New Meridian

MEA

Maine Educational Assessments

Technical Report Maine Science Assessment Spring 2024

Grade 5, Grade 8, and High School Maine Educational Assessments (MEA)

Table of Contents

List of Tables

List of Figures

Chapter 1. Overview of Maine Science Assessments

The Maine Science Assessment is administered to students in grade 5, grade 8, and the third year of high school (HS) via computer-based testing (CBT), with a wide range of accessibility features (e.g., color scheme, font size, and zoom) available for all students. Accommodated paper-based tests (PBT), which include standard font-size print, braille, and large print (LP), as well as response accommodations that allow students to respond to test items using different formats), are available for students with disabilities. The Maine Science Assessment was administered to 12,396 students in Grade 5, 12,754 students in Grade 8, and 13,740 students in high school for publicly funded students in May 2024.

The Spring 2024 operational and field-test assessments leveraged the items in the New Meridian Science Exchange, a licensable collection of content contributed by states from their Next Generation Science Standards (NGSS)–aligned assessments, as well as content specially developed by New Meridian. The items selected for the Maine Science Assessment measure the science standards of the Maine Learning Results (MLRs). To ensure item quality, the items in the Science Exchange are reviewed against The New Meridian Framework for Quality Review of NGSS Science Assessment Items, which New Meridian developed in partnership with experts in the field of science education to articulate the critical elements of quality science assessment. The items used on the Maine Science Assessment are continuously monitored for technical quality for Maine students.

1.1. Purpose of the Assessment

The Maine Science Assessment has three primary purposes:

1. To provide information to the public about school performance through the state's ESSA reporting system, the ESSA Data Dashboard.
2. To support school identification within the state's ESSA-compliant system of school identification and support.
3. To provide a source of information for ongoing local program evaluation.

Student results are reported according to academic achievement descriptors, utilizing cut scores established in standard setting for each of four achievement levels: Well Below State Expectations, Below State Expectations, At State Expectations, Above State Expectations.

The MLRs/NGSS that the Maine Science Assessment are designed to measure are three-dimensional learning standards that describe a vision of what it means to be proficient in science. They envision science as a body of knowledge, an evidence-based model, and a theory-building enterprise that continually extends, refines, and revises knowledge. Therefore, the standards weave together each of the following:

1. **Disciplinary Core Ideas (DCI)** – are science topics that have broad importance across multiple sciences or engineering disciplines that
   a. provide a key tool for understanding or investigating complex ideas and solving problems,
   b. relate to students' interests, life experiences, and societal concerns, and
   c. are teachable and learnable over multiple grades at increasing levels of complexity.

2. **Science and Engineering Practices (SEP)** – describe behaviors in which scientists engage as they investigate and build models and theories about the natural world and are the key set of engineering practices that engineers use as they design and build models and systems.
3. **Crosscutting Concepts (CCC)** – provide an organizational framework for connecting knowledge across science disciplines to form a coherent and scientifically based view of the world.

1.2. Current Year Updates

The administration window had a duration of two weeks, May 13–24, 2024, for all three grades.

Chapter 2. Test Design and Development

2.1. Test Specifications

Criterion-Referenced Test

All items on the Maine Science Assessment forms come directly from the New Meridian Science Exchange item bank. In 2019, New Meridian launched the Science Exchange, a participatory science assessment item bank that facilitates sharing of science content. The Science Exchange includes over 2,000 science items, all of which align to the NGSS, for grades 3–8 and high school. Most of the items from the exchange have been used operationally on other state forms, and all items have been reviewed for fairness, bias, sensitivity, three NGSS dimensions, sense-making, and technical quality.

Item Types

To support valid measurement of the depth and breadth of the Maine Learning Results (MLRs), a variety of item types were identified and used to best elicit evidence of a student's mastery of a DCI and an SEP. The range of item types used on the Maine Science Assessment was selected to ensure accessibility and fairness for all test takers while maintaining a tight alignment to the MLRs. Item types included selected-response, technology-enhanced, and constructed-response (i.e., prompts), which together provide a high level of reliability and validity in measuring student performance. Items on the Maine Science Assessment may appear as standalone items or be grouped together to form clusters based on a common stimulus.

A cluster includes two or more items that require students to actively use the SEPs while applying their knowledge of the CCCs and drawing on their understanding of the DCIs to explain a phenomenon or solve a science/engineering problem. This process requires students to engage in sense-making as they actively reason and think about a phenomenon/problem. The process of sense-making is central to measuring student understanding of the NGSS and is a conceptual process in which a learner actively engages with the natural or designed world, wonders about it, and then develops, tests, and refines ideas to make sense of a phenomenon.

Cluster Stimulus. The items in a cluster are linked together with a grade-appropriate common stimulus and are scaffolded to help students make sense of a novel phenomenon. Stimuli are developed around phenomena or scientific problems to engage students in intriguing, realistic, and meaningful scenarios. These scientific phenomena require test takers to engage in sense-making throughout the cluster and are purposefully chosen to support multiple items that require students to demonstrate their achievement across multiple dimensions. The stimuli provide sufficient information to measure multiple dimensions of the science standards without teaching the content. All stimuli are developed to avoid subject matter that could introduce bias or sensitivity issues in student responses.

For students taking the computer-based assessment, the common stimulus in each cluster is shown on the left side of the screen and appears with every item in the cluster. Paper forms contain the common stimulus on the left side of the booklet and the items on the right whenever possible. The right side of the paper booklet contains as many items as can reasonably fit in the space provided. If additional pages are required, the scenario is repeated so students do not have to flip back to a previous page to refer to images or data tables. The students do not need to reread the background each time, but it is there for them if they need to refer to it.

Cluster Items. The items within a cluster are closely tied to the stimuli to provide a valid measure of the MLRs. Within each cluster, the items cover the concepts and evidence that relate to a given Performance Expectation (PE), which are central to the phenomenon or problem presented in the scenario. PEs are statements of what students should know and be able to do within the NGSS. However, the primary focus of the items is on the more specific DCIs, SEPs, and CCCs that make up each PE. This focus allows items in the Science Exchange to measure all aspects of a given PE more carefully and not constrain the assessment to only one combination of DCIs, SEPs, and CCCs. Items within a given cluster may also assess several different SEPs, DCIs, and CCCs that are found in the NGSS and are best used to make sense of the phenomenon outlined in the scenario.

Multiple-Part Items. Some items include multiple questions presented in multiple parts for students to answer. In some items, the parts are independent of each other, and in others they are dependent. In both cases, the parts are included to assess a deeper understanding of the science concepts being tested. Many times, students will progress through these multiple-part items by using one or more of the three NGSS strands (DCI, SEP, and CCC) when making sense of a scenario. The first part typically asks students to make a claim or identify evidence of a claim. The second part often asks students to use scientific reasoning to support their claim or reasoning about the evidence that can be used to support their thinking. These items are generally worth two points.

Response Formats

The clusters and standalone items include three general response formats—selected-response, technology-enhanced, and constructed-response. See Appendix A for examples of item response formats.

Selected-Response. Selected-response (SR) items include both traditional multiple-choice (MC) (i.e., select one correct answer among four options) and multiple-select (MS) (i.e., select a specified number of correct options or all the correct options). Both are well-established, versatile item types that provide an objective, efficient, and reliable method for measuring all levels of content knowledge. Students can earn one point for each selected-response MC item and one or two points for each selected-response MS item.

Technology-Enhanced. Technology-enhanced items (TEIs) provide an objective, efficient, and reliable method of measuring students' readiness to engage with information of varying degrees of cognitive complexity. The range of TEIs can be used to assess the critical-thinking and problem-solving skills specified by the MLRs. Students can earn one or two points for each TEI.

A variety of TEIs were included in the test forms. These item response formats help provide an authentic and engaging experience for students.

Constructed-Response. Constructed-response (CR) prompts provide another dimension of depth by requiring students to generate a written response. Thus, CR prompts may be better suited to address standards that require more cognitively complex engagement with information, including synthesis, drawing conclusions, modeling, and application. Students can earn up to two points for each constructed-response item.

Test Design

In Spring 2024, the Maine Science Assessment consisted of linear fixed forms composed of three 60-minute sessions. New Meridian constructs forms based on Maine Department of Education (DOE)-approved blueprints and specifications for each grade. The number of items and score points per session may vary slightly, but each session is designed to be completed in the designated testing time.

Blueprints

Grade 5 Blueprint

Coverage of Science Topics

All items on the Maine Science Assessment for Grade 5 are aligned to a science topic and specific performance expectation. To ensure ample coverage of all grade-level science topics, the blueprint specifies targets for the minimum and maximum number of operational score points aligned to each topic. For Grade 5, two topics—Earth's Systems and Space Systems: Stars and the Solar System—are combined in the blueprint to support reporting at the topic level.

Table 1. Grade 5 Blueprint: The Number of Operational Score Points Aligned to Each Science Topic

Science Topic	Performance Expectations (PE)	Target Percentage	Target Operational Items		Target Operational Score Points	
			Min	Max	Min	Max
Matter and Energy in Organisms and Ecosystems	5-PS3-1 5-LS1-1 5-LS2-1	30%	10	12	13	15
Structure and Properties of Matter	5-PS1-1 5-PS1-2 5-PS1-3 5-PS1-4	30%	10	12	13	15
Earth's Systems and Space Systems: Stars and the Solar System	5-PS2-1 5-ESS1-1 5-ESS1-2 5-ESS2-1 5-ESS2-2 5-ESS3-1	40%	13	15	17	19
Total		**100%**	**36**		**45**	

Coverage of Science Practices

To ensure appropriate coverage of the science practices, the majority of the Maine Science items (at least 90%) are aligned to a Science and Engineering Practice (SEP). Items that do not measure a SEP are aligned to a Disciplinary Core Idea (DCI) and sometimes a Crosscutting Concept (CCC).

The SEPs are grouped into three more general science practices—Investigate, Evaluate, and Reason Scientifically—based on the skills they entail. The blueprint specifies the target percentage of operational score points aligned to the three science practices.

Table 2. Grade 5 Blueprint: Maine Science Item Alignment to Each Science Practice

Science Practice	Science and Engineering Practices (SEP)	Target Percentage	Target Operational Items		Target Operational Score Points	
			Min	Max	Min	Max
Investigate	SEP1 SEP3	30%	10	12	13	15
Evaluate	SEP4 SEP5 SEP7	30%	10	12	13	15
Reason Scientifically	SEP2 SEP6	30%	10	12	13	15
Total		**90%**	**32**		**41**	

Grade 8 Blueprint

Coverage of Science Disciplines

All items on the Maine Science Assessment for Grade 8 are aligned to a science discipline and to a specific performance expectation. To ensure ample coverage of all grade-level science topics, the blueprint specifies targets for the minimum and maximum number of operational score points aligned to each discipline.

Table 3. Grade 8 Blueprint: The Number of Operational Score Points Aligned to Each Science Discipline

		Target Operational Items		Target Operational Score Points			
Science Discipline	**Target Percent**	**Min**	**Max**	**Min**	**Max**	**Science Topic**	**Performance Expectations**
Physical Science	33%	12	14	16	18	Structure and Properties of Matter	MS-PS1-1 MS-PS1-3 MS-PS1-4
						Chemical Reactions	MS-PS1-2 MS-PS1-5 MS-PS1-6
						Forces and Interactions	MS-PS2-1 MS-PS2-2 MS-PS2-3 MS-PS2-4 MS-PS2-5
						Energy	MS-PS3-1 MS-PS3-2 MS-PS3-3 MS-PS3-4 MS-PS3-5
						Waves and Electromagnetic Radiation	MS-PS4-1 MS-PS4-2 MS-PS4-3

Science Discipline	Target Percent	Target Operational Items		Target Operational Score Points		Science Topic	Performance Expectations
		Min	Max	Min	Max		
Life Science	33%	12	14	16	18	Structure, Function, and Information Processing	MS-LS1-1 MS-LS1-2 MS-LS1-3 MS-LS1-8
						Matter and Energy in Organisms and Ecosystems	MS-LS1-6 MS-LS1-7 MS-LS2-1 MS-LS2-3 MS-LS2-4
						Interdependent Relationships in Ecosystems	MS-LS2-2 MS-LS2-6
						Growth, Development, and Reproduction of Organisms	MS-LS1-4 MS-LS1-5 MS-LS3-1 MS-LS3-2 MS-LS4-5
						Natural Selection and Adaptations	MS-LS4-1 MS-LS4-2 MS-LS4-3 MS-LS4-4 MS-LS4-6
Earth and Space Science	33%	12	14	16	18	Space Systems	MS-ESS1-1 MS-ESS1-2 MS-ESS1-3
						History of Earth	MS-ESS1-4 MS-ESS2-2 MS-ESS2-3
						Earth's Systems	MS-ESS2-1 MS-ESS2-4 MS-ESS3-1
						Weather and Climate	MS-ESS2-5 MS-ESS2-6 MS-ESS3-5
						Human Impacts	MS-ESS3-2 MS-ESS3-3 MS-ESS3-4
Total	**100%**	**40**		**50**			

Coverage of Science Practices

To ensure appropriate coverage of the science practices, the majority of the Maine Science items (at least 90%) are aligned to a Science and Engineering Practice (SEP). Items that do not measure a SEP are aligned to a Disciplinary Core Idea (DCI) and sometimes a Crosscutting Concept (CCC).

The SEPs are grouped into three more general science practices—Investigate, Evaluate, and Reason Scientifically—based on the skills they entail. The blueprint specifies the target percentage of operational score points aligned to the three science practices.

Table 4. Grade 8 Blueprint: Maine Science Item Alignment to Each Science Practice

Science Practice	Science and Engineering Practices (SEP)	Target Percent	Target Operational Items		Target Operational Score Points	
			Min	Max	Min	Max
Investigate	SEP1 SEP3	30%	11	13	14	16
Evaluate	SEP4 SEP5 SEP7	30%	11	13	14	16
Reason Scientifically	SEP2 SEP6	30%	11	13	14	16
Total		**90%**	**36**		**45**	

High School Blueprint

Coverage of Science Disciplines

All items on the Maine Science Assessment for High School are aligned to a science discipline and to a specific performance expectation. To ensure ample coverage of all grade-level performance expectations, the blueprint specifies targets for the minimum and maximum number of operational score points aligned to each science discipline.

Table 5. High School Blueprint: The Number of Operational Score Points Aligned to Each Science Discipline

Science Discipline	Target Percent	Target Operational Items		Target Operational Score Points		Science Topics	Performance Expectations
		Min	Max	Min	Max		
Physical Science	33%	14	16	17	19	Structure and Properties of Matter	HS-PS1-1 HS-PS1-3 HS-PS1-8 HS-PS2-6
						Chemical Reactions	HS-PS1-2 HS-PS1-4 HS-PS1-5 HS-PS1-6 HS-PS1-7
						Forces and Interactions	HS-PS2-1 HS-PS2-2 HS-PS2-3 HS-PS2-4 HS-PS2-5
						Energy	HS-PS3-1 HS-PS3-2 HS-PS3-3 HS-PS3-4 HS-PS3-5
						Waves and Electromagnetic Radiation	HS-PS4-1 HS-PS4-2 HS-PS4-3 HS-PS4-4 HS-PS4-5

Science Discipline	Target Percent	Target Operational Items		Target Operational Score Points		Science Topics	Performance Expectations
		Min	Max	Min	Max		
Life Science	33%	14	16	17	19	Structure and Function	HS-LS1-1 HS-LS1-2 HS-LS1-3
						Matter and Energy in Organisms and Ecosystems	HS-LS1-5 HS-LS1-6 HS-LS1-7 HS-LS2-3 HS-LS2-4 HS-LS2-5
						Interdependent Relationships in Ecosystems	HS-LS2-1 HS-LS2-2 HS-LS2-6 HS-LS2-7 HS-LS2-8 HS-LS4-6
						Inheritance and Variation of Traits	HS-LS1-4 HS-LS3-1 HS-LS3-2 HS-LS3-3
						Natural Selection and Evolution	HS-LS4-1 HS-LS4-2 HS-LS4-3 HS-LS4-4 HS-LS4-5

Science Discipline	Target Percent	Target Operational Items		Target Operational Score Points		Science Topics	Performance Expectations
		Min	Max	Min	Max		
Earth and Space Science	33%	14	16	17	19	Space Systems	HS-ESS1-1 HS-ESS1-2 HS-ESS1-3 HS-ESS1-4
						History of Earth	HS-ESS1-5 HS-ESS1-6 HS-ESS2-1
						Earth's Systems	HS-ESS2-2 HS-ESS2-3 HS-ESS2-5 HS-ESS2-6 HS-ESS2-7
						Weather and Climate	HS-ESS2-4 HS-ESS3-5
						Human Sustainability	HS-ESS3-1 HS-ESS3-2 HS-ESS3-3 HS-ESS3-4 HS-ESS3-6
Total	**100%**	**44**		**55**			

Coverage of Science Practices

To ensure appropriate coverage of the science practices, the majority of the Maine Science items (at least 90%) are aligned to a Science and Engineering Practice (SEP). Items that do not measure a SEP are aligned to a Disciplinary Core Idea (DCI) and sometimes a Crosscutting Concept (CCC).

The SEPs are grouped into three more general science practices—Investigate, Evaluate, and Reason Scientifically—based on the skills they entail. The blueprint specifies the target percentage of operational score points aligned to the three science practices.

Table 6. High School Blueprint: Maine Science Item Alignment to Each Science Practice

Science Practice	Science and Engineering Practices (SEP)	Target Percent	Target Operational Items		Target Operational Score Points	
			Min	Max	Min	Max
Investigate	SEP1 SEP3	30%	12	14	16	18

Science Practice	Science and Engineering Practices (SEP)	Target Percent	Target Operational Items		Target Operational Score Points	
			Min	Max	Min	Max
Evaluate	SEP4 SEP5 SEP7	30%	12	14	16	18
Reason Scientifically	SEP2 SEP6	30%	12	14	16	18
Total		**90%**	**40**		**50**	

The Operational/Field Test Form Planner in Appendix B provides additional details for the composition of each form administered in 2024 in terms of subdiscipline, performance expectation, DCIs, SEPs, and CCCs by item counts.

Depth of Knowledge/Cognitive Complexity

Traditional Depth of Knowledge (DOK) and Cognitive Complexity measures have limited utility on a multidimensional science assessment in which the focus of most of its item clusters is based in sense-making. For 2024, p-values, ITCs, and match-to-blueprint were considered when selecting items for the final form. Several items on the Maine science form were operationally administered in other Science Exchange states in previous years. Going forward, Cognitive Complexity will be determined using *The New Meridian Framework for Quality Review of NGSS Science Assessment Items* in Appendix C. This framework outlines how items are to be judged by a group of evaluators across multiple metrics to determine the cognitive load on each student as they are answering each item. As outlined in the NGSS, the primary driver of complexity in all NGSS-aligned items is the extent to which students engage in sense-making to complete an item or a task. This process of measuring complexity as an attribute of sense-making is aligned with *A Framework to Evaluate Cognitive Complexity in Science Assessments* (Achieve, Inc., 2019).

The New Meridian framework relies heavily on Achieve research, and this report serves as the backbone to quantifying the complexity of all Science Exchange items and tasks. Specifically, items are evaluated using the two-step process outlined in the report. The first step is an analysis at the individual item level. At this level, four indicators are considered:

1. Scenario contributions to complexity
2. DCI or disciplinary understanding contributions to sense-making
3. SEP contributions to complexity
4. CCC contributions to complexity

Using these indicators, evaluators determine the degree to which students engage with the item or scenario and how this contributes to the level of sense-making required by this item and in what ways students' use of this dimension supports sense-making.

The second step is an analysis at the task level. This analysis includes consideration of how multiple items come together to compose a complete task. Evaluators look at how many dimensions are a part of the task and the overall scaffolding that engages students in sense-making throughout the task. Evaluating items using the New Meridian and Achieve frameworks allows for a more detailed judgment of the level of thinking required for a specific science task and builds on the multi-dimensional and progressive nature of NGSS items and clusters.

2.2. Test Development Process

The test forms for the Maine Science Assessment were constructed from fully licensed content from the New Meridian Science Exchange item bank. Forms include a mix of operationally ready content contributed by two state partners in addition to items developed by New Meridian. The item development and review processes for items from these three sources are described and discussed here.

Item Development

The Maine Science Assessment consists of items developed by New Meridian and content contributed by two state partners denoted here as Contributing State A and Contributing State B. The contribution of items from these three sources is described below in Table 7.

Table 7. Operational Items by Contributor for Spring 2024 Forms

Grade	Contributor	Total Items	Clusters	Standalone Items
05	New Meridian	9	2	2
05	Contributing State A	16	4	3
05	Contributing State B	9	2	2
08	New Meridian	13	4	1
08	Contributing State A	9	2	5
08	Contributing State B	19	6	0
HS	New Meridian	22	5	5
HS	Contributing State A	6	2	0
HS	Contributing State B	13	3	0

Item Development Processes for New Meridian–Developed Items

New Meridian develops content for the Science Exchange to ensure the health of the item bank, including the quality of the items, while also ensuring that items in the bank meet specific state requirements. To fill gaps in content in the New Meridian Science Exchange item bank and to accommodate the Maine item release policy, New Meridian develops new content through the New Meridian Science Educator Cadre. The Cadre is a team of teachers, subject matter experts, and experienced contractors trained by New Meridian to develop and review science tasks.

For 2024, the clusters that were field tested in Maine came from the Science Exchange licensed item bank. To develop additional clusters to be field tested, New Meridian employed members of the Science Educator Cadre. These cadre members were current or former classroom teachers and were trained by New Meridian in cluster development. Fourteen cadre members contributed to item writing for the purpose of creating field-test sets for the 2024 forms. Ten of the cadre members hold a master's degree, two hold a bachelor's degree, and two hold a PhD. The cadre members reside in ten different states.

New Meridian's design principles include the following:

- Analyzing high-quality stimulus materials
- Researching, analyzing, synthesizing, organizing, and using information from multiple sources
- Elaborating on and extending understanding; reasoning from models and evidence
- Citing evidence in support of a response
- Applying content knowledge through discipline-specific practices

These design principles reflect a commitment to quality and to measuring what matters most for students' future success: critical thinking, deep understanding, and the ability to communicate ideas effectively.

Selection and Training of New Meridian Science Educator Cadre Members

All science cadre members involved in reviewing the new stimuli and items on the spring 2024 test hold at least a bachelor's degree, with over 80% holding a master's degree or PhD. Approximately 83% of the cadre members were active K–12 classroom teachers, with the remainder being involved in the educational assessment industry. All cadre members had prior experience with the NGSS.

New Meridian Science Exchange Item Reviews

To ensure item quality, New Meridian, along with several experts in the fields of science education and educational assessment, have developed a framework for reviewing science assessment items (Appendix C). Using this framework, the New Meridian Science Educator Cadre reviews each item that is submitted. They review items for scenario quality, NGSS multi-dimensional performance, and technical quality. With this approach, New Meridian ensures that all items in the bank meet the highest standards of quality and that these measures are transparent to our subscribers.

All cadre members participated in a professional development session in which they learned how to use *New Meridian's Item Review Framework for Quality Review of NGSS*. Multidimensional performance indicators determine the degree to which tasks and items require students to use the DCIs, SEPs, and CCCs to actively engage with the natural or designed world. Specifically, when reviewing multidimensional performance, New Meridian evaluates items and clusters based on evidence, models, and scientific principles (e.g., sense-making) and the extent to which items assess each dimension and multiple dimensions together.

Reasoning with evidence, models, and scientific principles (e.g., sense-making). This reasoning is the fundamental differentiator between three-dimensional tasks and more traditional science assessments when taken in concert with the specifics of the dimensions engaged.

- **Item level**. Individual items require students to engage in generating evidence, to apply evidence to claims with reasoning, or to reason about the validity of claims related to a phenomenon or problem.
- **Cluster level**. Assessment clusters require students to connect evidence (provided or student-generated) to claims, ideas, or problems (e.g., explanations, models, arguments, scientific questions, definitions of/solutions to a problem) by using the DCIs, SEPs, and CCCs as fundamental components of their reasoning.

Assessing each dimension and multiple dimensions together. For each dimension (DCI, SEP, CCC), alignment indicators include the element of the dimension that is required to respond to the item/cluster, at what grade band the dimension is engaged, and whether the dimension is engaged in service of sense-making (in contrast to rote information).

- **Item level**. Individual items require students to use each dimension at a grade level in service of sense-making.
- **Cluster level.** Across a task, students are required to use at least two dimensions together to make sense of phenomena and/or problems.

New Meridian Item Writer Training

New Meridian conducted two in-person item writer workshops in Austin, TX, based on grade level, in January 2023. Over the course of 5 days, the participants received professional development on the following:

- Unpacking the standards (DCIs, SEPs, and CCs)
- Phenomena brainstorming
- Equity and inclusion
- Scenario development
- Storyline development of items for a cluster that uses sensemaking

Peer reviews and content reviews were held throughout the development process. The final handoff was a stimulus with a phenomenon and 5–7 items aligned to a particular NGSS grade level and topic. These items then went through bias and sensitivity review, copyediting, and content review prior to field testing. They also were reviewed by other cadre members using *New Meridian's Item Review Framework for Quality Review of NGSS*.

The agenda for the training can be found in Appendix D.

Item Selection and Test Assembly

The items and tasks were selected from the New Meridian Science Exchange item bank to meet the approved blueprints. The items selected were based on operational or field-test performance on the Maine Science Assessment in 2021, 2022, 2023, or on performance in the contributing state as well as alignment to the blueprint.

The process of item selection begins with analyzing the content present in the Science Exchange item bank and comparing that to the Maine test blueprint to identify the content that needs to be developed and typically continues with the following:

- Selection and professional development of item reviewers (from the educator cadre)
- Full item review with at least two cadre members reviewing each item and stimulus, adjudication of those reviews, and results analysis of that adjudication (identification of what moves on to be eligible for Maine forms or for revisions)
- Accessibility, bias, and fairness review and copyediting
- Selection of items for forms
- Review by Maine DOE personnel

New Meridian's training for the cadre focuses on key design principles, including those described in *The New Meridian Framework for Quality Review of NGSS Science Assessment Items,* that ensure high-quality, well-aligned test items.

Accessibility, bias, and fairness review. Each item and task selected for the Maine Science Assessment went through an accessibility and bias and fairness review prior to use on the test forms. During this review, New Meridian reviewed the scenarios and items selected for the test forms to confirm that there were no accessibility or bias or sensitivity issues that would interfere with students' ability to achieve their best performance. Scenarios and items were reviewed to evaluate adherence to the *New Meridian Fairness Guidelines* and to ensure that they do not unfairly advantage or disadvantage one student or group of students. New Meridian made edits and modifications to the scenarios and items to eliminate sources of bias or sensitivity and to improve accessibility for all students.

Style and copyedit review. Each test form for the Maine Science Assessment passed through a final style and copyedit review. The following criteria were used during this review:

- Scenarios, images, and items are clear, correct, and formatted to adhere to the New Meridian style guide.
- Scenario language and stem questions for online and paper forms are identical.
- Alt tag language is correct, clear, and consistent with other tags where possible.
- Alt tag language is free of spelling, grammatical, and mechanical errors.
- Item directions for online forms use verbs such as "Select" and "Move" based on item type.
- Item directions for paper forms resemble the online versions as closely as possible and are clearly stated for paper administration (i.e., ask students to "Mark" instead of "Move").

Draft Test Forms Review

The Maine Science Assessment forms were constructed with items and tasks from the fully licensed New Meridian Science Exchange item bank. In November of 2023, New Meridian held a test form verification educator committee meeting for each of the three grades. The educators were given information about the Maine Science Assessment and the goals of the meeting. Then they took the assessment online in ADAM via a practice test, session by session, and filled out a checklist along with comments. Comments were discussed for each session. The checklist included the following:

- The form's overall structure and progression (flow) are logical. Clusters are distributed to provide a balanced and coherent flow throughout the form. (i.e., not all physical science clusters are grouped together, not all standalone items are grouped together, not all constructed response items are grouped together).
- The assessment form contains the appropriate number of points/items.
- Appropriate standard accessibility features are provided.
- The items do not contain significant key runs.
- The assessment form contains a distribution of topic and subject representation.
- The assessment form contains a variety of cognitive tasks.
- The assessment form contains a variety of item types i.e., technology enhanced items (TEI), multiple-choice (MC), multi-select (MS), and constructed response (CR).
- Each session takes roughly the same amount of time to complete.
- The blueprint has been satisfied in terms of the breakdown of item points by science discipline.

There were three educators for grade 5, three educators for Grade 8, and five educators for the high school assessment.

2.2.7 Alternative Presentations

Technology-enhanced items were converted to paper-based versions for use on accommodated forms. For example, a computer-based drag-and-drop item may have been accommodated to a matching item on paper, in which the student draws lines to the same response options as in the online version. The items and tasks were then arranged into paper-based forms that were comparable to the approved online forms, meaning the forms were designed to measure the same content but with an alternate response format. These paper-based forms were the foundation for producing large-print and braille accommodated forms.

2.3. Standards Rotational Plan

See Appendix E for more information.

A three-year (2023–2025) rotational plan based on content in the topic arrangement of the NGSS is proposed. Maine's blueprint is organized by topic. Per the blueprint, all topics should have a minimum of two score points and a maximum of five score points. The rotational plan was developed to cover all topics within a three-year period.

Following the intent of *A Framework for K–12 Science Education*, the DCIs can be mixed and matched with any SEP and any CCC; the PE is just one combination of these. For that reason, the topic organization of the DCIs are the focus and not the individual PEs that are in the topic. In the item review process, each individual dimension of the NGSS is aligned separately to the items. Although items are assigned a PE, the alignment of the PE is only based on the content that corresponds to a DCI within that PE. The SEP or CCC are not considered for that alignment.

The Maine Science Assessments consist of item clusters with a few standalone items sprinkled throughout. The experts recognize that standalone items do not elicit the in-depth sense-making intent of the NGSS very well, and therefore the Maine blueprint includes very few of these items. Each item cluster is designed to assess students' deeper understanding of the DCIs found within the NGSS. Using various SEPs and CCCs to assess the content of the DCIs found within the NGSS, each item cluster will assess unique combinations of DCIs, SEPs, and CCCs that are not necessarily found within the PE. This use is also consistent with the way Maine's achievement level descriptors (ALDs) are constructed.

Each cluster will assess at least one DCI, CCC, and SEP; however, no one item is expected to assess all three. In other words, items can be one- (only assess one dimension of the NGSS) to three-dimensional (assess all three dimensions). The clusters are developed to dig deeper into the content of the DCIs. Students are presented with a discrepant event (phenomenon) and are asked to make sense of the phenomenon as they work through the item cluster. Students need to apply their knowledge of the content, their knowledge of science and engineering practices and skills, and their ability to make connections across the different content areas through the crosscutting concepts to make sense of the phenomenon given.

Because most of the items are cluster-based and ask students to make sense of a phenomenon, most clusters can only assess one, possibly two, content pieces (DCIs). Each cluster consists of 2–7 items, which means there will be at least 2–7 score points for a given topic. The intent of the NGSS is to strive for deeper understanding of more complex content and skills, emphasizing depth over breadth. Item clusters allow for the assessment of deeper understanding by giving students the opportunity to apply multiple DCIs, SEPs, and CCCs to make sense of a phenomenon. (Student & Gong, 2012. Reference: *Recommendations to Support the Validity of Claims in NGSS Assessment*.) Therefore, the three-year rotational plan shows that all NGSS topics will be covered as outlined in the blueprint. It is also important to note that the SEPs are represented in three categories and that all categories will be equally represented in each administration.

Other considerations were made when developing the rotational plan. One is the size of the student population. Administration is to roughly 40,000 students throughout the state, which includes grades 5 and 8 and high school. For validity, Maine has two forms with approximately 13 field test (FT) score points each. With a total of approximately 26 FT score points in each administration, 5 or 6 new content pieces found in the topics lacking adequate coverage can be added each year.

Another consideration is the organization of PEs in middle school and high school. The PEs within middle school and high school are not broken out by grade level. For example, high school may have courses in biology, physics, chemistry, and/or Earth science. Maine assesses high school science only once, and a student will most likely take the assessment in grade 11. Therefore, there are 67 PEs to consider for inclusion in a high school assessment. According to the National Research Council, “Because externally developed assessments cannot, by design, assess the full range and breadth of the performance expectations in the Next Generation Science Standards (NGSS), they will have to focus on selected aspects of the NGSS (reflected as particular performance expectations or some other logical grouping structure)” (National Research Council, 2014). Instead of trying to assess every PE, which is not how items are aligned or written, Maine DOE has used the topic organization of the NGSS to ensure each topic is equally covered and equally emphasized. Some topics encompass more PEs than others, but equally weighing the content at the topic level helps support teachers in their understanding that each topic is equally important to teach.

Chapter 3. Assessment Administration

3.1. Responsibility for Administration

Principals and their designated assessment coordinators are primarily responsible for the assessment's overall security and ethical administration, scheduling logistics, materials handling, and training and supervision of all assessment administrators/proctors. Manuals were provided to ensure uniformity in assessment procedures across schools and districts.

3.2. Administration Procedures

Maine districts and schools were provided with eight types of manuals/guides conveying best practices and procedures to successfully administer the Spring 2024 Maine Science Assessment. The materials were available for download on the Maine Science Support site.

List of manuals/guides:

1. Principal and Assessment Coordinator Manual (PAC Manual)
2. Assessment Administrator Manuals (one for each grade)
3. Proctor User Guide
4. Accessibility Guide
5. Device, System, and Lockdown Browser Installation Guide
6. ADAM Platform User Guide
7. Quick Guide – Starting Your Maine Science Assessment
8. Practice Assessment Administration Manuals (one for each grade level)

The PAC and AAMs set expectations for assessment security and ethics and provided procedure checklists for use before, during, and after administration. These checklists were designed to assist with the logistics for preparing, administering, and cleaning up for the online and paper-based assessments.

The AAMs provided critical information for preparing to administer the Maine Science Assessment, including the materials to be provided for student use, the types of questions students encounter, instructions for assessing students who require accommodations, and final preparations. The AAMs also included scripts for administering the assessment and descriptions of the universal features, designated supports, and accommodations available for students.

The Proctor User Guide provided the procedures to proctor the online assessment in the Assessment Delivery and Management (ADAM) platform. It detailed how to log in to ADAM as a proctor; how to access and manage assessment session dashboards; how to confirm which students are in an assessment proctoring group; and how to start, pause, and end an assessment session.

For district and school assessment administrators and technology coordinators, the Accessibility Guide provided the necessary information for the embedded and non-embedded accessibility tools available for the Maine Science Assessment. The ADAM platform featured a range of onscreen tools that enhanced the accessibility of online assessments for all students, including those who required visual, auditory, and attention-focus supports. This guide explained the accessibility features and provided a brief tutorial for each tool, including where to find it within the assessment platform and how to use it.

Two of the guides, the Device, System, and Lockdown Browser Installation Guide and the ADAM Platform User Guide, contained procedures and information that helped districts and campuses prepare their networks, systems, and devices for the technology needs of the online assessment system.

In addition, the Quick Guide – Starting Your Maine Science Assessment was meant to be used along with the online tutorial in ADAM. From a student's perspective, the guide and the tutorial explained how to sign into the Maine Science Assessment in the lockdown browser, how to navigate the assessment from the welcome page to the review page, and how to use the universal tools.

The Practice Assessments in ADAM are an online set of scenarios and items meant to familiarize students with the types of questions they may encounter when they take the Maine Science Assessment. The practice test is not scored nor are the students' answers retained. Each online question can be answered and checked via the online interface.

The Practice Assessment Administration Manual is meant to be used in conjunction with the Practice Assessments for the Assessment Administrators Manual. The two manuals explain the uses of the practice assessments and contain the rationales and exemplars for those assessments.

Together, these manuals document the knowledge and procedures needed to support the successful administration of the Maine Science Assessment.

3.3. Participation Requirements and Documentation

The Maine Science Assessment assesses all publicly funded Maine students in grades 5, 8, and third year of high school. Students with significant cognitive disabilities who qualify for the alternate assessment to the Maine Science Assessment will participate in the MSAA-Science. The Maine Science Assessment does not need to be submitted for any student who was assessed through the alternate assessment. Publicly funded Maine students attending out-of-state schools, regional programs, and private schools were included in the assessment administration. Districts and/or schools could optionally offer the assessment for students who were homeschooled.

No assessments were allowed to be administered at home. All assessments had to be administered by trained assessment administrators at a school building unless the Maine DOE approved special considerations (i.e., medical exemptions). Students who answered at least 25% of the entire assessment (within any sessions) were considered participants, received scaled scores, and were included in the data matrix.

Appendix G. presents the participation in the Maine Science Assessment by publicly funded students by grade and demographic group.

3.4. Administrator Training

The Maine School Administrative Units (SAUs—districts) and schools were offered three avenues to obtain training for the Maine Science Assessment.

First, New Meridian, in partnership with Maine DOE, provided a prerecorded virtual session to introduce administrators to the Maine Science Assessment and to its alignment to the NGSS. This session was posted online and left open for viewing by the field at their leisure.

Second, three (3) live virtual question-and-answer sessions were provided for districts, schools, and coordinators to ask questions about three key topics—Maine Science Assessment – Install Lockdown Browser, Maine Science Assessment – Rostering in ADAM Platform & Accessibility, and Maine Science Assessment – Administering, Proctoring & Accessibility Review. These live sessions were recorded and made available online for districts and schools to view at a later time if they could not attend the live session.

Third, a support center was established to provide support for districts, campuses, principals, and coordinators before, during, and after the assessment administration. As this was the third year of the Maine Science Assessment, the support center opened for phone, chat, email, or Zendesk support requests starting Monday, April 22, 2024, from 7:30 a.m. to 4:00 p.m. EST. Starting Monday, May 6, 2024, until the last day of the administration (Friday, May 31, 2024), the support center was open from 6:30 a.m. to 6:00 p.m. EST. Then, for one week after the assessment, the support center was open Monday, June 3 through Friday, June 07, 2024, from 7:30 a.m. to 4:00 p.m. EST for phone, chat, email, or website support. The support center then resumed its normal operating hours of 7:30 a.m. to 4:00 p.m. EST for email or website support. No matter how the field contacted the support center, all contacts were documented as a Zendesk support ticket and tracked from first contact to resolution (as shown in Table 8).

Table 8. Support Center Operating Hours

Date Range	Hours	Support Type
June 10, 2023–April 19, 2024	7:30 a.m.–4:00 p.m. EST	Email or the support center website
April 22, 2024–April 26, 2024	7:30 a.m.–4:00 p.m. EST	Phone, chat, email, or the support center website
April 29, 2024–May 24, 2024	6:30 a.m.–6:00 p.m. EST	Phone, chat, email, or the support center website
May 27, 2024–May 31, 2024	7:30 a.m.–4:00 p.m. EST	Phone, chat, email, or the support center website
June 6, 2024–March 9, 2025	7:30 a.m.–4:00 p.m. EST	Email or the support center website

3.5. Universal Tools, Designated Supports/Features, and Accommodations

The Maine Science Assessment Accessibility Guide is structured into two sections, the *ADAM Accessibility Tools* and *Online Accessibility Tools User Guide*. The *ADAM Accessibility Tools* section is further categorized into three sections, catering to varying student needs: universal tools for all students, designated supports for some students, and accommodations requiring IEP/504 documentation. These tools, supports, and accommodations are available to all students. The decision regarding their use is made by the student's educational team and tailored to the individual's needs, irrespective of their disability status.

It is essential to ensure that tools, supports, and accommodations align with those utilized during the student's regular classroom instruction, including test-taking. In certain cases, a student who relies on these tools, supports, and accommodations may need to complete the test in a separate setting. This ensures minimal distractions for other students and safeguards the security and confidentiality of the test.

The responsibility for ensuring proper implementation of these resources is on principals and assessment coordinators, who are tasked with making sure that assessment administrators and proctors comprehend the correct usage and administration of these tools.

Before the administration of the Maine Science Assessment, principals and assessment coordinators were responsible for the following actions:

1. Assigning students to the appropriate embedded designated support (i.e., text-to-speech) within the ADAM assessment platform.
2. Assigning students to the relevant non-embedded designated supports and accommodations, ensuring that students had the correct paper-based version (standard font, large print, or braille) when necessary.
3. Verifying that the correct supports and accommodations were assigned to students who utilized them during the Maine Science Assessment administration.

Universal Tools for All Students

Universal tools are made available to all students. There were two types of universal tools available to students, embedded (shown in Table 9) and non-embedded (shown in Table 10).

- Embedded universal tools are provisions within the online assessment platform (ADAM) available to all students.
- Non-embedded universal tools are provisions outside the online assessment platform.

Table 9. Embedded Universal Tools—Provisions Within ADAM

Tool	Tool Icon	Description
Review	📅	Review page shows flagged items for review and items not attempted
Accessibility	♿	Accessibility options of Color Scheme / Font Size / Zoom enlargement
Flag or Bookmark	⚑	Ability to flag or bookmark an item to return to for review
Line Reader	▬	The line reader tool helps focus on reading one line of text at a time
Response Masking	ⱥ	Ability to hide/cover an answer choice – not available on all item types, such as technology enhanced

Table 10. Non-Embedded Universal Tools—Provisions Outside of ADAM

Tool	Description
Scrap/Scratch Paper	The student uses scratch paper, an individual erasable whiteboard, or an assistive technology device to make notes or record responses. Scratch paper can be lined, blank, or graph. All scratch paper must be collected and securely destroyed at the end of each test to maintain test security.

Designated Supports for Some Students

Supports outlined below may provide increased accessibility within the assessment. The provision of designated supports is not dependent upon disability status, and all students—regardless of disability status—are eligible for designated supports. Utilization and implementation of supports are determined on an individual basis by a team of two or more education professionals with knowledge of the student's performance, and supports must be consistent with the student's normal routine during classroom instruction and assessment. Provision of supports does not alter the construct of any test item.

For the Maine Science Assessment, all but one (See Table 11) of the designated supports were non-embedded (i.e., provisions outside the ADAM online assessment platform). (See Table 12.)

Table 11. Embedded Designated Support

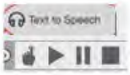
Tool	Tool Icon	Description
Provision within online platform that must be assigned to individual student by DAC/ITC/SAC		
Text-to-Speech (TTS)		Text is read aloud to the student via (embedded) TTS technology. **TTS should be consistent with the student's normal routine during classroom instruction.** Headphones or earbuds are necessary unless student is tested individually in a separate setting.

Table 12. Non-Embedded Designated Supports—Provisions Outside of ADAM

Support	Description
Breaks	Multiple or frequent breaks may be required by students whose attention span, distractibility, and physical and/or medical condition require shorter working periods.
Extended Time	Extended time is time beyond recommended/average of 60 minutes per session(s) 1, 2, and 3. Students with extended time must complete the assessment session on the day it was started; the session will auto-submit at 11:59 PM.
Small Group or Individual Setting	This designated support is used to minimize distractions for students whose test is administered out of the classroom or so that others will not be distracted by supports/accommodations being used.
Bilingual Word Glossary for MLs	A bilingual/dual language word-to-word glossary without definitions is provided to students who are multilingual learners as a language support as per their ILAP.

A bilingual word-to-word glossary without definitions is an available designated support for multilingual learners (MLs) with the support identified in an Individual Language Acquisition Plan, or ILAP.

Table 13. Bilingual Glossary and Multilingual Learners Counts

Grade	Bilingual Glossary Count	Multilanguage Learners Receiving Services Count
5	112	526
8	78	488
High School	179	426

Accommodations Requiring IEP/504 Documentation

Accommodations that required Individualized Education Program/Plan (IEP) or 504 Plan documentation were available for student use. These accommodations are changes in procedures or materials that do not alter what the assessment measures and are used to increase equitable access during the assessment for students for whom there is documentation of the need on an IEP or 504 Plan. Human reader and American Sign Language, braille, scribe, paper-based, or paper-based large print were provided as non-embedded accommodations. (See descriptions in Table 14.)

Table 14. Non-Embedded Accommodations—Provisions Outside of ADAM Based on IEP or 504 Plan

Accommodation	Description
Human Reader (Paper-Based Tests)	This accommodation is only allowed for students that have a documented need for paper/pencil. The student will have those parts of the test that have text-to-speech support in the computer-based version read by a qualified human reader in English.
American Sign Language	Test is translated via sign language interpreter to student by Test Administration as documented in the IEP/504 plan.
Braille	Both contracted and un-contracted braille (English braille, American Edition or Unified English braille) are available as indicated by a student's IEP/504 plan. Students who require a braille assessment will be sent a transcribed paper-based assessment.
Scribe	The student may dictate answers to a human scribe in an individual setting as indicated by a student's IEP/504 plan. Human scribe records verbatim what a student dictates and must give the student an opportunity to review scribed text. Scribed answers must be entered into the online testing platform – no paper submissions accepted.
Paper-Based Tests and Large Print	This accommodation is for students with an IEP/504 plan that requires assessments to be paper-based and not administered online.

Accommodations Frequency Tables

Table 15. Grade 5 Accommodation Frequencies

Accommodation Code	Description	Count	Percent
accomHumanReader	Human Reader	—	—

accomLargePrint	Large Print	—	—
accomBraille	Braille	—	—
accomScribe	Scribe	373	3.01
accomPaper	Paper	—	—
accomASL	ASL	—	—
IEP	Individualized Education Program	2,805	22.63
Plan504	Disabilities Plan	679	5.48

Table 16. Grade 8 Accommodation Frequencies

Accommodation Code	Description	Count	Percent
accomHumanReader	Human Reader	—	—
accomLargePrint	Large Print	[redacted]	[redacted]
accomBraille	Braille	—	—
accomScribe	Scribe	191	1.50
accomPaper	Paper	[redacted]	[redacted]
accomASL	ASL	[redacted]	[redacted]
IEP	Individualized Education Program	2,576	20.20
Plan504	Disabilities Plan	973	7.63

Table 17. High School Accommodation Frequencies

Accommodation Code	Description	Count	Percent
accomHumanReader	Human Reader	—	—
accomLargePrint	Large Print	—	—
accomBraille	Braille	—	—
accomScribe	Scribe	74	0.54
accomPaper	Paper	[redacted]	[redacted]
accomASL	ASL	—	—
IEP	Individualized Education Program	2,303	16.76
Plan504	Disabilities Plan	1,214	8.84

3.6. Assessment Security

The quality and usefulness of the assessment data generated by the Maine Science Assessment depend primarily on the uniformity of the assessment administration and the security of assessment materials. Valuable information about student achievement of the content standards and the effective measuring of Maine Learning Results (MLRs) would be seriously compromised if assessment security were not strictly implemented and maintained.

School principals are responsible for ensuring that the Maine Science Assessment administration takes place under these guidelines. Duplication of any portion of the Maine Science Assessment content is strictly forbidden, including but not limited to audio recording, video recording, photographing, photocopying, and handwritten copying. No assessment or record of student work or computer-generated responses may be retained, discarded, recycled, removed, or destroyed.

Principals and assessment coordinators were directed in the Principal and Assessment Coordinator (PAC) Manual to collect, inventory, and account for all secure assessment materials before, during, and after completion of the assessment administration, whether that administration was online or on paper. The principals and coordinators were to ensure that all assessment materials, including the Student Assessment Cards, Student Assessment Booklets, and Assessment Administrator Manuals, were returned by each assessment administrator/proctor and regional program.

In addition, the Maine DOE Assessment Security Handbook outlines best practices for prevention of assessment irregularities as well as processes for detection and investigation of irregularities.

All SAUs (districts) and schools were directed to call the Maine DOE in the event of a situation that could have caused the assessment administration to be compromised.

3.7. Assessment Administration Window

The test administration window for all grade 5, grade 8, and the third year of high school science assessments was scheduled for May 13–24, 2024.

3.8. Assessment and Administration Irregularities

Two irregularities were reported during this administration. The first irregularity pertained to a procedural issue in which a student with an IEP took the science assessment on paper, with the student's case manager subsequently inputting the answers into the ADAM platform. It is worth noting that this process deviated from the standard procedure for this assessment. The correct procedure entails shipping the paper assessment(s) to the paper scoring vendor, who then enters the student's responses into ADAM. Upon further investigation, it was confirmed that the answers in the test booklet matched the responses entered into ADAM. Consequently, the Maine DOE did not invalidate this student's assessment.

The second irregularity involved the use of an external resource to answer a constructed response question. Following a review by the Maine DOE Assessment Team Committee, this student's assessment was invalidated with no score reported.

3.9. Quality Assurance of Results

Rigorous quality control procedures were implemented throughout the test development, administration, scoring, and analyses phases.

Quality Control of Assessment Administration

Administrator Training sections (3.4 and 3.4.1) of this report provide details about the tutorials, training, administration manuals, and support center that supported the standardized administration and security of the Maine Science Assessment.

To ensure against loss of data during online administration, ADAM by default transmits student responses back to the MZD cloud-based servers every 15 seconds. As an additional precaution to minimize the impacts of interruptions in connectivity, the student session is also synchronized with the servers each time the student moves to a new question. Should an interruption occur, the student is prevented from moving to the next question. This limits the potential disruption impact to a single question. At the close of the test session, all temporarily stored content and data are automatically removed from the browser cache, meaning that the ADAM lockdown browser solution is also more secure than other local storage methods. Per an agreement with Maine DOE, all test sessions still open at 11:59 EST are considered idle. ADAM automatically submits the results and closes the session. Maine DOE authorization is required to reopen the session.

Paper-Based Assessments

With paper-based assessments, Strategic Measurement and Evaluation, Inc. (SME) used rigid and redundant secure materials–processing procedures to ensure that test booklets and the associated processing boxes remained secure throughout test administration, document processing, scanning, and storage. Material movement was constantly monitored as documents were shipped to schools, returned to SME, and processed through the scanning center. SME's processing procedures ensure 100 percent accounting of all materials.

SME created pre-ID labels based on student registration information and applied these labels to test booklets that were then sorted and shipped to the student's designated testing site. To facilitate tracking and security, SME pre-coded each test booklet with a unique sequential identification number. SME inventoried the test booklet numbers that were sent to each test administration site prior to shipping and audited return shipments to ensure all test booklets were returned. Specific packing and shipping instructions were included to support the distribution, collection, and return of test booklets to SME.

Materials were securely delivered in sealed boxes with a clear directive that only the test center supervisor was designated as a recipient. The delivery required the signature and printed name of the designated recipient. No package was allowed to be left at a school without a signature. A dedicated courier service picked up and returned test booklets directly from each testing site. Immediately upon receipt of return shipments, SME scanning staff inventoried the test materials.

After inventory, test booklets were boxed (by grade/subject) and placed in secure temporary storage at SME's scanning site. Access to the secure temporary storage area was restricted to authorized personnel. Processing boxes were numbered, inventoried, and recorded in the electronic inventory system. Handling and retrieval of boxes was limited to authorized personnel.

All student test booklets received by SME were scanned according to strict quality assurance (QA) procedures to ensure the accuracy of the data capture. QA procedures included the following:

- The preparation and testing of all scanning programs using test decks with known characteristics
- Constant monitoring of scanner operations, including scanner calibration, document alignment, scanner speed, clerical checks and monitoring of documents that generate an error code
- Verifying image file counts against expected page counts with deviations triggering an immediate alert

After being scanned, test booklets were stored at a secure warehouse for the time-period required by the contract and then securely recycled on site.

Chapter 4. Item-Level Scoring

The Maine Science Assessment consisted of a variety of item types, including selected-response, technology-enhanced, and constructed-response formats. Certain item types, such as selected-response and technology-enhanced, were configured for automatic, rule-based machine scoring in ADAM, the test administration platform. Scoring rules are documented on the test maps as part of the test form development process and verified during the initial key check validation process.

This section describes the scoring process for both the machine-scored items and the range-finding and hand-scoring processes for the constructed-response prompts. Along with the detailed description of the range-finding process, it provides information about scorer qualification, training, and monitoring.

4.1. Machine-Scored Items

Machine-scorable items were scored within the ADAM test delivery platform for both online and paper administration modalities.

Online Administrations: The ADAM platform technology is designed to automatically process scores upon student submission. Machine-scorable assessment questions include items for which the test taker response is an online interaction with the item, such as with multiple-choice, drag-and-drop, and other technology-enhanced items. These item interaction types include the programming necessary to correctly score student response(s) as designated by the item author. Programming also includes specific response-processing instructions, embedded in technical encoding, based on the item interaction type. ADAM uses the Question and Test Interoperability (QTI) technical encoding standard to render and score assessment items.

Paper-Based Administrations: Student responses for the machine-scorable items that were delivered on paper were entered into the ADAM system by SME during the processing of students' returned paper forms. Specifically, all responses were entered by the first transcriber. Then a second trained transcriber reviewed all the answer choices entered by the original transcriber. The second transcriber viewed all the items and the related transcription for 100% accuracy. If a discrepancy was found, the second transcriber logged the discrepancy and made the correction. A third transcriber reviewed each record as a final QA check.

Quality control systems and methodology included as part of the item review and approval process ensure that item scoring rules are configured correctly, and each item is properly machine scored when administered as part of an assessment. The ability to review simulated student responses to items and the machine-scored output is also native to the item authoring/review process and further ensures scoring accuracy.

4.2. Human-Scored Prompts

This section describes the complete range-finding and scoring process for all the constructed-response prompts on the Maine Science Assessment.

After the test administration, operational scoring was conducted for all the constructed-response prompts (i.e., two prompts for grade 5, three prompts for grade 8, and two prompts for high school). The operational scoring guides consisted of anchor papers, a practice set, and two qualification sets. The final scoring guides with Maine DOE–approved scores were used to train all scoring staff.

Student responses to constructed-response prompts were human-scored using MZ Development, Inc.'s (MZD) Online Scoring and Reporting (OSCAR) electronic scoring platform. Through OSCAR, qualified scorers accessed responses created by online test takers (i.e., in the ADAM online test administration platform). Scorers evaluated each response according to the scoring rubric and recorded its score via mouse entry through the OSCAR system. When a scorer finished evaluating one response, the next response immediately appeared on the computer screen.

Electronic responses in OSCAR were organized by grade and prompt. Access to student responses was controlled using role-based permissions to limit platform interaction and to enhance test security. Scorers were assigned to a specific grade-level team and were given a unique OSCAR login that allowed them to view specific prompts only once the qualification criteria were achieved. In the OSCAR system, scorers see only the student response; they do not have access to any student demographic information.

After test administration and operational scoring, range-finding was conducted for all constructed-response field-test prompts (i.e., one prompt for grade 5, four prompts for grade 8, and four prompts for high school) to finalize the scoring rubrics and identify samples of student responses for each score point. These samples were reviewed by the Maine educator range-finding committee and used to build field-test scoring guides. The field-test scoring guides with Maine DOE range-finding committee-approved scores were used to train all scoring staff.

Scoring Location and Staff

Scoring for the Maine Science Assessment was completed at SME's scoring center in Lafayette, Indiana. All training and scoring activities were conducted in person at the scoring center.

SME used a hierarchical structure to manage the Maine Science Assessment scoring project. The project had a designated scoring manager and scoring content specialist who reported directly to SME's scoring director. Based on experience and qualifications, scorers were assigned to a grade-level scoring team with a designated scoring supervisor and table leader. Scoring teams were physically separated into different rooms. These rooms were organized in such a way as to allow for constant supervision and monitoring of computer screens, facial expressions, and body language.

SME Staff Positions

The following SME staff positions were involved with scoring activities for the Maine Science Assessment:

- **Scoring Director** – oversaw program communication and coordination of all scoring activities.
- **Scoring Manager** – coordinated range-finding activities, oversaw daily scoring operations, managed scoring training, and monitored scoring supervisors and table leaders.
- **Scoring Content Specialist** – managed the range-finding team(s) and coordinated the creation of all scoring guides; with the scoring manager, trained grade-level scoring supervisors and table leaders and monitored their work.
- **Assessment Manager** – participated in range-finding activities and consulted with scoring leadership as needed.
- **Grade-Level Scoring Supervisors** – worked under direct supervision of the scoring manager to assist with range-finding, helped prepare scoring materials, and trained and supervised scorers and table leaders.

- **Table Leaders** – were experienced scorers who assisted scoring supervisors with read-behinds, answering questions, and monitoring group performance.
- **Scorers** – worked as part of grade-level teams based on qualifications and experience.

Range-Finding

Upon completion of operational scoring, SME facilitated range-finding for all field-test constructed response prompts that appeared on the Maine Science Assessment (i.e., one prompt for grade 5, four prompts for grade 8, and four prompts for high school).

SME and the Maine DOE completed range-finding for the Maine Science Assessment virtually (via Zoom) in July. Maine DOE sent out a general notice for recruitment to all Maine educators in early April. At that time of publication, the Maine DOE presented the educators with opportunities to participate in upcoming committees for the remainder of 2024 consisting of range-finding, standard setting, and data review. Once the Maine DOE educator recruitment form closed, the Maine DOE reviewed the list to confirm the educators were indeed certified science educators and then sent New Meridian the list with the educators' contact information. New Meridian then contacted the educators that signed up for range-finding to confirm their participation and provide them with further details, as well as the virtual Zoom meeting links and codes.

The educator participants' range-finding grade-level assignment was as follows: three educators for the grade 5 assessment, three for the grade 8 assessment, and four for the high school assessment. Of those ten certified educators, four identified their area of expertise as Life Science, three as Chemistry, two as Earth & Space Science, and one as Physics. All were certified in the grade levels for which they participated in range-finding. All educators were required to sign an NDA to participate.

Representatives from the New Meridian test development and program management teams also participated in the range-finding meetings.

To prepare for virtual range-finding meetings with Maine educators, an internal SME range-finding committee consisting of SME's assessment manager, scoring director, scoring manager, scoring content specialist, and experienced scoring supervisors reviewed the draft scoring rubrics. This review process included discussions with New Meridian's test development team.

After the scoring rubric discussions, SME's scoring manager and scoring content specialist reviewed samples of student responses and selected approximately 100 sample responses per prompt to be scored as part of the internal range-finding process. The final set of responses included for each field-test item was chosen to be representative of the types of answers students produced. OSCAR was used to select, organize, score, and annotate the sample responses for Maine educator review. These sample responses covered a range of score points and represented a variety of issues and patterns across student responses.

Next, SME's range-finding committee conducted initial range-finding for all prompts. The purpose of the initial SME range-finding sessions was to identify 30–45 representative samples for Maine DOE educator review and approval. The range-finding process for an individual item began with a review of the item prompt and the draft scoring rubric. Following this review, each SME range-finding participant (typically 4 or 5 per item) independently reviewed all the initial sample responses for a particular item and assigned a score. Participants were encouraged to also annotate each sample response to document questions, comments, patterns, or issues that needed further clarification.

After each participant independently reviewed and scored the sample responses, the range-finding facilitator (e.g., the scoring content specialist, scoring manager, or scoring director) led a group discussion of each sample response. The purpose of the group discussion was to assign a consensus score and an annotation to each sample response. If questions arose during the group discussion (e.g., regarding something that was not covered by the scoring rubric), the participants discussed the issue to identify the elements that led to the assignment of different scores. After discussion, the facilitator recorded the consensus score assigned to each sample response. Responses for which a score could not be agreed upon were noted as "do not use." The facilitator documented the discussions and noted any decisions that were made.

The range-finding targets (i.e., the number of sample responses pulled) for each score point appear in Table 18. The range-finding process described above was repeated, as needed, until the SME committee was satisfied with the sample responses and associated scores and notation.

Table 18. Targets for Initial Range-Finding

Item Max Points	Target		
	0 points	1 point	2 points
2	12–15	12–15	12–15

SME and the Maine DOE range-finding committee met virtually via Zoom for 3- to 6-hour sessions to discuss all sample responses and scores and finalize the scoring notes and rubrics for each prompt. These discussions included decisions about the defining characteristics and thresholds for responses at each score point.

Representatives from the New Meridian test development and program management teams also participated in the range-finding meetings and documented group feedback related to item development and content.

Following the Maine DOE/SME range-finding meetings, SME staff made any required edits to the scores, annotations, or rubrics as directed by the Maine educator range-finding committee. Then SME used the final committee-approved samples to construct item-level field-test scoring guides. These field-test scoring guides were used by senior scoring staff to score all the field-test responses.

After field-test scoring, the item data were reviewed to determine if a field-test item was eligible for inclusion on a future operational form. SME's internal range-finding committee followed the procedures described in this chapter to complete additional range-finding for all new items that were eligible for an operational form. This additional range-finding produced enough responses to build operational scoring guides. Each item-level operational scoring guide included one anchor set, one practice set, and two qualification sets. The approved specifications for the scoring guides are presented in Table 19. The master copies for each scoring guide (for table leaders, scoring supervisors, and scoring leadership) included the Maine DOE range-finding committee's comments and annotations.

Table 19. Specifications for Scoring Sets

Set Name	Set Specifications
Anchor Set	• 1 client-approved sample response per score point • May also include a second sample for each score point if there is more than one plausible way to illustrate the merits of a score point

Set Name	Set Specifications
Practice Set	• 1 set that includes a selection of 10–15 responses designed to help establish the full score point range and the range of possible responses within each score point
Qualification Sets 1 and 2	• 10 responses per set that are comparable to anchor sets

Flagging Criteria for Field-Test Items

Items were flagged as Red, Yellow, or Green based on their item statistics as seen in Table 20.

- Red items are expected to be rejected. The data review committee may review red items to inform future item development; however, these items typically do not progress past data review.
- Yellow items will be accepted or rejected based on the recommendations of the data review committee after their review.
- Green items have item statistics that are acceptable for all categories. The data review committee may use these items as a reference in reviewing Yellow- or Red-flagged items.

Table 20. Data Review Flagging Criteria for Field-Test Items

Analysis	Classification		
	Red	Yellow	Green
p-value	< 0.05	> 0.90 *or* 0.05 ≤ p-value < 0.25	0.25 ≤ p-value ≤ 0.90
Item-Total Correlation (ITC)	< 0.10	0.10 ≤ ITC < 0.25	≥ 0.25
Item Response Theory (IRT)	Could not compute	—	—
Differential Item Functioning (DIF)	—	C Flag	A or B Flag

Scorer Recruitment & Qualifications

For scoring the Maine Science Assessment, SME sought a diverse pool of scorers with a broad range of backgrounds, including current teachers, retired teachers, scientists, and graduate students. SME used a proprietary multistep recruiting and screening process to identify successful scoring candidates.

At a minimum, all scoring staff had a B.S. or B.A. degree and were experienced with the nuances of rubric application and scoring. The highest level of education for the scorers is presented in Table 21.

Table 21. Scorer Educational Degrees

Degree	Scoring Leadership	Scorers
Doctorate	1	1
Master's	3	10
Bachelor's	2	27

All scorers were required to sign confidentiality and nondisclosure agreements. Once selected as a potential scorer for the project, all scorers had to pass the project-specific training and qualification process before scoring live student responses. The qualification process is described in more detail in section 4.2.6.

Methodology for Scoring Polytomous Items

The range of possible score points for the polytomous items on the Maine Science Assessment was 0 to 2. Scoring procedures for polytomous items included both single-scoring and double-blind scoring. Each response received at least one score. As described in more detail in section 4.2.8, 10% of all responses were independently scored by a second scorer who was not provided any information regarding the first assigned score. The first and second scores were compared, and the results were used to calculate the level of inter-rater agreement.

Responses that could not be assigned a numeric score based on the rubric were assigned a non-score code. The list of valid non-score codes appear in Table 22[1]. Non-score codes assigned by scorers were included in read-behind protocols and were monitored by scoring leadership to ensure accuracy and consistency in scores. Responses that received a non-score code counted as zero points toward student scores.

Table 22. Non-Score Codes

Non-Score Code	Non-Score Code Explanation
B	Blank (no attempt to respond)
U	Unreadable (illegible, incoherent, random keystrokes [online only])

Scorer Training

Once selected as a potential scorer, all scorers were required to pass the initial training and qualification process before scoring live student responses. SME selected experienced scoring staff to serve as trainers, and their presentation of the scoring materials to potential scorers was closely monitored by the scoring director and scoring manager. The scorer training for each item was based on the scoring guides developed from the responses scored during prior range-finding sessions.

Scoring training began with an introduction of the scoring staff and an overview of the purpose, goals, and guidelines for the Maine program. This included a discussion about the security, confidentiality, and proprietary nature of all scoring materials, student responses, and procedures.

The scoring training was structured around a three-step process during which the trainer facilitated a review of anchor, practice, and qualification sets. SME administered practice and qualification sets electronically via OSCAR, which provided a concrete record of training and qualification outcomes and allowed trainers to identify challenging content and when scorers needed additional training.

Anchor Set. The scoring trainer used the anchor set to introduce the item and scoring rubric to scorers and to calibrate scorers to the response criteria required to achieve each score point. To accomplish this, the scoring trainer

[1] Other types of responses not meeting the rubric requirements, including off-topic, direct copy of the prompt, or language other than English, were assigned a score of zero.

reviewed the validated annotation for each response with the scorers and explained how the student response mapped onto the requirements listed in the scoring rubric.

Practice Set. Once the anchor set was reviewed and the trainer answered all content or rubric questions, the scoring trainees scored the responses in the practice set. The practice set responses were delivered electronically in OSCAR. Trainees logged into OSCAR and accessed the practice set associated with the item on which they just received training. The trainees independently read and scored each practice set response using the online system. After reviewing the group practice set scores to see if there were common errors or misunderstandings, the trainer facilitated a group discussion to review the practice set on a response-by-response basis. The trainer reviewed each response in detail and discussed the Maine DOE–approved score and annotation and explained why the response received the score it did.

Qualification Sets. Following the presentation of the anchor set responses and the scoring and discussion of the practice set, scoring trainees demonstrated their ability to apply the scoring criteria by qualifying (i.e., for operational scoring) with acceptable agreement to the true scores on the qualification sets. The selected qualification responses covered all score points on the targeted rubric and were representative of the range of possible responses. The specific qualifying criteria provided by Maine DOE were as follows:

- Responses were scored with at least 80% exact agreement and at least 90% exact or adjacent agreement on at least one qualifying set.
- Scorers were allowed 1 discrepant score (i.e., 1 score out of 10 that was more than one score point from the predetermined true score), provided they had at least 8 exact scores.

All scorers took both qualifying sets. Upon completion of Qualification Set 1, the trainer reviewed the group scores to see if there were common errors or misunderstandings. The trainer then facilitated a discussion of each response and explained the true score. This process was repeated for Qualification Set 2. A scorer had to qualify on at least one set to become eligible to score a particular item. The scoring platform was configured to lock out a user if the qualification criteria were not met. Trainees not meeting the qualification standard were either dismissed from the item and given the opportunity to train on a different item, or they were dismissed from the project scoring team.

Retraining. Individual scorers might receive retraining during the scoring process if deemed necessary by the table leader or scoring supervisor observations and/or from the results of various reports. More specifically, the need for retraining was identified if scorers had a large number of nonadjacent scores (e.g., on the 10% of responses requiring a second read), unsatisfactory exact agreement rates, or anomalies detected during the read-behind process.

Retraining by scoring leadership involved several techniques:

- Discussion of student response(s) and the scores involved in the resolution
- Discussion of specific responses identified during the read-behind process
- Review and discussion of anchor papers

Scoring Leadership Training

Prior to beginning the scoring process, SME's scoring manager conducted leadership training for scoring supervisors and table leaders. The scoring supervisors and table leaders were expert scorers who had experience in all facets of scoring. Scoring supervisors were assisted by table leaders, and both were responsible for carefully monitoring the scoring accuracy of all scorers on their assigned team. Scoring supervisors and table leaders are the next-level experts regarding the prompts and the scoring requirements and procedures for the project.

During the leadership training sessions, the logistics of the scoring sessions and scoring routines were discussed. This included the criteria by which scorers would qualify, procedures for monitoring accuracy and reliability, and procedures for retraining and evaluating scorers on their team.

Scoring Quality Control Methods

Scorers were required to demonstrate and maintain their ability to score student responses accurately and consistently throughout the scoring window. SME used several quality assurance techniques to ensure that scoring was valid and reliable for the duration of the scoring window:

- Creating small scoring teams
- Embedding validity papers
- Implementing read-behind protocols
- Implementing double-blind scoring
- Implementing recalibration sets

Small Scoring Teams. For scoring the Maine Science Assessment, the ratio of table leaders to scorers was approximately 1:8. Maintaining the ratio of table leaders to qualified scorers to below 1:10 allowed the table leader to meaningfully observe and interact with each member of their assigned scoring team and to intervene when questions or concerns arose.

Embedded Validity Papers. Embedded validity papers were reviewed by Maine DOE and SME during the range-finding process and assigned Maine DOE–approved scores. These validity responses were loaded into OSCAR and automatically inserted into the scoring queue so that they did not distinguish themselves from the live student responses.

Eight to ten embedded validity papers were distributed at random throughout the first full shift of scoring to ensure that scorers were sufficiently calibrated at the beginning of the scoring period. After submitting a score for an embedded validity paper, scorers received immediate confirmation or corrective feedback. The feedback included the true score and a brief annotation to highlight why the response received the score that it did. Embedded validity papers were used for all constructed-response prompts.

Read-Behind Protocols. Table leaders, under the supervision of a scoring supervisor, were responsible for reading behind each scorer on his or her scoring team. As an additional quality assurance check, scoring supervisors conducted additional read-behinds to monitor table leader performance.

Read-behinds were conducted at a rate of at least 5–10% per scoring shift. If a scorer was struggling or falling below the expected rate of agreement, additional read-behinds were conducted. The OSCAR scoring platform randomly selected responses scored by each scorer and directed those responses to the table leader or scoring supervisor for review. Table leaders could see the score assigned by the original scorer for each reviewed response. During read-behinds, table leaders looked for scoring patterns or issues requiring clarification and addressed issues on an individual or a group basis. Percentages of read-behinds conducted for each item are provided in Appendix H.

If the table leader determined that a response had been scored incorrectly, he or she provided the correct score, appropriate feedback, and/or retraining to the initial scorer. This retraining focused on using the language of the rubric and referring to the appropriate anchor, practice, or qualification papers. A score that was changed by a table leader (or any scoring leadership) became the new score of record.

The scoring director, scoring manager, and scoring content specialist monitored the status of read-behinds and monitored any score changes applied by table leaders or scoring supervisors to ensure consistency and accuracy across all scores.

Double-Blind Scoring. OSCAR was configured to automatically select 10% of student responses to all constructed-response prompts to be double scored. This double-blind scoring was used to calculate inter-rater agreement rates that scoring leadership used to monitor accuracy and consistency. In OSCAR, these second reads are tagged as “reliability papers” and are equally distributed across scorers throughout the scoring window for a particular item. Appendix H presents the percent exact and exact/adjacent agreement between scorers for each item by grade.

The reliability papers (i.e., second reads) with discrepant first and second scores were automatically flagged in OSCAR for a third score resolution. Resolution papers were reviewed on an ongoing basis by scoring leadership, and the scoring supervisor assigned the appropriate resolution score and provided immediate feedback to the scorer who assigned the discrepant score. The resolution score (e.g., third score) assigned by the scoring supervisor became the official score of record. The scoring manager and scoring content specialist monitored resolution scores applied by scoring supervisors to ensure consistency and accuracy across all scores.

If a scorer fell below the expected rate of agreement (e.g., 80%), the scorer was retrained or removed from the item. If a scorer was retrained, the scoring manager and scoring supervisor reviewed all scores assigned prior to retraining to determine if those scores should be deleted. If the scores were deleted, the responses were returned to the scoring queue and rescored by a different scorer. If a scorer was removed from an item, his or her scores were deleted, and the responses were returned to the scoring queue and rescored by a different scorer.

Recalibration Sets. If scoring for a particular item extended past one day, scorers were required to take an online recalibration set to determine if they were still calibrated to the scoring standards. Each recalibration set consisted of approximately five responses representing the entire range of possible scores. Any recalibration results that showed discrepant scores, or two or more adjacent scores, required a review with scoring leadership before the scorer could continue scoring. Recalibration sets were used, as needed, for all constructed-response prompts.

Scoring Quality Control Reports

OSCAR includes multiple quality control tools and reports that provide detailed data for scoring leadership. These scoring metrics, including scorer performance and reliability, were available in real-time for users with authorized roles and allowed staff to constantly monitor the accuracy, consistency, and productivity of scoring.

The reports, generated by individual scorer and the scoring team, provided the results of scoring on an ongoing basis. The information in these reports included the number of responses scored by the reader during a specific period, scorer agreement (or reliability) rates, score point distribution by item/prompt, and other useful metrics.

