, the scored operational part of the test consisted of the three segments corresponding to science disciplines. The embedded field-test segment consisted of two item clusters and four stand-alone items. In order to ensure that every student received exactly two item clusters and four stand-alone items as field-test items, the embedded field-test segment was split into two segments: one for field-test item clusters, and one for field-test stand-alone items.

The test was taken over two days. On the first day, half of the students received two operational segments, chosen at random from the three operational segments. The other half received one randomly chosen operational segment and the embedded field-test segments. The remaining segments were administered on the second day. Within one day, the order of the segments was randomized, with the restriction that the field-test segments for item clusters and stand-alone items were always administered right after each other.

4.2 TEST BLUEPRINTS

Test blueprints provide the following guidelines:

- Length of the test
- Science disciplines to be covered and the acceptable number of items across performance expectations within each science discipline and Disciplinary Core Idea (DCI)

The blueprint for science is given in Table 7 through Table 9 on the following pages.

Table 7. Science Test Blueprint, Grade 5

Grade 5	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline–Physical Sciences, PE Total = 17	**2**	**2**	**4**	**4**	**6**	**6**
DCI–Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
3-PS2-1: Forces-balanced and unbalanced forces	0	1	0	1	0	1
3-PS2-2: Forces-pattern predicts future motion	0	1	0	1	0	1
3-PS2-3: Forces-between objects not in contact	0	1	0	1	0	1
3-PS2-4: Forces-magnets*	0	1	0	1	0	1
5-PS2-1: Space Systems	0	1	0	1	0	1
DCI–Energy	**0**	**1**	**0**	**2**	**0**	**3**
4-PS3-1: Energy-relationship between speed and energy of object	0	1	0	1	0	1
4-PS3-2: Energy-transfer of energy	0	1	0	1	0	1
4-PS3-3: Energy-changes in energy when objects collide	0	1	0	1	0	1
4-PS3-4: Energy-converting energy from one form to another*	0	1	0	1	0	1
5-PS3-1: Matter and Energy	0	1	0	1	0	1
DCI–Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
4-PS4-1: Waves-waves can cause objects to move	0	1	0	1	0	1
4-PS4-2: Structure, Function, Information Processing	0	1	0	1	0	1
4-PS4-3: Waves-using patterns to transfer information*	0	1	0	1	0	1
DCI–Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
5-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
5-PS1-2: Structure and Properties of Matter	0	1	0	1	0	1
5-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
5-PS1-4: Structure and Properties of Matter	0	1	0	1	0	1

Grade 5	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline–Life Sciences, PE Total = 12	**2**	**2**	**4**	**4**	**6**	**6**
DCI–From Molecules to Organisms: Structure and Function	**0**	**1**	**0**	**2**	**0**	**3**
3-LS1-1: Inheritance	0	1	0	1	0	1
4-LS1-1: Structure, Function, Information Processing	0	1	0	1	0	1
4-LS1-2: Structure, Function, Information Processing	0	1	0	1	0	1
5-LS1-1: Matter and Energy	0	1	0	1	0	1
DCI–Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
3-LS2-1: Ecosystems	0	1	0	1	0	1
5-LS2-1: Matter and Energy	0	1	0	1	0	1
DCI–Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
3-LS3-1: Inheritance	0	1	0	1	0	1
3-LS3-2: Inheritance	0	1	0	1	0	1
DCI–Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
3-LS4-1: Ecosystems	0	1	0	1	0	1
3-LS4-2: Inheritance	0	1	0	1	0	1
3-LS4-3: Ecosystems	0	1	0	1	0	1
3-LS4-4: Ecosystems*	0	1	0	1	0	1
Discipline–Earth and Space Sciences, PE Total = 13	**2**	**2**	**4**	**4**	**6**	**6**
DCI–Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS2-1: Weather and Climate	0	1	0	1	0	1
3-ESS2-2: Weather and Climate	0	1	0	1	0	1
4-ESS2-1: Earth's Systems and Processes	0	1	0	1	0	1
4-ESS2-2: Earth's Systems and Processes	0	1	0	1	0	1
5-ESS2-1: Earth's Systems	0	1	0	1	0	1

Grade 5	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
5-ESS2-2: Earth's Systems	0	1	0	1	0	1
DCI–Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
3-ESS3-1: Weather and Climate*	0	1	0	1	0	1
4-ESS3-2: Earth's Systems and Processes*	0	1	0	1	0	1
4-ESS3-1: Energy	0	1	0	1	0	1
5-ESS3-1: Earth's Systems	0	1	0	1	0	1
DCI–Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
4-ESS1-1: Earth's Systems and Processes	0	1	0	1	0	1
5-ESS1-1: Space Systems	0	1	0	1	0	1
5-ESS1-2: Space Systems	0	1	0	1	0	1
PE Total = 42	**6**	**6**	**12**	**12**	**18**	**18**

Note. * These performance expectations (PEs) have an engineering component.

Table 8. Science Test Blueprint, Grade 8

Grade 8	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline–Physical Sciences, PE Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI–Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-2: Chemical Reactions	0	1	0	1	0	1
MS-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-4: Structure and Properties of Matter	0	1	0	1	0	1
MS-PS1-5: Chemical Reactions	0	1	0	1	0	1
MS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
DCI–Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS2-1: Forces and Interactions*	0	1	0	1	0	1
MS-PS2-2: Forces and Interactions	0	1	0	1	0	1
MS-PS2-3: Forces and Interactions	0	1	0	1	0	1
MS-PS2-4: Forces and Interactions	0	1	0	1	0	1
MS-PS2-5: Forces and Interactions	0	1	0	1	0	1
DCI–Energy	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS3-1: Energy	0	1	0	1	0	1
MS-PS3-2: Energy	0	1	0	1	0	1
MS-PS3-3: Energy*	0	1	0	1	0	1
MS-PS3-4: Energy	0	1	0	1	0	1
MS-PS3-5: Energy	0	1	0	1	0	1
DCI–Waves and Their Applications in Technologies for Information Transfer	**0**	**1**	**0**	**2**	**0**	**3**
MS-PS4-1: Waves and Electromagnetic Radiation	0	1	0	1	0	1
MS-PS4-2: Waves and Electromagnetic Radiation	0	1	0	1	0	1

Grade 8	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
MS-PS4-3: Waves and Electromagnetic Radiation	0	1	0	1	0	1
Discipline–Life Sciences, PE Total = 21	**2**	**2**	**4**	**4**	**6**	**6**
DCI–From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS1-1: Structure, Function, Information Processing	0	1	0	1	0	1
MS-LS1-2: Structure, Function, Information Processing	0	1	0	1	0	1
MS-LS1-3: Structure, Function, Information Processing	0	1	0	1	0	1
MS-LS1-4: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS1-5: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS1-6: Matter and Energy	0	1	0	1	0	1
MS-LS1-7: Matter and Energy	0	1	0	1	0	1
MS-LS1-8: Structure, Function, Information Processing	0	1	0	1	0	1
DCI–Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS2-1: Matter and Energy	0	1	0	1	0	1
MS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
MS-LS2-3: Matter and Energy	0	1	0	1	0	1
MS-LS2-4: Matter and Energy	0	1	0	1	0	1
MS-LS2-5: Interdependent Relationships in Ecosystems*	0	1	0	1	0	1
DCI–Hereditary: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS3-1: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS3-2: Growth, Development, Reproduction	0	1	0	1	0	1
DCI–Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
MS-LS4-1: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-2: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-3: Natural Selection and Adaptation	0	1	0	1	0	1
MS-LS4-4: Natural Selection and Adaptation	0	1	0	1	0	1

Grade 8	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
MS-LS4-5: Growth, Development, Reproduction	0	1	0	1	0	1
MS-LS4-6: Natural Selection and Adaptation	0	1	0	1	0	1
Discipline–Earth and Space Sciences, PE Total = 15	**2**	**2**	**4**	**4**	**6**	**6**
DCI–Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS1-1: Space Systems	0	1	0	1	0	1
MS-ESS1-2: Space Systems	0	1	0	1	0	1
MS-ESS1-3: Space Systems	0	1	0	1	0	1
MS-ESS1-4: History of Earth	0	1	0	1	0	1
DCI–Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS2-1: Earth's Systems	0	1	0	1	0	1
MS-ESS2-2: History of Earth	0	1	0	1	0	1
MS-ESS2-3: History of Earth	0	1	0	1	0	1
MS-ESS2-4: Earth's Systems	0	1	0	1	0	1
MS-ESS2-5: Weather and Climate	0	1	0	1	0	1
MS-ESS2-6: Weather and Climate	0	1	0	1	0	1
DCI–Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
MS-ESS3-1: Earth's Systems	0	1	0	1	0	1
MS-ESS3-2: Human Impacts	0	1	0	1	0	1
MS-ESS3-3: Human Impacts*	0	1	0	1	0	1
MS-ESS3-4: Human Impacts	0	1	0	1	0	1
MS-ESS3-5: Weather and Climate	0	1	0	1	0	1
PE Total = 55	**6**	**6**	**12**	**12**	**18**	**18**

Note. * These PEs have an engineering component.

Table 9. Science Test Blueprint, Grade 11

Grade 11	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Discipline–Physical Sciences, PE Total = 24	**2**	**2**	**4**	**4**	**6**	**6**
DCI–Matter and Its Interactions	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS1-1: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-2: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-3: Structure and Properties of Matter	0	1	0	1	0	1
HS-PS1-4: Chemical Reactions	0	1	0	1	0	1
HS-PS1-5: Chemical Reactions	0	1	0	1	0	1
HS-PS1-6: Chemical Reactions*	0	1	0	1	0	1
HS-PS1-7: Chemical Reactions	0	1	0	1	0	1
HS-PS1-8: Nuclear Processes	0	1	0	1	0	1
DCI–Motion and Stability: Forces and Interactions	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS2-1: Forces and Motion	0	1	0	1	0	1
HS-PS2-2: Forces and Motion	0	1	0	1	0	1
HS-PS2-3: Forces and Motion*	0	1	0	1	0	1
HS-PS2-4: Types of Interactions	0	1	0	1	0	1
HS-PS2-5: Types of Interactions	0	1	0	1	0	1
HS-PS2-6: Chemical Reactions*	0	1	0	1	0	1
DCI–Energy	**0**	**1**	**0**	**2**	**0**	**3**
HS-PS3-1: Energy	0	1	0	1	0	1
HS-PS3-2: Energy	0	1	0	1	0	1
HS-PS3-3: Energy*	0	1	0	1	0	1
HS-PS3-4: Energy	0	1	0	1	0	1
HS-PS3-5: Energy	0	1	0	1	0	1
DCI–Waves and Their Applications in Technologies for	**0**	**1**	**0**	**2**	**0**	**3**

Grade 11	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
Information Transfer						
HS-PS4-1: Wave Properties	0	1	0	1	0	1
HS-PS4-2: Wave Properties	0	1	0	1	0	1
HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	1	0	1	0	1
HS-PS4-4: Electromagnetic Radiation	0	1	0	1	0	1
HS-PS4-5: Electromagnetic Radiation*	0	1	0	1	0	1
Discipline–Life Sciences, PE Total = 24	**2**	**2**	**4**	**4**	**6**	**6**
DCI–From Molecules to Organisms: Structures and Processes	**0**	**1**	**0**	**2**	**0**	**3**
HS-LS1-1: Structure and Function	0	1	0	1	0	1
HS-LS1-2: Structure and Function	0	1	0	1	0	1
HS-LS1-3: Structure and Function	0	1	0	1	0	1
HS-LS1-4: Growth and Development of Organisms	0	1	0	1	0	1
HS-LS1-5: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
HS-LS1-6: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
HS-LS1-7: Organization for Matter and Energy Flow in Organisms	0	1	0	1	0	1
DCI–Ecosystems: Interactions, Energy, and Dynamics	**0**	**1**	**0**	**2**	**0**	**3**
HS-LS2-1: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
HS-LS2-2: Interdependent Relationships in Ecosystems	0	1	0	1	0	1
HS-LS2-3: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-4: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-5: Cycles of Matter and Energy Transfer in Ecosystems	0	1	0	1	0	1
HS-LS2-6: Ecosystem Dynamics, Functioning, and Resilience	0	1	0	1	0	1
HS-LS2-7: Ecosystem Dynamics, Functioning, and Resilience*	0	1	0	1	0	1
HS-LS2-8: Social Interactions and Group Behavior	0	1	0	1	0	1
DCI–Heredity: Inheritance and Variation of Traits	**0**	**1**	**0**	**2**	**0**	**3**

Grade 11	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
HS-LS3-1: Structure and Function	0	1	0	1	0	1
HS-LS3-2: Variation of Traits	0	1	0	1	0	1
HS-LS3-3: Variation of Traits	0	1	0	1	0	1
DCI–Biological Evolution: Unity and Diversity	**0**	**1**	**0**	**2**	**0**	**3**
HS-LS4-1: Evidence of Common Ancestry and Diversity	0	1	0	1	0	1
HS-LS4-2: Natural Selection	0	1	0	1	0	1
HS-LS4-3: Natural Selection	0	1	0	1	0	1
HS-LS4-4: Adaptation	0	1	0	1	0	1
HS-LS4-5: Adaptation	0	1	0	1	0	1
HS-LS4-6: Adaptation*	0	1	0	1	0	1
Discipline–Earth and Space Sciences, PE Total = 19	**2**	**2**	**4**	**4**	**6**	**6**
DCI–Earth's Place in the Universe	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS1-1: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-2: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-3: The Universe and Its Stars	0	1	0	1	0	1
HS-ESS1-4: Earth and the Solar System	0	1	0	1	0	1
HS-ESS1-5: The History of Planet Earth	0	1	0	1	0	1
HS-ESS1-6: The History of Planet Earth	0	1	0	1	0	1
DCI–Earth's Systems	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS2-1: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-2: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-3: Earth Materials and Systems	0	1	0	1	0	1
HS-ESS2-4: Weather and Climate	0	1	0	1	0	1
HS-ESS2-5: The Roles of Water in Earth's Surface Processes	0	1	0	1	0	1
HS-ESS2-6: Weather and Climate	0	1	0	1	0	1

Grade 11	Min Item Clusters	Max Item Clusters	Min Stand-Alone Items	Max Stand-Alone Items	Min Item Clusters + Min Stand-Alone Items	Max Item Clusters + Max Stand-Alone Items
HS-ESS2-7: Weather and Climate	0	1	0	1	0	1
DCI–Earth and Human Activity	**0**	**1**	**0**	**2**	**0**	**3**
HS-ESS3-1: Natural Resources	0	1	0	1	0	1
HS-ESS3-2: Natural Resources*	0	1	0	1	0	1
HS-ESS3-3: Human Impacts on Earth Systems	0	1	0	1	0	1
HS-ESS3-4: Human Impacts on Earth Systems*	0	1	0	1	0	1
HS-ESS3-5: Global Climate Change	0	1	0	1	0	1
HS-ESS3-6: Global Climate Change*	0	1	0	1	0	1
PE Total = 67	**6**	**6**	**12**	**12**	**18**	**18**

Note. *These PEs have an engineering component.

Main characteristics of the blueprint were that any PE could be tested only once (indicated by the values of 0 and 1 for the Min and Max values of the individual PEs in Table 7 through Table 9); in general, no more than one item cluster or two stand-alone items could be sampled from the same DCI, and no more than three total items could be sampled from the same DCI (as indicated by the Min and Max values in the rows representing DCIs).

While tests are not timed, the Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) published estimated testing times for the MSSA. Combined percentile 85 of testing times are presented in Table 10, Rhode Island percentile 85 of testing times are presented in Table 11, and Vermont percentile 85 of testing times are presented in Table 12, all below.

Table 10. Combined Percentile 85 Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	5	123.17
	8	110.16
	11	94.95

Table 11. Rhode Island Percentile 85 Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	5	123.60
	8	108.44
	11	93.31

Table 12. Vermont Percentile 85 Testing Times by Grade

Subject	Grade	85th Percentile Testing
Science	5	122.50
	8	113.18
	11	97.83

4.3 ONLINE TEST CONSTRUCTION

During fall 2018, AIR psychometricians and content experts worked with RIDE and VT AOE content specialists and leadership to build item pools for the spring 2019 administration. The MSSA test construction utilizes a structured test construction plan, explicit blueprints, and active collaborative participation from all parties.

The 2019 MSSA item pools were built by AIR test developers to match items exactly to the detailed test blueprints. Operational items were selected from eight item banks (AIRCore, Connecticut, Hawaii, MSSA, Oregon, Utah, West Virginia, and Wyoming) to fulfill the blueprint for that grade. Table 13 through Table 17 on the following pages summarize the 2019 MSSA item pool. Appendix I. Multi-State Assessment Item Pool provides the 2019 MSSA item pool by grade, PE, and origin.

Table 13. MSSA Spring 2019 Operational and Field-Test Item Pool

Grade	Item Type	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
5	Cluster	19	9	39	67
	Stand-Alone	25	8	28	61
8	Cluster	11	5	22	38
	Stand-Alone	31	7	40	78
11	Cluster	20	5	37	62
	Stand-Alone	29	7	25	61
Total		**135**	**41**	**191**	**367**

Note. [a]Other Memorandum of Understanding (MOU) states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 14. MSSA Spring 2019 Operational Item Pool

Grade	Item Type	Sp19 AIRCore OP Items	Sp19 MSSA OP Items	Sp19 Other MOU State OP Items[a]	Total Sp19 OP Items
5	Cluster	19	5	26	50
	Stand-Alone	20	2	14	36
8	Cluster	11	2	15	28
	Stand-Alone	25	2	15	42
11	Cluster	20	2	22	44
	Stand-Alone	23	4	15	42
Total		**118**	**17**	**107**	**242**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 15. MSSA Spring 2019 Field-Test Item Pool

Grade	Item Type	Sp19 AIRCore FT Items	Sp19 MSSA FT Items	Sp19 Other MOU State FT Items[a]	Total Sp19 FT Items
5	Cluster	0	4	13	17
	Stand-Alone	5	6	14	25
8	Cluster	0	3	7	10
	Stand-Alone	6	5	25	36
11	Cluster	0	3	15	18
	Stand-Alone	6	3	10	19
Total		**17**	**24**	**84**	**125**

Note. [a]Other MOU states include Connecticut Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 16. MSSA Spring 2019 Operational and Field-Test Item Pool by Science Discipline

Grade	Science Discipline	Item Type	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
5	Earth and Space Sciences	Cluster	6	3	10	19
		Stand-Alone	5	4	5	14
	Life Sciences	Cluster	6	3	15	24
		Stand-Alone	9	3	9	21
	Physical Sciences	Cluster	7	3	14	24
		Stand-Alone	11	1	14	26
8	Earth and Space Sciences	Cluster	6	2	9	17
		Stand-Alone	9	1	14	24
	Life Sciences	Cluster	2	1	8	11
		Stand-Alone	14	3	11	28
	Physical Sciences	Cluster	3	2	5	10
		Stand-Alone	8	3	15	26
11	Earth and Space Sciences	Cluster	3	3	8	14
		Stand-Alone	9	3	7	19
	Life Sciences	Cluster	10	1	18	29
		Stand-Alone	13	2	8	23

Grade	Science Discipline	Item Type	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
	Physical Sciences	Cluster	7	1	11	19
		Stand-Alone	7	2	10	19
Total			**135**	**41**	**191**	**367**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table 17. MSSA Spring 2019 Operational and Field-Test Item Pool by Disciplinary Core Idea

Grade	Science Discipline	Disciplinary Core Idea	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
5	Earth and Space Sciences	ESS1: Earth's Place in the Universe	3	2	4	9
		ESS2: Earth's Systems	7	3	10	20
		ESS3: Earth and Human Activity	1	2	1	4
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	4	3	11	18
		LS2: Ecosystems: Interactions, Energy, and Dynamics	4	0	3	7
		LS3: Heredity: Inheritance and Variation of Traits	1	3	4	8
		LS4: Biological Evolution: Unity and Diversity	6	0	6	12
	Physical Sciences	PS1: Matter and Its Interactions	5	2	3	10
		PS2: Motion and Stability: Forces and Interactions	3	1	8	12
		PS3: Energy	10	1	11	22
		PS4: Waves and Their Applications in Technologies for Information Transfer	0	0	6	6
8	Earth and Space Sciences	ESS1: Earth's Place in the Universe	4	2	7	13
		ESS2: Earth's Systems	5	0	12	17
		ESS3: Earth and Human Activity	6	1	4	11
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	4	1	9	14
		LS2: Ecosystems: Interactions, Energy, and Dynamics	5	2	6	13
		LS3: Heredity: Inheritance and Variation of Traits	2	0	1	3
		LS4: Biological Evolution: Unity and Diversity	5	1	3	9
	Physical Sciences	PS1: Matter and Its Interactions	2	3	7	12

Grade	Science Discipline	Disciplinary Core Idea	Sp19 AIRCore Items	Sp19 MSSA Items	Sp19 Other MOU State Items[a]	Total Sp19 Items
		PS2: Motion and Stability: Forces and Interactions	2	1	3	6
		PS3: Energy	5	1	5	11
		PS4: Waves and Their Applications in Technologies for Information Transfer	2	0	5	7
11	Earth and Space Sciences	ESS1: Earth's Place in the Universe	4	2	5	11
		ESS2: Earth's Systems	5	3	6	14
		ESS3: Earth and Human Activity	3	1	4	8
	Life Sciences	LS1: From Molecules to Organisms: Structure and Function	6	1	9	16
		LS2: Ecosystems: Interactions, Energy, and Dynamics	8	1	6	15
		LS3: Heredity: Inheritance and Variation of Traits	2	0	3	5
		LS4: Biological Evolution: Unity and Diversity	7	1	8	16
	Physical Sciences	PS1: Matter and Its Interactions	8	2	6	16
		PS2: Motion and Stability: Forces and Interactions	2	1	7	10
		PS3: Energy	4	0	6	10
		PS4: Waves and Their Applications in Technologies for Information Transfer	0	0	2	2
Total			**135**	**41**	**191**	**367**

Note. [a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia and Wyoming.

More information about *p*-values, biserial correlations, and item response theory (IRT) parameters can be found in Volume 1. The details on calibration, equating, and scoring of the MSSA can also be found in Volume 1.

4.4 PAPER-PENCIL ACCOMMODATION FORM CONSTRUCTION

Student scores should not depend upon the mode of administration or type of test form. Because the MSSA was primarily administered in an online test system in spring 2019, only 19 students took the paper-pencil form in grade 5, 15 in grade 8, and 10 in grade 11. Scores obtained via alternate modes of administration must be established as comparable to scores obtained through online testing. This section outlines the overall test development plans that ensured the comparability of online and paper-pencil tests.

To build paper-pencil forms, content specialists began with the online pool and removed any items that could not be rendered on paper. Next, content specialists constructed fixed forms adhering to the test blueprint. All overall and discipline (reporting category) level blueprint requirements were met; however, due to the availability of items in paper-pencil forms, some blueprint requirements at the DCI and PE levels were violated. For the grade 5 paper-pencil test, there were two items measuring the same PE in the "Earth's Place in the Universe" DCI, and two items measuring the same PE in the "Earth's Systems" DCI. In addition, two item clusters were selected from the "Motion and Stability: Forces and Interactions" DCI. For the grade 8 paper-pencil form, there were two item clusters from the "Earth's Systems" DCI, and two items measuring the same PE in the "Biological Evolution: Unity and Diversity" DCI. For the grade 11 paper-pencil test, the blueprint specifies that the two item clusters cannot pertain to the same DCI. The two Earth and Space Sciences item clusters were selected from the "Earth and Human Activity" DCI, and two Physical Sciences item clusters were selected from the "Matters and Its Interactions" DCI, violating the blueprint constraint that no two item clusters could come from the same DCI.

Future item development will focus on developing items that are amenable to paper delivery for those places where the blueprint was not met.

5. SIMULATION SUMMARY REPORT

This section describes the results of simulated test administrations used to configure and evaluate the adequacy of the item selection algorithm used to administer the 2018–2019 Multi-State Science Assessments (MSSA) assessments for grades 5, 8, and 11. Simulations were carried out to configure the settings of the algorithm and to evaluate whether individual tests adhered to the test blueprint.

Psychometricians reviewed the simulation results for the following key diagnostics:

- Match-to-test blueprint: Determines that the tests have the correct number of test items overall and the appropriate proportion by content categories at each level of the content hierarchy, as specified in the test blueprints for every science grade.
- Item exposure rate: Evaluates the utility of item pools and identifies overexposed and underexposed items.

These diagnostics are interrelated. For example, if the test pool for a particular content category is limited (i.e., there are only a few test items available), achieving a 100% match to the blueprint for this content level will lead to a high item exposure rate, which means that a large number of students are sharing items. The software system that performs the simulation allows the adjustment of setting parameters to attain the best possible balance among these diagnostics. The simulation involves an iterative process that reviews initial results, adjusts these system parameters, runs new simulations, reviews the new results, and repeats the exercise until an optimal balance is achieved. The final setting would then be applied for the operational tests.

5.1 FACTORS AFFECTING SIMULATION RESULTS

There are several factors that may influence simulation results for a linear-on-the-fly (LOFT) test administration. These include the following:

- The proportional relationship between the pool and the constraints to be met. Proportionally distributed pools tend to make better use of the pool (i.e., more uniform item exposure) and make it easier to meet blueprint and other constraints. For example, if the specifications call for at least one item cluster per Disciplinary Core Idea (DCI), but the pool has no item cluster for some DCIs, it may be impossible to meet this constraint.

- The correlational structure between constraints. It is easier to satisfy a constraint if there are instances of the constraint at all levels of another constraint. For example, if stand-alone items within a discipline are associated only with a specific DCI, it may be difficult to meet both the desired distribution of content and the desired distribution of item type.

- Whether or not there is a strict maximum on a given constraint. This means that the requirement must be met exactly in each test administration.

5.2 RESULTS OF SIMULATED TEST ADMINISTRATIONS: ENGLISH

This section presents the simulation results for the English online tests, which is the test taken by almost all students (98.93%). Simulations were evaluated for all content areas using 1,000 simulated cases per grade.

5.2.1 Summary of Blueprint Match

The simulation results showed no blueprint violations at all content levels for all three grades.

5.2.2 Item Exposure

The simulator output also reports the degree to which the constraints set forth in the blueprints may yield greater exposure of items to students. This is reported by examining the percentage of test administrations in which an item appears. For instance, in a fixed paper-pencil form, 100% of the items appear on 100% of the test administrations because every test taker takes the same form. In an adaptive test or a LOFT test with a sufficiently large item pool, we would expect that most of the items would appear on a relatively small percentage of the test administrations only.

When this condition holds, it suggests that test administrations between students are more or less unique. Therefore, we calculated the item exposure rate for each item across by dividing the total

number of test administrations in which an item appears by the total number of tests administered. Then we report the distribution of the item exposure rate (r) in six bins. The bins are r=0% (unused), 0%<r<=1%, 1%<r<=5%, 5%<r<=20%, 20%<r<=40%, 40%<r<=60%, 60%<r<=80%, and 80%<r<=100%. If global item exposure is minimal, we would expect the largest proportion of items to appear in the bins of 0%<r<=20%, an indication that most of the items appear on a very small percentage of the test forms.

Table 18 below presents the percentage of items that falls into each exposure bin for all grades. Most test items (95.35% or more) are administered in 5%–60% of the test administrations. No item has an exposure rate less than 5% and the minimum exposure rate is 5.42% in grade 11. A few items had an exposure rate higher than 60% because of the limitation of the current pool for some content categories.

Table 18. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All English Online Simulation Sessions

Grade	Total Items	[0,0]%	[0,1]%	[1,5]%	[5,20]%	[20,40] %	[40,60] %	[60,80] %	[80,100] %
5	86	0	0	0	59.30	27.91	8.14	3.49	1.16
8	70	0	0	0	42.86	41.43	15.71	0	0
11	86	0	0	0	56.98	38.37	1.16	2.33	1.16

5.3 RESULTS OF SIMULATED TEST ADMINISTRATIONS: SPANISH

This section presents the simulation results for the Spanish tests. The Spanish item pool consists of a subset of AIRCore items that has a Spanish translation available. Table 19 below presents the numbers of items available for the Spanish tests.

Table 19. Spring 2019 Spanish Operational Item Pool

Grade	Item Type	Total Number of Items
5	Cluster	10
	Stand-Alone	23
8	Cluster	9
	Stand-Alone	22
11	Cluster	10
	Stand-Alone	20
Total		**94**

Simulations were evaluated for all content areas using 1,000 simulated cases per grade.

5.3.1 Summary of Blueprint Match

There was no blueprint violation at the discipline level for all three grades. However, because of the limitation of the Spanish item pool, blueprint violations were observed in grade 8 for content levels below the discipline level.

In grade 8, students always received two item clusters from the same DCI ("Earth's Systems"), but the blueprint required no more than one item cluster from each DCI. The reason is that there was no item cluster available in other two DCIs ("Earth's Place in the Universe" and "Earth and Human Activity").

5.3.2 Item Exposure

Table 20 below presents the percentage of items that falls into each exposure bin for all grades. All test items were administered in more than 20% of the test administrations. Some items had an exposure rate of 100% because of the limited Spanish item pool. Only those items were available to satisfy the blueprint constraints.

Table 20. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spanish Simulation Sessions

Grade	Total Items	[0,0]%	[0,1]%	[1,5]%	[5,20]%	[20,40]%	[40,60]%	[60,80]%	[80,100]%
5	33	0	0	0	0	39.39	30.30	6.06	24.24
8	31	0	0	0	0	16.13	48.39	22.58	12.90
11	30	0	0	0	0	20	33.33	26.67	20

6. OPERATIONAL TEST ADMINISTRATION SUMMARY REPORT

This section presents the blueprint match reports and item exposure rates for the spring 2019 operational test administrations.