The following reports were generated and used each day by SME scoring leadership (including table leaders, scoring supervisors, the scoring content specialist, the scoring manager, and the scoring director). They were also posted daily for Maine DOE review:

- **Completion Report:** This report is designed to show the real status of every response loaded into the scoring system. A scoring supervisor or administrator can quickly see the state of all responses and how close an item/project is to completion. Included in this report is the total number of responses by grade and item. This report details responses that are unscored, withheld for supervisor review, waiting for a second read, in third-score resolution, flagged, complete, backread, and requiring resolution.
- **Scorers Report:** This report can be run by section (grade) or across all sections for one item or all prompts and for a specified date or date range. It lists the total score time, average score time, scoring rate, number of scores assigned, number of resolution responses (i.e., 1st and 2nd read discrepancies), number of validity and calibration response scores, and percentage of resolutions required.
- **Daily Report:** This report is run by section (grade) and item (or across all prompts) and includes additional filters for team, trait, user, or date range. For each scorer, it lists the total number of responses scored, the average scoring time, the score point distribution of assigned scores, and the percentage of exact and adjacent agreement (for reliability responses).
- **User Summary:** This report provides a detailed view of individual scorer performance for a particular item and includes a summary of practice, qualification, validity, and calibration set scores; the total number of scores assigned; scoring time and average scoring rate; inter-rater reliability (IRR); exact and exact/adjacent validity; exact and exact/adjacent agreement; percentage of resolutions required and changed; and a summary of recent activity in the platform.
- **Item Summary:** This report is run by grade and item across users or for a specific user and allows for the following comparisons of scores in an agreement matrix: 1st vs. 2nd, 1st vs. resolution, 2nd vs. resolution, 1st vs. backread, and 2nd vs. backread. The report also provides IRR by score and trait (if applicable).
- **User Agreement:** This report, used by table leaders and scoring supervisors, is run by grade and item and provides a summary by user and across all users of IRR (exact and exact/adjacent), validity (exact and exact/adjacent), and resolution (required and disagreed).
- **Project Agreement:** This report is used by scoring leadership to summarize IRR (exact and exact/adjacent) and validity scores (exact and exact/adjacent) for the project on an item-by-item basis.
- **QC Reports:** The QC reporting dashboard is utilized by scoring leadership to review scores for the sets used for training, qualification, validity, and calibration. The dashboard summarizes scorer performance at the item level, making it easy to identify patterns and responses that require further clarification.

4.3. Quality Assurance of Results

Rigorous quality control procedures were implemented throughout the test development, administration, scoring and analyses phases.

Quality Control of Scoring

Hand-scoring quality control processes were described in detail throughout section 4.2. Scorers were required to demonstrate and maintain their ability to score student responses accurately and consistently throughout the scoring window. OSCAR scoring metrics, including scorer performance and reliability, were available in real-time for users with authorized roles (i.e., scoring leadership) and allowed staff to constantly monitor the accuracy, consistency, and productivity of scoring.

Furthermore, for machine-scored items New Meridian conducted statistical key check and item-response adjudication reviews to verify that items were properly scored according to the rules in ADAM prior to item analyses and calibrations.

Chapter 5. Classical Item Analysis

This chapter describes the results of the classical item analyses conducted from the scored operational items that were selected after the preliminary item analyses and Maine DOE content review. Item analyses included data from the following item types: key-based selected-response items, rule-based machine-scored items, and hand-scored constructed-response prompts. For each item, analyses were conducted at both the item level and response option level. These analyses included difficulty (p-value and pseudo p-values), discrimination (item-total correlations), and frequencies (proportions of students selecting each option or obtaining each score point). Differential Item Functioning (DIF) analyses and student testing time analyses were also conducted.

Item analyses were conducted by test form based on administration mode.[2] Paper-based results are not included in the report due to the small sample size.

5.1. Final Item Analyses and Calibration Data Screening Criteria

Student results files were available in a single file layout. That file contained both student item-level data and test-level data. A single record contains all test information for a student, including demographic variables, form identification, item scores, and total raw scores, as well as the student responses and scores for each item and for separate parts of composite items (when applicable). Some parts of a composite item will have scores if there is a one-to-one relationship between the number of item parts and the overall score for the composite item.

For this administration, there was a limited number of paper-based forms (i.e., mode of administration for paper, large print, or braille). Rather than scanning the test documents into the published paper-based forms, student results were key-entered into the online version of the assessment. Therefore, the results for these assessments were recoded to correspond to the correct session form names, unique item number (UIN) for the paper version of TEIs, and any item time associated with the assessment stripped from the results file prior to analysis. The crosswalk for the correct UINs is generated from the approved test maps.

Analyses were performed on an Incomplete Data Matrix (IDM) that was generated from the results file. These analyses were conducted by form. Student records were removed prior to running the analyses if the records met any of the following criteria:[3]

Table 23. Data Exclusion Criteria

Exclusion	Criteria
Invalid Grade	Student Grade ≠ Form Name Grade
Invalid Test Status	SciTestStatus other than A
Invalid Test Session Status	Flagged SciInvSes1, SciInvSes2, or SciInvSes3
Duplicate Record[4]	Duplicate StateStudentID
Invalid Attemptedness	Did not complete at least 25% of test items

[2] Paper-based forms did not have sufficient volumes of students to generate meaningful statistical interpretation.

[3] Criteria for the key check/adjudication screening were slightly different since these analyses are intended as diagnostic of potential scoring issues.

[4] If a student has duplicate valid records, only the record with the higher raw score is retained.

Exclusion	Criteria

Invalid Mode of Administration	Paper: TestType = 2 but AccomPaper = 0;
	Large Print: TestType = 3 but AccomLargePrint = 0;
	Braille: TestType = 4 but AccomBraille = 0;
	Online: TestType = 1 but any of above accommodations = 1.

Items may not be scored due to a student omitting the item or the student not yet reaching an item within the test. "Omitted" (i.e., skipped) items are items for which a student did not provide a response when items coming before and after have student responses. These are designated with "?" in the response file. "Not administered" (i.e., "not reached") items are those at the end of the session for which no responses were provided—items that the student probably did not reach during the administration—or the items of an entire session that were not administered. Item response scores for "omits" are re-coded as '0' in the CTT analyses and IRT IDM files, whereas "Not reached" and "Not administered" items are considered missing and therefore do not contribute to the statistics.

5.2. Classical Difficulty and Discrimination Indices

Item p-value (Pseudo p-value for Polytomous Items)

The p-value represents the mean item score as a proportion of the maximum obtainable score points, indicating the item difficulty. Values range between 0 and 1. Higher values indicate easier items while lower values indicate more difficult items. For dichotomous items (item scored as either correct = 1 or incorrect = 0) the formula is

$$p\text{-}value = \bar{x}_i = \frac{1}{n} \cdot \sum_{1}^{n} x_i$$

where x_i are the individual student item scores on item *i* and *n* is the total number of students for whom the item was administered.

For polytomous items, the pseudo p-value is calculated by further dividing by the maximum obtainable points possible for the item:

$$pseudo\ p\text{-}value = \frac{\bar{x}}{T}$$

where $\bar{x}$ is the mean item score and *T* is the maximum obtainable points possible for the item.

Frequently, the p-value is reported as a percentage by multiplying by 100. For instance, a p-value of 0.67 means that 67 percent of the students answered a dichotomous item correctly.

Response Option or Score Point Proportions

A dichotomous item's alternate response options (i.e., distractors) are plausible but incorrect options that are included to test common misconceptions or miscalculations. Ideally, all response options should garner a proportion of student selections. These are calculated by the simple proportion formula:

$$proportion = \frac{N_O}{N_T}$$

where N_O is the number of students selecting the specific option and N_T is the total number of students for whom the question was administered, including those who did not record a response (i.e., omitted the item).

In the case of polytomous items, the numerator becomes the number of students obtaining the specific score point (N_{SP}):

$$proportion = \frac{N_{SP}}{N_T}.$$

Item-Total Correlations

The item-total correlation is the relationship between students' performance on the item and students' performance on the criterion.[5] Possible values range between −1 and +1. The correlation will be positive when the mean test score of the students answering the item correctly is greater than the mean test score of the students answering the item incorrectly. Negative values may indicate that an item has multiple correct answers or an incorrect answer key.

The point-biserial correlation (Crocker & Algina, 1986) is one possible item-total correlation for dichotomously scored items. However, the correlation will be spuriously high because the item of interest is also included in the total test score (i.e., correlating with itself; Henrysson, 1963). Therefore, a correction is made by using the means with the item deleted (i.e., the total operational test score not including the item of interest) from the calculation

$$r_{\text{pbis}} = \frac{(\bar{M}'_{+} - \overline{M'})}{S'} \sqrt{{}^{p}/_{(1-p)}}$$

where $\bar{M}'_{+}$ is the mean score with the item deleted for students who answered the item correctly, $\overline{M'}$ is the mean score with the item deleted for all students, S' is the standard deviation with the item deleted for all students, and p is the item p-value (difficulty).

The Pearson correlation (polyserial) with the item of interest deleted is typically calculated for polytomous items by this equation:

$$r = \frac{\sum(x_i - \bar{x})(y'_i - \bar{y}\prime)}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y'_i - \bar{y}\prime)^2}}$$

where x_i is the student score point on the item, $\bar{x}$ is the mean score for the item, y'_i is the total score with the item deleted for the student, and $\overline{y'}$ is the mean total score with the item deleted for all students (Lemke & Wiersma, 1976).

Response Option or Score Point Correlations

Like the overall item point-biserial correlation calculation, a correlation can be calculated for each incorrect response option (*O)* for multiple-choice single-response items, or the score point in the case of other item types, using the generalized formula

$$r_{pbis_O} = \frac{(\bar{M}_O - \overline{M'})}{S'} \sqrt{{}^{p_O}/_{(1-p_C)}}$$

[5] For the key check, it is the machine-scorable total raw score. Otherwise, it is the total test raw score.

where $\overline{M}_O$ is the mean score for students who selected the distractor, $\overline{M}$ is the mean score for all students with the item deleted, S' is the standard deviation of all students with the item deleted, p_O is the proportion of students selecting the distractor, and p_C is the proportion of students selecting the correct response.

Classical Item Analyses Results

Items are flagged during the analysis based on the criteria listed in Table 24. During the earlier preliminary analyses, the flagged items were provided to the science test development manager for review. The Maine DOE then reviewed all items to determine the final item status (i.e., scored, or unscored).

Table 24. Preliminary Item Analysis Flagging Criteria

Analysis	Criteria
p-value (pseudo p-value)	> 0.95 *or* < 0.20
Item-Total Correlation	> 0.25
Distractor-Total Correlation	< 0.00
Omits	> 5% omit rate for a dichotomous item > 15% omit rate for a polytomous item
Polytomous Item Score Distribution	A low percentage (<3%) of students obtaining a score point *or* no students obtaining a score point

Tables 25 through 27 present the aggregated results for the classical item analyses statistics by item type and for the operational form. Item-level statistics are provided in Appendix J.

Table 25. Grade 5 Summary of Item Difficulty and Discrimination Statistics

Form Name	Statistic	N	Minimum	Maximum	Mean	Std. Dev.
SP2024_GR05_SS1	p-value	13	0.175	0.797	0.492	0.173
SP2024_GR05_SS2	p-value	10	0.299	0.598	0.422	0.091
SP2024_GR05_SS3	p-value	11	0.179	0.720	0.444	0.156
Total Test	p-value	34	0.175	0.797	0.456	0.146
SP2024_GR05_SS1	Item-Total Correlation	13	0.173	0.449	0.361	0.071
SP2024_GR05_SS2	Item-Total Correlation	10	0.262	0.540	0.377	0.095
SP2024_GR05_SS3	Item-Total Correlation	11	0.143	0.500	0.375	0.103
Total Test	Item-Total Correlation	34	0.143	0.540	0.370	0.087

Table 25 provides statistics related to the grade 5 test, specifically focusing on item difficulty (p-value) and discrimination (item-total correlation).

The "Total Test" combines data from all three forms. There were 34 operational items on the Grade 5 test. The total test p-values range from 0.175 to 0.797, with a mean of 0.456 and a standard deviation of 0.146, demonstrating a range of item difficulties. The total test has a mean Item-Total Correlation of approximately 0.370 with a standard deviation of 0.087. These values reflect how strongly each individual item on the test is related to overall test performance.

Table 26. Grade 8 Summary of Item Difficulty and Discrimination Statistics

Form Name	Statistic	N	Minimum	Maximum	Mean	Std. Dev.
SP2024_GR08_SS1	p-value	17	0.157	0.791	0.464	0.161
SP2024_GR08_SS2	p-value	12	0.329	0.609	0.445	0.081
SP2024_GR08_SS3	p-value	12	0.149	0.662	0.430	0.167
Total Test	p-value	41	0.149	0.791	0.448	0.142
SP2024_GR08_SS1	Item-Total Correlation	17	0.160	0.569	0.345	0.116
SP2024_GR08_SS2	Item-Total Correlation	12	0.244	0.534	0.392	0.096
SP2024_GR08_SS3	Item-Total Correlation	12	0.208	0.516	0.354	0.110
Total Test	Item-Total Correlation	41	0.160	0.569	0.361	0.108

Table 26 provides statistics related to the grade 8 test, specifically focusing on item difficulty (p-value) and discrimination (item-total correlation).

There were 41 operational items on the Grade 8 test. Total test p-values range from 0.149 to 0.791, with a mean p-value of 0.448 and a standard deviation of 0.142. The Item-Total Correlation values range from 0.160 to 0.569, with a mean of 0.361 and a standard deviation of 0.108.

Table 27. High School Summary of Item Difficulty and Discrimination Statistics

Form Name	Statistic	N	Minimum	Maximum	Mean	Std. Dev.
SP2024_GRHS_SS1	p-value	17	0.165	0.708	0.523	0.134
SP2024_GRHS_SS2	p-value	13	0.223	0.700	0.415	0.163
SP2024_GRHS_SS3	p-value	11	0.123	0.633	0.367	0.163
Total Test	p-value	41	0.123	0.708	0.447	0.162
SP2024_GRHS_SS1	Item-Total Correlation	17	0.172	0.605	0.424	0.127
SP2024_GRHS_SS2	Item-Total Correlation	13	0.250	0.544	0.367	0.109
SP2024_GRHS_SS3	Item-Total Correlation	11	0.183	0.626	0.405	0.130
Total Test	Item-Total Correlation	41	0.172	0.626	0.401	0.122

Table 27 provides statistics related to the high school test, specifically focusing on item difficulty (p-value) and discrimination (item-total correlation).

There were 41 operational items on the high school test. Total test p-values range from 0.123 to 0.708, with a mean p-value of 0.447 and a standard deviation of 0.162. The item-total correlation for the total test ranges from 0.172 to 0.626, with a mean correlation of 0.401 and a standard deviation of 0.122.

5.3. Differential Item Functioning

Differential item functioning (DIF) is a procedure that matches students based on total test scores to compare the performance of two similarly performing groups of students. The procedure identifies two contrasting groups (i.e., focal and reference) for which differences in item performances are computed. Table 28 indicates potential comparison groups dependent on sufficient volumes of students (i.e., at least 100 in the focal group and at least 300 in the reference group). For the procedures described next, positive values indicate that, for students of similar ability, the focal group has a higher mean item score than the reference group. Negative DIF values indicate that, for students of similar ability, the focal group has a lower mean item score than the reference group.

Table 28. DIF Comparison Groups

Comparison Type	Focal Group (N≥100)	Reference Group (N≥300)
Gender	Female	Male
Ethnicity	African American	White
	Asian	White
	American Indian/Alaska Native	White
	Hispanic	White
	Pacific Islander	White
	Multiple	White
Economic Status	Economically Disadvantaged	Not Economically Disadvantaged
Multilingual Learners	Multilingual Learner	English Proficient (including former English learners)
Students with an IEP	IEP	No IEP

Dichotomous Items: Mantel-Haenszel

The Cochran-Mantel-Haenszel (*MH*) chi-square approach (Mantel & Haenszel, 1959) is used to detect DIF in dichotomously scored, one-point items. The range of total scores is divided into 10 stratifications (*S*), and those strata are used to match samples from each group based on student performance. Contingency tables (such as Table 29) for each stratum are constructed for the responses to the item in which *S* represents the strata, W_{rs} and W_{fs} represent the number of students (in the reference and focal groups, respectively) who answer the item incorrectly, R_{rs} and R_{fs} represent the number of students (in the reference and focal groups, respectively) who answer the item correctly, and N_{ts} represents the total number of students ($W_{rs} + R_{rs} + W_{fs} + R_{fs}$).

Table 29. Mantel-Haenszel Contingency Table

Score Stratum (S)	Incorrect/Wrong (0)	Correct/Right (1)	Total
Reference	W_{rs}	R_{rs}	$W_{rs} + R_{rs}$
Focal	W_{fs}	R_{fs}	$W_{fs} + R_{fs}$
Total	$W_{rs} + W_{fs}$	$R_{rs} + R_{fs}$	N_{ts}

A common odds ratio is computed across all intervals of matched groups using the following formula (Dorans & Holland, 1993):

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S} {R_{rs}W_{fs}}/{N_{ts}}}{\sum_{s=1}^{S} {R_{fs}W_{rs}}/{N_{ts}}}.$$

Furthermore, the Mantel-Haenszel delta statistic (*$MH_{D\text{-}DIF}$*) (Holland & Thayer, 1988) is computed to measure the degree and magnitude of DIF using the following formula:

$$MH_{D\text{-}DIF} = -2.35\ln(\hat{\alpha}_{MH}).$$

Polytomous Items: Standardized Mean Difference

For polytomous items, the *$MH_{D\text{-}DIF}$* is not calculated. Rather, a standardized mean difference (*SMD*) is calculated using a contingency table that extends the possible item scores beyond 1 point using this formula:

$$SMD = \sum_{s} w_{Fs} m_{Fs} - \sum_{s} w_{Fs} m_{Rs}$$

where $w_{Fs=}\ n_{F+s}/n_{F++}$ is the focal group proportion at the s^th^ stratification variable; $m_{Fs} = (1/n_{F+s})F_s$ is the focal group's mean item score in the s^th^ stratum; and $m_{Rs} = (1/n_{R+s})R_s$ is the reference group's mean item score in the s^th^ stratum. Because the focal group proportion is used in both terms of the equation, the reference group's item mean is weighted, whereas the focal group's item mean is unweighted.

The effect size (*ES*) is then computed by dividing by the total group standard deviation (*SD*) using this equation:

$$ES = \frac{SMD}{SD}.$$

By using Mantel's chi-square statistic (1963), the magnitude of the *ES* is interpreted using Golia's (2012) rules.

DIF Classification

Based on the DIF statistics and significance tests, items are classified into three categories: A, B, or C (as in Table 30). Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, while Category C items possess moderate to large DIF values. Items flagged with C-DIF during the preliminary analyses were provided to both the science test development manager and the accessibility, accommodations, and fairness (AAF) specialist as part of the preliminary analysis communication plan.

Table 30. DIF Classifications

Analysis	Criteria
Differential Item Functioning (DIF)	+ Favors the focal group – Favors the reference group
Mantel-Haenszel	A. Negligible – MH is not significantly different from 0 OR (MH is significantly different from 0 AND has a delta absolute value < 1). B. Slight to Moderate – MH is significantly different from 0 AND the absolute value of delta is < 1.5 AND has a delta absolute value greater than or equal to 1. C. Moderate to Large – MH is significantly different from 1 AND delta has an absolute value greater than or equal to 1.5.
Standardized Mean Difference	A. Negligible – is not significantly different from 0 OR has an absolute value ≤ 0.17. B. Slight to Moderate – is significantly different from 0 AND 0.17 < |ES| ≤ 0.25. C. Moderate to Large – is significantly different from 0 AND has an absolute value > 0.25.

Differential Item Functioning Results

Tables 31 through 33 present the C-DIF results for each assessment[6] Appendix K provides the DIF classification for items exhibiting B- or C-DIF by focal group. All operational items on the form were evaluated by educator panels at data review after field testing prior to becoming operational. Although there were some operational C-flagged items, reviews by New Meridian and Maine DOE staff did not find inherent bias in the way those items were written. Items that exhibited C-DIF will be limited in future form development to the extent possible.

Table 31. Grade 5 C-DIF

Form Name	No C-DIF Flags	At Least One C-DIF Flag	Percent with C-DIF
SP2024_GR05_SS1	13	—	0%
SP2024_GR05_SS2	10	—	0%
SP2024_GR05_SS3	10	1	9%
Total Test	33	1	3%

Table 32. Grade 8 C-DIF

Form Name	No C-DIF Flags	At Least One C-DIF Flag	Percent with C-DIF
SP2024_GR08_SS1	17	—	0%
SP2024_GR08_SS2	12	—	0%
SP2024_GR08_SS3	12	—	0%
Total Test	41	—	0%

Table 33. High School C-DIF

Form Name	No C-DIF Flags	At Least One C-DIF Flag	Percent with C-DIF
SP2024_GRHS_SS1	17	—	0%
SP2024_GRHS_SS2	11	2	15%
SP2024_GRHS_SS3	11	—	0%
Total Test	39	2	5%

5.4. Student Testing Time Analyses

As noted in Chapter 2, each assessment was administered as three 60-minute sessions. To ensure that students were provided sufficient time to answer all items and that their assessment experience was not speeded (i.e., not enough time provided for administration), time on items (both scored and unscored) was aggregated at the session level and compared to the allotted 60 minutes. Maine DOE required that 80 percent of students complete each session within the allotted time. Not all students attempted all three sessions. For this analysis, students who did not answer all items within a session were omitted to exclude extremely short session times due to few attempted items. For grade 5, this excluded between 4–5% of students across sessions; for grade 8, between 6–8% were

[6] The American Indian/Native American and Alaska Native/Pacific Islander focal groups did not reach the minimum threshold for analyses and were excluded from these tables.

excluded; and in high school, between 15–18% of students were excluded for not completing all items in a test session. Higher numbers in high school may reflect lower motivation in these grades.

Table 34. Summary of Time in Minutes to Complete Each Session

Grade	Session	N	Minimum	Median	Percentile 80	Maximum	Mean	Std. Dev.
05	1	11,779	2.0	21.0	30.1	100.5	22.8	9.9
05	2	11,673	1.4	23.0	32.8	133.4	24.7	11.1
05	3	11,653	1.9	25.1	35.8	155.2	26.7	11.9
08	1	11,833	1.1	25.7	34.4	121.0	26.6	10.7
08	2	11,804	0.8	22.7	31.9	175.3	23.9	11.2
08	3	11,647	1.1	19.6	27.4	107.7	20.7	9.5
HS	1	11,133	0.9	23.6	31.4	155.9	24.1	10.7
HS	2	10,888	1.3	20.3	28.0	150.3	20.8	10.2
HS	3	10,712	1.1	17.4	24.9	143.3	18.2	9.4

Table 34 shows the time taken in minutes to complete sessions categorized by grade levels (grade 5, grade 8, and high school). Each grade level has three sessions labeled as session 1, 2, or 3.

In Grade 5, the number of observations (N) was consistent across sessions (approximately 11,600 to 11,800) and the completion times varied from around 1.4 to 2.0 minutes (minimum) and 22.0 to 25.1 minutes (median). The maximum time taken for a session was around 100.5 to 155.2 minutes. The average time taken ranged from 22.8 to 26.7 minutes. 80 percent of students completed their sessions in 30.1 to 35.8 minutes.

In grade 8, as with grade 5, the number of observations was around 11,600 to 11,800 for each session. Completion times varied from 0.8 to 1.1 minutes (minimum) and 19.6 to 25.7 minutes (median). The maximum time taken ranges from 107.7 to 175.3 minutes. The average time taken ranged from 20.7 to 26.6 minutes. 80 percent of students completed their sessions 27.4 to 34.4 minutes.

The number of observations was around 10,700 to 11,100 for each session in high school. Completion times varied from 0.9 to 1.3 minutes (minimum) and 17.4 to 23.6 minutes (median). The maximum time taken ranges from 143.3 to 155.9 minutes. The average time taken ranged from 18.2 to 24.1 minutes. 80 percent of students completed their sessions in 24.9 to 31.4 minutes.

On average, testing times were fairly consistent across grades. Mean times were consistently lower than 30 minutes. 80 percent of students completed each session in 25 to 35 minutes, well below the 60-minute target desired. Across grades, the results presented in Table 34 provide evidence that speededness is not a significant concern for the test.

Chapter 6. Calibration and Scaling

This chapter describes the procedures used to calibrate and scale the Maine Science Assessment online forms. Calibration and scaling were conducted according to the processes outlined in the psychometric OPM (Appendix I). The Rasch and Partial Credit (PC) models were implemented using Winsteps Version 5.7.4.0 (Linacre, 2006), as noted in the OPM Addendum.

6.1. Item Response Theory

All Maine Science Assessment forms are calibrated using Item Response Theory (IRT) models. One advantage of using IRT models over classical test theory is that items and students are calibrated to a common scale (see Figure 1).

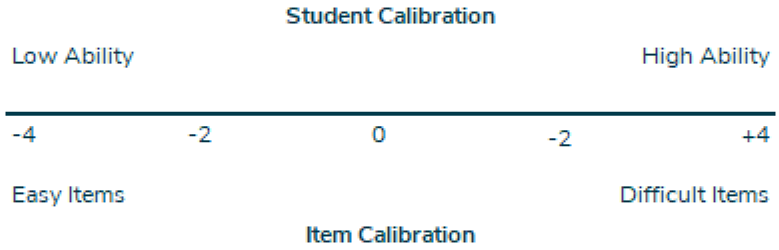


Figure 1. Item Response Theory Calibration

Dichotomous Items

Item response theory for dichotomous items (e.g., items with two score classifications such as 0 or 1) is commonly expressed as a three-parameter logistic model (3PL):

$$P_i(\theta) = c_i + (1 - c_i)\frac{exp[Da_i(\theta - b_i)]}{1 + exp[Da_i(\theta - b_i)]}$$

where $P_i(\theta)$ is the probability that a student gets item *i* correct, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the item lower asymptote (i.e., the guessing parameter), and *D* is a scaling factor approximately equal to 1.701 that generates the normal ogive function (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2014).

The 3PL model is designed for dichotomously scored multiple-choice items. For dichotomously scored items that are not multiple-choice, the guessing parameter is set to 0 because it is assumed that guessing does not occur. Lastly, the dichotomous IRT models are not appropriate for items with more than two score categories (i.e., polytomous items; Kolen & Brennan, 2014).

The Rasch model can be expressed as a reduced 3PL model since the discrimination parameter, a_i, is fixed to 1; the guessing parameter, c_i, is set to 0; and the value of D is set to one for all dichotomously scored multiple-choice items,

$$P_i(\theta) = \frac{exp[(\theta - b_i)]}{1 + exp[(\theta - b_i)]}$$

where $P_i(\theta)$ is the probability that a student gets item *i* correct, and b_i is the item difficulty parameter.

Polytomous Items

The Partial Credit Model (PCM; Masters, 1982) is used for the ordered categorization of responses when there are two or more ordered categories. The PCM is defined as:

$$P_{ih}(\theta) = \frac{exp\left[\sum_{v=1}^{h} Z_{iv}(\theta)\right]}{\sum_{c=1}^{m_i} exp\left[\sum_{v=1}^{c} Z_{iv}(\theta)\right]}$$

and

$$Z_{ih}(\theta) = (\theta - b_{ih}) = (\theta - b_i + d_h),$$

where b_{ih} is an item-category parameter, b_i is an item-location parameter, and d_h is a category parameter. Further, if the number of categories is m_i, then only $m_i - 1$ item-category parameters can be identified, and $b_{i1} \equiv 0$.

6.2. Calibration and Item Response Theory Results

Item and person parameters were estimated using the Joint Maximum Likelihood Estimation (JMLE) method. The default setting of Winsteps was used for all the estimations and calculations. Winsteps calibrations for the grade 5 and high school assessments ended via normal termination using the following parameter estimation controls: Both LCONV= for "logit change size" and RCONV= for "residual size" were controlled. Iteration stopped when the biggest logit change was less than or equal to LCONV=0.00001 and the biggest residual score was less or equal to RCONV=.001, or when both the biggest logit change size increased and the biggest residual size increased (divergence). For grade 8, however, an additional control for the maximum number of JMLE iterations needed to be specified in addition to the controls of LCONV=0.00001 and RCONV=0.001. The maximum number of JMLE iterations was set to 500 (i.e., MJMLE=500). IRT parameters are provided in Appendix L.

Appendix M presents the test characteristic curves (TCCs), test information functions (TIFs), and conditional standard error of measurements (CSEMs) for the assessment overall as well as for each session. In addition to the curves for the 2024 form, Appendix M presents the overall assessment curves overlaid for each administration of the test across years 2022–2024, allowing for direct comparison of the curve data across years.

Test Characteristic Curves (TCCs). The TCC represents the relationship between expected test performance and estimates of the science trait underlying test performance. The *x*-axis represents the underlying trait (referred to as *theta*), and the *y*-axis, which ranges from zero to the maximum possible raw score, represents expected performance on the assessment.

Test Information Function (TIF) Curves. The TIF indicates the amount of information about student ability (as measured by theta) provided by the assessment at different points along the continuum from low to high ability. When an assessment provides more information, reliability (measurement precision) is greater. The peak of the TIF indicates the ability level at which the assessment is the most reliable.

Conditional Standard Error of Measurement (CSEM) Curves. The CSEM indicates the amount of measurement error across the theta scale. Note that the CSEM is the lowest where the test information is greatest. (For a given level of theta, CSEM equals 1 divided by the square root of the test information at that level of theta.)

6.3. Model Assumption Analyses and Results

The Rasch and partial credit models are appropriate when the following assumptions are met:

- Unidimensionality
- Local independency
- The Model fit

Dimensionality

Unidimensionality is one of the essential assumptions of the IRT models commonly used in large-scale summative assessments (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). That is, all items on the assessment are measuring a single construct or a dominant dimension.

Dimensionality analysis was conducted on Spring 2022 data. The need to conduct a dimensionality analysis on Maine's data each year depends on the specific goals and objectives of the analysis, as well as the nature of the data and the changes that may occur over time. Since the 2024 forms have some items in common with the 2023 assessment administration, the percentage of common items between forms varies between 50% to 80% across grades. Since a drastic change also was not seen in the demographic features of the population, dimensionality analyses were not repeated on Spring 2024 data. The results from Spring 2022 showed that the Maine Science Assessments for grades 5 and 8 and high school are essentially unidimensional. For each academic level, while some individual assessment items loaded on two dimensions, there was clearly one single dominant factor, along with few additional small factors. A detailed explanation can be found in the 2022 Technical report.

Model Fit Index

Fit statistics indicate how accurately or predictably data fit the IRT model. Fit statistics for the Rasch model are calculated by comparing the observed empirical data with the data that the Rasch model would be expected to produce if the data fit the model perfectly. The outfit mean-square fit statistic is computed for all scored responses excluding responses in extreme total scores. This is a chi-square statistic divided by its degrees of freedom. The infit mean-square is an information-weighted fit statistic that is more sensitive to unexpected behavior affecting responses to items near the respondent's ability level. The expected value for both statistics is 1, meaning that values near 1 are of least concern and values less than 1 indicate that the response and rating patterns are too predictable and thus redundant, but not of great concern. High values are of greater concern. The interpretation guidelines according to Linacre (2002) for infit and outfit fit statistics for dichotomously and polytomously scored items are given below:

- Values greater than 2.0 "distort or degrade the measurement system."
- Values between 1.5 and 2.0 are "unproductive for construction of measurement, but not degrading."
- Values between 0.5 and 1.5 should be considered "productive for measurement."
- Values below 0.5 are "less productive for measurement, but not degrading."

Table 35. Range of Fit Statistics

	Infit		Outfit	
Grade	Min	Max	Min	Max
05	0.830	1.246	0.781	1.545
08	0.796	1.222	0.741	1.303
HS	0.792	1.687	0.706	1.580

According to Table 35, infit and outfit statistics range from 0.781 to 1.545 for grade 5, indicating that grade 5 science items mostly fit the model well. Although one item evidenced an outfit value slightly above 1.5, it would not degrade measurement. Infit and outfit statistics ranging from 0.741 to 1.303 for grade 8 also indicate good model fit. For high school, infit and outfit statistics ranged between a low of 0.706 to a high of 1.687, with one item having an infit statistic of 1.687 and another item having an outfit statistic of 1.580. Since the highest infit and outfit statistics are between 1.5 and 2.0, the items that have those values for the infit and outfit statistics do not degrade measurement. Given the observed minimum and maximum values of the infit and outfit statistics, it can be concluded that in general, items fit well across all grades.

6.4. Scaling

For each academic level, scaled scores were derived by applying a linear transformation to the students' IRT theta scores calculated using Winsteps based on students' scored item responses. Students' scaled scores are reported as integers. For grades 5 and 8 and high school respectively, the lowest obtainable scaled scores (LOSS) are 6, 3, and 5, and the highest obtainable scaled scores (HOSS) are 80, 90, and 90.

Table 36 reports the scaling constants used for linear transformation of theta scores at each grade level. For the linear transformation, the A constant is the slope, and the B constant is the intercept. Two sets of values, C1 & C2 and C3 & C4, are presented for each grade level. Each set was derived from the scaled scores corresponding to a given pair of ordered cut scores. For instance, the pair of the C1 & C2 cut scores were used to calculate the A and B scaling constants, then the calculation of the A and B constants was repeated based on the values of the pair of the C2 & C3 cut scores. For each grade, final scaled scores were calculated from the first set of scaling constants calculated, i.e., the set based on the C1 & C2 cut scores. The second set of scaling constants for each grade are presented here as evidence of the appropriateness of the scale derived via linear transformation.

Table 36. Reported Scaled score Transformation Constants

	C1 & C2 (Levels 2 & 3)		C2 & C3 (Levels 3 & 4)	
Grade	Slope (A)	Intercept (B)	Slope (A)	Intercept (B)
05	7.43494	36.84758	7.65864	36.75274
08	14.25178	45.48694	14.74926	45.67847
HS	7.94913	39.19714	7.76398	39.21584

Scoring tables are provided in Appendix N. Students were required to answer at least 25% of the assessment questions to receive a scaled score. The distributions of scaled scores are provided in Appendix O.

6.5. Calibration of Subscores

The Maine Science test comprises three subscores, each aligned with specific content areas. For grade 5, subscore 1 covers "Structure and Properties of Matter," subscore 2 focuses on "Matter and Energy in Organisms and Ecosystems," and subscore 3 delves into "Earth's Systems and Space Systems: Stars and the Solar System." For grade 8 and high school, subscore 1 pertains to "Physical Science," subscore 2 relates to "Life Science," and subscore 3 encompasses "Earth and Space Science."

In the individual student reports (ISR), only the achievement levels for each subscore are reported. The numerical raw scores are reported on the School Summary Reports, the SAU Summary Reports, and all CSV Roster Reports. Raw scores for subscores were computed by summing the scores of items within each subscore category. To determine the achievement level of subscores, items relevant to each subscore were calibrated. To calibrate subscores, only items relevant to each subscore were included in the IDM. The calibration process followed the procedure outlined in the OPM (Appendix I). Both Rasch and Partial Credit (PC) models were applied, utilizing Winsteps Version 5.7.4.0 (Linacre, 2006), as specified in the OPM Addendum.

Upon estimating the IRT true score, representing the expected score correct conditional on the person measure (i.e., theta score), these results were then compared to the theta cuts established by the 2022 standard setting process.

6.6. Linking 2024 and 2023 Assessment Forms

The 2024 assessment forms shared similarities with the 2023 forms, with the proportion of common items ranging from 50% to 80% depending on the grade level. A linking procedure was used to calibrate the results on new forms to the scales of the base forms from 2022 using common anchor items. The linking was done for the grades 5 and 8 and high school assessments. In this procedure, the estimated item parameters for the common anchor items that appear on the 2024 forms (and that were already calibrated to the scales of the base forms) were specified as fixed known values for the calibrations of the 2024 forms. This process of anchored calibration led to the calibration of the 2024 forms to the scales of the base forms, respectively for each grade level. Thus, students' performances on the 2024 forms can be interpreted using the same frame of reference established for the scales of the base forms. For each grade level, an item-parameter drift analysis was conducted for each common anchor item using the item displacement statistic. The item displacement statistic was used to evaluate for each common anchor item the possible "drift" (i.e., change) in item difficulty. This statistic shows the difference between the fixed value of item difficulty of the anchored item and what its difficulty value would have been had it not been fixed for 2024. Typically, anchor items with displacement values less than 0.5 logits are unlikely to have much impact on measurement in a test instrument (Linacre, n.d.). Hence, the banked item parameters for such items were specified as fixed known values.

Table 37. Comparison of Grade 5 Assessment Operational Items Across Years

Item	2023	2024	Displacement
1	0.3198	0.3198	-0.1804
2	0.3050	0.3050	-0.3382
3	0.5105	0.5105	-0.0680
4	-0.5733	-0.5733	0.1691
18	0.1926	0.1926	0.0191
19	0.7599	0.7599	-0.0103
20	0.7352	0.7352	-0.3413
21	-0.6673	-0.6673	-0.0572

Item	2023	2024	Displacement
5	-2.0061	-2.0061	0.1637
6	0.1009	0.1009	-0.3162
7	-0.2837	-0.2837	0.1355
8	0.4318	0.4318	-0.2340
9	-1.8259	-1.8259	0.4707
10	-1.5724	-1.5724	0.3756
11	-0.2597	-0.2597	0.1733
12	0.2661	0.2661	0.2920
13	-1.0912	-1.0912	0.0426
14	-0.5700	-0.5700	0.0566
15	-0.1425	-0.1425	0.0404
16	1.1946	1.1946	0.0208
17	0.6873	0.6873	0.1754
22	-0.5502	-0.5502	0.1176
23	-0.0094	-0.0094	-0.0001
24	-0.5602	-0.5602	0.2146
25	1.1138	1.1138	0.2604
26	-0.5893	-0.5893	-0.0006
27	0.1314	0.1314	0.0096
28	0.0302	0.0302	0.0987
29	-0.8583	-0.8583	-0.0838
30	0.4341	0.4341	-0.3846
31	0.2499	0.2499	0.0882
32	1.4548	1.4548	0.0567
33	0.4779	0.4779	-0.0391
34	0.1500	0.1500	0.0713

As shown in Table 37, all displacement values are less than 0.5, indicating that the banked item parameters of the items were specified as fixed known values for the calibration of the grade 5 2024 form.

Table 38. Comparison of Grade 8 Assessment Operational Items Across Years

Item	2023	2024	Displacement
1	-0.2342	-0.2342	0.0097
2	0.2451	0.2451	0.1470
3	-2.2308	-2.2308	0.1836
4	0.8577	0.8577	0.1880
5	-0.6490	-0.6490	0.2303
6	0.0630	0.0630	0.0791
7	-0.4707	-0.4707	0.0609
8	-0.2823	-0.2823	-0.1292
9	-0.2682	-0.2682	0.0622
10	-0.2640	-0.2640	0.1494
11	-0.9544	-0.9544	-0.2690
12	-1.1492	-1.1492	-0.1463
13	0.3835	0.3835	-0.2721
14	-0.8244	-0.8244	-0.0894
21	0.4171	0.4171	-0.1553
22	0.2020	0.2020	0.0708
23	1.7026	1.7026	-0.0802
24	-1.0142	-1.0142	0.0126
25	0.1406	0.1406	-0.0611
26	0.0998	0.0998	-0.2298
27	0.3766	0.3766	-0.0279
28	1.0715	1.0715	-0.0340
29	1.6490	1.6490	0.0836
30	-0.5974	-0.5974	0.3816
31	-0.5223	-0.5223	-0.0918
32	0.2507	0.2507	0.1727
33	-1.3329	-1.3329	0.0429
34	-0.9575	-0.9575	-0.0399

Item	2023	2024	Displacement
15	-0.2342	-0.2342	0.0097
16	0.2451	0.2451	0.1470
17	-2.2308	-2.2308	0.1836
18	0.8577	0.8577	0.1880
19	-0.6490	-0.6490	0.2303
20	0.0630	0.0630	0.0791
35	0.2814	0.2814	0.0116
36	-0.4664	-0.4664	-0.1600
37	-1.3594	-1.3594	0.2760
38	0.4921	0.4921	0.1700
39	-0.7062	-0.7062	-0.3194
40	0.0701	0.0701	-0.1528
41	0.1668	0.1668	-0.2571

Table 38 displays the item difficulty parameters of the grade 8 items administered for 2023 and 2024 administrations, along with their corresponding displacement values. As per the information presented in this table, it is evident that all items had a displacement value less than 0.5, indicating that the banked item parameters of the items were specified as fixed known values for the calibration of the grade 8 2024 form.

Table 39. Comparison of HS Assessment Operational Items Across Years

Item	2023	2024	Displacement
1	-0.0161	-0.0161	-0.2014
2	0.6807	0.6807	0.2151
3	-0.4148	-0.4148	-0.0869
4	-0.7302	-0.7302	0.2472
5	-1.6874	-1.6874	0.3544
6	**-0.9558**	**-0.3716***	**0.5742**
7	-0.5950	-0.5950	0.0065
8	0.6818	0.6818	0.0311
9	0.6802	0.6802	-0.0872
10	0.5324	0.5324	-0.1211
11	-1.1265	-1.1265	0.0320
12	-0.9239	-0.9239	-0.0382
13	-0.2455	-0.2455	-0.0346
14	-0.6466	-0.6466	-0.1110
15	-0.4624	-0.4624	0.0386
16	0.5473	0.5473	-0.0208
17	-1.2224	-1.2224	-0.0757
18	-0.9754	-0.9754	-0.0356
19	0.5277	0.5277	-0.1698
21	-0.6069	-0.6069	-0.0615
22	2.4854	2.4854	-0.1138
23	1.0494	1.0494	0.1031
24	-0.7321	-0.7321	0.1488
25	0.6824	0.6824	0.2510
26	-0.7582	-0.7582	0.2075
27	0.7524	0.7524	-0.0879
28	0.1830	0.1830	0.0308
29	2.2190	2.2190	-0.1492
30	1.5095	1.5095	-0.0938
31	0.8298	0.8298	0.0779
32	1.3336	1.3336	-0.0164
33	-0.4972	-0.4972	-0.1042
34	-1.0205	-1.0205	0.0038
35	0.7031	0.7031	0.0221
36	-0.6698	-0.6698	0.0762
37	-0.4435	-0.4435	-0.0763
38	-0.7666	-0.7666	-0.1089
39	0.9360	0.9360	-0.0235

Item	2023	2024	Displacement
20	0.2893	0.2893	0.0089
40	-0.0944	-0.0944	0.0544
41	1.0577	1.0577	0.0094

*The value displayed is the value of the item parameter estimate obtained after un-anchoring the item. The value of the displacement statistic is the value that resulted when the item was initially anchored.

As seen in Table 39, all items except for one had a displacement value less than 0.5. The item with a displacement value greater than 0.5 was unanchored during the calibration of the 2024 HS form. All other items functioned as common anchor items, with their banked item parameters specified as fixed known values.

Chapter 7. Reliability

Reliability focuses on the extent to which score differences reflect true differences in the knowledge, skills, and abilities being assessed rather than chance fluctuations. Thus, reliability measures the consistency of the scores across conditions that can be assumed to differ at random; for example, which form of the assessment the student is administered, or which raters are assigned to score constructed-response prompts. In statistical terms, the variance in the distributions of scores, essentially the differences among students, is partly due to real differences in the knowledge, skills, and abilities being assessed (true variance) and partly due to random errors in the measurement process (error variance). Reliability is an estimate of the proportion of the total observed variance that is true variance.

There are several different ways to estimate reliability. The type of raw score reliability estimate reported here is an internal-consistency measure, which is derived from analysis of the consistency of the performance of students across items within an assessment. It is used because it serves as a good estimate of alternate forms reliability, but it does not consider form-to-form variation due to lack of test form parallelism, nor is it responsive to day-to-day variation due to, for example, the student's state of health or the administration environment.

Reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain very similar scores upon repeated administrations if the students do not change in their level of the knowledge or skills measured by the assessment. Moderate to acceptable ranges of reliability tend to exceed 0.5 (Cortina, 1993; Schmitt, 1996). Estimates lower than 0.5 may indicate a lack of internal consistency.

Classically based standard error of measurement (SEM) quantifies the amount of error in the scores. SEM is the extent by which students' scores tend to differ from the scores they would receive if the test were perfectly reliable. As the SEM increases, the variability of students' observed scores is likely to increase across repeated administrations. Observed scores with large SEMs pose a challenge to the valid interpretation of a single score. Reliability and SEM estimates were calculated at the full assessment level.

7.1. Reliability and Standard Errors of Measurement

Cronbach's coefficient alpha (Cronbach, 1951) is a reliability measure for dichotomously or polytomously scored items (Brennan, 2001). The coefficient is calculated by substituting the variance of both items and total raw scores as follows:

$$\alpha_x = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_x^2}\right)$$

in which n is the number of items, σ_i^2 is the variance of scores on each item, and σ_x^2 is the variance of the total raw score. When other administration conditions are held constant, the more items the assessment includes, the greater the reliability coefficient. Conversely, when sample sizes become smaller and more homogeneous, lower reliability estimates are obtained.

The formula for the classical SEM is given as:

$$SEM = \sigma_x\sqrt{1-\alpha_x}$$

where σ_x is the standard deviation of the raw score and α_x is the estimated coefficient alpha computed above.

Student population descriptive statistics for the raw score, reliability (alpha), and SEM estimates are provided in Table 40.

Table 40. Raw Score Descriptive Statistics, Alpha, and Standard Errors of Measurement (SEM) by Grade

Grade	N	Maximum	Mean	Std. Dev.	Alpha	SEM
05	12,174	45	19.66	8.13	0.84	3.25
08	12,207	47	21.64	8.96	0.87	3.23
HS	11,583	52	23.53	10.38	0.88	3.60

Since each grade has a different blueprint (e.g., number of items) and there were variations in the percent of students assessed out of all students in the state, it is inappropriate to make inferences regarding the quality of the assessments by comparing the reliabilities across grades. As previously discussed, SEM is an estimate of the error associated with an individual's test score. It provides an estimate of the degree of measurement error associated with the scores. In Grade 5, the SEM is approximately 3.25; in Grade 8, it is about 3.23; and in high school, it is approximately 3.60. A higher SEM indicates greater measurement error. SEM remains relatively consistent across grades, suggesting similar precision in the scores obtained at each level.

7.2. Subgroup Reliability

As with the entire assessed student population, reliability and measurement error can be investigated for various subgroups of interest when the group size has a minimum of 25 students. Appendix P contains the computed reliability estimates by gender, ethnicity, multilingual learner status, Individual Education Plan (IEP) status, economically disadvantaged (SE) status, migrant status, and 504 Plan status. Because subgroup sample sizes vary considerably, results should be interpreted with caution.

7.3. Inter-rater Consistency

The level of inter-rater consistency in scoring CR items is reported in Appendix H. Inter-rater consistency is a measure of the level of consistency of the raters. Chapter 4 describes the processes SME used to monitor the quality of the hand-scored item prompt responses. Prompt level inter-rater reliability results are provided in Appendix H.

7.4. Accuracy and Consistency

Classification accuracy is defined as the extent to which the actual classifications of test takers (on the basis of their observed test scores) agree with those that would be made on the basis of their true scores if their true scores could somehow be known. The term consistency refers to the agreement between classifications based on two nonoverlapping, equally difficult forms of the test (i.e., parallel forms) (Livingston & Lewis, 1995).

We used Livingston and Lewis's (1995) approach, which is intended to handle situations where items are not equally weighted and/or some or all the items are polytomously scored. This method is formulated as

$$\tilde{n} = \frac{(\mu_x - X_{min})(X_{max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)},$$

where X_{min} is the lowest score for X, X_{max} is the highest score, μ_x is the mean, σ_x^2 is the variance, and r is the reliability. This method models the distribution of the true scores and of scores on a parallel form by using a four-parameter beta distribution.

As seen in the above formula, classification accuracy and consistency indices rely on the interaction between several different factors related to test design and standard-setting decisions. These factors include the number of cut scores, test reliability, measurement accuracy at the cut score(s), distance between adjacent cut scores, location(s) of the cut score(s) on the ability scale, and percentage of students around a cut score(s) (Ercikan & Julian, 2002; Lee, Hanson, & Brennan, 2002). Because these statistics are influenced by the interplay between a variety of factors, only a very limited number of studies to date have investigated the ideal or expected levels of classification consistency and accuracy needed for educational assessments.

Classification accuracy indices quantify the percentage of students who are accurately placed below and/or above a given cut score. For example, a classification accuracy index of 0.886 for the cut score demarcating achievement levels 1 and 2 means that were students to be classified twice, once according to their observed score and once according to their true score, 88% of those students would be classified in the same category both times.

Similarly, classification consistency indices give the percentage of students classified consistently below and/or above a given cut score. For example, a classification consistency index of 0.840 for the cut score demarcating achievement levels 1 and 2 means that if there were two parallel forms to be administered to students, 84% of those students would be consistently classified for both forms.

In this technical report, we summarized the range of classification accuracy and consistency of tests across grades. As shown in Table 41 classification accuracy ranged from 0.881 to 0.969; from 0.900 to 0.967; and from 0.917 to 0.970 for grade 5, grade 8, and high school, respectively. As seen in Table 42 classification consistency ranged from 0.834 to 0.957; from 0.860 to 0.955; and from 0.883 to 0.958 for grade 5, grade 8, and high school, respectively. Table 43 shows the values of CSEM for scaled scores at the cut scores. As demonstrated, even though the values of the CSEM are low in general, in grade 8 they are higher than the CSEM values of the other two grades.

Table 41. Classification Accuracy Indices at Achievement Level Cut

Grade	Achievement Level 1/2	Achievement Level 2/3	Achievement Level 3/4
05	0.881	0.912	0.969
08	0.900	0.906	0.967
HS	0.917	0.924	0.970

Table 42. Classification Consistency Indices at Achievement Level Cut

Grade	Achievement Level 1/2	Achievement Level 2/3	Achievement Level 3/4
05	0.834	0.877	0.957
08	0.860	0.869	0.955
HS	0.883	0.894	0.958

Table 43. Conditional Standard Errors of Measurement at Achievement Level Cut

Grade	Achievement Level 1/2	Achievement Level 2/3	Achievement Level 3/4
05	2	2	2
08	4	4	5
HS	2	2	2

Chapter 8. Score Reporting

Prior to spring 2023, score reporting was handled by other vendors contracted by Maine DOE. New Meridian Corporation took over all score reporting beginning with the spring 2023 administration. New Meridian Corporation now provides reports directly to SAUs (districts) and schools. This streamlined model enhances the efficiency of report delivery, ensuring that the reports reach their intended recipients in a more expeditious manner.

In 2023, Maine DOE, New Meridian, and Maine's Science TAC collaborated to design the new science assessment reports. A primary focus throughout this process was to provide actionable and valuable information to students, schools, and SAUs.

Once the report designs were finalized and received approval, they underwent development and creation before being integrated into a production-level reporting portal accessible to the educational community. These reports were generated at three distinct levels: individual student (ISR), school, and SAU. They were delivered through a dedicated reporting portal, which allowed District Assessment Coordinators (DACs) and School Assessment Coordinators (SACs) to access, download, and distribute the reports to their respective schools and students.

The subsequent sections will provide in-depth details about each report, while samples of these reports can be found in Appendix T for reference.

8.1. Business Requirements

To maintain the precision and reliability of reported results for the Maine Science Assessment, a comprehensive document outlining the processing and reporting business requirements was prepared ahead of the reporting cycle. These requirements serve as the foundation for the analysis of assessment data and the generation of results. Moreover, they provide critical guidance to data analysts when it comes to identifying students who should be excluded from summary computations at the school, SAU, and state levels.

8.2. Static Reports

The following deliverables were produced for the Maine Science Assessment:

- Individual Student Report–PDF file
- School Summary Report–PDF file
- SAU Summary Report–PDF file

- Student Score Data File (Roster report–CSV file)
    - School Student Score Data File (Roster Report–CSV File)
    - SAU Student Score Data File (Roster Report–CSV File)

All reports were made available for the SAUs and schools on the reporting platform. Each of these reporting deliverables are described in the following sections.

8.3. Individual Student Report (ISR)

The individual student report, prepared for each student, is a concise, single page double-sided color report. This report includes pertinent information, such as the scaled score, achievement level, and reporting category results for each assessed science area. It also presents a comparison of student performance by scaled score at the school, SAU, and state levels (For an illustrative example, please refer to Appendix T).

Each student should receive one report encompassing all their science assessment information. Moreover, guidance was provided to SAUs and schools on how to both interpret the reports and download the necessary files from the reporting portal.

The first page of the report provides the following information:

- Description of information presented in the report
- Description of the Maine Science Assessment
- "Questions for Your Student" related to application of science knowledge and understanding
- "Questions for the Teacher" related to assessment literacy

The second page of the report provides the following information:

- Overall student science performance, including
    - A graph showing how the student's scaled score relates to the state's achievement level
    - A bar graph score comparison of student scaled score to the school, SAU, and state averages
- An explanation of the four achievement levels for Maine
- A pill graph for each of the three science subscores relevant to the student's grade level, highlighting the specific achievement level attained by the student for each subscore within those areas:
    - Grade 5 – Structure and Properties of Matter; Matter and Energy in Organisms and Ecosystems; and Earth's Systems and Space Systems: Stars and the Solar System
    - Grade 8 – Physical Science; Life Science; and Earth and Space Science
    - High School – Physical Science; Life Science; and Earth and Space Science
- The science topic bundles for a grade level are listed out below each pill graph

8.4. School Summary Report

The school summary report includes score comparisons between the school, SAU, and state scaled score averages for each grade level the school has assessed. It contains a table showing the school aggregate data and additional table(s) for each grade level assessed by the school, indicating each of the following pieces of information:

- Total number of students assessed
- Overall average scaled score (data only present for grade level specific table[s])

- Overall average achievement level (data only present for grade level specific table[s])
- Percent borderline students – the percentage of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement
- Achievement level data – displays the n-count and percentage for each of the four achievement levels.
- Raw score averages for each of the three subscores (data only present for grade level specific table[s])

A pie graph visually depicts the score comparisons from the table.

8.5. SAU Summary Report

The SAU summary report includes score comparisons between the SAU and state scaled score averages for each grade level the assessed by the SAU. It contains a table showing the SAU aggregate data and additional table(s) for each grade level assessed by the SAU, along with each school within that grade level for the SAU indicating each of the following pieces of information:

- Total number of students assessed
- Overall average scaled score (data only present for grade level specific table[s])
- Overall average achievement level (data only present for grade level specific table[s])
- Percent borderline students – the percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement
- Achievement level data – displays the n-count and percentage for each of the four achievement levels
- Raw score averages for each of the three subscores (data only present for grade level specific table[s])

A pie graph visually depicts the score comparisons from the table.

8.6. Student Score Data File (Roster report–CSV file)

This extract comprises both the SAU Student Score Data File extract and the School Student Data File extract.

The Roster report comprises individual student scores and other scoring relevant information for a single administration. Users download the extract based on their organization (school, SAU, or statewide, subject to their access permissions) to obtain various demographic and score-related criteria.

8.7. Quality Control of Score Reporting

New Meridian conducts an annual score reporting quality-control process to verify the accuracy of all score reports (the Individual Student Report–ISR, the Student Roster Report, the School Summary Report, the School Roster Report, the SAU Summary Report, and the SAU Roster Report).

New Meridian's Integrated Quality (IQ) team collaborated closely with the Psychometrics team to ensure the capture and accurate delivery of high-quality data. Data used to populate the reports was computed and independently replicated by Psychometrics prior to hand-off to the IQ team. Utilizing a range of software tools, the IQ team conducted comprehensive quality control checks to ensure fidelity and accuracy as data transitions through various stages. These checks are instrumental in ensuring the precision and reliability of the data during subsequent analyses, computation, and formatting into tables and columns for delivery to Kansas University for report generation and posting to the reporting portal.

Additionally, quality control for report appearance included an assessment of overall structure and layout adherence to approved report mockups; the consistency of formatting properties like fonts and colors; the correctness of visual elements like logos, graphs, and diagrams; verification of static verbiage such as headers, footers, and body paragraphs (as found within the ISR); and the assessment of page breaks or other grouping functionality.

Quality assurance processes for reporting are customized for each report type, recognizing that the content and presentation are unique to each report.

Collaboration between IQ and Psychometrics facilitates a prompt and accurate response to any identified data anomalies. Test cases are systematically linked to a tracking system, ensuring that each required action was meticulously outlined and documented. After report generation, the IQ team executed test cases to validate student and summary printed reports as well as CSV roster reports, to ensure their alignment with specifications and design layouts. Once all test cases were successfully completed, the IQ team provided notification to the program team for final approval.

Chapter 9. Validity

The *Standards for Educational and Psychological Testing*, issued jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014), states that:

> …validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations (p. 11).

Therefore, the purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for specific uses. Test validation is not a quantifiable property but an ongoing process that begins at initial conceptualization and continues throughout the life cycle of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence of lack of validity), including design, content specifications, item development, and psychometric characteristics.

Validity was examined by looking at evidence based on assessment content, evidence based on internal structure, and evidence based on external variables. Test items were matched with blueprints and NGSS standards to ensure content validity. For construct validity, item-total correlations, subscore correlations, and factor structure were examined. In addition, relationships with external variables such as student questionnaire data were used. These data also support content and construct validity.

9.1. Evidence Based on Assessment Content

Evidence based on content of achievement tests is supported by the degree of correspondence between test items and content standards. The Maine Science Assessments adhere to the principles of evidence-centered design, in which the standards to be measured (the MLRs) are identified, and the performance a student needs to achieve to meet those standards is delineated in the performance expectations. As noted in Chapter 2, all assessment items have been thoroughly reviewed with the New Meridian Science Exchange Framework. Assessment items are further reviewed for adherence to universal design principles, which maximize the participation of the widest possible range of students prior to the items being selected for administration.

The form planners shown in Appendix B represent how the test forms matched the blueprints in terms of the reporting categories of science discipline and science and engineering practices. The number of items for each form is given for various dimensions of the NGSS. The science educator cadre performed reviews on the items field tested in 2024, and the average percentages across all field-test items for which the standard provided matched the educator cadre standard were as follows:

- Performance Expectation (PE): 64%
- Science and Engineering Practices (SEP): 59%

Note: the operational items were reviewed in prior years. See the 2022 and 2023 Technical Reports.

9.2. Evidence Based on Internal Structure

Analyses of the internal structure of an assessment typically involve studies of the relationships among items and/or test components (i.e., subclaims) in the interest of establishing the degree to which the items or components appear to reflect the construct on which a test score interpretation is based (AERA, APA, & NCME, 2014, p. 16). The term *construct* is used here to refer to the characteristics that a test is intended to measure; in the case of the operational tests, the characteristics of interest are the knowledge and skills defined by the test blueprint.

Evidence based on internal structure is discussed in detail in Chapter 5 and Chapter 6. Technical characteristics of the internal structure of the assessments are presented in terms of classical test theory statistics (item difficulty, item-total correlation), differential item functioning (DIF) analyses, item response theory (IRT) parameters and procedures, and dimensionality assumption analyses. In general, item difficulty and discrimination indices were in acceptable and expected ranges given the circumstances for this administration. Positive discrimination indices for the final operationally scored items indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall. Also, multidimensionality results, which are evidence for construct validity, proved that tests are unidimensional.

The reliability analyses presented in Chapter 7 provide information about the internal consistency of the assessments. Internal consistency is typically measured via correlations among the items on an assessment and provides an indication of how much the items measure the same general construct. Except for subgroups of the ML students, all reliabilities were greater than 0.80. As indicated in Chapter 7, the reason of low reliability for those subgroups is low sample size.

Consequences of testing refers to intended and unintended consequences associated with test result interpretation. Evidence for the consequences of testing is addressed with item analyses information in Chapter 5 and in the scaling information in Chapter 6. However, all chapters speak to the efforts undertaken to provide accurate and clear information regarding test scores. Evidence of the consequences of testing will also accrue with the continued implementation of the MLRs and the continued administration of the Maine Science Assessment.

Correlation between Subscores

Correlations between subscores can provide valuable validity evidence, especially in the context of construct validity, specifically convergent validity. When subscores measuring related constructs or aspects of the same trait are positively correlated, it suggests that the test is measuring the intended construct effectively. High correlations between subscores that are theoretically expected to be related indicate that the test is consistent with the underlying construct. This provides evidence of convergent validity.

These correlation coefficients help understand the degree to which these subscores are related to each other. A correlation coefficient of 1 indicates a perfect positive linear relationship, while a coefficient of -1 would indicate a perfect negative linear relationship. A coefficient of 0 suggests no linear relationship.

Table 44 shows the correlations among scaled scores of the three subscores for Grade 5. Subscore intercorrelations range from 0.668 and 0.709, indicating moderate to strong relationships between the subscores.

Table 44. Grade 5 Correlations Among Scaled Scores of Subscores

	Subscore 1	Subscore 2	Subscore 3
N	12,396	12,396	12,396
Subscore 1	1		
Subscore 2	0.668	1	
Subscore 3	0.709	0.668	1

Table 45 shows the correlations among scaled scores of the three subscores for Grade 8. Subscore intercorrelations range from 0.720 to 0.783, indicating moderate to strong relationships between the subscores.

Table 45. Grade 8 Correlations Among Scaled Scores of Subscores

	Subscore 1	Subscore 2	Subscore 3
N	12,754	12,754	12,754
Subscore 1	1		
Subscore 2	0.720	1	
Subscore 3	0.783	0.728	1

Table 46 shows the correlations among scaled scores of the three subscores for high school. Subscore intercorrelations range from 0.826 to 0.870, indicating strong relationships between the subscores.

Table 46. High School Correlations Among Scaled Scores of Subscores

	Subscore 1	Subscore 2	Subscore 3
N	13,740	13,740	13,740
Subscore 1	1		
Subscore 2	0.835	1	
Subscore 3	0.826	0.870	1

9.3. Evidence based on External Variables

External validity for the Maine Science Assessment may be described by the relationship of the test scores with situational variables such as student self-image, curriculum, and various instructional patterns. The situational variables were all based on student questionnaire data collected during Session 4 of the administration. The questions varied slightly by grade. The pertinent questions are presented in Appendix Q. Student performance on the assessment was broken out by performance quartiles for these analyses.

The relationship between students' performances with other variables known to be related (Crocker & Algina, 1986) is another way to examine validity. For example, it is known that there is a relationship between intelligence and school success or job performance. Although these are not two equivalent constructs, intelligence is a predictor of these variables, therefore it is interpreted as an indicator of validity. As seen in self-image results, the number of students who describe themselves as poor in science in all grades is higher in the first quartile (Q1) and lower in the fourth quartile (Q4), which are the categories with lower performing and higher performing students, respectively. This finding can be interpreted as evidence for construct validity since it is projecting students' thinking.

In contrast to the self-image results, students in all performance quartiles across grades agree that the questions on the MEA test reflect what they have learned in school about science. This finding can be interpreted as evidence for content and construct validity.

Self-Image

All students were asked how they would rate themselves as a student in science. Figures 2–4 provide the results by performance quartile for each grade separately.

Question: Which of the following best describes how you rate yourself as a student in science?

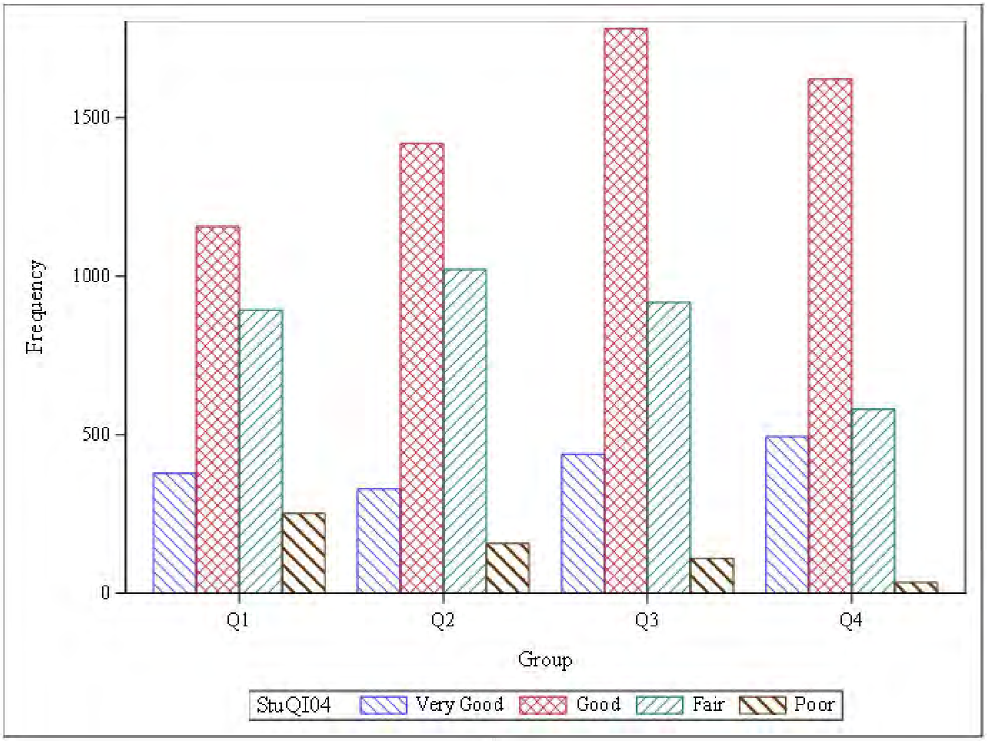


Figure 2. Grade 5 Self-Image

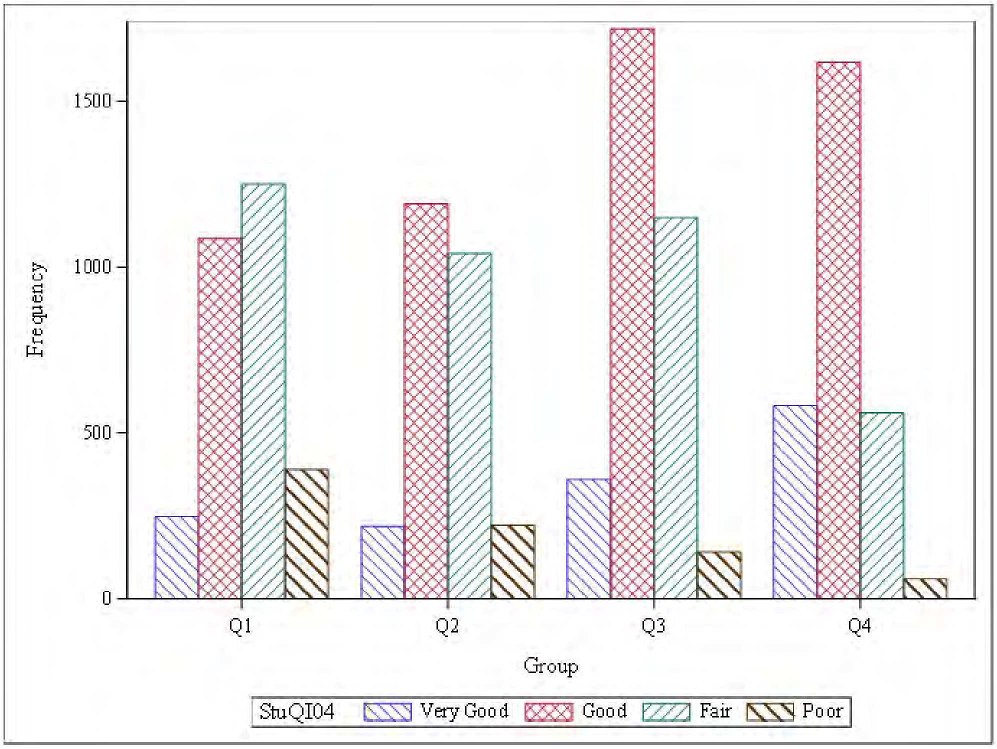


Figure 3. Grade 8 Self-Image

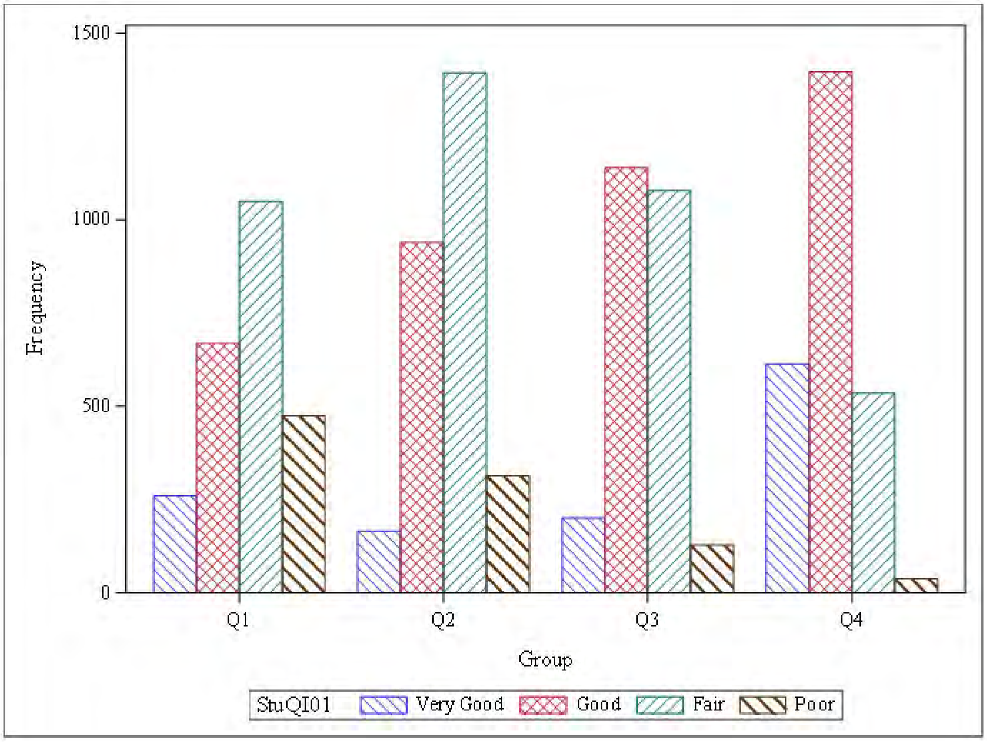


Figure 4. High School Self-Image

Match with Curriculum

All students were asked how well the content of the assessment matched what they learned in the classroom. Figures 5–7 provide the results by performance quartile for each grade separately.

Question: How well do the questions that you have just been given on this MEA test match what you have learned in school about science?

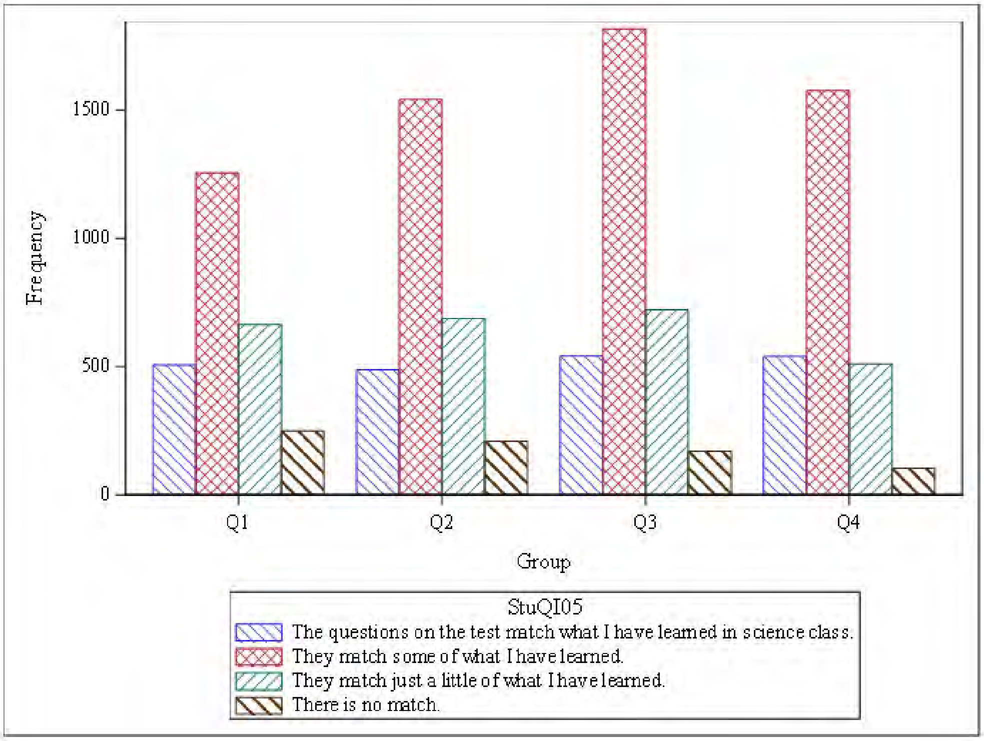


Figure 5. Grade 5 Assessment Match with Curriculum

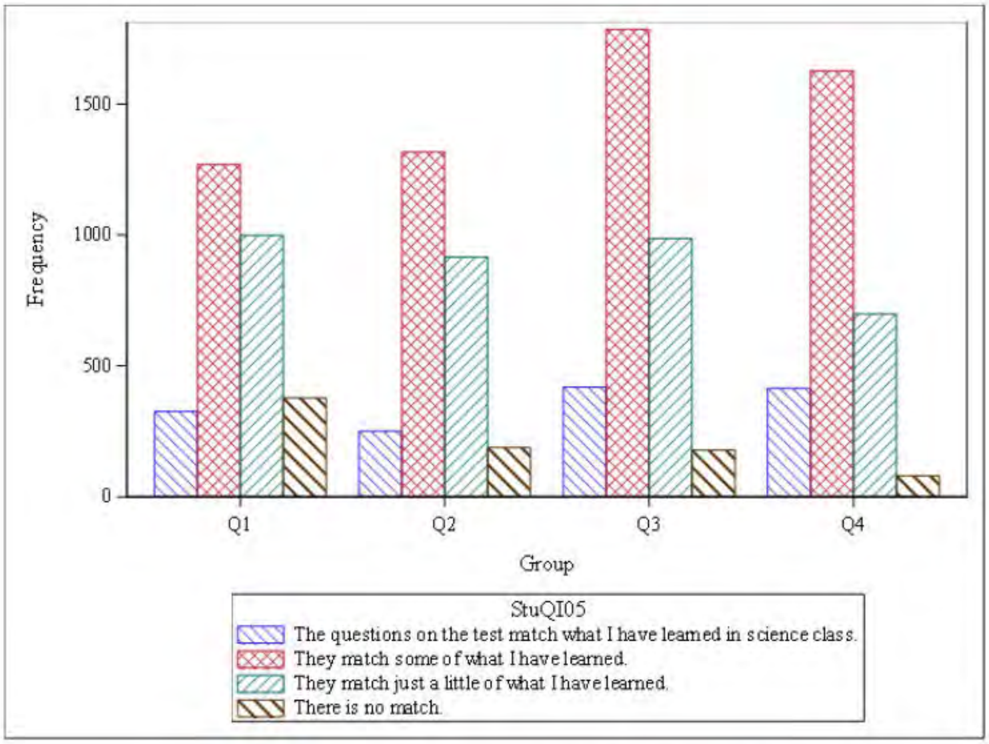


Figure 6. Grade 8 Assessment Match with Curriculum

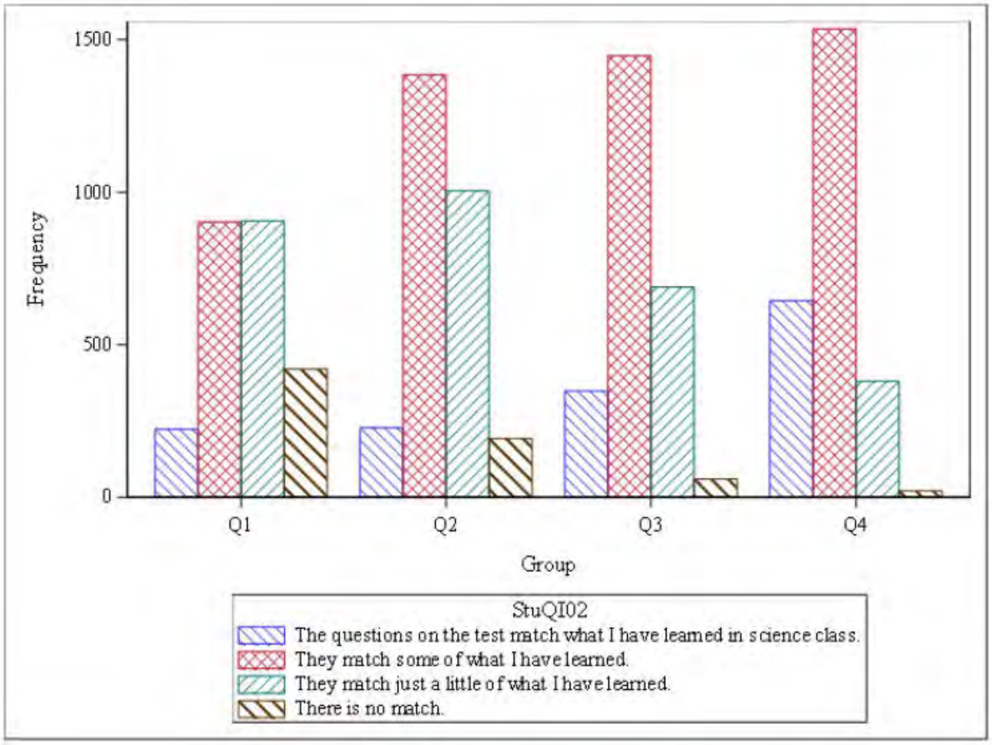


Figure 7. High School Assessment Match with Curriculum

Chapter 10. Student Performance

10.1. Raw Scores

Table 47 provides descriptive statistics for raw scores by grade level. Across 12,174 students in grade 5, the average raw score was 19.66 points, with a standard deviation of 8.13 points, indicating some variability in scores among students. The raw score distribution shows some spread, though is generally normally distributed. Raw score distribution plots for all grades are presented in Appendix O.

Across 12,207 students in grade 8, the average raw score was 21.64 points, with a standard deviation of 8.96 points, indicating some variation in scores among the students. The raw score distribution shows some spread, and a slight skew towards lower total scores relative to a normal distribution.

Among 11,583 students in high school, the average raw score was 23.53 points, with a standard deviation of 10.38 points, indicating a slightly larger amount of variation in scores among the students relative to grade 5 or grade 8. The raw score distribution shows some spread, and a slight skew towards lower total scores relative to a normal distribution.

Table 47. Raw Score Descriptive Statistics

Grade	N	Minimum	Maximum	Mean	Std. Dev
05	12,174	0	45	19.66	8.13
08	12,207	0	47	21.64	8.96
HS	11,583	1	52	23.53	10.38

10.2. Scaled Scores

Table 48 provides descriptive statistics for scaled scores by grade level. Across 12,174 students in grade 5, the average scaled score was 34.25, with a standard deviation of 6.94, indicating some variability in scores among students. The scaled score distribution is normally distributed and shows very little spread. Scaled score distribution plots for all grades are presented in Appendix O.

Across 12,207 students in grade 8, the average scaled score was 37.62, with a standard deviation of 13.69. The larger standard deviation indicates more variability in scores among students. This is reflected in a larger spread of scores in the distribution plots, with a slight skew towards lower scaled scores.

Across 11,583 students in high school, the average scaled score was 36.46, with a standard deviation of 8.24, indicating some variability in scores among students. However, scores were generally normally distributed.

Table 48. Scaled Score Descriptive Statistics

Grade	N	Minimum	Maximum	Mean	Std. Dev
05	12,174	6	80	34.25	6.94
08	12,207	3	89	37.62	13.69
HS	11,583	5	68	36.46	8.24

10.3. Achievement level

Table 49 provides the percent of students in each achievement level by grade. Between 41–48 percent of students were Well Below State Expectations. The Below and At State Expectations levels showed the greatest variability, with 30 percent of students in grade 5 Below State Expectations compared with 15 and 21 percent in grade 8 and high school, respectively. Conversely, only 18 percent of students were categorized as At State Expectations in grade 5, compared to 37 and 30 percent in grade 8 and high school.

Table 49. Percent Achievement Level

Grade	Well Below	Below	At	Above
05	48%	30%	18%	4%
08	41%	15%	37%	6%
HS	43%	21%	30%	6%

Chapter 11. References

Achieve, Inc. (2019). A framework to evaluate cognitive complexity in science assessments. Retrieved from https://www.nextgenscience.org/sites/default/files/Science%20Cognitive%20Complexity%20Framework_Final_093019.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38(*4), 295–317.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(*3), 297–334.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Ercikan, K, & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269–294.

Golia, S. (2012). Differential item functioning classification for polytomously scored items. *Electronic Journal of Applied Statistical Analysis, 5*(3), 367–373.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Henrysson, S. (1963). Correction of Item-Total Correlations in item analysis. *Psychometrika, 28*(2), 211–218.

Holland, P. W. & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* (3rd ed.). New York, NY: Springer-Verlag.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412–432.

Lemke, E., & Wiersma, W. (1976). *Principles of psychological measurement*. Chicago, Ill: McNally.

Linacre, J. M. (2002). What Do Infit and Outfit, Mean-Square and Standardized Mean?" *Rasch Measurement Transactions 16*, 878. https://www.rasch.org/rmt/rmt162f.htm.

Linacre, J.M. (2006). WINSTEPS: Facets Rasch Measurement Computer Program. Chicago: Winsteps. Com.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690–700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*(2), 149–174.

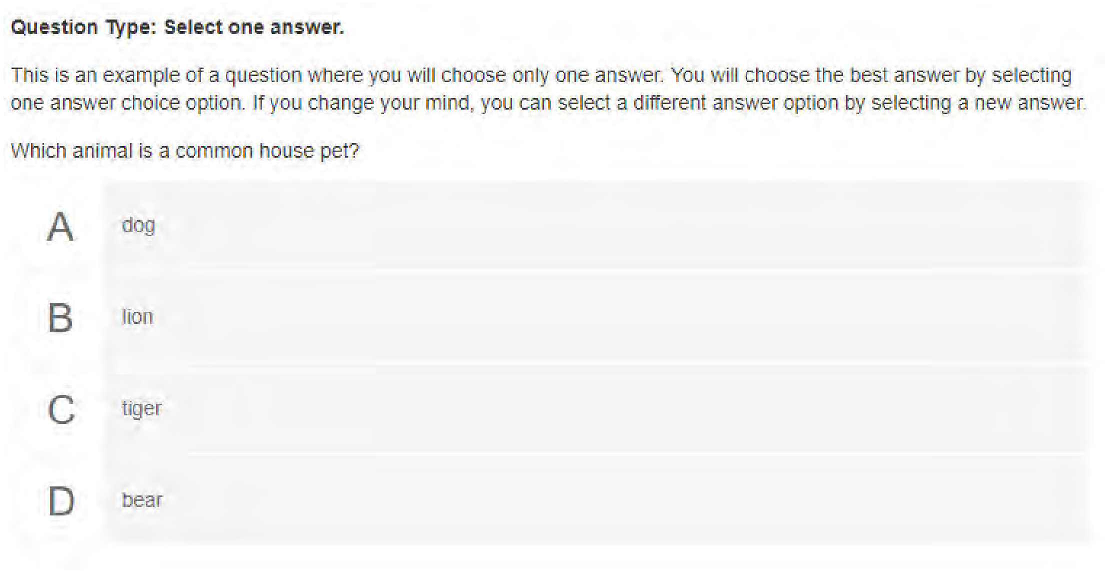
National Research Council, 2014. *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18409.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.

Student, S. & Gong, B. (2020). *Recommendations to support the validity of claims in NGSS assessment*. Center from Assessment. https://www.nciea.org/blog/recommendations-to-support-the-validity-of-claims-in-ngss-assessment/.

Appendix A. Item Response Type Examples

Multiple-Choice



Question Type: Select one answer.

This is an example of a question where you will choose only one answer. You will choose the best answer by selecting one answer choice option. If you change your mind, you can select a different answer option by selecting a new answer.

Which animal is a common house pet?

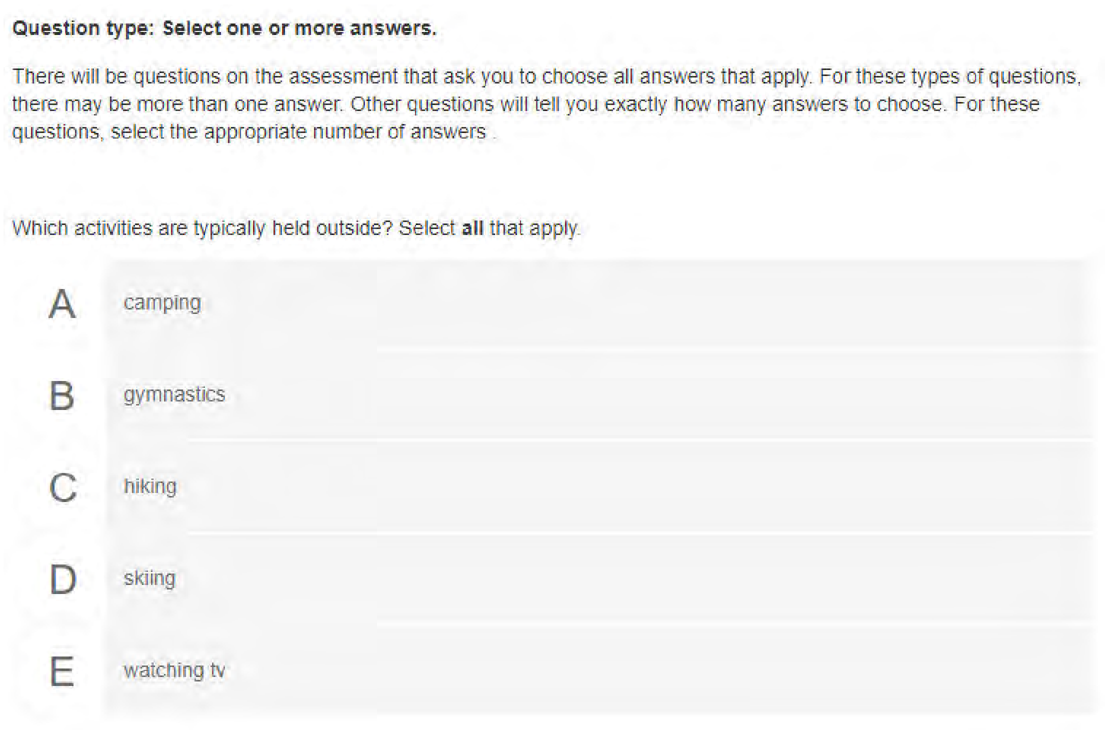
A dog

B lion

C tiger

D bear

Multiple-Select



Question type: Select one or more answers.

There will be questions on the assessment that ask you to choose all answers that apply. For these types of questions, there may be more than one answer. Other questions will tell you exactly how many answers to choose. For these questions, select the appropriate number of answers .

Which activities are typically held outside? Select **all** that apply.

A camping

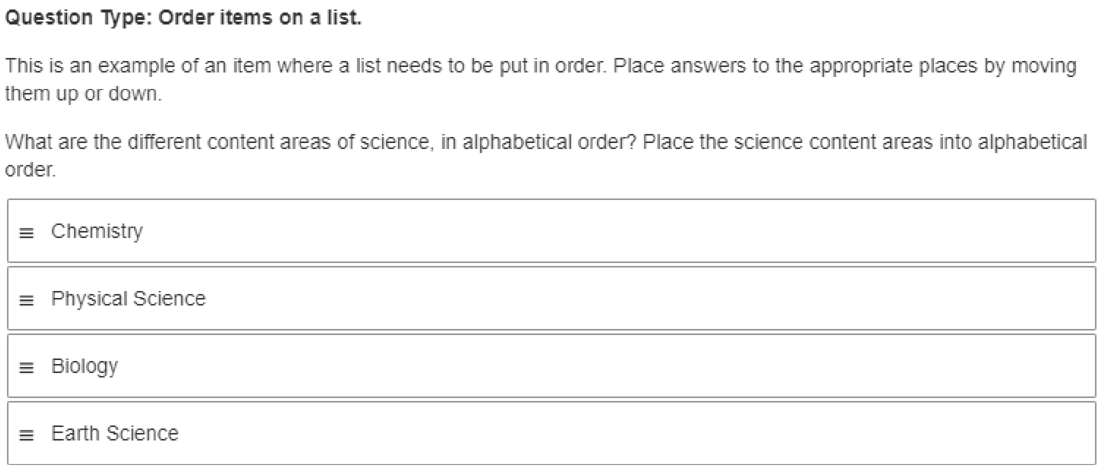
B gymnastics

C hiking

D skiing

E watching tv

Order



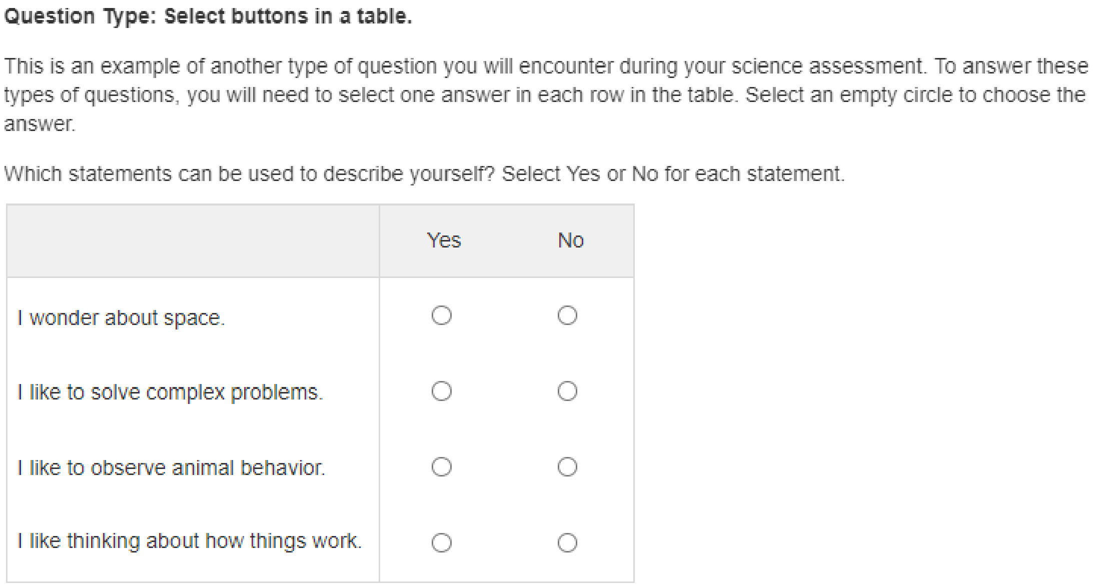
Question Type: Order items on a list.

This is an example of an item where a list needs to be put in order. Place answers to the appropriate places by moving them up or down.

What are the different content areas of science, in alphabetical order? Place the science content areas into alphabetical order.

- ≡ Chemistry
- ≡ Physical Science
- ≡ Biology
- ≡ Earth Science

Matrix



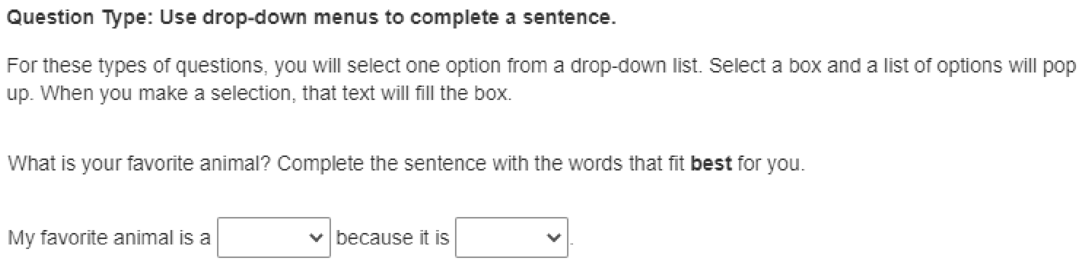
Question Type: Select buttons in a table.

This is an example of another type of question you will encounter during your science assessment. To answer these types of questions, you will need to select one answer in each row in the table. Select an empty circle to choose the answer.

Which statements can be used to describe yourself? Select Yes or No for each statement.

	Yes	No
I wonder about space.	○	○
I like to solve complex problems.	○	○
I like to observe animal behavior.	○	○
I like thinking about how things work.	○	○

Drop-Down Menu



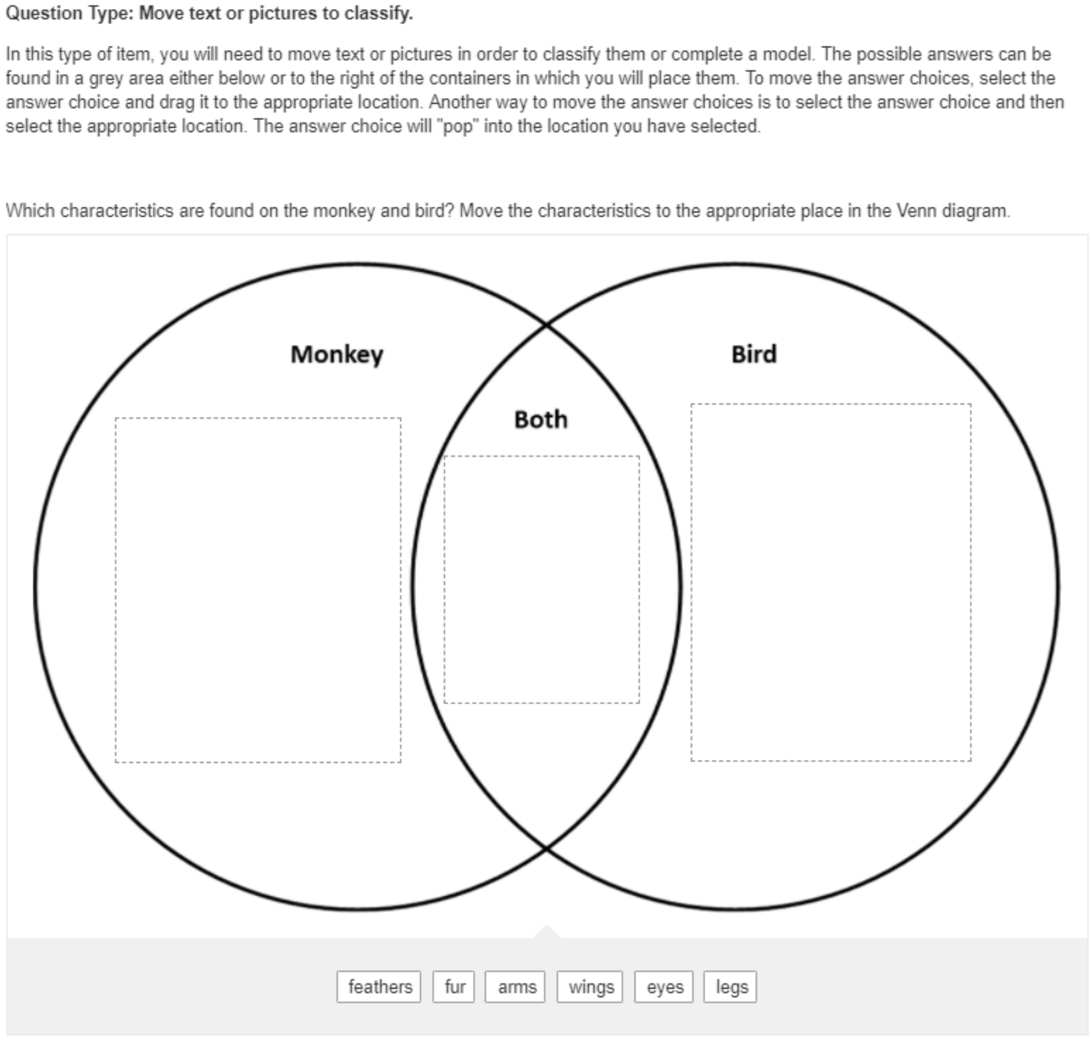
Question Type: Use drop-down menus to complete a sentence.

For these types of questions, you will select one option from a drop-down list. Select a box and a list of options will pop up. When you make a selection, that text will fill the box.

What is your favorite animal? Complete the sentence with the words that fit **best** for you.

My favorite animal is a ▾ because it is ▾.

Gap Match

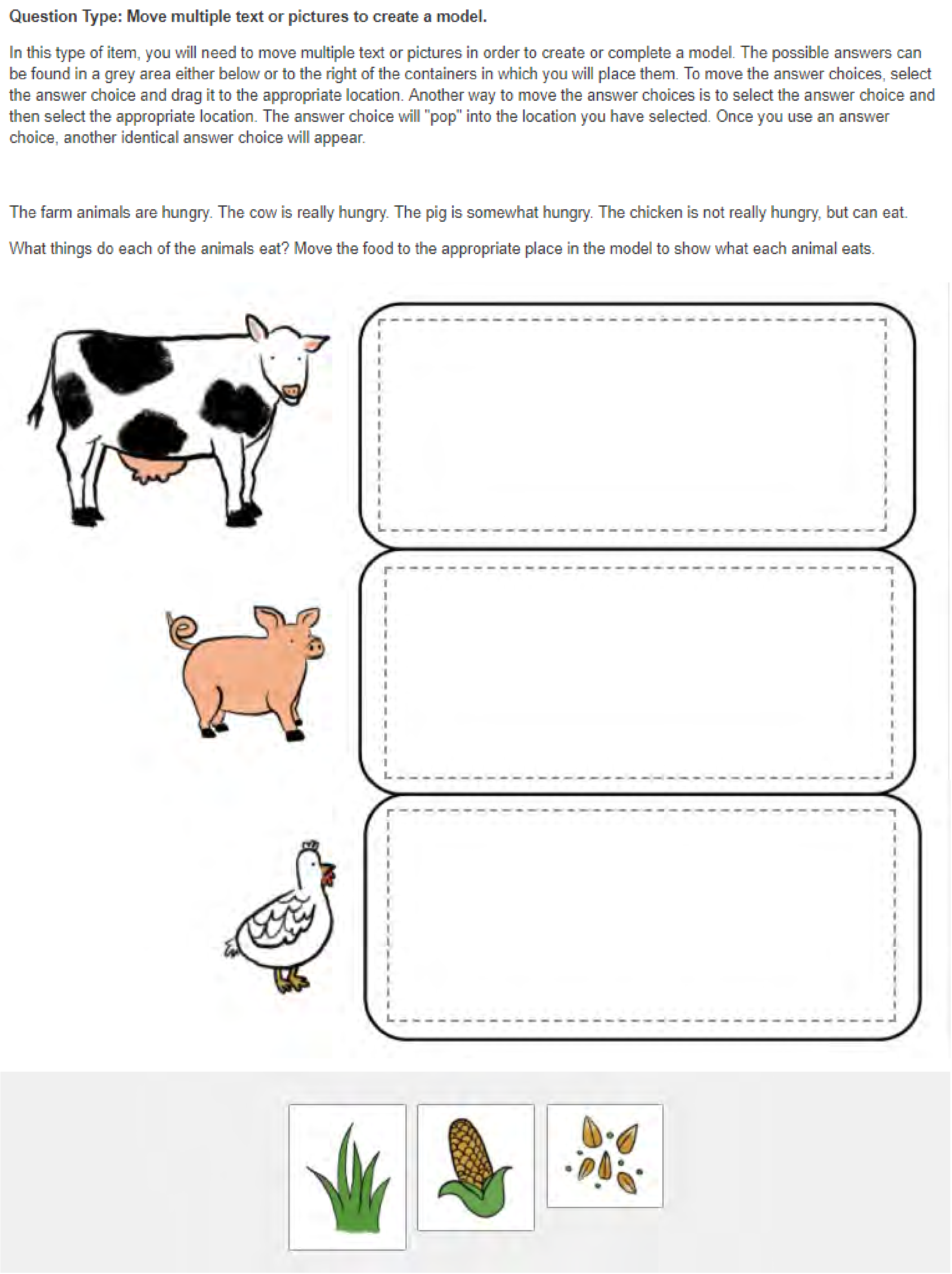


Question Type: Move text or pictures to classify.

In this type of item, you will need to move text or pictures in order to classify them or complete a model. The possible answers can be found in a grey area either below or to the right of the containers in which you will place them. To move the answer choices, select the answer choice and drag it to the appropriate location. Another way to move the answer choices is to select the answer choice and then select the appropriate location. The answer choice will "pop" into the location you have selected.

Which characteristics are found on the monkey and bird? Move the characteristics to the appropriate place in the Venn diagram.

Graphic Gap Match: Multiple Answer Choices in One Container



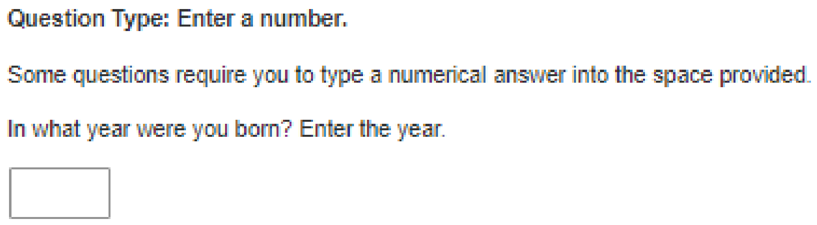
Question Type: Move multiple text or pictures to create a model.

In this type of item, you will need to move multiple text or pictures in order to create or complete a model. The possible answers can be found in a grey area either below or to the right of the containers in which you will place them. To move the answer choices, select the answer choice and drag it to the appropriate location. Another way to move the answer choices is to select the answer choice and then select the appropriate location. The answer choice will "pop" into the location you have selected. Once you use an answer choice, another identical answer choice will appear.

The farm animals are hungry. The cow is really hungry. The pig is somewhat hungry. The chicken is not really hungry, but can eat.

What things do each of the animals eat? Move the food to the appropriate place in the model to show what each animal eats.

Short Text Entry

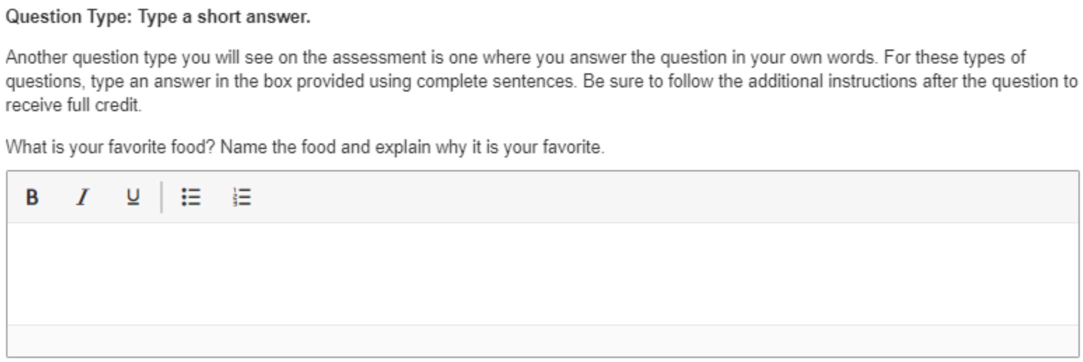


Question Type: Enter a number.

Some questions require you to type a numerical answer into the space provided.

In what year were you born? Enter the year.

Constructed-Response



Question Type: Type a short answer.

Another question type you will see on the assessment is one where you answer the question in your own words. For these types of questions, type an answer in the box provided using complete sentences. Be sure to follow the additional instructions after the question to receive full credit.

What is your favorite food? Name the food and explain why it is your favorite.

Appendix B. Spring 2024 Form Planner

(Document begins on next page.)

Maine Science Form Planner

Spring 2024 Online Forms

Grades 5, 8, and High School

Contents

New Meridian

Grade 5

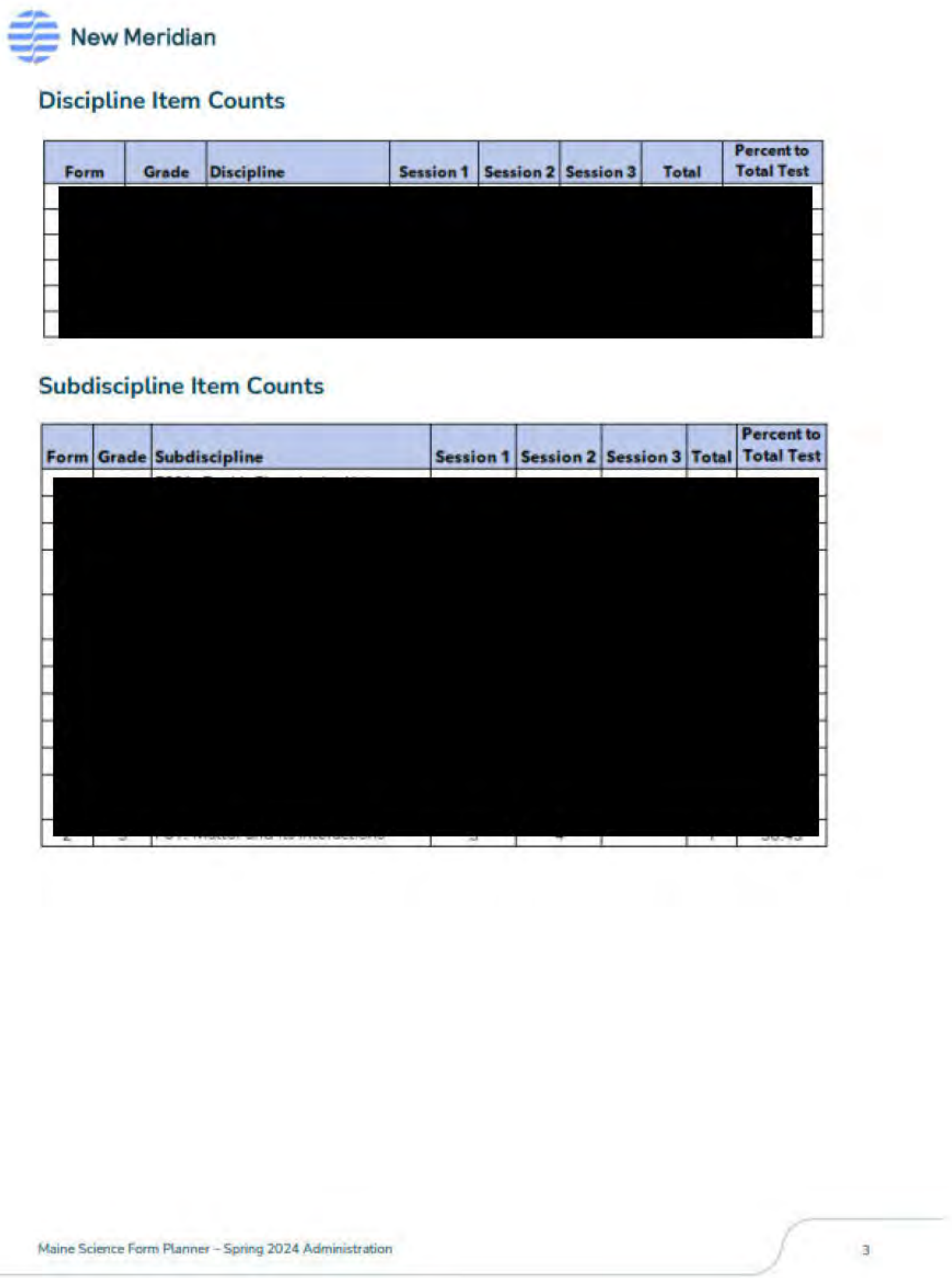
Blueprint Item Counts

Form	Grade	Group	Type	Counts	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

New Meridian

Blueprint Point Totals

Form	Grade	Group	Type	Counts	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



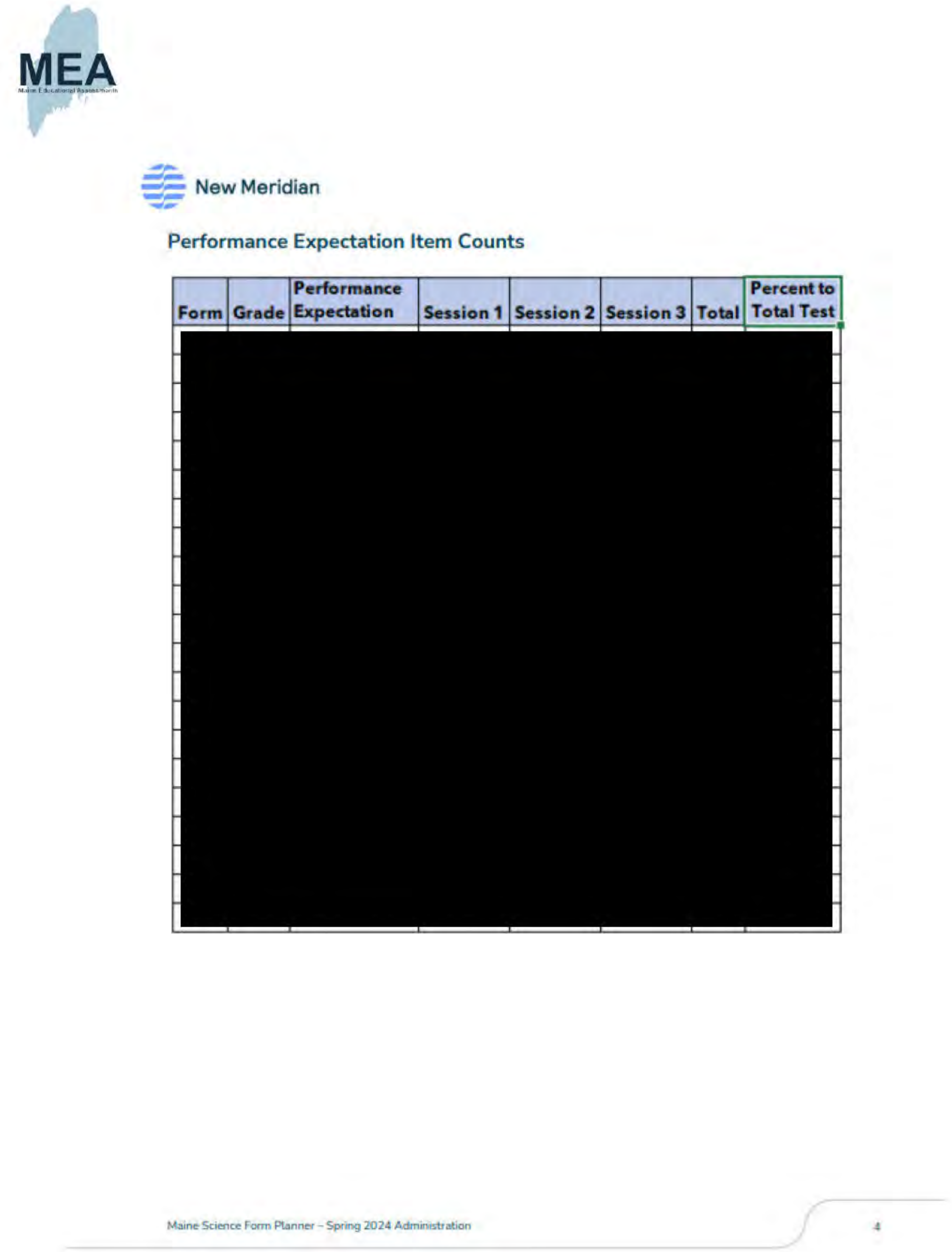
New Meridian

Discipline Item Counts

Form	Grade	Discipline	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Subdiscipline Item Counts

Form	Grade	Subdiscipline	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



New Meridian

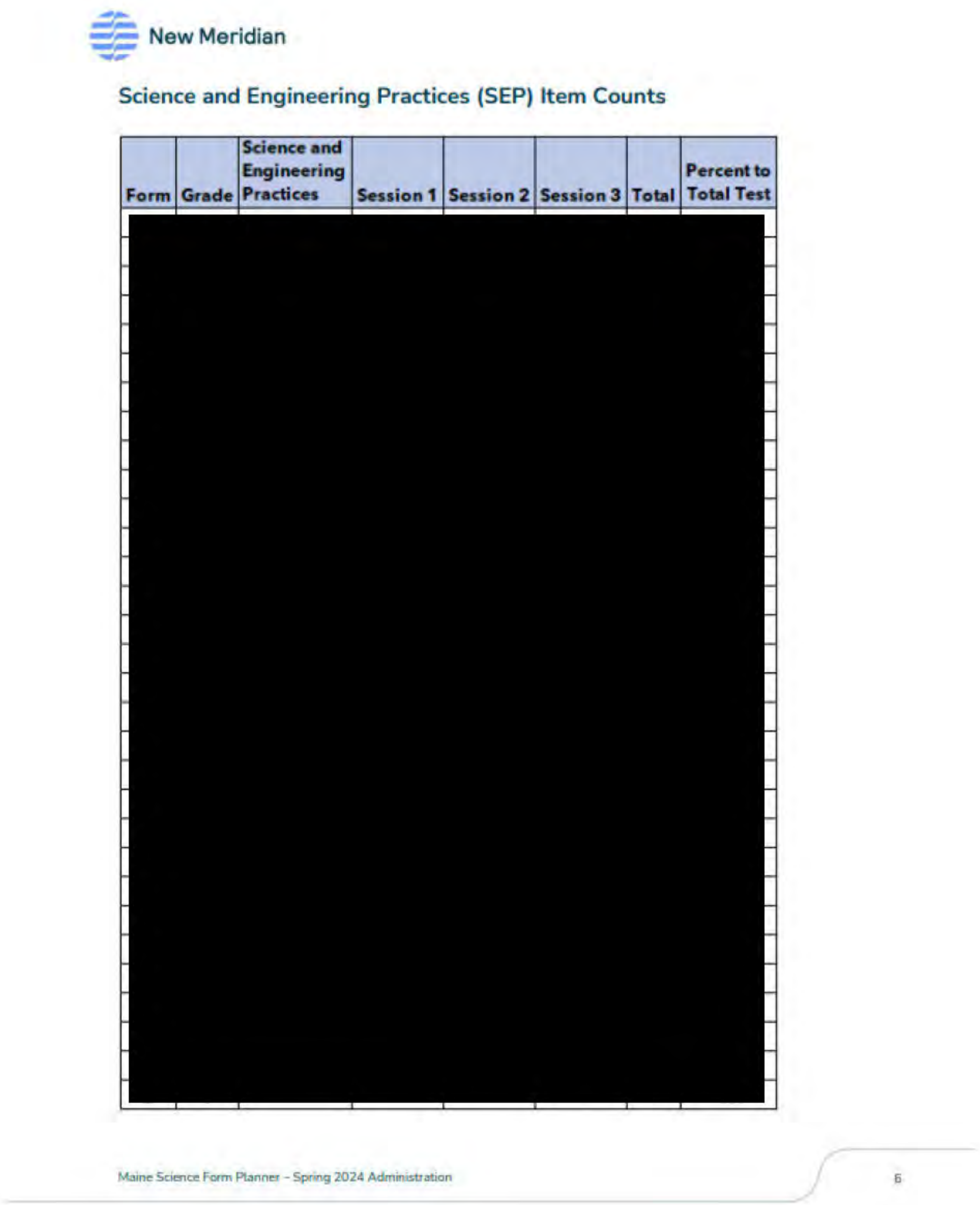
Performance Expectation Item Counts

Form	Grade	Performance Expectation	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

New Meridian

Disciplinary Core Ideas (DCI) Item Counts

Form	Grade	Disciplinary Core Ideas	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



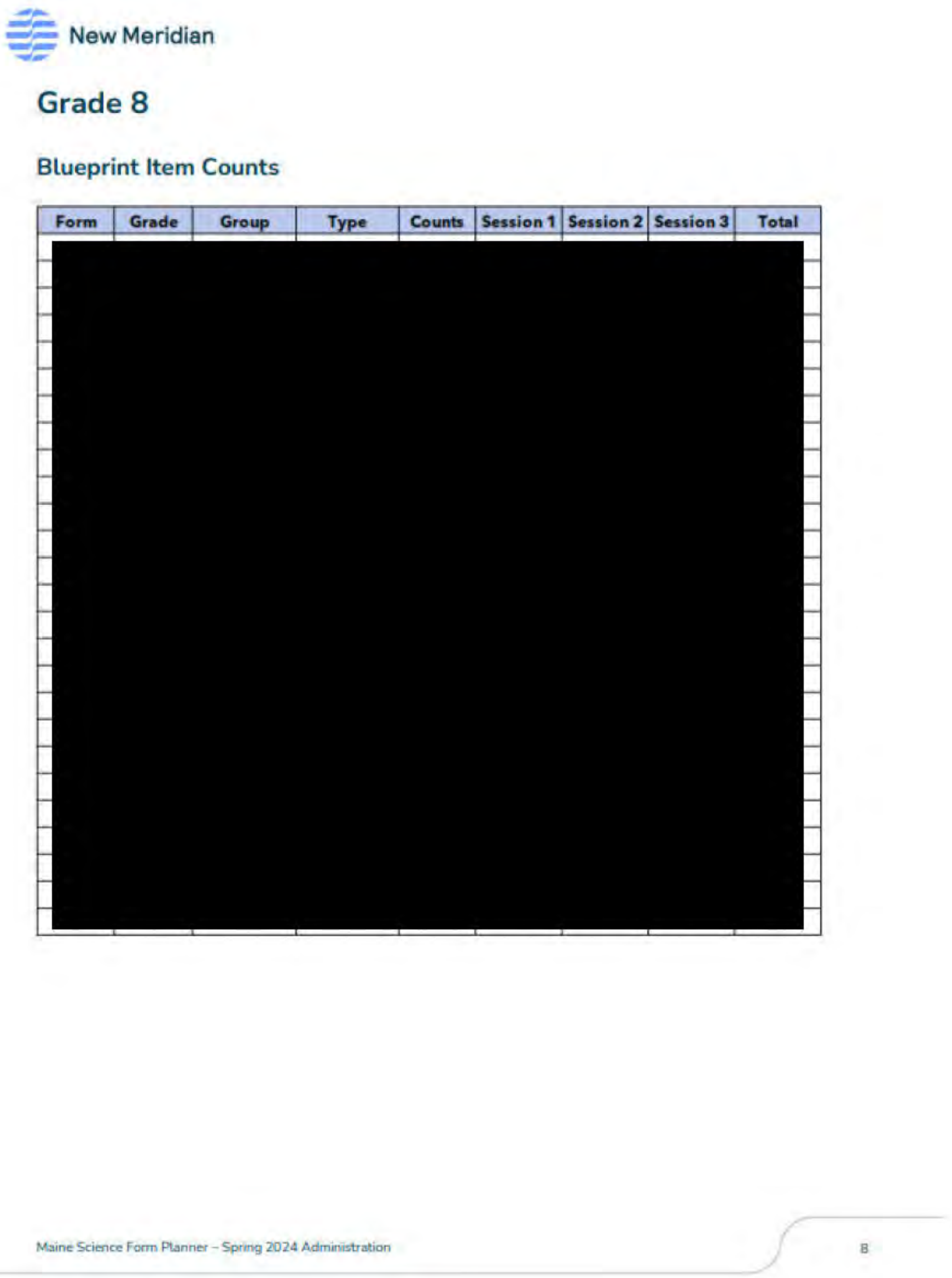
New Meridian

Science and Engineering Practices (SEP) Item Counts

Form	Grade	Science and Engineering Practices	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Cross Cutting Concepts (CCC) Item Counts

Form	Grade	Cross Cutting Concepts	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



New Meridian

Grade 8

Blueprint Item Counts

Form	Grade	Group	Type	Counts	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

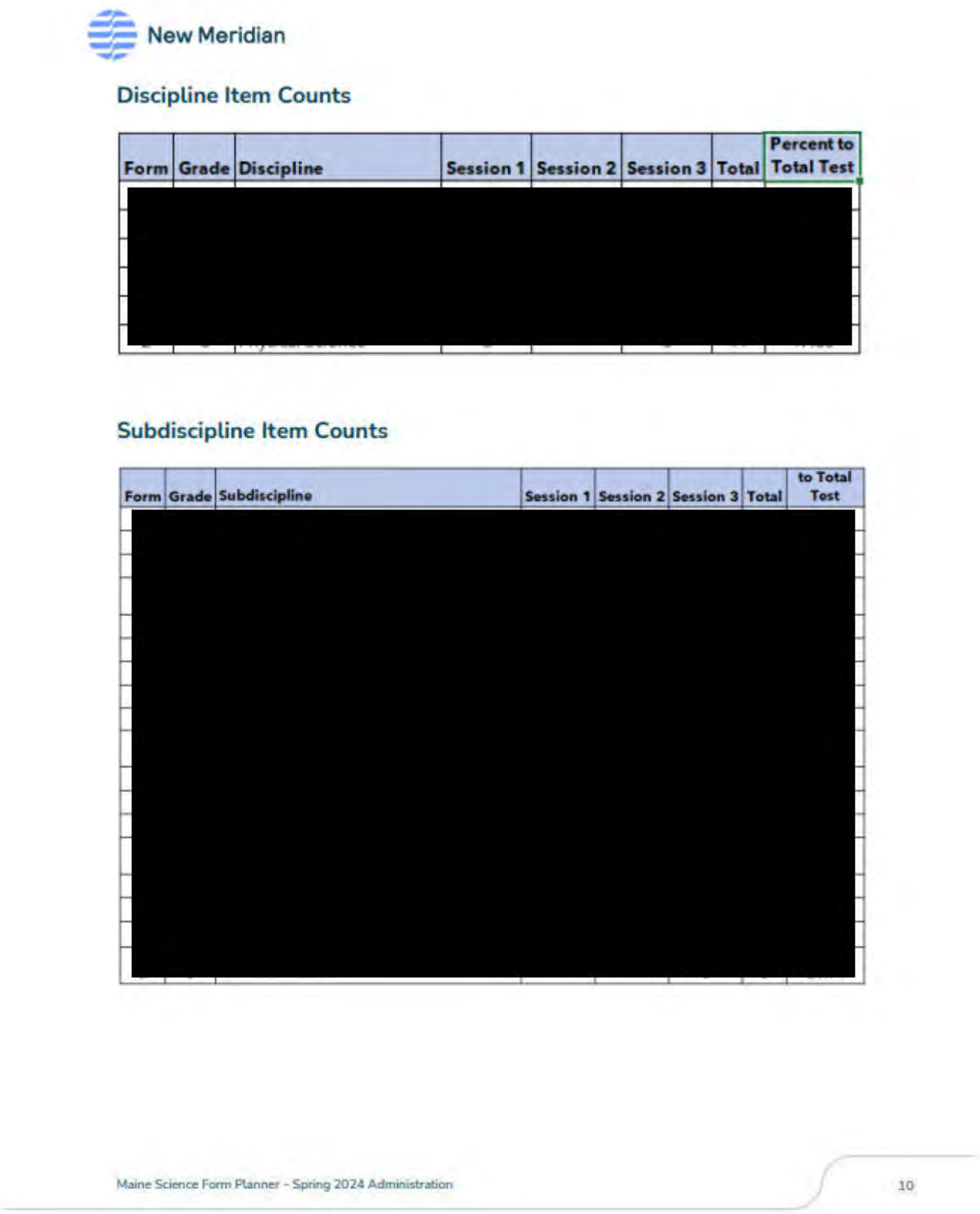
Maine Science Form Planner – Spring 2024 Administration

8

New Meridian

Blueprint Point Totals

Form	Grade	Group	Type	Cluster Counts or Item Points	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



Discipline Item Counts

Form	Grade	Discipline	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Subdiscipline Item Counts

Form	Grade	Subdiscipline	Session 1	Session 2	Session 3	Total	to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

New Meridian

Performance Expectation Item Counts

Form	Grade	Performance Expectation	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

New Meridian

Disciplinary Core Ideas (DCI) Item Counts

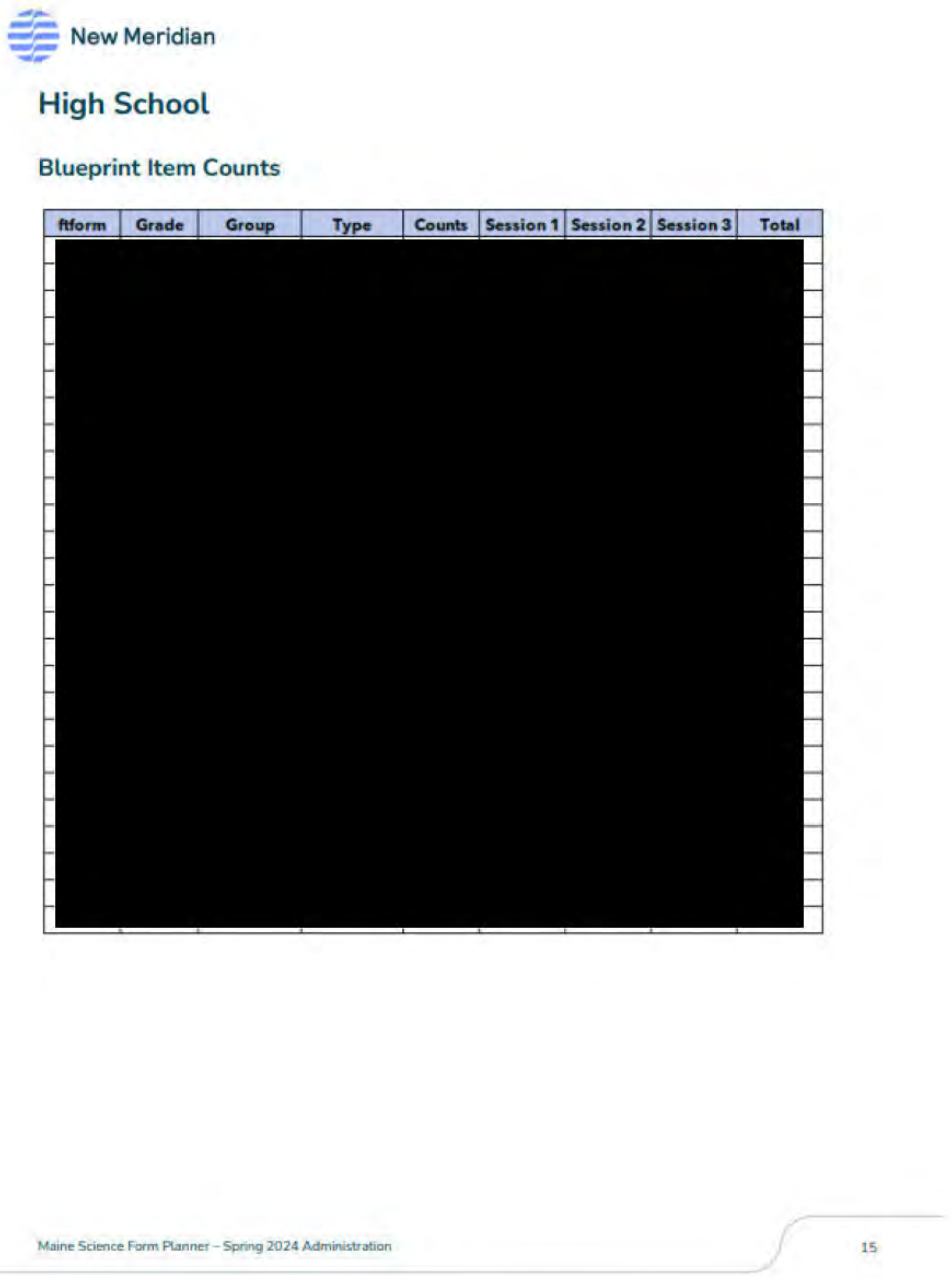
Form	Grade	Disciplinary Core Ideas	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Science and Engineering Practices (SEP) Item Counts

Form	Grade	Science and Engineering Practices	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Cross Cutting Concepts (CCC) Item Counts

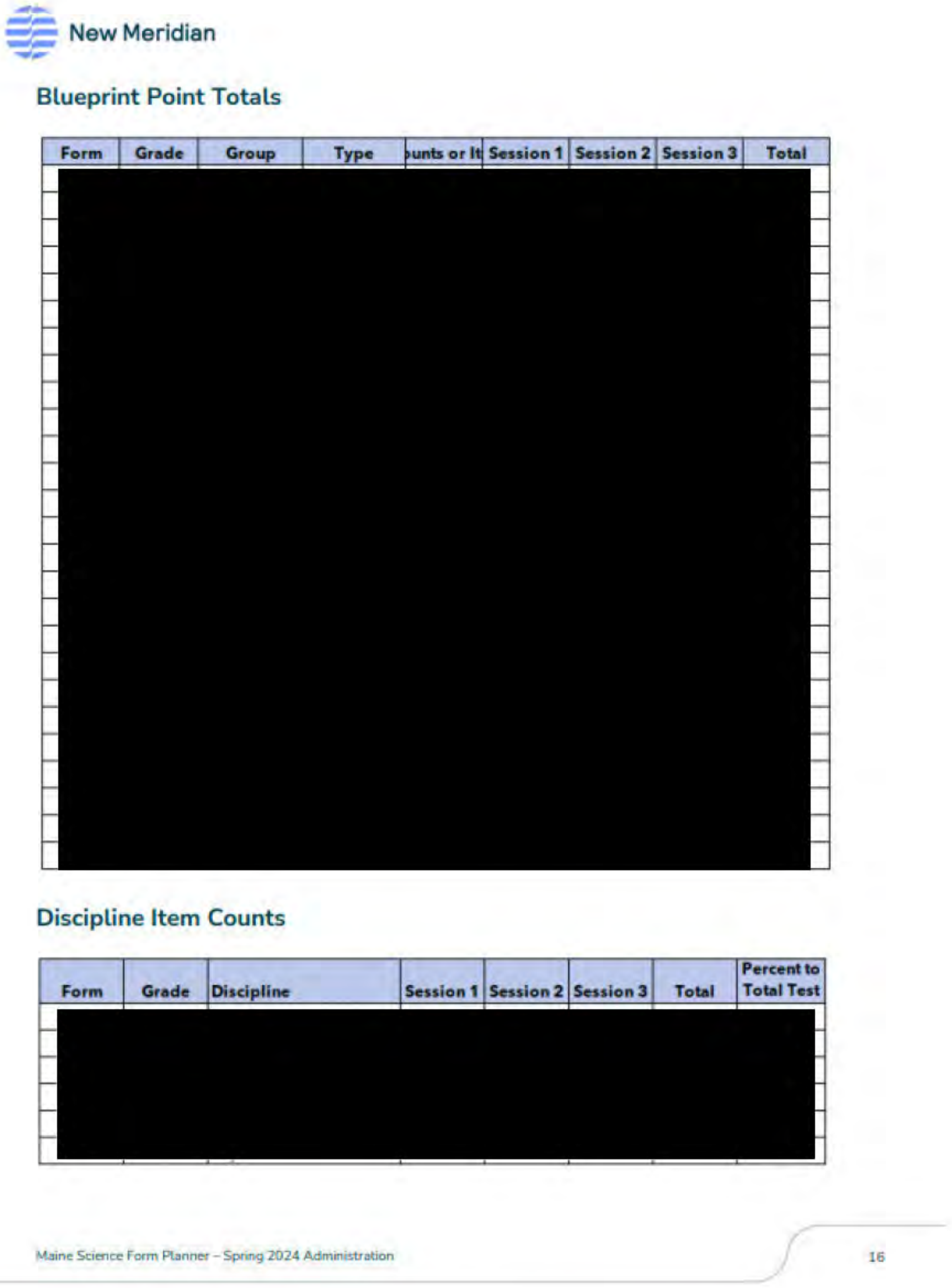
Form	Grade	Cross Cutting Concepts	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



High School

Blueprint Item Counts

ftform	Grade	Group	Type	Counts	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



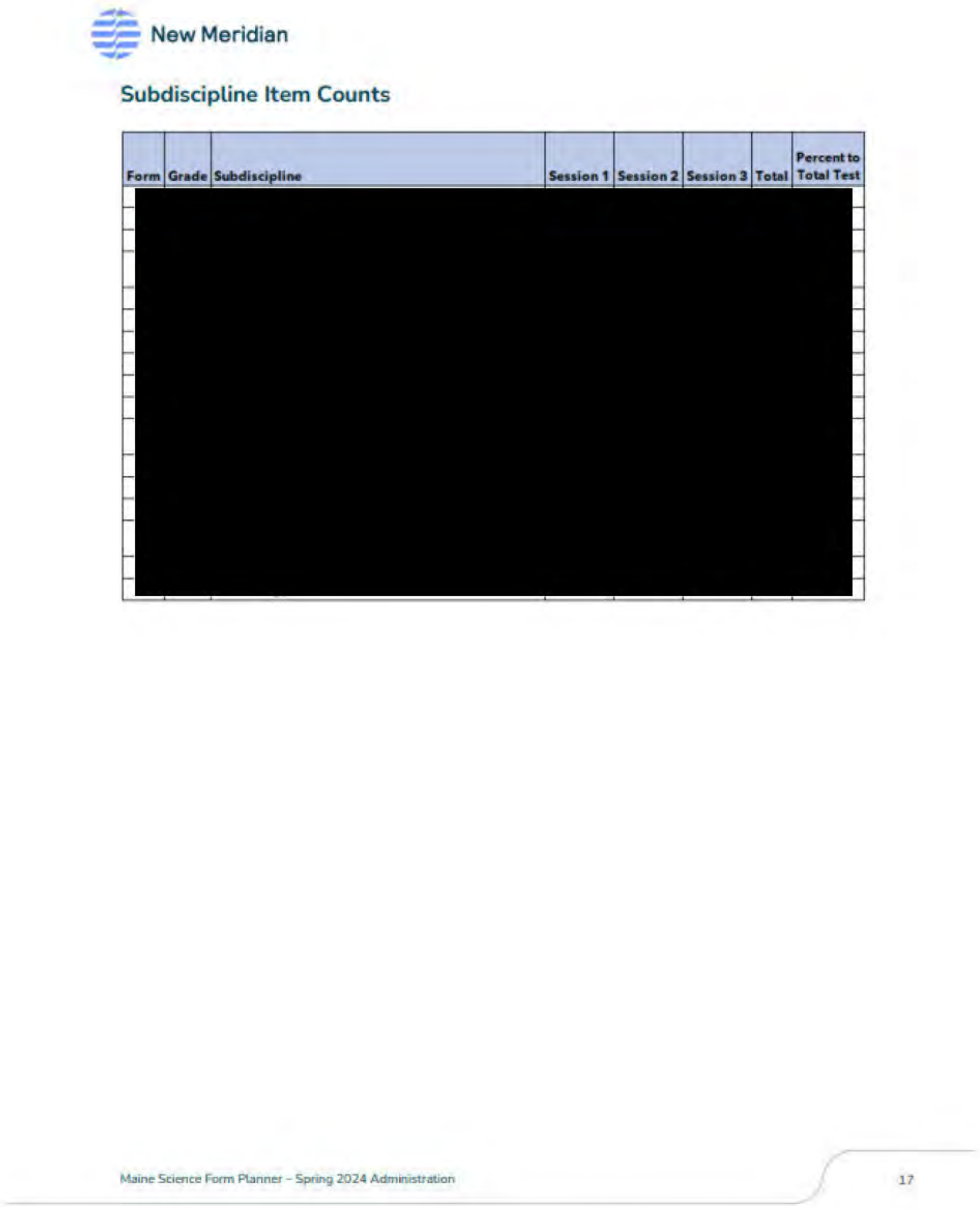
New Meridian

Blueprint Point Totals

Form	Grade	Group	Type	unts or It	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Discipline Item Counts

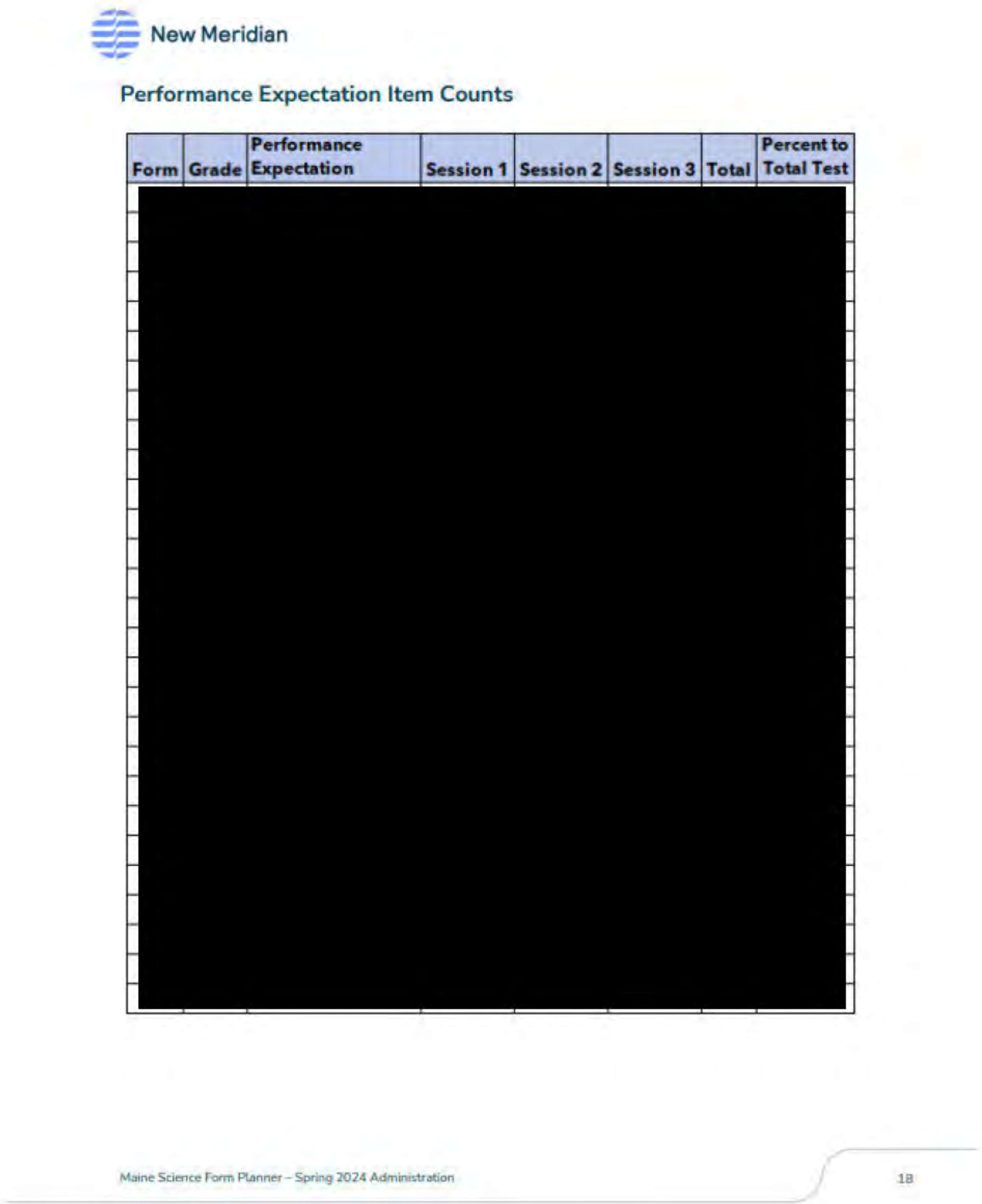
Form	Grade	Discipline	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



New Meridian

Subdiscipline Item Counts

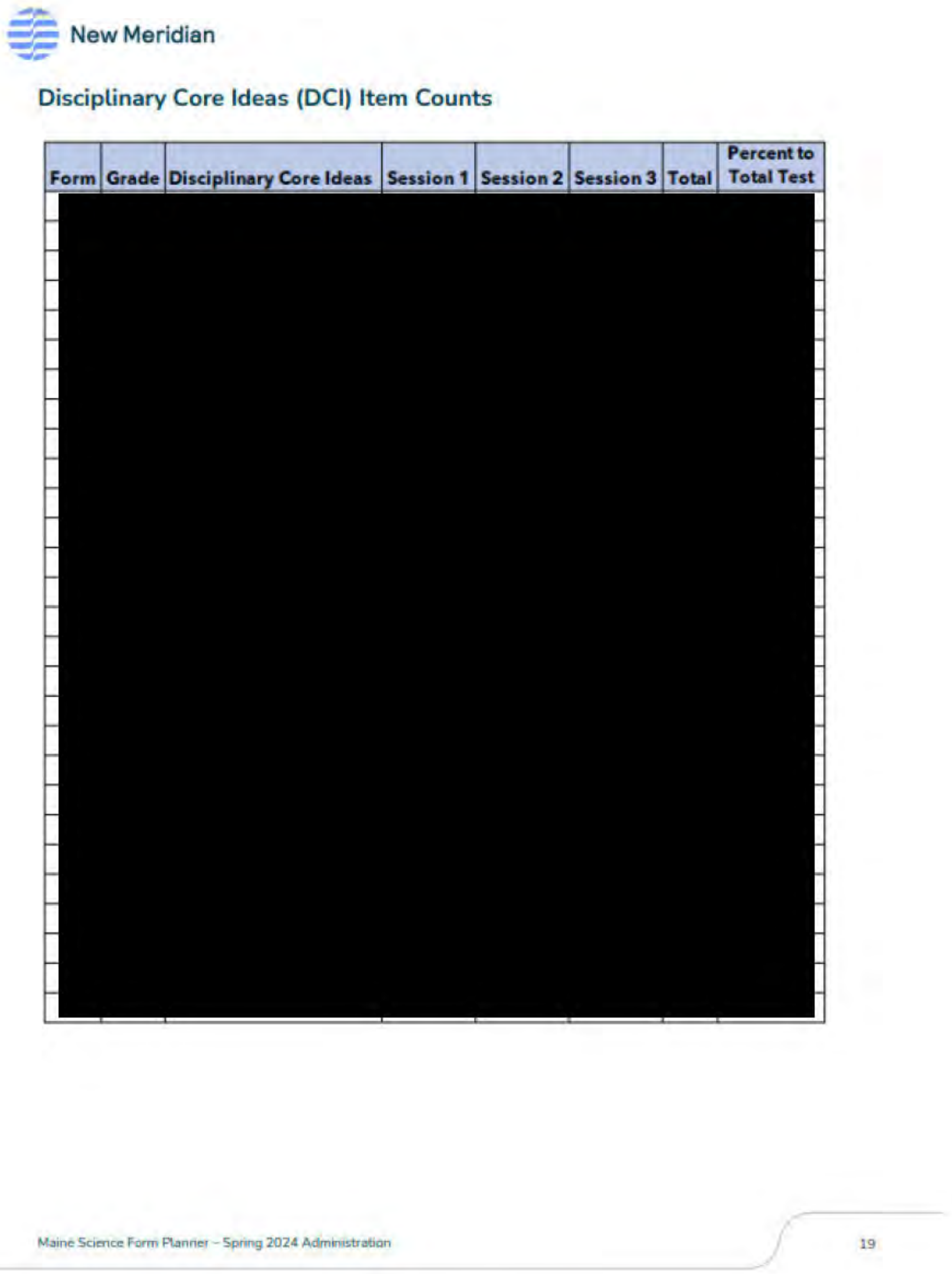
Form	Grade	Subdiscipline	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



New Meridian

Performance Expectation Item Counts

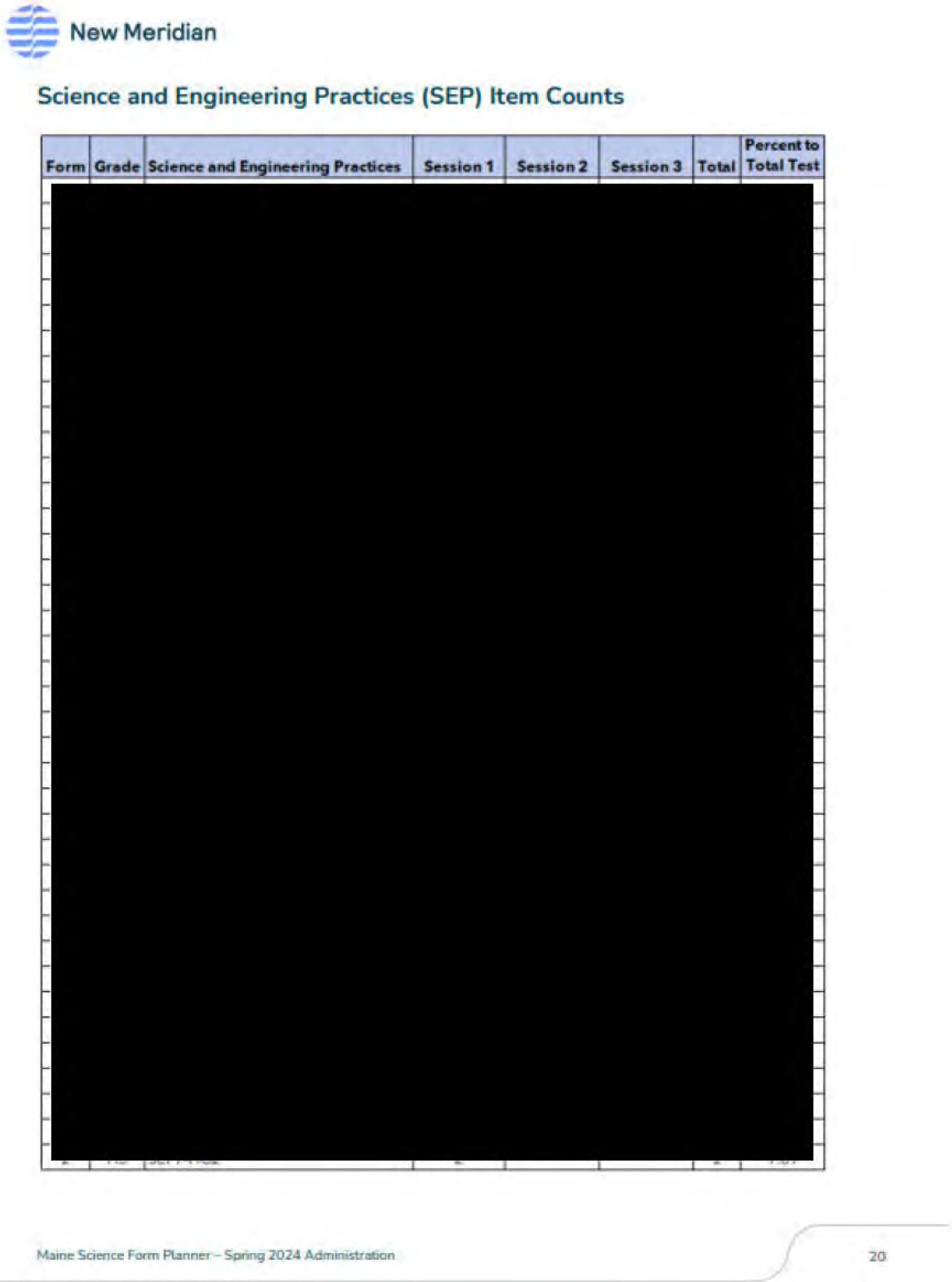
Form	Grade	Performance Expectation	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]



New Meridian

Disciplinary Core Ideas (DCI) Item Counts

Form	Grade	Disciplinary Core Ideas	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

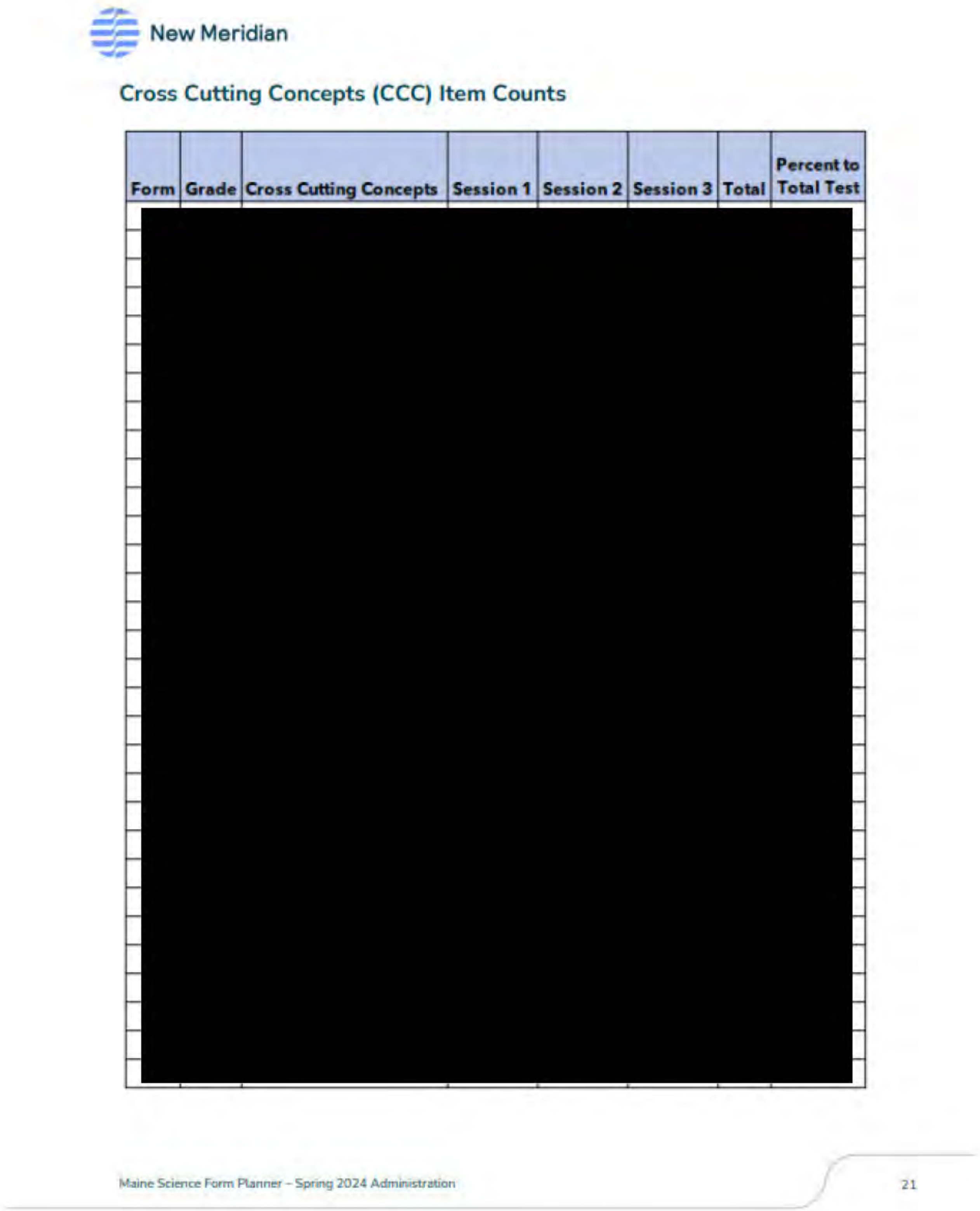


New Meridian

Science and Engineering Practices (SEP) Item Counts

Form	Grade	Science and Engineering Practices	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Maine Science Form Planner – Spring 2024 Administration 20



New Meridian

Cross Cutting Concepts (CCC) Item Counts

Form	Grade	Cross Cutting Concepts	Session 1	Session 2	Session 3	Total	Percent to Total Test
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

The Science Disciplines

Table 50.Science Discipline Distribution of Items

Grade	Discipline	Session 1	Session 2	Session 3	Total	Percent
05	Earth and Space Science	[redacted]	[redacted]	[redacted]	12	35.29
05	Life Science	[redacted]	[redacted]	[redacted]	8	23.53
05	Physical Science	[redacted]	[redacted]	[redacted]	14	41.18
08	Earth and Space Science	[redacted]	[redacted]	[redacted]	15	36.59
08	Life Science	[redacted]	[redacted]	[redacted]	11	26.83
08	Physical Science	[redacted]	[redacted]	[redacted]	15	36.59
HS	Earth and Space Science	[redacted]	[redacted]	[redacted]	12	29.27
HS	Life Science	[redacted]	[redacted]	[redacted]	13	31.71
HS	Physical Science	[redacted]	[redacted]	[redacted]	16	39.02

Number of Items by Item Type

Table 51. Number of Items by Item Type

Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Maximum Score Points by Item Type

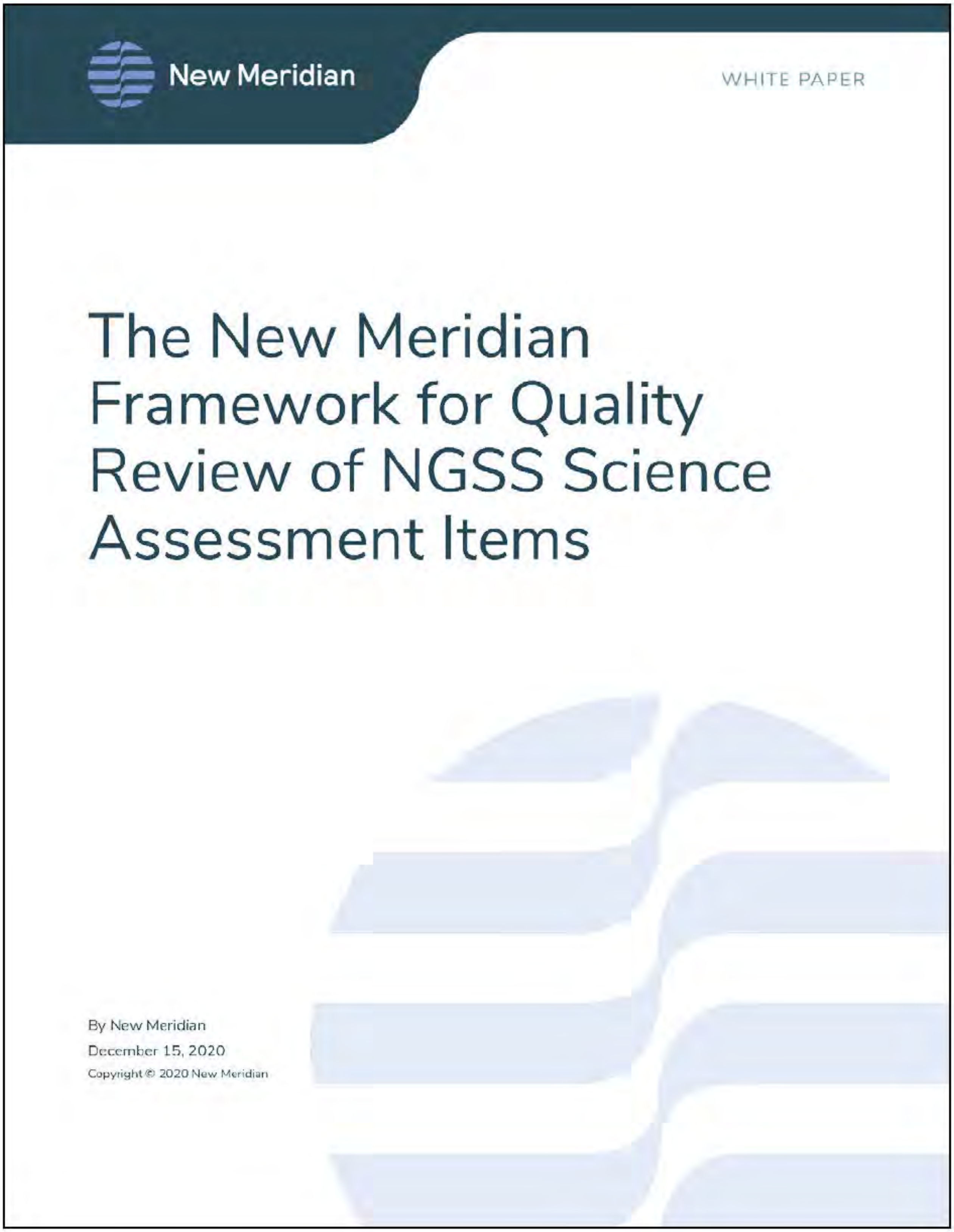
Table 52. Maximum Score Points by Item Type

Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]							
Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]							
Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total

Grade	Group	Type	Data	Session 1	Session 2	Session 3	Total
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Appendix C. New Meridian Framework for Quality Review of NGSS Science Assessment Items

(Document begins on next page.)