6.1 BLUEPRINT MATCH

Table 21 and Table 22 on the following pages present the percentages of the spring 2019 tests that violated the blueprint requirement in Rhode Island and Vermont, respectively. The English online tests in all grades met the blueprint specifications with a 100% match at all content levels, except for five grade 5 students out of 10,798 in Rhode Island, three grade 8 students out of 10,546 in Rhode Island, and two grade 5 students out of 6,069 in Vermont.

In Rhode Island, all eight students with blueprint violations received two items from the same performance expectation (PE), while the blueprint requires no more than one item from the same

PE. In addition, among the five students in grade 5, four received three stand-alone items from the "Earth's Systems" Disciplinary Core Idea (DCI), while the blueprint requires at most two stand-alone items from each DCI. One student received four items in total from the "Earth's Systems" DCI, while the blueprint requires at most three items from the same DCI. Similarly, the two students with blueprint violations in Vermont also received two items from the same PE and three stand-alone items from the "Earth's Systems" DCI. One of these students also received four items in total from the "Earth's Systems" DCI.

These types of violations did not happen during simulations. The reason why they happened in the operational test administrations is that these students have seen the items before the test administration that were supposed to meet blueprint requirements. The item selection algorithm automatically filtered out the items they had taken so the students would not see the same items twice. There are two possible scenarios for this. First, the students saw the items in a previous attempt in spring 2019 and reset the test. Second, these students took the science test at the same grade in both spring 2018 and spring 2019. Therefore, the pool became shallower for these students. At the end of the test, the algorithm had no choice left to satisfy the blueprint requirement but to pick one that caused a blueprint violation. Note that these violations were all below the discipline level.

For the Spanish tests, the blueprint violations and percentages were similar to the findings from the simulations due to the limited Spanish pool.

Table 21. Rhode Island Spring 2019 Blueprint Match for Test Delivered, Science

Grade	Content Level	Min Items	Max Items	% of Cases Violating Blueprint			
				+1	+2	-1	-2
English							
5	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	0.009	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	0.037	–	–	–
	PE	0	1	0.046	–	–	–
8	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	0.028	–	–	–
11	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–

Grade	Content Level	Min Items	Max Items	% of Cases Violating Blueprint			
				+1	+2	-1	-2
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
Spanish							
5	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
8	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	100	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
11	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–

Table 22. Vermont Spring 2019 Blueprint Match for Test Delivered, Science

Grade	Content Level	Min Items	Max Items	% of Cases Violating Blueprint			
				+1	+2	-1	-2
English							
5	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	0.033	–	–	–
8	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
11	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
Spanish							
5	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
8	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	100	–	–	–

Grade	Content Level	Min Items	Max Items	% of Cases Violating Blueprint			
				+1	+2	-1	-2
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–
11	Discipline	6	6	–	–	–	–
	Discipline – Cluster	2	2	–	–	–	–
	Discipline – Stand-Alone	4	4	–	–	–	–
	DCI	0	3	–	–	–	–
	DCI – Cluster	0	1	–	–	–	–
	DCI – Stand-Alone	0	2	–	–	–	–
	PE	0	1	–	–	–	–

6.2 ITEM EXPOSURE

Table 23 and Table 24 below present the item exposure rates of the spring 2019 test administration for Rhode Island and Vermont, respectively. The exposure rates were very similar to the simulation results described in Section 5.2.2, Item Exposure, for the English test administrations. The item exposure rate for field-test items ranged from 10% to 22% for all three grades. For the Spanish tests, more items had high exposure rates compared to the English tests because of a smaller item pool. Also, the operational exposure rates were slightly different from the simulation results because of small population sizes in all three grades. In spring 2019, less than 200 students took the Spanish test in each grade in Rhode Island, and less than five students took the Spanish test in each grade in Vermont.

Table 23. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spring 2019 Test Administrations in Rhode Island

Grade	Total Items	[0,0]%	[0,1]%	[1,5]%	[5,20]%	[20,40]%	[40,60]%	[60,80]%	[80,100]%
English									
5	86	0	0	0	59.30	27.91	8.14	4.65	0
8	70	0	0	0	41.43	44.29	14.29	0	0
11	86	0	0	0	60.47	34.88	1.16	2.33	1.16
Spanish									
5	33	0	0	0	3.03	42.42	18.18	12.12	24.24
8	31	0	0	0	0	12.90	51.61	19.35	16.13
11	30	0	0	0	0	20.00	33.33	26.67	20.00

Table 24. Item Exposure Rates by Grade: Percentage of Items by Exposure Rate, Across All Spring 2019 Test Administrations in Vermont

Grade	Total Items	[0,0]%	[0,1]%	[1,5]%	[5,20]%	[20,40]%	[40,60]%	[60,80]%	[80,100]%
English									
5	86	0	0	0	59.30	27.91	8.14	4.65	0
8	70	0	0	0	37.14	50.00	12.86	0	0
11	86	0	0	0	60.47	36.05	0	2.33	1.16
Spanish									
5	33	0	0	0	0	12.12	30.30	12.12	27.27
8	31	0	0	0	0	0	38.71	0	38.71
11	30	0	0	0	0	30.00	0	30.00	30.00

7. REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014*). Standards for educational and psychological testing*.

Calisir, F., & Gurel, Z. (2003). Influence of text structure and prior knowledge of the learner on reading comprehension, browsing and perceived control. *Computers in Human Behavior, 19*(2), 135–145.

Fisher, D., Frey, N., & Lapp, D. (2012). *Text complexity: Raising rigor in reading*. Newark, DE: International Reading Association.

Freebody, P., & Anderson, R.C. (1983). Effects on Text Comprehension of Differing Proportions and Locations of Difficult Vocabulary. *Journal of Reading Behavior, 15*(3), 19–39.

Gillioz, C., Gygax, P., & Tapiero, I. (2012). Individual differences and emotional inferences during reading comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(4), 239–250.

Kucer, S.B. (2010). Going beyond the author: What retellings tell us about comprehending narrative and expository texts. *Literacy, 45*(2), 62–69.

Long, D.L., & De Ley, L. (2000). Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language, 42*(4), 545–570.

McConaughy, S. (1985). Good and Poor Readers' Comprehension of Story Structure across Different Input and Output Modalities. *Reading Research Quarterly, 20*(2), 219–232. doi:10.2307/747757.

Rapp, D.N., & Mensink, M.C. (2011). Focusing effects from online and offline reading tasks. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 141–164). Charlotte, NC, US: IAP Information Age Publishing.

Rich, S.S., & Taylor, H.A. (2000). Not all narrative shifts function equally. *Memory & Cognition, 28*(7), 1257–1266.

Riding, R.J., & Taylor, E.M. (1976). Imagery performance and prose comprehension in seven-year-old children. *Educational Studies, 2*(1), 21–2.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience, 25*(5), 762–776.

Sadoski, M., Goetz, E.T., & Fritz, J.B. (1993). A causal model of sentence recall: Effects of familiarity, concreteness, comprehensibility, and interestingness. *Journal of Reading Behavior, 25*(1), 5–16.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal*, *95*(1), 26–43.

Sparks, J.R., & Rapp, D.N. (2011). Readers reliance on source credibility in the service of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 230–247.

Thompson, S.J., Johnstone, C.J., & Thurlow, M.L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 15, 2012, from http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis44.html.

Webb, N.L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

Appendix A

Item Writer Training Materials

Exhibit A-1. LABS Guidelines

LABS Guidelines

1. STEREOTYPING

Testing materials should not present persons stereotyped according to the following characteristics:

- Age
- Disability
- Gender
- Race/Ethnicity
- Sexual orientation

2. SENSITIVE OR CONTROVERSIAL SUBJECTS

Controversial or potentially distressing subjects should be avoided or treated sensitively. For example, a passage discussing the historical importance of a battle is acceptable whereas a graphic description of a battle would not be. Controversial subjects include:

- Death and Disease
- Gambling*
- Politics (Current)
- Race relations
- Religion
- Sexuality
- Superstition
- War

**References to gambling should be avoided in mathematics items related to probability.*

3. ADVICE

Testing materials should not advocate specific lifestyles or behaviors except in the most general or universally agreed-upon ways. For example, a recipe for a healthful fruit snack is acceptable but a passage recommending a specific diet is not. The following categories of advice should be avoided:

- Religion
- Sexual preference
- Exercise
- Diet

4. DANGEROUS ACTIVITY

Tests should not contain content that portrays people engaged in or explains how to engage in dangerous activities. Examples of dangerous activities include:

- Deep-sea diving
- Stunts
- Parachuting
- Smoking
- Drinking

5. POPULATION DIVERSITY AND ETHNOCENTRISM

Testing materials should:

- Reflect the diversity of the testing population
- Use stimulus materials (such as works of literature) produced by members of minority communities
- Use personal names from different ethnic origin communities
- Use pictures of people from different ethnic origin communities
- Avoid *ethnocentrism*, or the attitude that all people should share a particular group's language, beliefs, culture, or religion

6. DIFFERENTIAL FAMILIARITY AND ELITISM

Specialized concepts and terminology extraneous to the core content of test questions should be avoided. This caveat applies to terminology from the fields of:

- Construction
- Finance
- Sports
- Law
- Machinery
- Military topics
- Politics
- Science
- Technology
- Agriculture

7. LANGUAGE USE

Language should be as inclusive as possible.

- Avoid masculine-coded words like mankind, manmade, and the generic "he"
- Use equal pairs such as husband and wife rather than man and wife

8. LANGUAGE ACCESSIBILITY

The grammar and vocabulary should be clear, concise, and appropriate for the intended grade level. The following should be avoided or used with care:

- Passive constructions
- Idioms
- Multiple subordinate clauses
- Pronouns with unclear antecedents
- Multiple-meaning words
- Non-standard grammar
- Dialect
- Jargon

9. ILLUSTRATIONS AND GRAPHICS

Illustrations and graphics should embody all of the previously referenced LABS Guidelines.

Exhibit A-2. LABS Checklist

LABS–Checklist

STEREOTYPING CONSIDERATIONS

- ☐ Does the material negatively represent, or stereotype people based on gender or sexual preference?
- ☐ Does the material portray one or more people with disabilities in a negative or stereotypical manner?
- ☐ Does the material portray one or more religious groups as aggressive or violent?
- ☐ Does the material romanticize or demean people based on socioeconomic status?
- ☐ Does the material portray one or more ethnic groups or cultures participating in certain stereotypical activities or occupations?
- ☐ Does the material portray one or more age groups in a negative or stereotypical manner?

SENSITIVE/CONTROVERSIAL MATERIAL CONSIDERATIONS

- ☐ Does the material require a student to take a position that challenges authority?
- ☐ Does the material present war or violence in an overly graphic manner?
- ☐ Does the material present sensitive or highly controversial subjects, such as death, war, abortion, euthanasia, or natural disasters, except where they are needed to meet State Content Standards?
- ☐ Does the material require test takers to disclose values that they would rather hold confidential?
- ☐ Does the material present sexual innuendoes?
- ☐ Does the material trivialize significant or tragic human experiences?
- ☐ Does the material require the parent, teacher, or test taker to support a position that is contrary to their religious beliefs?

ADVICE CONSIDERATIONS

- Does the material contain advice pertaining to health and well-being about which there is not a universal agreement?

POPULATION DIVERSITY

- Is the material written by members of diverse groups?
- Does the material reflect the experiences of diverse groups?
- Does the material portray people in positive nontraditional roles?
- Does test material represent the racial and ethnic composition of the testing population?
- Does the material reflect ethnocentrism?
- Does the material refer to population subgroups accurately?
- Does test material reflect diversity through the use of names, cultural references, pictures, and roles?

DIFFERENTIAL FAMILIARITY/ELITISM

- Does the material contain phrases, concepts, and beliefs that are irrelevant to testing domain and are likely to be more familiar to specific groups that others?
- Does the material require knowledge of individuals, events, or groups that is not familiar to all groups of students?
- Does the material suggest that affluence is related to merit or intelligence?
- Does the material suggest that poverty is related to increased negative behaviors in society?
- Does the material use language, content, or context that is offensive to people of a particular economic status?
- Does success with the material assume that the test taker has experience with a certain type of family structure?
- Does the material favor one socioeconomic group over another?
- Does the material assume values not shared by all test takers?

LINGUISTIC FEATURES/LANGUAGE ACCESSIBILITY/GRAPHICS

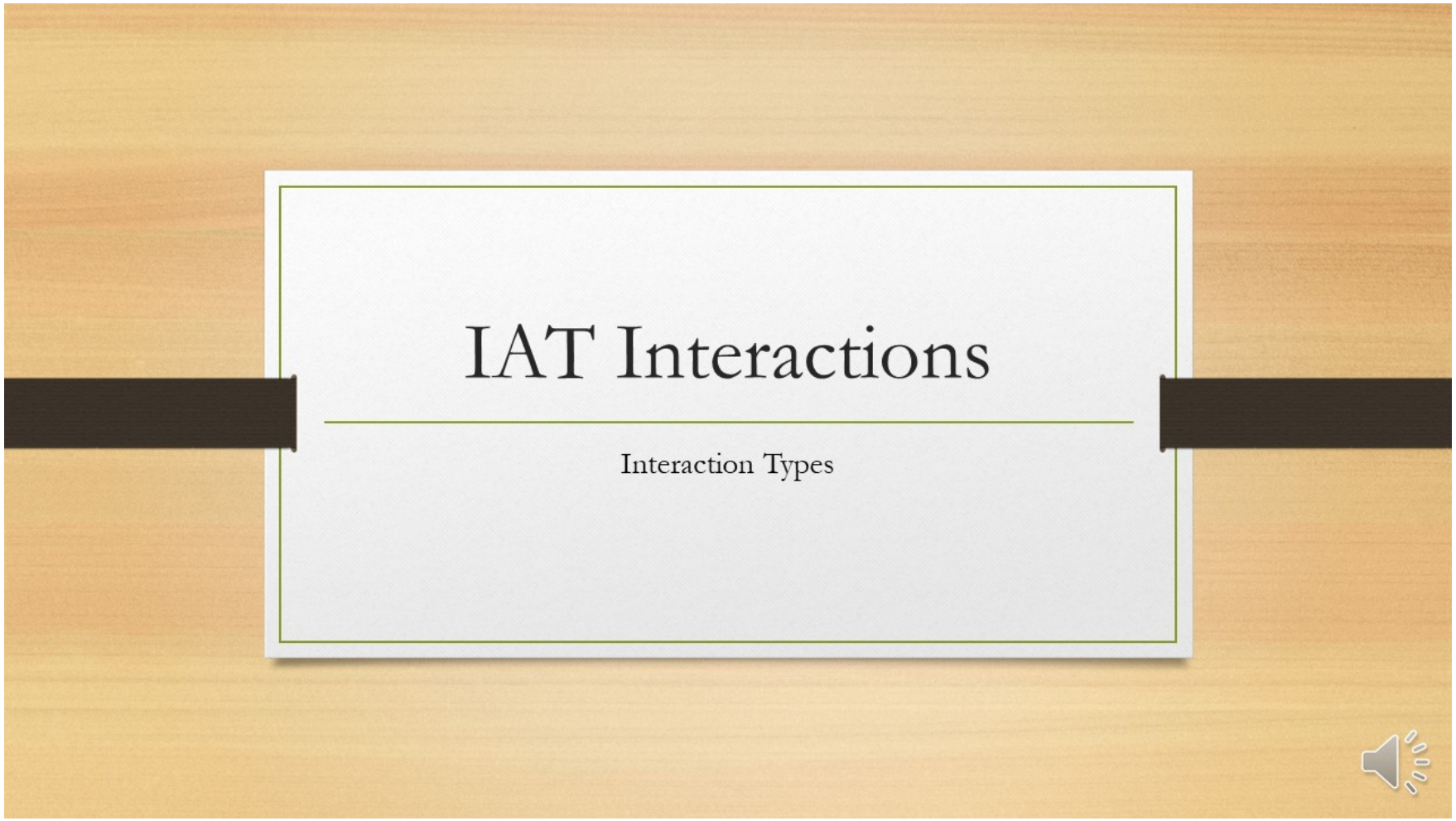
- Is grammar and vocabulary used in the items clear, concise, and appropriate for the intended grade level?
- Are passages at a difficulty level that is appropriate for the intended grade level?

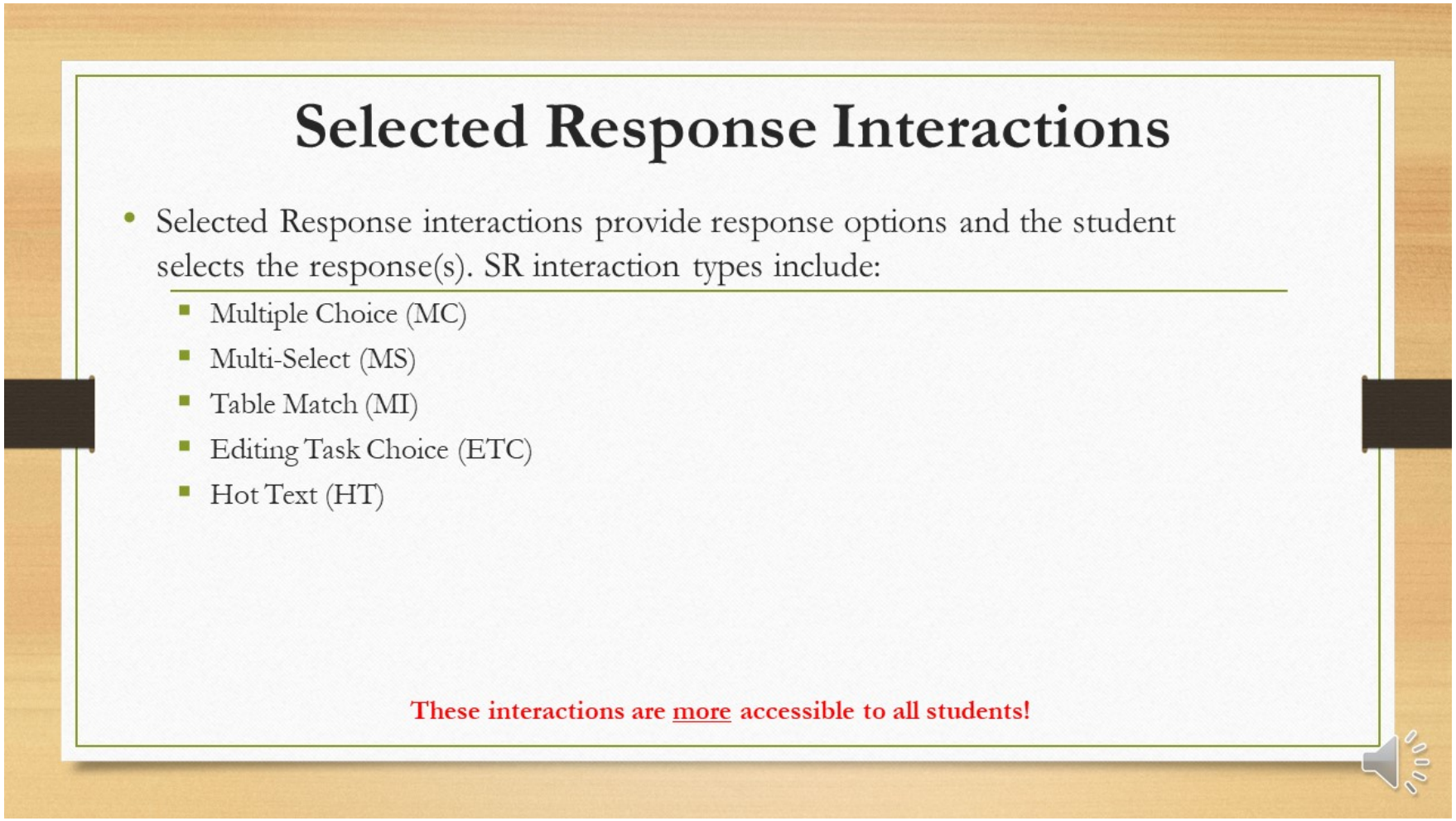
- Do the illustrations and graphics embody all of the previously referenced LABS Guidelines?

OTHER QUESTIONS TO CONSIDER

- Does the material favor one age group over others except in a context where experience or maturation is relevant?
- Does the material use language, content, or context that is not accessible to one or more of the age groups tested?
- Does the material contain language or content that contradicts values held by a certain culture?
- Does the material favor one racial or ethnic group over others?
- Does the material degrade people based on physical appearance or any physical, cognitive, or emotional challenge?
- Does the material focus only on a person's disability rather than portraying the whole person?
- Does the material favor one religion and/or demean others?

Exhibit A-3. An Overview of Interaction Types


IAT Interactions
Interaction Types


Selected Response Interactions
Selected Response interactions provide response options and the student selects the response(s). SR interaction types include:
Multiple Choice (MC)
Multi-Select (MS)
Table Match (MI)
Editing Task Choice (ETC)
Hot Text (HT)
These interactions are more accessible to all students!

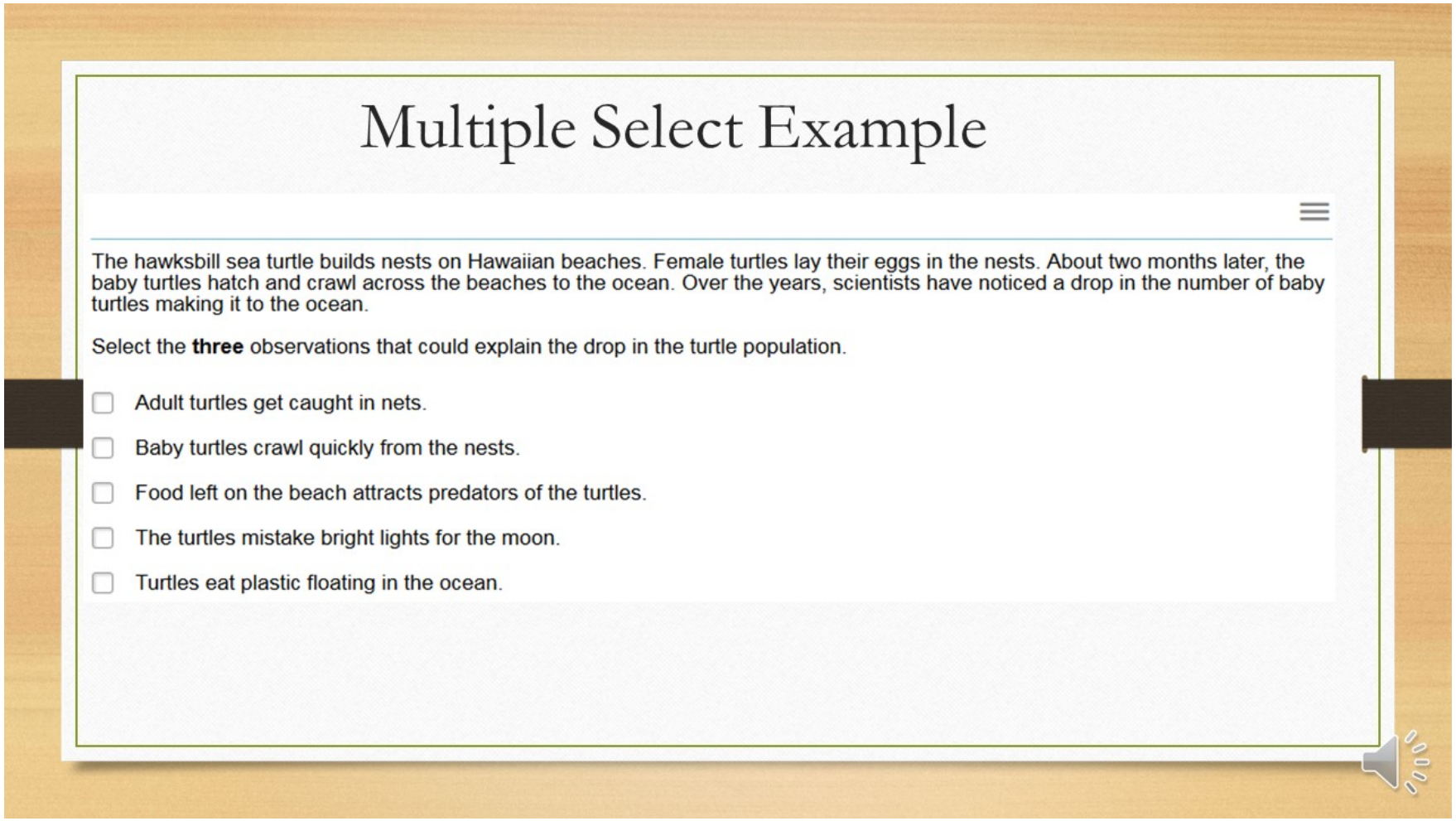
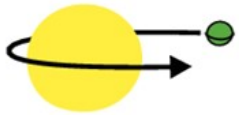

Multiple Select Example
The hawksbill sea turtle builds nests on Hawaiian beaches. Female turtles lay their eggs in the nests. About two months later, the baby turtles hatch and crawl across the beaches to the ocean. Over the years, scientists have noticed a drop in the number of baby turtles making it to the ocean.
Select the three observations that could explain the drop in the turtle population.
Adult turtles get caught in nets.
Baby turtles crawl quickly from the nests.
Food left on the beach attracts predators of the turtles.
The turtles mistake bright lights for the moon.
Turtles eat plastic floating in the ocean.

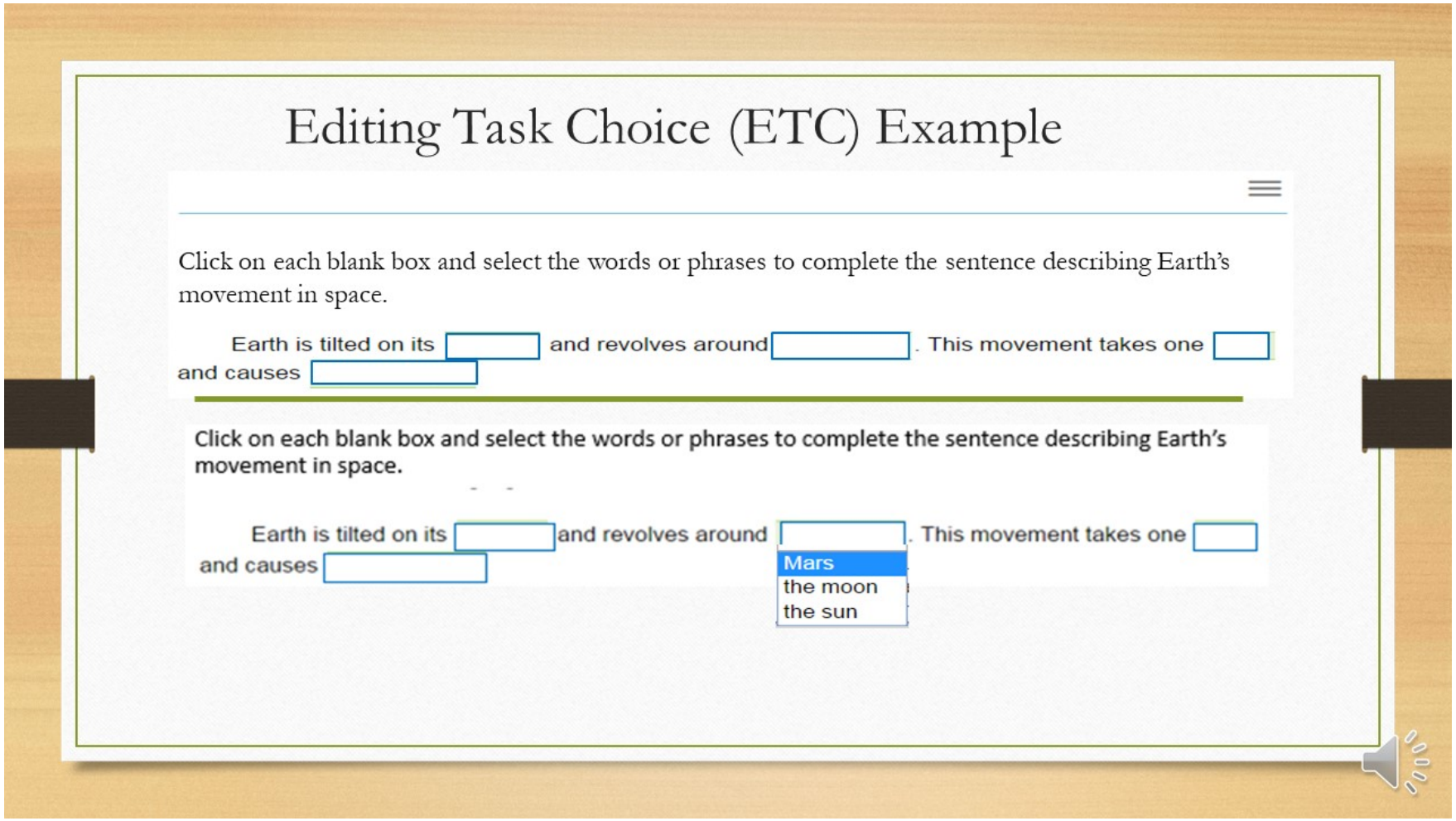
Table Match (MI) Example

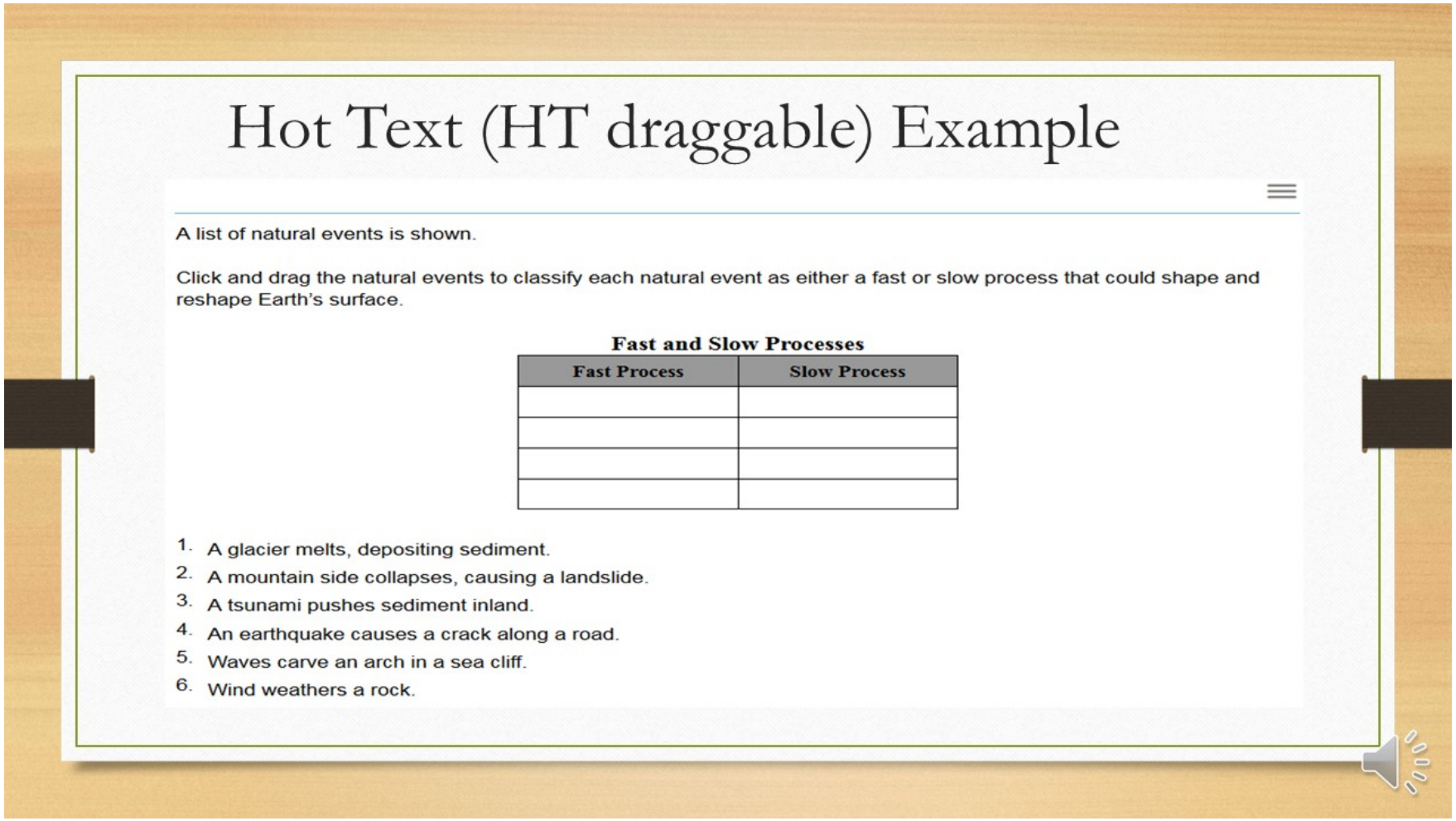
Students use a large yellow ball and a small green ball to model the sun and Earth. They use the balls to explain the cause of day and night, to model the length of a year, and to explain the cause of the seasons.