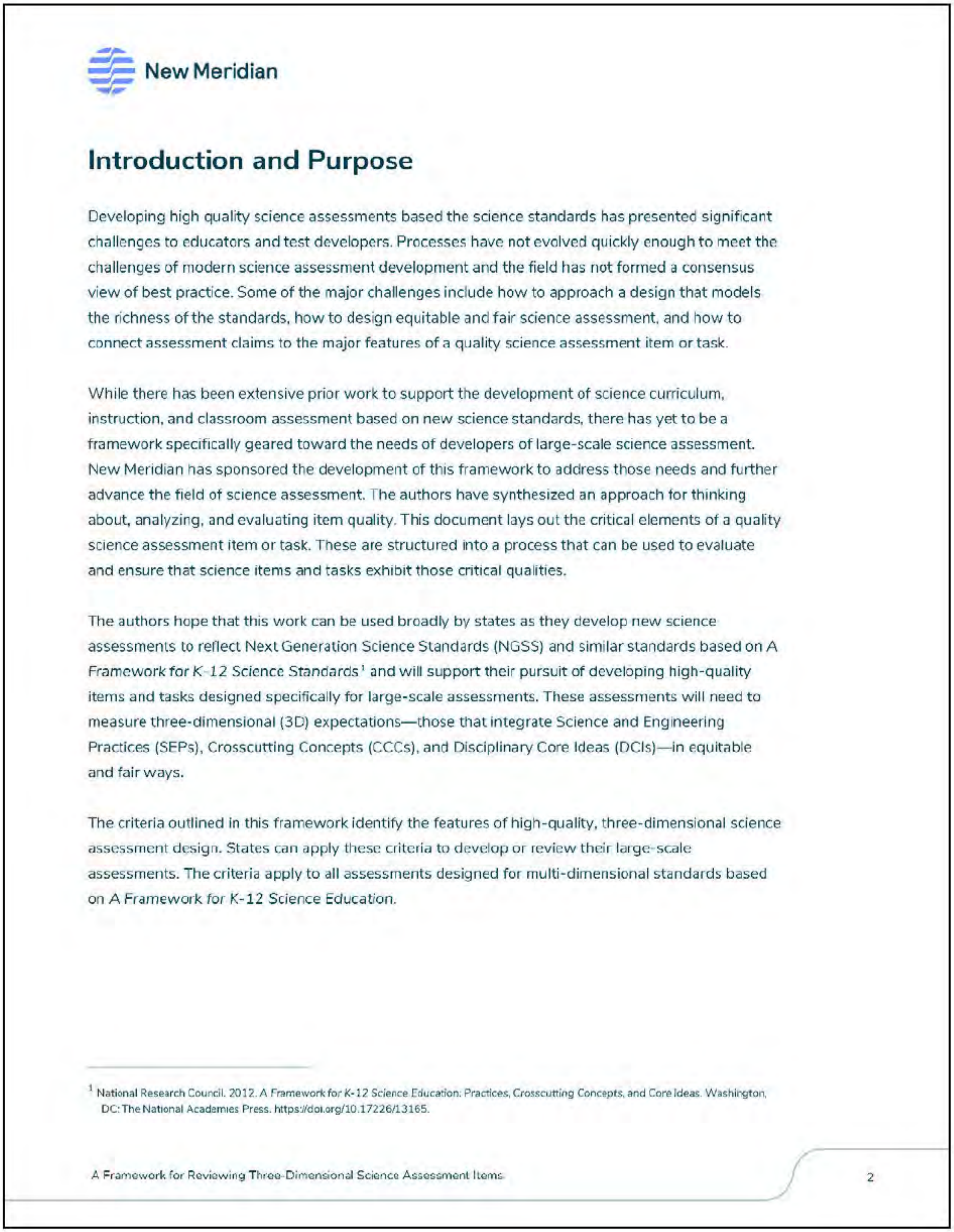
New Meridian

WHITE PAPER

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

By New Meridian

December 15, 2020



Introduction and Purpose

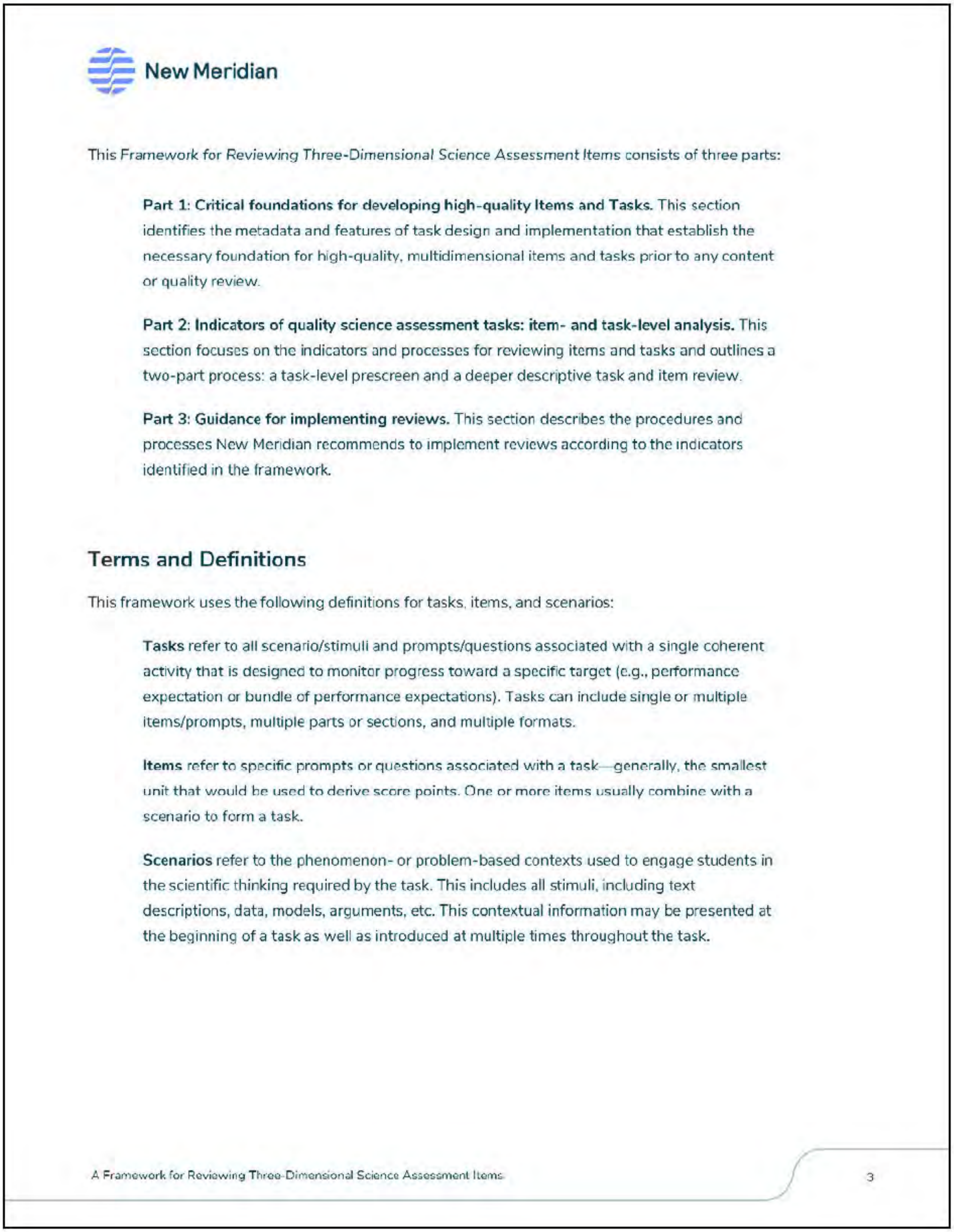
Developing high quality science assessments based the science standards has presented significant challenges to educators and test developers. Processes have not evolved quickly enough to meet the challenges of modern science assessment development and the field has not formed a consensus view of best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

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

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

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

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



New Meridian

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

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

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

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

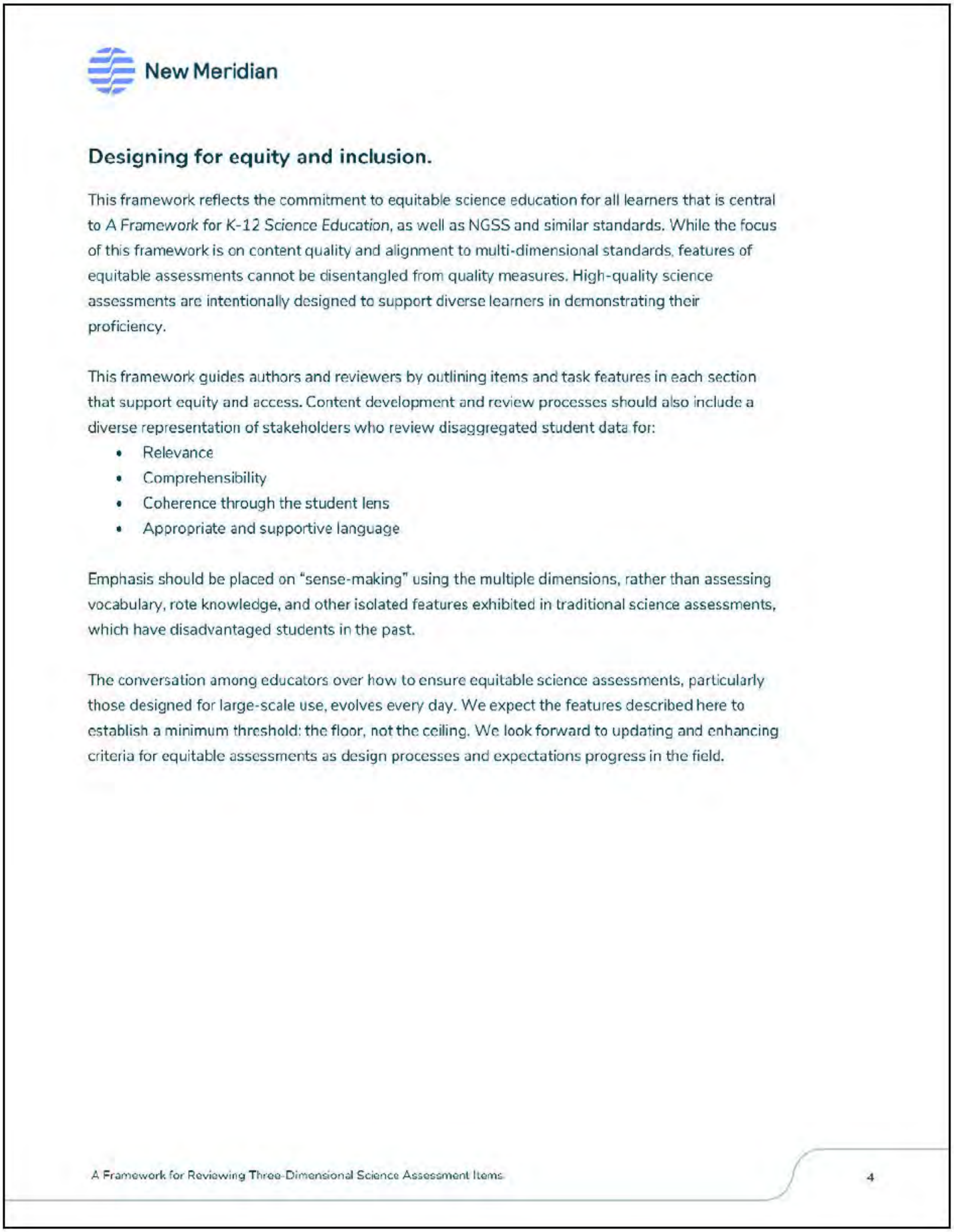
Terms and Definitions

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

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

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

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



New Meridian

Designing for equity and inclusion.

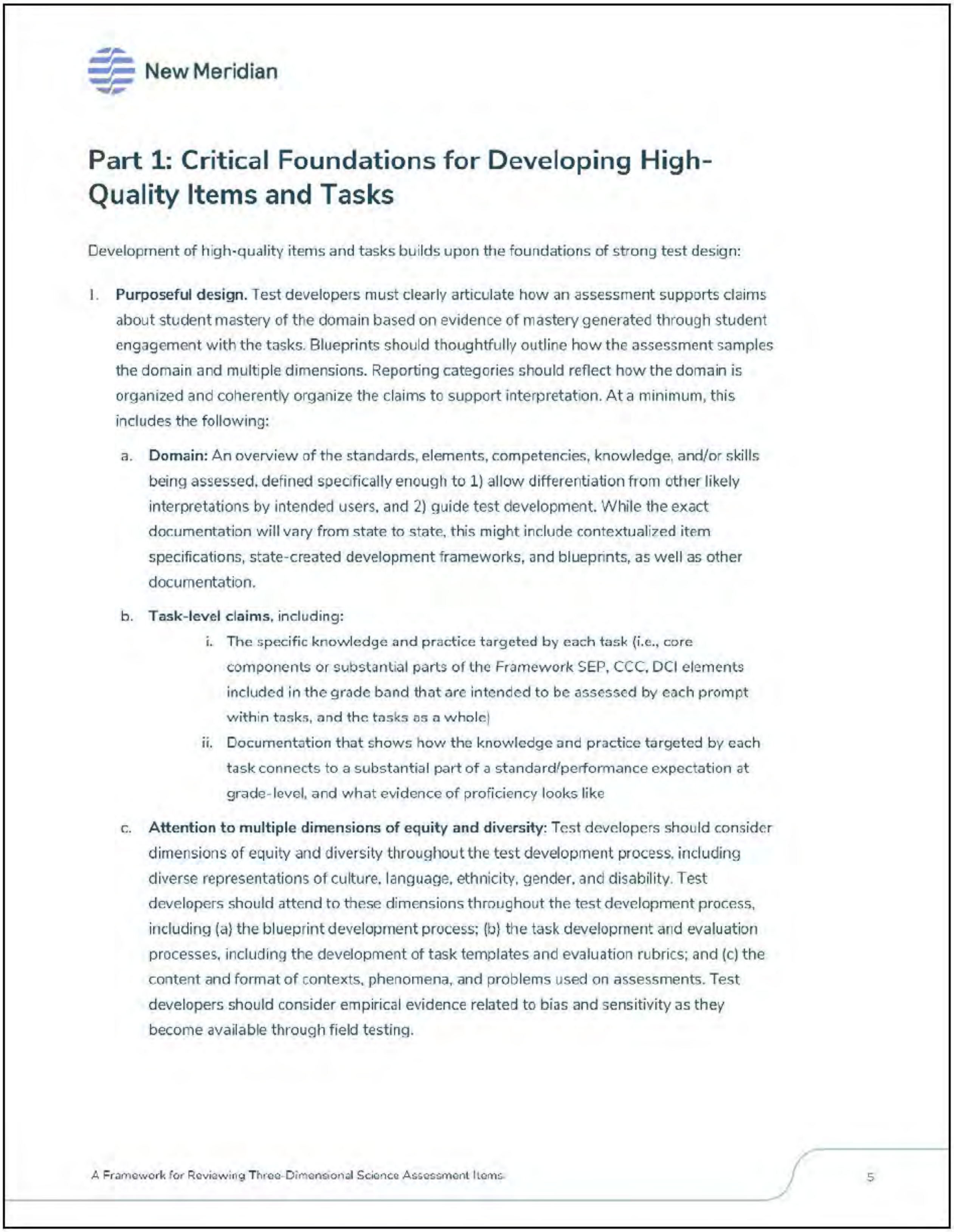
This framework reflects the commitment to equitable science education for all learners that is central to *A Framework for K-12 Science Education*, as well as NGSS and similar standards. While the focus of this framework is on content quality and alignment to multi-dimensional standards, features of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

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

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

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

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



New Meridian

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

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

1. **Purposeful design.** Test developers must clearly articulate how an assessment supports claims about student mastery of the domain based on evidence of mastery generated through student engagement with the tasks. Blueprints should thoughtfully outline how the assessment samples the domain and multiple dimensions. Reporting categories should reflect how the domain is organized and coherently organize the claims to support interpretation. At a minimum, this includes the following:

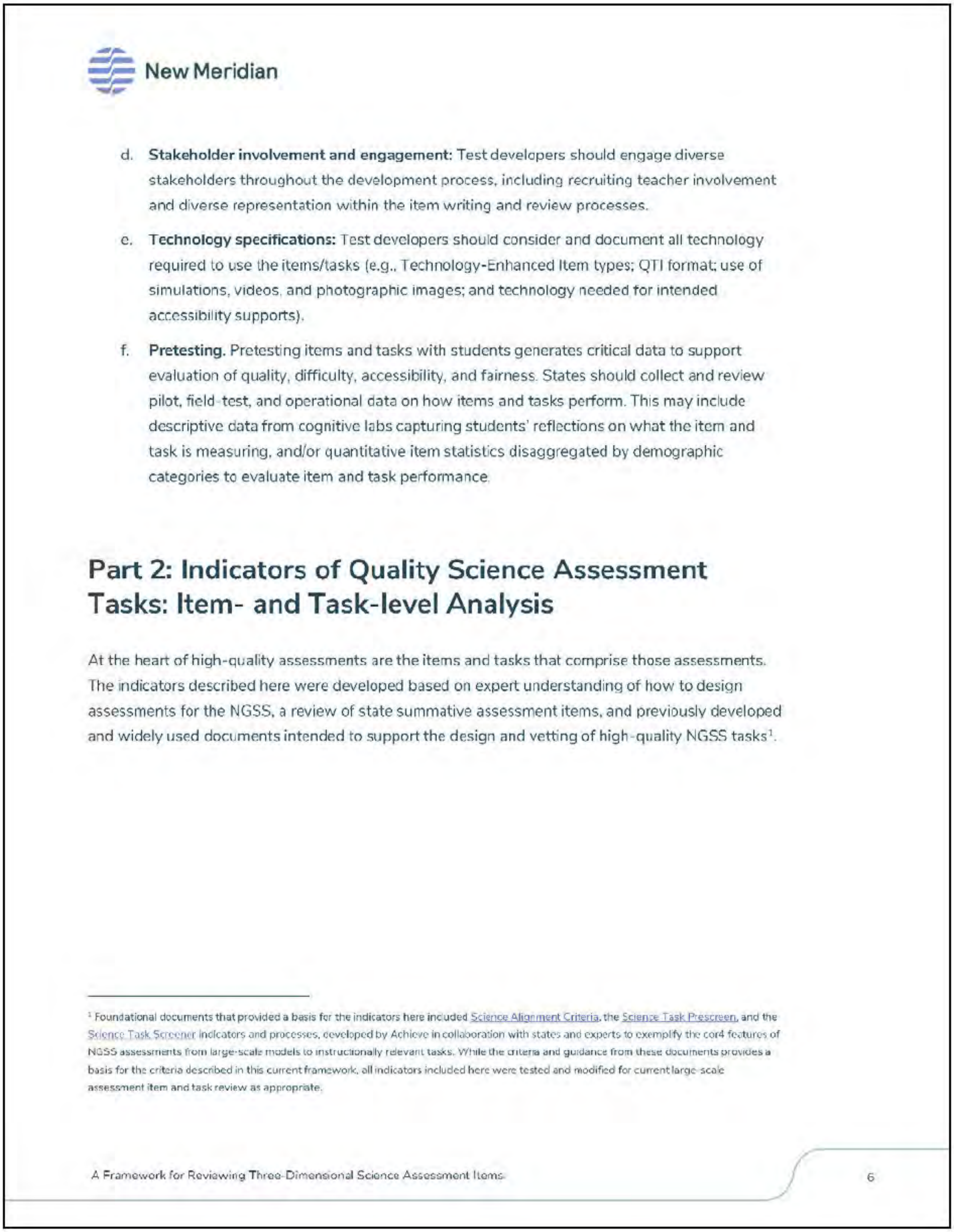
    a. **Domain:** An overview of the standards, elements, competencies, knowledge, and/or skills being assessed, defined specifically enough to 1) allow differentiation from other likely interpretations by intended users, and 2) guide test development. While the exact documentation will vary from state to state, this might include contextualized item specifications, state-created development frameworks, and blueprints, as well as other documentation.

    b. **Task-level claims**, including:

        i. The specific knowledge and practice targeted by each task (i.e., core components or substantial parts of the Framework SEP, CCC, DCI elements included in the grade band that are intended to be assessed by each prompt within tasks, and the tasks as a whole)

        ii. Documentation that shows how the knowledge and practice targeted by each task connects to a substantial part of a standard/performance expectation at grade-level, and what evidence of proficiency looks like

    c. **Attention to multiple dimensions of equity and diversity:** Test developers should consider dimensions of equity and diversity throughout the test development process, including diverse representations of culture, language, ethnicity, gender, and disability. Test developers should attend to these dimensions throughout the test development process, including (a) the blueprint development process; (b) the task development and evaluation processes, including the development of task templates and evaluation rubrics; and (c) the content and format of contexts, phenomena, and problems used on assessments. Test developers should consider empirical evidence related to bias and sensitivity as they become available through field testing.



New Meridian

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

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

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

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

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

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

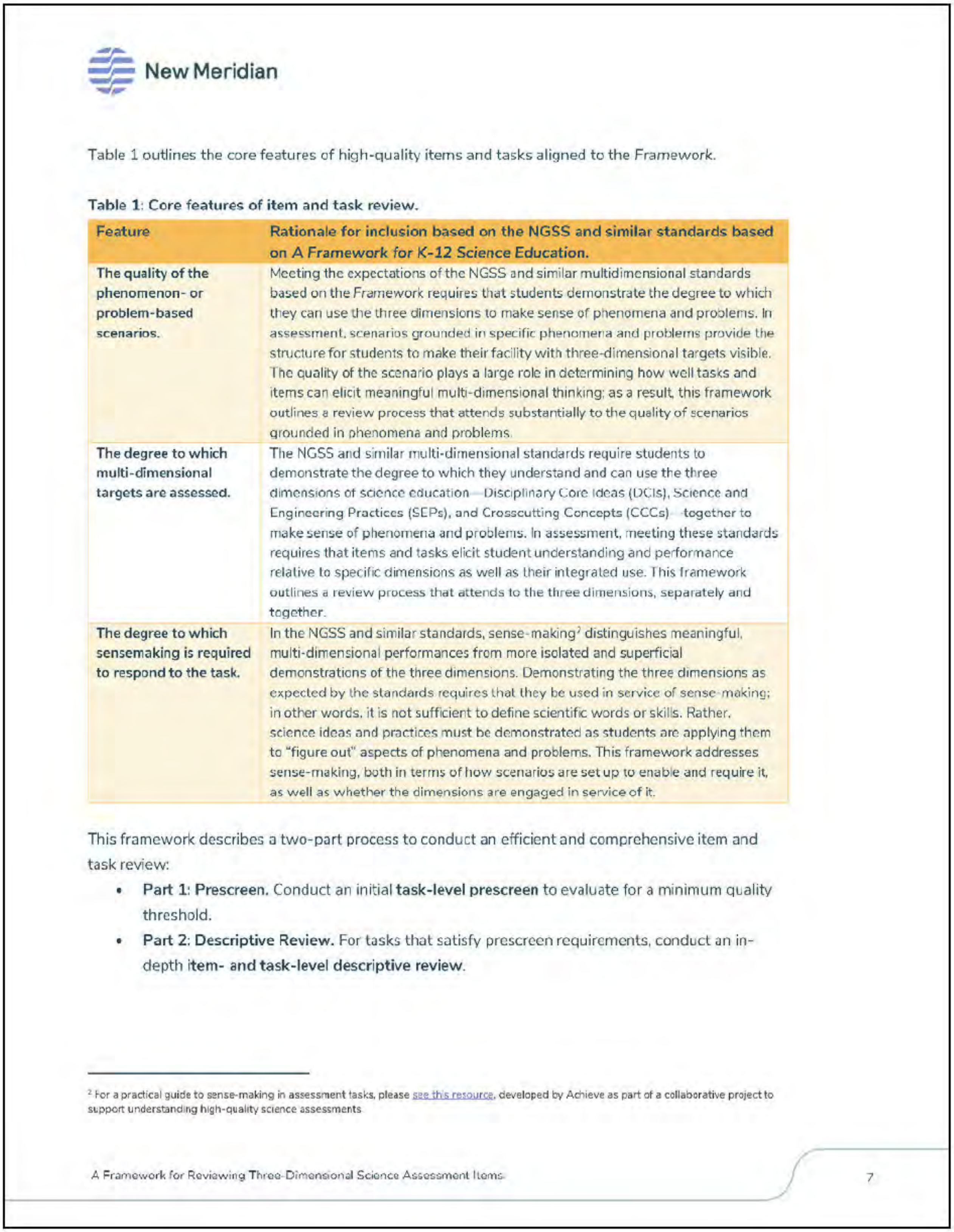


Table 1 outlines the core features of high-quality items and tasks aligned to the *Framework*.

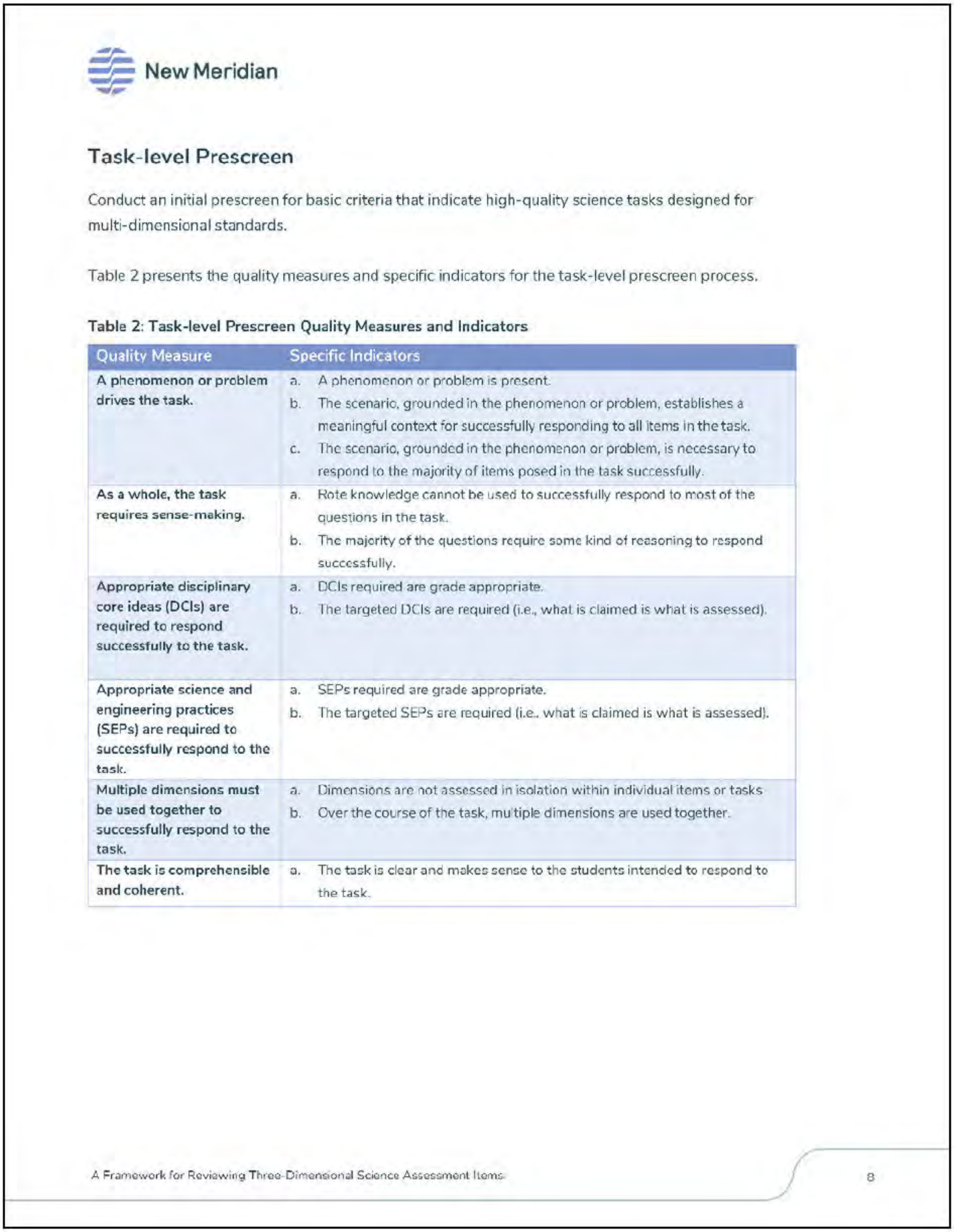
Table 1: Core features of item and task review.

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

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

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

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



New Meridian

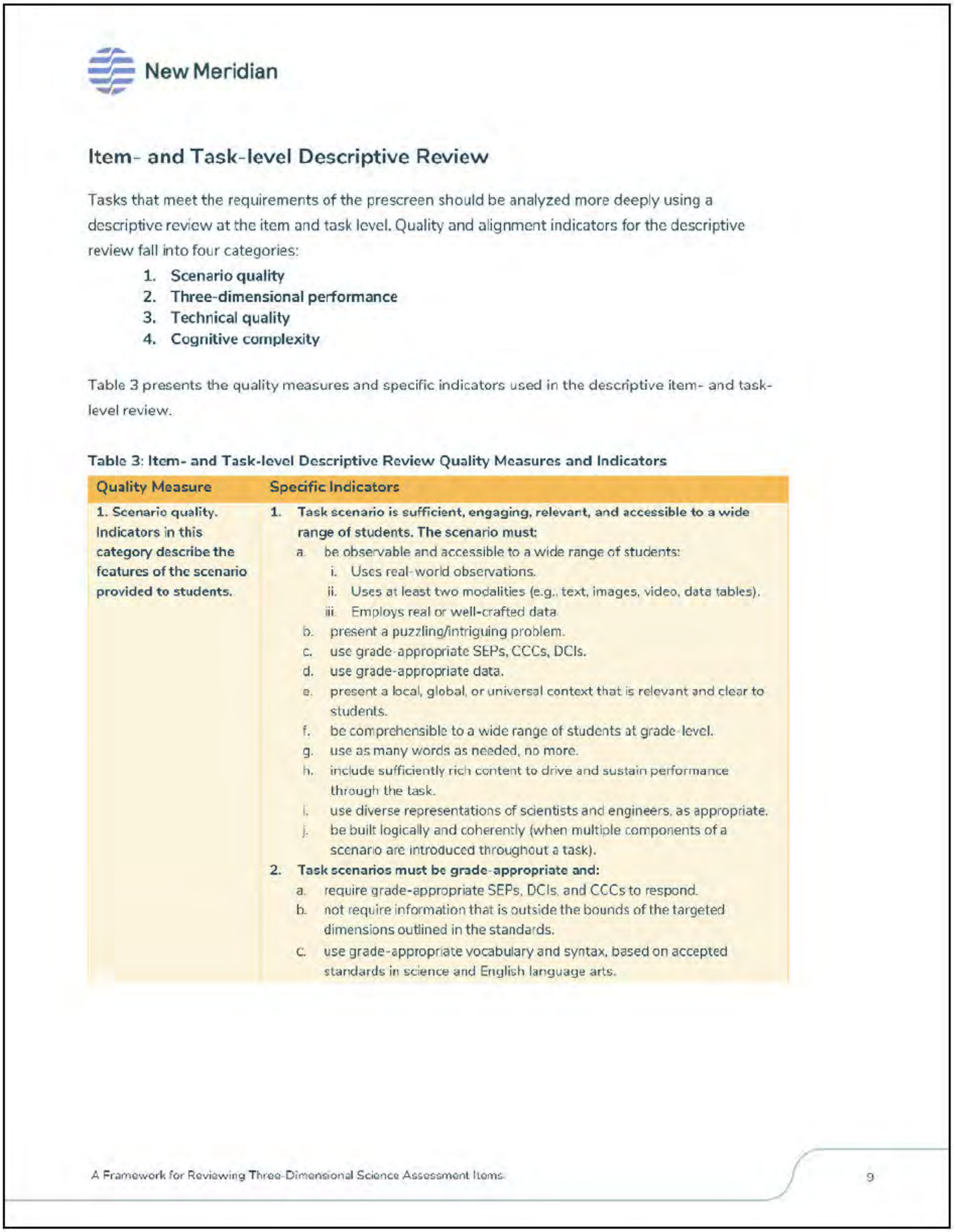
Task-level Prescreen

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

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

Table 2: Task-level Prescreen Quality Measures and Indicators

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



New Meridian

Item- and Task-level Descriptive Review

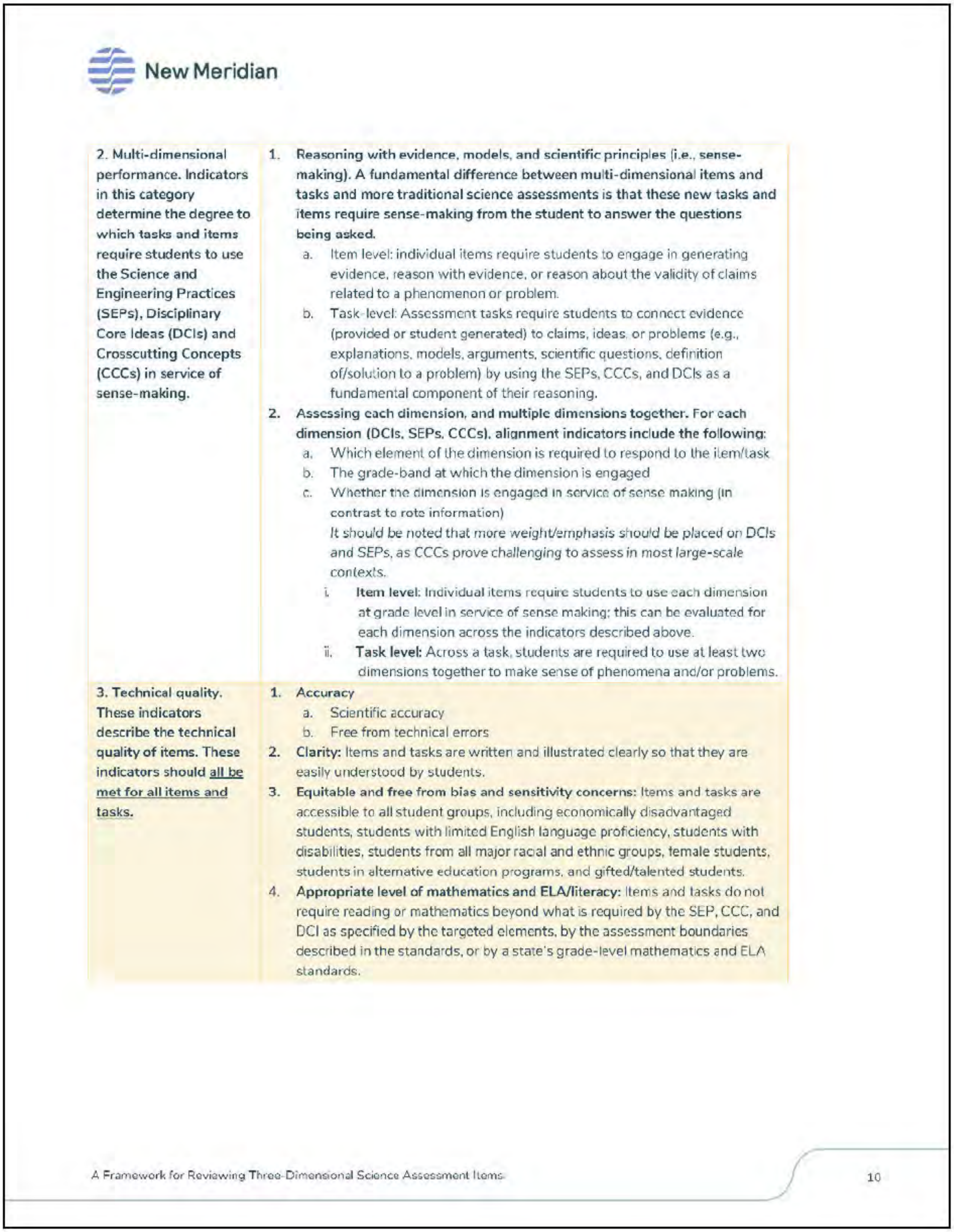
Tasks that meet the requirements of the prescreen should be analyzed more deeply using a descriptive review at the item and task level. Quality and alignment indicators for the descriptive review fall into four categories:

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

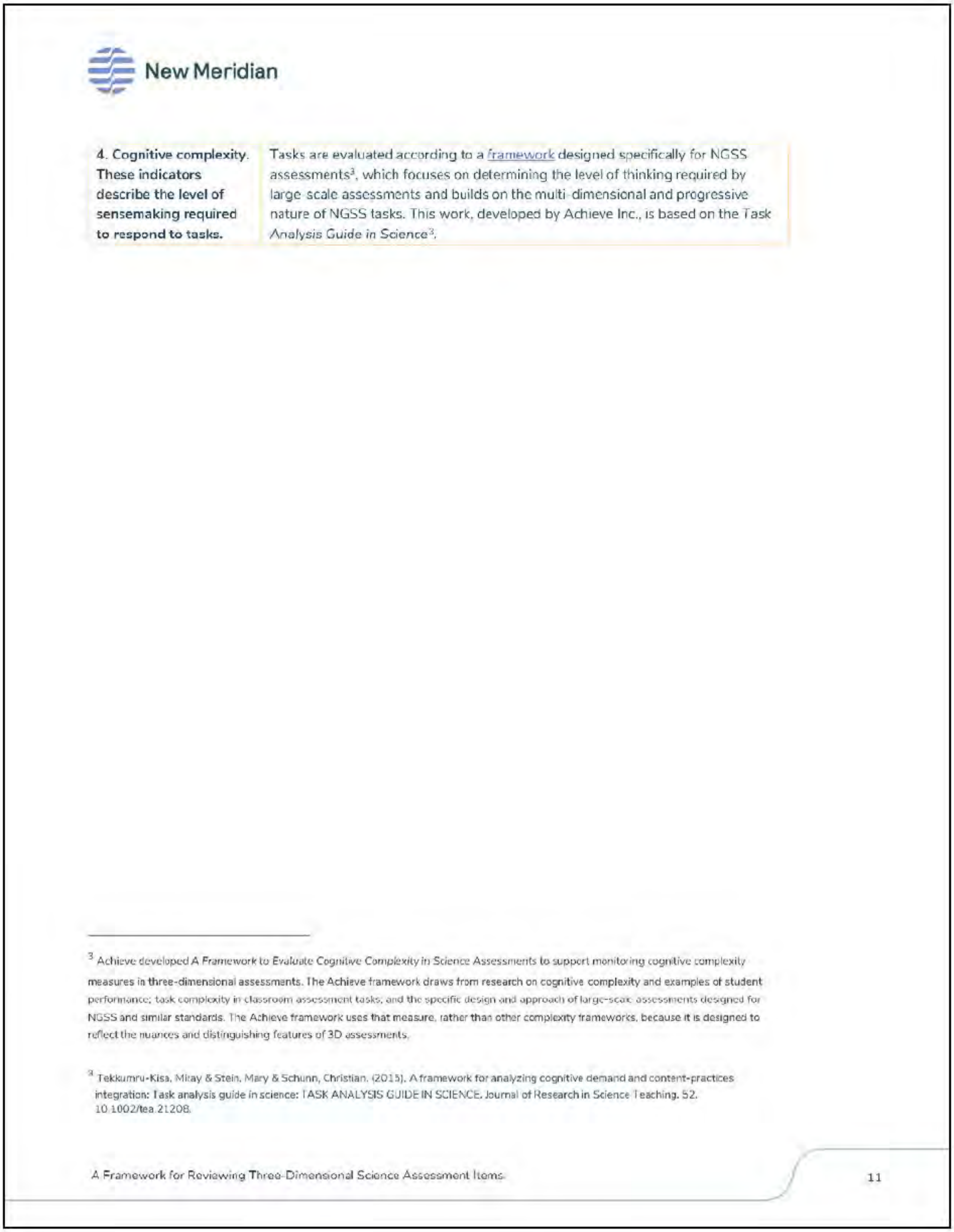
Table 3 presents the quality measures and specific indicators used in the descriptive item- and task-level review.

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

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



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



New Meridian

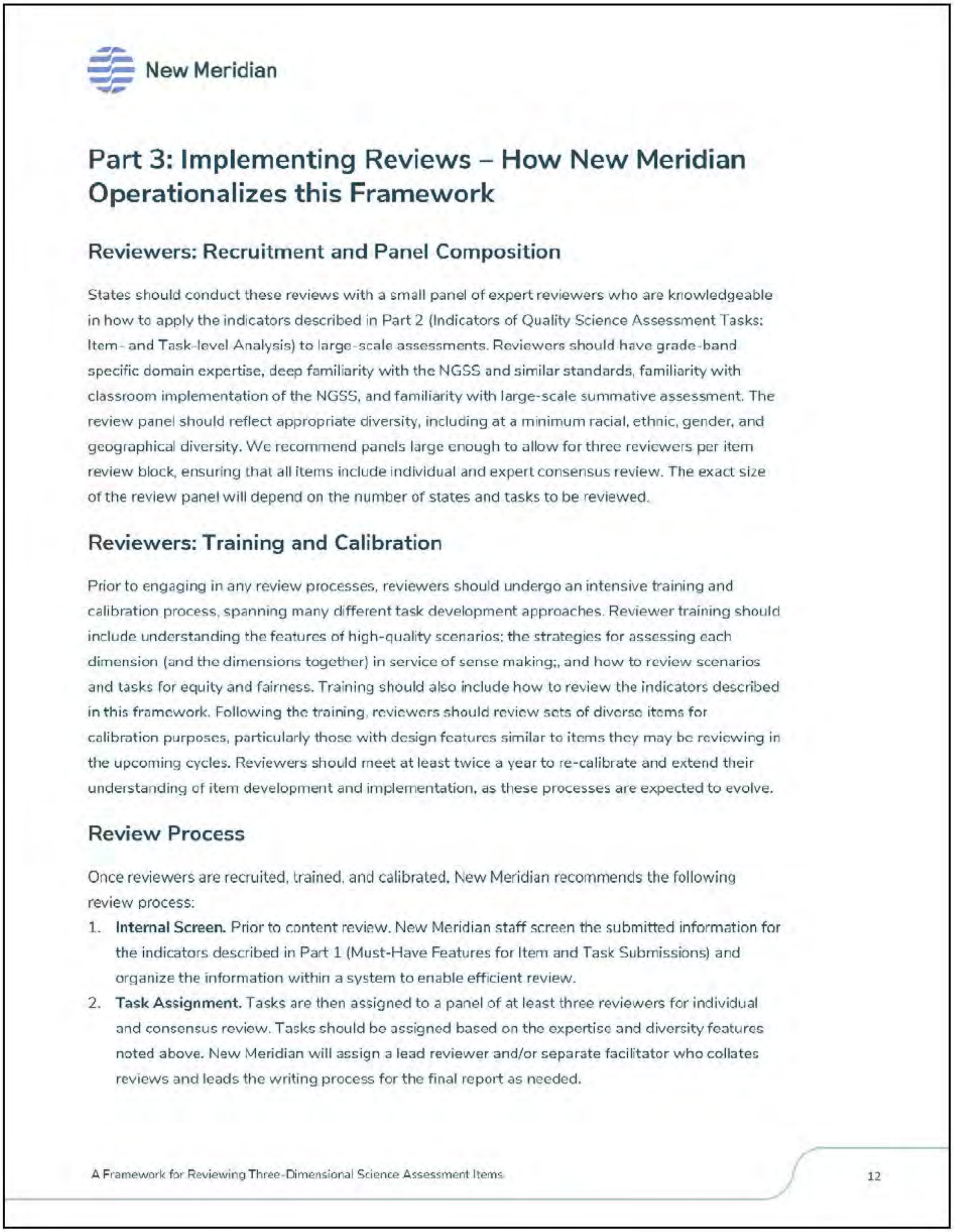
4. Cognitive complexity. These indicators describe the level of sensemaking required to respond to tasks.	Tasks are evaluated according to a framework designed specifically for NGSS assessments[3], which focuses on determining the level of thinking required by large-scale assessments and builds on the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

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

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

A Framework for Reviewing Three-Dimensional Science Assessment Items

11



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

Reviewers: Recruitment and Panel Composition

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

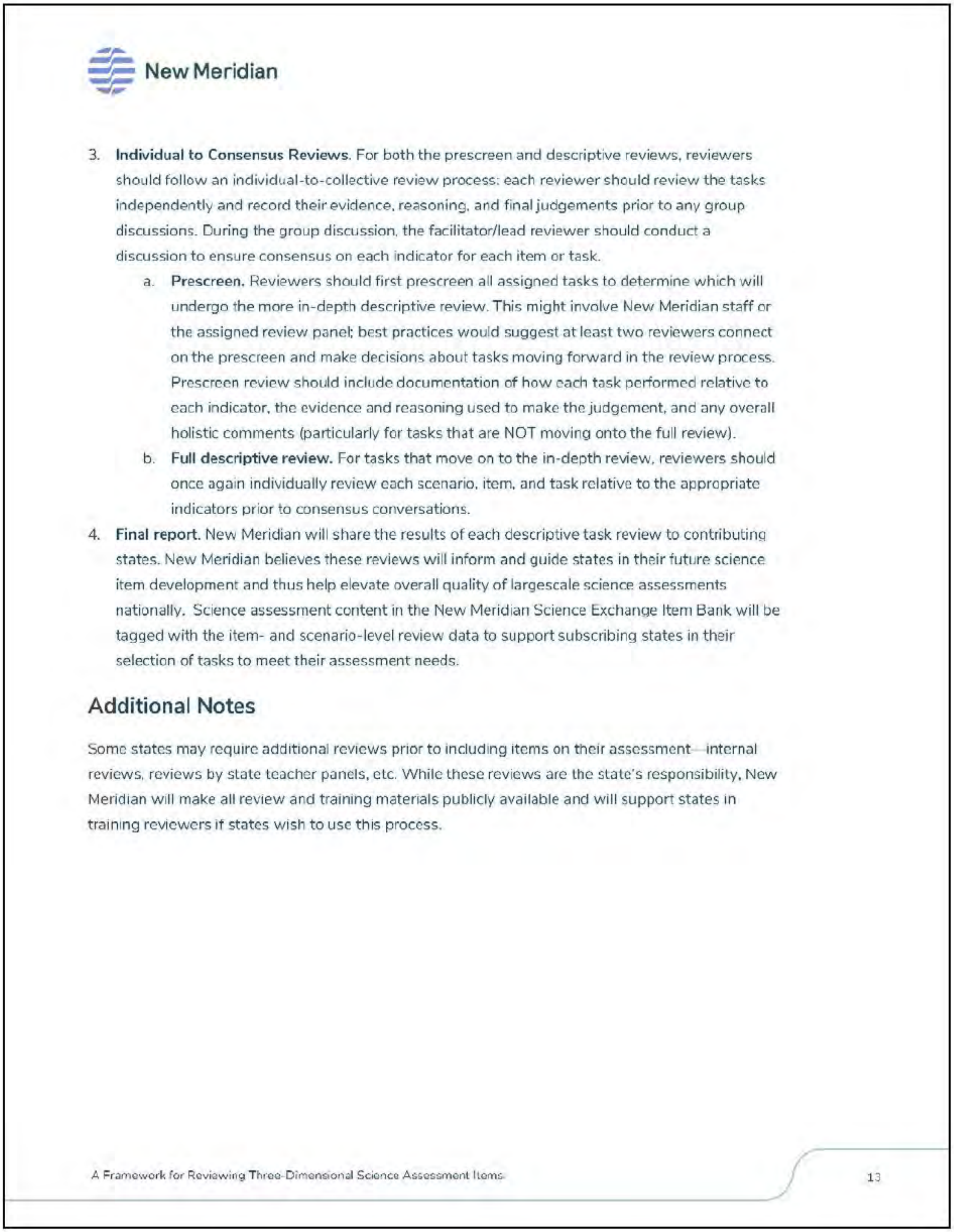
Reviewers: Training and Calibration

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

Review Process

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

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



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

Additional Notes

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

Appendix D. Item Writing Training Agenda

(Document begins on next page.)

New Meridian Item Development January 2023
Agenda

Day	Time	Activity	Structure
Wed	9-10:30	Introductions and Getting Started	Whole Group
	10:30-12:00	Unpacking DCIs, SEP, CCC	Individual
	12:30-1:00	Lunch	
	1:30-2:30	Sharing the Unpacking	Whole Group
	2:30-3:30	Focus on Topic Unpacking	Individual
	3:30 – 4:00	Begin Phenomena Search	Whole Group
	4:00-5:00	Phenomenon Feedback	Whole Group
Thursday	9:00-10:00	Sense-Making of the Phenomenon*	Whole group
	10:30-12:00	Storyline Development	Writing Teams
	12:00-1:00	Lunch	
	1:00-1:45	Share Storyline	Whole Group
	2:00-5:00	Cluster Development/ 1x1 with each item writer	Writing Teams
Friday	9:00-10:00	Individual work time	Individual
	9:30 - 12:00	Cluster Development	Writing Teams
	12:00 - 1:00	Lunch	
	1:00- 2:00	Peer Review Process: use prescreen of Review Framework	Whole Group
	1:30-3:00	Peer Review	Paired Writing Teams
	3:00-5:00	Cluster Development	Writing Teams
Saturday	9:00-9:30	Content Review Process	Whole Group
	9:30-12:00	Content Review and Cluster Development	Content Review Teams
	12:00-1:00	Lunch	
	1:00-4:00	Cluster Development	Whole group
	4:00-5:00	Cluster Development Share-out	Writing Teams
Sunday	9:00-9:30	Submission Process	Whole Group
	9:30-12:00	Cluster Development	Writing Teams
	12:00-1:00	Lunch	
	1:00-5:00	Complete and Submit Cluster	Individual

Appendix E. Asset Development Plan

Breakdown is by NGSS Topic

Item Writing assignments are determined by low number of points operationally in a topic.

Table 53. Asset Development Plan for Grade 5

[redacted]

Table 54. Asset Development Plan for Grade 8

[redacted]

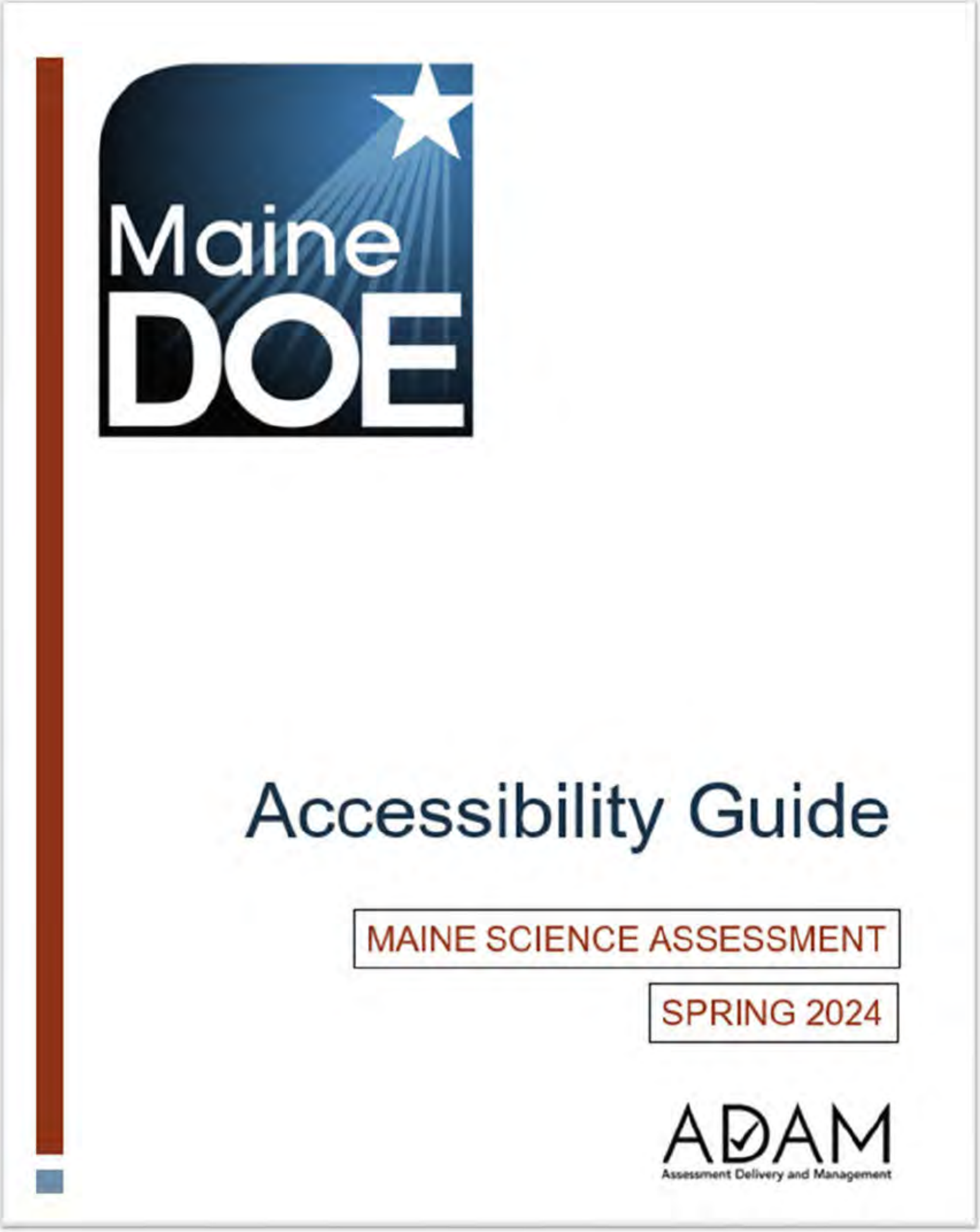
Table 55. Asset Development Plan for High School

[redacted]

Appendix F. Manuals

Accessibility Guide

(Document begins on next page.)



Maine DOE

Accessibility Guide

MAINE SCIENCE ASSESSMENT

SPRING 2024

ADAM
Assessment Delivery and Management

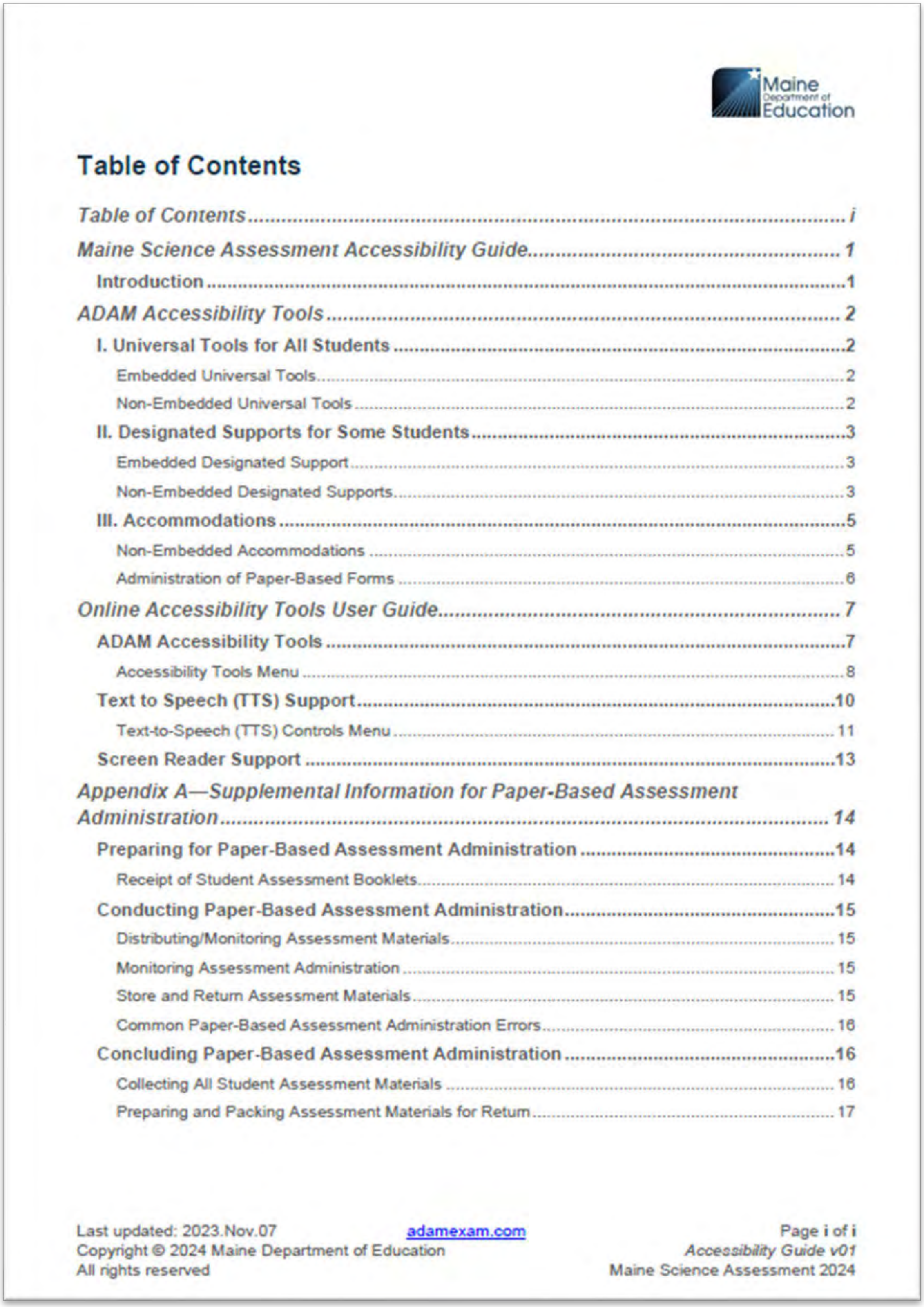
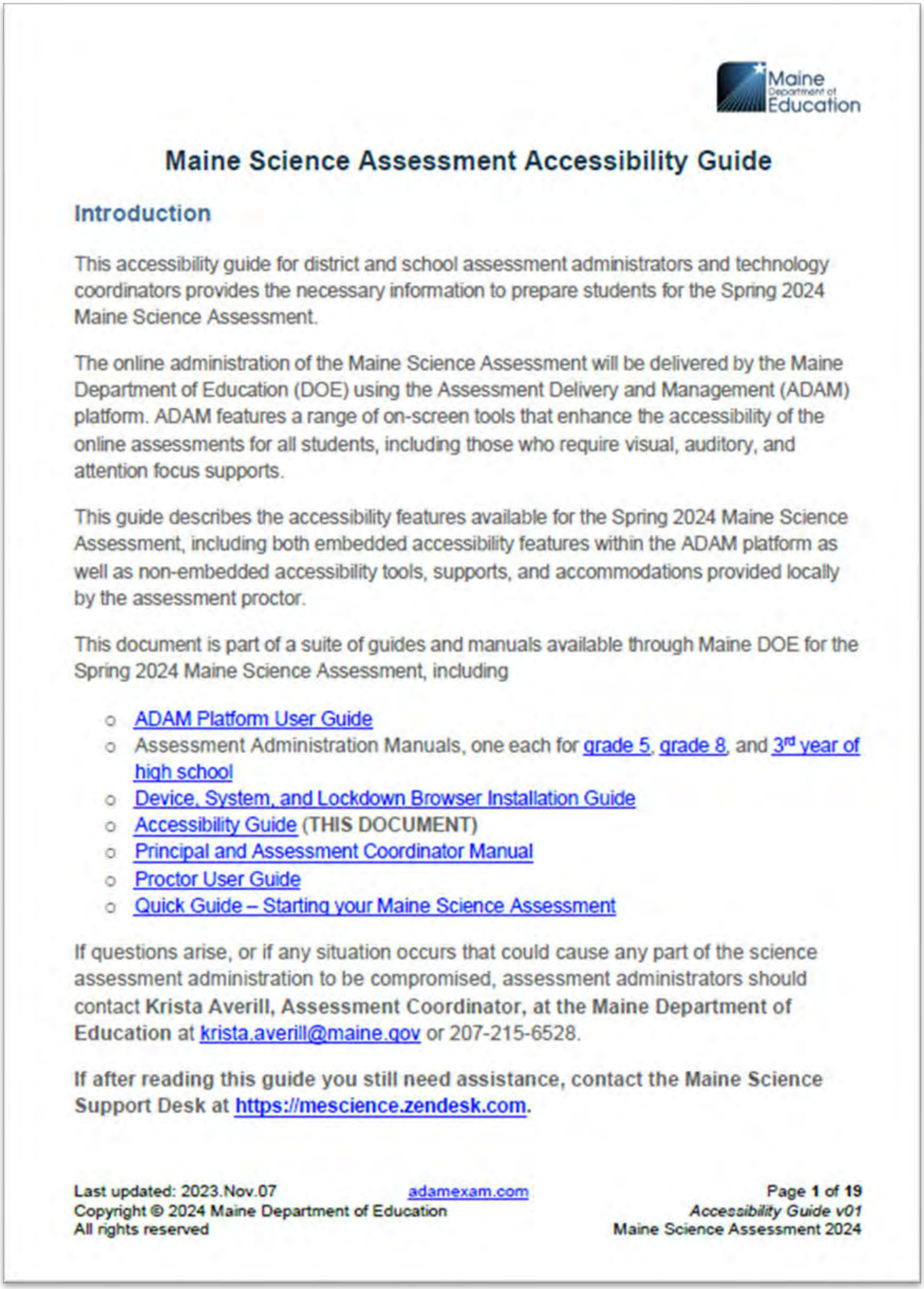


Table of Contents

Last updated: 2023.Nov.07 adamexam.com Page i of i
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



Maine Department of Education

Maine Science Assessment Accessibility Guide

Introduction

This accessibility guide for district and school assessment administrators and technology coordinators provides the necessary information to prepare students for the Spring 2024 Maine Science Assessment.

The online administration of the Maine Science Assessment will be delivered by the Maine Department of Education (DOE) using the Assessment Delivery and Management (ADAM) platform. ADAM features a range of on-screen tools that enhance the accessibility of the online assessments for all students, including those who require visual, auditory, and attention focus supports.

This guide describes the accessibility features available for the Spring 2024 Maine Science Assessment, including both embedded accessibility features within the ADAM platform as well as non-embedded accessibility tools, supports, and accommodations provided locally by the assessment proctor.