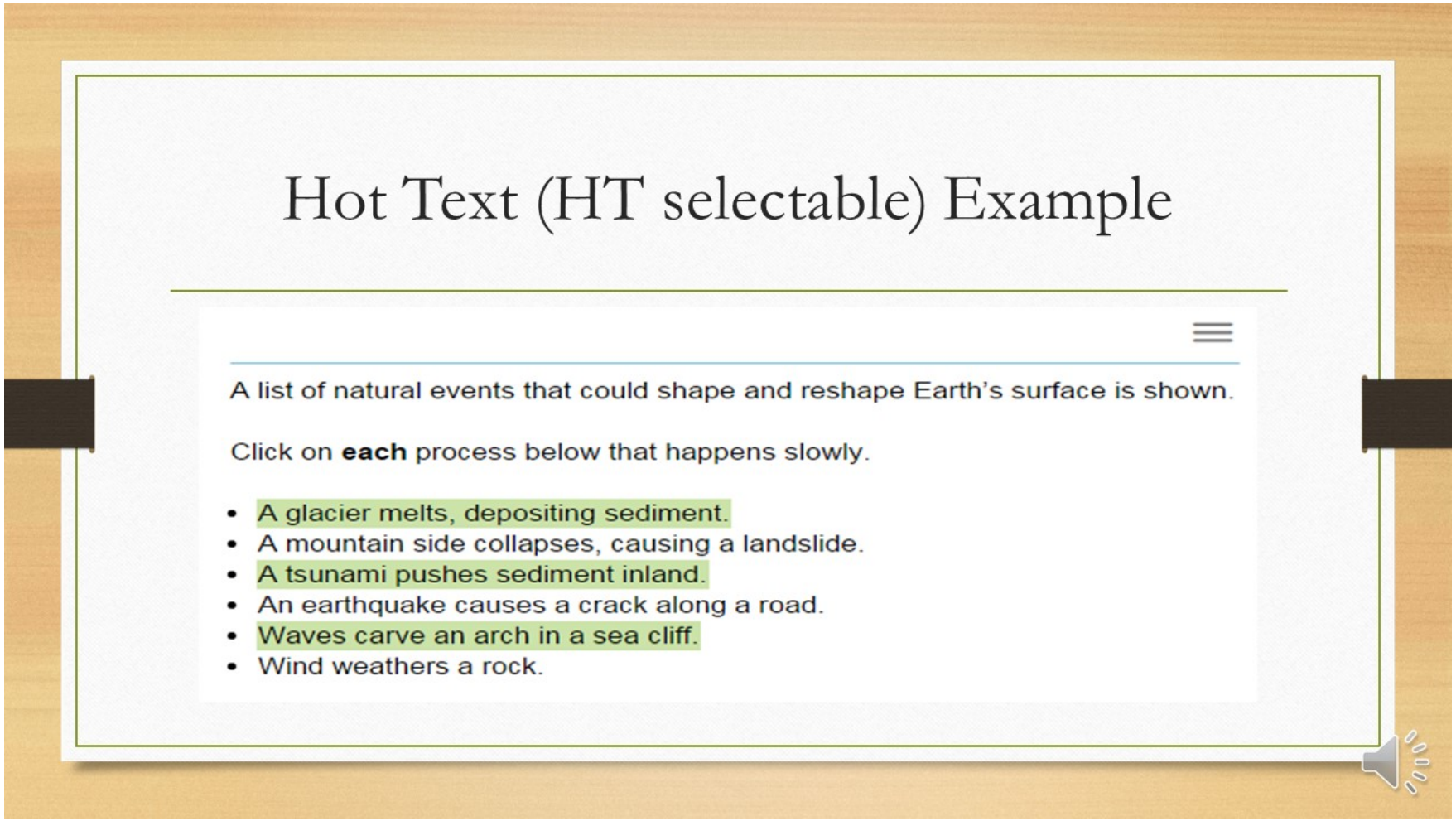
Select **each** box to identify which movements of the balls are needed to explain each phenomenon.

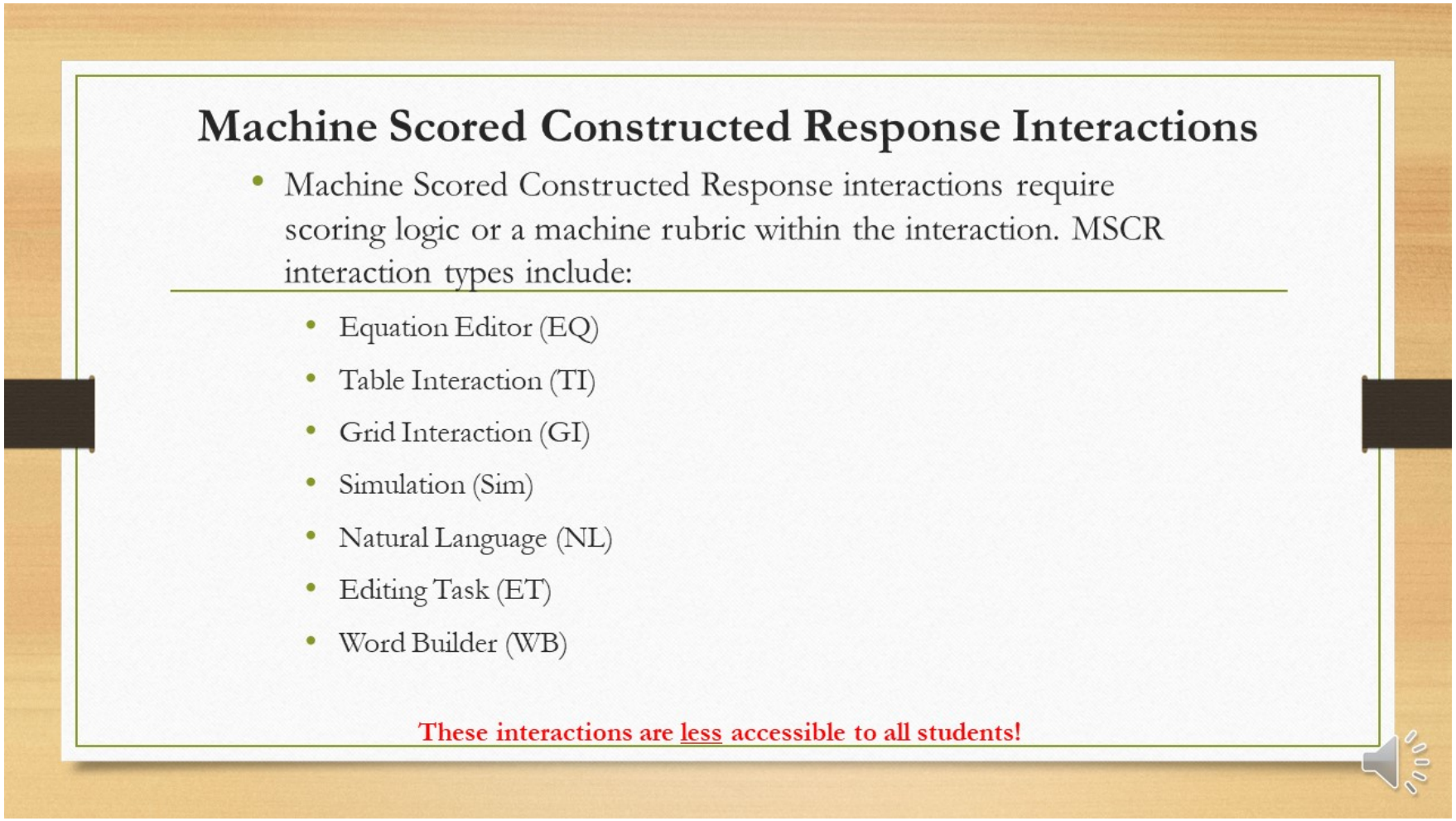
- You can select more than one box for each statement.

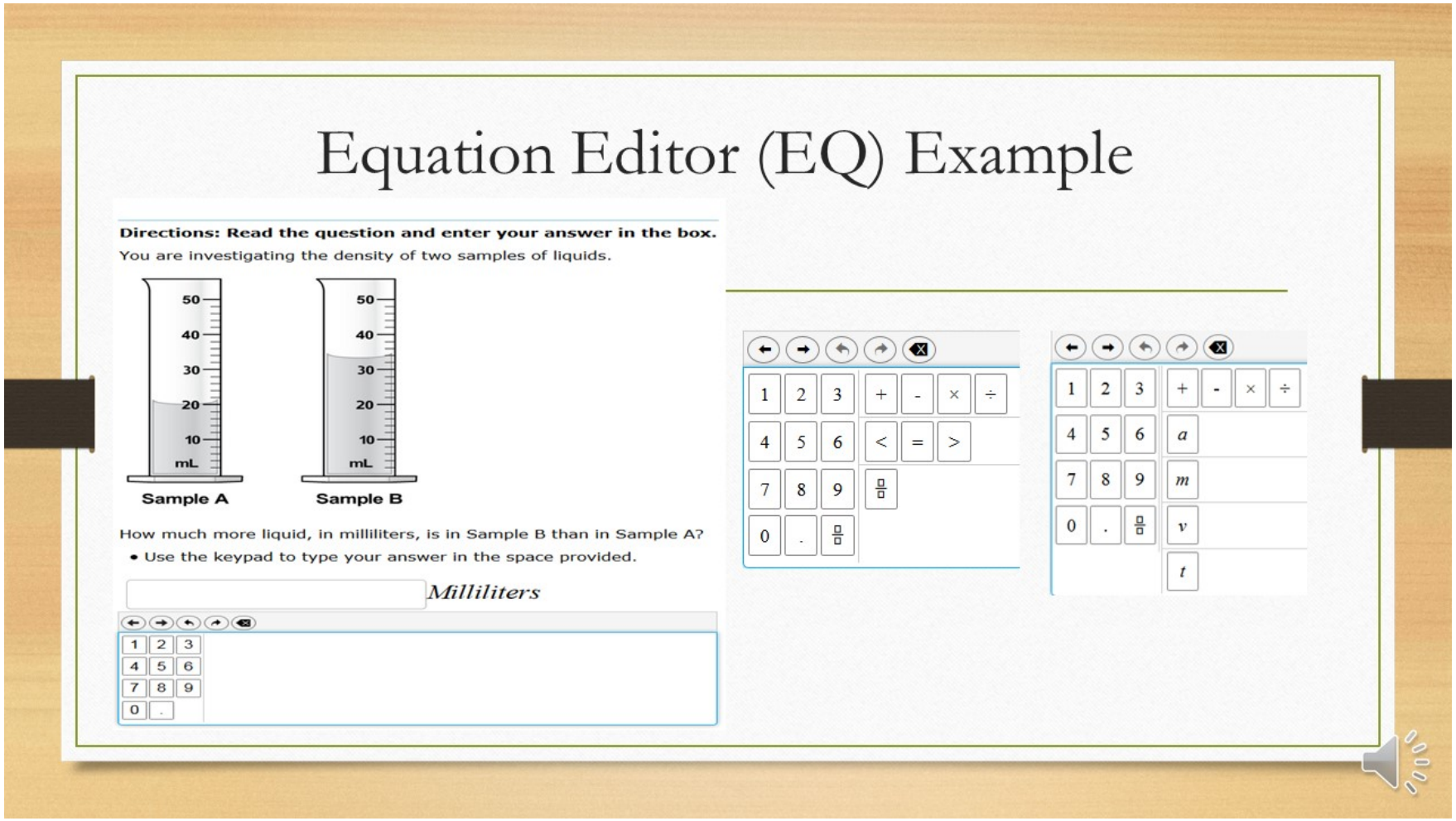
	 **Large yellow ball is stationary, while small green ball spins.**	 **Large yellow ball is stationary, while small green ball is tilted.**	 **Large yellow ball is stationary, while small green ball moves around it.**
The cause of day and night	☐	☐	☐
The length of a year	☐	☐	☐
The cause of the seasons	☐	☐	☐


Editing Task Choice (ETC) Example
Click on each blank box and select the words or phrases to complete the sentence describing Earth's movement in space.
Earth is tilted on its and revolves around . This movement takes one and causes
Click on each blank box and select the words or phrases to complete the sentence describing Earth's movement in space.
Earth is tilted on its and revolves around . This movement takes one and causes
Mars
the moon
the sun


Hot Text (HT draggable) Example
A list of natural events is shown.
Click and drag the natural events to classify each natural event as either a fast or slow process that could shape and reshape Earth's surface.
Fast and Slow Processes
Fast Process
Slow Process
1. A glacier melts, depositing sediment.
2. A mountain side collapses, causing a landslide.
3. A tsunami pushes sediment inland.
4. An earthquake causes a crack along a road.
5. Waves carve an arch in a sea cliff.
6. Wind weathers a rock.


Hot Text (HT selectable) Example
A list of natural events that could shape and reshape Earth's surface is shown.
Click on **each** process below that happens slowly.
A glacier melts, depositing sediment.
A mountain side collapses, causing a landslide.
A tsunami pushes sediment inland.
An earthquake causes a crack along a road.
Waves carve an arch in a sea cliff.
Wind weathers a rock.


Machine Scored Constructed Response Interactions
Machine Scored Constructed Response interactions require scoring logic or a machine rubric within the interaction. MSCR interaction types include:
Equation Editor (EQ)
Table Interaction (TI)
Grid Interaction (GI)
Simulation (Sim)
Natural Language (NL)
Editing Task (ET)
Word Builder (WB)
These interactions are less accessible to all students!


Equation Editor (EQ) Example
Directions: Read the question and enter your answer in the box.
You are investigating the density of two samples of liquids.
50
40
30
20
10
mL
Sample A
Sample B
How much more liquid, in milliliters, is in Sample B than in Sample A?
• Use the keypad to type your answer in the space provided.
Milliliters
1 2 3 + - × ÷
4 5 6 < = >
7 8 9
0 .
a m v t

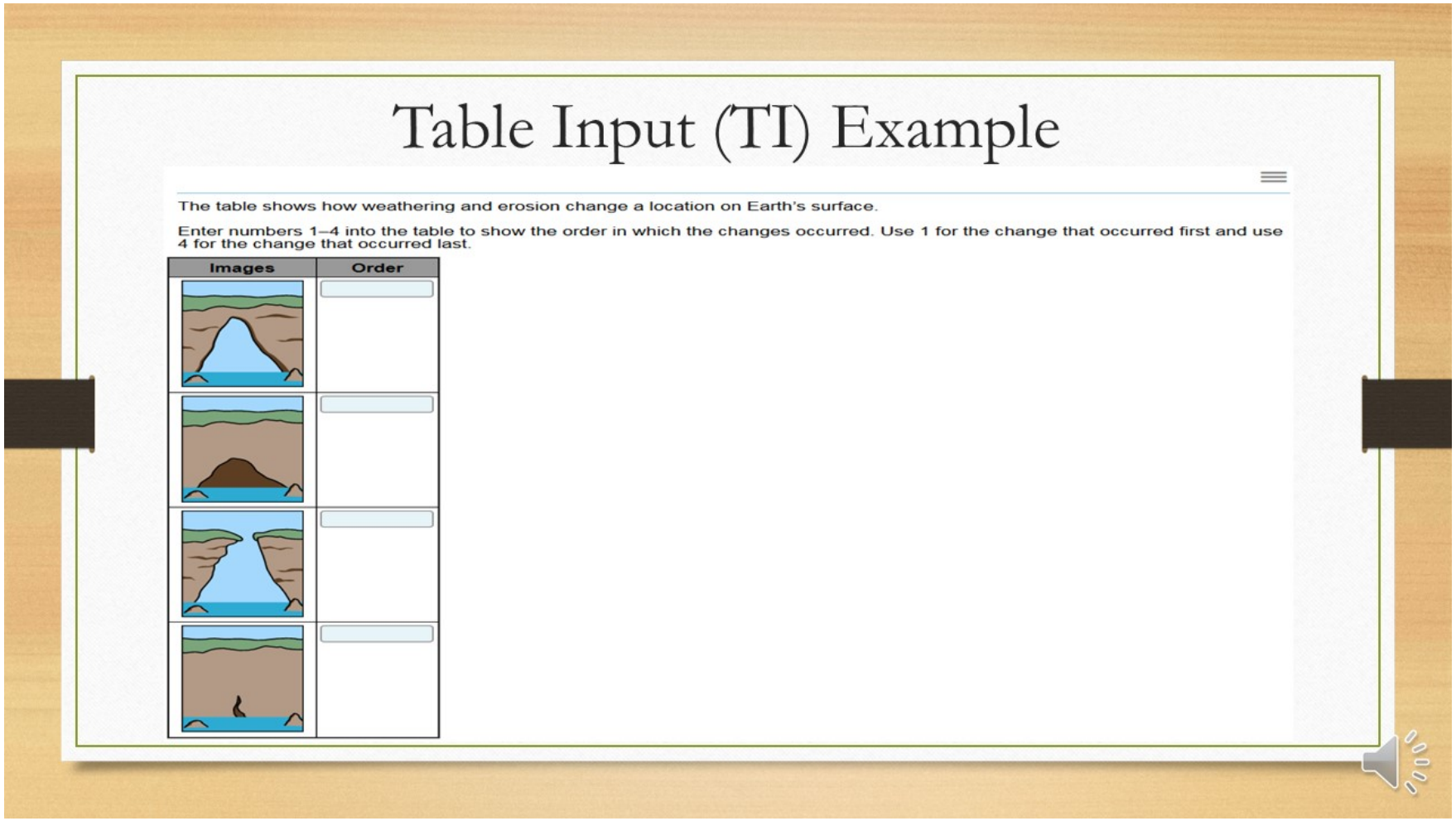
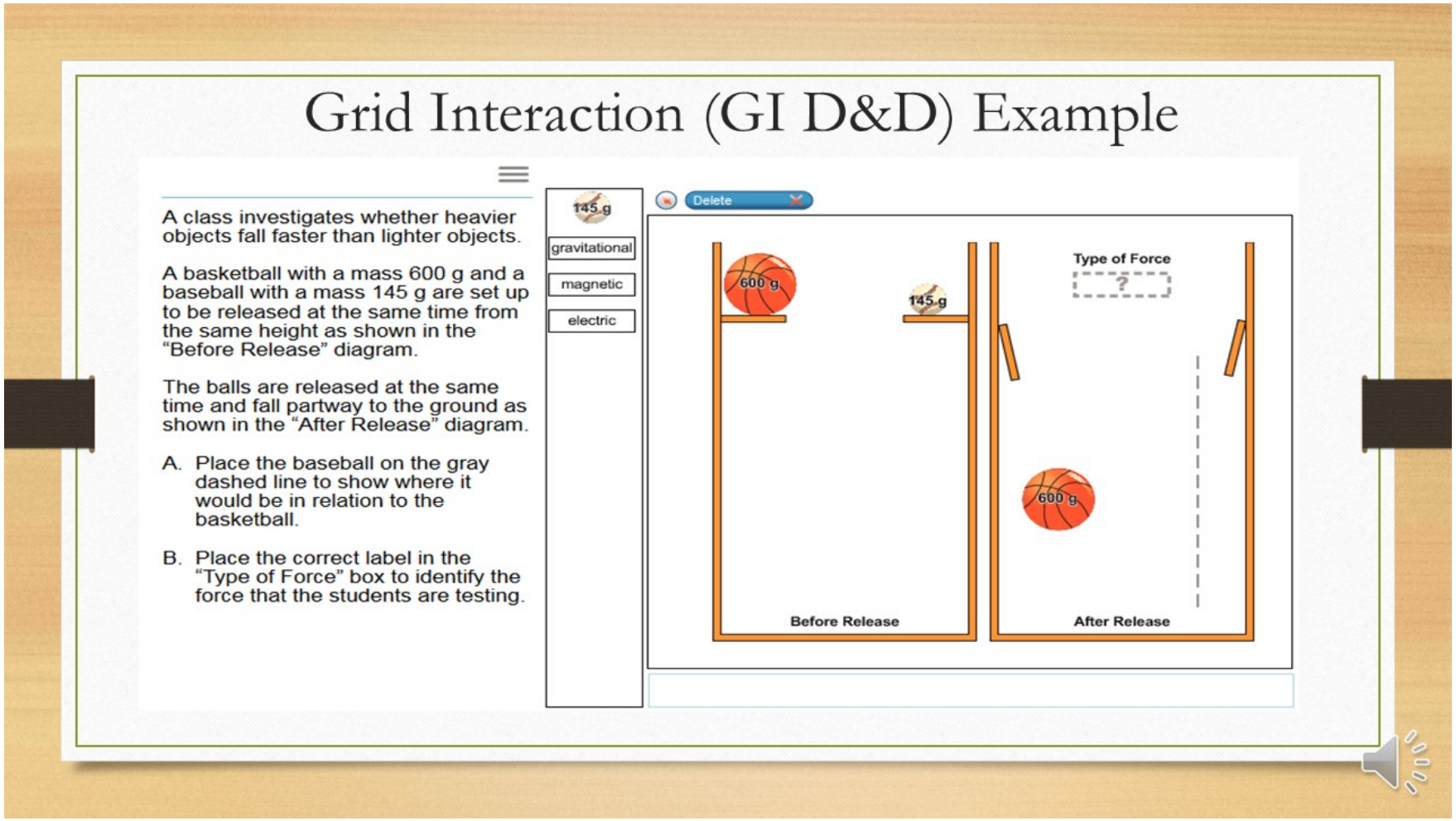
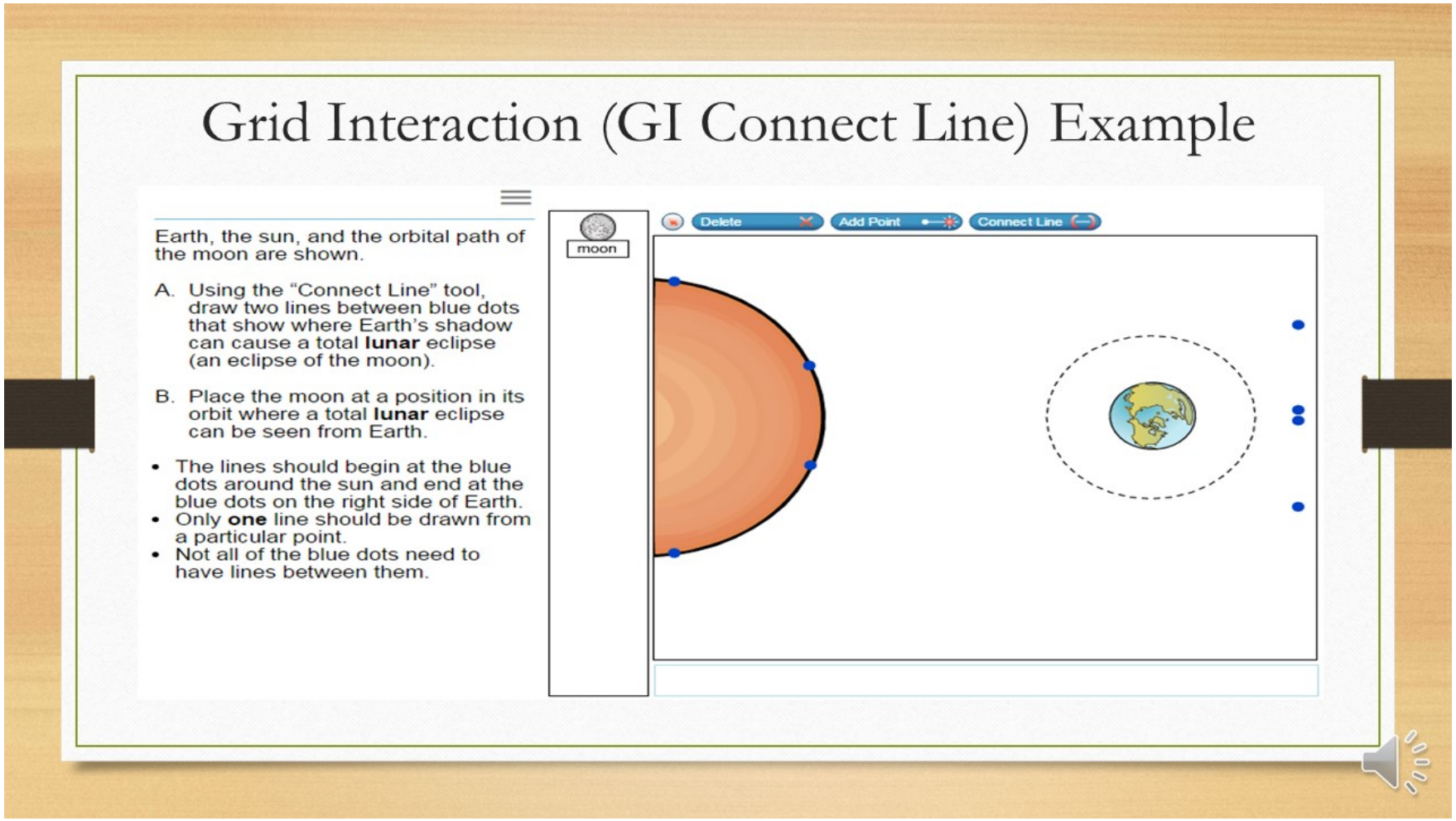


Table Input (TI) Example
The table shows how weathering and erosion change a location on Earth's surface.
Enter numbers 1–4 into the table to show the order in which the changes occurred. Use 1 for the change that occurred first and use 4 for the change that occurred last.
Images
Order


Grid Interaction (GI D&D) Example
A class investigates whether heavier objects fall faster than lighter objects.
A basketball with a mass 600 g and a baseball with a mass 145 g are set up to be released at the same time from the same height as shown in the "Before Release" diagram.
The balls are released at the same time and fall partway to the ground as shown in the "After Release" diagram.
A. Place the baseball on the gray dashed line to show where it would be in relation to the basketball.
B. Place the correct label in the "Type of Force" box to identify the force that the students are testing.
145 g
gravitational
magnetic
electric
Delete
600 g
145 g
Type of Force
?
600 g
Before Release
After Release


Grid Interaction (GI Connect Line) Example
Earth, the sun, and the orbital path of the moon are shown.
A. Using the "Connect Line" tool, draw two lines between blue dots that show where Earth's shadow can cause a total lunar eclipse (an eclipse of the moon).
B. Place the moon at a position in its orbit where a total lunar eclipse can be seen from Earth.
• The lines should begin at the blue dots around the sun and end at the blue dots on the right side of Earth.
• Only one line should be drawn from a particular point.
• Not all of the blue dots need to have lines between them.
moon
Delete
Add Point
Connect Line

Grid Interaction (GI Click up/Add Arrow) Example

Students investigate how the direction of light affects plant growth. They grow three plants in individual cardboard boxes using light from lamps. The picture shows the growth of Plant 1 with light coming from directly above the plant.