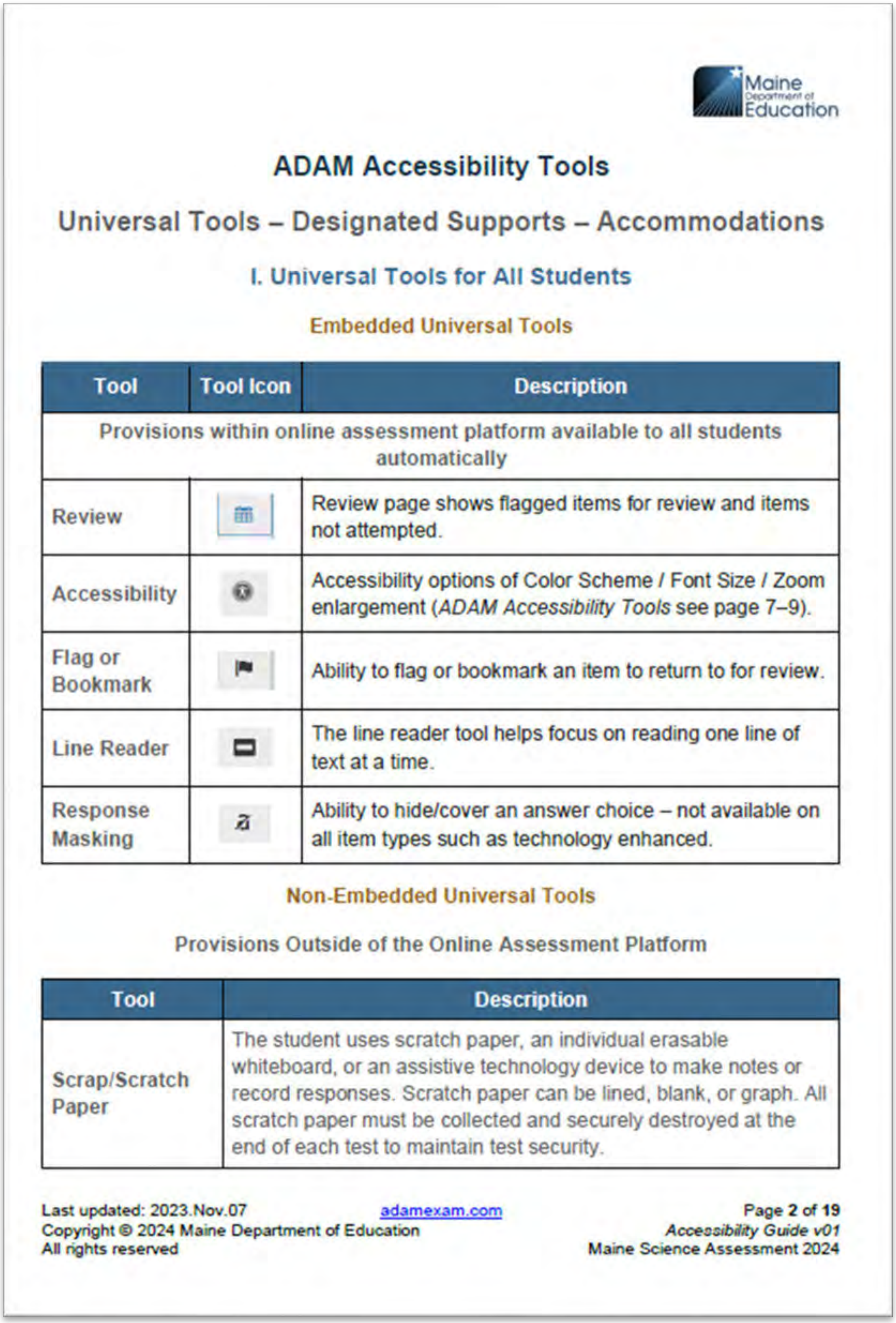
This document is part of a suite of guides and manuals available through Maine DOE for the Spring 2024 Maine Science Assessment, including

- ADAM Platform User Guide
- Assessment Administration Manuals, one each for grade 5, grade 8, and 3rd year of high school
- Device, System, and Lockdown Browser Installation Guide
- Accessibility Guide (THIS DOCUMENT)
- Principal and Assessment Coordinator Manual
- Proctor User Guide
- Quick Guide – Starting your Maine Science Assessment

If questions arise, or if any situation occurs that could cause any part of the science assessment administration to be compromised, assessment administrators should contact **Krista Averill, Assessment Coordinator, at the Maine Department of Education** at krista.averill@maine.gov or 207-215-6528.

If after reading this guide you still need assistance, contact the Maine Science Support Desk at https://mescience.zendesk.com.

Last updated: 2023.Nov.07 adamexam.com Page **1** of **19**
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



Maine Department of Education

ADAM Accessibility Tools

Universal Tools – Designated Supports – Accommodations

I. Universal Tools for All Students

Embedded Universal Tools

Tool	Tool Icon	Description
Provisions within online assessment platform available to all students automatically		
Review	▦	Review page shows flagged items for review and items not attempted.
Accessibility	●	Accessibility options of Color Scheme / Font Size / Zoom enlargement (*ADAM Accessibility Tools* see page 7–9).
Flag or Bookmark	⚑	Ability to flag or bookmark an item to return to for review.
Line Reader	▬	The line reader tool helps focus on reading one line of text at a time.
Response Masking	ā	Ability to hide/cover an answer choice – not available on all item types such as technology enhanced.

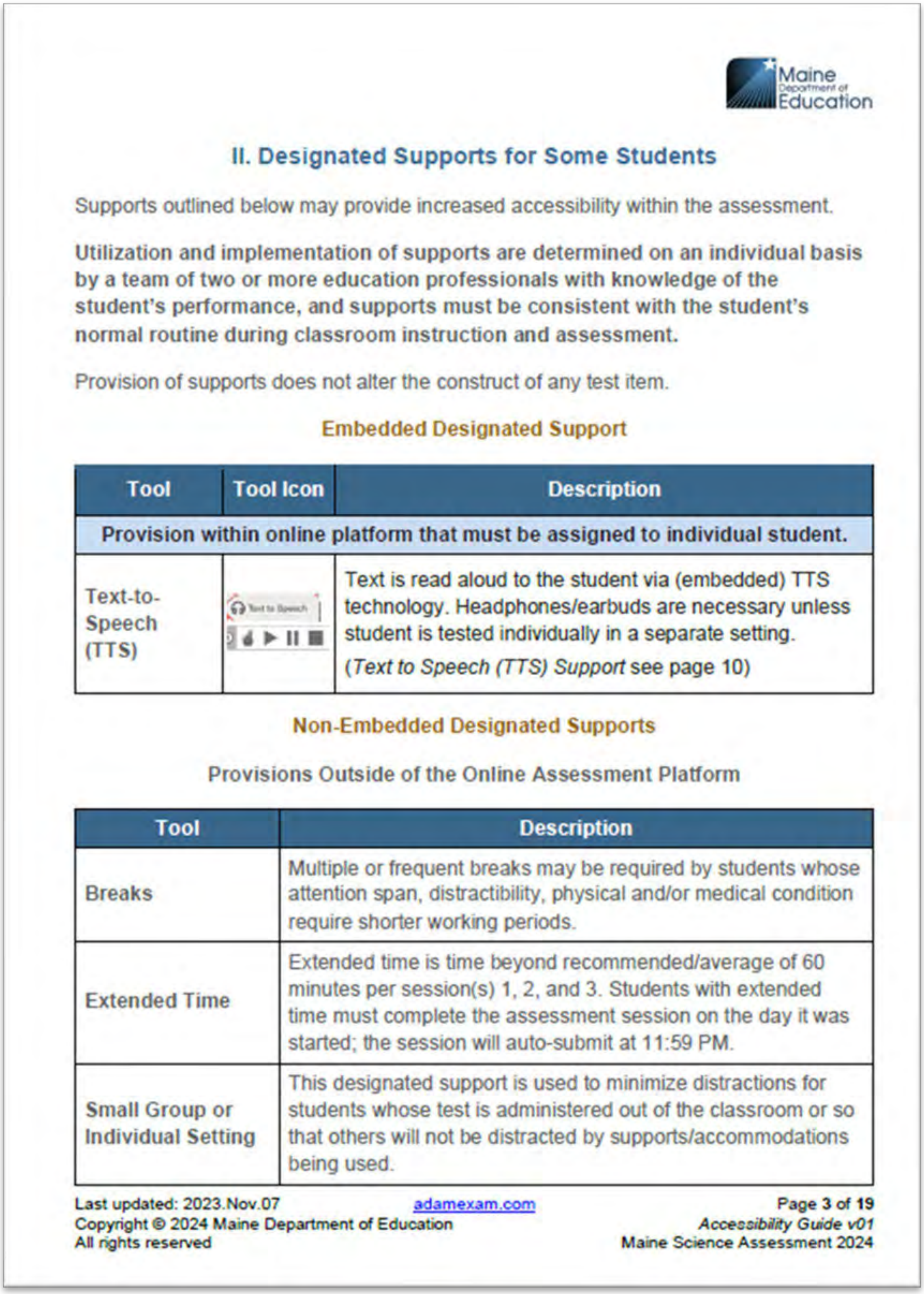
Non-Embedded Universal Tools

Provisions Outside of the Online Assessment Platform

Tool	Description
Scrap/Scratch Paper	The student uses scratch paper, an individual erasable whiteboard, or an assistive technology device to make notes or record responses. Scratch paper can be lined, blank, or graph. All scratch paper must be collected and securely destroyed at the end of each test to maintain test security.

Last updated: 2023.Nov.07 adamexam.com Page 2 of 19

Accessibility Guide v01
Maine Science Assessment 2024



II. Designated Supports for Some Students

Supports outlined below may provide increased accessibility within the assessment.

Utilization and implementation of supports are determined on an individual basis by a team of two or more education professionals with knowledge of the student's performance, and supports must be consistent with the student's normal routine during classroom instruction and assessment.

Provision of supports does not alter the construct of any test item.

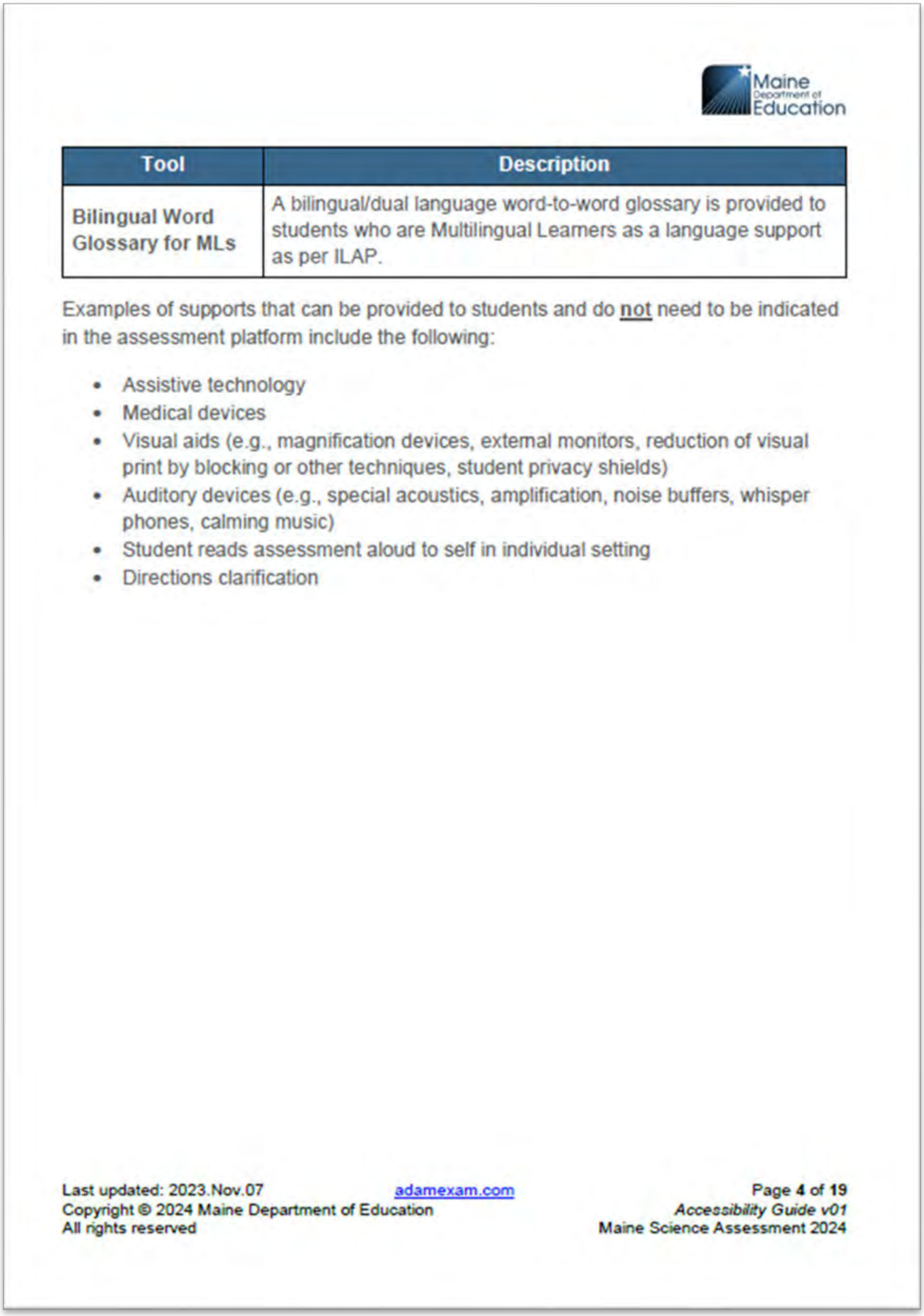
Embedded Designated Support

Tool	Tool Icon	Description
Provision within online platform that must be assigned to individual student.		
Text-to-Speech (TTS)	Text to Speech ▶ ⏸ ■	Text is read aloud to the student via (embedded) TTS technology. Headphones/earbuds are necessary unless student is tested individually in a separate setting. (*Text to Speech (TTS) Support* see page 10)

Non-Embedded Designated Supports

Provisions Outside of the Online Assessment Platform

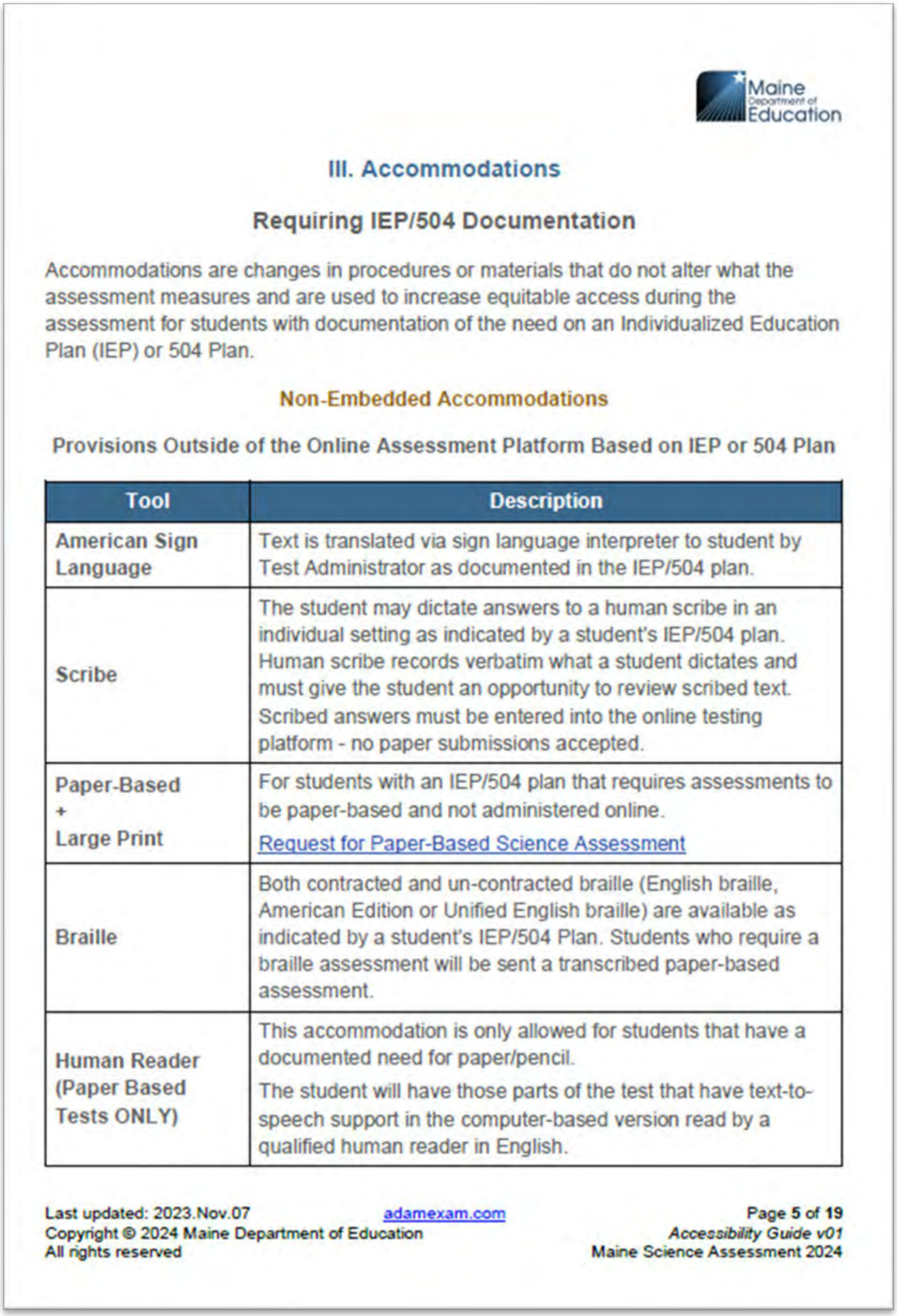
Tool	Description
Breaks	Multiple or frequent breaks may be required by students whose attention span, distractibility, physical and/or medical condition require shorter working periods.
Extended Time	Extended time is time beyond recommended/average of 60 minutes per session(s) 1, 2, and 3. Students with extended time must complete the assessment session on the day it was started; the session will auto-submit at 11:59 PM.
Small Group or Individual Setting	This designated support is used to minimize distractions for students whose test is administered out of the classroom or so that others will not be distracted by supports/accommodations being used.



Tool	Description
Bilingual Word Glossary for MLs	A bilingual/dual language word-to-word glossary is provided to students who are Multilingual Learners as a language support as per ILAP.

Examples of supports that can be provided to students and do not need to be indicated in the assessment platform include the following:

- Assistive technology
- Medical devices
- Visual aids (e.g., magnification devices, external monitors, reduction of visual print by blocking or other techniques, student privacy shields)
- Auditory devices (e.g., special acoustics, amplification, noise buffers, whisper phones, calming music)
- Student reads assessment aloud to self in individual setting
- Directions clarification



Maine Department of Education

III. Accommodations

Requiring IEP/504 Documentation

Accommodations are changes in procedures or materials that do not alter what the assessment measures and are used to increase equitable access during the assessment for students with documentation of the need on an Individualized Education Plan (IEP) or 504 Plan.

Non-Embedded Accommodations

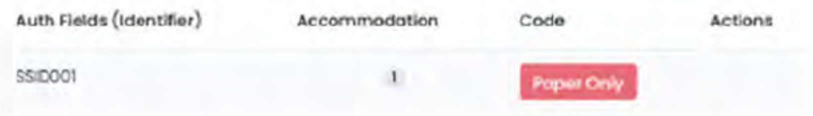
Provisions Outside of the Online Assessment Platform Based on IEP or 504 Plan

Tool	Description
American Sign Language	Text is translated via sign language interpreter to student by Test Administrator as documented in the IEP/504 plan.
Scribe	The student may dictate answers to a human scribe in an individual setting as indicated by a student's IEP/504 plan. Human scribe records verbatim what a student dictates and must give the student an opportunity to review scribed text. Scribed answers must be entered into the online testing platform - no paper submissions accepted.
Paper-Based **+** **Large Print**	For students with an IEP/504 plan that requires assessments to be paper-based and not administered online. Request for Paper-Based Science Assessment
Braille	Both contracted and un-contracted braille (English braille, American Edition or Unified English braille) are available as indicated by a student's IEP/504 Plan. Students who require a braille assessment will be sent a transcribed paper-based assessment.
Human Reader (Paper Based Tests ONLY)	This accommodation is only allowed for students that have a documented need for paper/pencil. The student will have those parts of the test that have text-to-speech support in the computer-based version read by a qualified human reader in English.

Administration of Paper-Based Forms

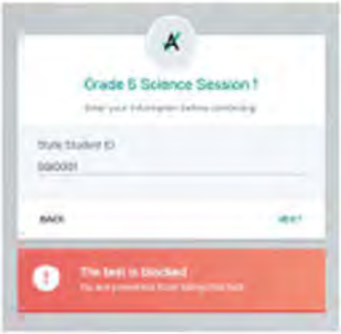
Students will complete their responses on the paper-based forms, and the school will return the paper test booklets to the Maine Science scoring vendor according to the procedures for handling paper testing materials. It is important to note that local test administrators and/or proctors will not enter student responses into an online form.

Students assigned a paper-based form will have the Paper-Based Form accommodation indicated in the test administration dashboards. They will also be indicated with a Paper Only code.



Auth Fields (Identifier)	Accommodation	Code	Actions
SSID001	1	Paper Only	

Should a student receiving a paper-based accommodation attempt to login to the online platform the following alert message will appear on the screen: *The test is blocked, you are prevented from taking this test.*



Grade 6 Science Session 1

Enter your information before continuing

State Student ID

SSID001

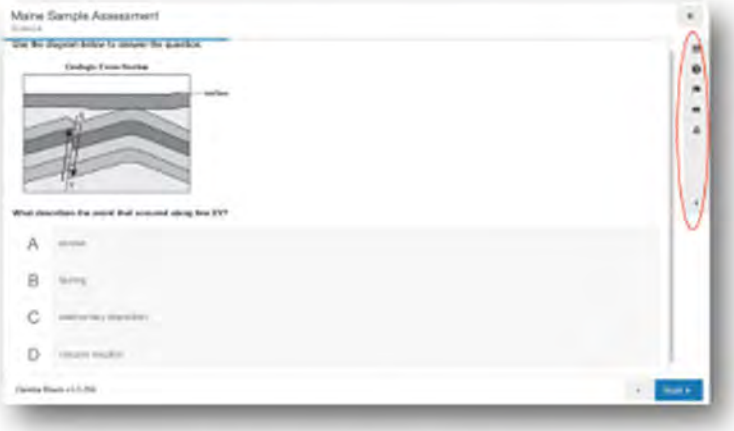
The test is blocked

You are prevented from taking this test

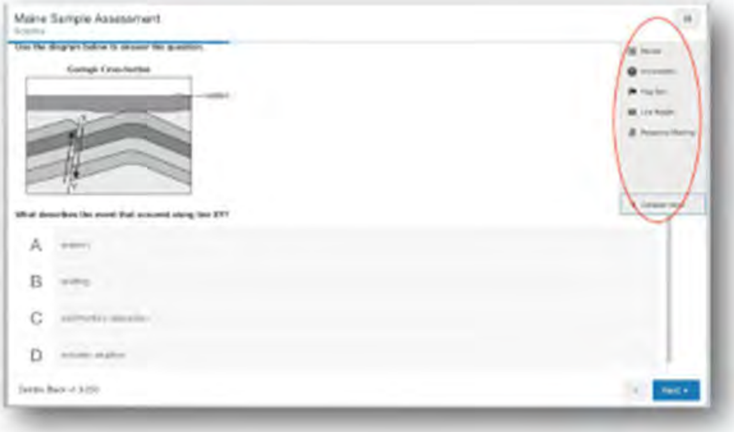
Online Accessibility Tools User Guide

ADAM Accessibility Tools

The accessibility tools menu is located on the right side of each assessment screen within the test supports toolbar. The student can access this menu at any time during the assessment session.



The accessibility tools menu can be expanded or collapsed by selecting the arrow icon at the toolbar's bottom.



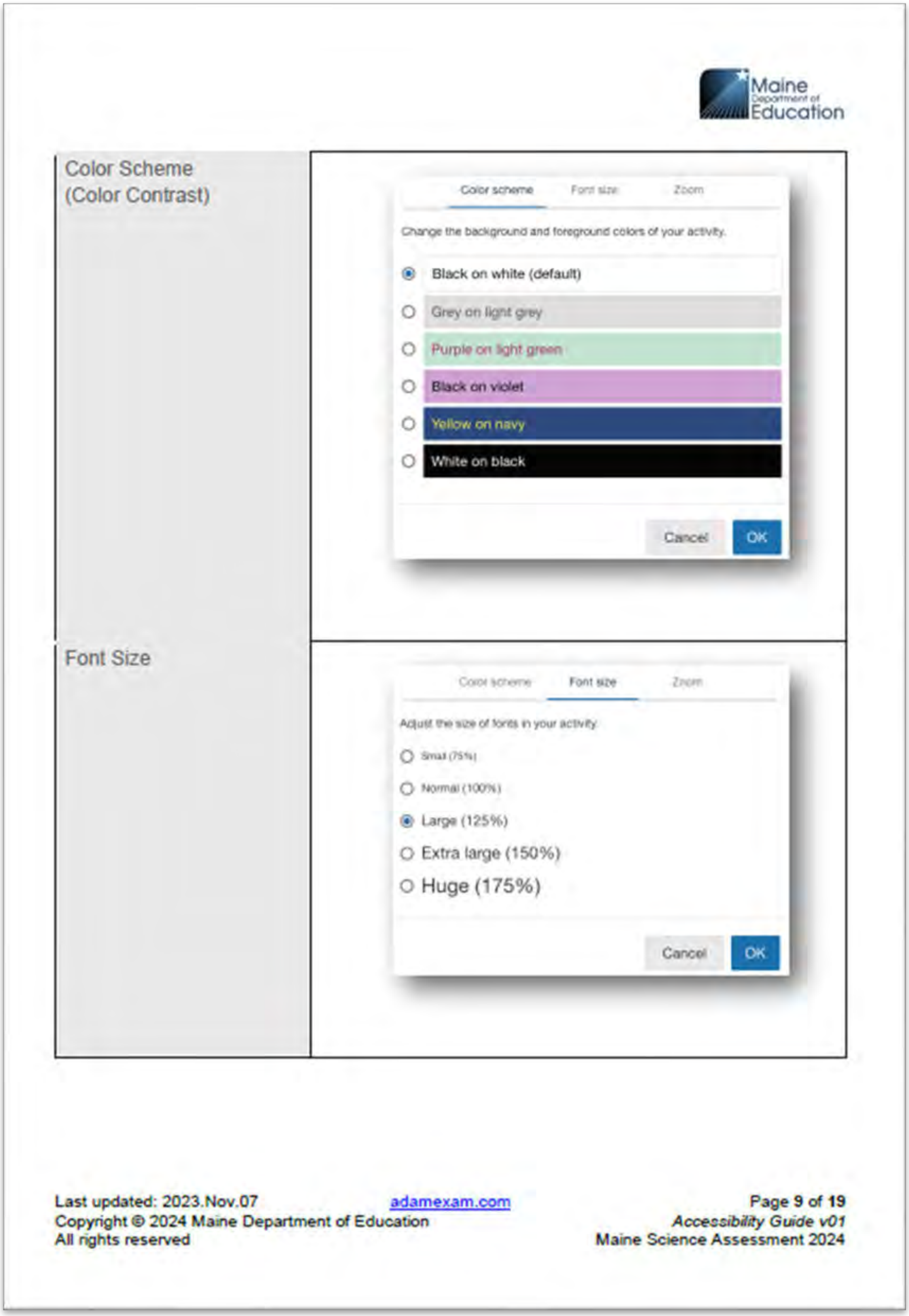
Accessibility Tools Menu

Students can set preferences for many of the tools that will persist from screen to screen (question to question) during an assessment session.

Students must reselect their preferred settings each time they log in to a new session or rejoin a session.

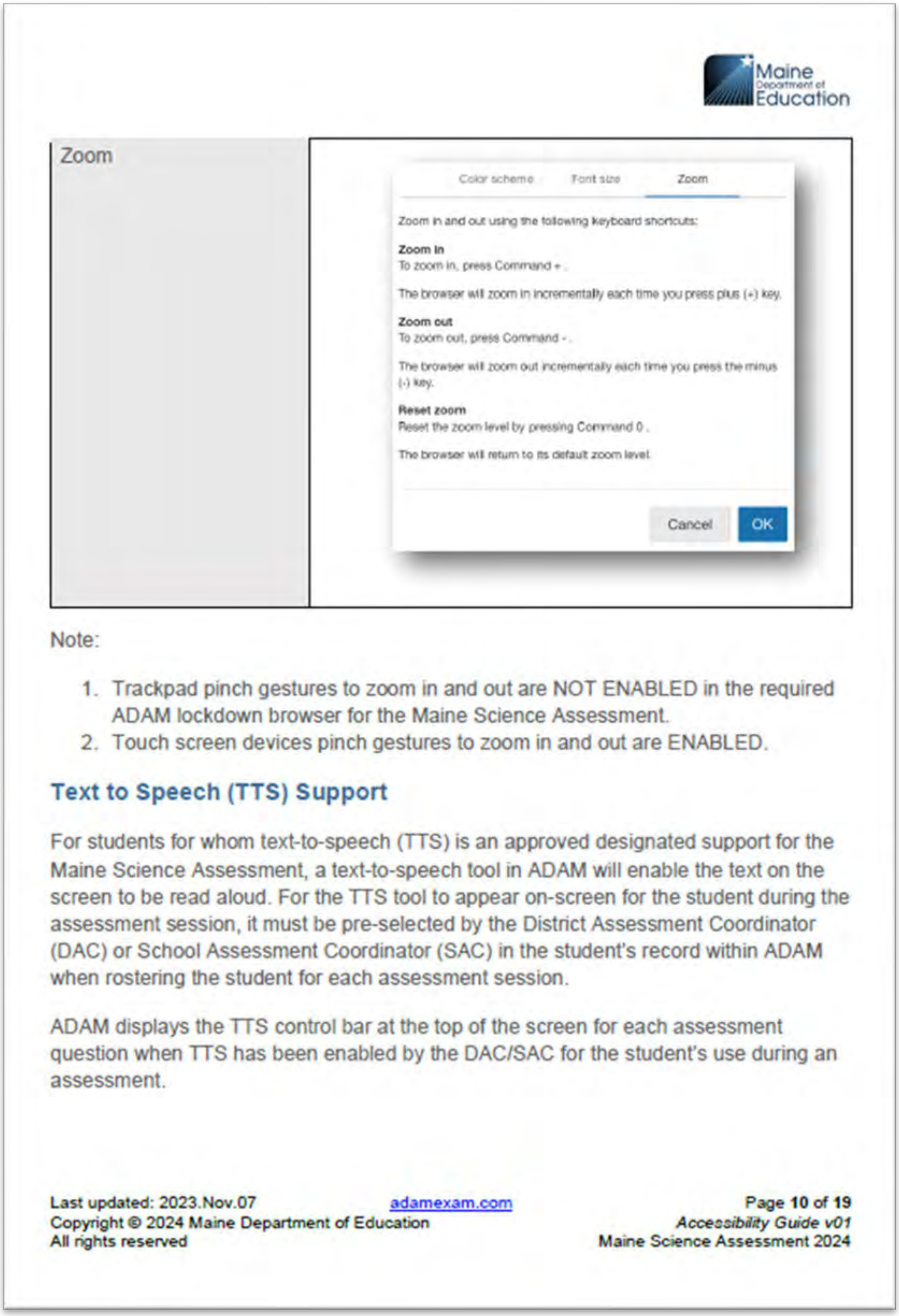
Accessibility	The accessibility menu provides options to • Change the color scheme (color contrast) of the background and text. • Change the font size. • View instructions for using the zoom functionality built into the web browser.

Last updated: 2023.Nov.07 adamexam.com Page 8 of 19
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



Color Scheme (Color Contrast)	Color scheme · Font size · Zoom Change the background and foreground colors of your activity. ● Black on white (default) ○ Grey on light grey ○ Purple on light green ○ Black on violet ○ Yellow on navy ○ White on black Cancel · OK
Font Size	Color scheme · Font size · Zoom Adjust the size of fonts in your activity. ○ Small (75%) ○ Normal (100%) ● Large (125%) ○ Extra large (150%) ○ Huge (175%) Cancel · OK

Last updated: 2023.Nov.07 adamexam.com Page 9 of 19
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



Zoom

Zoom in and out using the following keyboard shortcuts:

Zoom in
To zoom in, press Command + .

The browser will zoom in incrementally each time you press plus (+) key.

Zoom out
To zoom out, press Command - .

The browser will zoom out incrementally each time you press the minus (-) key.

Reset zoom
Reset the zoom level by pressing Command 0 .

The browser will return to its default zoom level.

Note:

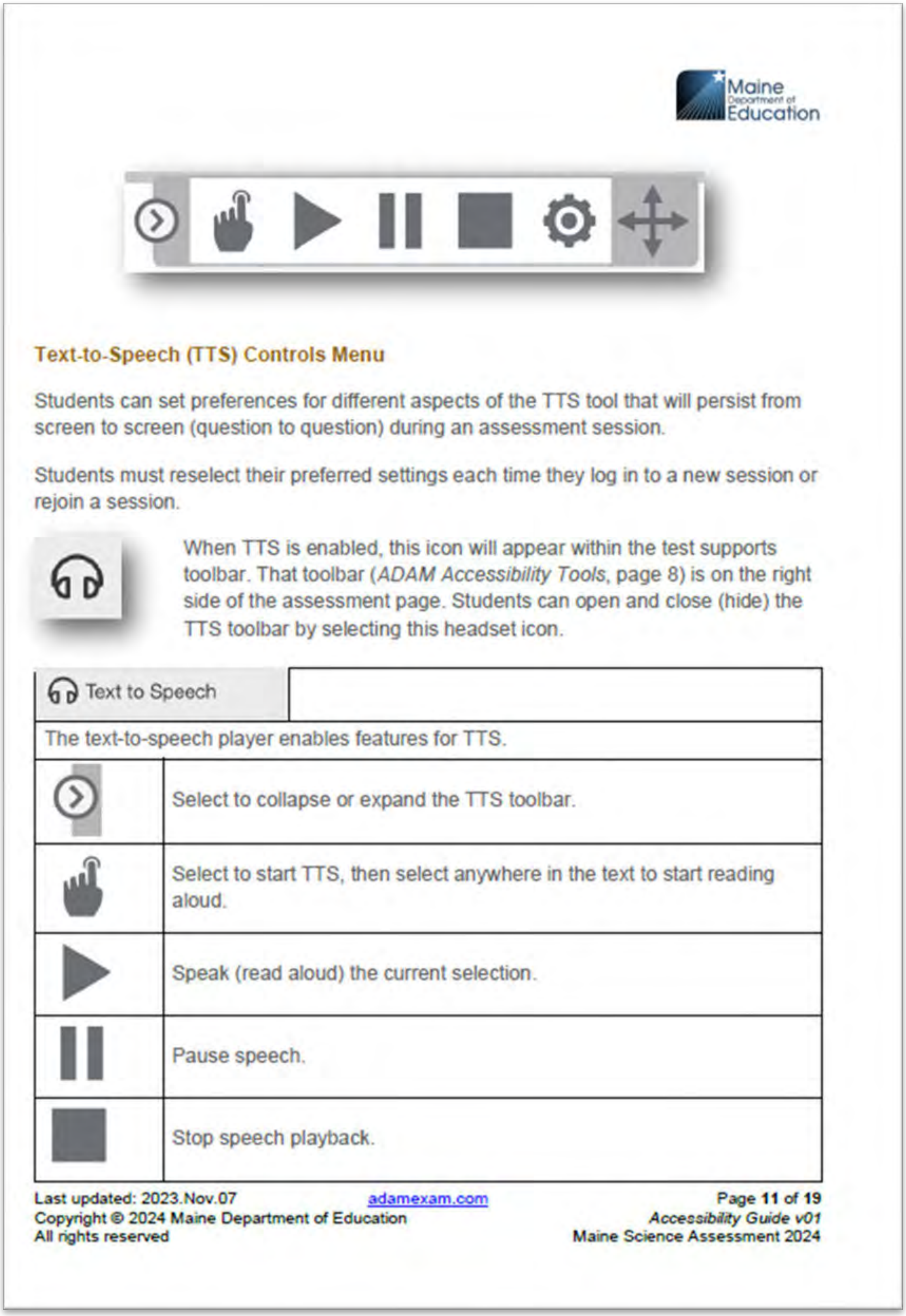
1. Trackpad pinch gestures to zoom in and out are NOT ENABLED in the required ADAM lockdown browser for the Maine Science Assessment.
2. Touch screen devices pinch gestures to zoom in and out are ENABLED.

Text to Speech (TTS) Support

For students for whom text-to-speech (TTS) is an approved designated support for the Maine Science Assessment, a text-to-speech tool in ADAM will enable the text on the screen to be read aloud. For the TTS tool to appear on-screen for the student during the assessment session, it must be pre-selected by the District Assessment Coordinator (DAC) or School Assessment Coordinator (SAC) in the student's record within ADAM when rostering the student for each assessment session.

ADAM displays the TTS control bar at the top of the screen for each assessment question when TTS has been enabled by the DAC/SAC for the student's use during an assessment.

Last updated: 2023.Nov.07 adamexam.com **Page 10 of 19**
Copyright © 2024 Maine Department of Education ***Accessibility Guide v01***
All rights reserved **Maine Science Assessment 2024**



Maine Department of Education

Text-to-Speech (TTS) Controls Menu

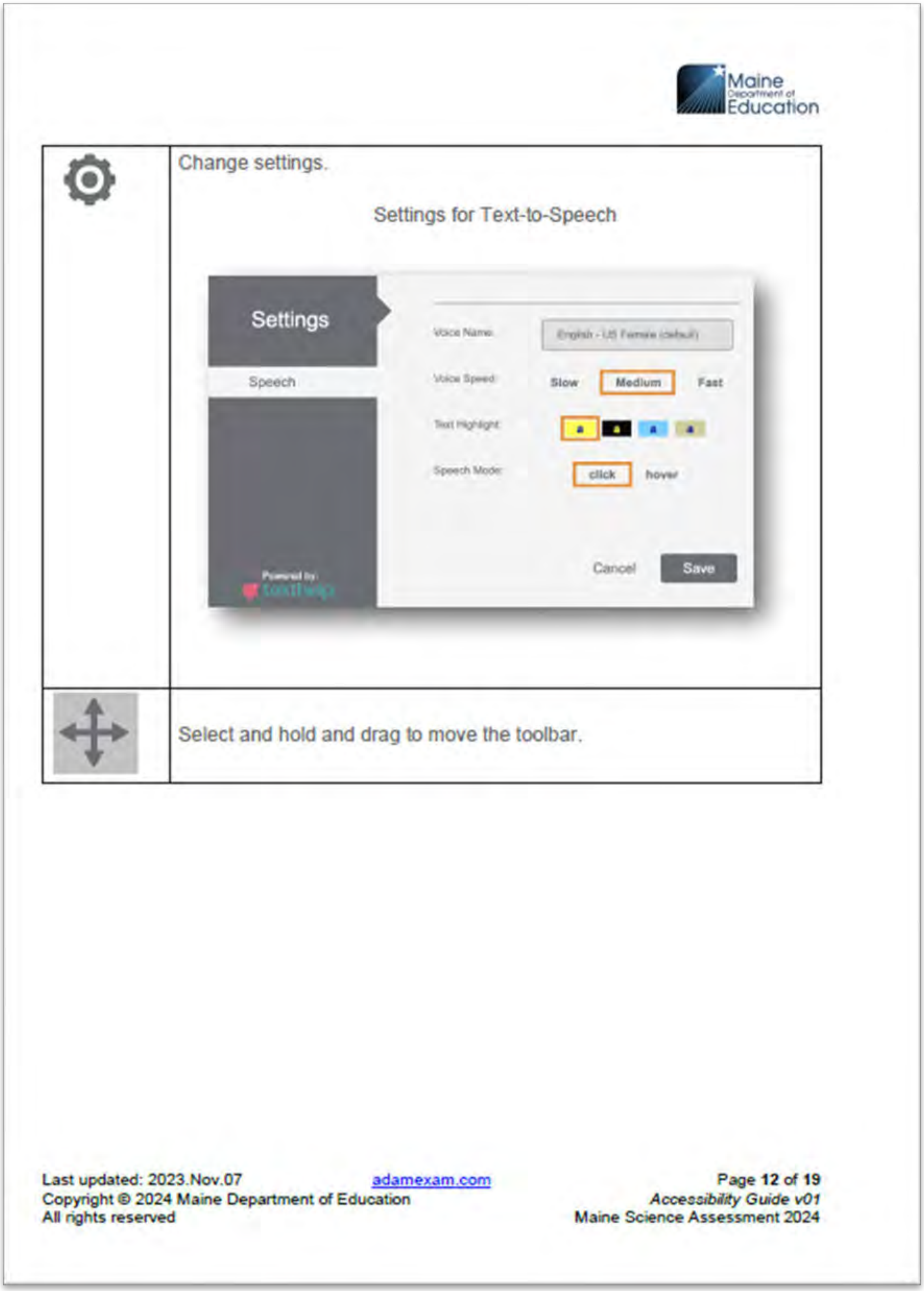
Students can set preferences for different aspects of the TTS tool that will persist from screen to screen (question to question) during an assessment session.

Students must reselect their preferred settings each time they log in to a new session or rejoin a session.

🎧 When TTS is enabled, this icon will appear within the test supports toolbar. That toolbar (*ADAM Accessibility Tools*, page 8) is on the right side of the assessment page. Students can open and close (hide) the TTS toolbar by selecting this headset icon.

🎧 Text to Speech	
The text-to-speech player enables features for TTS.	
›	Select to collapse or expand the TTS toolbar.
☝	Select to start TTS, then select anywhere in the text to start reading aloud.
▶	Speak (read aloud) the current selection.
❚❚	Pause speech.
■	Stop speech playback.

Last updated: 2023.Nov.07 adamexam.com Page 11 of 19
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024

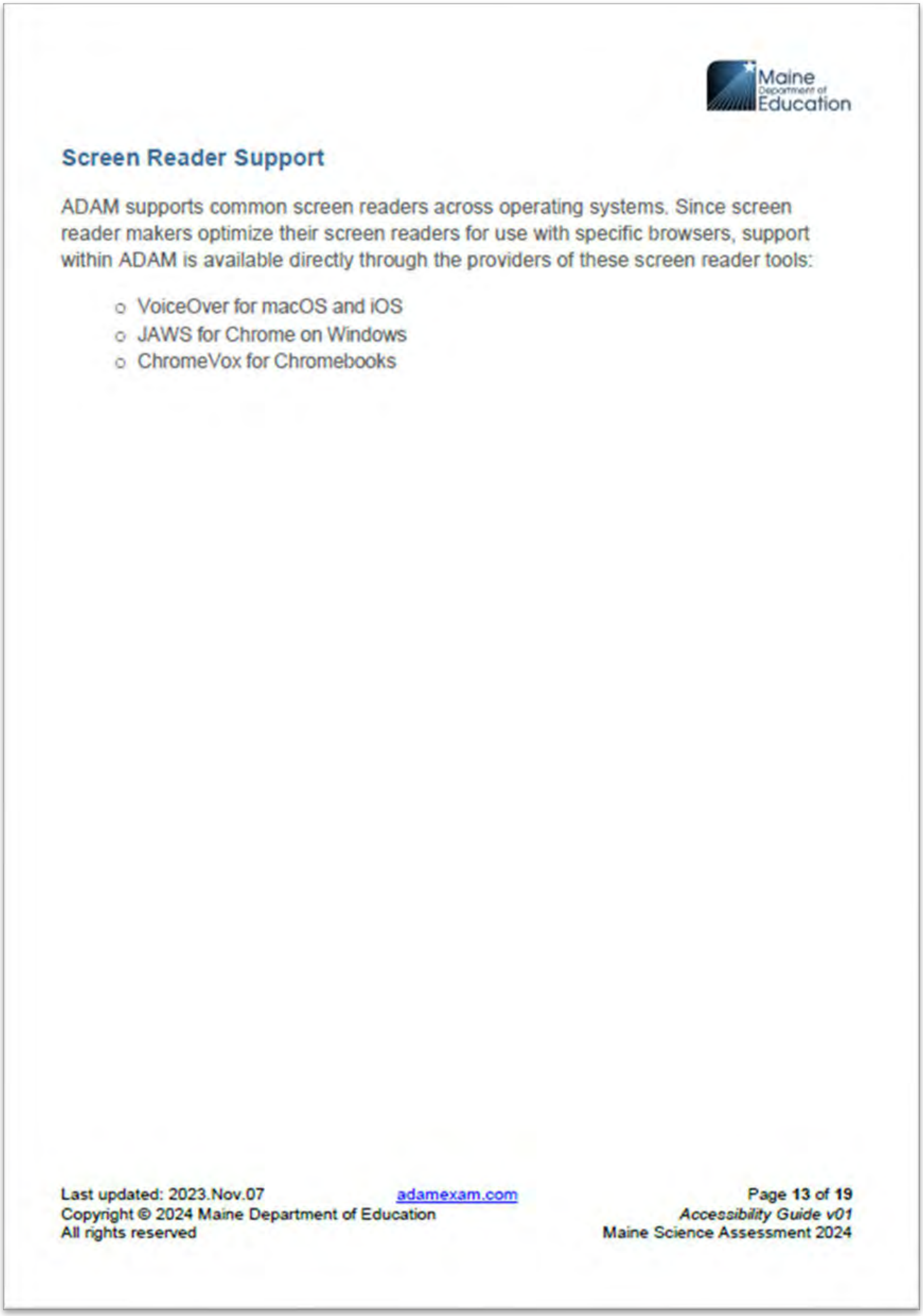


Maine Department of Education

⚙	Change settings. Settings for Text-to-Speech Settings Speech Voice Name: English - US Female (default) Voice Speed: Slow [✓ Medium] Fast Text Highlight: Speech Mode: [✓ click] hover Cancel Save Powered by texthelp
✥	Select and hold and drag to move the toolbar.

Last updated: 2023.Nov.07 adamexam.com
Copyright © 2024 Maine Department of Education
All rights reserved

Page **12** of **19**
Accessibility Guide v01
Maine Science Assessment 2024



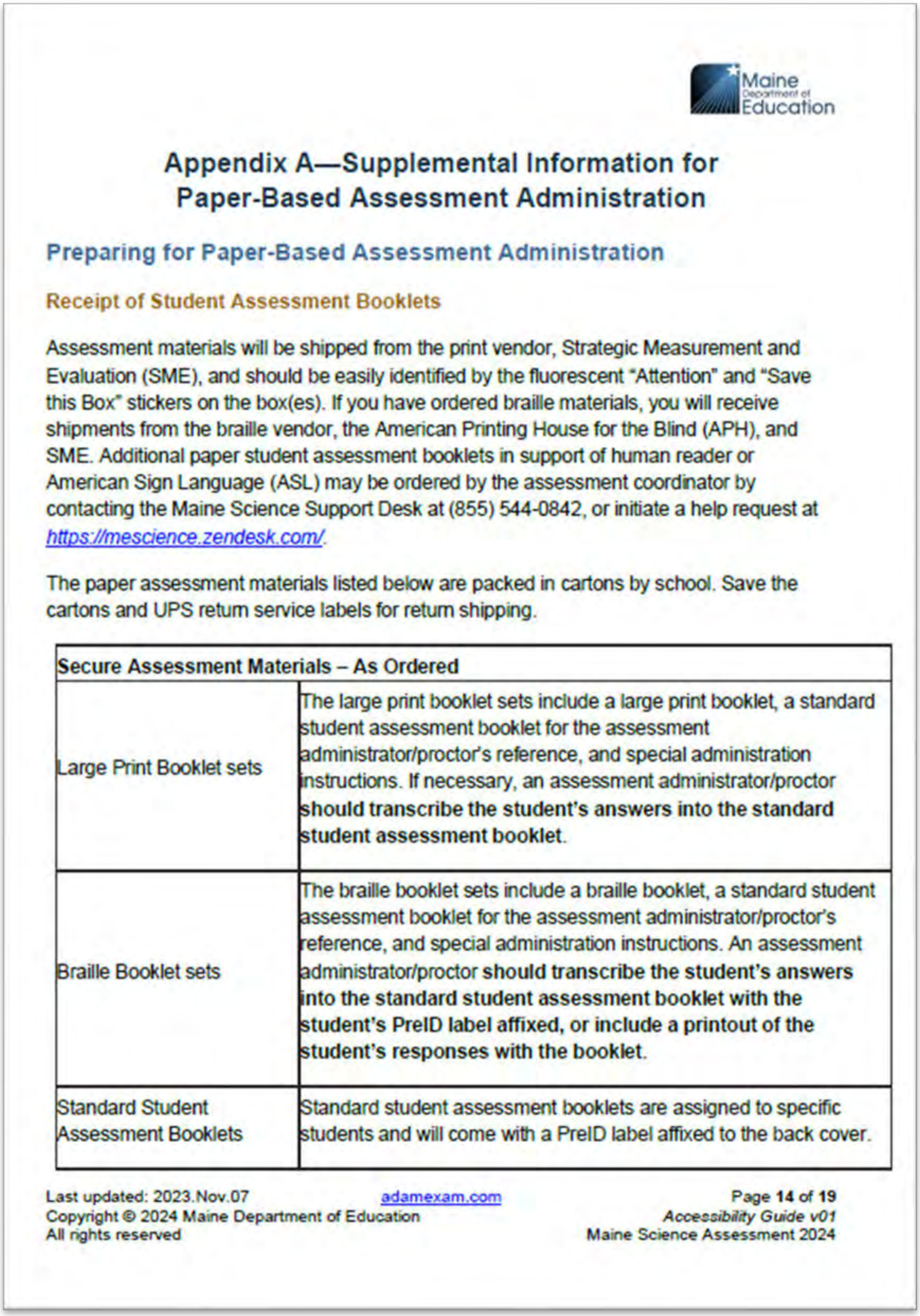
Maine Department of Education

Screen Reader Support

ADAM supports common screen readers across operating systems. Since screen reader makers optimize their screen readers for use with specific browsers, support within ADAM is available directly through the providers of these screen reader tools:

- VoiceOver for macOS and iOS
- JAWS for Chrome on Windows
- ChromeVox for Chromebooks

Last updated: 2023.Nov.07 adamexam.com Page 13 of 19
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



Appendix A—Supplemental Information for Paper-Based Assessment Administration

Preparing for Paper-Based Assessment Administration

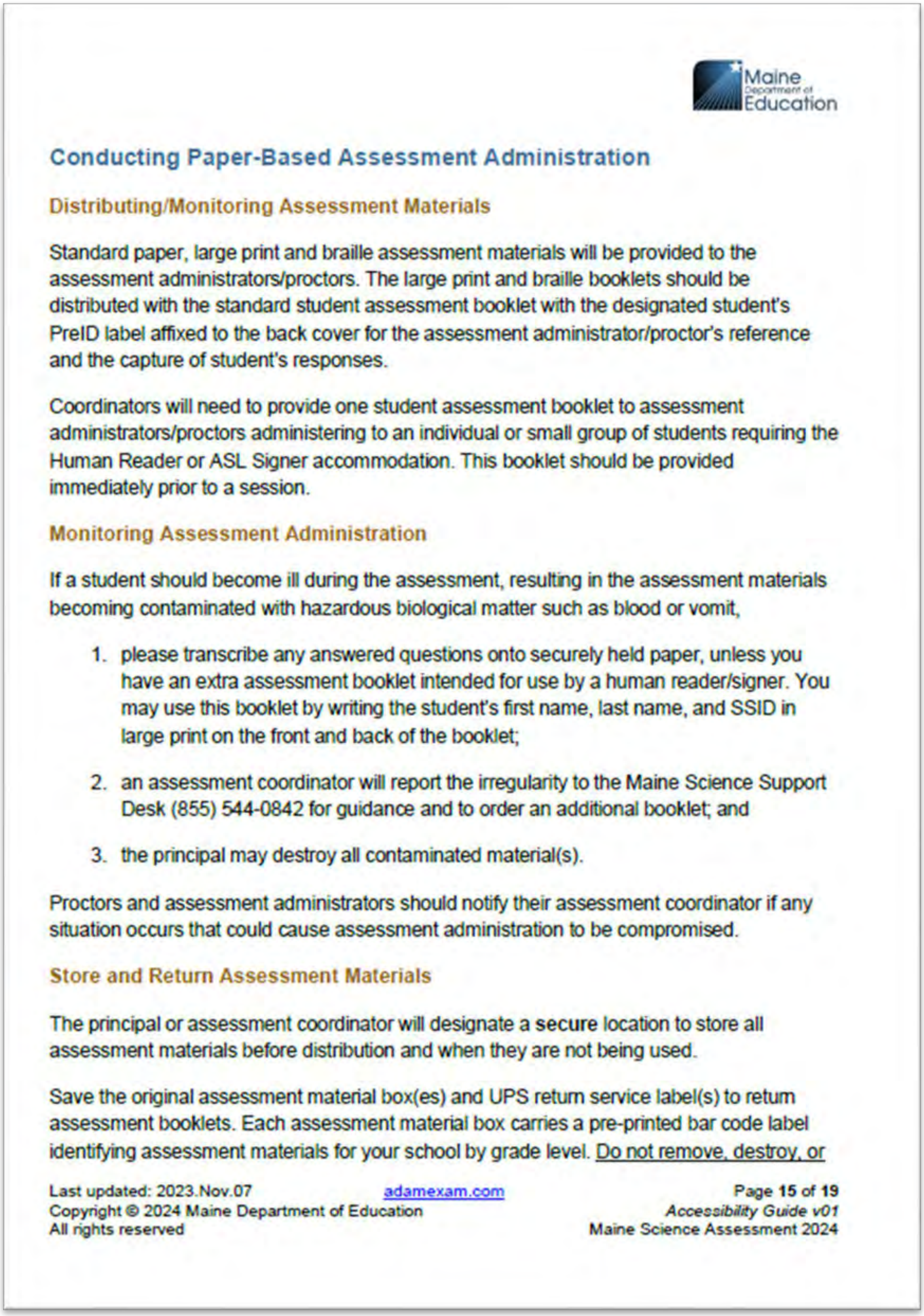
Receipt of Student Assessment Booklets

Assessment materials will be shipped from the print vendor, Strategic Measurement and Evaluation (SME), and should be easily identified by the fluorescent "Attention" and "Save this Box" stickers on the box(es). If you have ordered braille materials, you will receive shipments from the braille vendor, the American Printing House for the Blind (APH), and SME. Additional paper student assessment booklets in support of human reader or American Sign Language (ASL) may be ordered by the assessment coordinator by contacting the Maine Science Support Desk at (855) 544-0842, or initiate a help request at https://mescience.zendesk.com/.

The paper assessment materials listed below are packed in cartons by school. Save the cartons and UPS return service labels for return shipping.

Secure Assessment Materials – As Ordered	
Large Print Booklet sets	The large print booklet sets include a large print booklet, a standard student assessment booklet for the assessment administrator/proctor's reference, and special administration instructions. If necessary, an assessment administrator/proctor **should transcribe the student's answers into the standard student assessment booklet**.
Braille Booklet sets	The braille booklet sets include a braille booklet, a standard student assessment booklet for the assessment administrator/proctor's reference, and special administration instructions. An assessment administrator/proctor **should transcribe the student's answers into the standard student assessment booklet with the student's PreID label affixed, or include a printout of the student's responses with the booklet**.
Standard Student Assessment Booklets	Standard student assessment booklets are assigned to specific students and will come with a PreID label affixed to the back cover.

Last updated: 2023.Nov.07 adamexam.com Page 14 of 19
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



Conducting Paper-Based Assessment Administration

Distributing/Monitoring Assessment Materials

Standard paper, large print and braille assessment materials will be provided to the assessment administrators/proctors. The large print and braille booklets should be distributed with the standard student assessment booklet with the designated student's PreID label affixed to the back cover for the assessment administrator/proctor's reference and the capture of student's responses.

Coordinators will need to provide one student assessment booklet to assessment administrators/proctors administering to an individual or small group of students requiring the Human Reader or ASL Signer accommodation. This booklet should be provided immediately prior to a session.

Monitoring Assessment Administration

If a student should become ill during the assessment, resulting in the assessment materials becoming contaminated with hazardous biological matter such as blood or vomit,

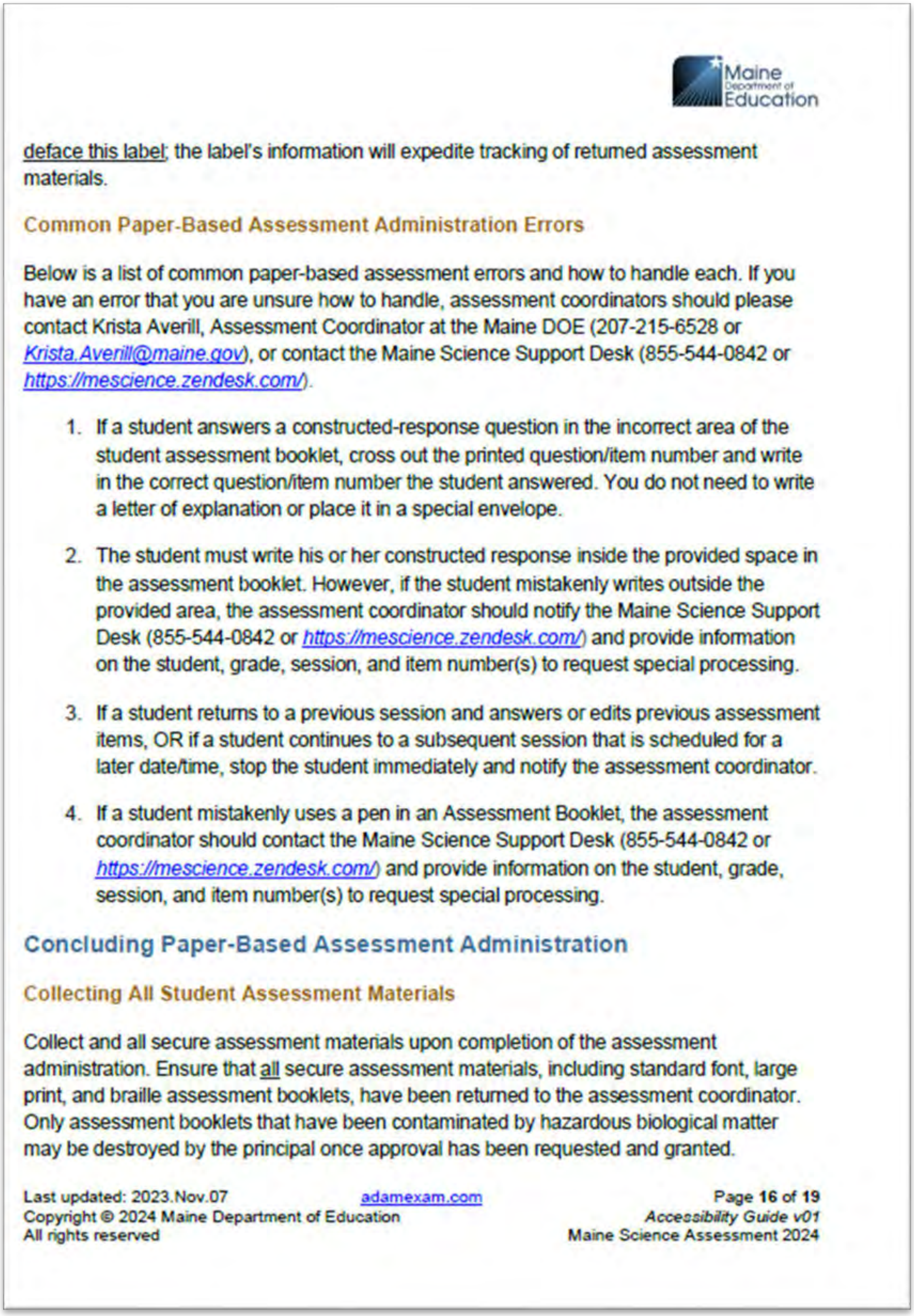
1. please transcribe any answered questions onto securely held paper, unless you have an extra assessment booklet intended for use by a human reader/signer. You may use this booklet by writing the student's first name, last name, and SSID in large print on the front and back of the booklet;
2. an assessment coordinator will report the irregularity to the Maine Science Support Desk (855) 544-0842 for guidance and to order an additional booklet; and
3. the principal may destroy all contaminated material(s).

Proctors and assessment administrators should notify their assessment coordinator if any situation occurs that could cause assessment administration to be compromised.

Store and Return Assessment Materials

The principal or assessment coordinator will designate a **secure** location to store all assessment materials before distribution and when they are not being used.

Save the original assessment material box(es) and UPS return service label(s) to return assessment booklets. Each assessment material box carries a pre-printed bar code label identifying assessment materials for your school by grade level. Do not remove, destroy, or



deface this label; the label's information will expedite tracking of returned assessment materials.

Common Paper-Based Assessment Administration Errors

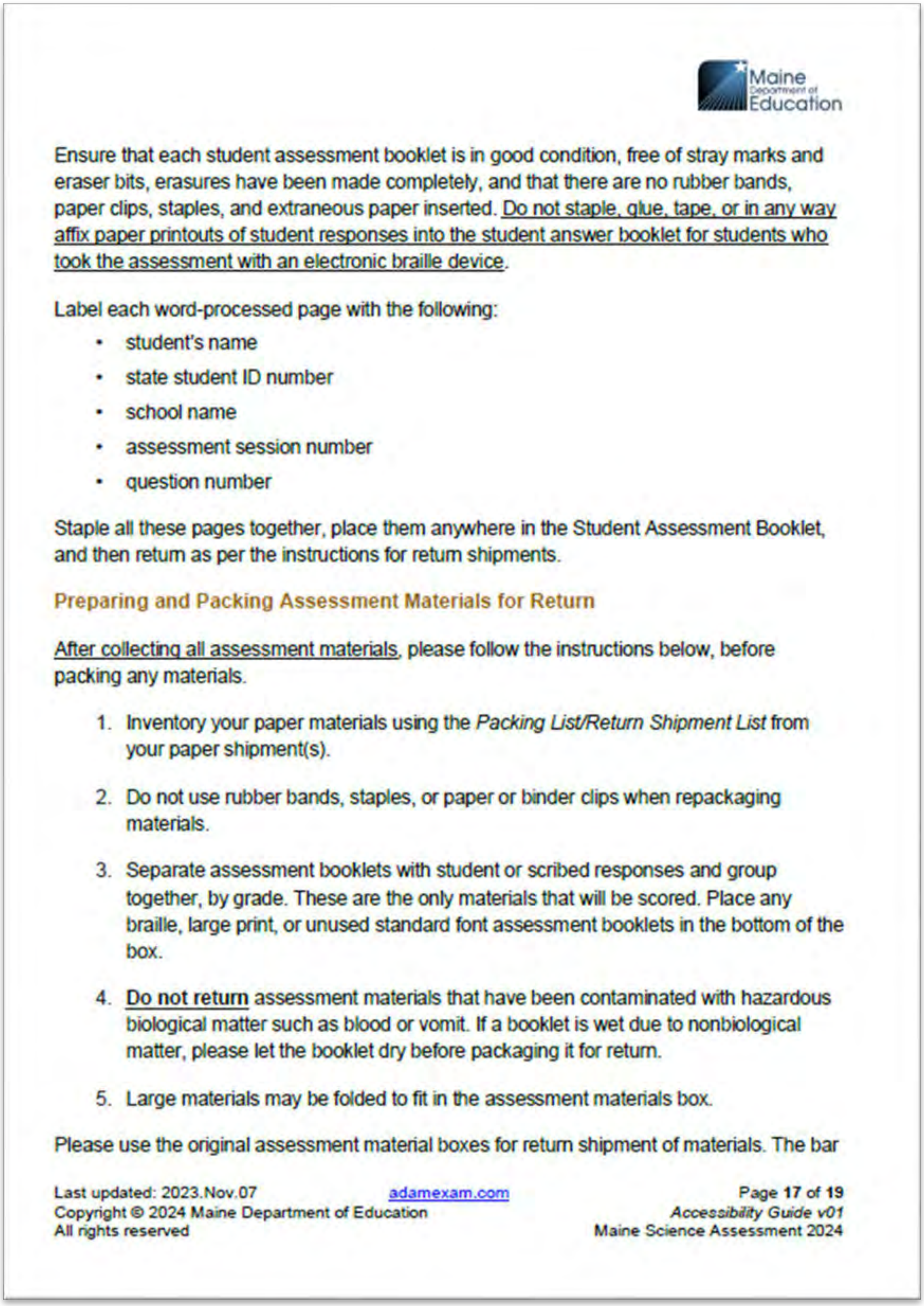
Below is a list of common paper-based assessment errors and how to handle each. If you have an error that you are unsure how to handle, assessment coordinators should please contact Krista Averill, Assessment Coordinator at the Maine DOE (207-215-6528 or Krista.Averill@maine.gov), or contact the Maine Science Support Desk (855-544-0842 or https://mescience.zendesk.com/).

1. If a student answers a constructed-response question in the incorrect area of the student assessment booklet, cross out the printed question/item number and write in the correct question/item number the student answered. You do not need to write a letter of explanation or place it in a special envelope.
2. The student must write his or her constructed response inside the provided space in the assessment booklet. However, if the student mistakenly writes outside the provided area, the assessment coordinator should notify the Maine Science Support Desk (855-544-0842 or https://mescience.zendesk.com/) and provide information on the student, grade, session, and item number(s) to request special processing.
3. If a student returns to a previous session and answers or edits previous assessment items, OR if a student continues to a subsequent session that is scheduled for a later date/time, stop the student immediately and notify the assessment coordinator.
4. If a student mistakenly uses a pen in an Assessment Booklet, the assessment coordinator should contact the Maine Science Support Desk (855-544-0842 or https://mescience.zendesk.com/) and provide information on the student, grade, session, and item number(s) to request special processing.

Concluding Paper-Based Assessment Administration

Collecting All Student Assessment Materials

Collect and all secure assessment materials upon completion of the assessment administration. Ensure that all secure assessment materials, including standard font, large print, and braille assessment booklets, have been returned to the assessment coordinator. Only assessment booklets that have been contaminated by hazardous biological matter may be destroyed by the principal once approval has been requested and granted.



Ensure that each student assessment booklet is in good condition, free of stray marks and eraser bits, erasures have been made completely, and that there are no rubber bands, paper clips, staples, and extraneous paper inserted. Do not staple, glue, tape, or in any way affix paper printouts of student responses into the student answer booklet for students who took the assessment with an electronic braille device.

Label each word-processed page with the following:

- student's name
- state student ID number
- school name
- assessment session number
- question number

Staple all these pages together, place them anywhere in the Student Assessment Booklet, and then return as per the instructions for return shipments.

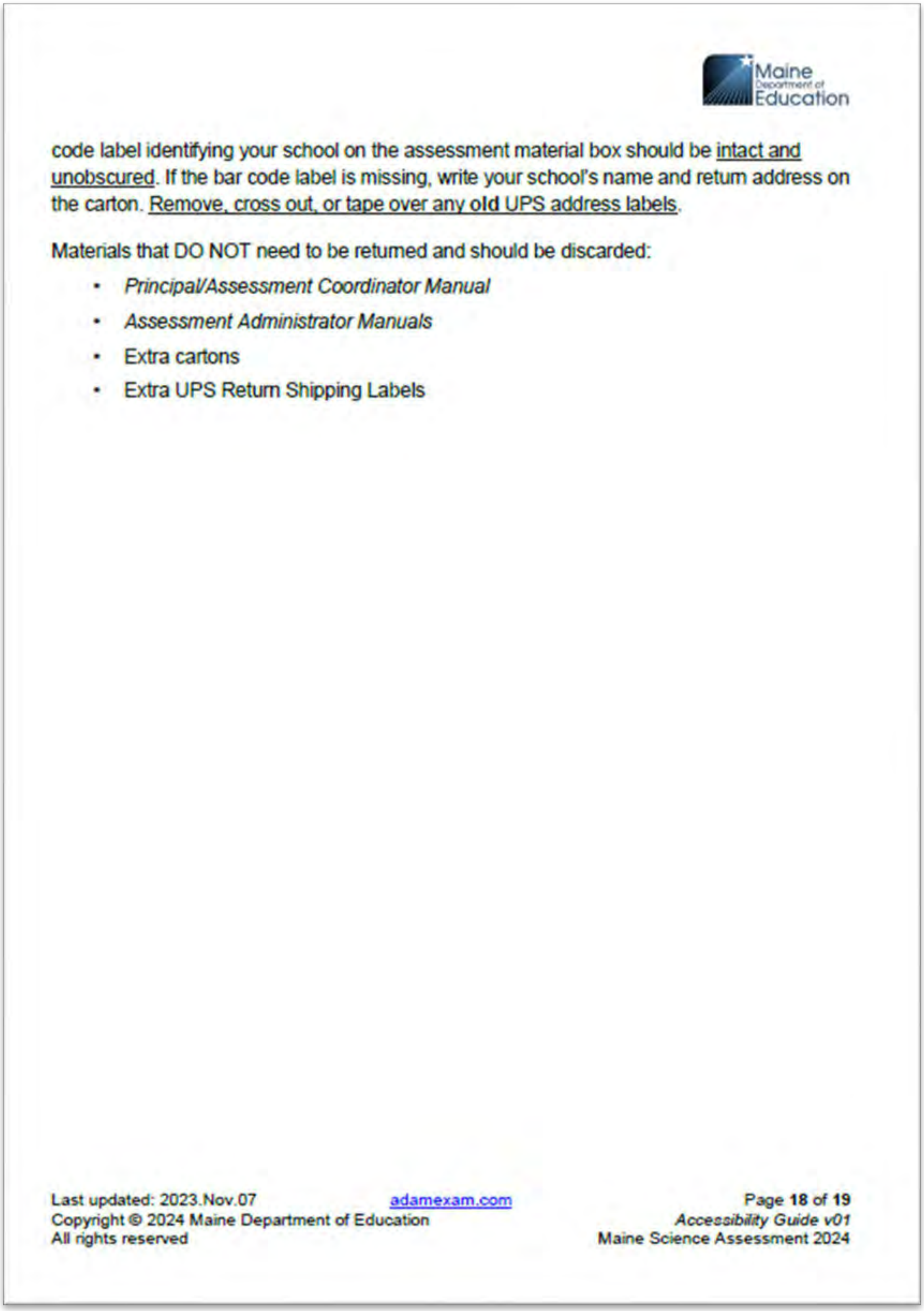
Preparing and Packing Assessment Materials for Return

After collecting all assessment materials, please follow the instructions below, before packing any materials.

1. Inventory your paper materials using the *Packing List/Return Shipment List* from your paper shipment(s).
2. Do not use rubber bands, staples, or paper or binder clips when repackaging materials.
3. Separate assessment booklets with student or scribed responses and group together, by grade. These are the only materials that will be scored. Place any braille, large print, or unused standard font assessment booklets in the bottom of the box.
4. **Do not return** assessment materials that have been contaminated with hazardous biological matter such as blood or vomit. If a booklet is wet due to nonbiological matter, please let the booklet dry before packaging it for return.
5. Large materials may be folded to fit in the assessment materials box.

Please use the original assessment material boxes for return shipment of materials. The bar

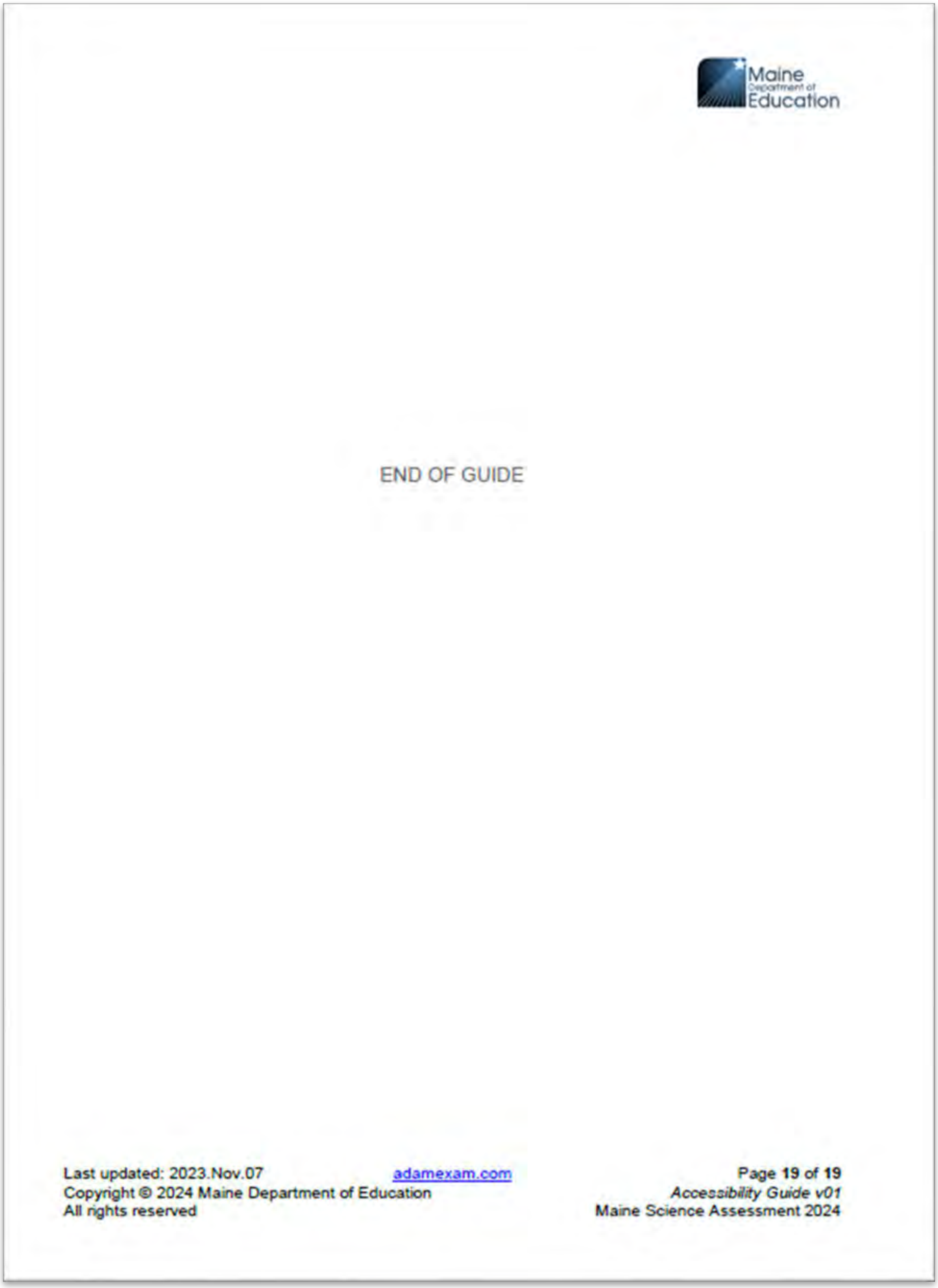
Last updated: 2023.Nov.07 adamexam.com Page **17** of **19**
Copyright © 2024 Maine Department of Education *Accessibility Guide v01*
All rights reserved Maine Science Assessment 2024



code label identifying your school on the assessment material box should be intact and unobscured. If the bar code label is missing, write your school's name and return address on the carton. Remove, cross out, or tape over any **old** UPS address labels.

Materials that DO NOT need to be returned and should be discarded:

- *Principal/Assessment Coordinator Manual*
- *Assessment Administrator Manuals*
- Extra cartons
- Extra UPS Return Shipping Labels



Maine Department of Education

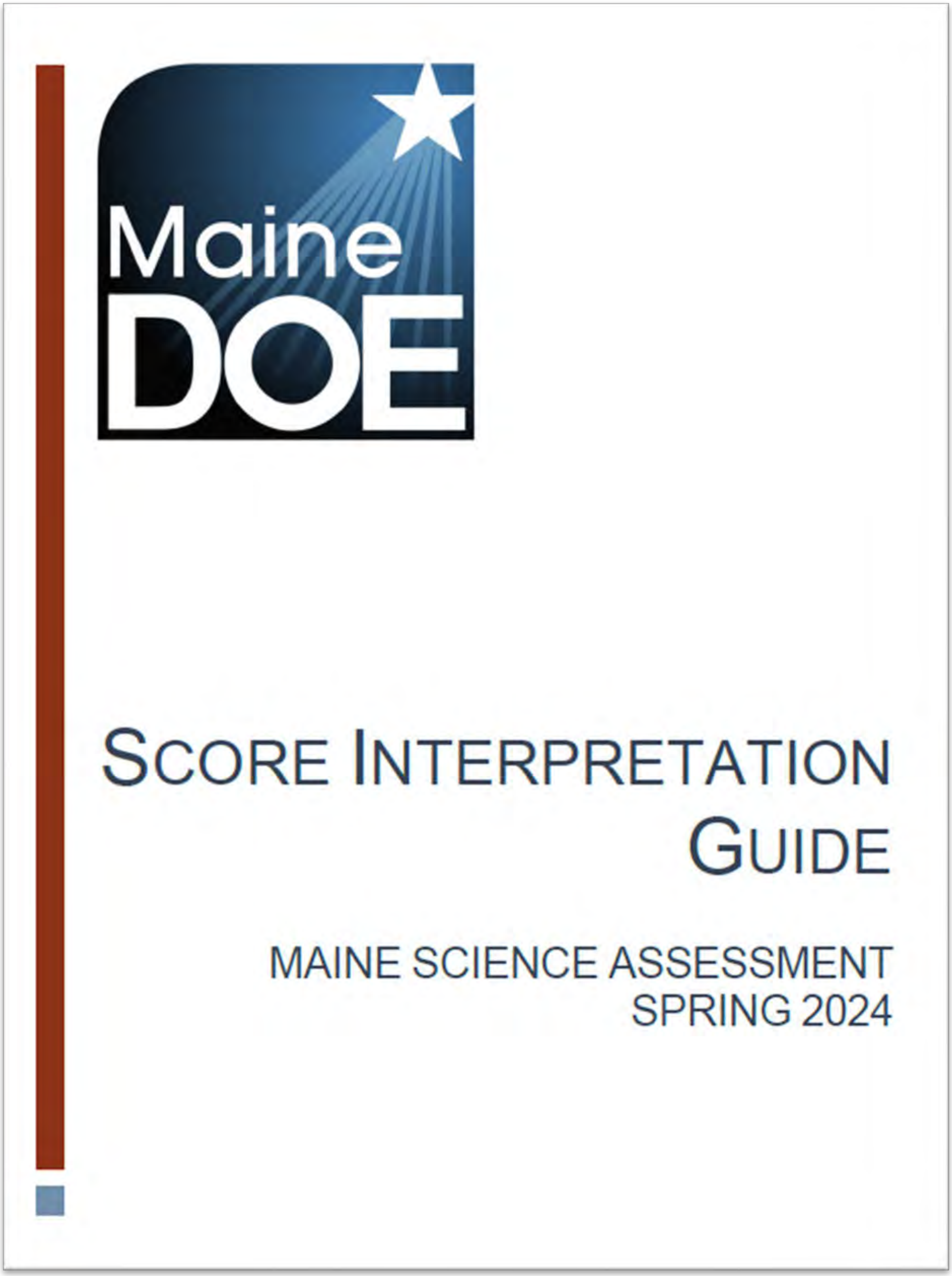
END OF GUIDE

Last updated: 2023.Nov.07 adamexam.com
Copyright © 2024 Maine Department of Education
All rights reserved

Page 19 of 19
Accessibility Guide v01
Maine Science Assessment 2024

Score Interpretation Guide

(Document begins on next page.)



Maine DOE

SCORE INTERPRETATION GUIDE

MAINE SCIENCE ASSESSMENT
SPRING 2024

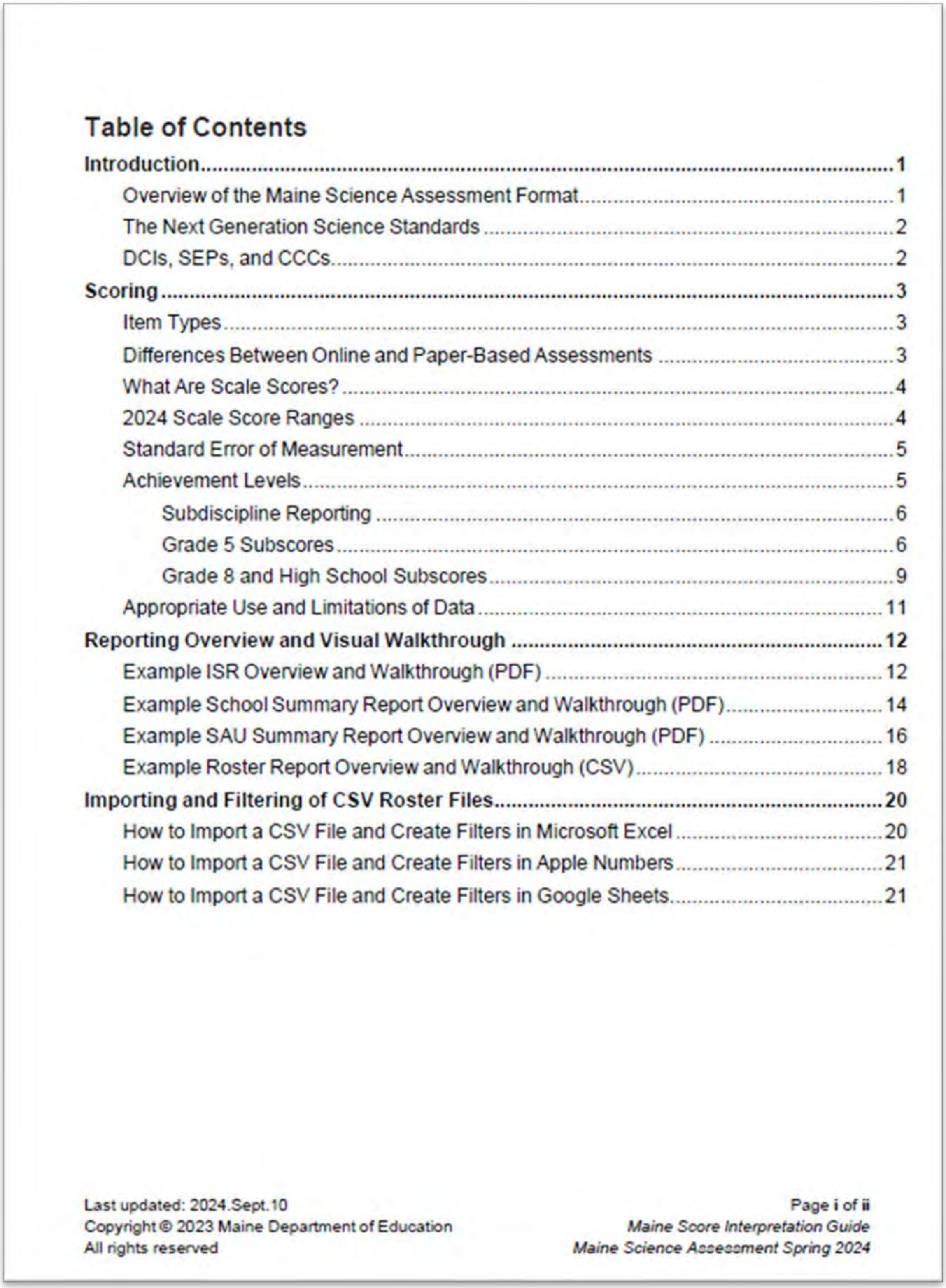


Table of Contents

Last updated: 2024.Sept.10

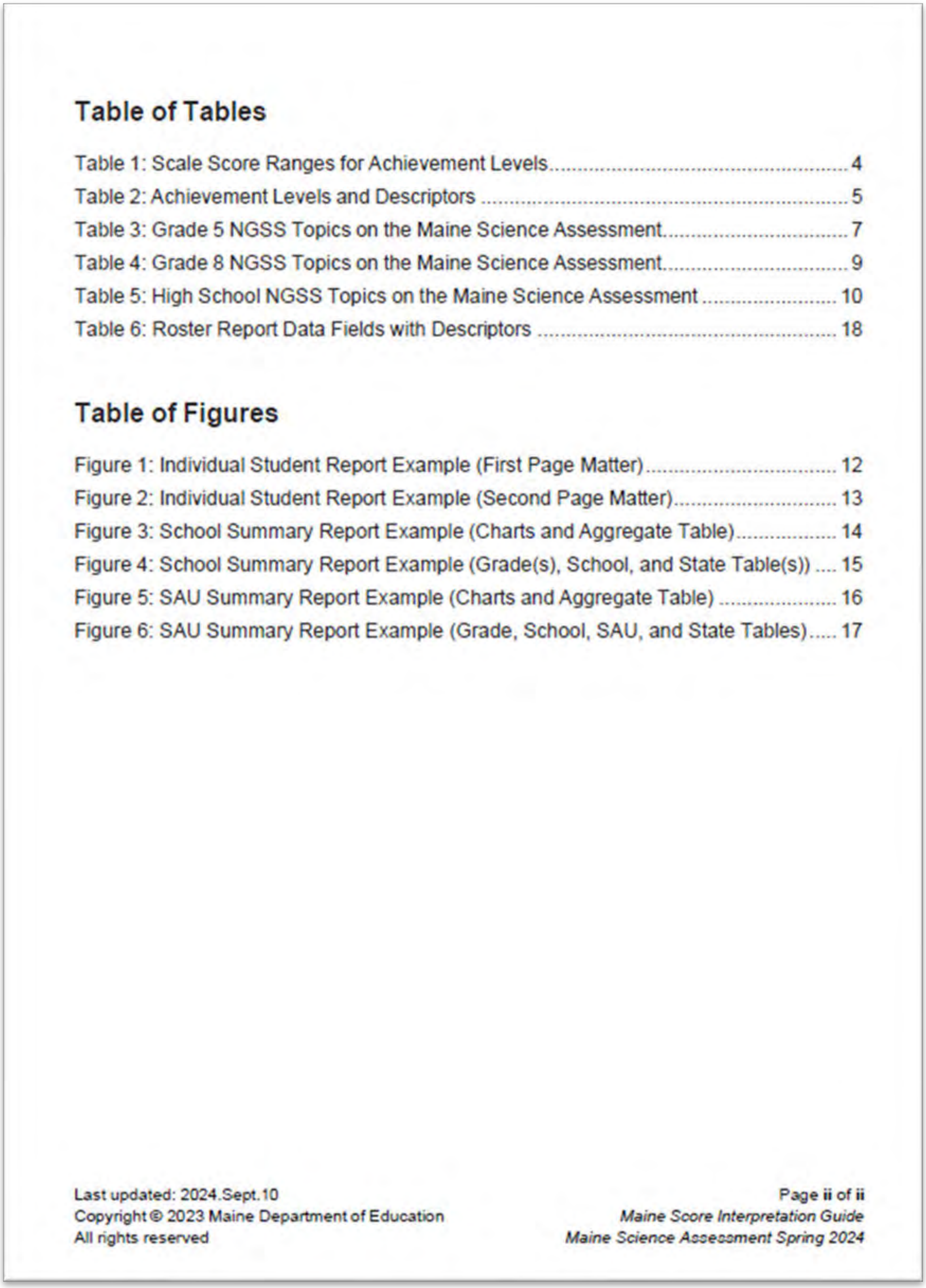
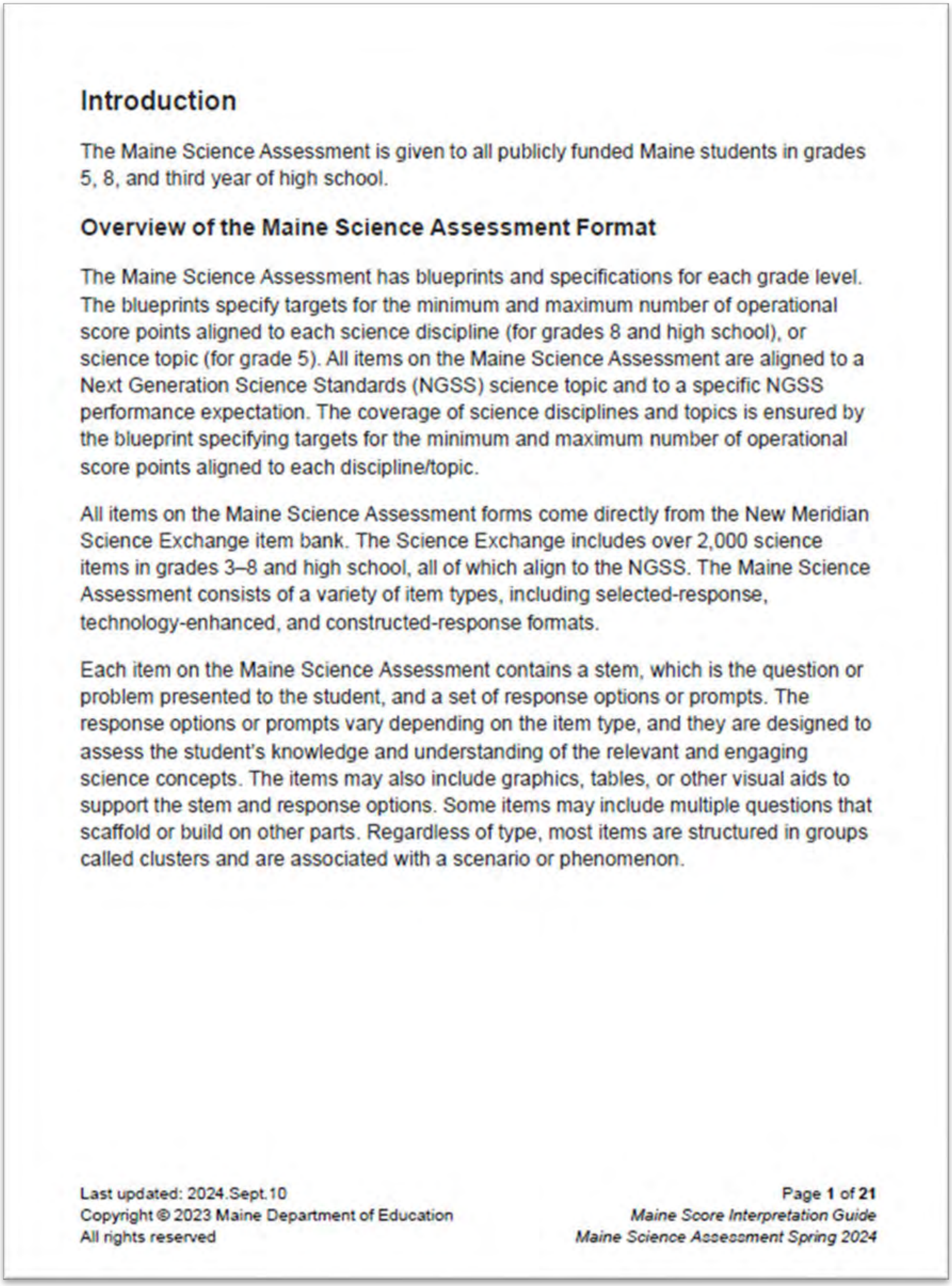


Table of Tables

Table of Figures

Last updated: 2024.Sept.10

Page **ii** of **ii**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



Introduction

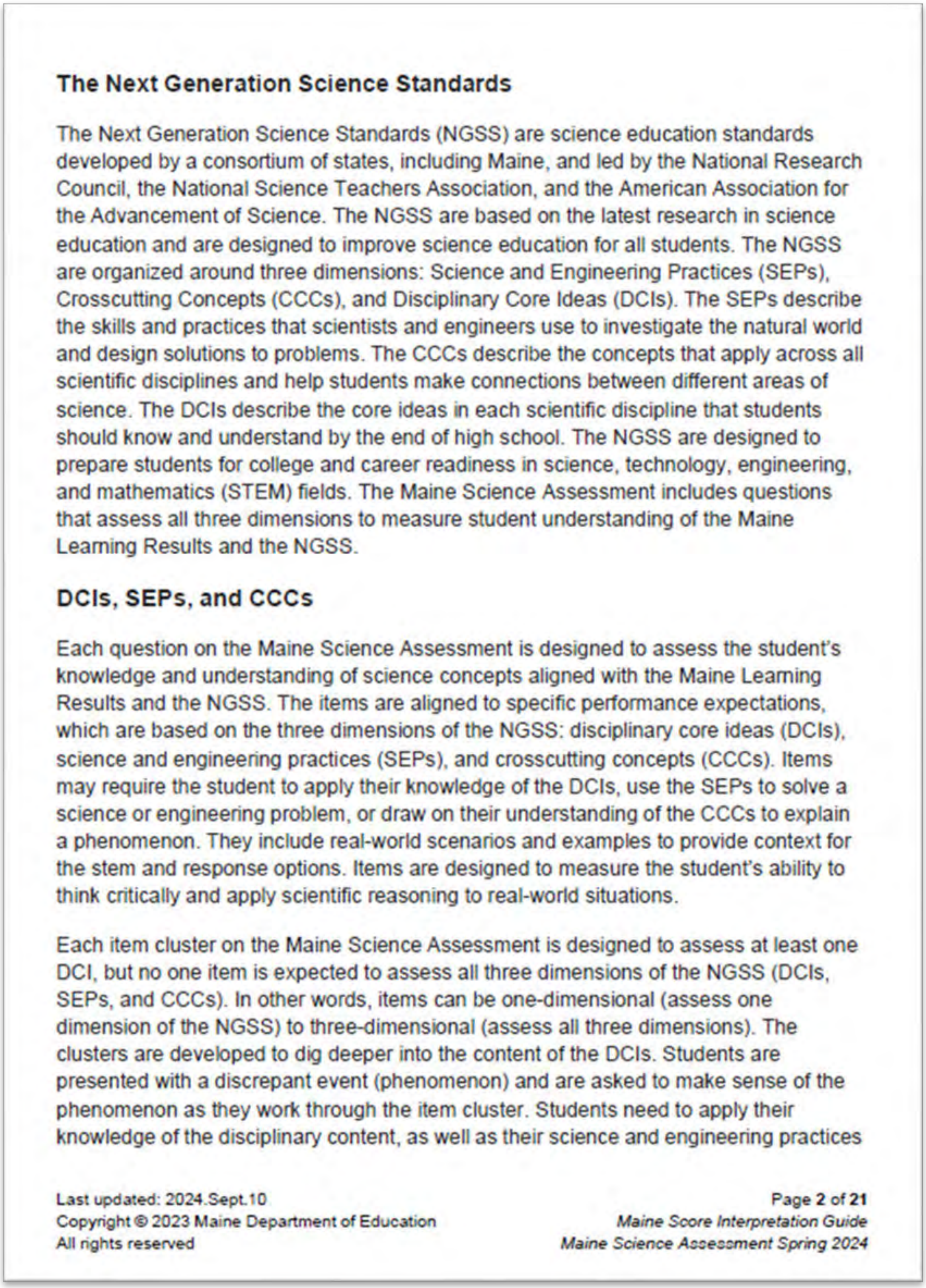
The Maine Science Assessment is given to all publicly funded Maine students in grades 5, 8, and third year of high school.

Overview of the Maine Science Assessment Format

The Maine Science Assessment has blueprints and specifications for each grade level. The blueprints specify targets for the minimum and maximum number of operational score points aligned to each science discipline (for grades 8 and high school), or science topic (for grade 5). All items on the Maine Science Assessment are aligned to a Next Generation Science Standards (NGSS) science topic and to a specific NGSS performance expectation. The coverage of science disciplines and topics is ensured by the blueprint specifying targets for the minimum and maximum number of operational score points aligned to each discipline/topic.

All items on the Maine Science Assessment forms come directly from the New Meridian Science Exchange item bank. The Science Exchange includes over 2,000 science items in grades 3–8 and high school, all of which align to the NGSS. The Maine Science Assessment consists of a variety of item types, including selected-response, technology-enhanced, and constructed-response formats.

Each item on the Maine Science Assessment contains a stem, which is the question or problem presented to the student, and a set of response options or prompts. The response options or prompts vary depending on the item type, and they are designed to assess the student's knowledge and understanding of the relevant and engaging science concepts. The items may also include graphics, tables, or other visual aids to support the stem and response options. Some items may include multiple questions that scaffold or build on other parts. Regardless of type, most items are structured in groups called clusters and are associated with a scenario or phenomenon.



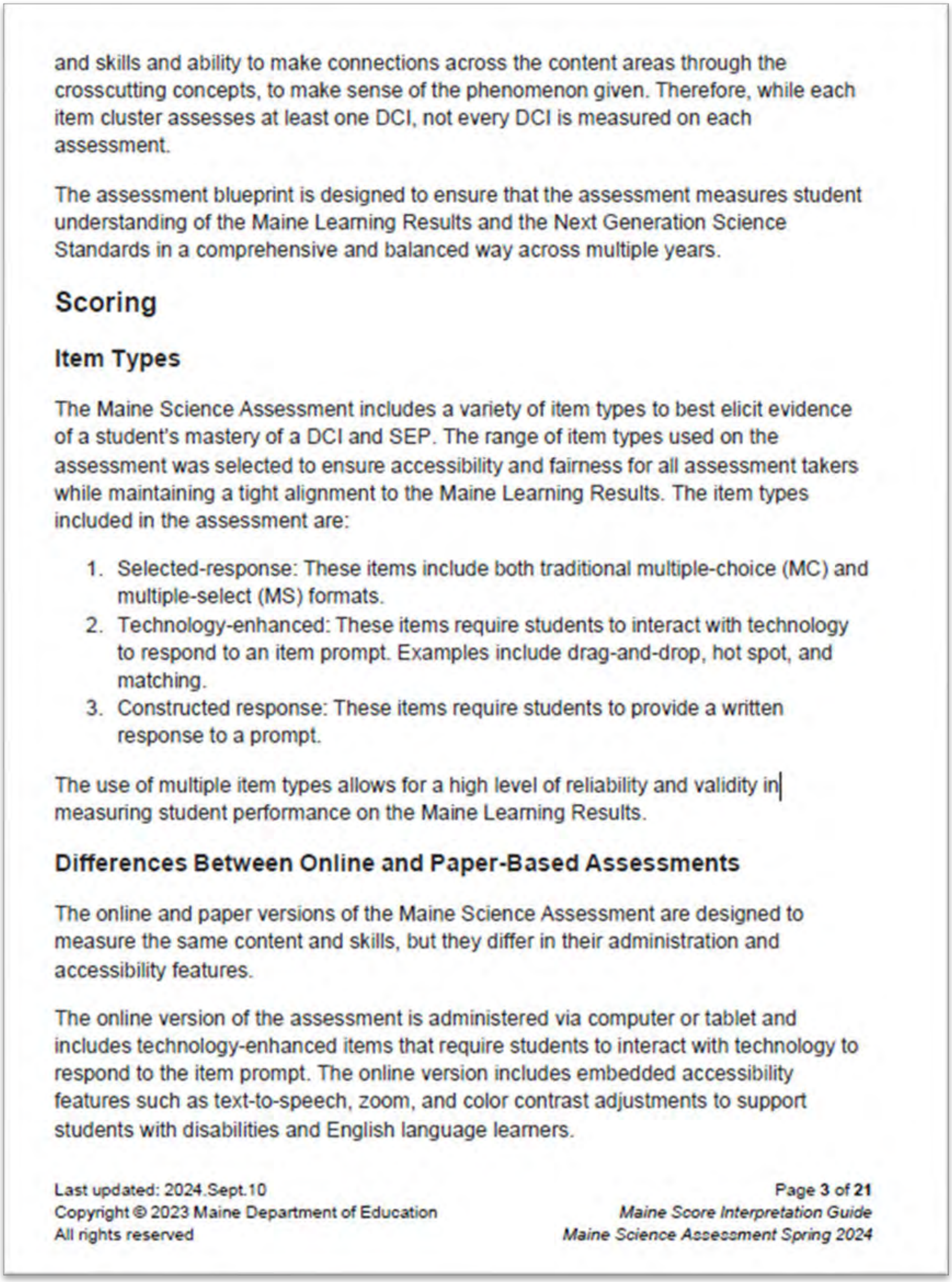
The Next Generation Science Standards

The Next Generation Science Standards (NGSS) are science education standards developed by a consortium of states, including Maine, and led by the National Research Council, the National Science Teachers Association, and the American Association for the Advancement of Science. The NGSS are based on the latest research in science education and are designed to improve science education for all students. The NGSS are organized around three dimensions: Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs). The SEPs describe the skills and practices that scientists and engineers use to investigate the natural world and design solutions to problems. The CCCs describe the concepts that apply across all scientific disciplines and help students make connections between different areas of science. The DCIs describe the core ideas in each scientific discipline that students should know and understand by the end of high school. The NGSS are designed to prepare students for college and career readiness in science, technology, engineering, and mathematics (STEM) fields. The Maine Science Assessment includes questions that assess all three dimensions to measure student understanding of the Maine Learning Results and the NGSS.

DCIs, SEPs, and CCCs

Each question on the Maine Science Assessment is designed to assess the student's knowledge and understanding of science concepts aligned with the Maine Learning Results and the NGSS. The items are aligned to specific performance expectations, which are based on the three dimensions of the NGSS: disciplinary core ideas (DCIs), science and engineering practices (SEPs), and crosscutting concepts (CCCs). Items may require the student to apply their knowledge of the DCIs, use the SEPs to solve a science or engineering problem, or draw on their understanding of the CCCs to explain a phenomenon. They include real-world scenarios and examples to provide context for the stem and response options. Items are designed to measure the student's ability to think critically and apply scientific reasoning to real-world situations.

Each item cluster on the Maine Science Assessment is designed to assess at least one DCI, but no one item is expected to assess all three dimensions of the NGSS (DCIs, SEPs, and CCCs). In other words, items can be one-dimensional (assess one dimension of the NGSS) to three-dimensional (assess all three dimensions). The clusters are developed to dig deeper into the content of the DCIs. Students are presented with a discrepant event (phenomenon) and are asked to make sense of the phenomenon as they work through the item cluster. Students need to apply their knowledge of the disciplinary content, as well as their science and engineering practices



and skills and ability to make connections across the content areas through the crosscutting concepts, to make sense of the phenomenon given. Therefore, while each item cluster assesses at least one DCI, not every DCI is measured on each assessment.

The assessment blueprint is designed to ensure that the assessment measures student understanding of the Maine Learning Results and the Next Generation Science Standards in a comprehensive and balanced way across multiple years.

Scoring

Item Types

The Maine Science Assessment includes a variety of item types to best elicit evidence of a student's mastery of a DCI and SEP. The range of item types used on the assessment was selected to ensure accessibility and fairness for all assessment takers while maintaining a tight alignment to the Maine Learning Results. The item types included in the assessment are:

1. Selected-response: These items include both traditional multiple-choice (MC) and multiple-select (MS) formats.
2. Technology-enhanced: These items require students to interact with technology to respond to an item prompt. Examples include drag-and-drop, hot spot, and matching.
3. Constructed response: These items require students to provide a written response to a prompt.

The use of multiple item types allows for a high level of reliability and validity in measuring student performance on the Maine Learning Results.

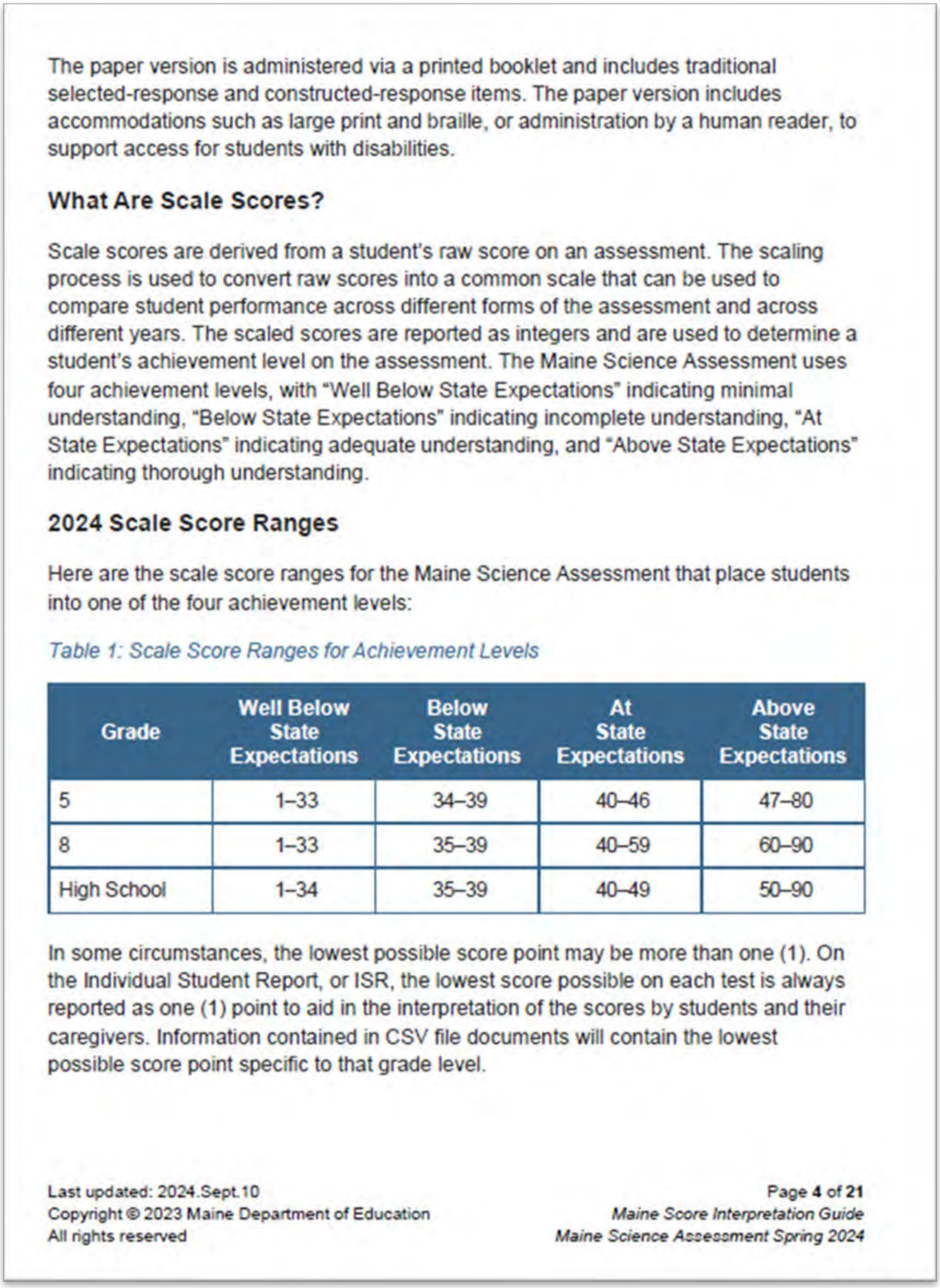
Differences Between Online and Paper-Based Assessments

The online and paper versions of the Maine Science Assessment are designed to measure the same content and skills, but they differ in their administration and accessibility features.

The online version of the assessment is administered via computer or tablet and includes technology-enhanced items that require students to interact with technology to respond to the item prompt. The online version includes embedded accessibility features such as text-to-speech, zoom, and color contrast adjustments to support students with disabilities and English language learners.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page 3 of 21
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



The paper version is administered via a printed booklet and includes traditional selected-response and constructed-response items. The paper version includes accommodations such as large print and braille, or administration by a human reader, to support access for students with disabilities.

What Are Scale Scores?

Scale scores are derived from a student's raw score on an assessment. The scaling process is used to convert raw scores into a common scale that can be used to compare student performance across different forms of the assessment and across different years. The scaled scores are reported as integers and are used to determine a student's achievement level on the assessment. The Maine Science Assessment uses four achievement levels, with "Well Below State Expectations" indicating minimal understanding, "Below State Expectations" indicating incomplete understanding, "At State Expectations" indicating adequate understanding, and "Above State Expectations" indicating thorough understanding.

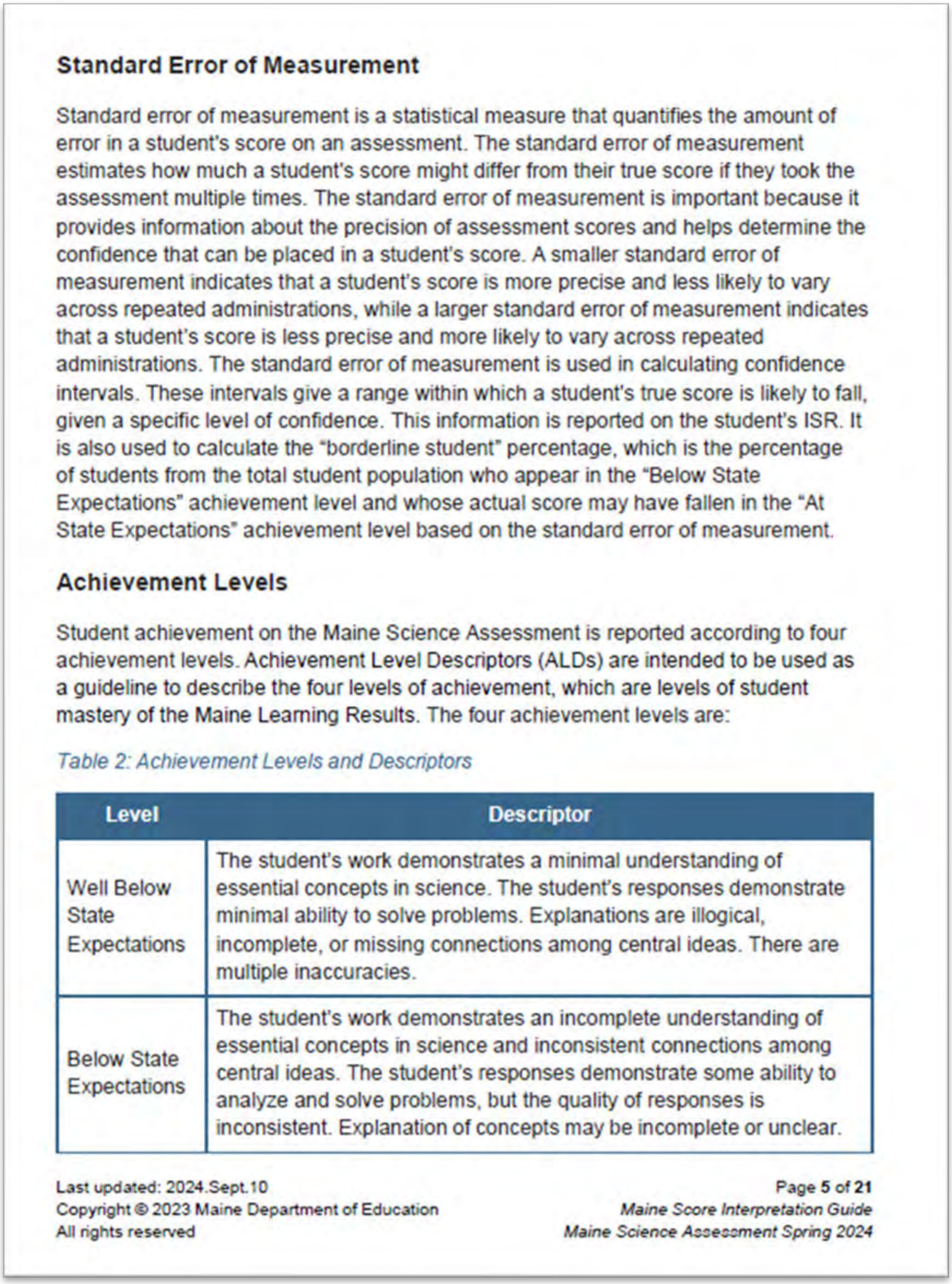
2024 Scale Score Ranges

Here are the scale score ranges for the Maine Science Assessment that place students into one of the four achievement levels:

Table 1: Scale Score Ranges for Achievement Levels

Grade	Well Below State Expectations	Below State Expectations	At State Expectations	Above State Expectations
5	1–33	34–39	40–46	47–80
8	1–33	35–39	40–59	60–90
High School	1–34	35–39	40–49	50–90

In some circumstances, the lowest possible score point may be more than one (1). On the Individual Student Report, or ISR, the lowest score possible on each test is always reported as one (1) point to aid in the interpretation of the scores by students and their caregivers. Information contained in CSV file documents will contain the lowest possible score point specific to that grade level.



Standard Error of Measurement

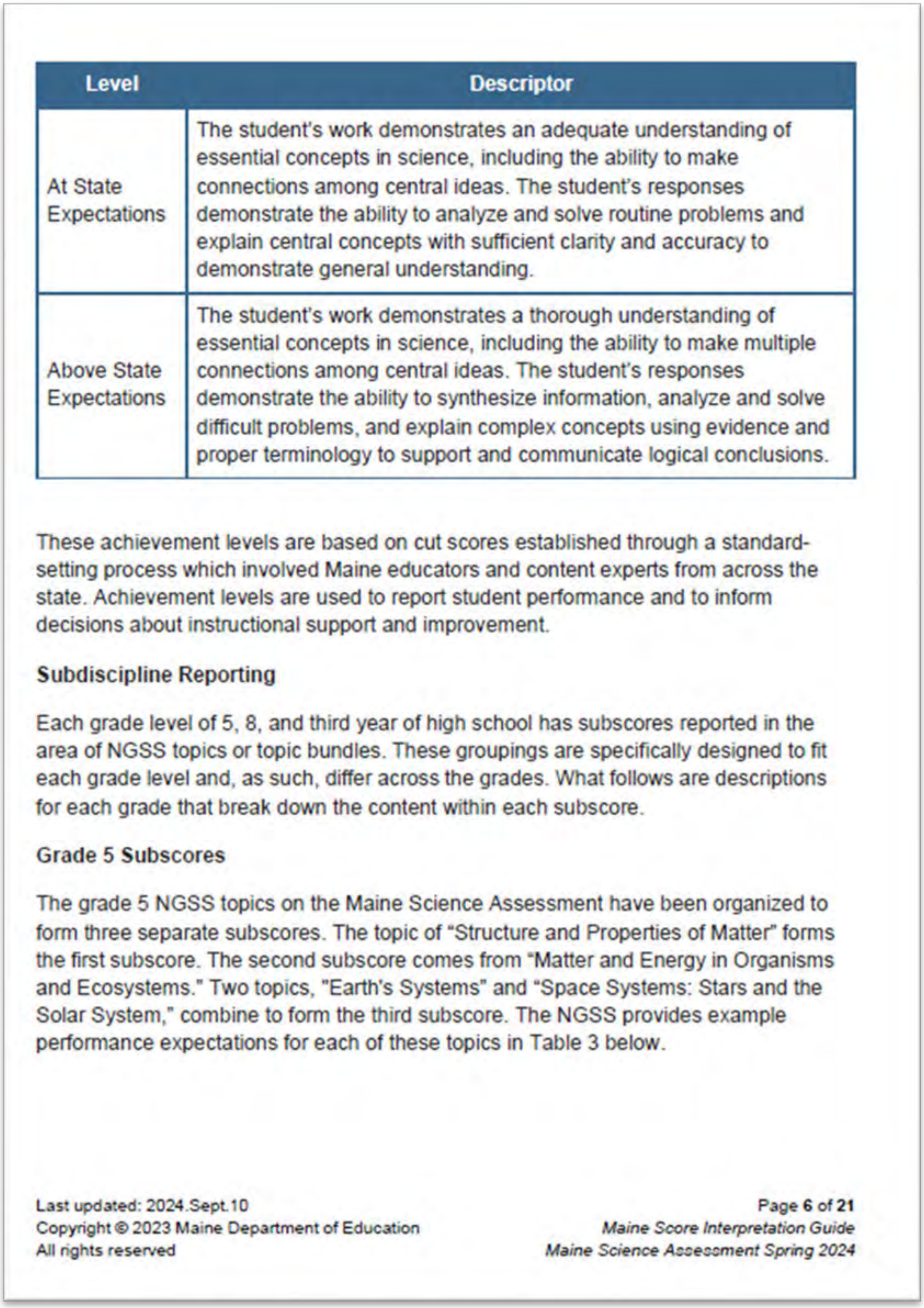
Standard error of measurement is a statistical measure that quantifies the amount of error in a student's score on an assessment. The standard error of measurement estimates how much a student's score might differ from their true score if they took the assessment multiple times. The standard error of measurement is important because it provides information about the precision of assessment scores and helps determine the confidence that can be placed in a student's score. A smaller standard error of measurement indicates that a student's score is more precise and less likely to vary across repeated administrations, while a larger standard error of measurement indicates that a student's score is less precise and more likely to vary across repeated administrations. The standard error of measurement is used in calculating confidence intervals. These intervals give a range within which a student's true score is likely to fall, given a specific level of confidence. This information is reported on the student's ISR. It is also used to calculate the "borderline student" percentage, which is the percentage of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

Achievement Levels

Student achievement on the Maine Science Assessment is reported according to four achievement levels. Achievement Level Descriptors (ALDs) are intended to be used as a guideline to describe the four levels of achievement, which are levels of student mastery of the Maine Learning Results. The four achievement levels are:

Table 2: Achievement Levels and Descriptors

Level	Descriptor
Well Below State Expectations	The student's work demonstrates a minimal understanding of essential concepts in science. The student's responses demonstrate minimal ability to solve problems. Explanations are illogical, incomplete, or missing connections among central ideas. There are multiple inaccuracies.
Below State Expectations	The student's work demonstrates an incomplete understanding of essential concepts in science and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems, but the quality of responses is inconsistent. Explanation of concepts may be incomplete or unclear.



Level	Descriptor
At State Expectations	The student's work demonstrates an adequate understanding of essential concepts in science, including the ability to make connections among central ideas. The student's responses demonstrate the ability to analyze and solve routine problems and explain central concepts with sufficient clarity and accuracy to demonstrate general understanding.
Above State Expectations	The student's work demonstrates a thorough understanding of essential concepts in science, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and explain complex concepts using evidence and proper terminology to support and communicate logical conclusions.

These achievement levels are based on cut scores established through a standard-setting process which involved Maine educators and content experts from across the state. Achievement levels are used to report student performance and to inform decisions about instructional support and improvement.

Subdiscipline Reporting

Each grade level of 5, 8, and third year of high school has subscores reported in the area of NGSS topics or topic bundles. These groupings are specifically designed to fit each grade level and, as such, differ across the grades. What follows are descriptions for each grade that break down the content within each subscore.

Grade 5 Subscores

The grade 5 NGSS topics on the Maine Science Assessment have been organized to form three separate subscores. The topic of "Structure and Properties of Matter" forms the first subscore. The second subscore comes from "Matter and Energy in Organisms and Ecosystems." Two topics, "Earth's Systems" and "Space Systems: Stars and the Solar System," combine to form the third subscore. The NGSS provides example performance expectations for each of these topics in Table 3 below.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **6** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024

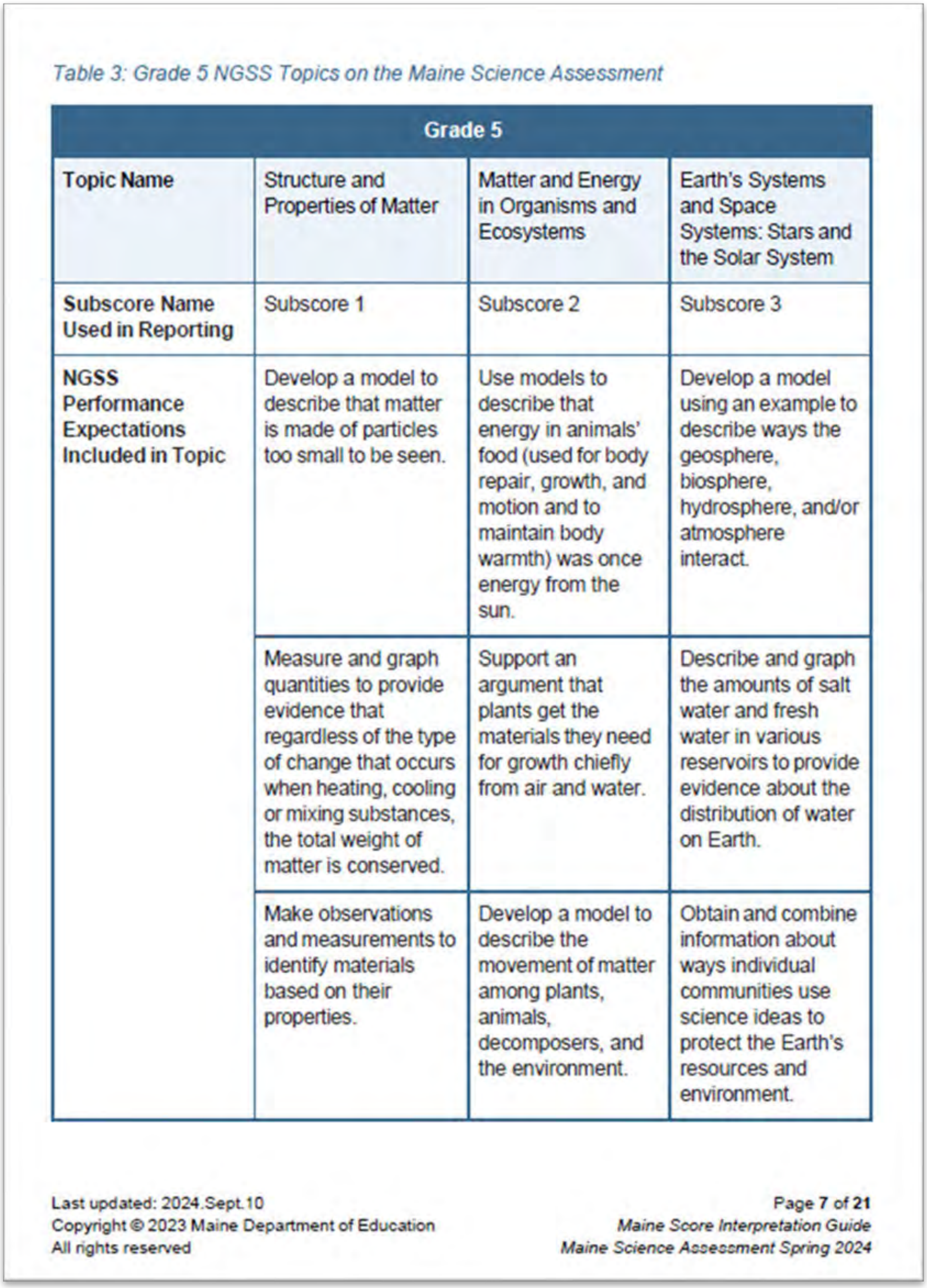
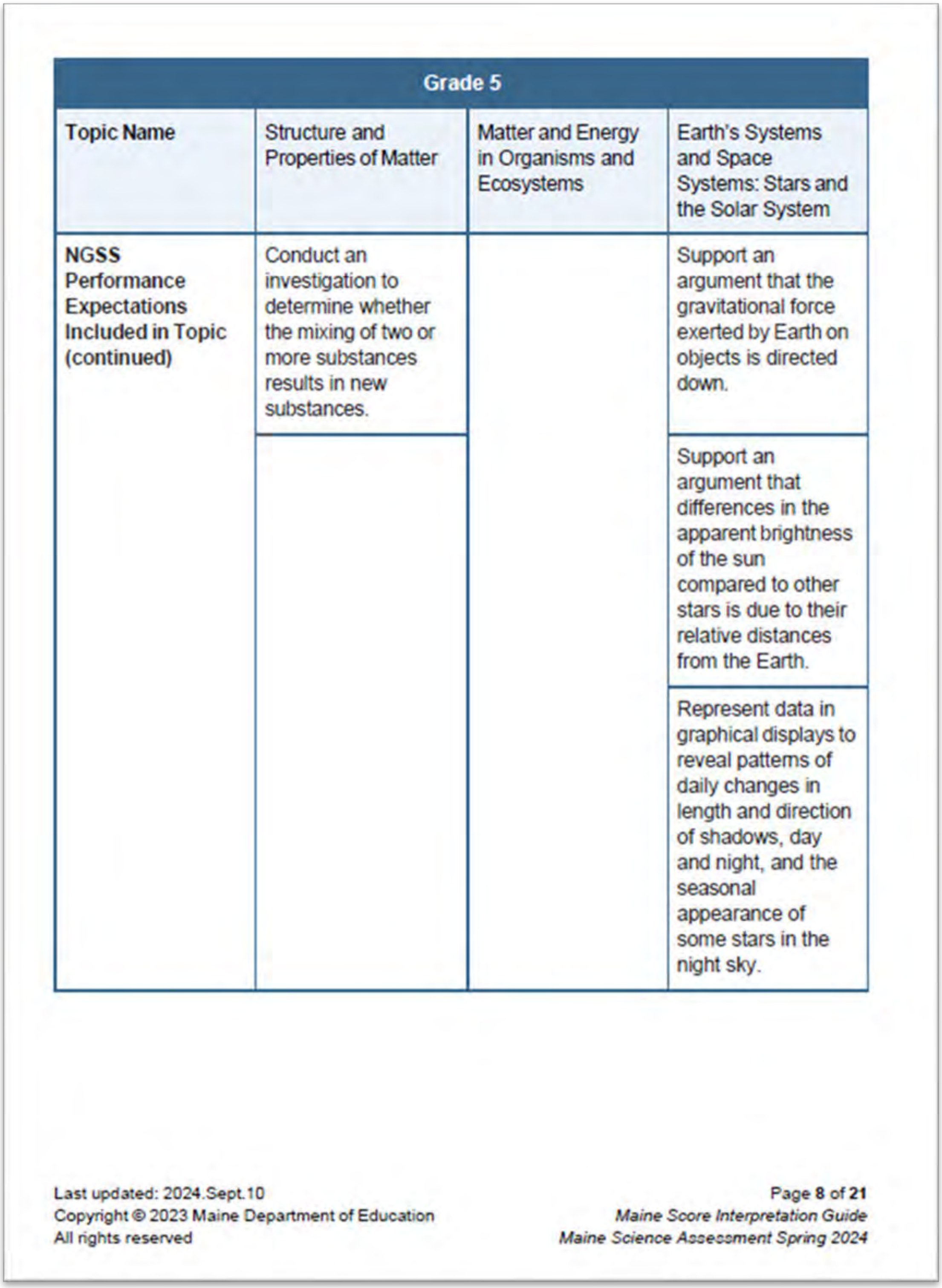


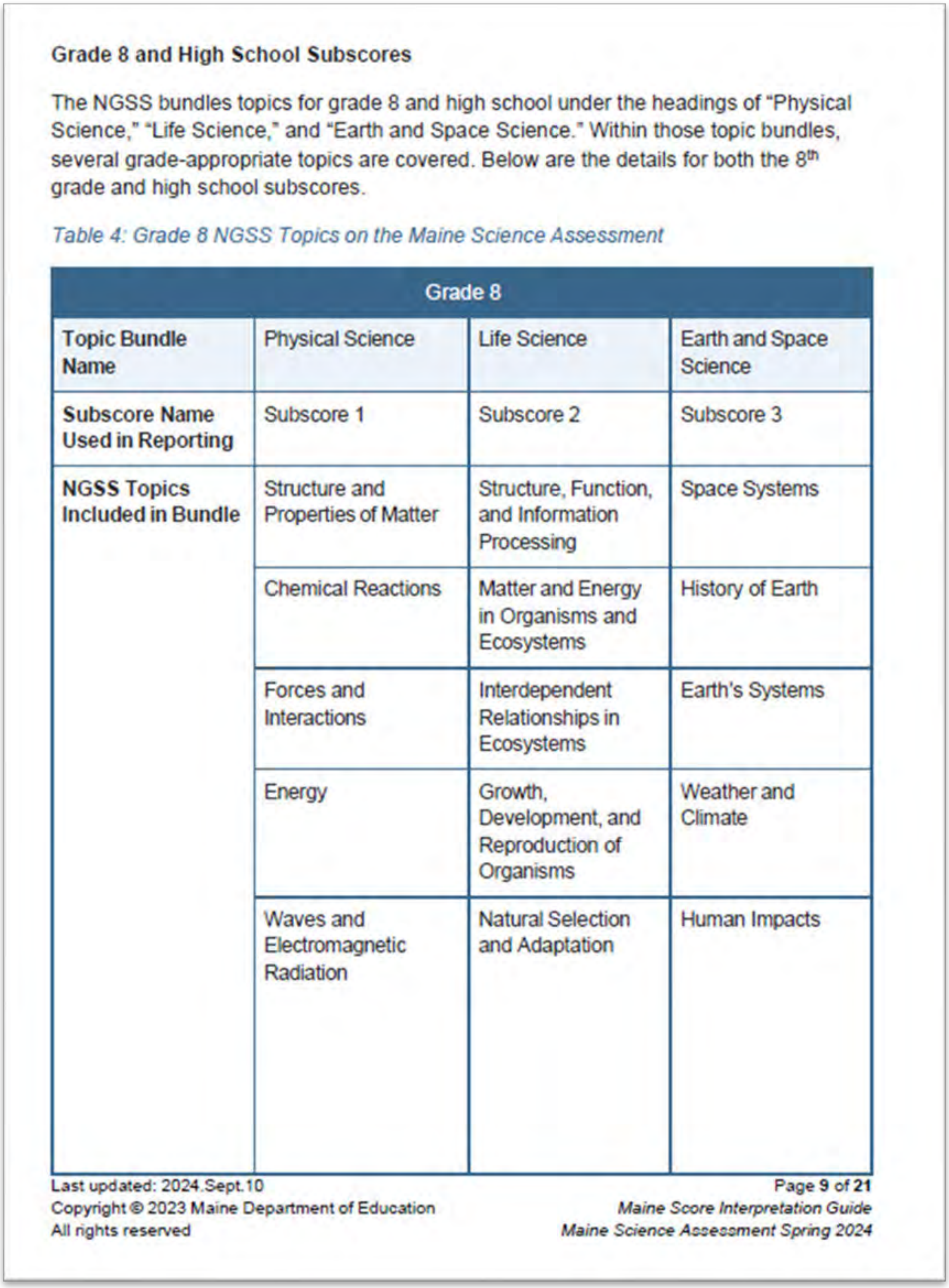
Table 3: Grade 5 NGSS Topics on the Maine Science Assessment

Grade 5			
Topic Name	Structure and Properties of Matter	Matter and Energy in Organisms and Ecosystems	Earth's Systems and Space Systems: Stars and the Solar System
Subscore Name Used in Reporting	Subscore 1	Subscore 2	Subscore 3
NGSS Performance Expectations Included in Topic	Develop a model to describe that matter is made of particles too small to be seen.	Use models to describe that energy in animals' food (used for body repair, growth, and motion and to maintain body warmth) was once energy from the sun.	Develop a model using an example to describe ways the geosphere, biosphere, hydrosphere, and/or atmosphere interact.
	Measure and graph quantities to provide evidence that regardless of the type of change that occurs when heating, cooling or mixing substances, the total weight of matter is conserved.	Support an argument that plants get the materials they need for growth chiefly from air and water.	Describe and graph the amounts of salt water and fresh water in various reservoirs to provide evidence about the distribution of water on Earth.
	Make observations and measurements to identify materials based on their properties.	Develop a model to describe the movement of matter among plants, animals, decomposers, and the environment.	Obtain and combine information about ways individual communities use science ideas to protect the Earth's resources and environment.

Last updated: 2024.Sept.10



Grade 5			
Topic Name	Structure and Properties of Matter	Matter and Energy in Organisms and Ecosystems	Earth's Systems and Space Systems: Stars and the Solar System
NGSS Performance Expectations Included in Topic (continued)	Conduct an investigation to determine whether the mixing of two or more substances results in new substances.		Support an argument that the gravitational force exerted by Earth on objects is directed down.
			Support an argument that differences in the apparent brightness of the sun compared to other stars is due to their relative distances from the Earth.
			Represent data in graphical displays to reveal patterns of daily changes in length and direction of shadows, day and night, and the seasonal appearance of some stars in the night sky.



Grade 8 and High School Subscores

The NGSS bundles topics for grade 8 and high school under the headings of "Physical Science," "Life Science," and "Earth and Space Science." Within those topic bundles, several grade-appropriate topics are covered. Below are the details for both the 8th grade and high school subscores.

Table 4: Grade 8 NGSS Topics on the Maine Science Assessment

Grade 8			
Topic Bundle Name	Physical Science	Life Science	Earth and Space Science
Subscore Name Used in Reporting	Subscore 1	Subscore 2	Subscore 3
NGSS Topics Included in Bundle	Structure and Properties of Matter	Structure, Function, and Information Processing	Space Systems
	Chemical Reactions	Matter and Energy in Organisms and Ecosystems	History of Earth
	Forces and Interactions	Interdependent Relationships in Ecosystems	Earth's Systems
	Energy	Growth, Development, and Reproduction of Organisms	Weather and Climate
	Waves and Electromagnetic Radiation	Natural Selection and Adaptation	Human Impacts

Last updated: 2024.Sept.10

Page 9 of 21
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024

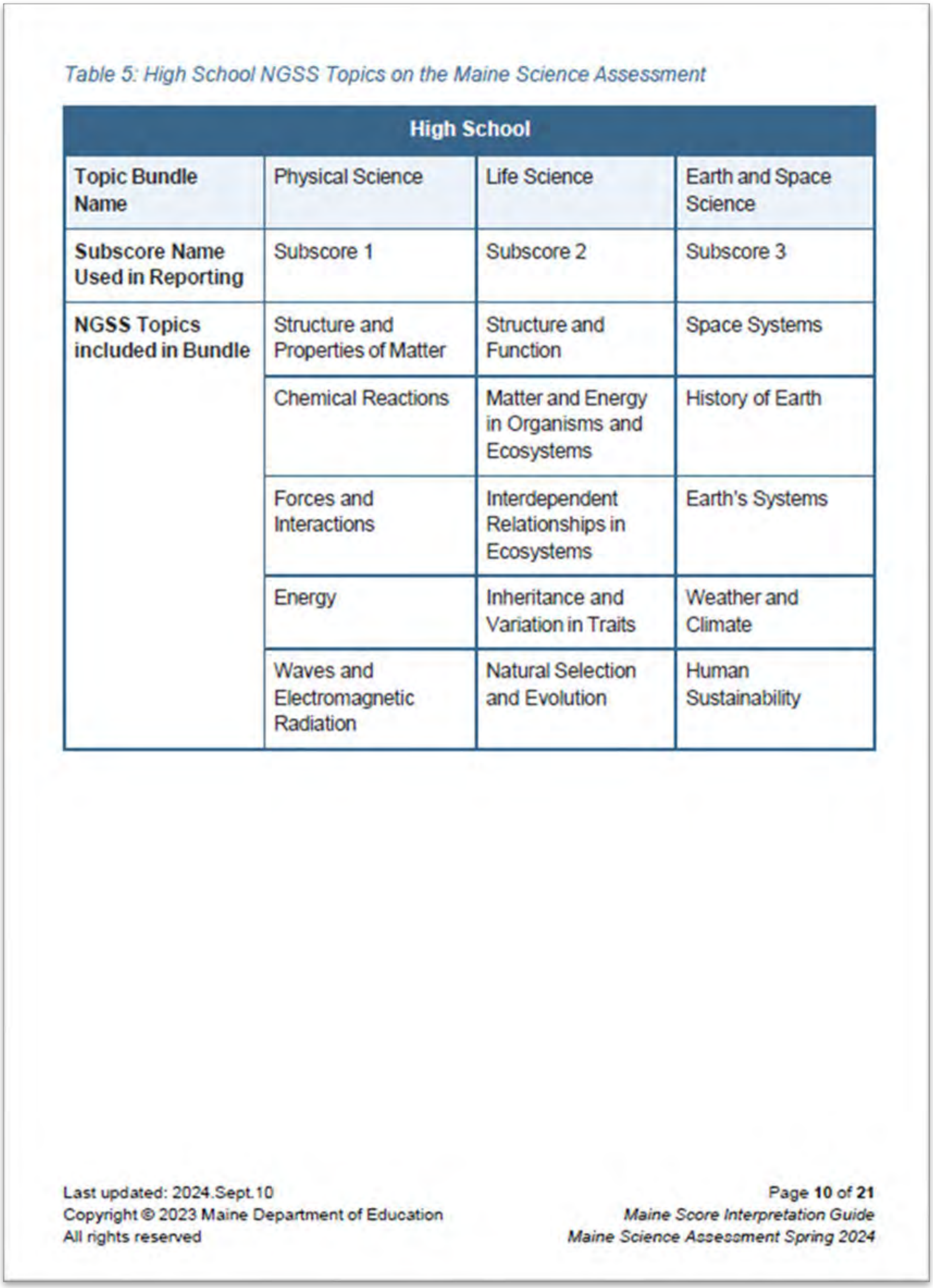
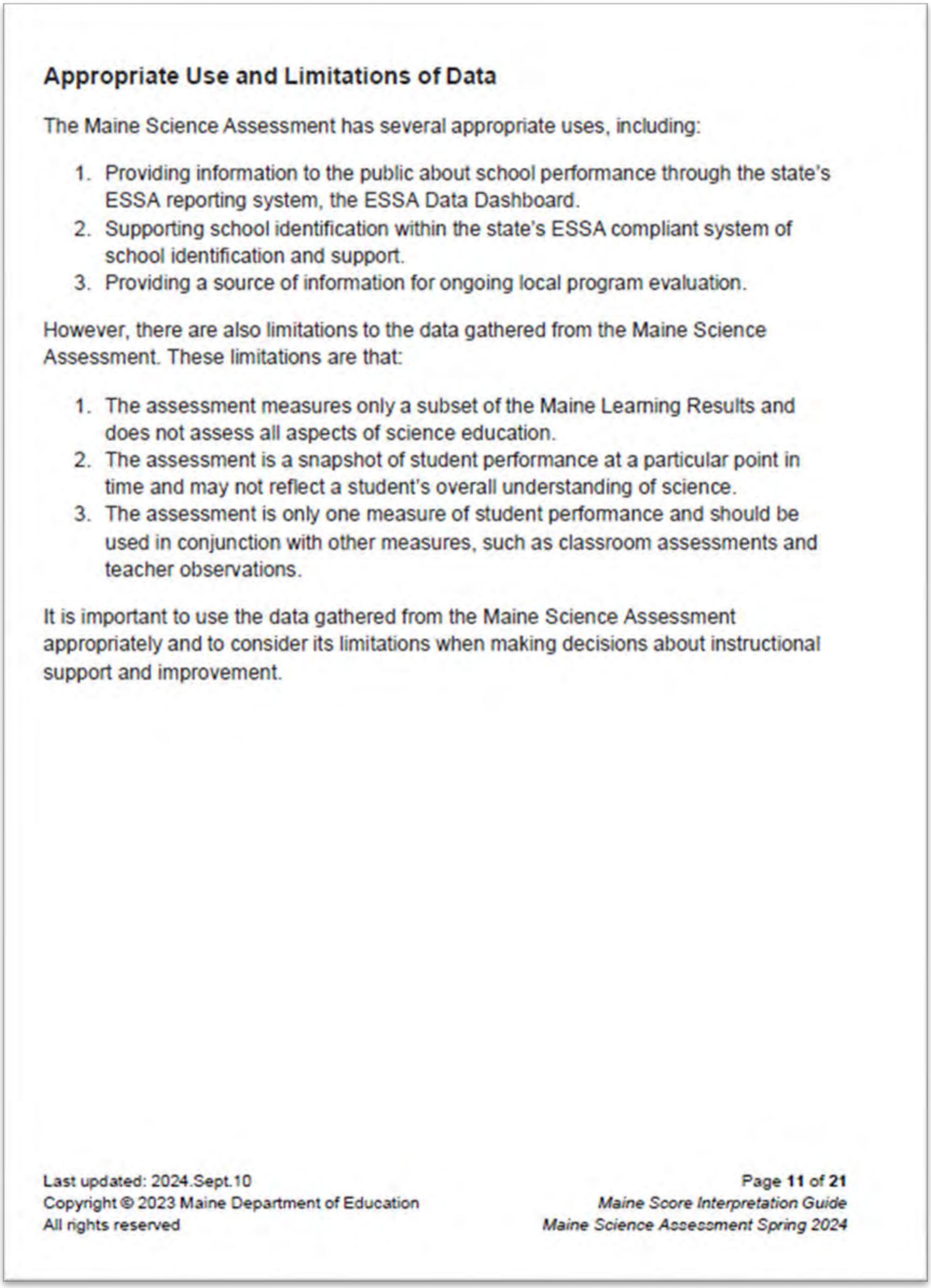


Table 5: High School NGSS Topics on the Maine Science Assessment

High School			
Topic Bundle Name	Physical Science	Life Science	Earth and Space Science
Subscore Name Used in Reporting	Subscore 1	Subscore 2	Subscore 3
NGSS Topics included in Bundle	Structure and Properties of Matter	Structure and Function	Space Systems
	Chemical Reactions	Matter and Energy in Organisms and Ecosystems	History of Earth
	Forces and Interactions	Interdependent Relationships in Ecosystems	Earth's Systems
	Energy	Inheritance and Variation in Traits	Weather and Climate
	Waves and Electromagnetic Radiation	Natural Selection and Evolution	Human Sustainability

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **10** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



Appropriate Use and Limitations of Data

The Maine Science Assessment has several appropriate uses, including:

1. Providing information to the public about school performance through the state's ESSA reporting system, the ESSA Data Dashboard.
2. Supporting school identification within the state's ESSA compliant system of school identification and support.
3. Providing a source of information for ongoing local program evaluation.

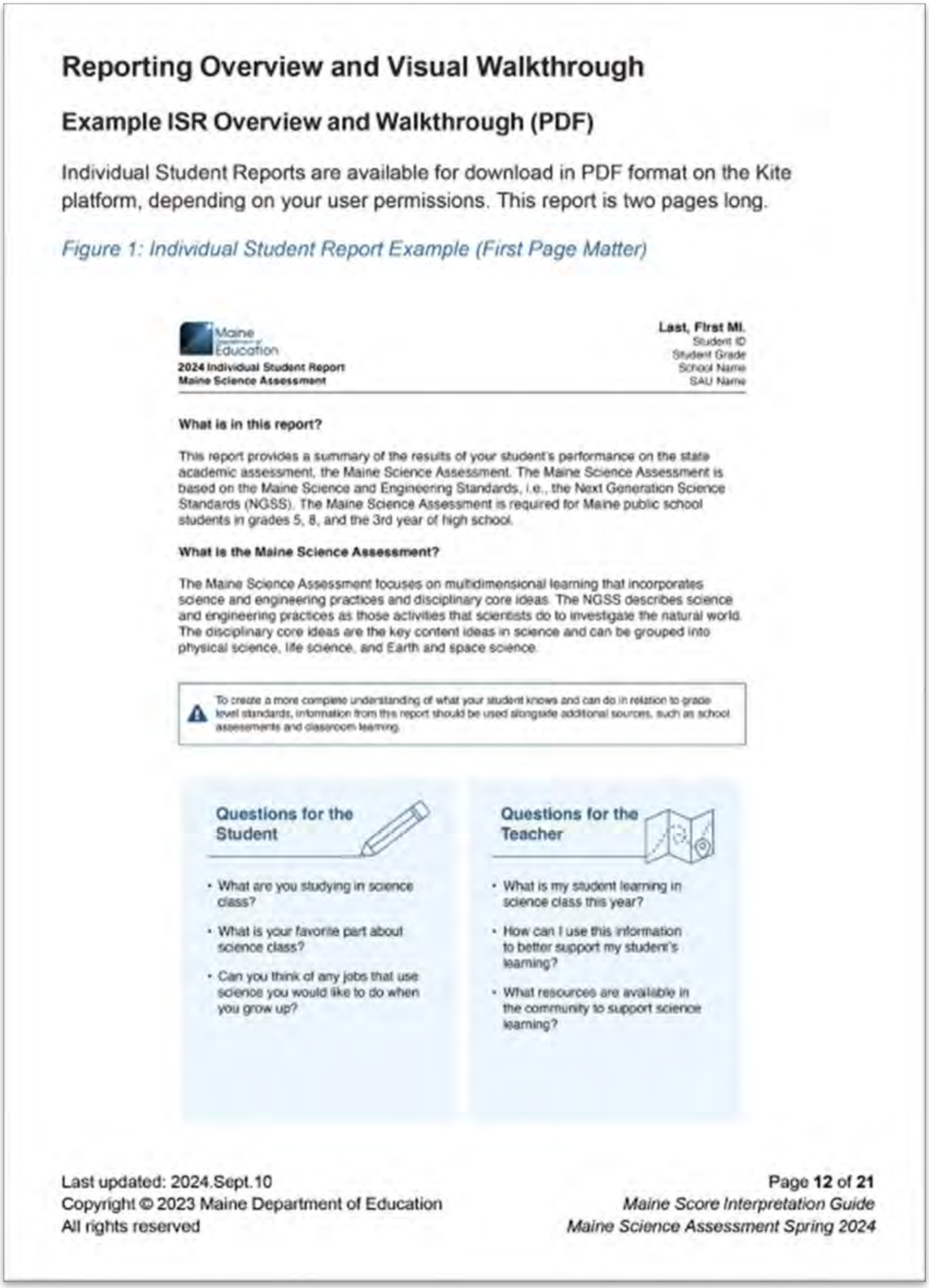
However, there are also limitations to the data gathered from the Maine Science Assessment. These limitations are that:

1. The assessment measures only a subset of the Maine Learning Results and does not assess all aspects of science education.
2. The assessment is a snapshot of student performance at a particular point in time and may not reflect a student's overall understanding of science.
3. The assessment is only one measure of student performance and should be used in conjunction with other measures, such as classroom assessments and teacher observations.

It is important to use the data gathered from the Maine Science Assessment appropriately and to consider its limitations when making decisions about instructional support and improvement.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **11** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



Reporting Overview and Visual Walkthrough

Example ISR Overview and Walkthrough (PDF)

Individual Student Reports are available for download in PDF format on the Kite platform, depending on your user permissions. This report is two pages long.

Figure 1: Individual Student Report Example (First Page Matter)

Maine Department of Education

2024 Individual Student Report
Maine Science Assessment

Last, First MI.
Student ID
Student Grade
School Name
SAU Name

What is in this report?