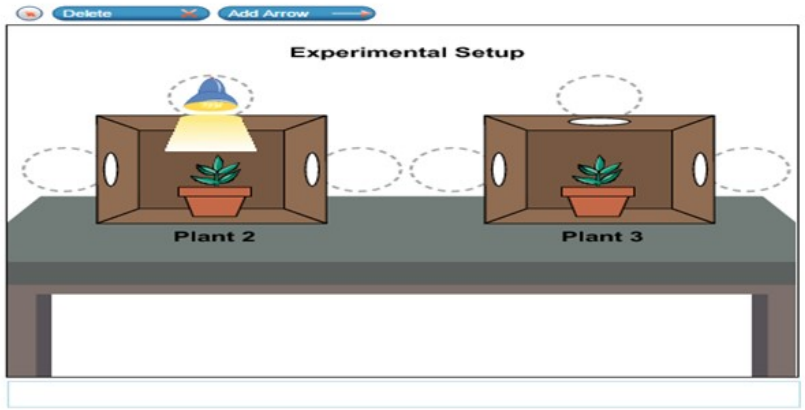
Plant 1

The students want to set up Plant 2 and Plant 3 with a light source to complete the investigation.

A. Click on one blank circle for Plant 2 and one blank circle for Plant 3 to show the direction of the light source for each plant to complete the investigation.

B. Use the Add Arrow button to draw an arrow showing the predicted growth of Plant 2 and Plant 3 based on the light source on each plant.

- Draw only **one** arrow for Plant 2.
- Draw only **one** arrow for Plant 3.
- There may be more than one correct answer.



Simulation (SIM Nonscoring) Example

12

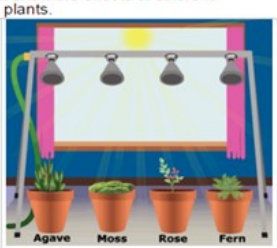
Students are studying different kinds of plants and the conditions that they grow in. They have planted four kinds of young plants.

Design and run an experiment that will show the effects of different amounts of sunlight and water on the plants.

Amount of Water: Little

Amount of Light: Direct Sun

Start



Amount of Water	Amount of Light	Agave	Moss	Rose	Fern

13

Which of the plants would grow *best* in a desert environment?

- Agave
- Ⓑ Fern
- Ⓒ Moss
- Ⓓ Rose

14

Which two kinds of plants could grow in the same environment based on the data from the experiment?

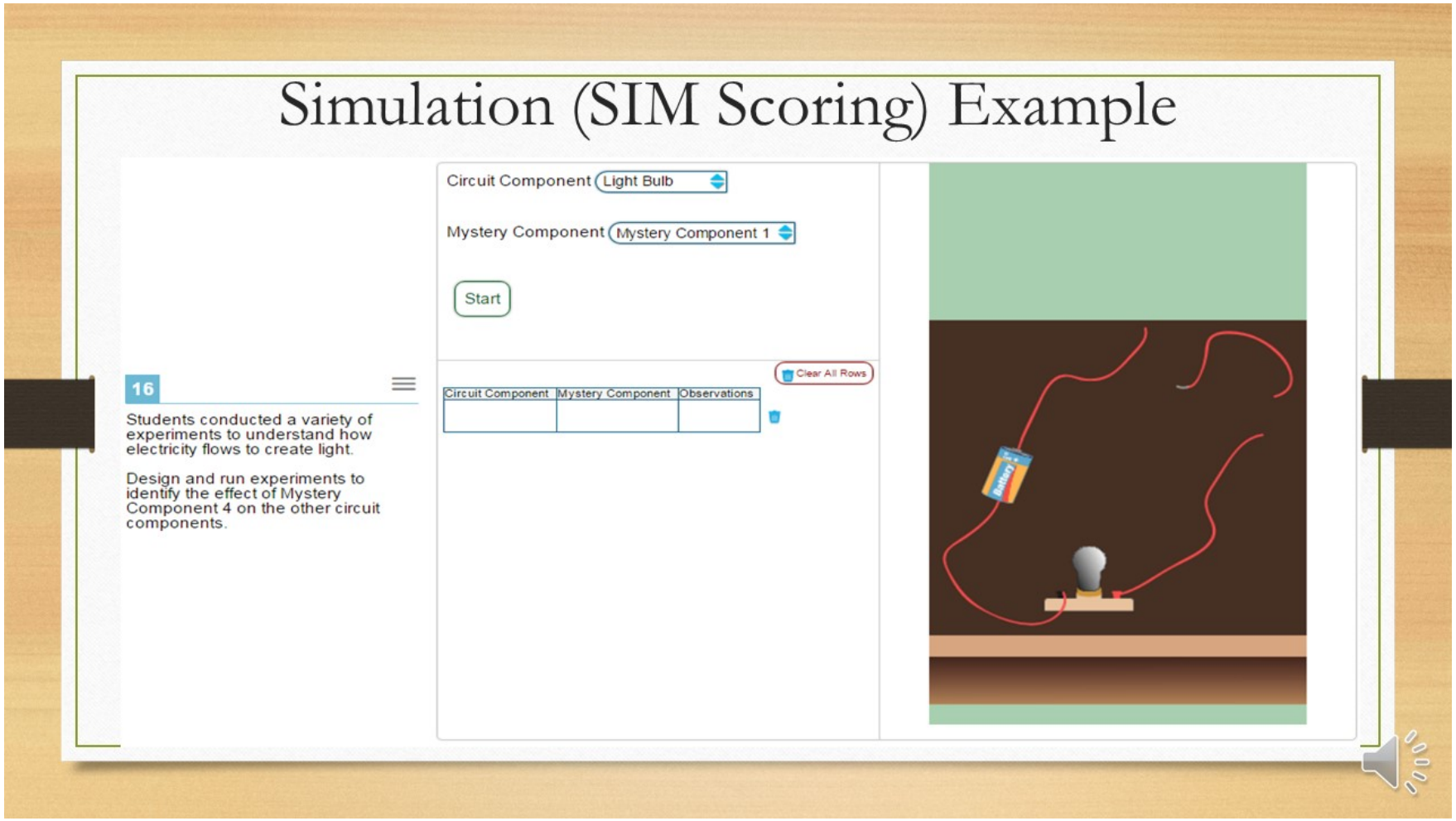
- Ⓐ Agave and fern
- Ⓑ Fern and moss
- Ⓒ Moss and rose
- Ⓓ Rose and agave

15

A student records some notes in a notebook during the experiment. Some of the notes are observations and some are inferences.

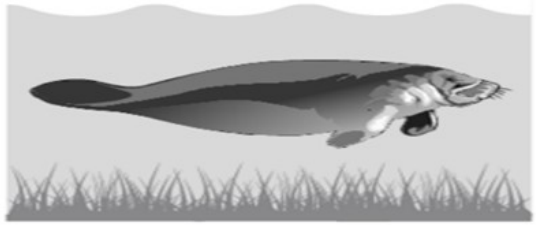
Select a box to identify whether each note is an observation or an inference.

	Observation	Inference
Agave is a desert plant.	☐	☐
No type of fern can survive in direct sun.	☐	☐
The rose did not grow taller in the shade.	☐	☐
The fern turned brown when there was little water.	☐	☐


Simulation (SIM Scoring) Example
16
Students conducted a variety of experiments to understand how electricity flows to create light.
Design and run experiments to identify the effect of Mystery Component 4 on the other circuit components.
Circuit Component
Light Bulb
Mystery Component
Mystery Component 1
Start
Clear All Rows
Circuit Component
Mystery Component
Observations
Battery

Natural Language (NL) Example

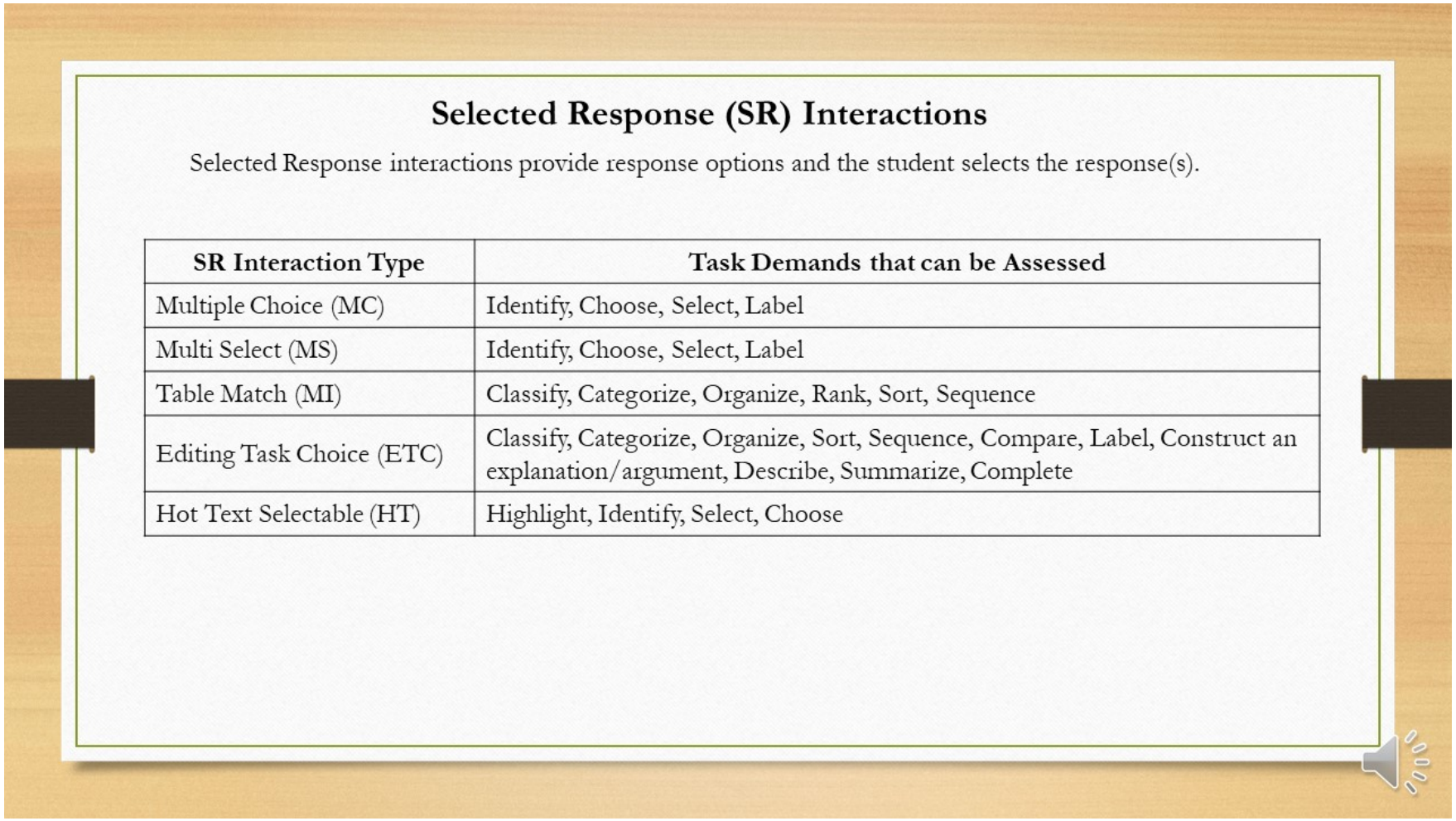
The picture shows a manatee.



A. State one observation that can be made about the manatee from this picture. Be sure to identify it as an observation.

B. State one inference that can be made about the manatee from this picture. Be sure to identify it as an inference.

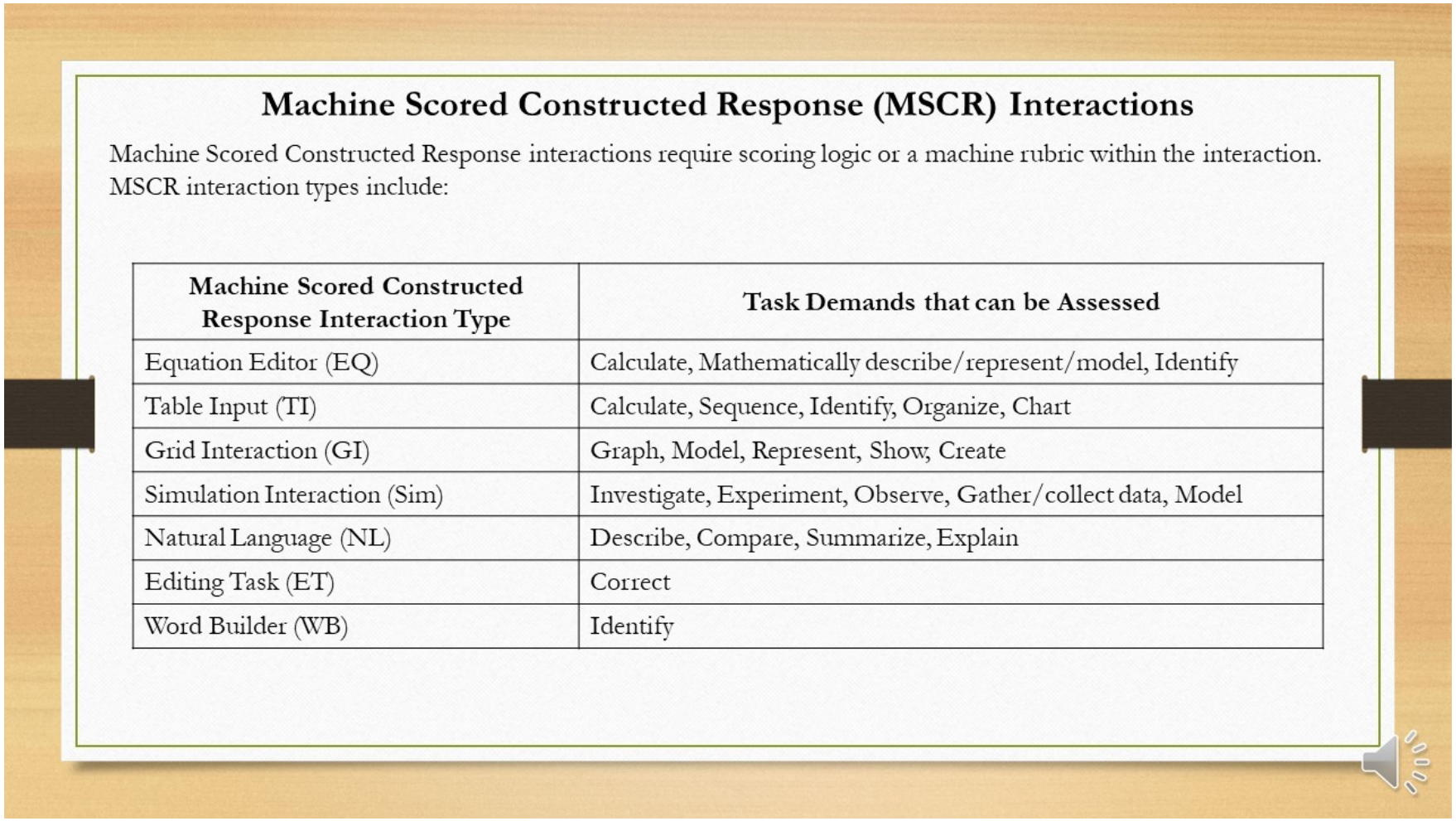
Type your answer in the space provided.



Selected Response (SR) Interactions

Selected Response interactions provide response options and the student selects the response(s).

SR Interaction Type	Task Demands that can be Assessed
Multiple Choice (MC)	Identify, Choose, Select, Label
Multi Select (MS)	Identify, Choose, Select, Label
Table Match (MI)	Classify, Categorize, Organize, Rank, Sort, Sequence
Editing Task Choice (ETC)	Classify, Categorize, Organize, Sort, Sequence, Compare, Label, Construct an explanation/argument, Describe, Summarize, Complete
Hot Text Selectable (HT)	Highlight, Identify, Select, Choose



Machine Scored Constructed Response (MSCR) Interactions

Machine Scored Constructed Response interactions require scoring logic or a machine rubric within the interaction. MSCR interaction types include:

Machine Scored Constructed Response Interaction Type	Task Demands that can be Assessed
Equation Editor (EQ)	Calculate, Mathematically describe/represent/model, Identify
Table Input (TI)	Calculate, Sequence, Identify, Organize, Chart
Grid Interaction (GI)	Graph, Model, Represent, Show, Create
Simulation Interaction (Sim)	Investigate, Experiment, Observe, Gather/collect data, Model
Natural Language (NL)	Describe, Compare, Summarize, Explain
Editing Task (ET)	Correct
Word Builder (WB)	Identify

Appendix B

Item Review Checklist

Item Review Checklist

Tier 1 – Sufficiency/Appropriateness of the Phenomenon to Assess the Performance Expectation

The elements in this tier are critical

- Is the phenomenon based on a specific real-world scenario and focused enough to get the student to investigate what the Performance Expectation (PE) intends for them to investigate (i.e., the students' application of the Practice in the context of the Disciplinary Core Idea [DCI] and Crosscutting Concepts [CCC] as intended by the PE is sufficient to make sense of the phenomena)?
- Is there an appropriate science-related activity that is puzzling and/or intriguing for students to engage in? Is the scenario focused on real-world observations that students can connect with or have direct experience with?
- Is the context and complexity of the phenomenon grade-appropriate?
- Cluster Task Statement: Does the "call to action" reflect the end goal of the interactions to be answered? Does the statement make sense? Is this an engaging and reasonable outcome to work towards?
- Is the phenomenon presented in way(s) that all students can access and comprehend it based on information provided (including text, graphics, data, images, animations, etc.)? Is the phenomenon free of cultural bias, insensitivity or depreciation of unsafe situations?

Tier 2 – Review of Specific Elements by Component

Stimulus

Reading Load/Readability/Style

- Is the reading load appropriate for the grade (i.e., the amount of text minimized to reduce cognitive load)?
- Is the language and vocabulary appropriate for the grade?
- Non-specific vocabulary should be one grade level lower than the tested grade.
- Science vocabulary should be part of the "Science Vocabulary Students Are Expected to Know" in the item specifications.
- Is all of the information in the stimulus necessary for the student to complete the item interactions?
- Is language consistent throughout the cluster (i.e., does not switch between steam and vapor)?

- Is everything in the active voice (i.e., avoids unnecessary and unclear passive construction)?

Measurement/Units

- Are the data in SI units? Check style guide for exceptions.
- Are units of measurement introduced or defined before they are used in graphs/tables?
- Are the dependent/independent variables on the correct axes or in the correct columns?
- Are the graphs/tables/pictures free of extraneous information and appropriate for the grade level?
- Is there information included in graphs/pictures/tables that is not necessary and can be removed?
- Do the graphs/tables/pictures depend on color? Is there another way to represent the difference in the data other than by color (e.g., using patterns)?

Data Source and Scientific Reference

- Is content both accurate and appropriate in its context?
- Are the data sources appropriate for the subject/grade and taken from reliable academic sources?
- Does the item use the most up-to-date explanation?

Formatting

- Is everything presented within the browser dimensions (1024x768) without horizontal scrolling?
- Are the tables/graphs/etc. laid out in a way that is easy to read?
- Are details and text in animations easy to see? Are labels in diagrams easy to read?
- Is the average file size appropriate for test delivery (approximately 100KB, 250KB maximum)?

Item

Interaction and Alignment to Specifications

- Does the item make sense if you are responding to the interactions as if you are the student in the intended grade-level?
- Does the interaction require the student to demonstrate the science practice and/or content that the PE is assessing them on?
- Are the interactions grade level/developmentally appropriate and do they follow a logical progression? Do the interactions use appropriate scaffolding to guide students in making sense of the phenomena?
- Do the interactions align with the task demands?

- Do the interactions avoid redundancy? Do the student interactions follow a coherent progression?
- Do the student interactions follow a coherent progression? Does the order of the interactions allow students to make sense of the phenomenon or problem?
- Is the item stem worded in a way that makes the intent of the interaction clear to the student?
- Is it clear to the student what they will be scored on in the interaction?
- Is the language (e.g., words, phrases) consistent throughout the stimulus and items?

Grade Appropriate

- Is the content within the item accurate and grade appropriate?
- Are the correct units used? Are the units grade appropriate? Where necessary, are the abbreviations of the units introduced?
- Is the number of item parts/scoring assertions appropriate for the grade level?
- Is the mathematics level appropriate for the grade being tested?

Formatting

- Is everything presented within the browser frame without horizontal scrolling?
- Are the tables/graphs/etc. easy to read? Are the images created in an appropriate color palette per the Style Guide?
- Are details and text in animations easy to see?

Tier 3 – Review of the Scoring and Assertion(s)

Scoring Accuracy

- Do the interactions/task provide clear guidance on how student responses will be scored/interpreted?
- Are scores assigned appropriately as correct or incorrect?
- Are the dependencies logical?
- Are any of the scoring assertions exclusive (i.e., the student can get only one assertion correct and not another at any given time)?
- Is the correct answer clear and distinct from the distractors?
- Does the scoring result in an appropriate distribution of points?

Scoring Assertions

- Is the appropriate wording used for each scoring assertion (e.g., <Feature of response> providing some evidence of <what we want to infer about the student>)?
- Does the inference follow from the data?
- Are the assertions specific to the individual interactions (i.e., does not just repeat the PE)?
- Are the scoring assertions in the same order as the interactions?
- Does the wording of the scoring assertion make it very clear which interaction and action it refers to?

Strategies for Editing Text to Produce Plain Language

- Reduce excessive length
- Use common words
- Avoid ambiguous words
- Limit irregularly spelled words
- Avoid inconsistent naming and graphic conventions
- Avoid multiple terms for the same concept
- Limit the use of embedded clauses and phrases
- Avoid the passive voice

Appendix C

Content Advisory Committee Participant Details

Content Advisory Committee Participant Details

Table C-1. Content Advisory Committee Participants, Science

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Science Items Reviewed	Number of Science Items Approved by Teacher Committees
AIRCore	**March 2018**	Virtual	Elementary School Middle School High School	26[a]	**Gender**: Male 27%, Female 73% **Ethnicity**: Not collected **State**: Connecticut 46%, Hawaii 8%, Maryland 4%, Oregon 12%, West Virginia 27%, Utah 4% **Teaching Experience**: General Education 31%, General Ed and Other 12%, Science Curriculum Specialist 15%, Science Department Head 8%, STEM Consultant 8%, No response 27%	152	N/A[b]
Connecticut	**February 2017**	Cromwell, Connecticut	Elementary School	11	**Gender:** Male 22%, Female 78% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	45	31
			Middle School	14			
			High School	16			
	May 2017	New Britain, Connecticut	Elementary School	12	**Gender:** Male 26%, Female 74% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	40	N/A[b]
			Middle School	15			
			High School	15			
	October 2017	New Britain, Connecticut	Elementary School	11	**Gender:** Male 20%, Female 80% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	75	64
			Middle School	12			
			High School	18			
	November 2017	New Britain, Connecticut	Elementary School	7	**Gender:** Male 17%, Female 83% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	41	32
			Middle School	11			
			High School	17			
	January 2018	New Britain, Connecticut	Elementary School	11	**Gender:** Male 18%, Female 82% **Ethnicity:** Not collected **Region:** Not collected	42	25
			Middle School	14			

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Science Items Reviewed	Number of Science Items Approved by Teacher Committees
			High School	8	**Teaching Experience:** Not collected		
	October 2018	New Britain, Connecticut	Elementary School	13	**Gender:** Male 16%, Female 84% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	84	54
			Middle School	16			
			High School	16			
	November 2018	New Britain, Connecticut	Elementary School	10	**Gender:** Male 14%, Female 86% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	235	200
			Middle School	18			
			High School	21			
	December 2018	New Britain, Connecticut	Elementary School	10	**Gender:** Male 19%, Female 81% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	56	55
			Middle School	7			
			High School	15			
	January 2019	New Britain, Connecticut	Elementary School	13	**Gender:** Male 18%, Female 82% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	65	59
			Middle School	13			
			High School	18			
	September 2019	Rocky Hill, Connecticut	Elementary School	14	**Gender:** Male 18%, Female 82% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	60	57
			Middle School	16			
			High School	20			
Hawaii	**July 2017**	Honolulu, Hawaii	Elementary School	7	**Gender**: Male 36%, Female 64% **Ethnicity**: Black 5%, Chinese and White 5%, Filipino 9%, Hawaiian 14%, Hispanic 9%, Japanese 14%, White 41%, No response 5% **Region**: Not collected **Teaching Experience**: General Education 64%, General Education w/SPED Certification 5%, SPED Teacher 5%, Other 23%, No response 5%	25	N/A[b]
			Middle School	8			
			High School	7			
	September 2017	Honolulu, Hawaii	Elementary School	6	**Gender**: Male 25%, Female 75%	65	N/A[b]

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Science Items Reviewed	Number of Science Items Approved by Teacher Committees
			Middle School	8	**Ethnicity**: Black 5%, Filipino 10%, Hispanic 10%, Japanese 15%, White 50%, No response 10% **Region**: Not collected **Teaching Experience**: General Education 65%, General Education w/SPED Certification 15%, Other 20%		
			High School	6			
	October 2018	Honolulu, Hawaii	Elementary School	10	**Gender:** Male 17.24%, Female 82.76% **Ethnicity:** White 27.59%, N/A 10.34%, Hispanic 10.34%, Asian 31.03%, Hawaiian 3.45%, Asian Pacific Islander 6.9%, Two or More: 10.34% **Region:** Not collected **Teaching Experience:** General Education 82.76%, SPED Teacher 0%, ELL Teacher 0%, General Education w/ SPED Certification 0%, Other 24.14%	85	79
			Middle School	6			
			High School	12			
	February 2019	Honolulu, Hawaii	Elementary School	8	**Gender:** Male 20%, Female 80% **Ethnicity:** White 35%, Asian 50%, Two or More: 15% **Region:** Not collected **Teaching Experience:** General Education 65%, SPED Teacher 5%, General Education w/ SPED Certification 5%, Other 25%	44	44
			Middle School	6			
			High School	7			
Idaho	**December 2018**	N/A[d]	Elementary School	21[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	111	N/A[b]
			Middle School				
MSSA[c]	**January 2018**	N/A[d]	Elementary School	15	**Gender**: Not collected **Ethnicity**: Not collected **State**: 90% Rhode Island, 10% Vermont **Teaching Experience**: General Education 69%, Bilingual Education 2%, Science Coordinator 14%, Other 14%	73	N/A[b]
			Middle School	14			
			High School	13			
	March 2018	N/A[d]	Elementary School	12	**Gender**: Not collected **Ethnicity**: Not collected **State**: Rhode Island 25%, Vermont 75%	100	N/A[b]
			Middle School	13			

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Science Items Reviewed	Number of Science Items Approved by Teacher Committees
			High School	9	**Teaching Experience**: Not collected		
	January 2019	N/A[d]	Elementary School Middle School High School	21[a]	**Gender:** Male 25.71%, Female 74.29% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** General Education 68.57%, Special Education 2.86%, Bilingual Education 0%, Administration 0%, Other 28.57%, N/A 5.71%	116	N/A[b]
Oregon	**August 2017**	Salem, Oregon	Elementary School Middle School High School	4 3 3	**Gender**: Male 10%, Female 90% **Ethnicity**: Not collected **Region**: Urban 50%, Suburban 0%, Rural 50% **Teaching Experience**: Regular Education 100%, Bilingual Education 10%, Special Education 10%, Administration 20%, Other 0%	235	142
	August 2018	Salem, Oregon	Elementary School Middle School High School	4 8 6	**Gender:** Male 20%, Female 80% **Ethnicity:** Other 5%, White 95% **Region:** Urban 56%, Suburban 0%, Rural 44% **Teaching Experience:** Bilingual Education 65%, Special Education 65%, Other 55%	257	200
	December 2018	Virtual	Elementary School Middle School High School	6 5 5	**Gender:** Male 38%, Female 63% **Ethnicity:** Asian 6%, White 94% **Region:** Urban 50%, Suburban 50%, Rural 0% **Teaching Experience:** General Education 38%, Bilingual Education 63%, Special Education 25%	62	48
Utah	**July 2017**	Park City, Utah	Grade 6 Grade 7 Grade 8	6 6 6	**Gender:** Male 26.09%, Female 73.91% **Ethnicity:** White 91.3%, Native American 4.35%, Other 4.35% **Region:** Not collected **Teaching Experience:** General Education 100%, Special Education 4.35%, Bilingual Education 0%, Administration 0%, Other 4.35%	55	51
	December 2017	Salt Lake City, Utah	Grade 6 Grade 7	12 12	**Gender:** Male 16%, Female 83.87% **Ethnicity:** American Indian or Alaska Native and White 3.23%, Other 3.23%, White 93.55% **Region:** Not collected	64	62

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Science Items Reviewed	Number of Science Items Approved by Teacher Committees
			Grade 8	12	**Teaching Experience:** General Education 87.09%, General Education and Other 9.68%, General Education and ESOL 3.23%		
West Virginia	**January 2017**	N/A[d]	Elementary School	28[a, e]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	39	N/A[b]
			Middle School				
	October 2018	N/A[d]	Elementary School	10[a]	**Gender:** Male 11.11%, Female 88.89% **Ethnicity:** White 88.89%, Black 11.11% **Region:** Rural 100%, Urban 0%, Suburban 0% **Teaching Experience:** General Education 100%, Special Education 0%, Bilingual Education 0%, Administration 0%, Other 0%	191	N/A[b]
			Middle School				
	July 2019	N/A[d]	Elementary School	6	**Gender:** Male 13.04%, Female 86.96% **Ethnicity:** White 86.96%, Asian 4.35%, Black 4.35%, N/A 4.35% **Region:** Rural 69.57%, Urban 30.43%, Suburban 0%, N/A 4.35% **Teaching Experience:** General Education 71.74%, Special Education 4.35%, Bilingual Education 0%, Administration 0%, Other 13.04%, N/A 13.04%	50	N/A[b]
			Middle School	6			
Wyoming	**December 2017**	Cheyenne, Wyoming	Elementary School	6	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	51	N/A[b]
			Middle School	8			
			High School	4			
	October 2018	Cheyenne, Wyoming	Elementary School	14[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	37	N/A[b]
			Middle School				
			High School				

Note. [a]Number of Committee Members by grade band is not available.

[b]Number of science items approved by teacher committees is unavailable at the time of writing this report.

[c]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.

[d]Location of Content Advisory Committee Meeting is unavailable at the time of writing this report.

[e]Number of Committee Members includes total committee members for ELA, math, and science. The number for science only committee members is not available.

Appendix D

Fairness Committee Participant Details

Fairness Committee Participant Details

Table D-1. Fairness Committee Participants, Science

State	Date	Location	Total Number of Committee Members	Teacher Demographic Summary	Number of Items Reviewed
AIRCore	**March 2018**	Virtual	13	**Gender**: Male 15%, Female 85% **Ethnicity**: Not collected **State**: Connecticut 46%, Indiana 8%, Utah 15%, West Virginia 23%, Wyoming 8% **Teaching Experience**: General Education 8%, General Education and Other 15%, EL Instructional Coach 8%, No response 69%	152
Connecticut	**February 2017**	Cromwell, Connecticut	6	**Gender:** Male 17%, Female 83% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	45
	December 2017	New Britain, Connecticut	9	**Gender:** Male 22%, Female 78% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	75
	December 2017	Cromwell, Connecticut	10	**Gender:** Male 30%, Female 70% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	41
	February 2018	New Britain, Connecticut	3	**Gender:** Male 33%, Female 67% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	42
	November 2018	New Britain, Connecticut	11	**Gender:** Male 9%, Female 91% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	319
	December 2018	New Britain, Connecticut	10	**Gender:** Male 20%, Female 80% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	56
	January 2019	New Britain, Connecticut	9	**Gender:** Male 22%, Female 78% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	65

State	Date	Location	Total Number of Committee Members	Teacher Demographic Summary	Number of Items Reviewed
	September 2019	Cromwell, Connecticut	9	**Gender:** Male 11%, Female 89% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	48
Hawaii	**July 2017**	Honolulu, Hawaii	22	**Gender**: Male 36%, Female 64% **Ethnicity**: Black 5%, Chinese and White 5%, Filipino 9%, Hawaiian 14%, Hispanic 9%, Japanese 14%, White 41%, No response 5% **Region**: Not collected **Teaching Experience**: General Education 64%, General Education w/SPED Certification 5%, SPED Teacher 5%, Other 23%, No response 5%	25
	September 2017	Honolulu, Hawaii	20	**Gender**: Male 25%, Female 75% **Ethnicity**: Black 5%, Filipino 10%, Hispanic 10%, Japanese 15%, White 50%, No response 10% **Region**: Not collected **Teaching Experience**: General Education 65%, General Education w/SPED Certification 15%, Other 20%	65
	October 2018	Honolulu, Hawaii	29	**Gender:** Male 20.69%, Female 79.31% **Ethnicity:** White 27.59%, Japanese 10.34%, N/A 10.34%, Hispanic 10.34%, Chinese 6.9%, Asian 6.9%, Hawaiian 3.45%, Asian Pacific Islander 6.9%, Filipino 3.45%, Multi-Racial/Ethnic 13.8% **Region:** Not collected	85
	February 2019	Honolulu, Hawaii	21	**Gender:** Male 20%, Female 80% **Ethnicity:** White 35%, Asian 50%, Two or More: 15% **Region:** Not collected **Teaching Experience:** General Education 65%, SPED Teacher 5%, General Education w/ SPED Certification 5%, Other 25%	44
Idaho	**December 2018**	N/A[a]	15	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	111
MSSA[b]	**January 2018**	N/A[a]	21	**Gender**: Not collected **Ethnicity**: Not collected **State**: Rhode Island 100%, Vermont 0% **Teaching Experience**: General Education 67%, Bilingual Education 14%, Special Equation 5%, Science Coordinator 5%, Other 10%	73
	March 2018	N/A[a]	11	**Gender**: Not collected **Ethnicity**: Not collected **State**: Rhode Island 55%, Vermont 45% **Teaching Experience**: Not collected	100

State	Date	Location	Total Number of Committee Members	Teacher Demographic Summary	Number of Items Reviewed
	January 2019	N/A[a]	14	**Gender:** Male 22.86%, Female 62.86% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** General Education 68.57%, Special Education 2.86%, Bilingual Education 0%, Administration 0%, Other 17.14%, Coach 11.43%	116
Oregon	**August 2017**	Salem, Oregon	5	**Gender**: Male 0%, Female 100% **Ethnicity**: Not collected **Region**: Urban 80%, Suburban 20%, Rural 0% **Teaching Experience**: Regular Education 40%, Bilingual Education 20%, Special Education 20%, Administration 60%, Other 20%	110
	August 2018	Salem, Oregon	39	**Gender:** Male 26%, Female 74% **Ethnicity:** Asian 3%, Hispanic 8%, Native American 3%, White 82%, Other 10% **Region:** Urban 56%, Suburban 0%, Rural 44% **Teaching Experience:** General Education 15%, Bilingual Education 72%, Special Education 33%, Other 33%	257
	December 2018	Virtual	11	**Gender:** Male 9%, Female 91% **Ethnicity:** Hispanic 9%, White 91% **Region:** Urban 55%, Suburban 0%, Rural 45% **Teaching Experience:** General Education 27%, Bilingual Education 64%, Special Education 18%, Administration 9%, Other 64%	62
Utah	**August 2017**	Park City, Utah	6	**Gender**: Male 0%, Female 100% **Ethnicity**: American Indian or Alaska Native 33%, Hispanic 33%, White 33% **Region**: Urban 0%, Suburban 0%, Rural 17%, Unknown/No response/Not applicable 83% **Teaching Experience**: General Education 17%, Special Education 17%, Administrator 33%, Other 33%	44
	December 2017	Salt Lake City, Utah	6	**Gender:** Male 16.67%, Female 83.33% **Ethnicity:** Black 33.33%, Native American 33.33%, Hispanic 16.67, White 0%, N/A 16.67% **Region:** Not collected **Teaching Experience:** General Education 0%, Special Education 0%, Bilingual Education 0%, Administration 33.33%, Other 83.33%	48
West Virginia	**January 2017**	N/A[a]	28[c]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	34

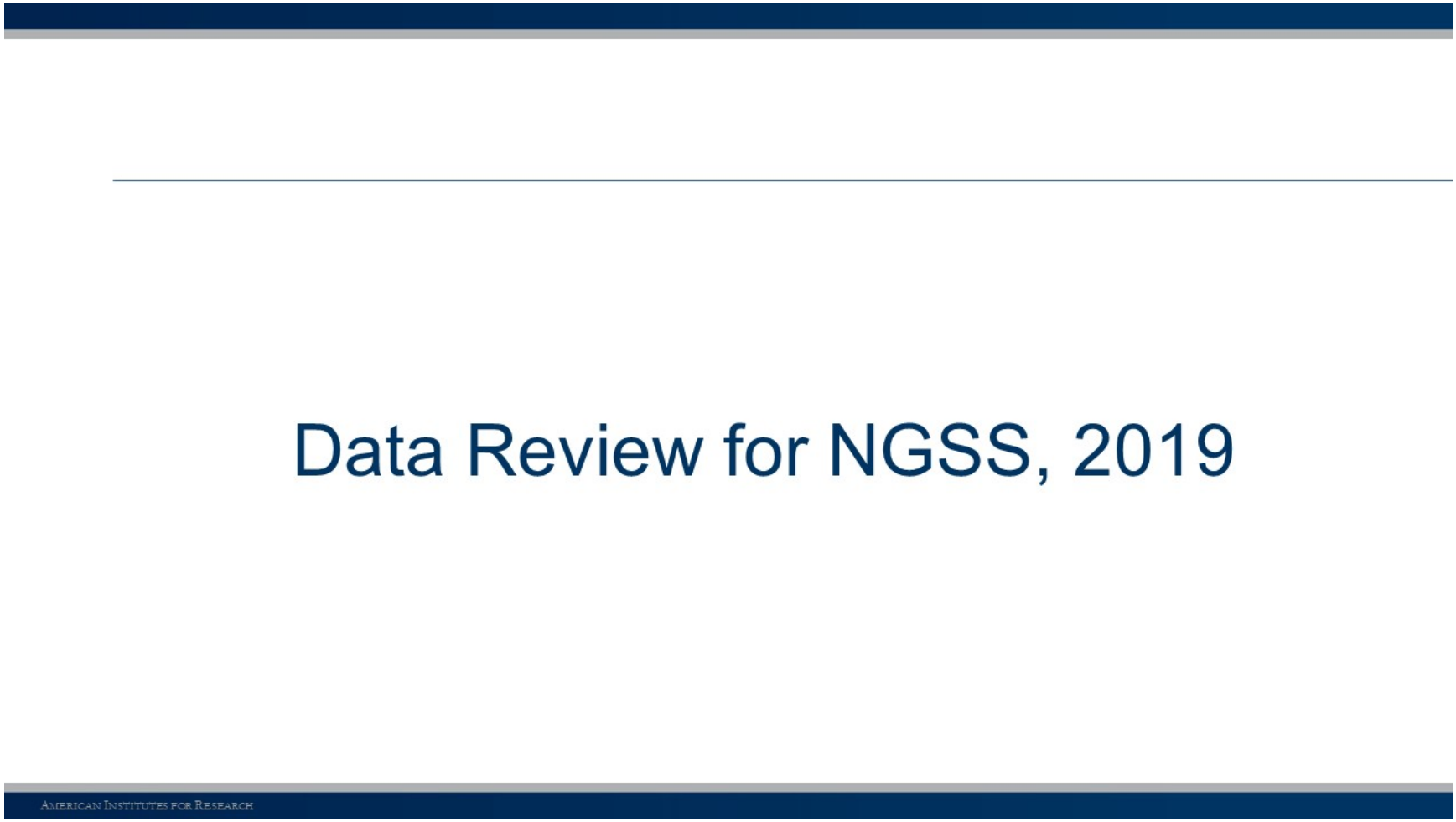
State	Date	Location	Total Number of Committee Members	Teacher Demographic Summary	Number of Items Reviewed
	January 2019	N/A[a]	10	**Gender:** Male 11.11%, Female 88.89% **Ethnicity:** Black 11.11%, White 88.89% **Region:** Rural 100%, Urban 0%, Suburban 0% **Teaching Experience:** General Education 100%, Special Education 0%, Bilingual Education 0%, Administration 0%, Other 0%	191
Wyoming	**December 2017**	Cheyenne, Wyoming	5	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	51
	October 2018	Cheyenne, Wyoming	5	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	37

Note. [a]Location of Fairness Committee Meeting is unavailable at the time of writing this report.
[b]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[c]Number of Committee Members includes total committee members for ELA, math, and science. The number for science only committee members is not available.

Appendix E

Sample Data Review Training Materials

Sample Data Review Training Materials



Read and Sign Non-Disclosure

- Read and Sign Non-Disclosure
- Turn in to AIR Facilitator

Overview of Training

- Steps in the Development Process
- Describe the structure of Three-Dimensional clusters
- Describe scoring assertions
- Role of the Data Review Committee
- Data Review Process
- Participant Guidelines

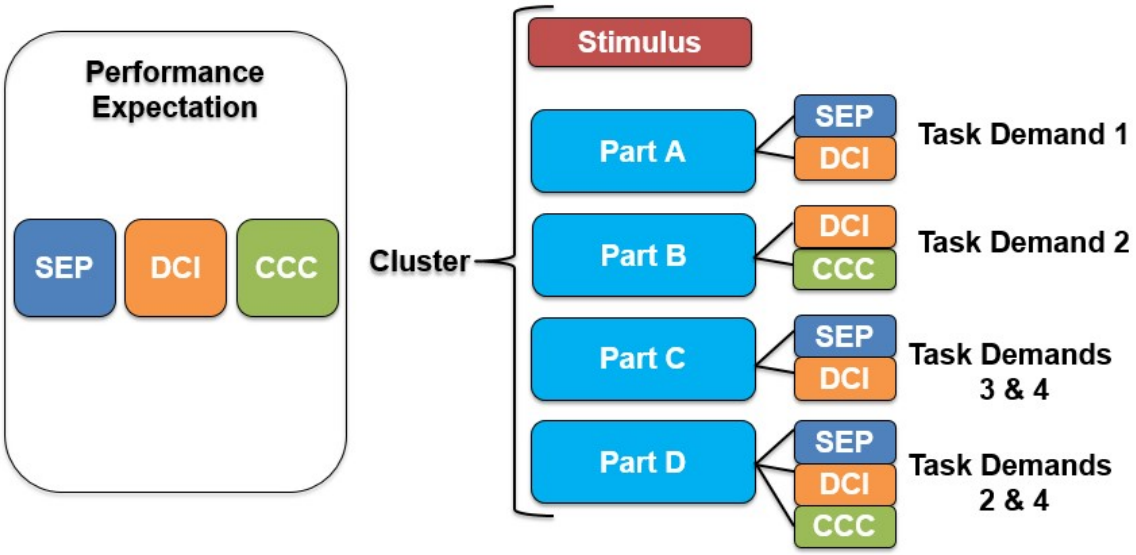
Steps in the Development Process

- AIR Writes Clusters & Standalones
- AIR Internal Review (Content & Fairness)
- Client & Educator Review (Content & Fairness)
- Field Test with Students
- Rubric Validation Process
- Update Scores & Generate Field Test Data
- **Review of Field Test Data**

NGSS in Hawaii

- A new Science Assessment has been developed to assess how well students master the NGSS
- The items of the new assessment look very different
  - Focus on item clusters
    - » Aligned to a single performance expectation
    - » Consisting of multiple interactions

Structure of AIR Clusters



Willow populations in Yellowstone National Park have increased since wolves were reintroduced to the park in 1995.

Willows are small trees that grow best in marshlike environments. After studying the Yellowstone food web shown in Diagram 1 and the population data for the park shown in Table 1, students arrive at two different hypotheses.

Diagram 1. Yellowstone Food Web

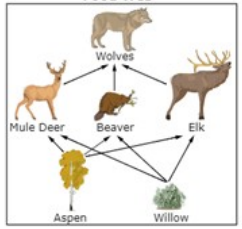


Table 1. Yellowstone Population Data

	Wolves	Elk	Beaver	Mule Deer
1995	31	16,791	10	2,014
2004	171	8,335	120	2,014

Note: These data are approximate.

Hypothesis 1:
When wolves were reintroduced to Yellowstone, the wolves preyed upon the elk, which allowed the beavers to eat more willow. This led to more beavers and beaver dams. Beaver dams create marsh environments that willows do well in, allowing the willow's population to increase.

Part A

Click on each box and select a word/phrase that completes the table with the Yellowstone population data from 1995 and 2004 and the hypothesis those data support.

Table 2. Summary of Yellowstone Population Data and Supported Hypotheses

Data	Hypothesis Supported
Elk population ▾	▾
Beaver population ▾	▾
Mule deer population ▾	▾

Part B

Which hypothesis is best supported by the evidence?

Ⓐ All of the evidence is consistent with Hypothesis 1.

Ⓑ All of the evidence is consistent with Hypothesis 2.

Ⓒ Most of the evidence is consistent with Hypothesis 1.

Ⓓ Most of the evidence is consistent with Hypothesis 2.

Ⓔ The evidence does not favor either hypothesis.

Part C

Aspen trees are shown in Diagram 1. Moose and bison are two plant-eating animal species that are not shown in Diagram 1 but are also part of the Yellowstone food web.

Based on Hypothesis 2, click on each box to select a word/phrase to make a prediction about what would happen to the moose, bison, and aspen tree populations after the reintroduction of wolves.

Table 3. Population Predictions

Species	Population after Wolf Reintroduction	Reason for Impact on Population
Moose	▾	▾
Bison	▾	▾
Aspen tree	▾	▾

American Institutes for Research 7

Scoring Assertions

- Scoring
  - Within each item cluster, a series of explicit assertions can be made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses
  - Scoring assertions can be supported based on students' responses in *one or more* interactions within an item cluster.
  - For example:
    - » A student correctly graphs data points indicating that (s)he can construct a graph showing the relationship between two variables,
    - » Makes an incorrect inference about the relationship between the two variables, thereby not supporting the assertion that the student can interpret relationships expressed graphically

2019 Test Administrations and Rubric Validation

- Items were embedded as field test items in Spring 2019
- This past June, Items went through rubric validation
  - To check whether assertions were scored correctly
    - » Looking at actual student responses
  - It was determined for two items that student facing changes were necessary, so they will be updated and re-field tested this next year
  - Some assertions were modified (deleted/added)

Data Review

- After rubric validation, statistics were computed at the assertion-level
  - Assertions can only be evaluated in the context of the entire item
  - Inclusion in data review will be decided at the item level, not at the assertion level
  - Inclusion is based on statistical flags that rely on assertion level statistics but are evaluated for the entire item

AMERICAN INSTITUTES FOR RESEARCH

Data Review

- Flagging is based on business rules related to:
  - Difficulty of the cluster
  - Relation between the score on cluster and the overall student's score
  - Response time of the cluster
  - Statistical flags for differential item functioning
- These items may be perfectly fine, but we want your input
  - Is this a good item and set of assertions?
  - Do you see any reason for why the item was flagged from a content perspective?

AMERICAN INSTITUTES FOR RESEARCH

Flagging Rules

- *p*-value
  - The *p*-value is the proportion of students for which the assertion is TRUE
  - Corresponds to the difficulty of an item in a traditional assessment
  - Across an item bank, we want to see assertions with *p*-values across the full range to be able to precisely measure proficiency across all proficiency levels
    - » A low *p*-value is not bad per se
  - However, we want to make sure the low *p*-value is not a result of an item being misleading

Flagging Rules

- *p*-value
  - Criteria for clusters:
    - » average *p*-value < .30 (across the assertions within a cluster)
    - » average *p*-value > .85 (across the assertions within a cluster)
  - Criteria for stand-alone items (typically has 1-3 assertions):
    - » average *p*-value < . 15 (across the assertions within a stand-alone)
    - » average *p*-value > .95 (across the assertions within a stand-alone)

AMERICAN INSTITUTES FOR RESEARCH

Flagging Rules

- Item-total correlation
  - We expect students who do well on the test overall to have a higher probability of doing well on individual assertions
  - The item-total correlation describes that relation
  - Criterion
    - » Average item-total (biserial) correlation < .25
    - » One or more assertions with an item-total correlation < 0

Flagging Rules

- **Differential item functioning**
  - Fair items behave similar across groups
  - Probability of answering correctly is the same for all students of similar ability regardless of group membership
- **Groups are defined by**
  - Gender
  - Ethnicity
  - Economically disadvantaged vs. not
  - LEP vs. not
  - Special Education vs. not

Flagging Rules

- Severity of possible bias
  - "A" No statistical evidence of DIF
  - "B" Evidence for potential mild DIF
  - "C" Evidence for potential severe DIF
- Direction of possible bias
  - "–" assertion favors reference groups (whites/females/non LEPs)
  - "+" assertion favors focal group

Flagging Rules

- Criteria
  - For clusters: 2 or more assertions show 'C' DIF in the same direction
  - For stand-alone items: 1 or more assertions show 'C' DIF in the same direction

Flagging Rules

- Timing
  - We want a good balance between the amount of information an item provides, and the time students spend on the item
- Criteria
  - For clusters: percentile 80 > 15 minutes
    - A percentile 80 of *x* minutes: 80% of the students spent *x* minutes or less on the cluster
  - For stand-alone items: percentile 80 > 3 minutes
  - Assertions per minute < .5 for clusters and stand-alone items

American Institutes for Research

Data Review Process

AMERICAN INSTITUTES FOR RESEARCH

Process

- Item is presented with information on
  - » Grade
  - » Discipline
  - » Topic
  - » Performance Expectation
- Facilitator will present the cluster or stand-alone item
- Statistics on the assertions of the item are presented
  - – Including the reason for flagging
- Evaluation of item (Stimulus, interactions, assertions)
- For every item, one of the following decisions is made
  - – Reject
  - – Accept as is

AMERICAN INSTITUTES FOR RESEARCH

Participant Guidelines

- Keep phones turned off & stowed while in the meeting room.
  - If needed, please take the call outside of meeting room
- Keep your name tent visible.
- Do not keep personal items on the table with secure materials.
  - No personal laptop or tablet use is allowed in the meeting rooms.
- Do not speak to other panelists about specific items outside of the meeting rooms.
- To limit disruptions, try to take breaks at designated break times.
- If you have any questions about the review or procedures, feel free to talk to AIR or DOE staff during breaks or at lunch.

Questions?

Appendix F

Data Review Committee Participant Details

Data Review Committee Participant Details

Table F-1. Data Review Committee Participants, Science

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Items Reviewed
AIRCore	**July 2018**	Virtual	Elementary School	18[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	84
			Middle School			
			High School			
	August 2019	N/A	Elementary School	N/A[b]	N/A[b]	43
			Middle School			
			High School			
Connecticut	**August 2018**	New Britain, Connecticut	Elementary School	10	**Gender:** Male 12%, Female 88% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	18
			Middle School	8		
			High School	8		
	August 2019	Cromwell, Connecticut	Elementary School	7	**Gender:** Male 17%, Female 83% **Ethnicity:** Not collected **Region:** Not collected **Teaching Experience:** Not collected	53
			Middle School	10		
			High School	6		
Hawaii	**August 2018**	Honolulu, Hawaii	Elementary School	18[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	32
			Middle School			
			High School			
	August 2019	Honolulu, Hawaii	Elementary School	6	**Gender:** Male 29%, Female 71%	37

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Items Reviewed
			Middle School	7	**Ethnicity:** American Indian and White 12%, Asian 41%, Asian and White 6%, Hispanic and White 12%, Native Hawaiian or Pacific Islander 18%, White 12% **Region:** Not collected **Teaching Experience:** General Education 53%, General Education with SPED Certification 6%, Bilingual Education 0%, Administration 0%, Other 29%, Special Education 12%	
			High School	5		
Idaho	**August 2019**	N/A[c]	Elementary School	10[a]	**Gender:** Male 20%, Female 70%, Did not specify 1% **Ethnicity:** White 100% **Region:** Rural 60%, Urban 0%, Suburban 40% **Teaching Experience:** General Education 60%, Administration 2%, Coach 20%	12
			Middle School			
MSSA[d]	**August 2018**	N/A[e]	N/A[e]	N/A[e]	N/A[e]	9
	August 2019	N/A[e]	N/A[e]	N/A[e]	N/A[e]	14
Oregon	**September 2018**	Salem, Oregon	Elementary School	3	**Gender**: Male 18%, Female 82% **Ethnicity**: White 100% **Region**: Urban 27%, Suburban 0%, Rural 73% **Teaching Experience**: Regular Education 64%, Bilingual Education 55%, Special Education 36%, Administration 18%, Other 18%	44
			Middle School	4		
			High School	4		
	August 2019	Remote	Elementary School	1	**Gender**: Male 50%, Female 50% **Ethnicity**: White 100% **Region**: Urban 50%, Suburban 0%, Rural 50% **Teaching Experience**: Regular Education 50%, Bilingual Education 25%, Special Education 25%, Administration 25%, Other 75%	8
			Middle School	2		
			High School	1		
Utah	**August 2018**	Salt Lake City, Utah	Grade 6	6	**Gender**: Male 7%, Female 93% **Ethnicity**: White 87%, Unknown 13% **Region**: Urban 0%, Suburban 13%, Rural 27%, Unknown/no response 60% **Teaching Experience**: General Education 100%	40
			Grade 7	5		
			Grade 8	5		
West Virginia	**July 2018**	N/A[c]	Elementary School	4[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	3
			Middle School			

State	Date	Location	Grade Band	Number of Teachers in Each Group	Teacher Demographic Summary	Number of Items Reviewed
	September 2019	N/A[c]	Elementary School	4[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	7
			Middle School			
Wyoming	**October 2018**	Cheyenne, Wyoming	Elementary School	11[a]	**Gender**: Not collected **Ethnicity**: Not collected **Region**: Not collected **Teaching Experience**: Not collected	16
			Middle School			
			High School			
	August 2019	Cheyenne, Wyoming	Elementary School	3	**Gender**: Male 10%, Female 90% **Ethnicity**: N/A **Region**: Urban 0% Suburban 40%, Rural 60% **Teaching Experience:** 90% Regular Education, 10% Administration	16
			Middle School	4		
			High School	3		

Note. [a]Number of Committee Members by grade band is not available.
[b]In summer 2019, AIRCore field-test items were taken to Connecticut, Hawaii, and Idaho for committee review.
[c]Location of Data Review Committee Meeting is unavailable at the time of writing this report.
[d]MSSA = Rhode Island and Vermont's Multi-State Science Assessment.
[e]Conducted by the Rhode Island Department of Education and the Vermont Agency of Education science content experts.

Appendix G

Example Item Interactions

Interaction Types Available in the Multi-State Science Assessment (MSSA)

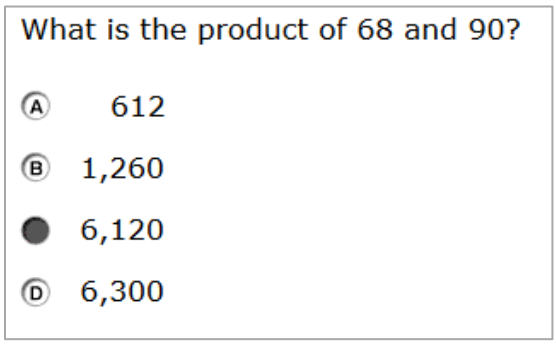
Review of Different Interaction Types

Interaction Type	Associated Sub-Types	Legacy Item Types Supported
Choice	Multiple Choice	MC
	Multiple Select	MS
	Scaffolding	ASI2, ASI3
Text Entry	Simple Text Entry	EA, ECR, LA, OE, SA, SR, WCR, RW, SCR
	Embedded Text Entry	CL, FI
	Natural Language	NL
	Extended Response	ER
Table	Table Match	MI
	Table Input	TI
	Column Match	MI
Edit Task	Edit Task	ET
	Edit Task with Choice	ETC
	Edit Task Inline Choice	ETC
Hot Text	Selectable	HTQ
	Re-orderable	HT
	Drag-from-Palette	DnD
	Custom	HTQ, HT, DnD
Equation	N/A	EQN
Grid	Grid	GI
	Hot Spot	GI
	Graphic Gap Match	GI
Simulation*	N/A	SIM

Note. the abbreviations correlate to the attributes used in AIR's Item Tracking System

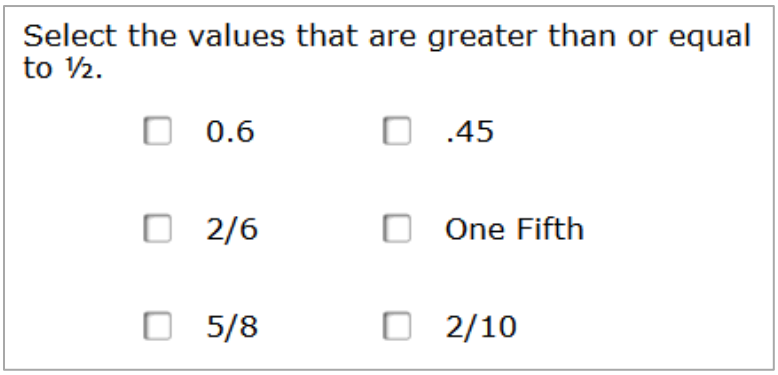
Multiple-Choice Interactions

Multiple-Choice (MC) interactions require students to select a single option from a list of possible answer options. The number and orientation of answer options in a multiple-choice interaction are configurable. Answer options may appear vertically, horizontally, vertically-stacked (in a specified number of columns), or horizontally-stacked (in a specified number of rows).



Multiple-Select Interactions

Multiple-Select interactions require students to select one or more options from a list of possible answer options. The number and orientation of answer options in a multiple-select interaction are configurable. Answer options may appear vertically, horizontally, horizontally-stacked (in a specified number of rows), or vertically-stacked (in a specified number of columns).