This report provides a summary of the results of your student's performance on the state academic assessment, the Maine Science Assessment. The Maine Science Assessment is based on the Maine Science and Engineering Standards, i.e., the Next Generation Science Standards (NGSS). The Maine Science Assessment is required for Maine public school students in grades 5, 8, and the 3rd year of high school.

What is the Maine Science Assessment?

The Maine Science Assessment focuses on multidimensional learning that incorporates science and engineering practices and disciplinary core ideas. The NGSS describes science and engineering practices as those activities that scientists do to investigate the natural world. The disciplinary core ideas are the key content ideas in science and can be grouped into physical science, life science, and Earth and space science.

⚠ To create a more complete understanding of what your student knows and can do in relation to grade level standards, information from this report should be used alongside additional sources, such as school assessments and classroom learning.

Questions for the Student

- What are you studying in science class?
- What is your favorite part about science class?
- Can you think of any jobs that use science you would like to do when you grow up?

Questions for the Teacher

- What is my student learning in science class this year?
- How can I use this information to better support my student's learning?
- What resources are available in the community to support science learning?

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **12** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024

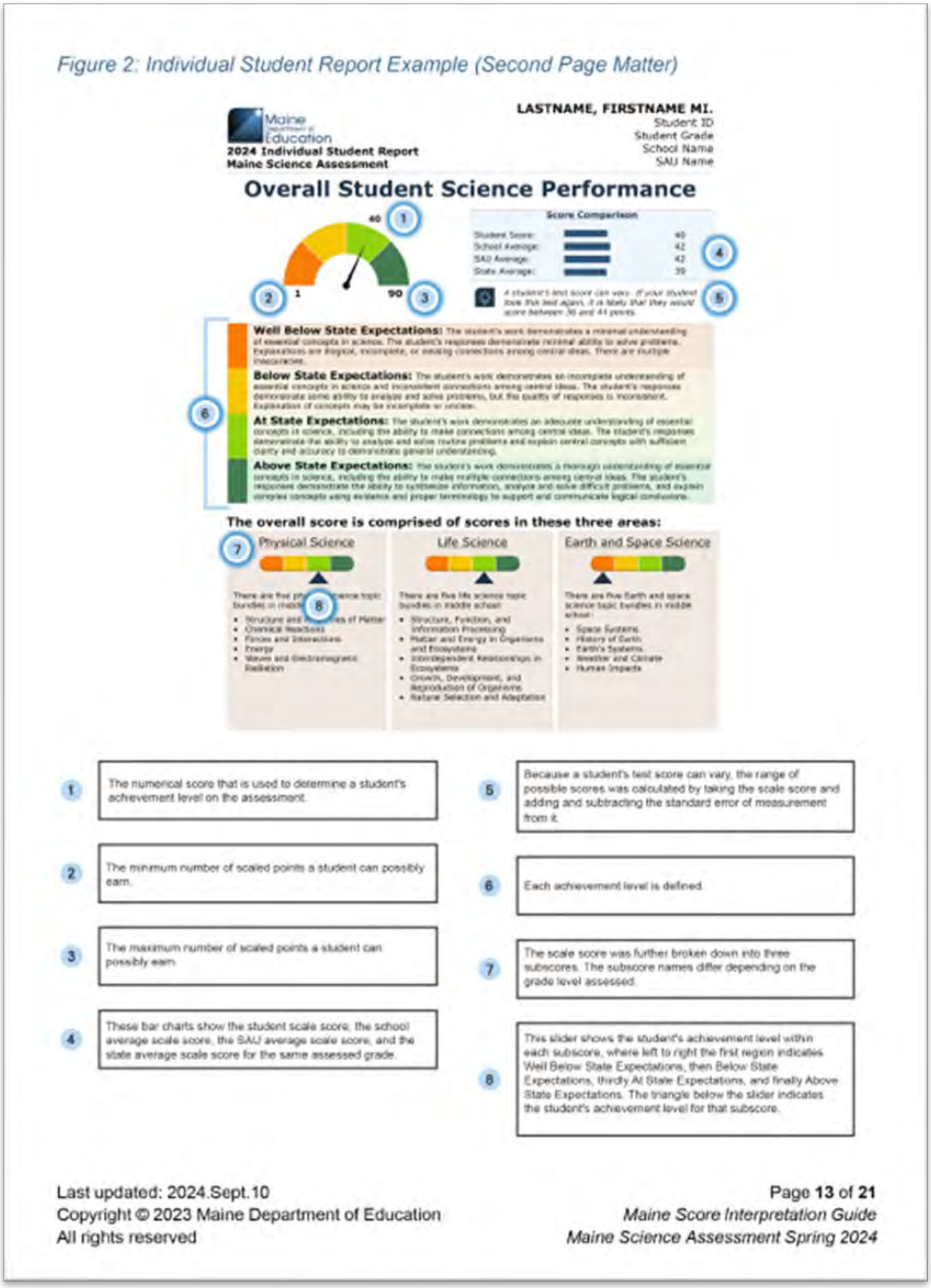


Figure 2: Individual Student Report Example (Second Page Matter)

Maine Department of Education
2024 Individual Student Report
Maine Science Assessment

LASTNAME, FIRSTNAME MI.
Student ID
Student Grade
School Name
SAU Name

Overall Student Science Performance

Score Comparison

Student Score:	40
School Average:	42
SAU Average:	42
State Average:	39

A student's test score can vary. If your student took this test again, it is likely that they would score between 36 and 44 points.

Well Below State Expectations: The student's work demonstrates a minimal understanding of essential concepts in science. The student's responses demonstrate minimal ability to solve problems. Explanations are illogical, incomplete, or missing connections among central ideas. There are multiple inaccuracies.

Below State Expectations: The student's work demonstrates an incomplete understanding of essential concepts in science and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems, but the quality of responses is inconsistent. Explanation of concepts may be incomplete or unclear.

At State Expectations: The student's work demonstrates an adequate understanding of essential concepts in science, including the ability to make connections among central ideas. The student's responses demonstrate the ability to analyze and solve routine problems and explain central concepts with sufficient clarity and accuracy to demonstrate general understanding.

Above State Expectations: The student's work demonstrates a thorough understanding of essential concepts in science, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and explain complex concepts using evidence and proper terminology to support and communicate logical conclusions.

The overall score is comprised of scores in these three areas:

Physical Science

There are five physical science topic bundles in middle school:

- Structure and Properties of Matter
- Chemical Reactions
- Forces and Interactions
- Energy
- Waves and Electromagnetic Radiation

Life Science

There are five life science topic bundles in middle school:

- Structure, Function, and Information Processing
- Matter and Energy in Organisms and Ecosystems
- Interdependent Relationships in Ecosystems
- Growth, Development, and Reproduction of Organisms
- Natural Selection and Adaptation

Earth and Space Science

There are five Earth and space science topic bundles in middle school:

- Space Systems
- History of Earth
- Earth's Systems
- Weather and Climate
- Human Impacts

1 The numerical score that is used to determine a student's achievement level on the assessment.

2 The minimum number of scaled points a student can possibly earn.

3 The maximum number of scaled points a student can possibly earn.

4 These bar charts show the student scale score, the school average scale score, the SAU average scale score, and the state average scale score for the same assessed grade.

5 Because a student's test score can vary, the range of possible scores was calculated by taking the scale score and adding and subtracting the standard error of measurement from it.

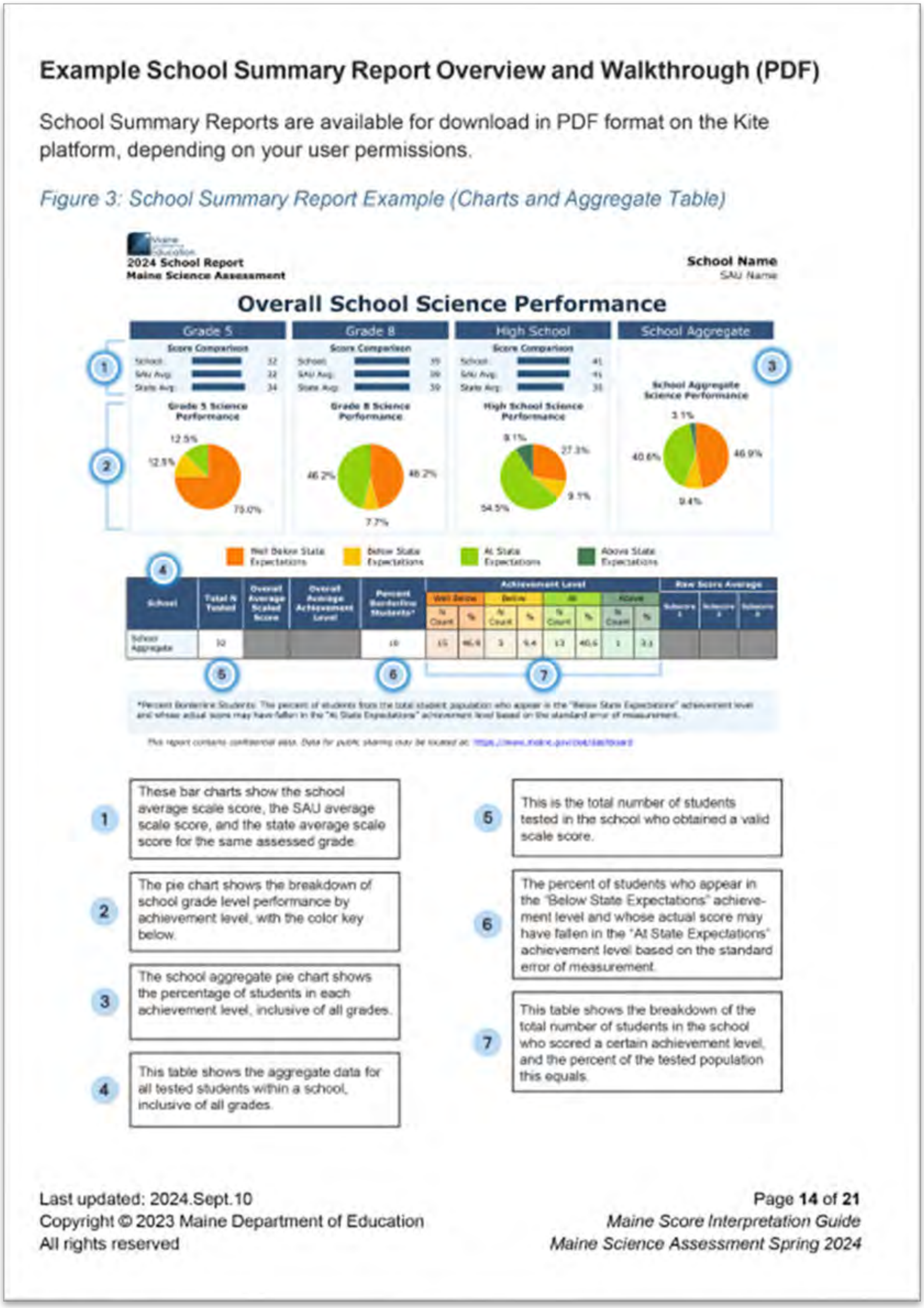
6 Each achievement level is defined.

7 The scale score was further broken down into three subscores. The subscore names differ depending on the grade level assessed.

8 This slider shows the student's achievement level within each subscore, where left to right the first region indicates Well Below State Expectations, then Below State Expectations, thirdly At State Expectations, and finally Above State Expectations. The triangle below the slider indicates the student's achievement level for that subscore.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **13** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



Example School Summary Report Overview and Walkthrough (PDF)

School Summary Reports are available for download in PDF format on the Kite platform, depending on your user permissions.

Figure 3: School Summary Report Example (Charts and Aggregate Table)

School	Total N Tested	Overall Average Scaled Score	Overall Average Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School Aggregate	32			10	15	46.9	3	9.4	13	40.6	1	3.1			

*Percent Borderline Students: The percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

This report contains confidential data. Data for public sharing may be located at: https://www.maine.gov/doe/dashboard

1 These bar charts show the school average scale score, the SAU average scale score, and the state average scale score for the same assessed grade.

2 The pie chart shows the breakdown of school grade level performance by achievement level, with the color key below.

3 The school aggregate pie chart shows the percentage of students in each achievement level, inclusive of all grades.

4 This table shows the aggregate data for all tested students within a school, inclusive of all grades.

5 This is the total number of students tested in the school who obtained a valid scale score.

6 The percent of students who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

7 This table shows the breakdown of the total number of students in the school who scored a certain achievement level, and the percent of the tested population this equals.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **14** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024

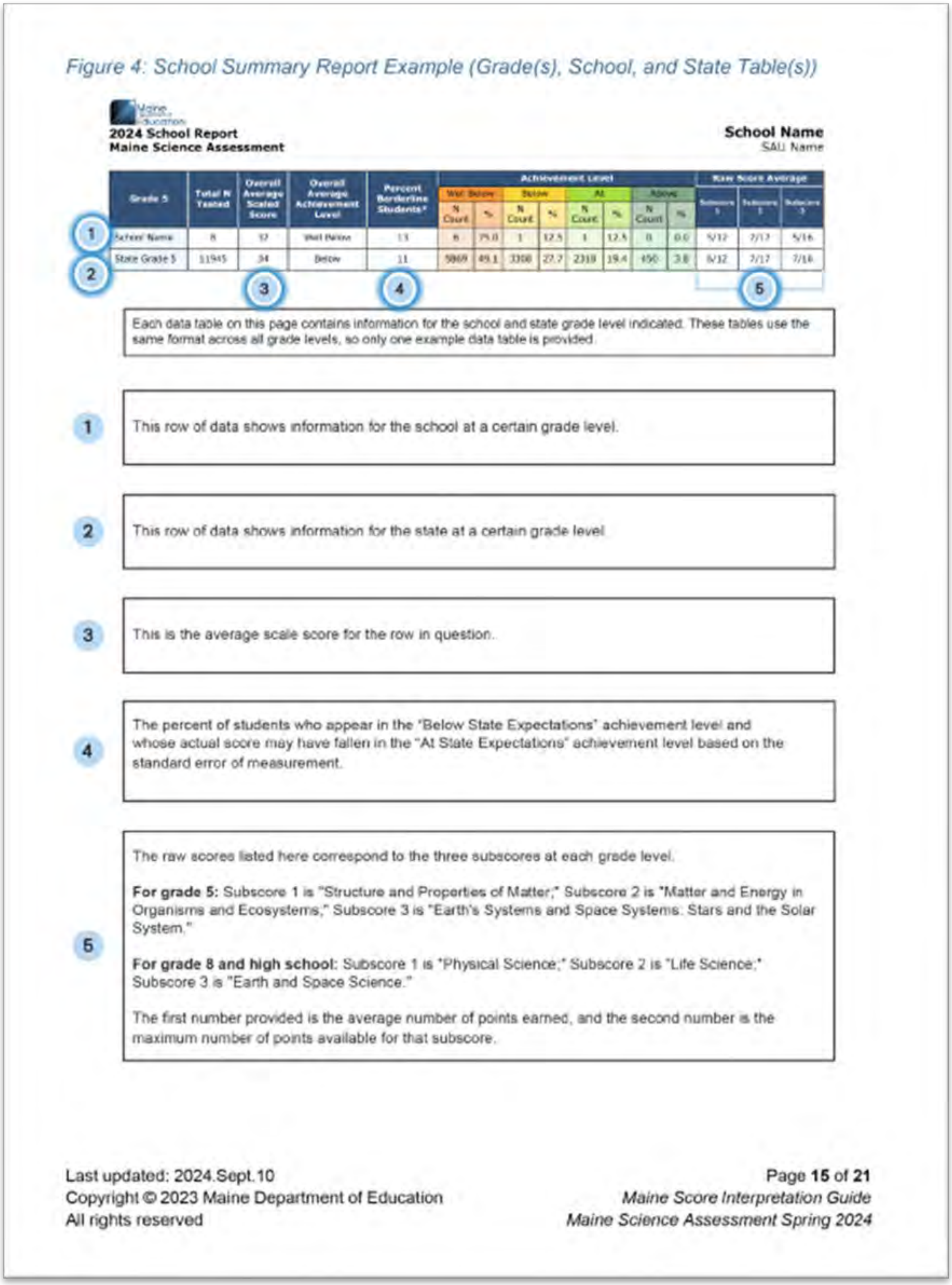


Figure 4: School Summary Report Example (Grade(s), School, and State Table(s))

Maine Education

2024 School Report
Maine Science Assessment

School Name
SAU Name

Grade 5	Total N Tested	Overall Average Scaled Score	Overall Average Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above				
					N Count	%	N Count	%	N Count	%	N Count	%	Subscore 1	Subscore 2	Subscore 3
School Name	8	17	Well Below	13	6	75.0	1	12.5	1	12.5	0	0.0	5/12	2/17	5/16
State Grade 5	11945	34	Below	11	5869	49.1	3308	27.7	2319	19.4	450	3.8	6/12	7/17	7/16

Each data table on this page contains information for the school and state grade level indicated. These tables use the same format across all grade levels, so only one example data table is provided.

1 This row of data shows information for the school at a certain grade level.

2 This row of data shows information for the state at a certain grade level.

3 This is the average scale score for the row in question.

4 The percent of students who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

5 The raw scores listed here correspond to the three subscores at each grade level.

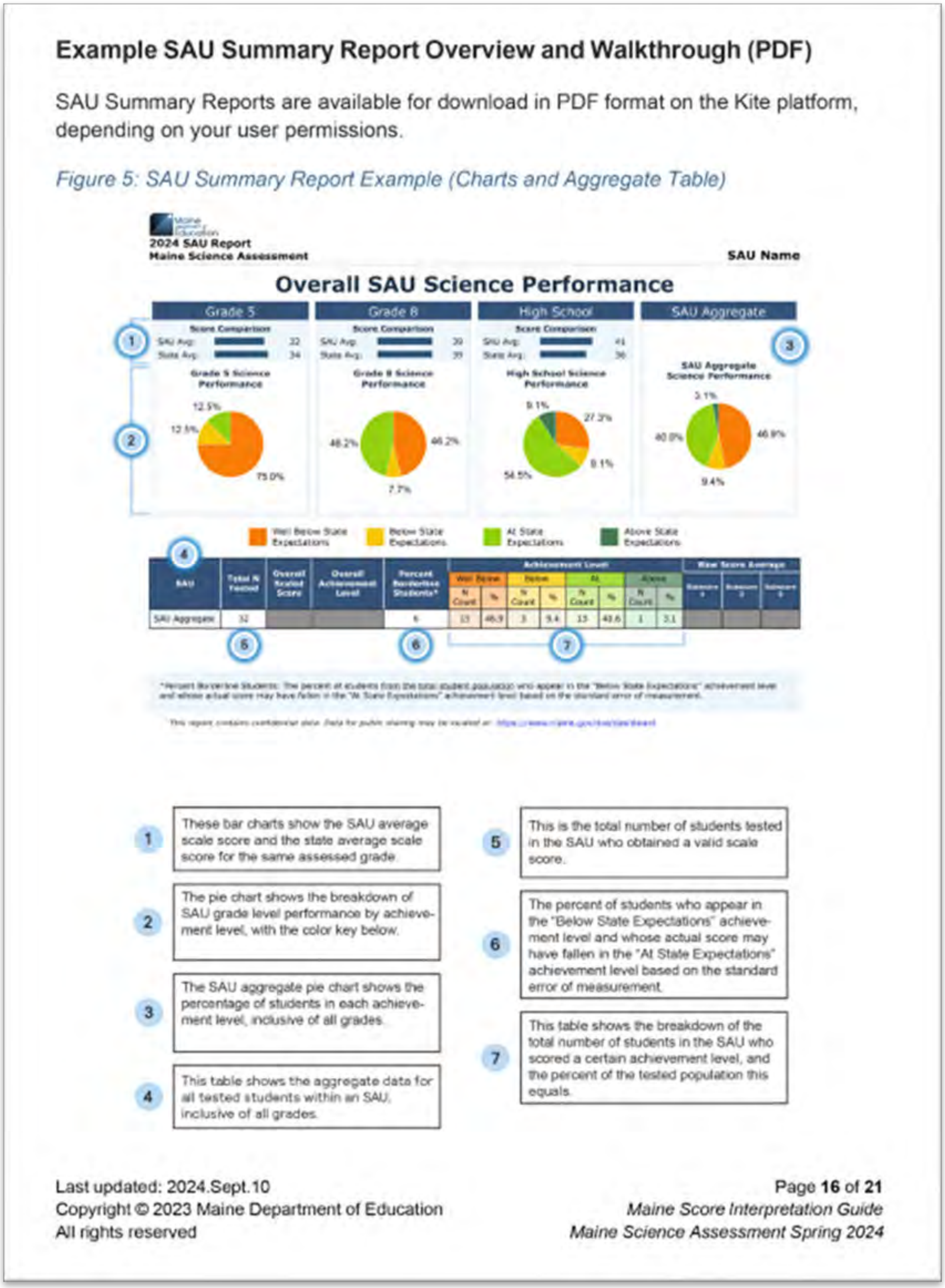
For grade 5: Subscore 1 is "Structure and Properties of Matter;" Subscore 2 is "Matter and Energy in Organisms and Ecosystems;" Subscore 3 is "Earth's Systems and Space Systems: Stars and the Solar System."

For grade 8 and high school: Subscore 1 is "Physical Science;" Subscore 2 is "Life Science;" Subscore 3 is "Earth and Space Science."

The first number provided is the average number of points earned, and the second number is the maximum number of points available for that subscore.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **15** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



Example SAU Summary Report Overview and Walkthrough (PDF)

SAU Summary Reports are available for download in PDF format on the Kite platform, depending on your user permissions.

Figure 5: SAU Summary Report Example (Charts and Aggregate Table)

Maine Department of Education
2024 SAU Report
Maine Science Assessment

SAU Name

Overall SAU Science Performance

SAU	Total N Tested	Overall Scaled Score	Overall Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above				
					N Count	%	N Count	%	N Count	%	N Count	%	Subscore 1	Subscore 2	Subscore 3
SAU Aggregate	32			6	15	46.9	3	9.4	13	40.6	1	3.1			

*Percent Borderline Students: The percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

This report contains confidential data. Data for public sharing may be located at: https://www.maine.gov/doe/dashboard

1. These bar charts show the SAU average scale score and the state average scale score for the same assessed grade.
2. The pie chart shows the breakdown of SAU grade level performance by achievement level, with the color key below.
3. The SAU aggregate pie chart shows the percentage of students in each achievement level, inclusive of all grades.
4. This table shows the aggregate data for all tested students within an SAU, inclusive of all grades.
5. This is the total number of students tested in the SAU who obtained a valid scale score.
6. The percent of students who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.
7. This table shows the breakdown of the total number of students in the SAU who scored a certain achievement level, and the percent of the tested population this equals.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **16** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024

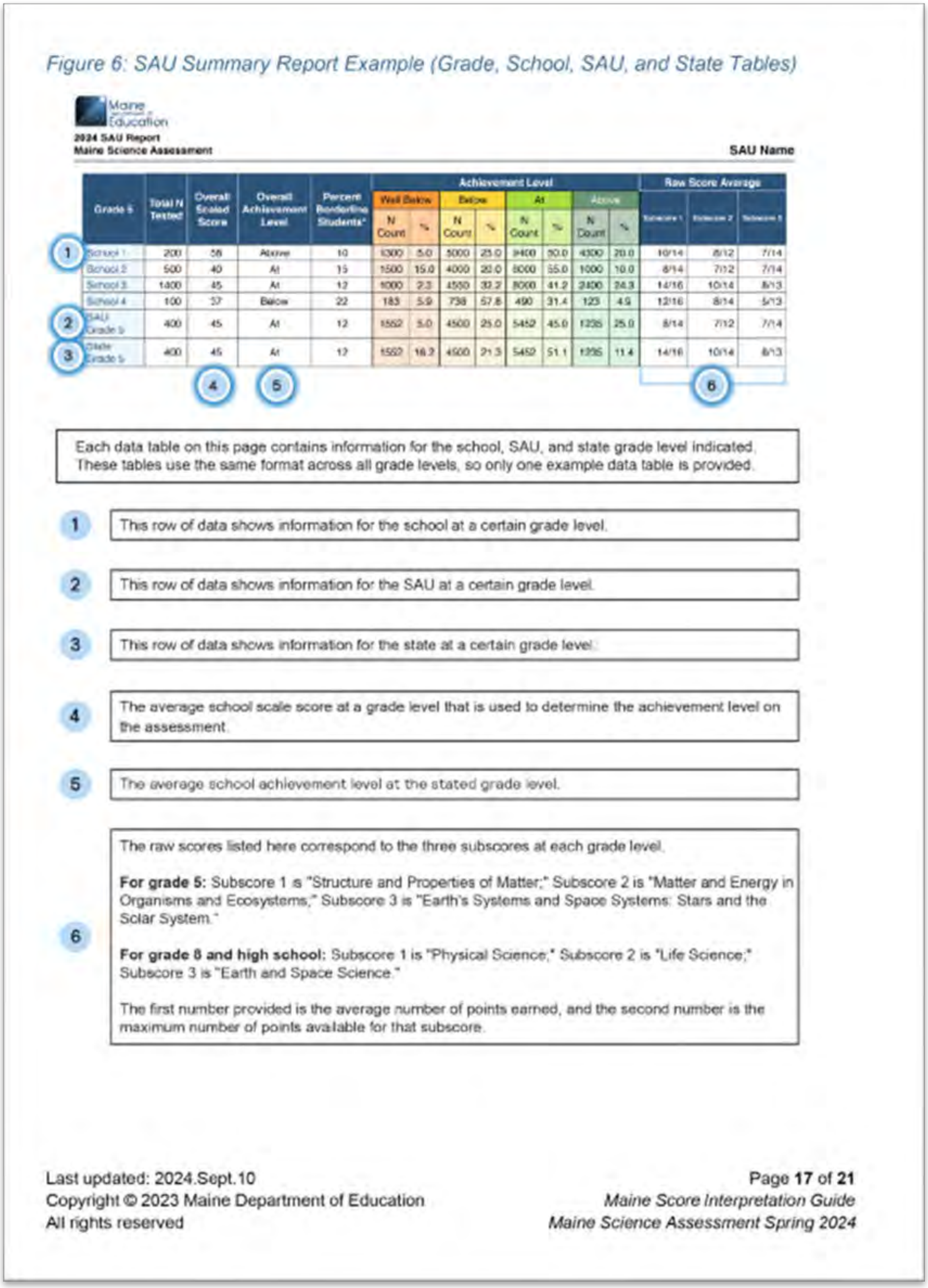


Figure 6: SAU Summary Report Example (Grade, School, SAU, and State Tables)

Maine Education
2024 SAU Report
Maine Science Assessment **SAU Name**

Grade 5	Total N Tested	Overall Scaled Score	Overall Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School 1	200	58	Above	10	1300	5.0	5000	25.0	9400	50.0	4300	20.0	10/14	8/12	7/14
School 2	500	40	At	15	1500	15.0	4000	20.0	8000	55.0	1000	10.0	8/14	7/12	7/14
School 3	1400	45	At	12	1000	2.3	4550	32.2	8000	41.2	2400	24.3	14/16	10/14	8/13
School 4	100	37	Below	22	183	5.9	738	57.8	490	31.4	123	4.9	12/16	8/14	5/13
SAU Grade 5	400	45	At	12	1552	5.0	4500	25.0	5452	45.0	1235	25.0	8/14	7/12	7/14
State Grade 5	400	45	At	12	1552	16.2	4500	21.3	5452	51.1	1235	11.4	14/16	10/14	8/13

Each data table on this page contains information for the school, SAU, and state grade level indicated. These tables use the same format across all grade levels, so only one example data table is provided.

1 This row of data shows information for the school at a certain grade level.

2 This row of data shows information for the SAU at a certain grade level.

3 This row of data shows information for the state at a certain grade level.

4 The average school scale score at a grade level that is used to determine the achievement level on the assessment.

5 The average school achievement level at the stated grade level.

6 The raw scores listed here correspond to the three subscores at each grade level.

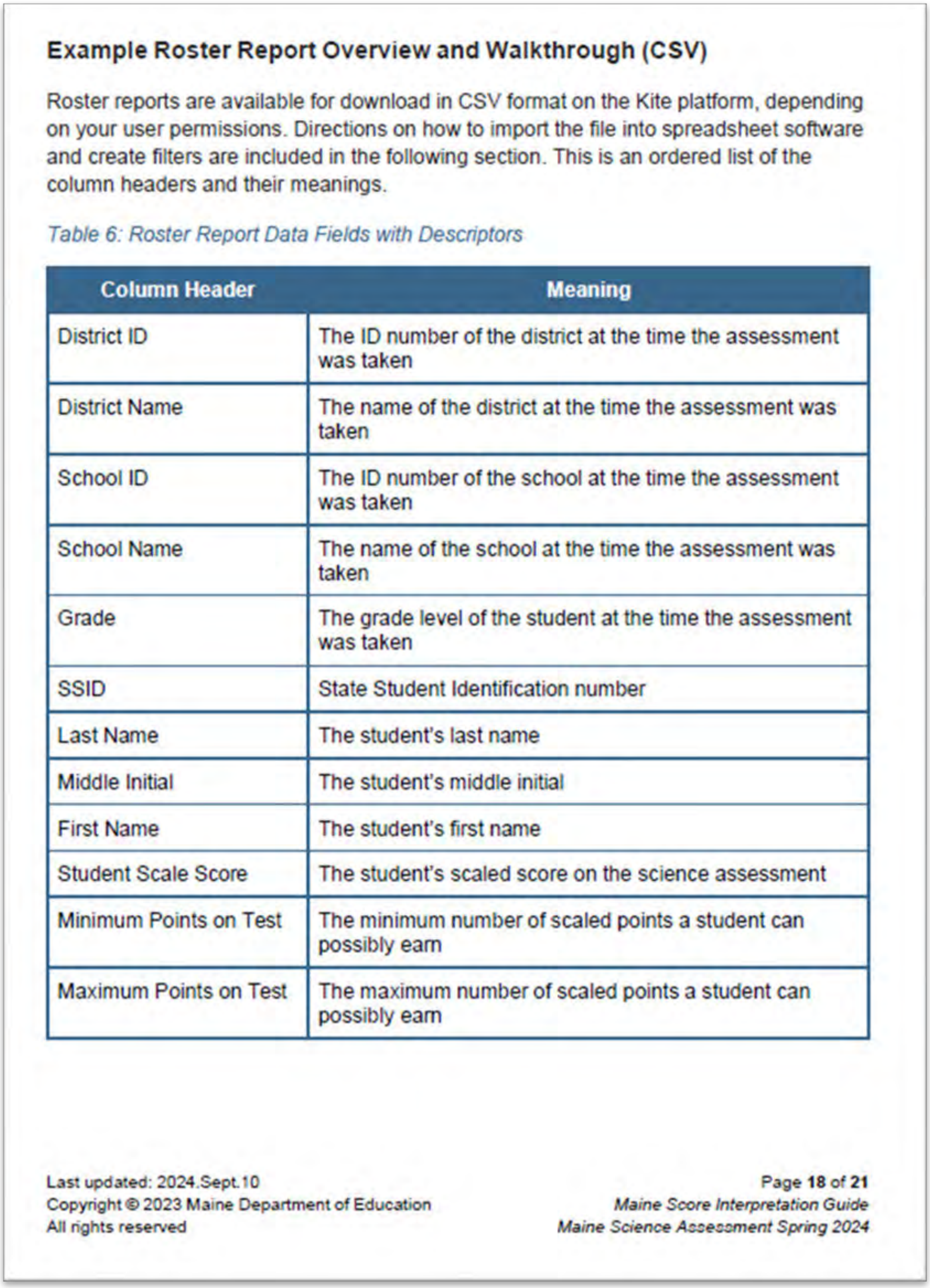
For grade 5: Subscore 1 is "Structure and Properties of Matter;" Subscore 2 is "Matter and Energy in Organisms and Ecosystems;" Subscore 3 is "Earth's Systems and Space Systems: Stars and the Solar System."

For grade 8 and high school: Subscore 1 is "Physical Science;" Subscore 2 is "Life Science;" Subscore 3 is "Earth and Space Science."

The first number provided is the average number of points earned, and the second number is the maximum number of points available for that subscore.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **17** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



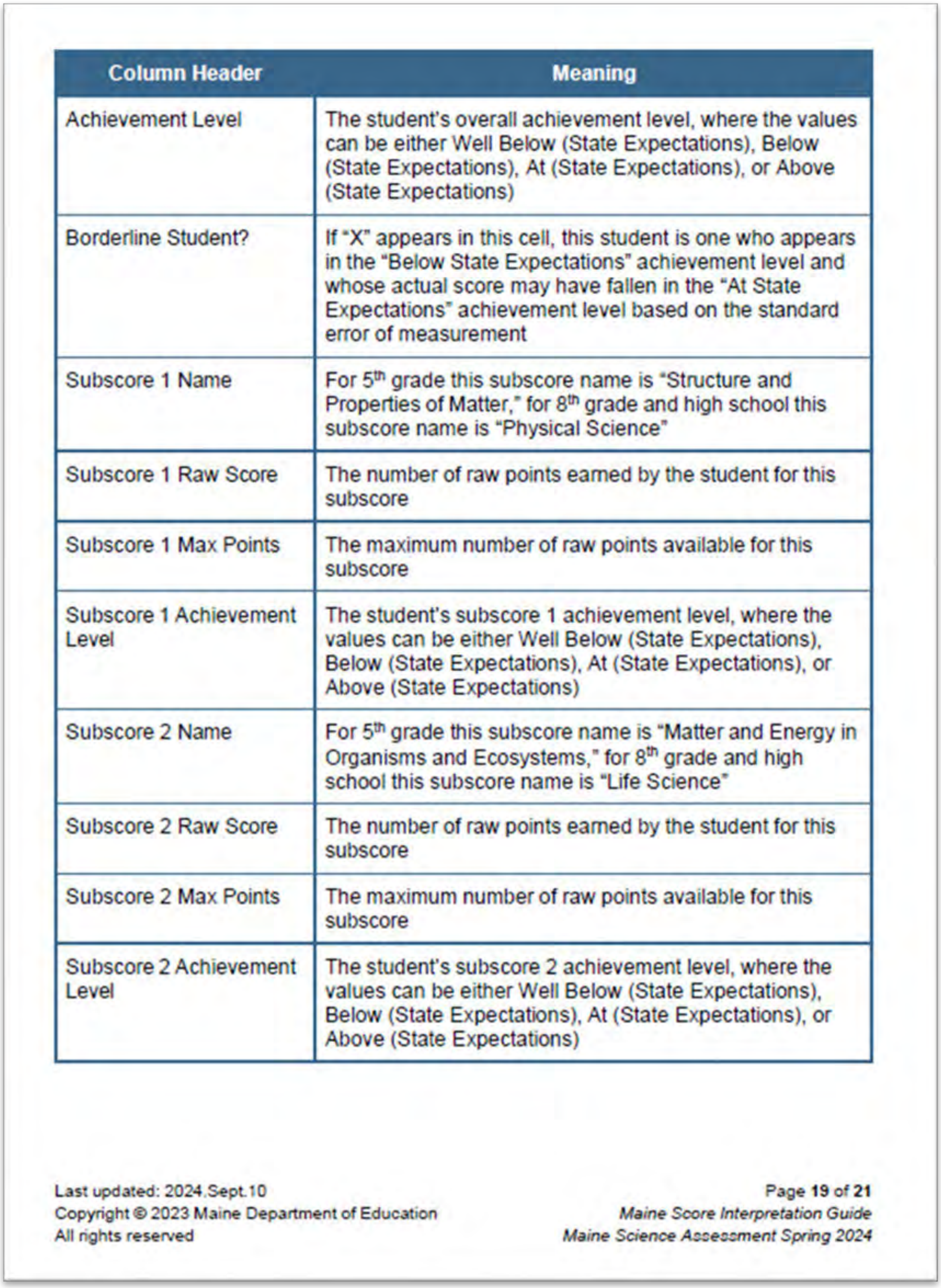
Example Roster Report Overview and Walkthrough (CSV)

Roster reports are available for download in CSV format on the Kite platform, depending on your user permissions. Directions on how to import the file into spreadsheet software and create filters are included in the following section. This is an ordered list of the column headers and their meanings.

Table 6: Roster Report Data Fields with Descriptors

Column Header	Meaning
District ID	The ID number of the district at the time the assessment was taken
District Name	The name of the district at the time the assessment was taken
School ID	The ID number of the school at the time the assessment was taken
School Name	The name of the school at the time the assessment was taken
Grade	The grade level of the student at the time the assessment was taken
SSID	State Student Identification number
Last Name	The student's last name
Middle Initial	The student's middle initial
First Name	The student's first name
Student Scale Score	The student's scaled score on the science assessment
Minimum Points on Test	The minimum number of scaled points a student can possibly eam
Maximum Points on Test	The maximum number of scaled points a student can possibly eam

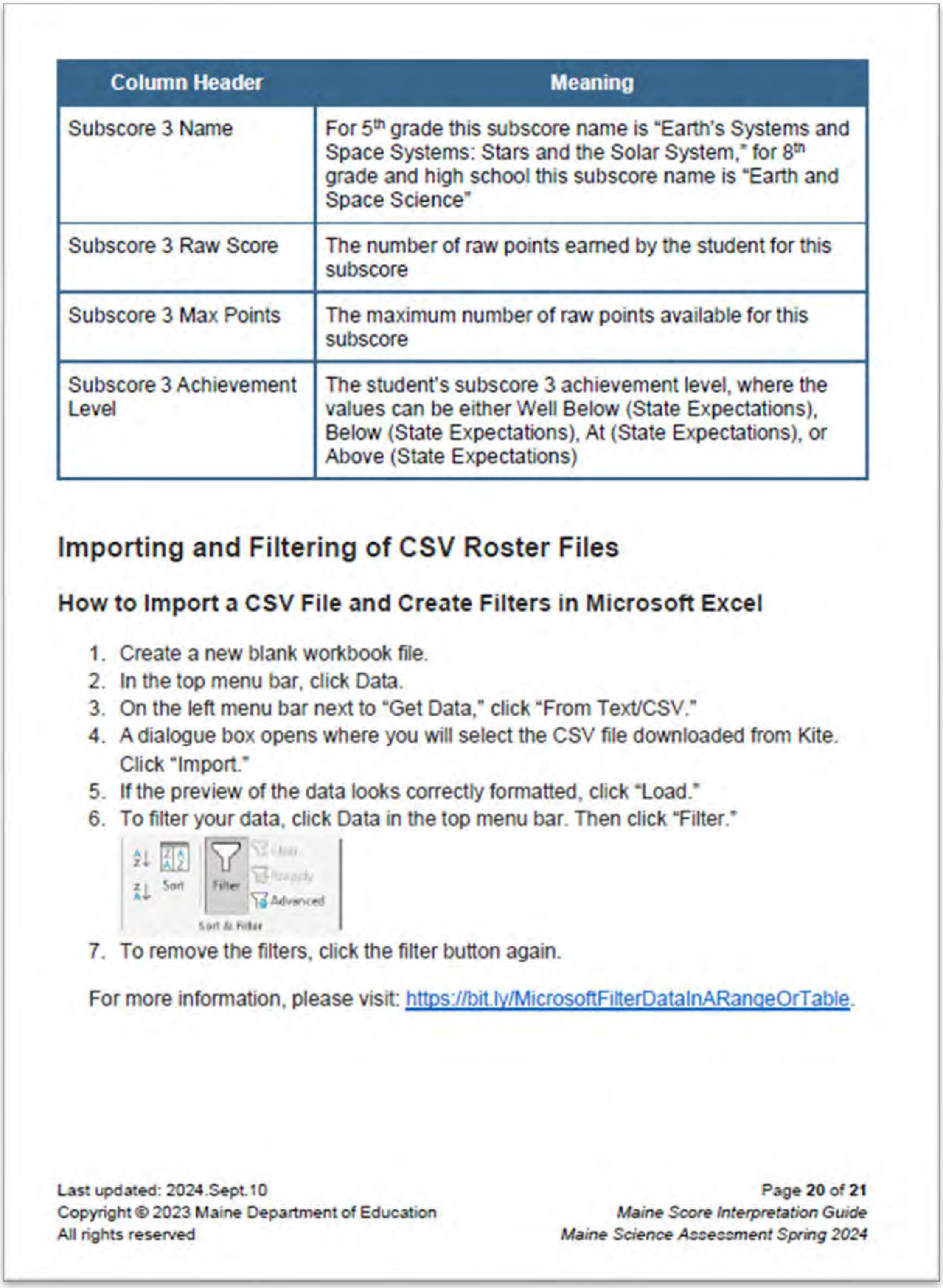
Last updated: 2024.Sept.10



Column Header	Meaning
Achievement Level	The student's overall achievement level, where the values can be either Well Below (State Expectations), Below (State Expectations), At (State Expectations), or Above (State Expectations)
Borderline Student?	If "X" appears in this cell, this student is one who appears in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement
Subscore 1 Name	For 5th grade this subscore name is "Structure and Properties of Matter," for 8th grade and high school this subscore name is "Physical Science"
Subscore 1 Raw Score	The number of raw points earned by the student for this subscore
Subscore 1 Max Points	The maximum number of raw points available for this subscore
Subscore 1 Achievement Level	The student's subscore 1 achievement level, where the values can be either Well Below (State Expectations), Below (State Expectations), At (State Expectations), or Above (State Expectations)
Subscore 2 Name	For 5th grade this subscore name is "Matter and Energy in Organisms and Ecosystems," for 8th grade and high school this subscore name is "Life Science"
Subscore 2 Raw Score	The number of raw points earned by the student for this subscore
Subscore 2 Max Points	The maximum number of raw points available for this subscore
Subscore 2 Achievement Level	The student's subscore 2 achievement level, where the values can be either Well Below (State Expectations), Below (State Expectations), At (State Expectations), or Above (State Expectations)

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **19** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024



Column Header	Meaning
Subscore 3 Name	For 5th grade this subscore name is "Earth's Systems and Space Systems: Stars and the Solar System," for 8th grade and high school this subscore name is "Earth and Space Science"
Subscore 3 Raw Score	The number of raw points earned by the student for this subscore
Subscore 3 Max Points	The maximum number of raw points available for this subscore
Subscore 3 Achievement Level	The student's subscore 3 achievement level, where the values can be either Well Below (State Expectations), Below (State Expectations), At (State Expectations), or Above (State Expectations)

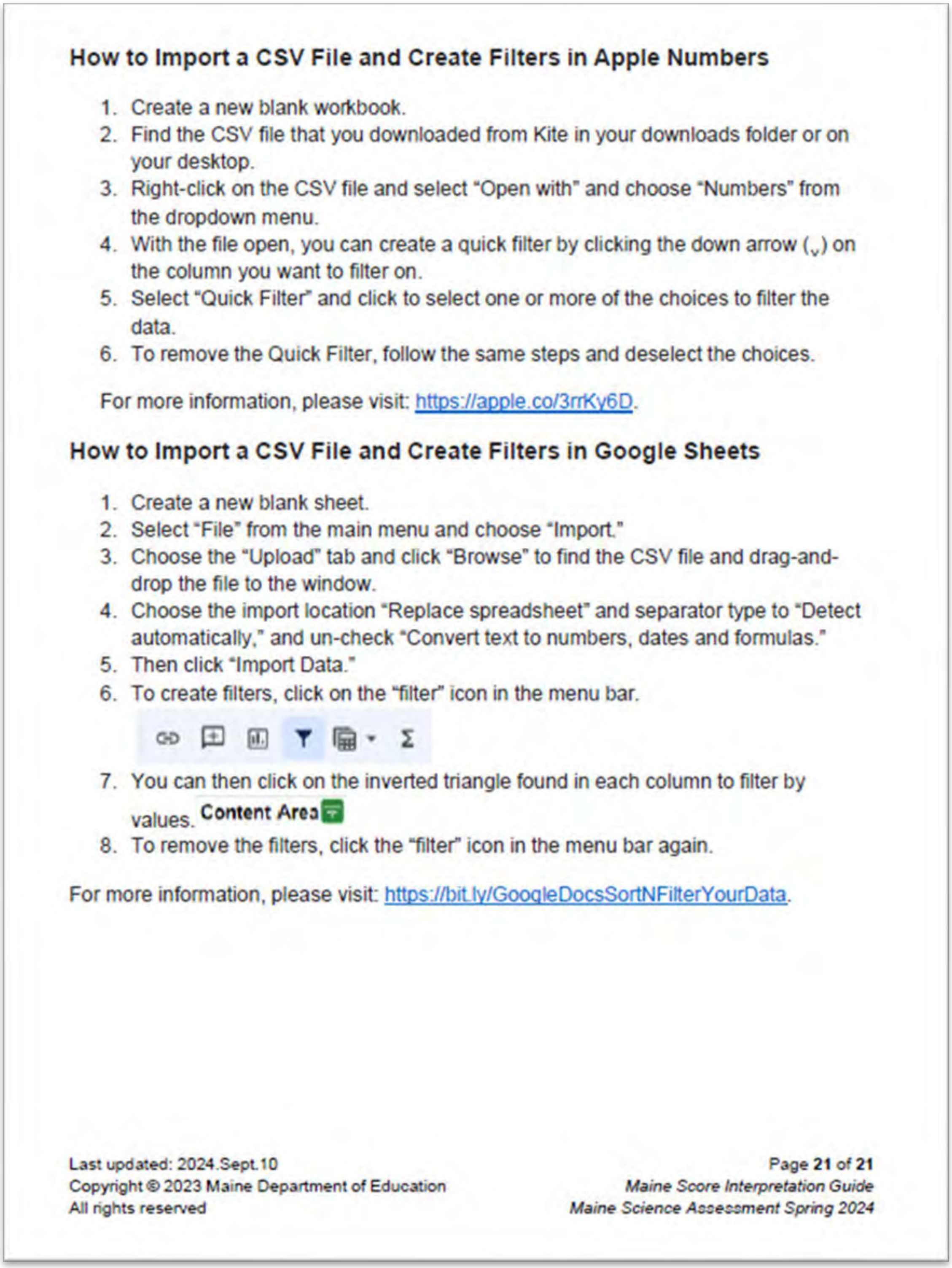
Importing and Filtering of CSV Roster Files

How to Import a CSV File and Create Filters in Microsoft Excel

1. Create a new blank workbook file.
2. In the top menu bar, click Data.
3. On the left menu bar next to "Get Data," click "From Text/CSV."
4. A dialogue box opens where you will select the CSV file downloaded from Kite. Click "Import."
5. If the preview of the data looks correctly formatted, click "Load."
6. To filter your data, click Data in the top menu bar. Then click "Filter."

7. To remove the filters, click the filter button again.

For more information, please visit: https://bit.ly/MicrosoftFilterDataInARangeOrTable.



How to Import a CSV File and Create Filters in Apple Numbers

1. Create a new blank workbook.
2. Find the CSV file that you downloaded from Kite in your downloads folder or on your desktop.
3. Right-click on the CSV file and select "Open with" and choose "Numbers" from the dropdown menu.
4. With the file open, you can create a quick filter by clicking the down arrow (˅) on the column you want to filter on.
5. Select "Quick Filter" and click to select one or more of the choices to filter the data.
6. To remove the Quick Filter, follow the same steps and deselect the choices.

For more information, please visit: https://apple.co/3rrKy6D.

How to Import a CSV File and Create Filters in Google Sheets

1. Create a new blank sheet.
2. Select "File" from the main menu and choose "Import."
3. Choose the "Upload" tab and click "Browse" to find the CSV file and drag-and-drop the file to the window.
4. Choose the import location "Replace spreadsheet" and separator type to "Detect automatically," and un-check "Convert text to numbers, dates and formulas."
5. Then click "Import Data."
6. To create filters, click on the "filter" icon in the menu bar.
7. You can then click on the inverted triangle found in each column to filter by values. Content Area
8. To remove the filters, click the "filter" icon in the menu bar again.

For more information, please visit: https://bit.ly/GoogleDocsSortNFilterYourData.

Last updated: 2024.Sept.10
Copyright © 2023 Maine Department of Education
All rights reserved

Page **21** of **21**
Maine Score Interpretation Guide
Maine Science Assessment Spring 2024

Appendix F. Student Population[7]

Table 56. Grade 5 Summary of Participation by Demographic Category

Description	Count	Percent
All Students	12,173	100.00
Male	6,300	51.75
Female	5,873	48.25
Hispanic or Latino	416	3.42
American Indian or Alaskan Native	103	0.85
Asian	168	1.38
Black or African American	598	4.91
Native Hawaiian or Other Pacific Islander	—	—
White	10,409	85.51
Two or More races	463	3.80
English Speaking Students	11,461	94.15
Multilingual Learners: Currently receiving EL services	526	4.32
Former EL student - monitoring year 1	152	1.25
Former EL student - monitoring year 2	—	—
Former EL student - monitoring year 3	—	—
IEP: All Other Students	9,478	77.86
Students with an IEP	2,695	22.14
SES: All Other Students	7,330	60.22
Economically Disadvantaged Students	4,843	39.78
Migrant: All Other Students	12,166	99.94
Migrant Students	—	—
Plan 504	672	5.52
Plan 504: All Other Students	11,501	94.48

[7] Data is not presented for groups with less than 25 students.

Table 57. Grade 8 Summary of Participation by Demographic Category

Description	Count	Percent
All Students	12,201	100.00
Male	6,358	52.11
Female	5,843	47.89
Hispanic or Latino	408	3.34
American Indian or Alaskan Native	94	0.77
Asian	159	1.30
Black or African American	606	4.97
Native Hawaiian or Other Pacific Islander	—	—
White	10,493	86.00
Two or More races	425	3.48
English Speaking Students	11,552	94.68
Multilingual Learners: Currently receiving EL services	488	4.00
Former EL student - monitoring year 1	—	—
Former EL student - monitoring year 2	29	0.24
Former EL student - monitoring year 3	111	0.91
IEP: All Other Students	9,830	80.57
Students with an IEP	2,371	19.43
SES: All Other Students	7,556	61.93
Economically Disadvantaged Students	4,645	38.07
Migrant: All Other Students	12,190	99.91
Migrant Students	—	—
Plan 504	910	7.46
Plan 504: All Other Students	11,291	92.54

Table 58. High School Summary of Participation by Demographic Category

Description	Count	Percent
All Students	11,574	100.00
Male	5,973	51.61
Female	5,601	48.39
Hispanic or Latino	400	3.46
American Indian or Alaskan Native	76	0.66
Asian	219	1.89
Black or African American	537	4.64
Native Hawaiian or Other Pacific Islander	—	—
White	9,938	85.86
Two or More races	388	3.35
English Speaking Students	11,084	95.77
Multilingual Learners: Currently receiving EL services	426	3.68
Former EL student - monitoring year 1	36	0.31
Former EL student - monitoring year 2	—	—
Former EL student - monitoring year 3	—	—
IEP: All Other Students	9,789	84.58
Students with an IEP	1,785	15.42
SES: All Other Students	8,074	69.76
Economically Disadvantaged Students	3,500	30.24
Migrant: All Other Students	11,565	99.92
Migrant Students	—	—
Plan 504	1,011	8.74
Plan 504: All Other Students	10,563	91.26

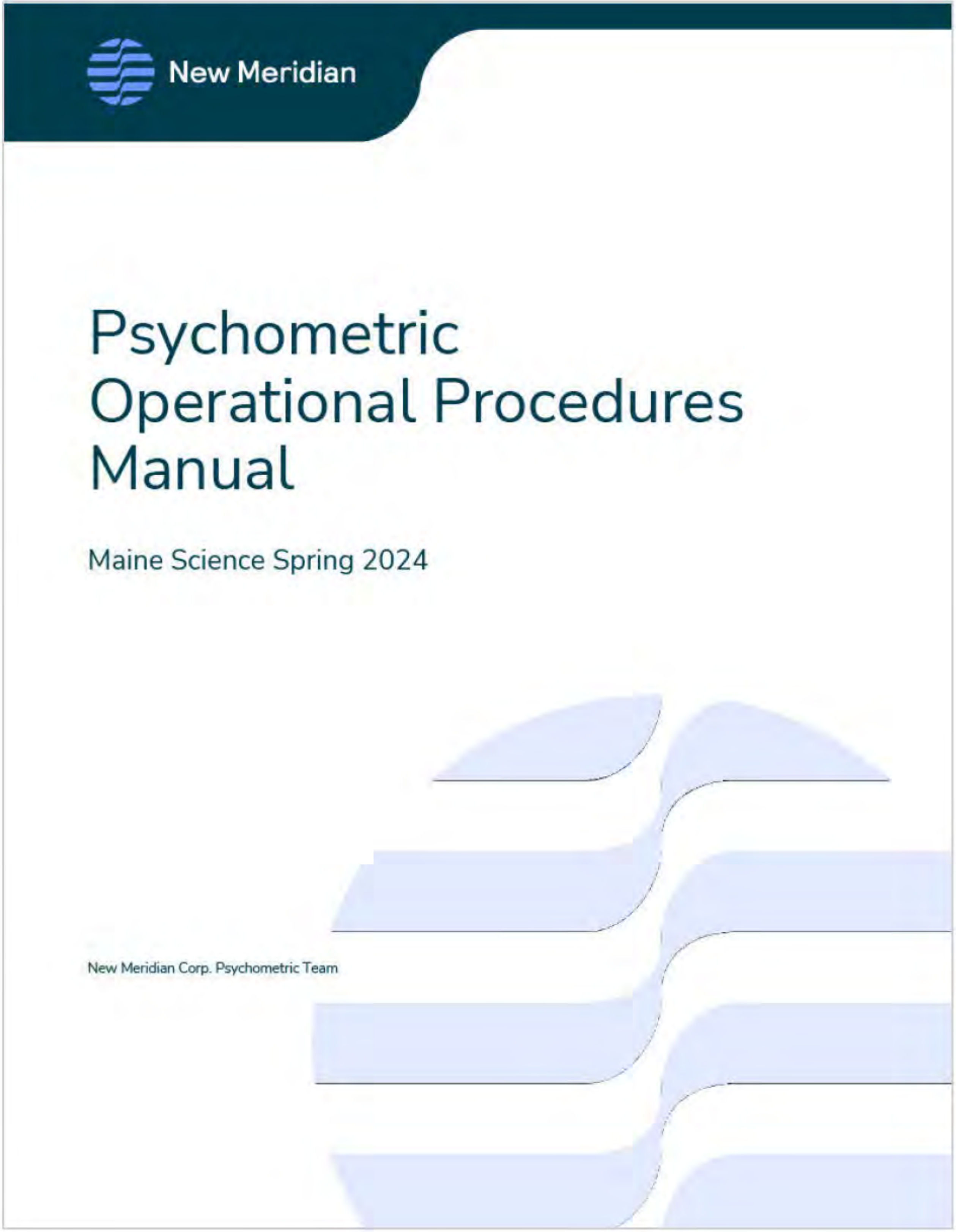
Appendix H. Inter-Rater Agreement

Table 59. Inter-Rater Agreement by Grade

Grade	ItemCode	Total	Read-behinds	Read-behinds Percent	Resolutions	Resolutions Percent	Exact Agreement	Exact/Adjacent Agreement
05	[redacted]	12,127	1,503	12.4	61	0.5	95	100
05	[redacted]	12,138	1,724	14.2	97	0.8	92	99.8
08	[redacted]	12,179	1,213	10	167	1.4	86.3	99.8
08	[redacted]	12,197	1,233	10.1	90	0.7	92.6	99.9
08	[redacted]	12,028	1,285	10.7	80	0.7	93.3	100
HS	[redacted]	11,563	1,420	12.3	78	0.7	93.3	99.5
HS	[redacted]	11,406	1,366	12	109	1	90.4	100

Appendix I. Psychometric Operational Procedures Manual

(Document begins on next page.)

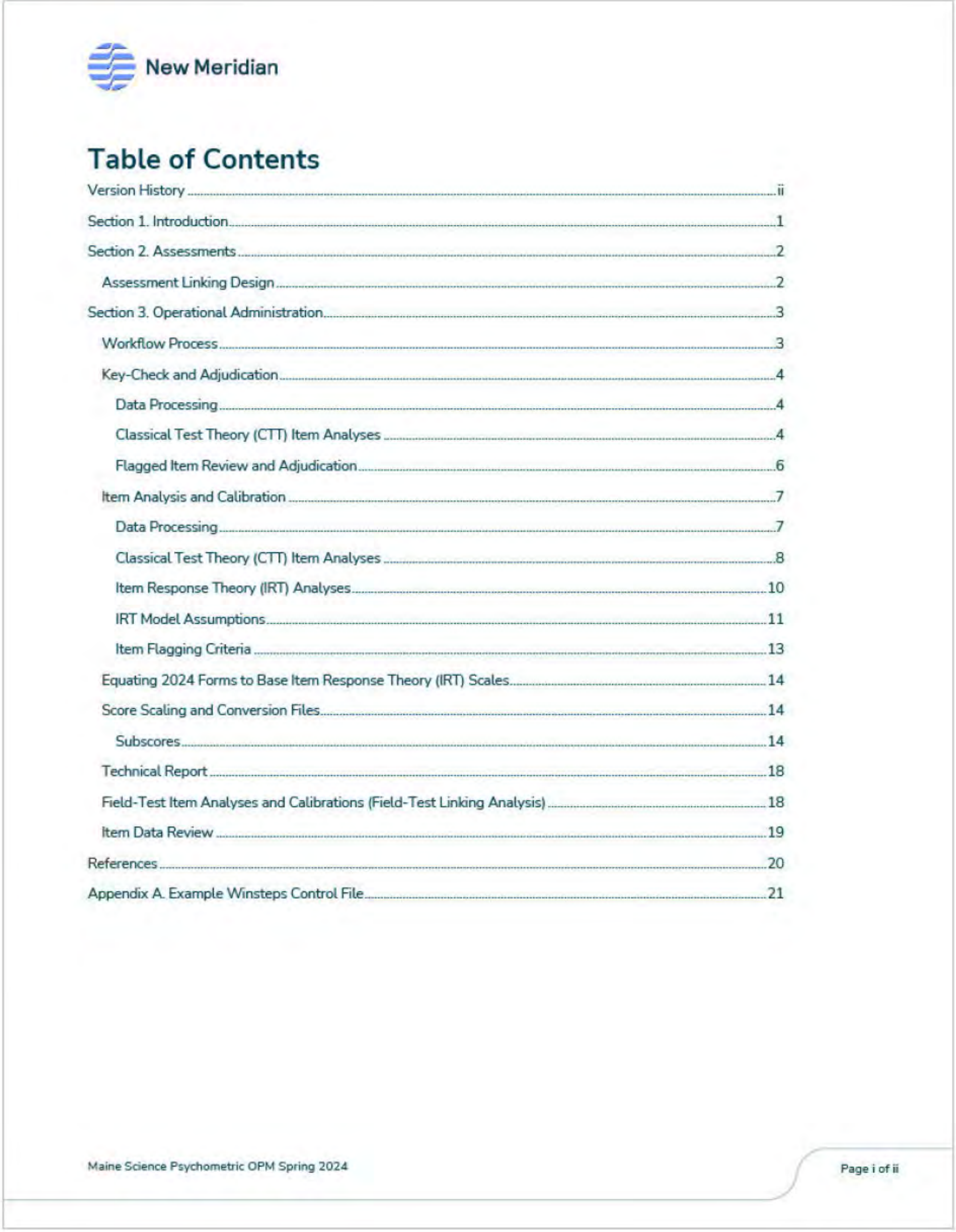


New Meridian

Psychometric Operational Procedures Manual

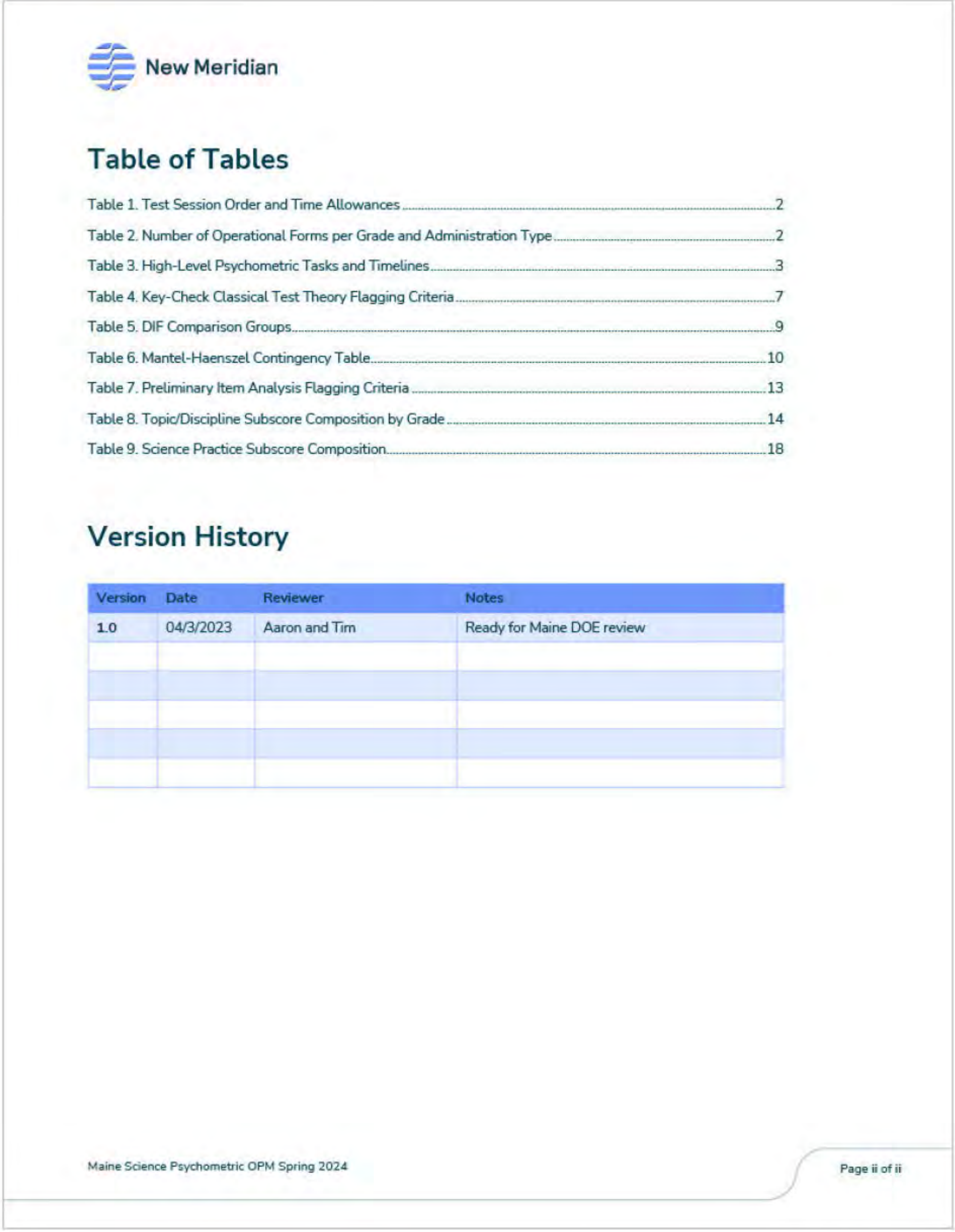
Maine Science Spring 2024

New Meridian Corp. Psychometric Team



New Meridian

Table of Contents

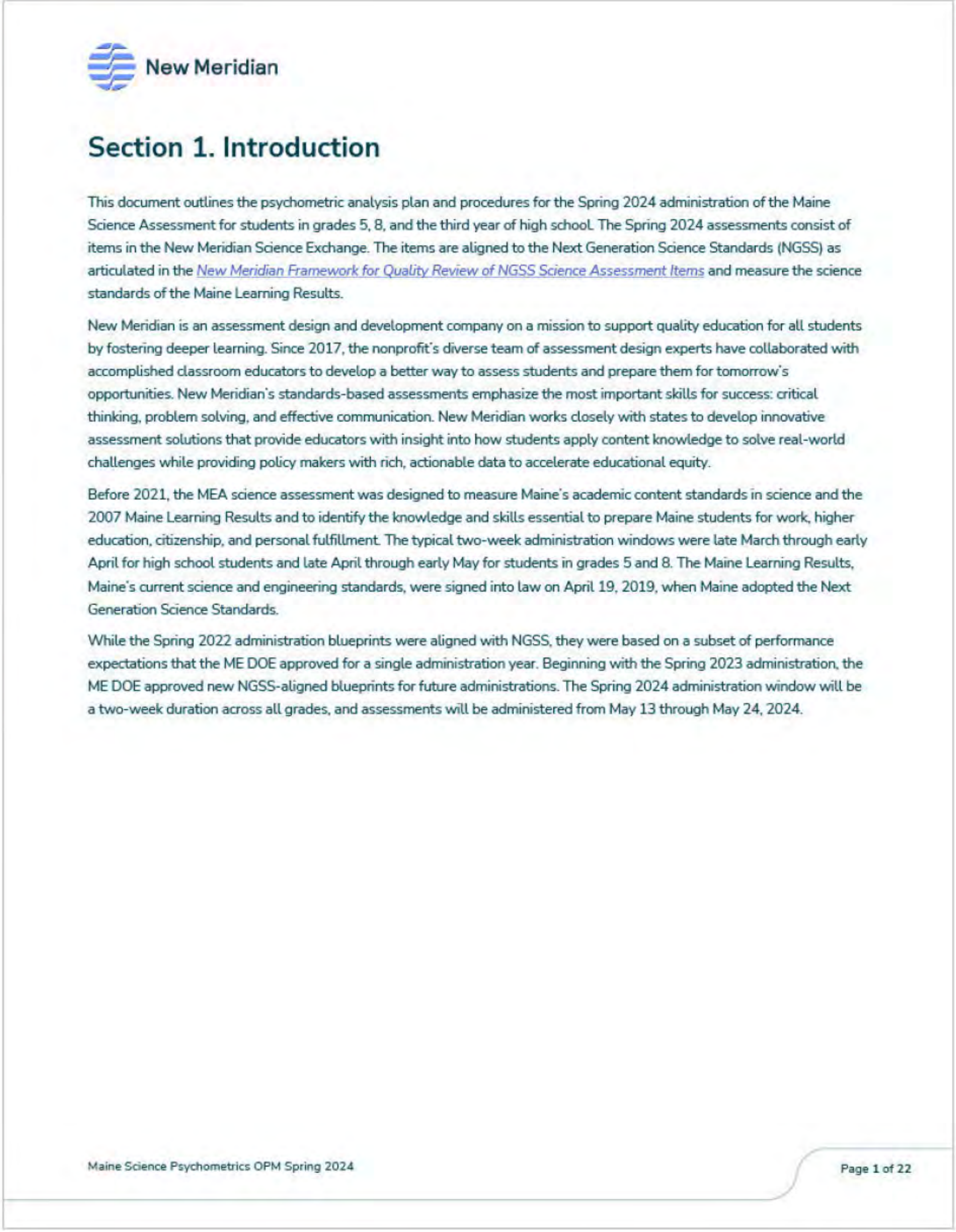


New Meridian

Table of Tables

Version History

Version	Date	Reviewer	Notes
1.0	04/3/2023	Aaron and Tim	Ready for Maine DOE review



New Meridian

Section 1. Introduction

This document outlines the psychometric analysis plan and procedures for the Spring 2024 administration of the Maine Science Assessment for students in grades 5, 8, and the third year of high school. The Spring 2024 assessments consist of items in the New Meridian Science Exchange. The items are aligned to the Next Generation Science Standards (NGSS) as articulated in the *New Meridian Framework for Quality Review of NGSS Science Assessment Items* and measure the science standards of the Maine Learning Results.

New Meridian is an assessment design and development company on a mission to support quality education for all students by fostering deeper learning. Since 2017, the nonprofit's diverse team of assessment design experts have collaborated with accomplished classroom educators to develop a better way to assess students and prepare them for tomorrow's opportunities. New Meridian's standards-based assessments emphasize the most important skills for success: critical thinking, problem solving, and effective communication. New Meridian works closely with states to develop innovative assessment solutions that provide educators with insight into how students apply content knowledge to solve real-world challenges while providing policy makers with rich, actionable data to accelerate educational equity.

Before 2021, the MEA science assessment was designed to measure Maine's academic content standards in science and the 2007 Maine Learning Results and to identify the knowledge and skills essential to prepare Maine students for work, higher education, citizenship, and personal fulfillment. The typical two-week administration windows were late March through early April for high school students and late April through early May for students in grades 5 and 8. The Maine Learning Results, Maine's current science and engineering standards, were signed into law on April 19, 2019, when Maine adopted the Next Generation Science Standards.

While the Spring 2022 administration blueprints were aligned with NGSS, they were based on a subset of performance expectations that the ME DOE approved for a single administration year. Beginning with the Spring 2023 administration, the ME DOE approved new NGSS-aligned blueprints for future administrations. The Spring 2024 administration window will be a two-week duration across all grades, and assessments will be administered from May 13 through May 24, 2024.

Section 2. Assessments

The science assessments consist of dichotomous and polytomous items that are machine- or hand-scored. Items may be stand-alone or a part of a "cluster" that has a common phenomenon or stimulus. The assessments contain selected-response (i.e., multiple-choice and multiple-select), brief constructed-response, and technology-enhanced items (e.g., inline choice, order, hot spot, match table grid, and graphic gap match).

The assessments are administered online as computer-based tests (CBT) with a wide range of accessibility features for all students (e.g., color scheme, font size, and zoom) and as paper-based tests (PBT) with accommodations for students with disabilities (e.g., braille and large-print [LP] versions of the assessment, as well as response accommodations that allow students to respond to assessment items using different formats). Each assessment is composed of three administration sessions followed by a session for the student questionnaire. Table 1 provides the time allocation for the assessment administration at grades 5 and 8, and high school.

Table 1. Test Session Order and Time Allowances

Session	Total Time
Session 1: Science Assessment	60 minutes
Session 2: Science Assessment	60 minutes
Session 3: Science Assessment	60 minutes
Session 4: Student Questionnaire	15–20 minutes
TOTAL (not including distribution/instructions)	200 minutes

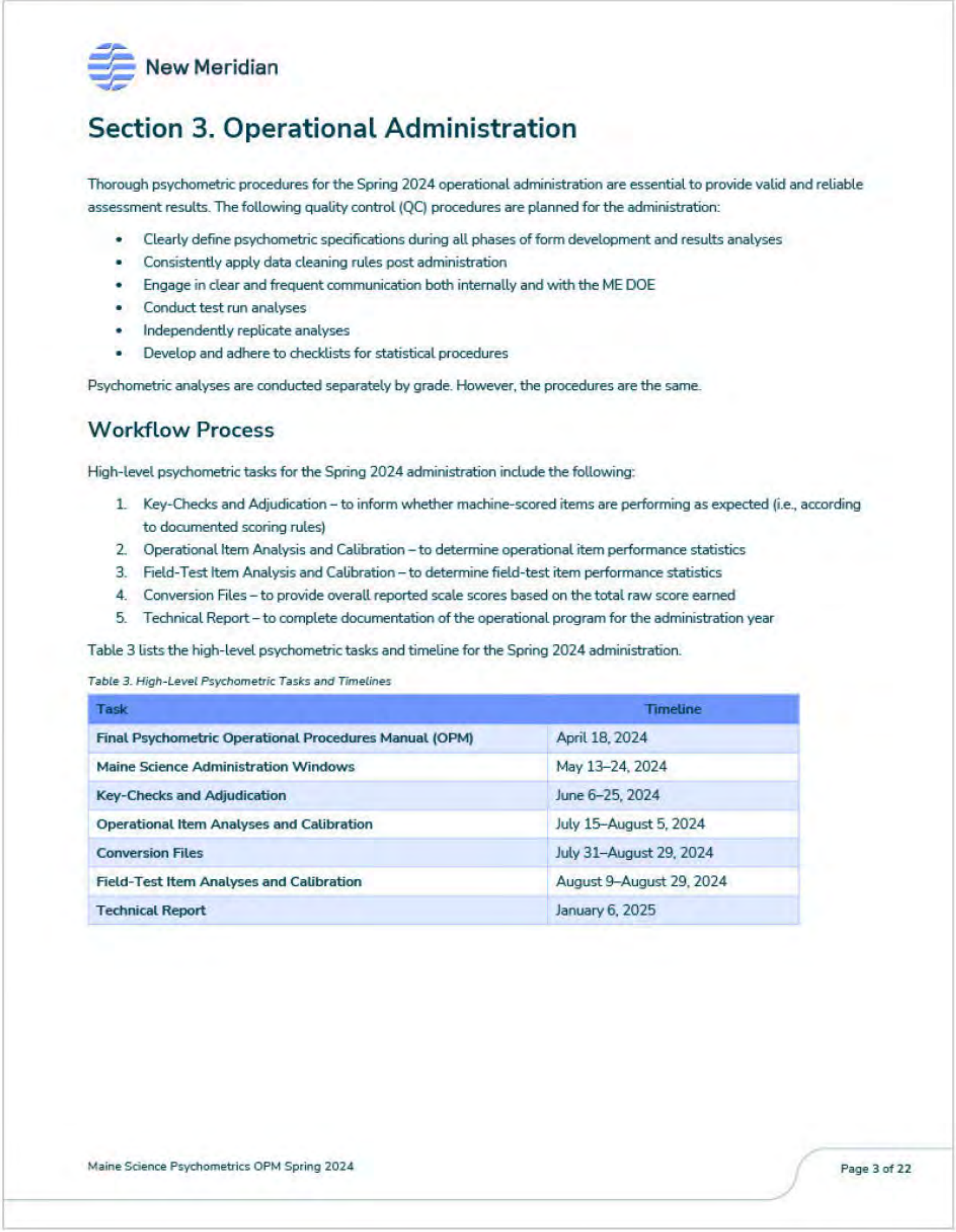
Assessment Linking Design

The Spring 2024 online assessments only consist of one operational base form per grade. Some technology-enhanced items cannot be administered efficiently to students requiring a paper-based form (i.e., large print and braille). Therefore, paper friendly versions of technology-enhanced items are included on paper-based forms. The paper friendly items assess the same standards as their TEI counterparts, with only the item interaction changing (i.e., instead of the students dragging to place steps in the correct order, they write numbers to show the order in which the steps occur).