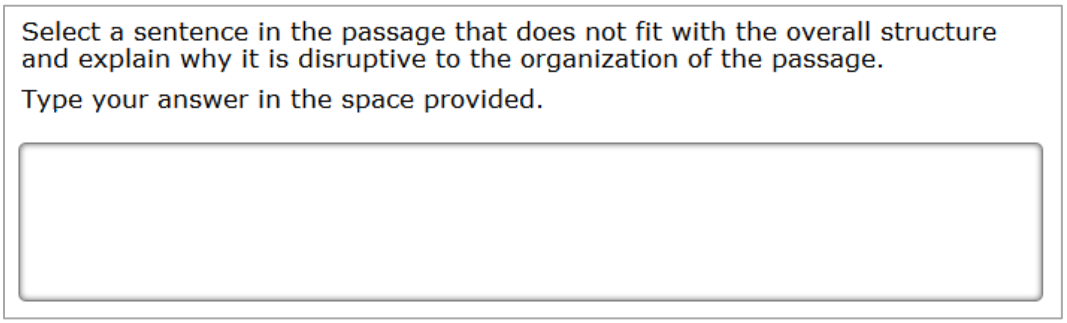
Text Entry Interactions

The Text Entry Interaction Editor allows you to create content for the following interaction types:

- Simple Text Entry Interactions
- Embedded Text Entry Interactions
- Natural Language Interactions
- Extended Response Interactions

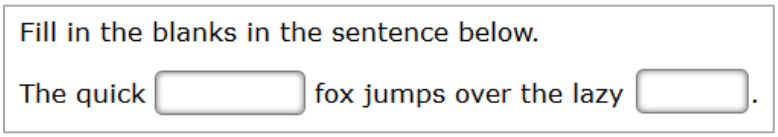
Simple Text Entry Interactions

Simple Text Entry interactions require students to type a response in a text box. For Simple Text Entry interactions, we can allow you to specify the maximum response length for the text box and the type of text editor available to students.



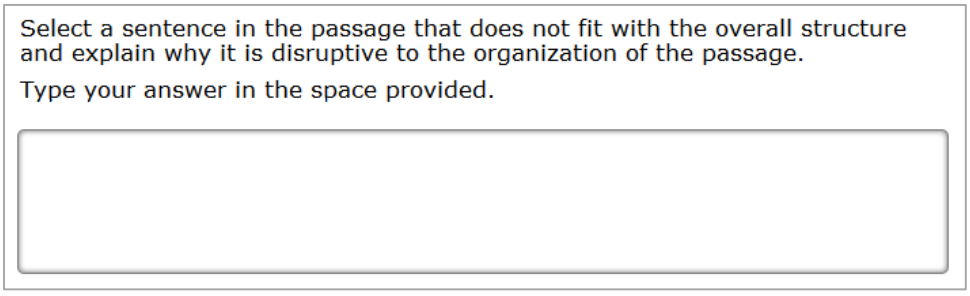
Embedded Text Entry Interactions

Embedded Text Entry interactions require students to type their response in one or more text boxes that are embedded in a section of read-only text.



Extended Response Interactions

Extended Response interactions require students to type a response in a text box. Extended Response interactions are scored by an uploaded essay scoring model that analyzes the student's response to identify variations of acceptable key words and phrases. For Extended Text Entry interactions, we can allow you to specify the maximum response length for the text box and the type of text editor available to students.



Alert: Extended Response interactions cannot be combined with any other interactions in the item.

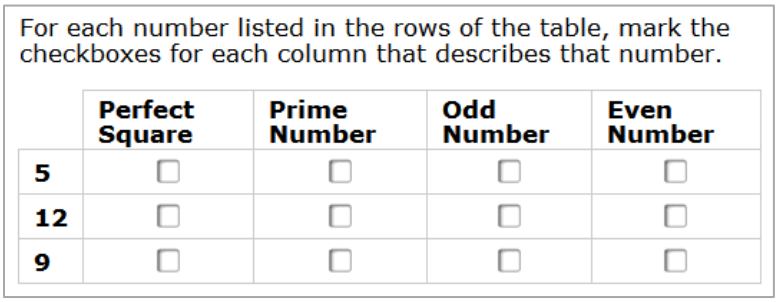
Table Entry Interaction

The Table Entry Interaction Editor allows you to create content for the following interaction types:

- Table Match Interactions
- Table Input Interactions
- Column Match Interactions

Table Match Interactions

Table Match interactions arrange two sets of match options in a table, with one set listed in columns and the other set listed in rows. Students match options in the columns to options in the rows by marking checkboxes in the cells where the columns and rows intersect.



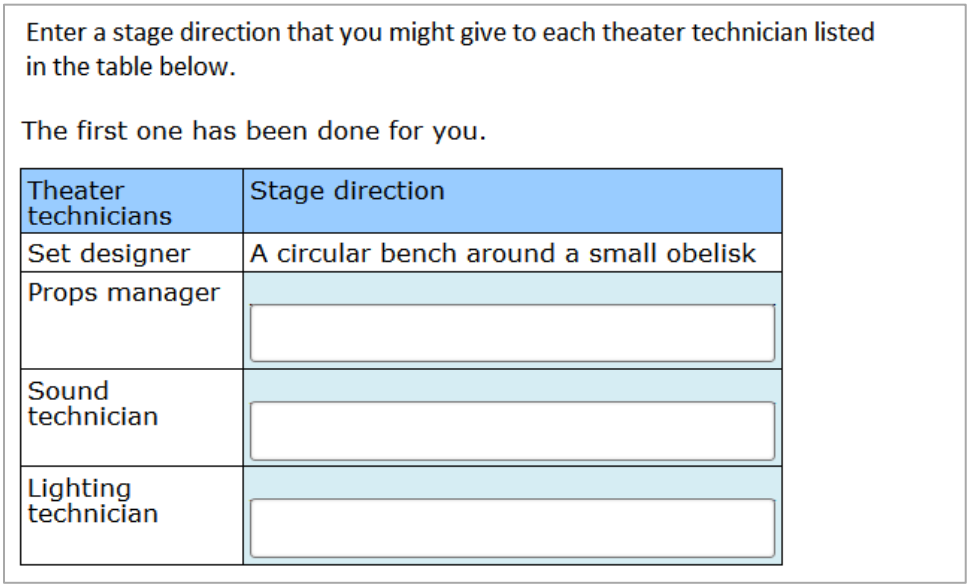
For each number listed in the rows of the table, mark the checkboxes for each column that describes that number.

	Perfect Square	Prime Number	Odd Number	Even Number
5	☐	☐	☐	☐
12	☐	☐	☐	☐
9	☐	☐	☐	☐

Table Match interactions allow you to customize the number of match options in each set and enter the content for each match option. You can also set restrictions on the number of matches students can make. By default, the panel includes a basic table consisting of three rows and columns (including the row header and column header).

Table Input Interactions

Table Input interactions provide students with a table that includes one or more blank cells. Each blank cell displays a text box in which students can type their response.



Enter a stage direction that you might give to each theater technician listed in the table below.

The first one has been done for you.

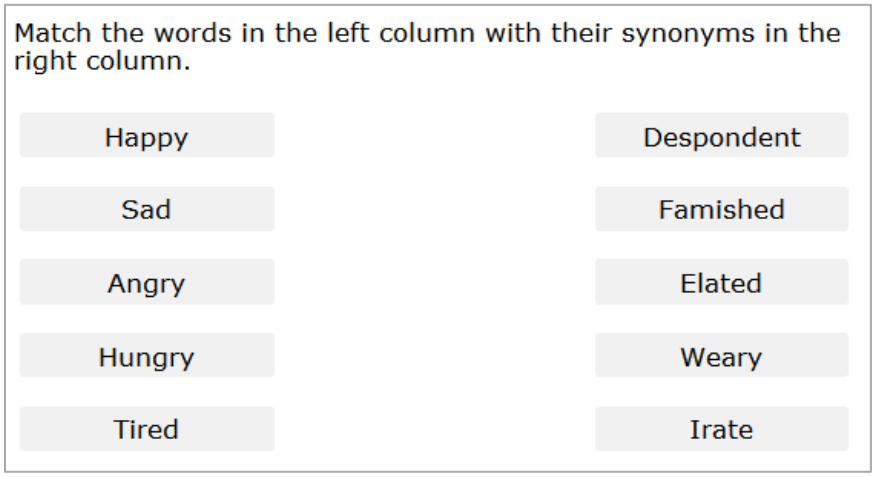
Theater technicians	Stage direction
Set designer	A circular bench around a small obelisk
Props manager	
Sound technician	
Lighting technician	

Table Input interactions allows you to customize the number of rows and columns in the table, specify which cells display text boxes, and enter content for the read-only cells. By default, the panel includes a basic table consisting of three rows and columns (including the row header and column header).

Alert: If a table does not include row headers, then it must include column headers. If a table does not include column headers, then it must include row headers.

Column Match Interactions

Column Match interactions provide students with two columns that each contain a set of match options. Students respond to the interaction by selecting a match option in the left column and then selecting the corresponding match option in the right column. A match option in one set may have one, multiple, or no matches in the other set.



Column Match interactions allows you to customize the number of match options in each set and enter the content for each match option. By default, the panel includes two single-column tables, each of which includes two match options. You can also set restrictions on the number of matches students can make.

Edit Task Interactions

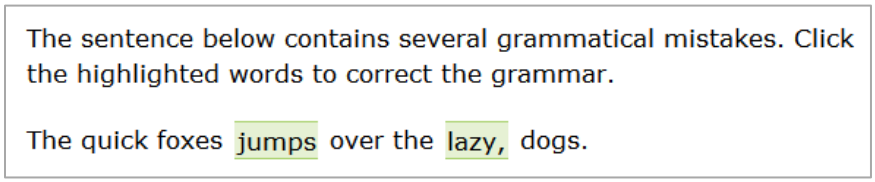
The Edit Task Interaction Editor allows you to create content for the following interaction types:

- Edit Task Interactions
- Edit Task with Choice Interactions
- Edit Task Inline Choice Interactions

Edit Task Interactions

Edit Task interactions provide students with a sentence or paragraph containing one or more tagged text elements. Tagged elements usually contain an error, such as improper spelling or grammar.

To respond to these interactions, students click a tagged element and enter corrected text in an editing window. The entered text replaces the original tagged text.



Edit Task interactions allow you to enter the text that appears in the response area and tag elements within the text that students can edit.

Warning: You cannot include hand-scored and machine-scored interactions in the same item.

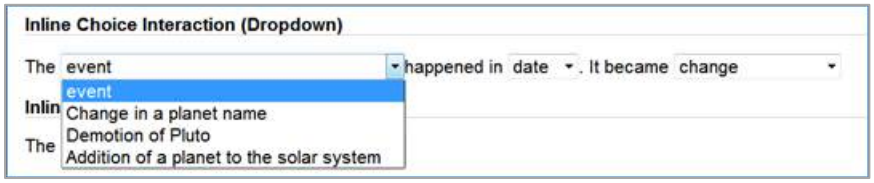
Edit Task with Choice Interactions

Edit Task with Choice interactions are similar to Edit Task interactions. The only difference is that when responding to Edit Task with Choice interactions, students replace the tagged text elements with options selected from a drop-down list.

Edit Task with Choice interactions allow you to enter the text that appears in the response area and tag elements within the text that students can edit.

Edit Task Inline Choice Interactions

Edit Task Inline Choice interactions are similar to Edit Task with Choice interactions. The only difference is that students select replacement options from a drop-down list embedded within the read-only text, rather than accessing the drop-down list via a pop-up window.



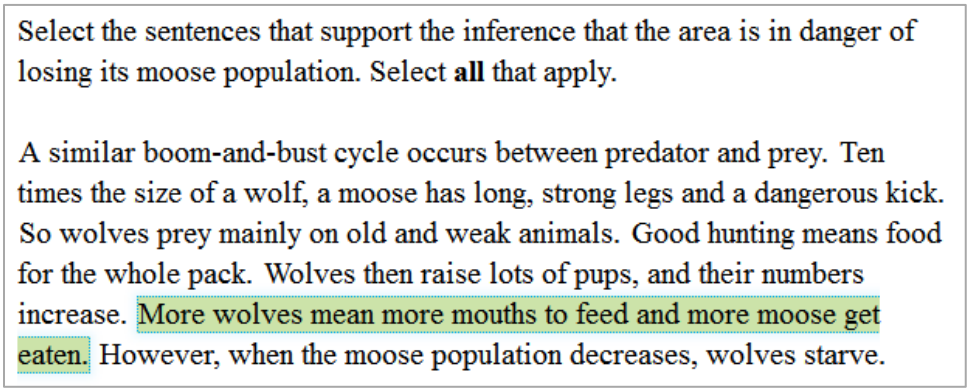
Hot Text Interactions

The Hot Text Interaction Editor allows you to create content for the following interaction types:

- Selectable Hot Text Interactions
- Re-orderable Hot Text Interactions
- Drag-from-Palette Hot Text Interactions
- Custom Hot Text Interactions

Selectable Hot Text Interactions

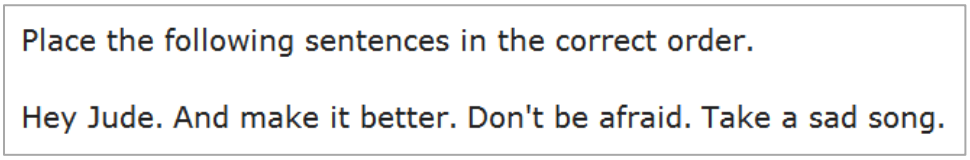
Selectable Hot Text interactions require students to select one or more text elements in the response area.



Selectable Hot Text interactions allows you to set the minimum and maximum number of elements students can select, enter the text that appears in the response area, and tag the text elements that will be selectable.

Re-orderable Hot Text Interactions

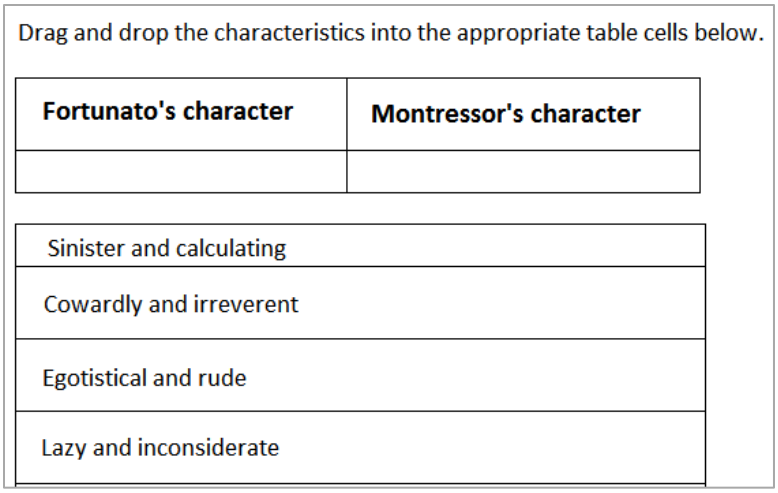
Re-orderable Hot Text interactions require students to click and drag hot text elements into a different order.



Re-orderable Hot Text interactions allow you to enter the re-orderable text elements in the response area. You can specify the elements' orientation and set them to appear in random order to students.

Drag-from-Palette Hot Text Interactions a.k.a. Hot Text Gap Match

Drag-from-Palette Hot Text interactions require students to drag elements from a palette into the available blank table cells or "gaps" (text boxes) in the response area. Palette elements may consist of text and/or images. Students may be able to drag the same palette element into multiple gaps, depending on the interaction's configuration.



Drag-from-Palette Hot Text interactions allow you to enter the elements that appear in the palette, enter static text for the response area, and create the gap targets where students can drag the text elements. You can enter all of the elements in a single text box or enter each segment in its own text box.

- Can set a minimum/maximum number of times a student is required/allowed to use a specific palette object
- Only supports drag-and-drop of palette items (images or plain text) onto pre-defined drop targets ("gaps" or "blanks") in the body text
    - These palette items are always confined to a special palette region (no "preplacing" them)
    - There is some control over palette placement
    - The items can only be placed in predefined "target" regions

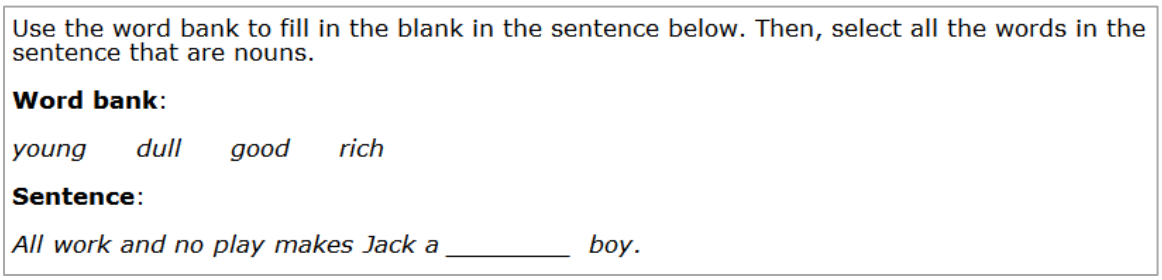
Custom Hot Text Interactions

Custom Hot Text interactions combine the functionality of the other Hot Text interaction sub-types. Students responding to a Custom Hot Text interaction may need to select text elements, rearrange text elements, and/or drag text elements from a palette to blank table cells or drop targets in the response area. In many ways, this is the grid of the text-interaction world. In practice, it is typically used to do drag-and-drop with text, but it can technically do more:

- Supports dragging and dropping text elements onto drop target areas
    - Text elements can originally be placed anywhere in the interaction (there's no dedicated palette)
    - Multiple elements can be dropped onto a target
        - this constitutes a "group"
        - much like grid hotspots, you can set constraints on the group

- Supports selectable text elements

- Like grid hotspots, these too can be grouped



Custom Hot Text interactions allow you to create groups of text elements, as well as the drop targets and static text that appear in the response area. When you create a group of text elements, you must assign a Hot Text functionality to that group. The following functionalities are available:

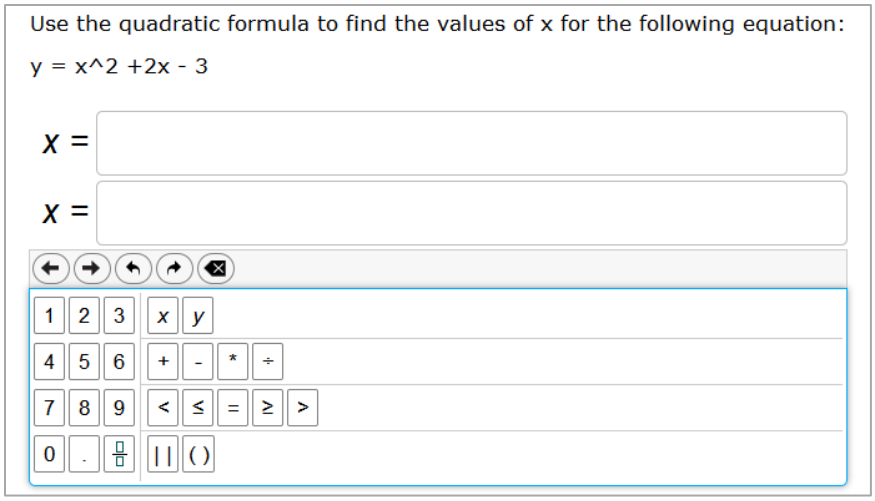
- **Selectable:** When you assign this functionality to a group, the text elements in the group behave like elements in a Selectable Hot Text interaction. You cannot add drop target elements to this kind of group.

- **Draggable:** When you assign this functionality to a group, the text elements in the group behave like elements in a Re-Orderable Hot Text interaction. If you assign this functionality to a group and also add drop targets to the group, the text elements in the group behave like elements in a Drag-from-Palette Hot Text interaction.

You can create as many groups as you wish, but you can only assign one Hot Text functionality to each group.

Equation Interaction Editor

The Equation Interaction Editor allows you to create content for Equation interactions only. Equation interactions require students to enter a response into input boxes using an on-screen keypad, which may consist of special mathematics characters. Students can also enter their response via a physical keyboard, but they cannot enter any characters that are not included in the on-screen keyboard.



Equation interactions allow you to select the buttons to include in the on-screen keypad, enter static text in the response area, and specify the number of input boxes to include in the response area. When selecting buttons to include in the keypad, you can add individual buttons or an entire row or tab of buttons.

Grid Interactions

The Grid Interaction Editor allows you to create content for the following interaction types:

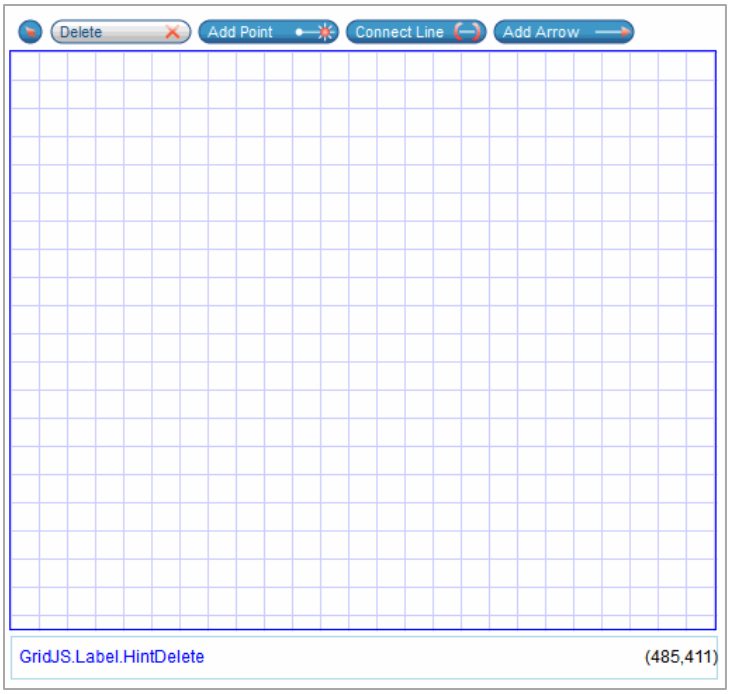
- Grid Interactions
- Hot Spot Interactions
- Graphic Gap Match Interactions

Note: Although there are three options available in the **Interaction Type** drop-down list, the generic **Grid** option allows you to create interactions with functionality similar to Hot Spot and Graphic Gap Match sub-types.

Grid Interactions Types

Grid interactions require students to enter a response by interacting with a grid area in the answer space. There are three general ways in which students can interact with the grid area.

- **Graphing Functionality:** Students can use various tool buttons to add points, lines, and other geometric shapes to the grid area. Only the Grid interaction sub-type allows you to create interactions with this functionality.



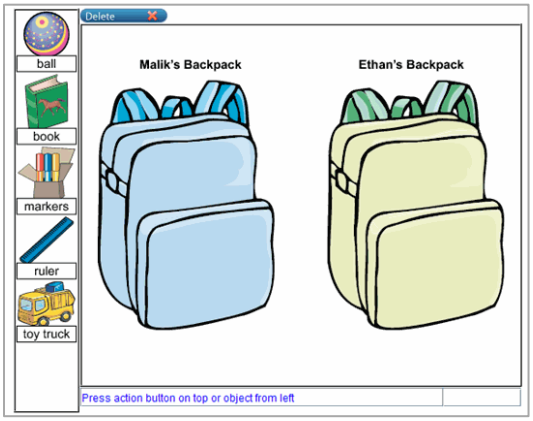
- **Hot Spot Functionality:** Students can click or hover over interactive regions in the grid area (hot spots) in order to activate them. Activated hot spots become highlighted, become outlined, or display an image. The Grid and Hot Spot interaction sub-types allow you to create interactions with this functionality.
    - Hotspots can be defined in groups, each of which can have its own selection constraints
    - These regions support events so clicking a hotspot might change the appearance of the interaction by showing/hiding other images, for example

School regulations include a requirement for the ration of fat to protein. Select the box in appropriate column next to each ingredient to show whether it has:

- Less than 1 gram of protein for every 3 grams of fat.
- 1 – 2 grams of protein for every 3 grams of fat.
- More than 2 grams of protein for every 3 grams of fat.

	Less than 1 gram of protein for every 3 grams of fat	Between 1 and 2 grams of protein for every 3 grams of fat	More than 2 gram of protein for every 3 grams of fat
Pretzels			
Sesame sticks			
Chocolate bits			
Almonds			
Sunflower seeds			
Raisins			
Banana chips			

- **Drag-and-Drop Functionality:** Students can click image or text objects and drag them into various locations in the grid area. The objects for these interactions are either provided in a palette beside the grid area or pre-placed within the grid area itself. The Grid and Graphic Gap Match interaction sub-types allow you to create interactions with this functionality; however, only Graphic Gap Match interactions allow text objects.
    - These palette items can be "preplaced" on the canvas or listed in a separate palette
    - The items can be placed anywhere on the canvas or guided to specific regions with snap points



Note: The functionalities of these interaction types are not mutually exclusive. A single Grid interaction may require students to select hot spots and place objects, or graph lines and select hot spots, and so on. However, a Grid interaction cannot include preplaced objects if it also includes the **Delete** tool button above the grid area.

Grid Hot Spot Interactions

Hot Spot interaction sub-types allow you to create Grid interactions with hot spot functionality. These interactions require students to select hot spot regions in the grid area.

- Only supports click-to-select "hotspots"
    - No visual side-effect events are supported
    - No hotspot groups are supported

Grid Graphic Gap Match Interactions

Graphic Gap Match interactions allow you to create Grid interactions with both hot spot and drag-and-drop functionality. These interactions require students to drag image objects from a palette to hot spot regions (gaps) in the grid area.

- Only supports drag-and-drop of palette items (images or plain text) onto the canvas/background

- These palette items are always confined to a special palette region (no "preplacing" them on the canvas)
- The items can only be placed in predefined "target" regions

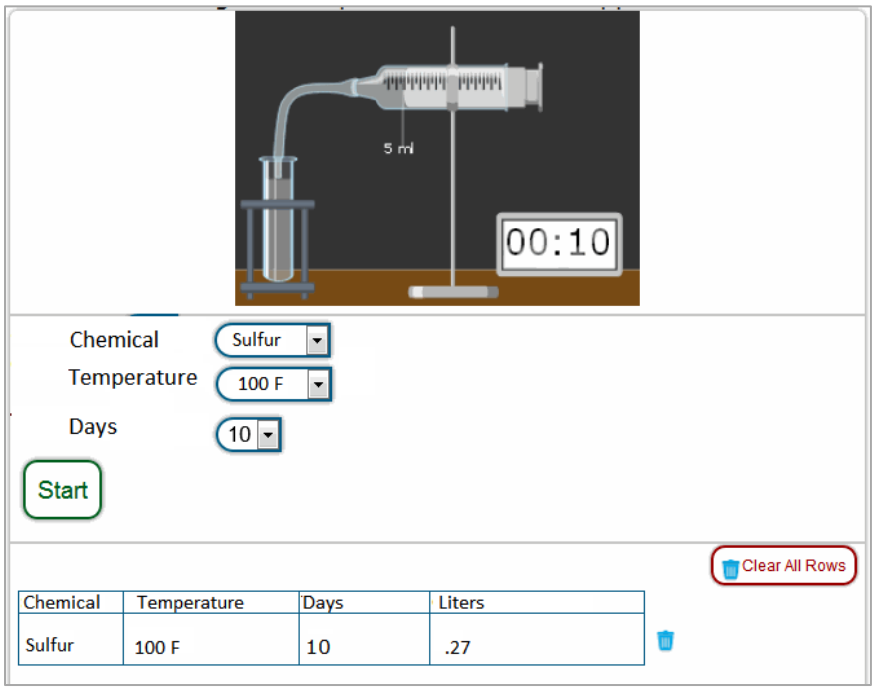
Alert: Graphic Gap Match interactions do not allow you to enable Snap-to-Point or Snap-to-Grid mode. You cannot pre-place image or text objects in the grid area with Graphic Gap Match Interactions.

Basically, graphic gap match and hotspot are dedicated interactions that don't support all the features of a grid. The trade-off here is:

- Graphic gap match and hotspot interactions are rendered differently (more simplistically)
- In some ways, graphic gap match and hotspot are easier to author and maintain
- Grid interactions need to use the "grid rubric tool," which is quite complicated

Simulation Interaction Editor

The Simulation Interaction Editor allows you to create content for Simulation interactions only. Simulation interactions consist of an animation tool, a set of input tools, and an output table. Students select parameters from the input tools to influence the animation. After the animation runs, the simulation results appear in the output table. Students can run multiple trials with different parameters to insert additional rows into this table.



Appendix H

Science Item Bank

Science Item Bank

Table H-1. Spring 2019 Shared Science Assessment Operational and Field-Test Item Bank by Performance Expectation, Elementary School

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
Earth and Space Sciences	ESS1	4-ESS1-1: Earth's Systems & Processes	2	0	6	8
		5-ESS1-1: Space Systems	1	2	4	7
		5-ESS1-2: Space Systems	4	0	3	7
	ESS2	3-ESS2-1: Weather & Climate	3	1	4	8
		3-ESS2-2: Weather & Climate	2	0	6	8
		4-ESS2-1: Earth's Systems & Processes	1	0	5	6
		4-ESS2-2: Earth's Systems & Processes	1	0	4	5
		5-ESS2-1: Earth's Systems	0	1	3	4
		5-ESS2-2: Earth's Systems	3	1	3	7
	ESS3	3-ESS3-1: Weather & Climate*	1	0	2	3
		4-ESS3-1: Energy	2	0	2	4
		4-ESS3-2: Earth's Systems & Processes*	1	2	1	4
		5-ESS3-1: Earth's Systems	2	0	3	5
Life Sciences	LS1	3-LS1-1: Inheritance	1	2	4	7
		4-LS1-1: Structure, Function, Information Processing	6	0	5	11
		4-LS1-2: Structure, Function, Information Processing	1	1	5	7
		5-LS1-1: Matter & Energy	2	0	6	8
	LS2	3-LS2-1: Ecosystems	3	0	7	10
		5-LS2-1: Matter & Energy	1	0	4	5
	LS3	3-LS3-1: Inheritance	1	2	4	7

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		3-LS3-2: Inheritance	1	1	2	4
	LS4	3-LS4-1: Ecosystems	2	0	6	8
		3-LS4-2: Inheritance	6	0	5	11
		3-LS4-3: Ecosystems	3	0	2	5
		3-LS4-4: Ecosystems*	1	0	3	4
Physical Sciences	PS1	5-PS1-1: Structure & Properties of Matter	2	0	5	7
		5-PS1-2: Structure & Properties of Matter	2	1	4	7
		5-PS1-3: Structure & Properties of Matter	2	0	3	5
		5-PS1-4: Structure & Properties of Matter	0	1	1	2
	PS2	3-PS2-1: Forces-balanced and unbalanced forces	2	1	6	9
		3-PS2-2: Forces-pattern predicts future motion	3	0	2	5
		3-PS2-3: Forces-between objects not in contact	1	0	4	5
		3-PS2-4: Forces-magnets*	0	0	2	2
		5-PS2-1: Space Systems	1	0	4	5
	PS3	4-PS3-1: Energy-relationship between speed and energy of object	4	0	6	10
		4-PS3-2: Energy-transfer of energy	4	0	2	6
		4-PS3-3: Energy-changes in energy when objects collide	3	0	4	7
		4-PS3-4: Energy-converting energy from one form to another*	0	0	2	2
		5-PS3-1: Matter & Energy	2	1	2	5
	PS4	4-PS4-1: Waves-waves can cause objects to move	0	0	3	3
		4-PS4-2: Structure, Function, Information Processing	1	0	6	7

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		4-PS4-3: Waves-using patterns to transfer information*	1	0	1	2

Note. *These PEs have an engineering component.

[a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table H-2. Spring 2019 Shared Science Assessment Operational and Field-Test Item Bank by Performance Expectation, Middle School

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
Earth and Space Sciences	ESS1	MS-ESS1-1: Space Systems	5	0	4	9
		MS-ESS1-2: Space Systems	3	1	1	5
		MS-ESS1-3: Space Systems	2	0	5	7
		MS-ESS1-4: History of Earth	2	1	5	8
	ESS2	MS-ESS2-1: Earth's Systems	0	0	5	5
		MS-ESS2-2: History of Earth	2	0	5	7
		MS-ESS2-3: History of Earth	1	0	6	7
		MS-ESS2-4: Earth's Systems	1	0	4	5
		MS-ESS2-5: Weather & Climate	1	0	6	7
		MS-ESS2-6: Weather & Climate	0	0	2	2
	ESS3	MS-ESS3-1: Earth's Systems	3	0	3	6
		MS-ESS3-2: Human Impacts	2	0	7	9
		MS-ESS3-3: Human Impacts*	0	0	1	1
		MS-ESS3-4: Human Impacts	1	1	9	11
		MS-ESS3-5: Weather & Climate	3	0	3	6
Engineering and Technology	ETS1	MS-ETS1-1: Engineering Design	0	0	1	1
Life Sciences	LS1	MS-LS1-1: Structure, Function, Information Processing	0	0	2	2
		MS-LS1-2: Structure, Function, Information Processing	0	1	5	6
		MS-LS1-3: Structure, Function, Information Processing	0	0	3	3

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		MS-LS1-4: Growth, Development, Reproduction	2	0	6	8
		MS-LS1-5: Growth, Development, Reproduction	0	1	3	4
		MS-LS1-6: Matter & Energy	1	0	4	5
		MS-LS1-7: Matter & Energy	1	0	3	4
		MS-LS1-8: Structure, Function, Information Processing	1	0	5	6
	LS2	LS2-MS-4: Cycle of Matter and Energy Transfer in Ecosystems	0	0	1	1
		MS-LS2-1: Matter & Energy	4	0	5	9
		MS-LS2-2: Interdependent Relationships in Ecosystems	3	0	3	6
		MS-LS2-3: Matter & Energy	1	1	3	5
		MS-LS2-4: Matter & Energy	5	0	6	11
		MS-LS2-5: Interdependent Relationships in Ecosystems*	2	1	4	7
	LS3	MS-LS3-1: Growth, Development, Reproduction	0	0	3	3
		MS-LS3-2: Growth, Development, Reproduction	2	0	7	9
	LS4	LS4-MS-3: Classification of Organisms	0	0	1	1
		MS-LS4-1: Natural Selection & Adaptation	4	0	4	8
		MS-LS4-2: Natural Selection & Adaptation	0	0	6	6
		MS-LS4-3: Natural Selection & Adaptation	1	0	3	4
		MS-LS4-4: Natural Selection & Adaptation	2	0	4	6
		MS-LS4-5: Growth, Development, Reproduction	1	1	2	4
		MS-LS4-6: Natural Selection & Adaptation	1	0	1	2
Physical Sciences	PS1	MS-PS1-1: Structure & Properties of Matter	1	0	3	4
		MS-PS1-2: Chemical Reactions	3	1	6	10

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		MS-PS1-3: Structure & Properties of Matter	0	1	3	4
		MS-PS1-4: Structure & Properties of Matter	1	0	10	11
		MS-PS1-5: Chemical Reactions	0	0	6	6
		MS-PS1-6: Chemical Reactions*	1	1	2	4
	PS2	MS-PS2-1: Forces & Interactions*	2	0	4	6
		MS-PS2-2: Forces & Interactions	0	0	5	5
		MS-PS2-3: Forces & Interactions	1	0	2	3
		MS-PS2-4: Forces & Interactions	0	0	4	4
		MS-PS2-5: Forces & Interactions	0	1	6	7
	PS3	MS-PS3-1: Energy	2	1	1	4
		MS-PS3-2: Energy	1	0	5	6
		MS-PS3-3: Energy*	0	0	3	3
		MS-PS3-4: Energy	1	0	3	4
		MS-PS3-5: Energy	4	0	4	8
	PS4	MS-PS4-1: Waves & Electromagnetic Radiation	1	0	6	7
		MS-PS4-2: Waves & Electromagnetic Radiation	4	0	4	8
		MS-PS4-3: Waves & Electromagnetic Radiation	0	0	1	1

Note. *These PEs have an engineering component.

[a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Table H-3. Spring 2019 Shared Science Assessment Operational and Field-Test Item Bank by Performance Expectation, High School

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
Earth and Space Sciences	ESS1	HS-ESS1-1: The Universe and Its Stars	1	0	2	3
		HS-ESS1-2: The Universe and Its Stars	1	0	1	2
		HS-ESS1-3: The Universe and Its Stars	1	1	1	3
		HS-ESS1-4: Earth and the Solar System	1	0	1	2
		HS-ESS1-5: The History of Planet Earth	1	0	4	5
		HS-ESS1-6: The History of Planet Earth	1	1	1	3
	ESS2	HS-ESS2-1: Earth Materials and Systems	0	1	2	3
		HS-ESS2-2: Earth Materials and Systems	2	1	2	5
		HS-ESS2-3: Earth Materials and Systems	1	0	0	1
		HS-ESS2-4: Weather and Climate	0	0	3	3
		HS-ESS2-5: The Roles of Water in Earth's Surface Processes	0	0	0	0
		HS-ESS2-6: Weather and Climate	1	1	0	2
		HS-ESS2-7: Weather and Climate	1	0	1	2
	ESS3	HS-ESS3-1: Natural Resources	2	0	1	3
		HS-ESS3-2: Natural Resources*	1	0	0	1
		HS-ESS3-3: Human Impacts on Earth Systems	0	1	1	2
		HS-ESS3-4: Human Impacts on Earth Systems*	0	0	0	0
		HS-ESS3-5: Global Climate Change	2	0	2	4
		HS-ESS3-6: Global Climate Change*	1	0	1	2
Life Sciences	LS1	HS-LS1-1: Structure and Function	0	0	4	4
		HS-LS1-2: Structure and Function	2	0	6	8

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		HS-LS1-3: Structure and Function	0	0	3	3
		HS-LS1-4: Growth and Development of Organisms	4	0	2	6
		HS-LS1-5: Organization for Matter and Energy Flow in Organisms	0	1	4	5
		HS-LS1-6: Organization of Matter and Energy Flow in Organisms	3	0	1	4
		HS-LS1-7: Organization of Matter and Energy Flow in Organisms	2	0	3	5
	LS2	HS-LS2-1: Interdependent Relationships in Ecosystems	2	0	4	6
		HS-LS2-2: Interdependent Relationships in Ecosystems	1	0	1	2
		HS-LS2-3: Cycles of Matter and Energy Transfer in Ecosystems	2	0	3	5
		HS-LS2-4: Cycles of Matter and Energy Transfer in Ecosystems	5	1	3	9
		HS-LS2-5: Cycles of Matter and Energy Transfer in Ecosystems	0	0	2	2
		HS-LS2-6: Ecosystem Dynamics, Functioning, and Resilience	2	0	3	5
		HS-LS2-7: Ecosystem Dynamics, Functioning and Resilience*	2	0	3	5
		HS-LS2-8 Social Interactions and Group Behavior	1	0	2	3
	LS3	HS-LS3-1: Structure and Function	2	0	1	3
		HS-LS3-2: Variation of Traits	3	0	1	4
		HS-LS3-3: Variation of Traits	3	0	2	5
	LS4	HS-LS4-1: Evidence of Common Ancestry and Diversity	6	0	3	9
		HS-LS4-2: Natural Selection	3	0	3	6
		HS-LS4-3: Natural Selection	2	0	3	5
		HS-LS4-4: Adaptation	2	1	3	6
		HS-LS4-5: Adaptation	4	0	4	8

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		HS-LS4-6: Adaptation*	0	0	1	1
Physical Sciences	PS1	HS-PS1-1: Structure and Properties of Matter	1	0	2	3
		HS-PS1-2: Structure and Properties of Matter	2	0	2	4
		HS-PS1-3: Structure and Properties of Matter	2	1	3	6
		HS-PS1-4: Chemical Reactions	0	0	1	1
		HS-PS1-5: Chemical Reactions	0	0	2	2
		HS-PS1-6: Chemical Reactions*	1	0	2	3
		HS-PS1-7: Chemical Reactions	2	0	1	3
		HS-PS1-8: Nuclear Processes	0	1	2	3
	PS2	HS-PS2-1: Forces and Motion	2	0	4	6
		HS-PS2-2: Forces and Motion	1	1	3	5
		HS-PS2-3: Forces and Motion*	0	0	1	1
		HS-PS2-4: Types of Interactions	1	0	1	2
		HS-PS2-5: Types of Interactions	0	0	1	1
		HS-PS2-6: Chemical Reactions*	0	0	0	0
	PS3	HS-PS3-1: Energy	1	0	0	1
		HS-PS3-2: Energy	1	0	3	4
		HS-PS3-3: Energy*	0	0	3	3
		HS-PS3-4: Energy	0	0	4	4
		HS-PS3-5: Energy	2	0	1	3
	PS4	HS-PS4-1: Wave Properties	0	0	2	2
		HS-PS4-2: Wave Properties	0	0	2	2
		HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	0	0	0

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		HS-PS4-4: Electromagnetic Radiation	0	0	1	1
		HS-PS4-5: Electromagnetic Radiation*	0	0	0	0

Note. *These PEs have an engineering component.

[a]Other MOU states include Connecticut, Hawaii, Idaho, Oregon, Utah, West Virginia, and Wyoming.

Appendix I

Multi-State Science Assessment (MSSA) Item Pool

Multi-State Science Assessment Item Pool

Table I-1. Spring 2019 MSSA Operational and Field-Test Item Pool by Performance Expectation, Grade 5

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
Earth and Space Sciences	ESS1	4-ESS1-1: Earth's Systems & Processes	1	0	2	3
		5-ESS1-1: Space Systems	0	2	1	3
		5-ESS1-2: Space Systems	2	0	1	3
	ESS2	3-ESS2-1: Weather & Climate	2	1	2	5
		3-ESS2-2: Weather & Climate	2	0	2	4
		4-ESS2-1: Earth's Systems & Processes	1	0	1	2
		4-ESS2-2: Earth's Systems & Processes	1	0	2	3
		5-ESS2-1: Earth's Systems	0	1	2	3
		5-ESS2-2: Earth's Systems	1	1	1	3
	ESS3	3-ESS3-1: Weather & Climate*	0	0	0	0
		4-ESS3-1: Energy	0	0	0	0
		4-ESS3-2: Earth's Systems & Processes*	0	2	0	2
		5-ESS3-1: Earth's Systems	1	0	1	2
Life Sciences	LS1	3-LS1-1: Inheritance	0	2	2	4
		4-LS1-1: Structure, Function, Information Processing	2	0	2	4
		4-LS1-2: Structure, Function, Information Processing	1	1	3	5
		5-LS1-1: Matter & Energy	1	0	4	5
	LS2	3-LS2-1: Ecosystems	3	0	2	5
		5-LS2-1: Matter & Energy	1	0	1	2
	LS3	3-LS3-1: Inheritance	0	2	2	4
		3-LS3-2: Inheritance	1	1	2	4
	LS4	3-LS4-1: Ecosystems	2	0	2	4

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		3-LS4-2: Inheritance	3	0	2	5
		3-LS4-3: Ecosystems	1	0	1	2
		3-LS4-4: Ecosystems*	0	0	1	1
Physical Sciences	PS1	5-PS1-1: Structure & Properties of Matter	1	0	1	2
		5-PS1-2: Structure & Properties of Matter	2	1	2	5
		5-PS1-3: Structure & Properties of Matter	2	0	0	2
		5-PS1-4: Structure & Properties of Matter	0	1	0	1
	PS2	3-PS2-1: Forces-balanced and unbalanced forces	0	1	3	4
		3-PS2-2: Forces-pattern predicts future motion	2	0	2	4
		3-PS2-3: Forces-between objects not in contact	1	0	2	3
		3-PS2-4: Forces-magnets*	0	0	0	0
		5-PS2-1: Space systems	0	0	1	1
	PS3	4-PS3-1: Energy-relationship between speed and energy of object	2	0	5	7
		4-PS3-2: Energy-transfer of energy	3	0	1	4
		4-PS3-3: Energy-changes in energy when objects collide	3	0	3	6
		4-PS3-4: Energy-converting energy from one form to another*	0	0	2	2
		5-PS3-1: Matter & Energy	2	1	0	3
	PS4	4-PS4-1: Waves-waves can cause objects to move	0	0	1	1
		4-PS4-2: Structure, function, information processing	0	0	5	5
		4-PS4-3: Waves-using patterns to transfer information*	0	0	0	0

Note. *These PEs have an engineering component.

[a]Other MOU states include Connecticut, Hawaii, Oregon, Utah, West Virginia, and Wyoming.

Table I-2. Spring 2019 MSSA Operational and Field-Test Item Pool by Performance Expectation, Grade 8

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
Earth and Space Sciences	ESS1	MS-ESS1-1: Space Systems	1	0	2	3
		MS-ESS1-2: Space Systems	0	1	1	2
		MS-ESS1-3: Space Systems	2	0	2	4
		MS-ESS1-4: History of Earth	1	1	2	4
	ESS2	MS-ESS2-1: Earth's Systems	0	0	2	2
		MS-ESS2-2: History of Earth	2	0	2	4
		MS-ESS2-3: History of Earth	1	0	4	5
		MS-ESS2-4: Earth's Systems	1	0	2	3
		MS-ESS2-5: Weather & Climate	1	0	1	2
		MS-ESS2-6: Weather & Climate	0	0	1	1
	ESS3	MS-ESS3-1: Earth's Systems	1	0	0	1
		MS-ESS3-2: Human Impacts	1	0	3	4
		MS-ESS3-3: Human Impacts*	0	0	0	0
		MS-ESS3-4: Human Impacts	1	1	1	3
		MS-ESS3-5: Weather & Climate	3	0	0	3
Life Sciences	LS1	MS-LS1-1: Structure, Function, Information Processing	0	0	0	0
		MS-LS1-2: Structure, Function, Information Processing	0	0	1	1
		MS-LS1-3: Structure, Function, Information Processing	0	0	1	1
		MS-LS1-4: Growth, Development, Reproduction	2	0	3	5
		MS-LS1-5: Growth, Development, Reproduction	0	1	0	1
		MS-LS1-6: Matter & Energy	0	0	1	1
		MS-LS1-7: Matter & Energy	1	0	2	3

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		MS-LS1-8: Structure, Function, Information Processing	1	0	1	2
	LS2	MS-LS2-1: Matter & Energy	3	0	3	6
		MS-LS2-2: Interdependent Relationships in Ecosystems	0	0	0	0
		MS-LS2-3: Matter & Energy	0	1	0	1
		MS-LS2-4: Matter & Energy	1	0	3	4
		MS-LS2-5: Interdependent Relationships in Ecosystems*	1	1	0	2
	LS3	MS-LS3-1: Growth, Development, Reproduction	0	0	0	0
		MS-LS3-2: Growth, Development, Reproduction	2	0	1	3
	LS4	MS-LS4-1: Natural Selection & Adaptation	1	0	2	3
		MS-LS4-2: Natural Selection & Adaptation	0	0	1	1
		MS-LS4-3: Natural Selection & Adaptation	1	0	0	1
		MS-LS4-4: Natural Selection & Adaptation	2	0	0	2
		MS-LS4-5: Growth, Development, Reproduction	1	1	0	2
		MS-LS4-6: Natural Selection & Adaptation	0	0	0	0
Physical Sciences	PS1	MS-PS1-1: Structure & Properties of Matter	1	0	1	2
		MS-PS1-2: Chemical Reactions	1	1	2	4
		MS-PS1-3: Structure & Properties of Matter	0	1	1	2
		MS-PS1-4: Structure & Properties of Matter	0	0	0	0
		MS-PS1-5: Chemical Reactions	0	0	3	3
		MS-PS1-6: Chemical Reactions*	0	1	0	1
	PS2	MS-PS2-1: Forces & Interactions*	1	0	0	1
		MS-PS2-2: Forces & Interactions	0	0	2	2
		MS-PS2-3: Forces & Interactions	1	0	0	1

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		MS-PS2-4: Forces & Interactions	0	0	0	0
		MS-PS2-5: Forces & Interactions	0	1	1	2
	PS3	MS-PS3-1: Energy	1	1	0	2
		MS-PS3-2: Energy	0	0	1	1
		MS-PS3-3: Energy*	0	0	1	1
		MS-PS3-4: Energy	0	0	1	1
		MS-PS3-5: Energy	4	0	2	6
	PS4	MS-PS4-1: Waves & Electromagnetic Radiation	0	0	3	3
		MS-PS4-2: Waves & Electromagnetic Radiation	2	0	2	4
		MS-PS4-3: Waves & Electromagnetic Radiation	0	0	0	0

Note. *These PEs have an engineering component.

[a]Other MOU states include Connecticut, Hawaii, Oregon, Utah, West Virginia, and Wyoming.

Table I-3. Spring 2019 MSSA Operational and Field-Test Item Pool by Performance Expectation, Grade 11

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
Earth and Space Sciences	ESS1	HS-ESS1-1: The Universe and Its Stars	0	0	2	2
		HS-ESS1-2: The Universe and Its Stars	1	0	0	1
		HS-ESS1-3: The Universe and Its Stars	1	1	0	2
		HS-ESS1-4: Earth and the Solar System	0	0	0	0
		HS-ESS1-5: The History of Planet Earth	1	0	3	4
		HS-ESS1-6: The History of Planet Earth	1	1	0	2
	ESS2	HS-ESS2-1: Earth Materials and Systems	0	1	0	1
		HS-ESS2-2: Earth Materials and Systems	2	1	2	5
		HS-ESS2-3: Earth Materials and Systems	1	0	0	1
		HS-ESS2-4: Weather and Climate	0	0	3	3
		HS-ESS2-5: The Roles of Water in Earth's Surface Processes	0	0	0	0
		HS-ESS2-6: Weather and Climate	1	1	0	2
		HS-ESS2-7: Weather and Climate	1	0	1	2
	ESS3	HS-ESS3-1: Natural Resources	0	0	0	0
		HS-ESS3-2: Natural Resources*	1	0	0	1
		HS-ESS3-3: Human Impacts on Earth Systems	0	1	1	2
		HS-ESS3-4: Human Impacts on Earth Systems*	0	0	0	0
		HS-ESS3-5: Global Climate Change	1	0	2	3
		HS-ESS3-6: Global Climate Change*	1	0	1	2
Life Sciences	LS1	HS-LS1-1: Structure and Function	0	0	1	1
		HS-LS1-2: Structure and Function	1	0	2	3
		HS-LS1-3: Structure and Function	0	0	3	3
		HS-LS1-4: Growth and Development of Organisms	2	0	2	4

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		HS-LS1-5: Organization for Matter and Energy Flow in Organisms	0	1	1	2
		HS-LS1-6: Organization for Matter and Energy Flow in Organisms	2	0	0	2
		HS-LS1-7: Organization for Matter and Energy Flow in Organisms	1	0	0	1
	LS2	HS-LS2-1: Interdependent Relationships in Ecosystems	1	0	2	3
		HS-LS2-2: Interdependent Relationships in Ecosystems	1	0	1	2
		HS-LS2-3: Cycles of Matter and Energy Transfer in Ecosystems	0	0	0	0
		HS-LS2-4: Cycles of Matter and Energy Transfer in Ecosystems	3	1	1	5
		HS-LS2-5: Cycles of Matter and Energy Transfer in Ecosystems	0	0	1	1
		HS-LS2-6: Ecosystem Dynamics, Functioning and Resilience	2	0	1	3
		HS-LS2-7: Ecosystem Dynamics, Functioning and Resilience*	0	0	0	0
		HS-LS2-8: Social Interactions and Group Behavior	1	0	0	1
	LS3	HS-LS3-1: Structure and Function	0	0	1	1
		HS-LS3-2: Variation of Traits	1	0	1	2
		HS-LS3-3: Variation of Traits	1	0	1	2
	LS4	HS-LS4-1: Evidence of Common Ancestry and Diversity	3	0	1	4
		HS-LS4-2: Natural Selection	2	0	1	3
		HS-LS4-3: Natural Selection	0	0	2	2
		HS-LS4-4: Adaptation	2	1	2	5
		HS-LS4-5: Adaptation	0	0	2	2
		HS-LS4-6: Adaptation*	0	0	0	0
Physical Sciences	PS1	HS-PS1-1: Structure and Properties of Matter	1	0	1	2
		HS-PS1-2: Structure and Properties of Matter	2	0	1	3

Science Discipline	Disciplinary Core Idea	Performance Expectation	AIRCore Items	MSSA Items	MOU Items[a]	Total Items
		HS-PS1-3: Structure and Properties of Matter	2	1	1	4
		HS-PS1-4: Chemical Reactions	0	0	0	0
		HS-PS1-5: Chemical Reactions	0	0	0	0
		HS-PS1-6: Chemical Reactions*	1	0	1	2
		HS-PS1-7: Chemical Reactions	2	0	1	3
		HS-PS1-8: Nuclear Processes	0	1	1	2
	PS2	HS-PS2-1: Forces and Motion	0	0	3	3
		HS-PS2-2: Forces and Motion	1	1	2	4
		HS-PS2-3: Forces and Motion*	0	0	0	0
		HS-PS2-4: Types of Interactions	1	0	1	2
		HS-PS2-5: Types of Interactions	0	0	1	1
		HS-PS2-6: Chemical Reactions*	0	0	0	0
	PS3	HS-PS3-1: Energy	1	0	0	1
		HS-PS3-2: Energy	1	0	2	3
		HS-PS3-3: Energy*	0	0	1	1
		HS-PS3-4: Energy	0	0	2	2
		HS-PS3-5: Energy	2	0	1	3
	PS4	HS-PS4-1: Wave Properties	0	0	2	2
		HS-PS4-2: Wave Properties	0	0	0	0
		HS-PS4-3: Wave Properties/Electromagnetic Radiation	0	0	0	0
		HS-PS4-4: Electromagnetic Radiation	0	0	0	0
		HS-PS4-5: Electromagnetic Radiation*	0	0	0	0

Note. *These PEs have an engineering component.

[a]Other MOU states include Connecticut, Hawaii, Oregon, Utah, West Virginia, and Wyoming.

Appendix J

Adaptive Algorithm Design

TABLE OF CONTENTS

Adaptive Item Selection Algorithm

1. INTRODUCTION, BACKGROUND, AND DEFINITIONS

This document describes the adaptive item selection algorithm. The item selection algorithm is designed to cover a standards-based blueprint, which may include content, cognitive complexity, and item type constraints. The item selection algorithm will also include:

- the ability to customize an item pool based on access constraints and screen items that have been previously viewed or may not be accessible for a given individual;
- a mechanism for inserting embedded field-test items; and
- a mechanism for delivering "segmented" tests in which separate parts of the test are administered in a fixed order.

This document describes the algorithm and the design for its implementation for the test delivery system (TDS). The implementation builds extensively on the algorithm implemented in the American Institutes for Research (AIR)'s TDS and incorporates substantial AIR intellectual property. AIR will release the algorithm and the implementation described here under the same open-source license under which the rest of the open-source system is released.

The general approach described here is based on a highly parameterized multiple-objective utility function. The objective function includes:

- a measure of content match to the blueprint;
- a measure of overall test information; and
- measures of test information for each reporting category on the test.

We define an objective function that measures an item's contribution to each of these objectives, weighting them to achieve the desired balance among them. Equation (1) sketches this objective function for a single item.

$$f_{ijt} = w_2 \frac{1}{\sum_{r=1}^{R} d_{rj}} \sum_{r=1}^{R} s_{rit} p_r d_{rj} + w_1 \sum_{k=1}^{K} q_k h_{1k}(v_{kijt}, V_{kit}, t_k) + w_0 h_0(u_{ijt}, U_{it}, t_0) \qquad (1)$$

where the term w represents user-supplied weights that assign relative importance to meeting each of the objectives d_{rj} indicates whether item j has the blueprint-specified feature r, and p_r is the user-supplied priority weight for feature r. The term s_{rit} is an adaptive control parameter that is described. In general, s_{rit} increases for features that have not met their designated minimum as the end of the test approaches.

The remainder of the terms represents an item's contribution to measurement precision:

- v_{kijt} is the value of item j toward reducing the measurement error for reporting category k for examinee i at selection t; and
- u_{ijt} is the value of item j in terms of reducing the overall measurement error for examinee i at selection t.

The terms U_{it} and V_{kit} represent the total information overall and on reporting category k, respectively.

The term q_k is a user-supplied priority weight associated with the precision of the score estimate for reporting category k. The terms t represent precision targets for the overall score (t_0) and each score reporting category score. The functions $h(.)$ are given by:

$$h_0(u_{ijt}, U_{it}, t_0) = \begin{cases} au_{ijt} \text{ if } U_{it} < t_0 \\ bu_{ijt} \text{ otherwise} \end{cases}$$

$$h_{1k}(v_{kijt}, V_{kit}, t_k) = \begin{cases} c_k v_{kijt} \text{ if } V_{kit} < t_k \\ d_k v_{kijt} \text{ otherwise} \end{cases}$$

Items can be selected to maximize the value of this function. This objective function can be manipulated to produce a pure, standards-free adaptive algorithm by setting w_2 to zero or a completely blueprint-driven test by setting $w_1 = w_0 = 0$. Adjusting the weights to optimize performance for a given item pool will enable users to maximize information subject to the constraint that the blueprint is virtually always met.

We note that the computations of the content values and information values generate values on very different scales, and that the scale of the content value varies as the test progresses. Therefore, we normalize both the information and content values before computing the value of Equation (1). This normalization is given by $x = \begin{cases} 1 \; if \; min = max \\ \frac{v-min}{max-min} \; otherwise \end{cases}$, where min and max represent the minimum and maximum, respectively, of the metric computed over the current set of items or item groups.

The remainder of this section describes the overall program flow, the form of the blueprint, and the various value calculations employed in the objective function. Subsequent sections describe the details of the selection algorithm.

1.1 BLUEPRINT

Each test will be described by a single blueprint for each segment of the test and will identify the order in which the segments appear. The blueprint will include:

- an indicator of whether the test is adaptive or fixed form;
- termination conditions for the segment, which are described in a subsequent section;
- a set of nested content constraints, each of which is expressed as:

  - the minimum number of items to be administered within the content category;
  - the maximum number of items to be administered within the content category;
  - an indication of whether the maximum should be deterministically enforced (a "strict" maximum);
  - a priority weight for the content category p_r;
  - an explicit indicator as to whether this content category is a reporting category; and
  - an explicit precision-priority weight (q_k) for each group identified as a reporting category.

- a set of non-nested content constraints, which are represented as:
  - a name for the collection of items meeting the constraint;
  - the minimum number of items to be administered from this group of items;
  - the maximum number of items to be administered from this group of items;
  - an indication of whether the maximum should be deterministically enforced (a "strict" maximum);
  - a priority weight for the group of items p_r;
  - an explicit indicator as to whether this named group will make up a reporting category; and
  - an explicit precision-priority weight (q_k) for each group identified as a reporting category.
  - The priority weights, p_r on the blueprint, can be used to express values in the blueprint match. Large weights on reporting categories paired with low (or zero) weights on the content categories below them may allow more flexibility to maximize information in a content category covering fewer fine-grained targets, while the reverse would mitigate toward more reliable coverage of finer-grained categories, with less content flexibility within reporting categories.

An example of a blueprint specification appears in Appendix J-1.

1.2 CONTENT VALUE

Each item or item group will be characterized by its contribution to meeting the blueprint, given the items that have already been administered at any point. The contribution is based on the presence or absence of features specified in the blueprint and denoted by the term d in Equation (1). This section describes the computation of the content value.

1.2.1 Content Value for Single Items

For each constraint appearing in the blueprint (r), an item i either does or does not have the characteristic described by the constraint. For example, a constraint might require a minimum of four and a maximum of six algebra items. An item measuring algebra has the described characteristic, and an item measuring geometry, but algebra does not. To capture this constraint, we define the following:

- d_j is a feature vector in which the elements are d_{rj}, summarizing item j's contribution to meeting the blueprint. This feature vector includes content categories such as claims and targets as well as other features of the blueprint, such as Depth of Knowledge (DOK) and item type.
- S_{it} is a diagonal matrix, the diagonal elements of which are the adaptive control parameters s_{rit}.
- p is the vector containing the user-supplied priority weights p_r.

The scalar content value for an item is given by $C_{ijt} = d_j^{'} S_{it} p$.

Letting z_{rit} represent the number of items with feature r administered to student i by iteration t, the value of the adaptive control parameters is:

$$s_{rit} = \begin{cases} m_{it}\left(2 - \dfrac{z_{rit}}{Min_r}\right) & if\ z_r < Min_r \\ 1 - \dfrac{z_{rit} - Min_r}{Max_r - Min_r} & if\ Min_r < z_{rit} < Max_r \\ \left(Max_r - z_{rit}\right) - 1 & if\ Max_r \le z_{rit} \end{cases}$$

The blueprint defines the minimum (Min_r) and maximum (Max_r) number of items to be administered with each characteristic (r).

The term $m_{it} = \frac{T}{T-t}$ where T is the total test length. This has the effect of increasing the algorithm's preference for items that have not yet met their minimums as the end of the test nears and the opportunities to meet the minimum diminish.

This increases the likelihood of selecting items for content that has not met its minimum as the opportunities to do so are used up. The value s is highest for items with content that has not met its minimum, declines for items representing content for which the minimum number of items has been reached but the maximum has not, and turns negative for items representing content that has met the maximum.

1.2.2 Content Value for Sets of Items

Calculation of the content value of sets of items is complicated by two factors:

1. The desire to allow more items to be developed for each set and to have the most advantageous set of items administered.

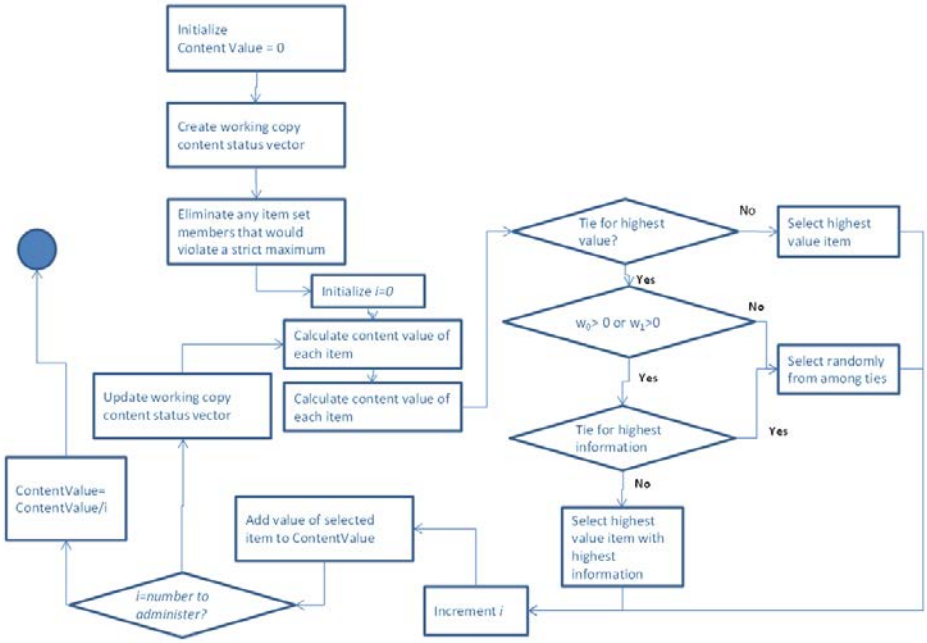
2. The design objective of characterizing the information contribution of a set of items as the expected information over the working theta distribution for the examinee.

The former objective is believed to enhance the ability to satisfy highly constrained blueprints while still adapting to obtain good measurement for a broad range of students. The latter arises from the recognition that English Language Arts (ELA) tests will select one set of items at a time, without an opportunity to adapt once the passage has been selected.

The general approach involves successive selection of the highest content value item in the set until the indicated number of items in the set have been selected. Because the content value of an item changes with each selection, a temporary copy of the already-administered content vector for the examinee is updated with each selection such that subsequent selections reflect the items selected in previous iterations.

Exhibit A on the following page presents a flowchart for this calculation. Readers will note the check to determine whether $w_0 > 0$ or $w_1 > 0$. These weights, defined with Equation (1), identify the user-supplied importance of information optimization relative to blueprint optimization. In cases such as independent field tests, this weight may be set to zero, as it may not be desirable to make item administration dependent on the match to student performance. In more typical adaptive cases where item statistics will not be recalculated, favoring more informative items is generally better. The final measure of content value for the set of selected set of items is divided by the number of items selected to avoid a bias toward selection of sets with more items.

Exhibit A. Content Value Calculation for Item Sets



1.3 INFORMATION VALUE

Each item or item group also has value in terms of maximizing information, both overall and on reporting categories.

1.3.1 Individual Information Value

The information value associated with an item will be an approximation of information. The system will be designed to use generalized Item Response Theory (IRT) models; however, it will treat all items as though they offer equal measurement precision. This is the assumption made by the Rasch model, but in more general models, items known to offer better measurement are given preference by many algorithms. Subsequent algorithms are then required to control the exposure of the items that measure best. Ignoring the differences in slopes serves to eliminate this bias and help equalize exposure.

1.3.2 Binary Items

The approximate information value of a binary item will be characterized as $I_j(\theta) = p_j(\theta)(1 - p_j(\theta))$, where the slope parameters are artificially replaced with a constant.

1.3.3 Polytomous Items

In terms of information, the best polytomous item in the pool is the one that maximizes the expected information, $I_j(\theta)$. Formally, $I_j(\theta) > I_k(\theta)$ for all items $k \neq j$. The true value θ,

however, remains unknown and is accessed only through an estimate, $\hat{\theta} \sim N(\bar{\theta}, \sigma_\theta)$. By definition of an expectation, the expected information $I_j(\theta) = \int I_j(t) f(t|\bar{\theta}, \sigma_\theta) dt$.

The intuition behind this result is illustrated in Exhibit B. In Exhibit B, each panel graphs the distribution of the estimate of θ for an examinee. The top panel assumes a polytomous item in which one step threshold (A1) matches the mean of the θ estimate distribution. In the bottom panel, neither step threshold matches the mean of the θ estimate distribution. The shaded area in each panel indicates the region in which the hypothetical item depicted in the panel provides more information. We see that approximately 2/3 of the probability density function is shaded in the lower panel, while the item depicted in the upper panel dominates in only about 1/3 of the cases. In this example, the item depicted in the lower panel has a much greater probability of maximizing the information from the item, despite the fact that the item in the upper panel has a threshold exactly matching the mean of the estimate distribution and the item in the lower panel does not.

Exhibit B. Two Example Items, with the Shaded Region Showing the Probability that the Item Maximizes Information for the Examinee Depicted

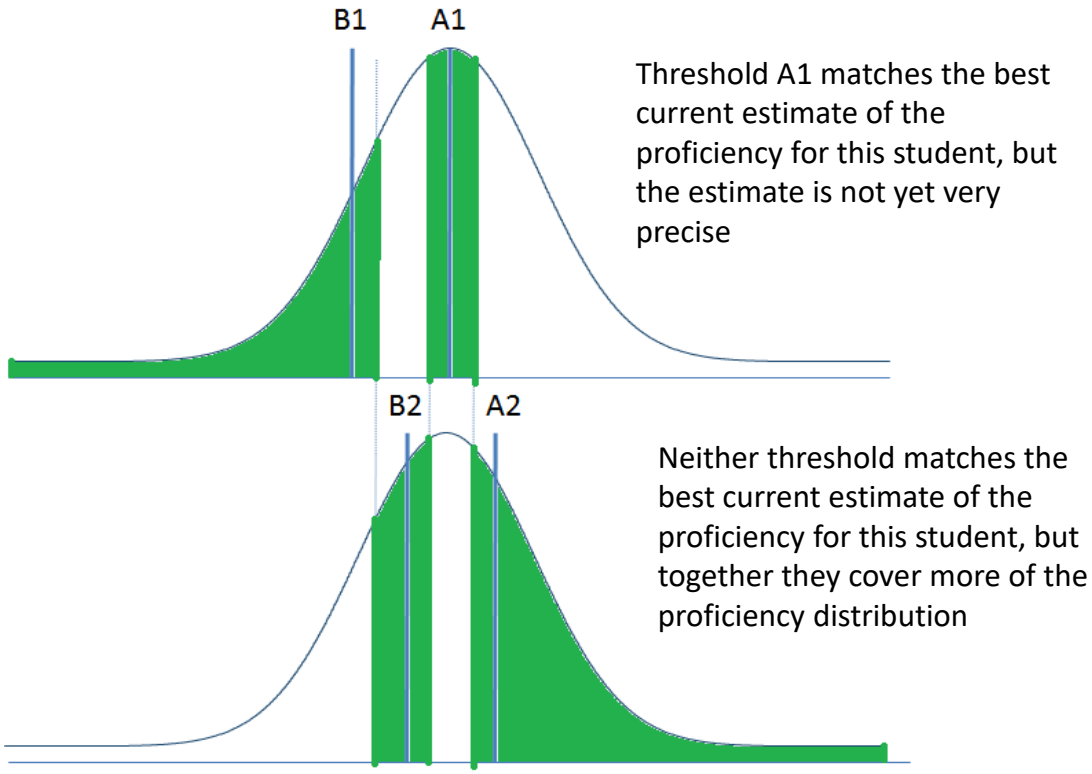
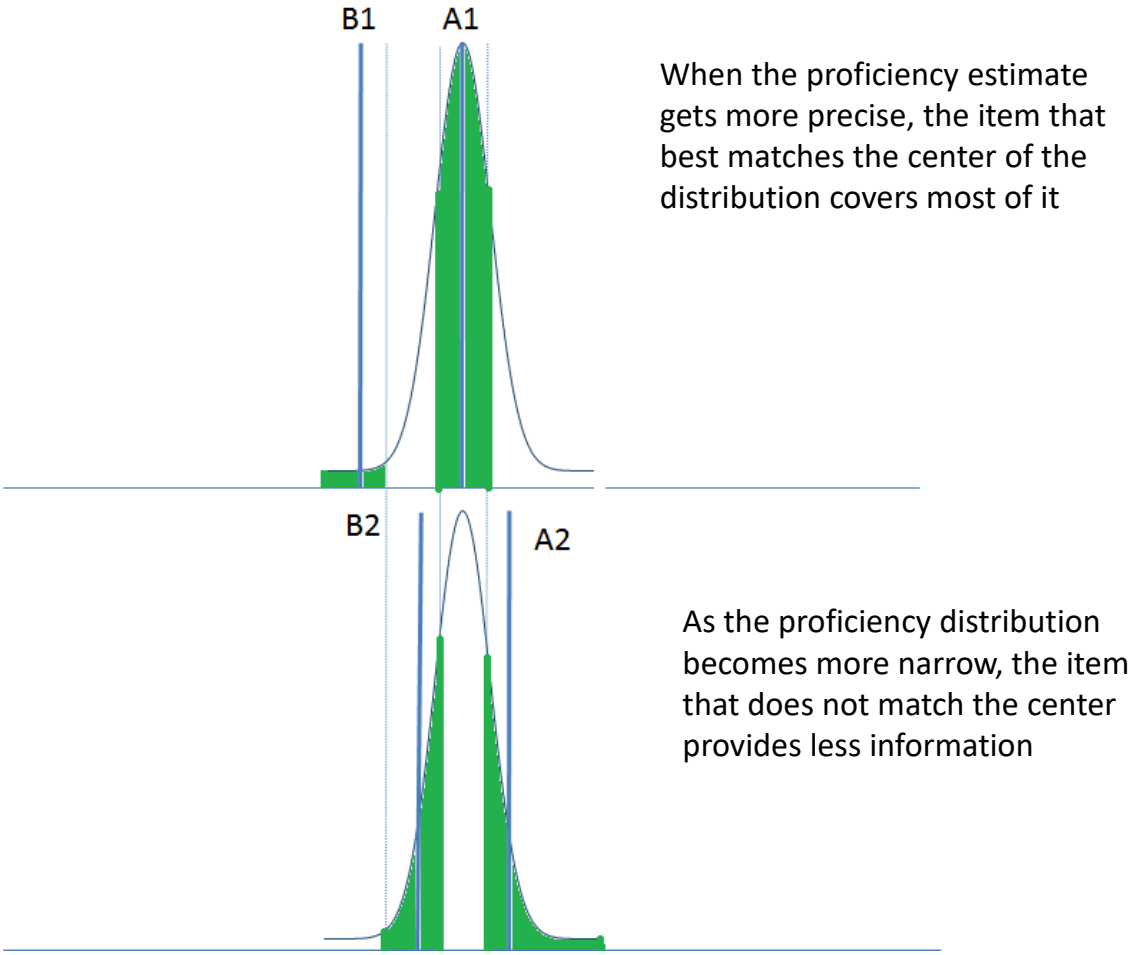


Exhibit C on the following page shows what happens to information as the estimate of this student's proficiency becomes more precise (later in the test). In this case, the item depicted in the top panel maximizes information about 65 to70 percent of the time, compared to about 30 to 35 percent for the item depicted in the lower panel. These are the same items depicted in the Exhibit B, but in this case, we are considering information for a student with a more precise current proficiency estimate.

Exhibit C. Two Example Items, with the Shaded Region Showing the Probability that the Item Maximizes Information for the Examinee Depicted



The approximate information value of polytomous items will be characterized as the expected information, specifically $E\left[I_j(\theta)|m_i, s_i\right] = \int \sum_{k=1}^{K} I_{jk}(t)\, p_j(k|t)\phi(t; m_i, s_i)dt$, where $I_{jk}(t)$ represents the information at t of response k to item j, $p_j(k|t)$ is the probability of response k to item j (artificially holding slope constant), given proficiency t, $\phi(.)$ represents the normal probability density function, and m_i and s_i represent the mean and standard deviation of examinee i's current estimated proficiency distribution.

We propose to use Gauss-Hermite quadrature with a small number of quadrature points (approximately five). Experiments show that we can complete this calculation for 1,000 items in fewer than 5 milliseconds, making it computationally reasonable.

As with the binary items, we propose to ignore the slope parameters to even exposure and avoid a bias toward the items with better measurement.

1.3.4 Item Group Information Value

Item groups differ from individual items in that a set of items will be selected for administration. Therefore, the goal is to maximize information across the working theta distribution. As with the polytomous items, we propose to use Gauss-Hermite quadrature to estimate the expected information of the item group.

In the case of multiple-item groups

$$E\left[I_g(\theta)|m_i, s_i\right] = \frac{1}{J_g}\int \sum_{j=1}^{J_g} I_{g(j)}(t)\,\phi(t; m_i, s_i)dt$$

Where $I_g(.)$ is the information from item group g, $I_{g(j)}$ is the information associated with item $j \in g$, for the J_g items in set g. In the case of polytomous items, we use the expected information, as described above.

2. ENTRY AND INITIALIZATION

At startup, the system will

- create a custom item pool;
- initialize theta estimates for the overall score and each score point; and
- insert embedded field-test items.

2.1 ITEM POOL

At test startup, the system will generate a *custom item pool*, a string of item IDs for which the student is eligible. This item pool will include all items that

- are active in the system at test startup; and
- are not flagged as "access limited" for attributes associated with this student.

The list will be stored in ascending order of ID.

2.2 ADJUST SEGMENT LENGTH

Custom item pools run the risk of being unable to meet segment blueprint minimums. To address this special case, the algorithm will adjust the blueprint to be consistent with the custom item pool. This capability becomes necessary when an accommodated item pool systematically excludes some content.

Let

> $\boldsymbol{S}$ be the set of top-level content constraints in the hierarchical set of constraints, each consisting of the tuple (*name*, *min*, *max*, *n*);
>
> $\boldsymbol{C}$ be the custom item pool, each element consisting of a set of content constraints $\boldsymbol{B}$;
>
> $\boldsymbol{f}$, $\boldsymbol{p}$ integers represent item shortfall and pool count, respectively; and
>
> $\boldsymbol{t}$ be the minimum required items on the segment.

For each s in $\boldsymbol{S}$, compute n as the sum of active operational items in $\boldsymbol{C}$ classified on the constraint.

> $\boldsymbol{f}$ = summation over S $(min - n)$
>
> $\boldsymbol{p}$ = summation over S (n)
>
> if $t - f < p$, then $t = t - f$

2.3 INITIALIZATION OF STARTING THETA ESTIMATES

The user will supply five pieces of information in the test configuration:

1. A default starting value if no other information is available

2. An indication whether prior scores on the same test should be used, if available
3. Optionally, the test ID of another test that can supply a starting value, along with
4. Slope and intercept parameters to adjust the scale of the value to transform it to the scale of the target test
5. A constant prior variance for use in calculation of working EAP scores

2.4 INSERTION OF EMBEDDED FIELD-TEST ITEMS

Each blueprint will specify

- the number of field-test items to be administered on each test;
- the first item position into which a field-test item may be inserted; and
- the last item position into which a field-test item may be inserted.

Upon startup, select randomly from among the field-test items or item sets until the system has selected the specified number of field-test items. If the items are in sets, the sets will be administered as a complete set, and this may lead to more than the specified number of items administered.

The probability of selection will be given by $p_j = \frac{\sum_{j=1}^{K} K_j}{\sum_{j=1}^{K} a_j K_j} a_j K_j \frac{m}{N_J}$, where

p_j represents the probability of selecting the item;

m is the targeted number of field-test items;

N_j is the total number of active items in the field-test pool;

K_j is the number of items in item set j; and

a_j is a user-supplied weight associated with each item (or item set) to adjust the relative probability of selection.

The a_j variables are included to allow for operational cases in which some items must complete field testing sooner or enter field testing later. While using this parameter presents some statistical risk, not doing so poses operational risks.

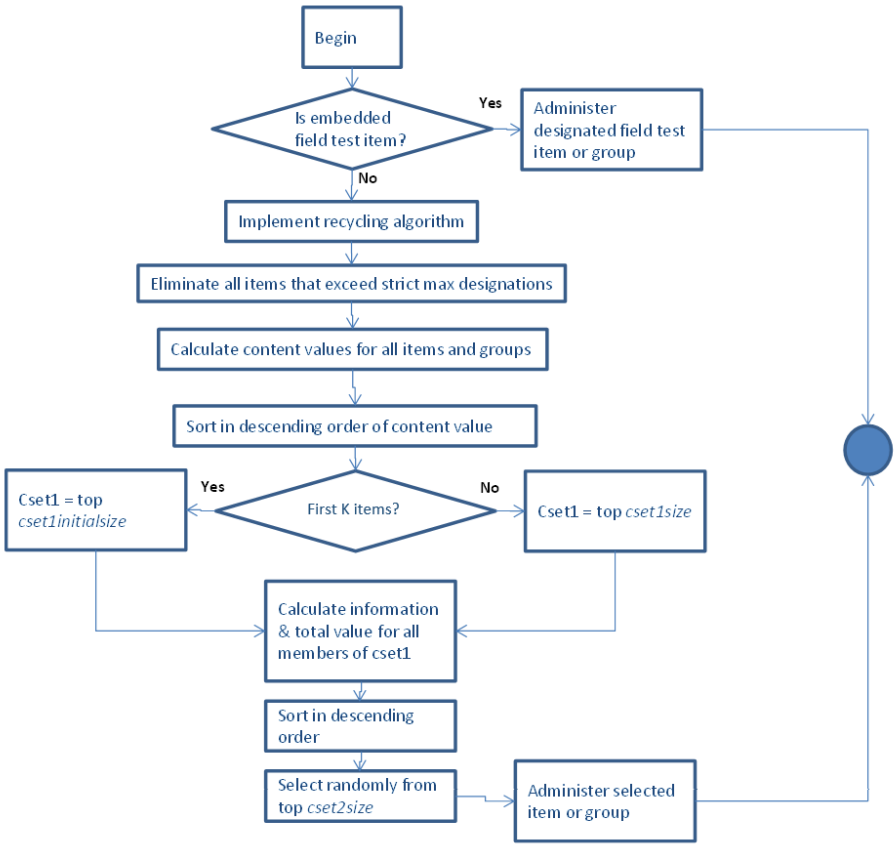
For each item set, generate a uniform random number r_j on the interval $\{0,1\}$. Sort the items in ascending order by $\frac{r_j}{p_j}$. Sequentially select items, summing the number of items in the set. Stop the selection of field-test items once $FTNMin \leq m \leq FTNMax = \sum_{j=0} K_j$.

Next, each item is assigned to a position on the test. To do so, select a starting position within $f - FTMax - FTMin$ positions from $FTMin$, where $FTMax$ is the maximum allowable position for field-test items and $FTMin$ is the minimum allowable position for field-test items. $FTNMin$ and $FTNMax$ refer to the minimum and maximum number of field-test items, respectively. Distribute the items evenly within these positions.

3. ITEM SELECTION

Exhibit D summarizes the item selection process. If the item position has been designated for a field-test item, administer that item. Otherwise, the adaptive algorithm kicks in.

Exhibit D. Summary of Item Selection Process



This approach is a "content first" approach designed to optimize match to blueprint. An alternative, "information first" approach, is possible. Under an information first approach, all items within a specified information range would be selected as the first set of candidates, and subsequent selection within that set would be based, in part, on content considerations. The engine is being designed so that future development could build such an algorithm using many of the calculations already available.

3.1 TRIMMING THE CUSTOM ITEM POOL

At each item selection, the active item pool is modified in four steps:

1. The custom item pool is intersected with the active item pool, resulting in a custom active item pool.
2. Items already administered on this test are removed from the custom active item pool.

3. Items that have been administered on prior tests are tentatively removed (see Section 3.2, Recycling Algorithm).
4. Items that measure content that has already exceeded a strict maximum are tentatively removed from the pool, removing entire sets containing items that meet this criterion.

3.2 RECYCLING ALGORITHM

When students are offered multiple opportunities to test, or when prior tests have been started and invalidated, students will have seen some of the items in the pool. The trimming of the item pool eliminates these items from the pool. It is possible that in such situations, the pool may no longer contain enough items to meet the blueprint.

Hence, items that have been seen on previous administrations may be returned to the pool. If there are not enough items remaining in the pool, the algorithm will recycle items (or item groups) with the required characteristic that is found in insufficient numbers. Working from the least recently administered group, items (or item groups) are reintroduced into the pool until the number of items with the required characteristics meets the minimum requirement. When item groups are recycled, the entire group is recycled rather than an individual item. Items administered on the current test are never recycled.

3.3 ADAPTIVE ITEM SELECTION

Selection of items will follow a common logic, whether the selection is for a single item or an item group. Item selection will proceed in the following three steps:

1. Select Candidate Set 1 (*cset1*).
    a. Calculate the content value of each item or item group.
    b. Sort the item groups in descending order of content value.
    c. Select the top *cset1size*, a user-supplied value that may vary by test.
2. Select Candidate Set 2 (*cset2*).
    a. Calculate the information values for each item group in *cset1*.
    b. Calculate the overall value of each item group in *cset1* as defined in Equation (1).
    c. Sort *cset2* in descending order of value.
    d. Select the top *cset2size* item groups, where *cset2size* is a user-supplied value that may vary by test.
3. Select the item or item group to be administered.
    a. Select randomly from *cset2* with uniform probability.

Note that a "pure adaptive" test, without regard to content constraints, can be achieved by setting *cset1size* to the size of the item pool and w_2, the weight associated with meeting content constraints

in Equation (1), to zero. Similarly, linear on-the-fly tests can be constructed by setting w_0 and w_1 to zero.

3.4 SELECTION OF THE INITIAL ITEM

Selection of the initial item can affect item exposure. At the start of the test, all tests have no content already administered, so the items and item groups have the same content value for all examinees. In general, it is a good idea to spread the initial item selection over a wider range of content values. Therefore, we define an additional user-settable value, *cset1initialsize*, which is the size of Candidate Set 1 on the first K items only, where K is the number of reporting categories. Similarly, we define *cset2initialisize*.

3.5 EXPOSURE CONTROL

This algorithm uses randomization to control exposure and offers several parameters that can be adjusted to control the tradeoff between optimal item allocation and exposure control. The primary mechanism for controlling exposure is the random selection from $CSET2$, the set of items or item groups that best meet the content and information criteria. These represent the "top k" items, where k can be set. Larger values of k provide more exposure control at the expense of optional selection.

In addition to this mechanism, we avoid a bias toward items with higher measurement precision by treating all items as though they measured with equal precision by ignoring variation in the slope parameter. This has the effect of randomizing over items with differing slope parameters. Without this step, it would be necessary to have other *post hoc* explicit controls to avoid the overexposure of items with higher slope parameters, an approach that could lead to different test characteristics over the course of the testing window.

4. TERMINATION

The algorithm will have configurable termination conditions. These may include

- administering a minimum number of items in each reporting category and overall;
- achieving a target level of precision on the overall test score;
- achieving a target level of precision on all reporting categories; and
- achieving a score insufficiently distant from a specified score with sufficient precision (e.g., less than two standard errors below proficient). American Institutes for Research (AIR) envisions this being used in conjunction with other termination conditions to allow very high or very low achieving students to continue on to a segment that contains items from adjacent grades but barring other students from those segments.

We will define four user-defined flags indicating whether each of these is to be considered in the termination conditions (*TermCount*, *TermOverall*, *TermReporting*, *TermTooClose*). A fifth user-supplied value will indicate whether these are taken in conjunction or if satisfaction of any one of them will suffice (*TermAnd*). Reaching the minimum number of items is always a necessary condition for termination.

In addition, two conditions will each individually and independently cause termination of the test:

1. Administering the maximum number of items specified in the blueprint
2. Having no items in the pool left to administer

APPENDIX 1. DEFINITIONS OF USER-SETTABLE PARAMETERS

This appendix summarizes the user-settable parameters in the adaptive algorithm.

Parameter Name	Description	Entity Referred to by Subscript Index
w_0	Priority weight associated with match to blueprint	N/A
w_1	Priority weight associated with reporting category information	N/A
w_2	Priority weight associated with overall information	N/A
q_k	Priority weight associated with a specific reporting category	reporting categories
p_r	Priority weight associated with a feature specified in the blueprint (These inputs appear as a component of the blueprint.)	features specified in the blueprint
a	Parameter of the function $h(.)$ that controls the overall information weight when the information target has not yet been hit	N/A
b	Parameter of the function $h(.)$ that controls the overall information weight after the information target has been hit	N/A
c_k	Parameter of the function $h(.)$ that controls the information weight when the information target has not yet been hit for reporting category k	reporting categories
d_k	Parameter of the function $h(.)$ that controls the information weight after the information target has been hit for reporting category k	reporting categories
cset1size	Size of candidate pool based on contribution to blueprint match	N/A
cset1initialsize	Size of candidate pool based on contribution to blueprint match for the first K items or item sets selected	N/A
cset2size	Size of final candidate pool from which to select randomly	N/A
cset2initialsize	Size of candidate pool based on contribution to blueprint match and information for the first item or item set selected	
t_0	Target information for the overall test	N/A
t_k	Target information for reporting categories	reporting categories
startTheta	A default starting value if no other information is available	N/A
startPrevious	An indication of whether previous scores on the same test should be used, if available	N/A
startOther	The test ID of another test that can supply a starting value, along with startOtherSlope	N/A
startOtherSlope	Slope parameter to adjust the scale of the value to transform it to the scale of the target test	N/A

Parameter Name	Description	Entity Referred to by Subscript Index
startOtherInt	Intercept parameter to adjust the scale of the value to transform it to the scale of the target test	N/A
$FTMin$	Minimum position in which field-test items are allowed	N/A
$FTMax$	Maximum position in which field-test items are allowed	N/A
$FTNMin$	Target minimum number of field-test items	N/A
$FTNMax$	Target maximum number of field-test items	N/A
a_j	Weight adjustment for individual embedded field-test items used to increase or decrease their probability of selection	field-test items
AdaptiveCut	The overall score cutscore, usually proficiency, used in consideration of *TermTooClose*	
TooCloseSEs	The number of standard errors below which the difference is considered "too close" to the adaptive cut to proceed. In general, this will signal proceeding to a final segment that contains off-grade items.	
TermOverall	Flag indicating whether to use the overall information target as a termination criterion	N/A
TermReporting	Flag to indicate whether to use reporting category information target as a termination criterion	N/A
TermCount	Flag to indicate whether to use minimum test size as a termination condition	N/A
TermTooClose	Terminate if you are not sufficiently distant from the specified adaptive cut	
TermAnd	Flag to indicate whether the other termination conditions are to be taken separately or conjunctively	N/A

APPENDIX 2. SUPPORTING DATA STRUCTURES

American Institutes for Research (AIR) Cautions and Caveats

- Use of standard error termination conditions will likely cause inconsistencies between the blueprint content specifications, and the information criteria will cause unpredictable results, likely leading to failures to meet blueprint requirements.
- The field-test positioning algorithm outlined here is very simple and will lead to deterministic placement of field-test items.

ADDENDUM. ADJUSTMENTS TO THE USE OF ITEM CLUSTERS

American Institutes for Research (AIR) adjusted the adaptive algorithm to the use of item clusters as follows:

- Using marginal maximum likelihood estimator (MMLE) to update proficiency estimates, marginalizing out cluster effects.
- Normalizing the information by the number of assertions within an item, to avoid over-selection of item clusters and stand-alone items with more assertions.