Table 2 lists the number of forms per grade and administration type. The volume of paper-based assessments is expected to be very small compared to online testing.

Table 2. Number of Operational Forms per Grade and Administration Type

Grade	Online	Paper
5	1	1
8	1	1
HS	1	1



New Meridian

Section 3. Operational Administration

Thorough psychometric procedures for the Spring 2024 operational administration are essential to provide valid and reliable assessment results. The following quality control (QC) procedures are planned for the administration:

- Clearly define psychometric specifications during all phases of form development and results analyses
- Consistently apply data cleaning rules post administration
- Engage in clear and frequent communication both internally and with the ME DOE
- Conduct test run analyses
- Independently replicate analyses
- Develop and adhere to checklists for statistical procedures

Psychometric analyses are conducted separately by grade. However, the procedures are the same.

Workflow Process

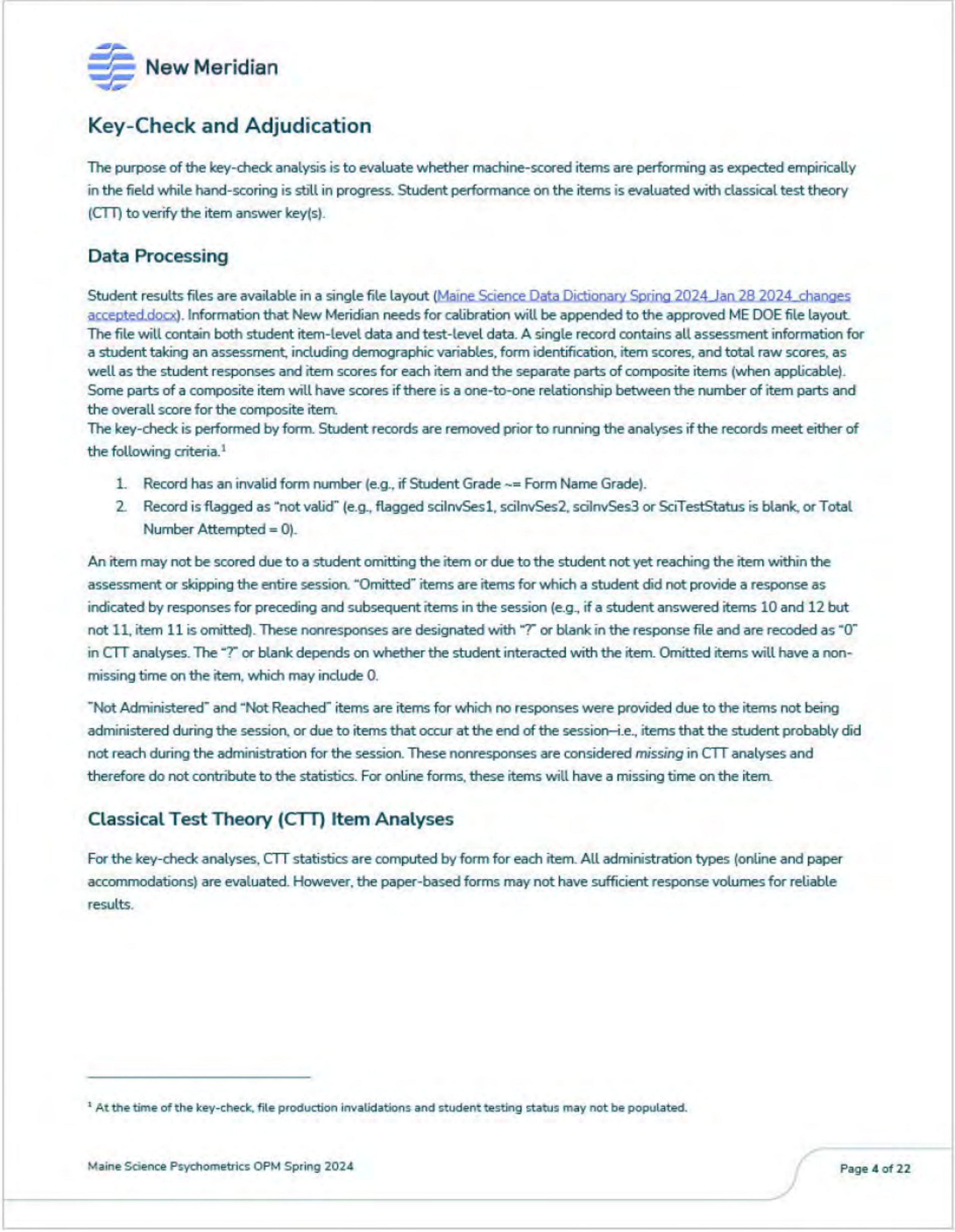
High-level psychometric tasks for the Spring 2024 administration include the following:

1. Key-Checks and Adjudication – to inform whether machine-scored items are performing as expected (i.e., according to documented scoring rules)
2. Operational Item Analysis and Calibration – to determine operational item performance statistics
3. Field-Test Item Analysis and Calibration – to determine field-test item performance statistics
4. Conversion Files – to provide overall reported scale scores based on the total raw score earned
5. Technical Report – to complete documentation of the operational program for the administration year

Table 3 lists the high-level psychometric tasks and timeline for the Spring 2024 administration.

Table 3. High-Level Psychometric Tasks and Timelines

Task	Timeline
Final Psychometric Operational Procedures Manual (OPM)	April 18, 2024
Maine Science Administration Windows	May 13–24, 2024
Key-Checks and Adjudication	June 6–25, 2024
Operational Item Analyses and Calibration	July 15–August 5, 2024
Conversion Files	July 31–August 29, 2024
Field-Test Item Analyses and Calibration	August 9–August 29, 2024
Technical Report	January 6, 2025



New Meridian

Key-Check and Adjudication

The purpose of the key-check analysis is to evaluate whether machine-scored items are performing as expected empirically in the field while hand-scoring is still in progress. Student performance on the items is evaluated with classical test theory (CTT) to verify the item answer key(s).

Data Processing

Student results files are available in a single file layout (Maine Science Data Dictionary Spring 2024_Jan 28 2024_changes accepted.docx). Information that New Meridian needs for calibration will be appended to the approved ME DOE file layout. The file will contain both student item-level data and test-level data. A single record contains all assessment information for a student taking an assessment, including demographic variables, form identification, item scores, and total raw scores, as well as the student responses and item scores for each item and the separate parts of composite items (when applicable). Some parts of a composite item will have scores if there is a one-to-one relationship between the number of item parts and the overall score for the composite item.

The key-check is performed by form. Student records are removed prior to running the analyses if the records meet either of the following criteria.[1]

1. Record has an invalid form number (e.g., if Student Grade ~= Form Name Grade).
2. Record is flagged as "not valid" (e.g., flagged sciInvSes1, sciInvSes2, sciInvSes3 or SciTestStatus is blank, or Total Number Attempted = 0).

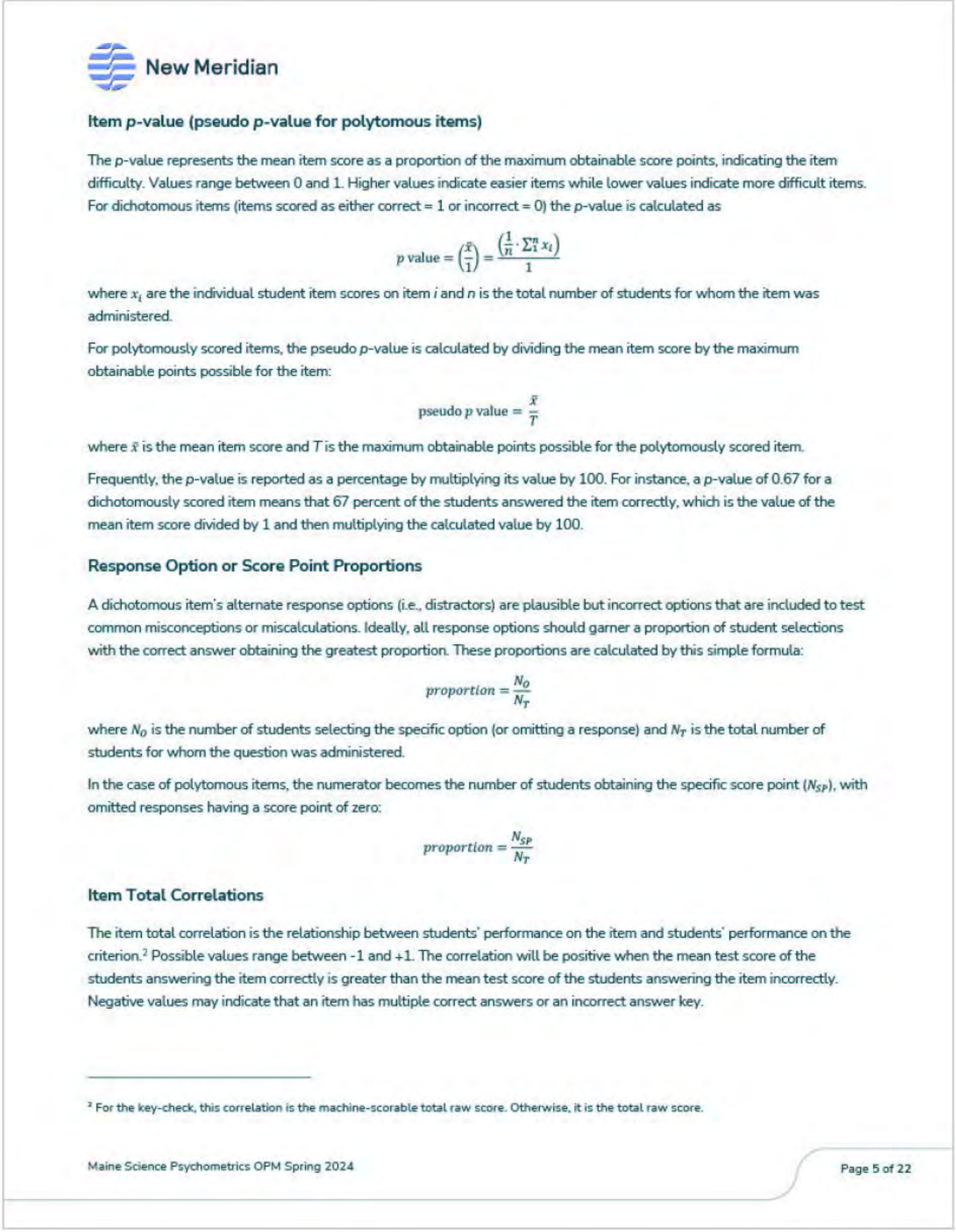
An item may not be scored due to a student omitting the item or due to the student not yet reaching the item within the assessment or skipping the entire session. "Omitted" items are items for which a student did not provide a response as indicated by responses for preceding and subsequent items in the session (e.g., if a student answered items 10 and 12 but not 11, item 11 is omitted). These nonresponses are designated with "?" or blank in the response file and are recoded as "0" in CTT analyses. The "?" or blank depends on whether the student interacted with the item. Omitted items will have a non-missing time on the item, which may include 0.

"Not Administered" and "Not Reached" items are items for which no responses were provided due to the items not being administered during the session, or due to items that occur at the end of the session–i.e., items that the student probably did not reach during the administration for the session. These nonresponses are considered *missing* in CTT analyses and therefore do not contribute to the statistics. For online forms, these items will have a missing time on the item.

Classical Test Theory (CTT) Item Analyses

For the key-check analyses, CTT statistics are computed by form for each item. All administration types (online and paper accommodations) are evaluated. However, the paper-based forms may not have sufficient response volumes for reliable results.

[1] At the time of the key-check, file production invalidations and student testing status may not be populated.



Item *p*-value (pseudo *p*-value for polytomous items)

The *p*-value represents the mean item score as a proportion of the maximum obtainable score points, indicating the item difficulty. Values range between 0 and 1. Higher values indicate easier items while lower values indicate more difficult items. For dichotomous items (items scored as either correct = 1 or incorrect = 0) the *p*-value is calculated as

$$p \text{ value} = \left(\frac{\bar{x}}{1}\right) = \frac{\left(\frac{1}{n} \cdot \sum_1^n x_i\right)}{1}$$

where x_i are the individual student item scores on item *i* and *n* is the total number of students for whom the item was administered.

For polytomously scored items, the pseudo *p*-value is calculated by dividing the mean item score by the maximum obtainable points possible for the item:

$$\text{pseudo } p \text{ value} = \frac{\bar{x}}{T}$$

where $\bar{x}$ is the mean item score and *T* is the maximum obtainable points possible for the polytomously scored item.

Frequently, the *p*-value is reported as a percentage by multiplying its value by 100. For instance, a *p*-value of 0.67 for a dichotomously scored item means that 67 percent of the students answered the item correctly, which is the value of the mean item score divided by 1 and then multiplying the calculated value by 100.

Response Option or Score Point Proportions

A dichotomous item's alternate response options (i.e., distractors) are plausible but incorrect options that are included to test common misconceptions or miscalculations. Ideally, all response options should garner a proportion of student selections with the correct answer obtaining the greatest proportion. These proportions are calculated by this simple formula:

$$proportion = \frac{N_O}{N_T}$$

where N_O is the number of students selecting the specific option (or omitting a response) and N_T is the total number of students for whom the question was administered.

In the case of polytomous items, the numerator becomes the number of students obtaining the specific score point (N_{SP}), with omitted responses having a score point of zero:

$$proportion = \frac{N_{SP}}{N_T}$$

Item Total Correlations

The item total correlation is the relationship between students' performance on the item and students' performance on the criterion.[2] Possible values range between -1 and +1. The correlation will be positive when the mean test score of the students answering the item correctly is greater than the mean test score of the students answering the item incorrectly. Negative values may indicate that an item has multiple correct answers or an incorrect answer key.

[2] For the key-check, this correlation is the machine-scorable total raw score. Otherwise, it is the total raw score.

The point-biserial correlation (Crocker & Algina, 1986) is one possible item total correlation for dichotomously scored items. However, the correlation will be spuriously high because the item of interest is also included in the total test score (i.e., correlating with itself; Henrysson, 1963). Therefore, a correction is made by using the means with the item deleted (i.e., the total operational test score not including the item of interest) from the calculation

$$r_{\text{pbis}} = \frac{(\bar{M}'_{+} - \overline{M'})}{S''} \sqrt{{}^{p}/_{(1-p)}}$$

where $\bar{M}'_{+}$ is the mean score with the item deleted for students who answered the item correctly, $\overline{M'}$ is the mean score with the item deleted for all students, S'' is the standard deviation with the item deleted for all students, and p is the item p-value (difficulty).

The Pearson correlation (polyserial) with the item of interest deleted is typically calculated for polytomous items by using this formula:

$$r = \frac{\Sigma(x_i - \bar{x})(y'_i - \bar{y}')}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y'_i - \bar{y}')^2}}$$

where x_i is the student score point on the item, $\bar{x}$ is the mean score for the item, y'_i is the total score with the item deleted for the student, and $\bar{y}'$ is the mean total score with the item deleted for all students (Lemke & Wiersma, 1976).

Response Option or Score Point Correlations

Similar to the overall item point-biserial correlation calculation, a correlation can be calculated for each incorrect response option (O) for multiple-choice single-response items or for the score point in the case of other item types by using this generalized formula:

$$r_{pbis_O} = \frac{(\bar{M}_O - \overline{M'})}{S''} \sqrt{{}^{p_O}/_{(1-p_C)}}$$

where $\bar{M}_O$ is the mean score for students who selected the distractor/received the number of score points, $\bar{M}$ is the mean score for all students with the item deleted, S'' is the standard deviation of all students with the item deleted, p_O is the proportion of students selecting the distractor/receiving the number of score points, and p_C is the proportion of students selecting the correct response/receiving full credit.

Flagged Item Review and Adjudication

Items flagged for issues during the key-check analyses are provided to the science test-development manager for review and/or adjudication. The items' scoring rules and responses (correct and partially correct) are reviewed. If items are confirmed to have a scoring issue, recommendations are provided for rectifying scoring, if possible. Otherwise, items may be recommended for an unscored status.

Table 4 presents the classical test theory criteria for flagging an item.

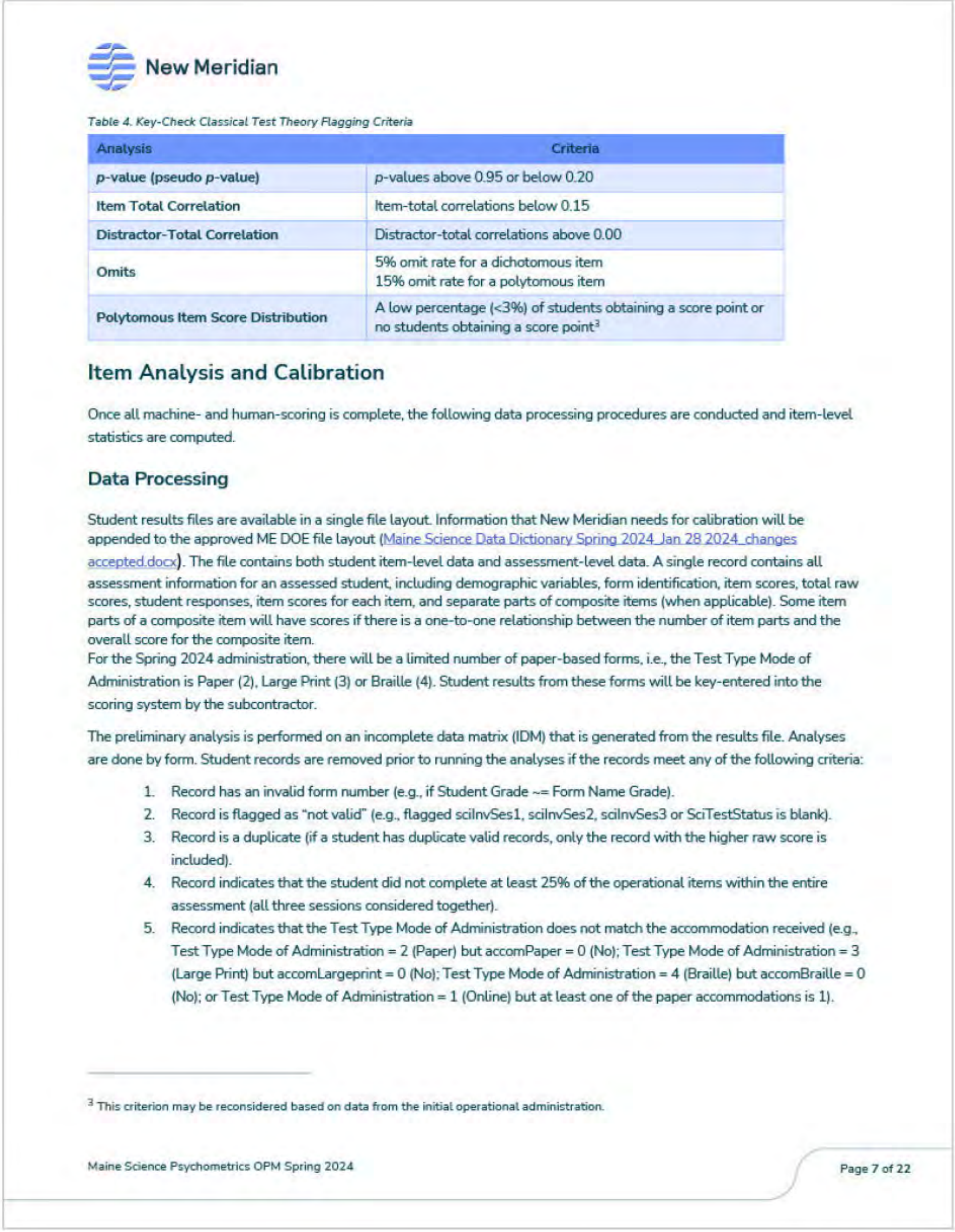


Table 4. Key-Check Classical Test Theory Flagging Criteria

Analysis	Criteria
***p*-value (pseudo *p*-value)**	*p*-values above 0.95 or below 0.20
Item Total Correlation	Item-total correlations below 0.15
Distractor-Total Correlation	Distractor-total correlations above 0.00
Omits	5% omit rate for a dichotomous item 15% omit rate for a polytomous item
Polytomous Item Score Distribution	A low percentage (<3%) of students obtaining a score point or no students obtaining a score point[3]

Item Analysis and Calibration

Once all machine- and human-scoring is complete, the following data processing procedures are conducted and item-level statistics are computed.

Data Processing

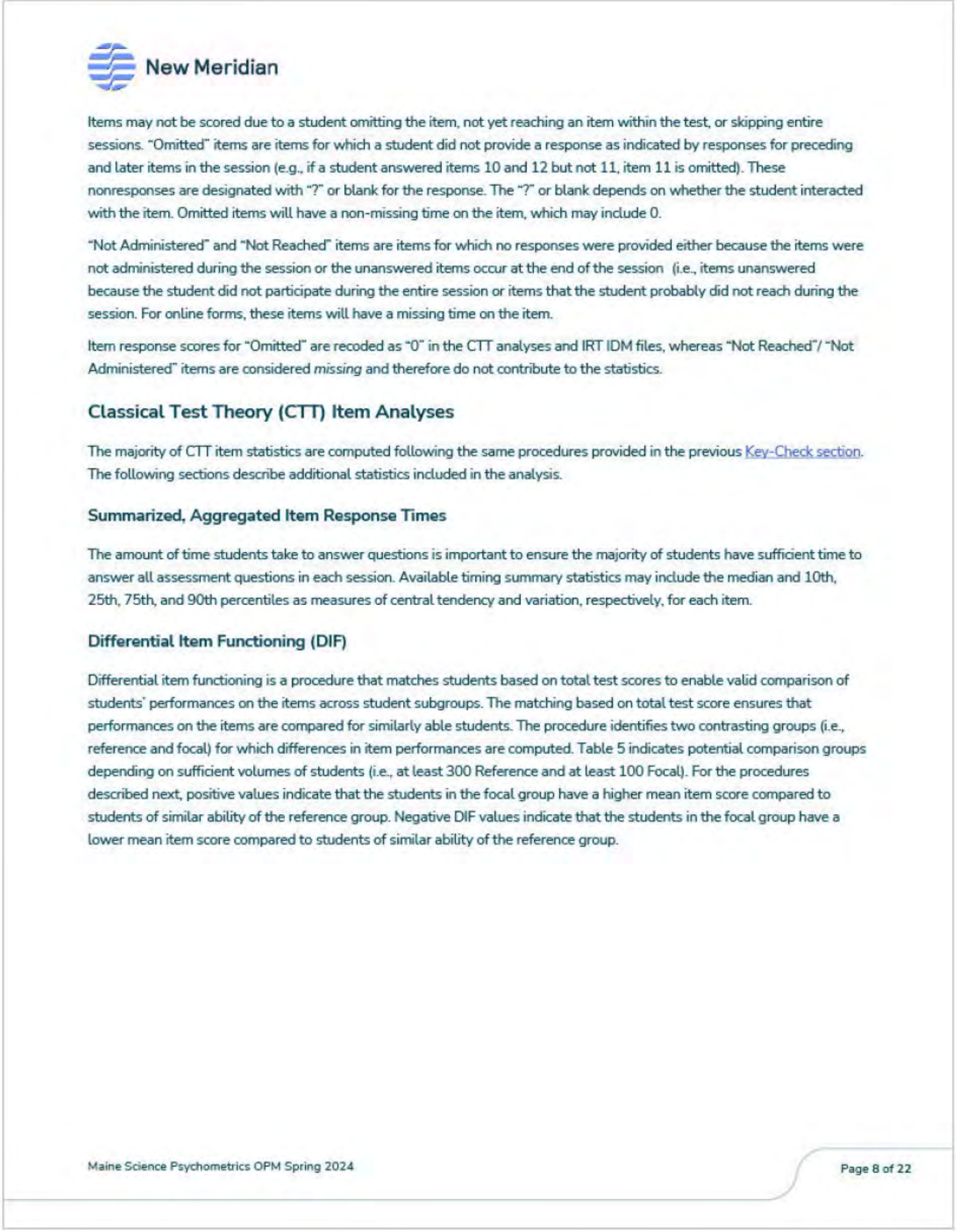
Student results files are available in a single file layout. Information that New Meridian needs for calibration will be appended to the approved ME DOE file layout (Maine Science Data Dictionary Spring 2024_Jan 28 2024_changes accepted.docx). The file contains both student item-level data and assessment-level data. A single record contains all assessment information for an assessed student, including demographic variables, form identification, item scores, total raw scores, student responses, item scores for each item, and separate parts of composite items (when applicable). Some item parts of a composite item will have scores if there is a one-to-one relationship between the number of item parts and the overall score for the composite item.

For the Spring 2024 administration, there will be a limited number of paper-based forms, i.e., the Test Type Mode of Administration is Paper (2), Large Print (3) or Braille (4). Student results from these forms will be key-entered into the scoring system by the subcontractor.

The preliminary analysis is performed on an incomplete data matrix (IDM) that is generated from the results file. Analyses are done by form. Student records are removed prior to running the analyses if the records meet any of the following criteria:

1. Record has an invalid form number (e.g., if Student Grade ~= Form Name Grade).
2. Record is flagged as "not valid" (e.g., flagged sciInvSes1, sciInvSes2, sciInvSes3 or SciTestStatus is blank).
3. Record is a duplicate (if a student has duplicate valid records, only the record with the higher raw score is included).
4. Record indicates that the student did not complete at least 25% of the operational items within the entire assessment (all three sessions considered together).
5. Record indicates that the Test Type Mode of Administration does not match the accommodation received (e.g., Test Type Mode of Administration = 2 (Paper) but accomPaper = 0 (No); Test Type Mode of Administration = 3 (Large Print) but accomLargeprint = 0 (No); Test Type Mode of Administration = 4 (Braille) but accomBraille = 0 (No); or Test Type Mode of Administration = 1 (Online) but at least one of the paper accommodations is 1).

[3] This criterion may be reconsidered based on data from the initial operational administration.



New Meridian

Items may not be scored due to a student omitting the item, not yet reaching an item within the test, or skipping entire sessions. "Omitted" items are items for which a student did not provide a response as indicated by responses for preceding and later items in the session (e.g., if a student answered items 10 and 12 but not 11, item 11 is omitted). These nonresponses are designated with "?" or blank for the response. The "?" or blank depends on whether the student interacted with the item. Omitted items will have a non-missing time on the item, which may include 0.

"Not Administered" and "Not Reached" items are items for which no responses were provided either because the items were not administered during the session or the unanswered items occur at the end of the session (i.e., items unanswered because the student did not participate during the entire session or items that the student probably did not reach during the session. For online forms, these items will have a missing time on the item.

Item response scores for "Omitted" are recoded as "0" in the CTT analyses and IRT IDM files, whereas "Not Reached"/ "Not Administered" items are considered *missing* and therefore do not contribute to the statistics.

Classical Test Theory (CTT) Item Analyses

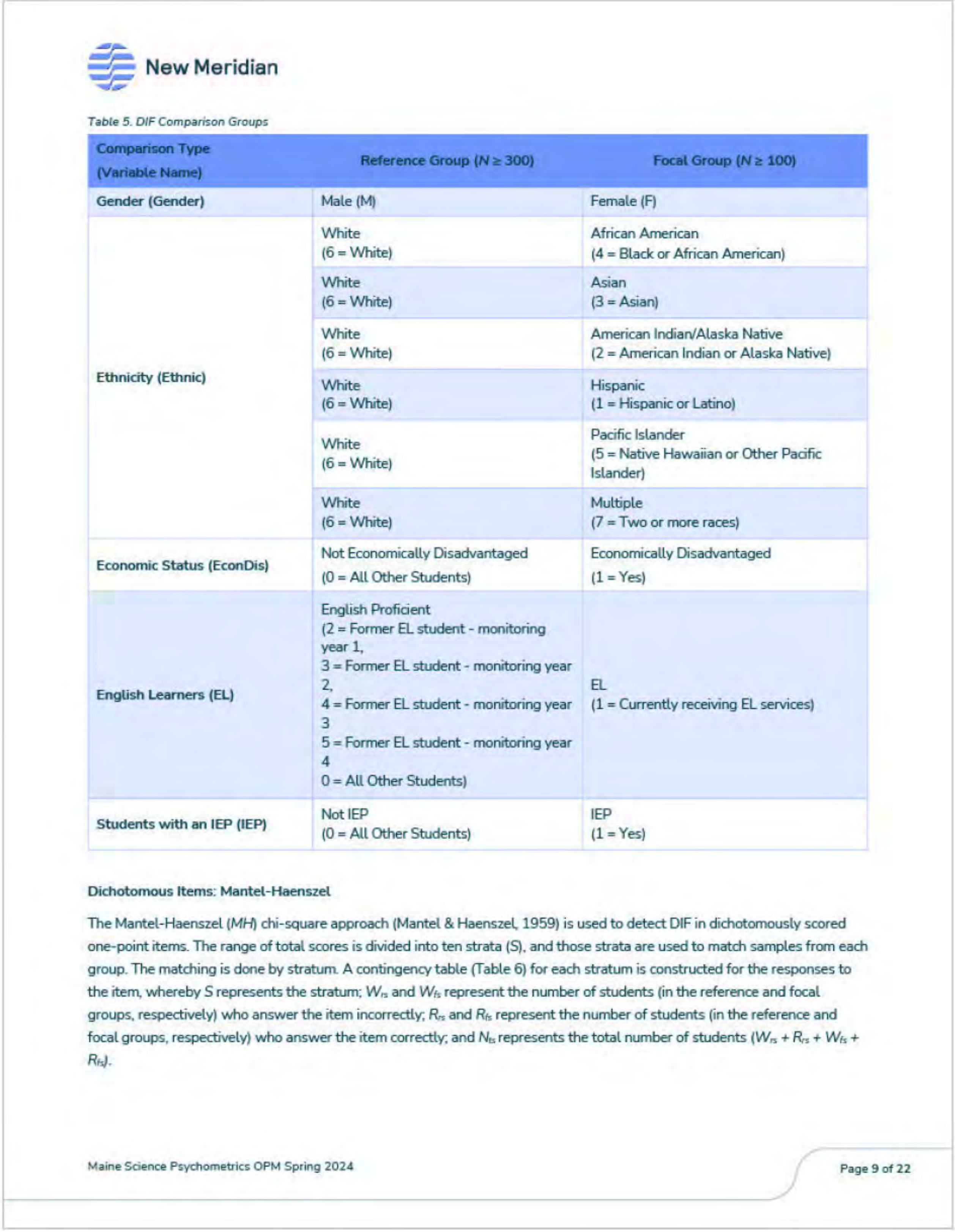
The majority of CTT item statistics are computed following the same procedures provided in the previous Key-Check section. The following sections describe additional statistics included in the analysis.

Summarized, Aggregated Item Response Times

The amount of time students take to answer questions is important to ensure the majority of students have sufficient time to answer all assessment questions in each session. Available timing summary statistics may include the median and 10th, 25th, 75th, and 90th percentiles as measures of central tendency and variation, respectively, for each item.

Differential Item Functioning (DIF)

Differential item functioning is a procedure that matches students based on total test scores to enable valid comparison of students' performances on the items across student subgroups. The matching based on total test score ensures that performances on the items are compared for similarly able students. The procedure identifies two contrasting groups (i.e., reference and focal) for which differences in item performances are computed. Table 5 indicates potential comparison groups depending on sufficient volumes of students (i.e., at least 300 Reference and at least 100 Focal). For the procedures described next, positive values indicate that the students in the focal group have a higher mean item score compared to students of similar ability of the reference group. Negative DIF values indicate that the students in the focal group have a lower mean item score compared to students of similar ability of the reference group.



New Meridian

Table 5. DIF Comparison Groups

Comparison Type (Variable Name)	Reference Group ($N \geq 300$)	Focal Group ($N \geq 100$)
Gender (Gender)	Male (M)	Female (F)
Ethnicity (Ethnic)	White (6 = White)	African American (4 = Black or African American)
	White (6 = White)	Asian (3 = Asian)
	White (6 = White)	American Indian/Alaska Native (2 = American Indian or Alaska Native)
	White (6 = White)	Hispanic (1 = Hispanic or Latino)
	White (6 = White)	Pacific Islander (5 = Native Hawaiian or Other Pacific Islander)
	White (6 = White)	Multiple (7 = Two or more races)
Economic Status (EconDis)	Not Economically Disadvantaged (0 = All Other Students)	Economically Disadvantaged (1 = Yes)
English Learners (EL)	English Proficient (2 = Former EL student - monitoring year 1, 3 = Former EL student - monitoring year 2, 4 = Former EL student - monitoring year 3 5 = Former EL student - monitoring year 4 0 = All Other Students)	EL (1 = Currently receiving EL services)
Students with an IEP (IEP)	Not IEP (0 = All Other Students)	IEP (1 = Yes)

Dichotomous Items: Mantel-Haenszel

The Mantel-Haenszel (*MH*) chi-square approach (Mantel & Haenszel, 1959) is used to detect DIF in dichotomously scored one-point items. The range of total scores is divided into ten strata (*S*), and those strata are used to match samples from each group. The matching is done by stratum. A contingency table (Table 6) for each stratum is constructed for the responses to the item, whereby *S* represents the stratum; W_{rs} and W_{fs} represent the number of students (in the reference and focal groups, respectively) who answer the item incorrectly; R_{rs} and R_{fs} represent the number of students (in the reference and focal groups, respectively) who answer the item correctly; and N_{ts} represents the total number of students ($W_{rs} + R_{rs} + W_{fs} + R_{fs}$).

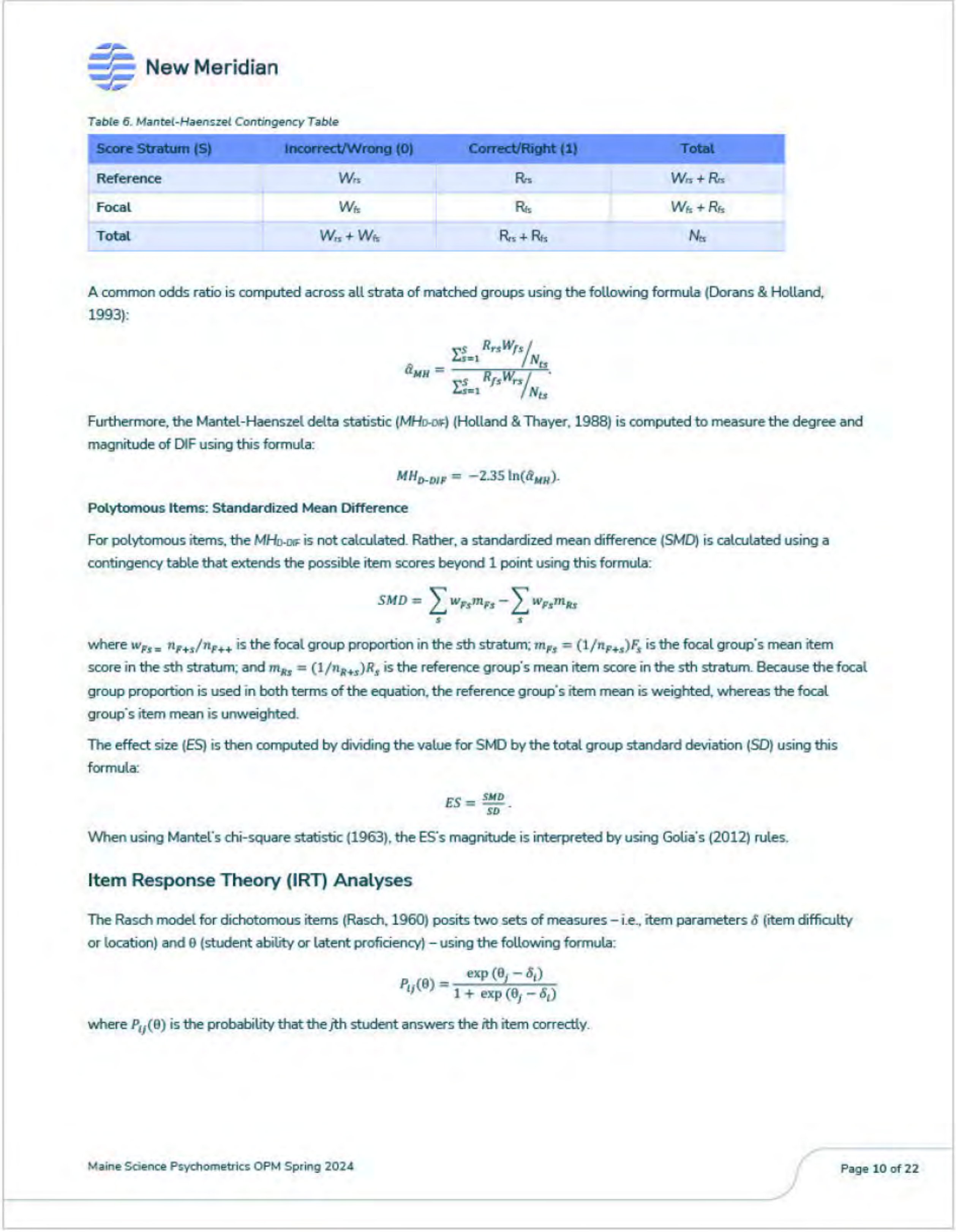


Table 6. Mantel-Haenszel Contingency Table

Score Stratum (S)	Incorrect/Wrong (0)	Correct/Right (1)	Total
Reference	W_{rs}	R_{rs}	$W_{rs} + R_{rs}$
Focal	W_{fs}	R_{fs}	$W_{fs} + R_{fs}$
Total	$W_{rs} + W_{fs}$	$R_{rs} + R_{fs}$	N_{ts}

A common odds ratio is computed across all strata of matched groups using the following formula (Dorans & Holland, 1993):

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S} {R_{rs}W_{fs}}/{N_{ts}}}{\sum_{s=1}^{S} {R_{fs}W_{rs}}/{N_{ts}}}.$$

Furthermore, the Mantel-Haenszel delta statistic (*MH*$_{D\text{-}DIF}$) (Holland & Thayer, 1988) is computed to measure the degree and magnitude of DIF using this formula:

$$MH_{D\text{-}DIF} = -2.35 \ln(\hat{\alpha}_{MH}).$$

Polytomous Items: Standardized Mean Difference

For polytomous items, the *MH*$_{D\text{-}DIF}$ is not calculated. Rather, a standardized mean difference (*SMD*) is calculated using a contingency table that extends the possible item scores beyond 1 point using this formula:

$$SMD = \sum_{s} w_{Fs}m_{Fs} - \sum_{s} w_{Fs}m_{Rs}$$

where $w_{Fs=} n_{F+s}/n_{F++}$ is the focal group proportion in the sth stratum; $m_{Fs} = (1/n_{F+s})F_s$ is the focal group's mean item score in the sth stratum; and $m_{Rs} = (1/n_{R+s})R_s$ is the reference group's mean item score in the sth stratum. Because the focal group proportion is used in both terms of the equation, the reference group's item mean is weighted, whereas the focal group's item mean is unweighted.

The effect size (*ES*) is then computed by dividing the value for SMD by the total group standard deviation (*SD*) using this formula:

$$ES = \frac{SMD}{SD}.$$

When using Mantel's chi-square statistic (1963), the ES's magnitude is interpreted by using Golia's (2012) rules.

Item Response Theory (IRT) Analyses

The Rasch model for dichotomous items (Rasch, 1960) posits two sets of measures – i.e., item parameters δ (item difficulty or location) and θ (student ability or latent proficiency) – using the following formula:

$$P_{ij}(\theta) = \frac{\exp(\theta_j - \delta_i)}{1 + \exp(\theta_j - \delta_i)}$$

where $P_{ij}(\theta)$ is the probability that the *j*th student answers the *i*th item correctly.

Masters (1982) proposes the Partial Credit Model (PCM) for use when there are two or more ordered categories of responses (i.e., polytomous items). The PCM is a generalization of the simple Rasch model for dichotomous items; and it is considered to be just one of many types of Rasch models. The PCM is defined as $P_{ijx}(\theta) = \frac{\exp \sum_{k=0}^{x}(\theta_j - \delta_{ik})}{\sum_{h=0}^{m_j} exp \sum_{k=0}^{h}(\theta_j - \delta_{ik})}$, $x = 1,2, \dots, m_i$,

where $P_{ijx}(\theta)$ is the probability that the *j*th student gets a score of *x* on the *i*th item.

Furthermore, for notational convenience:

$$\sum_{k=0}^{0}(\theta_j - \delta_{ik}) \equiv 0$$

and

$$\sum_{k=0}^{h}(\theta_j - \delta_{ik}) \equiv \sum_{k=1}^{h}(\theta_j - \delta_{ik}).$$

The program Winsteps 5.6.1.0 (Linacre, 2023) is used to estimate the item parameters and student ability/latent proficiency parameters. To ensure quality, two members of the New Meridian psychometrics team independently replicate the results of the item calibrations and estimations of student ability/latent proficiency parameters, and the two team member's results must match exactly. An example of the Winsteps control file used to conduct item calibration and ability/proficiency estimation is provided in Appendix A.

Incomplete Data Matrix (IDM)

For the IRT item calibrations and student ability/proficiency estimation, items are excluded if their CTT statistics appear flawed or problematic based on the following criteria:

1. Item has an item total correlation less than 0.
2. Item has 100% of the students with same item score, such as 100% omitted the item or received the same score.

New Meridian will document flagged items and notify the ME DOE regarding the recommendation to exclude them from calibration.

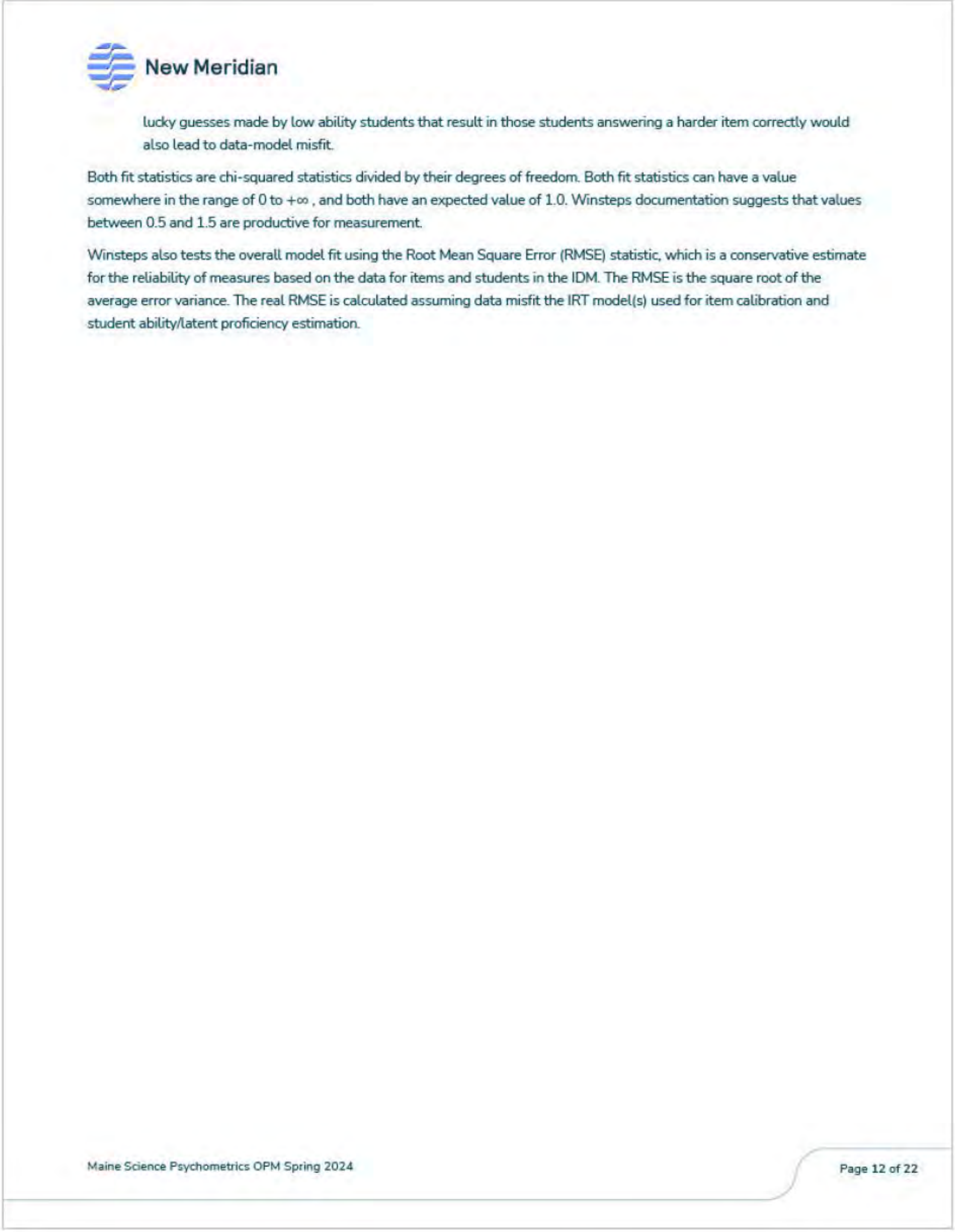
IRT Model Assumptions

Internal structure validity studies investigate "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, & NCME, 2014, p. 16).

Fit Indices

One assumption of IRT models is that the data fit the model. The following two fit indices are calculated for the items in Winsteps to evaluate the assumption regarding model fit:

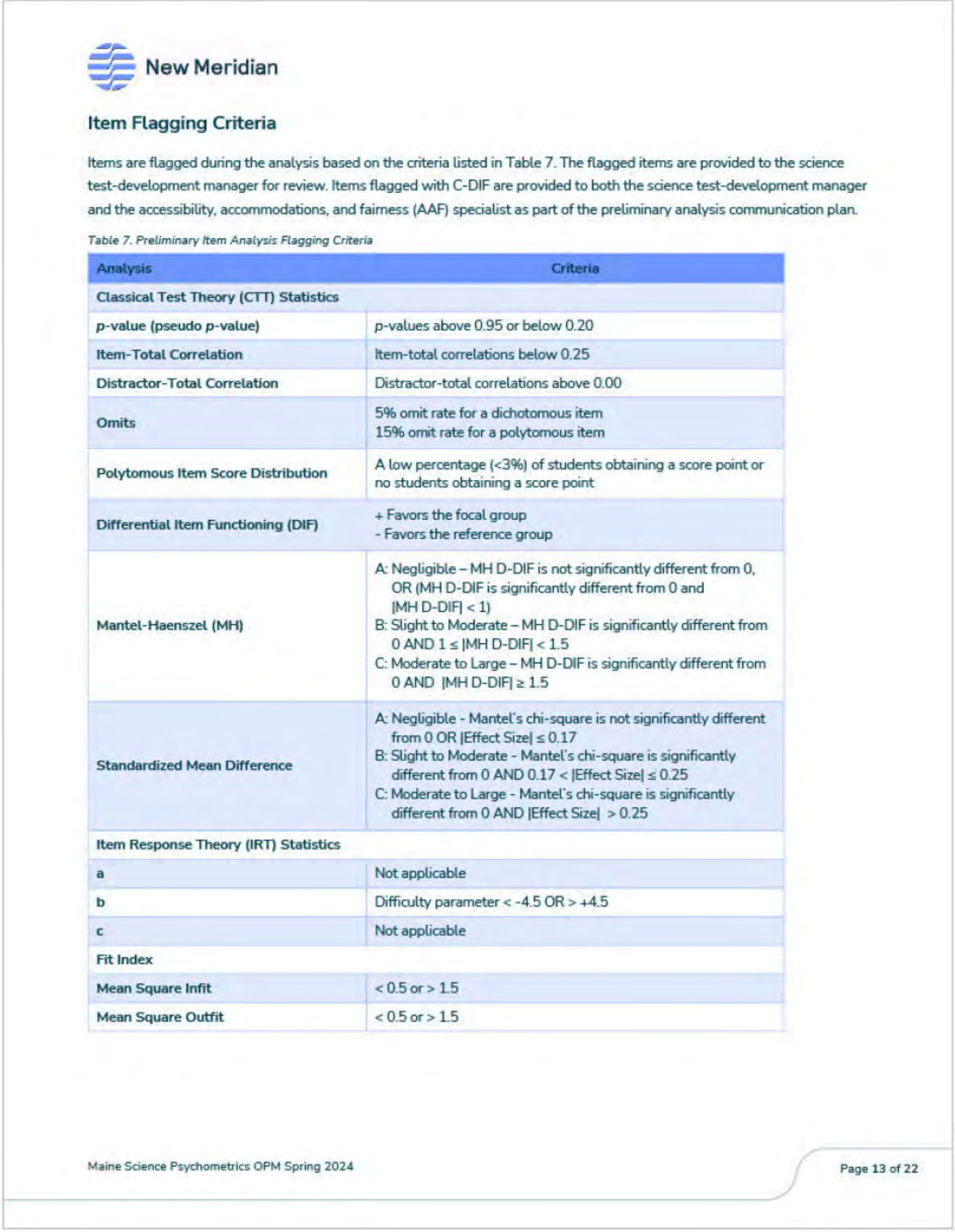
- **Mean Square Infit (IN.MSQ)** – An information-weighted statistic based on response patterns, which is more sensitive to a student responding unexpectedly to an item having a level of difficulty near the student's ability level.
- **Mean Square Outfit (OUT.MSQ)** – An unweighted statistic that is more sensitive to a student responding unexpectedly to an item that has a level of difficulty far from the student's ability level. For example, one would expect students demonstrating a high level of ability to correctly answer an item with a low level of difficulty. Careless mistakes by the high ability student, however, would lead to data-model misfit for the easy item. Equally,



lucky guesses made by low ability students that result in those students answering a harder item correctly would also lead to data-model misfit.

Both fit statistics are chi-squared statistics divided by their degrees of freedom. Both fit statistics can have a value somewhere in the range of 0 to $+\infty$, and both have an expected value of 1.0. Winsteps documentation suggests that values between 0.5 and 1.5 are productive for measurement.

Winsteps also tests the overall model fit using the Root Mean Square Error (RMSE) statistic, which is a conservative estimate for the reliability of measures based on the data for items and students in the IDM. The RMSE is the square root of the average error variance. The real RMSE is calculated assuming data misfit the IRT model(s) used for item calibration and student ability/latent proficiency estimation.



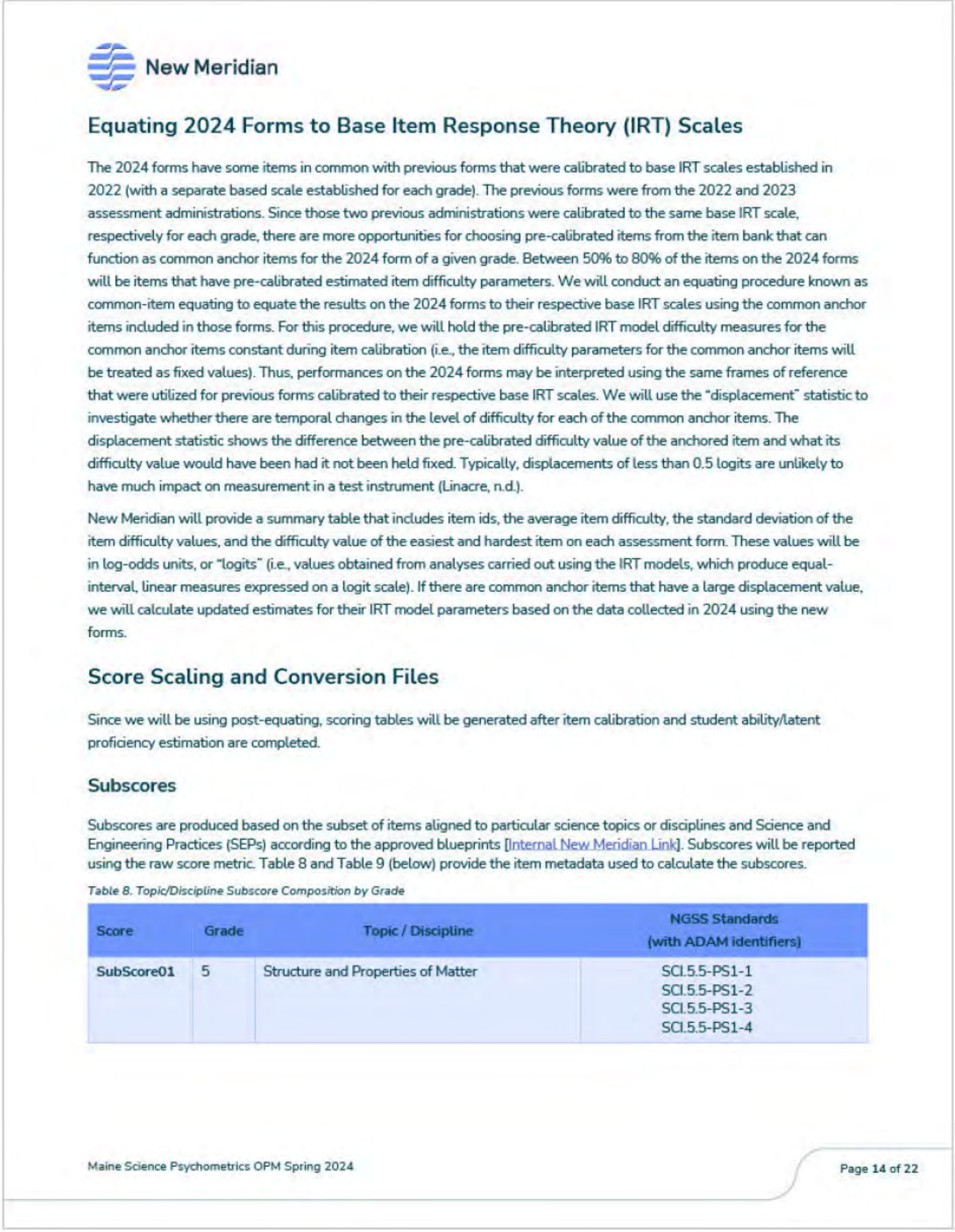
New Meridian

Item Flagging Criteria

Items are flagged during the analysis based on the criteria listed in Table 7. The flagged items are provided to the science test-development manager for review. Items flagged with C-DIF are provided to both the science test-development manager and the accessibility, accommodations, and fairness (AAF) specialist as part of the preliminary analysis communication plan.

Table 7. Preliminary Item Analysis Flagging Criteria

Analysis	Criteria
Classical Test Theory (CTT) Statistics	
***p*-value (pseudo *p*-value)**	*p*-values above 0.95 or below 0.20
Item-Total Correlation	Item-total correlations below 0.25
Distractor-Total Correlation	Distractor-total correlations above 0.00
Omits	5% omit rate for a dichotomous item 15% omit rate for a polytomous item
Polytomous Item Score Distribution	A low percentage (<3%) of students obtaining a score point or no students obtaining a score point
Differential Item Functioning (DIF)	+ Favors the focal group - Favors the reference group
Mantel-Haenszel (MH)	A: Negligible – MH D-DIF is not significantly different from 0, OR (MH D-DIF is significantly different from 0 and |MH D-DIF| < 1) B: Slight to Moderate – MH D-DIF is significantly different from 0 AND 1 ≤ |MH D-DIF| < 1.5 C: Moderate to Large – MH D-DIF is significantly different from 0 AND |MH D-DIF| ≥ 1.5
Standardized Mean Difference	A: Negligible - Mantel's chi-square is not significantly different from 0 OR |Effect Size| ≤ 0.17 B: Slight to Moderate - Mantel's chi-square is significantly different from 0 AND 0.17 < |Effect Size| ≤ 0.25 C: Moderate to Large - Mantel's chi-square is significantly different from 0 AND |Effect Size| > 0.25
Item Response Theory (IRT) Statistics	
a	Not applicable
b	Difficulty parameter < -4.5 OR > +4.5
c	Not applicable
Fit Index	
Mean Square Infit	< 0.5 or > 1.5
Mean Square Outfit	< 0.5 or > 1.5



New Meridian

Equating 2024 Forms to Base Item Response Theory (IRT) Scales

The 2024 forms have some items in common with previous forms that were calibrated to base IRT scales established in 2022 (with a separate based scale established for each grade). The previous forms were from the 2022 and 2023 assessment administrations. Since those two previous administrations were calibrated to the same base IRT scale, respectively for each grade, there are more opportunities for choosing pre-calibrated items from the item bank that can function as common anchor items for the 2024 form of a given grade. Between 50% to 80% of the items on the 2024 forms will be items that have pre-calibrated estimated item difficulty parameters. We will conduct an equating procedure known as common-item equating to equate the results on the 2024 forms to their respective base IRT scales using the common anchor items included in those forms. For this procedure, we will hold the pre-calibrated IRT model difficulty measures for the common anchor items constant during item calibration (i.e., the item difficulty parameters for the common anchor items will be treated as fixed values). Thus, performances on the 2024 forms may be interpreted using the same frames of reference that were utilized for previous forms calibrated to their respective base IRT scales. We will use the "displacement" statistic to investigate whether there are temporal changes in the level of difficulty for each of the common anchor items. The displacement statistic shows the difference between the pre-calibrated difficulty value of the anchored item and what its difficulty value would have been had it not been held fixed. Typically, displacements of less than 0.5 logits are unlikely to have much impact on measurement in a test instrument (Linacre, n.d.).

New Meridian will provide a summary table that includes item ids, the average item difficulty, the standard deviation of the item difficulty values, and the difficulty value of the easiest and hardest item on each assessment form. These values will be in log-odds units, or "logits" (i.e., values obtained from analyses carried out using the IRT models, which produce equal-interval, linear measures expressed on a logit scale). If there are common anchor items that have a large displacement value, we will calculate updated estimates for their IRT model parameters based on the data collected in 2024 using the new forms.

Score Scaling and Conversion Files

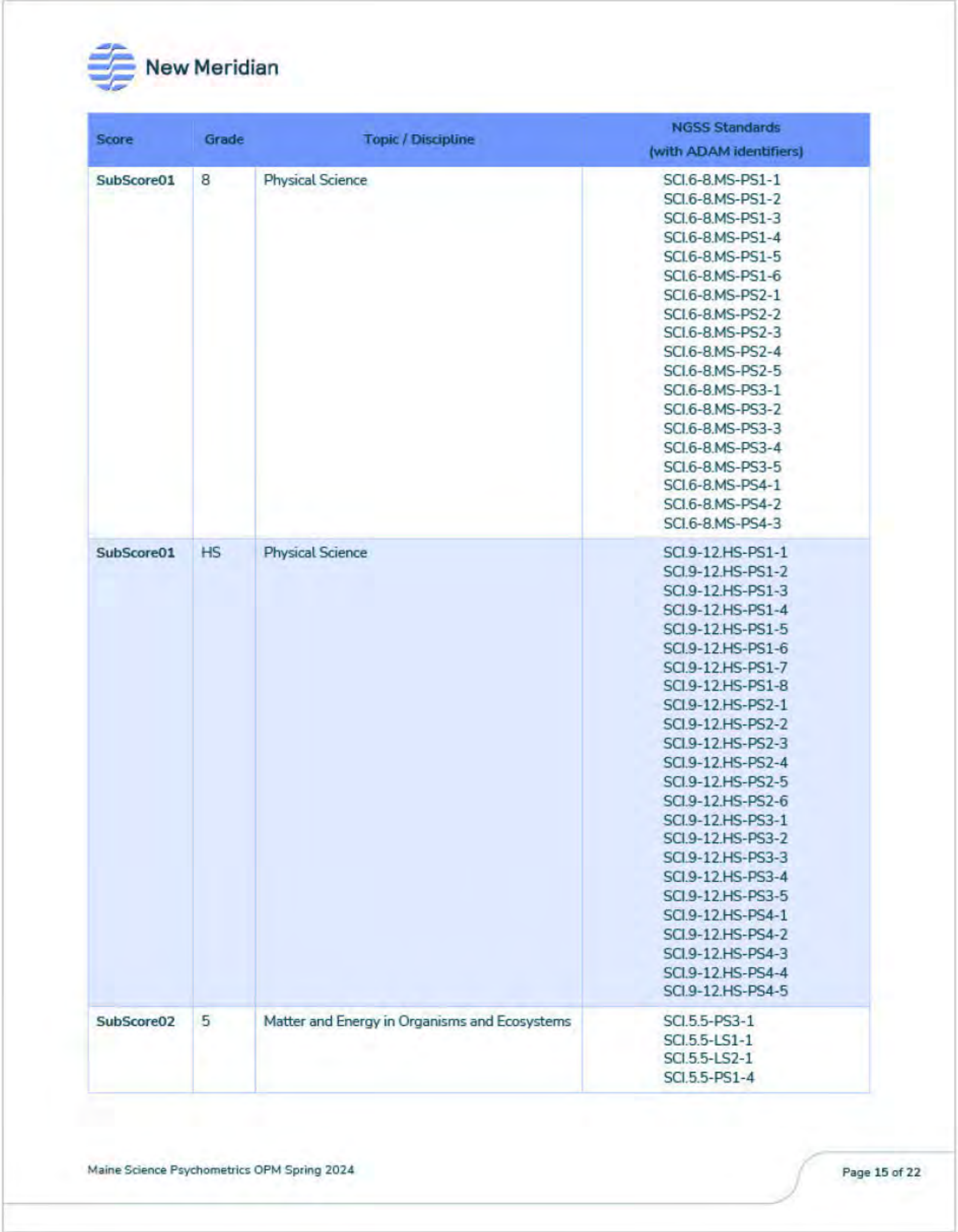
Since we will be using post-equating, scoring tables will be generated after item calibration and student ability/latent proficiency estimation are completed.

Subscores

Subscores are produced based on the subset of items aligned to particular science topics or disciplines and Science and Engineering Practices (SEPs) according to the approved blueprints [Internal New Meridian Link]. Subscores will be reported using the raw score metric. Table 8 and Table 9 (below) provide the item metadata used to calculate the subscores.

Table 8. Topic/Discipline Subscore Composition by Grade

Score	Grade	Topic / Discipline	NGSS Standards (with ADAM identifiers)
SubScore01	5	Structure and Properties of Matter	SCI.5.5-PS1-1 SCI.5.5-PS1-2 SCI.5.5-PS1-3 SCI.5.5-PS1-4

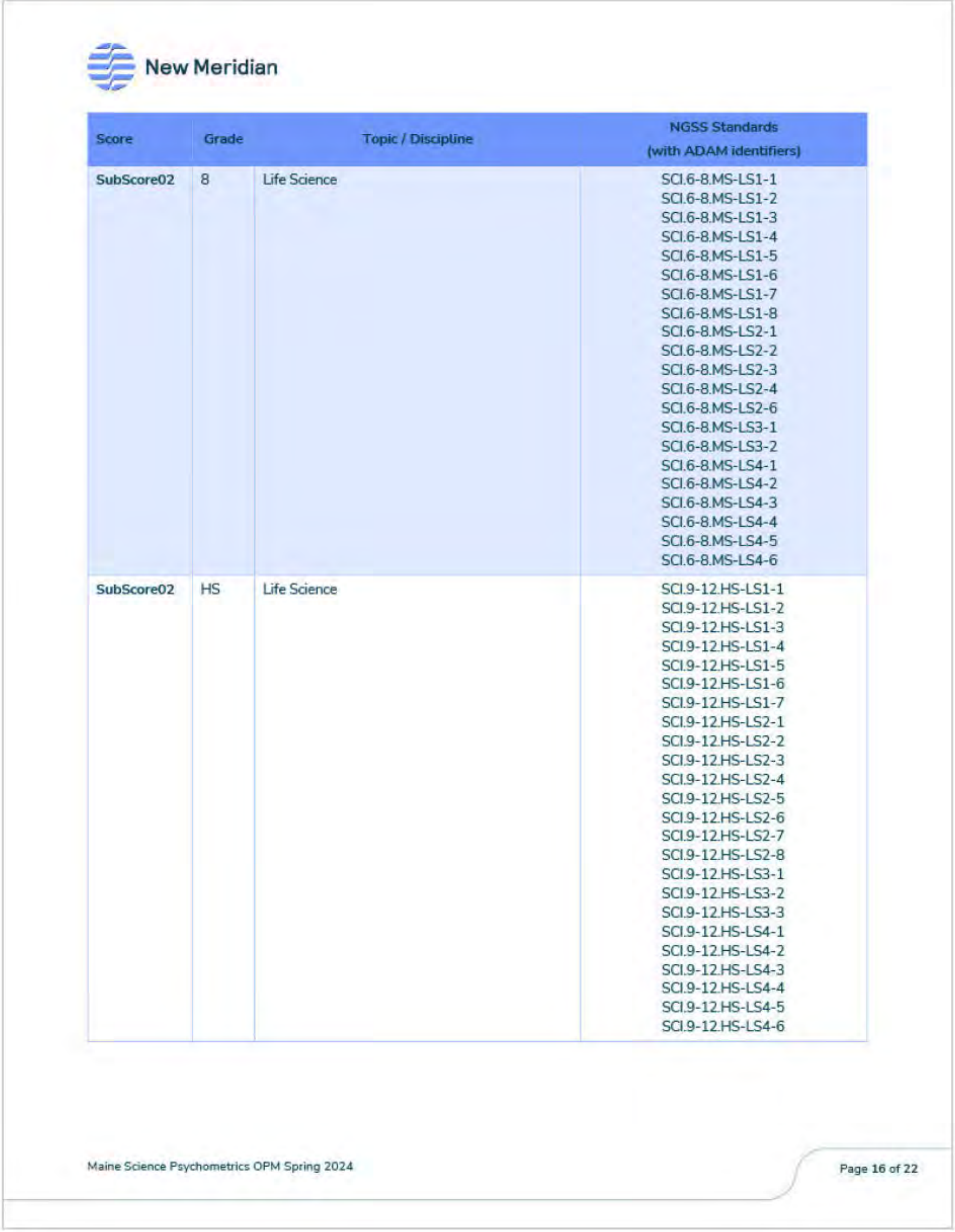


New Meridian

Score	Grade	Topic / Discipline	NGSS Standards (with ADAM identifiers)
SubScore01	8	Physical Science	SCI.6-8.MS-PS1-1 SCI.6-8.MS-PS1-2 SCI.6-8.MS-PS1-3 SCI.6-8.MS-PS1-4 SCI.6-8.MS-PS1-5 SCI.6-8.MS-PS1-6 SCI.6-8.MS-PS2-1 SCI.6-8.MS-PS2-2 SCI.6-8.MS-PS2-3 SCI.6-8.MS-PS2-4 SCI.6-8.MS-PS2-5 SCI.6-8.MS-PS3-1 SCI.6-8.MS-PS3-2 SCI.6-8.MS-PS3-3 SCI.6-8.MS-PS3-4 SCI.6-8.MS-PS3-5 SCI.6-8.MS-PS4-1 SCI.6-8.MS-PS4-2 SCI.6-8.MS-PS4-3
SubScore01	HS	Physical Science	SCI.9-12.HS-PS1-1 SCI.9-12.HS-PS1-2 SCI.9-12.HS-PS1-3 SCI.9-12.HS-PS1-4 SCI.9-12.HS-PS1-5 SCI.9-12.HS-PS1-6 SCI.9-12.HS-PS1-7 SCI.9-12.HS-PS1-8 SCI.9-12.HS-PS2-1 SCI.9-12.HS-PS2-2 SCI.9-12.HS-PS2-3 SCI.9-12.HS-PS2-4 SCI.9-12.HS-PS2-5 SCI.9-12.HS-PS2-6 SCI.9-12.HS-PS3-1 SCI.9-12.HS-PS3-2 SCI.9-12.HS-PS3-3 SCI.9-12.HS-PS3-4 SCI.9-12.HS-PS3-5 SCI.9-12.HS-PS4-1 SCI.9-12.HS-PS4-2 SCI.9-12.HS-PS4-3 SCI.9-12.HS-PS4-4 SCI.9-12.HS-PS4-5
SubScore02	5	Matter and Energy in Organisms and Ecosystems	SCI.5.5-PS3-1 SCI.5.5-LS1-1 SCI.5.5-LS2-1 SCI.5.5-PS1-4

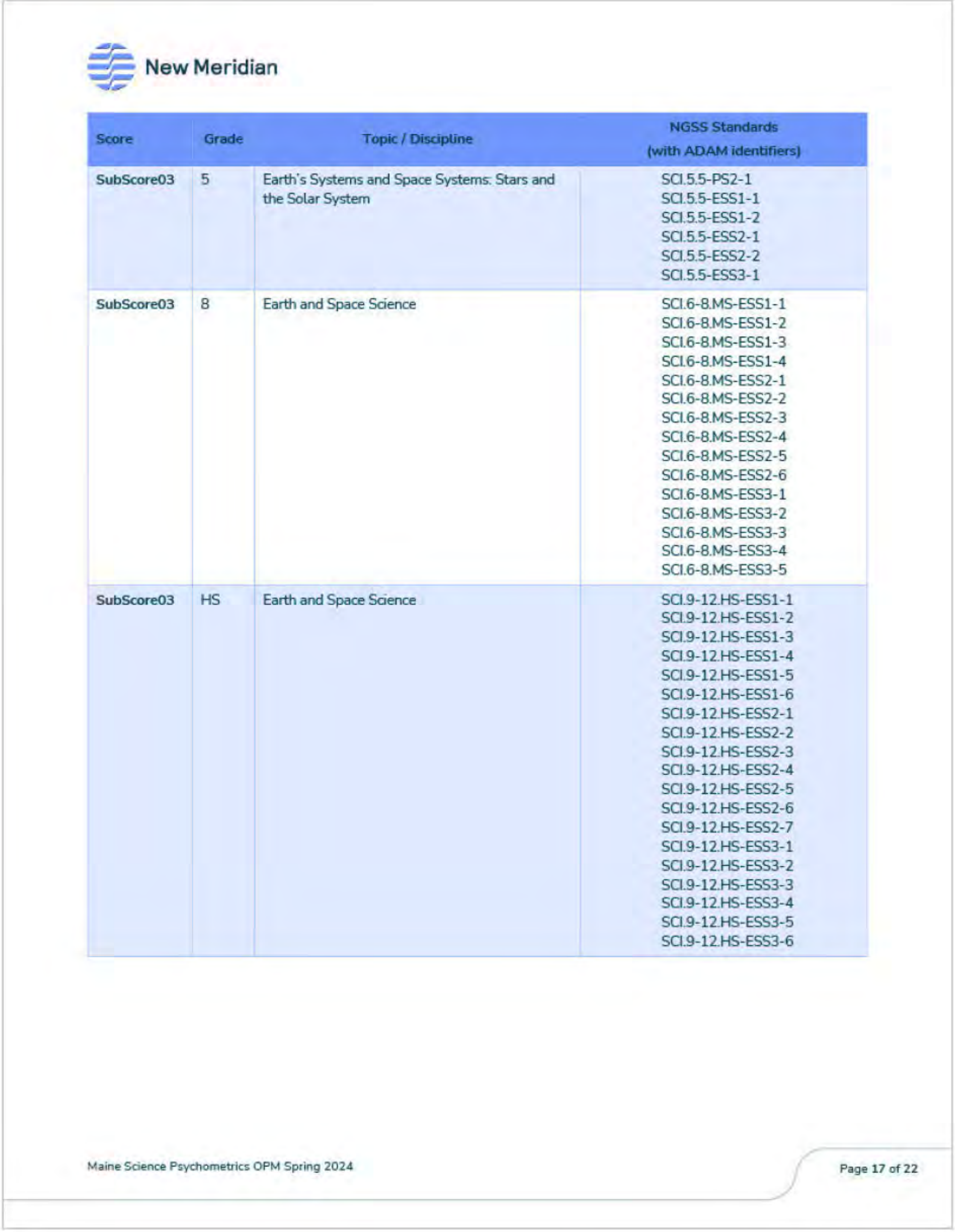
Maine Science Psychometrics OPM Spring 2024

Page 15 of 22



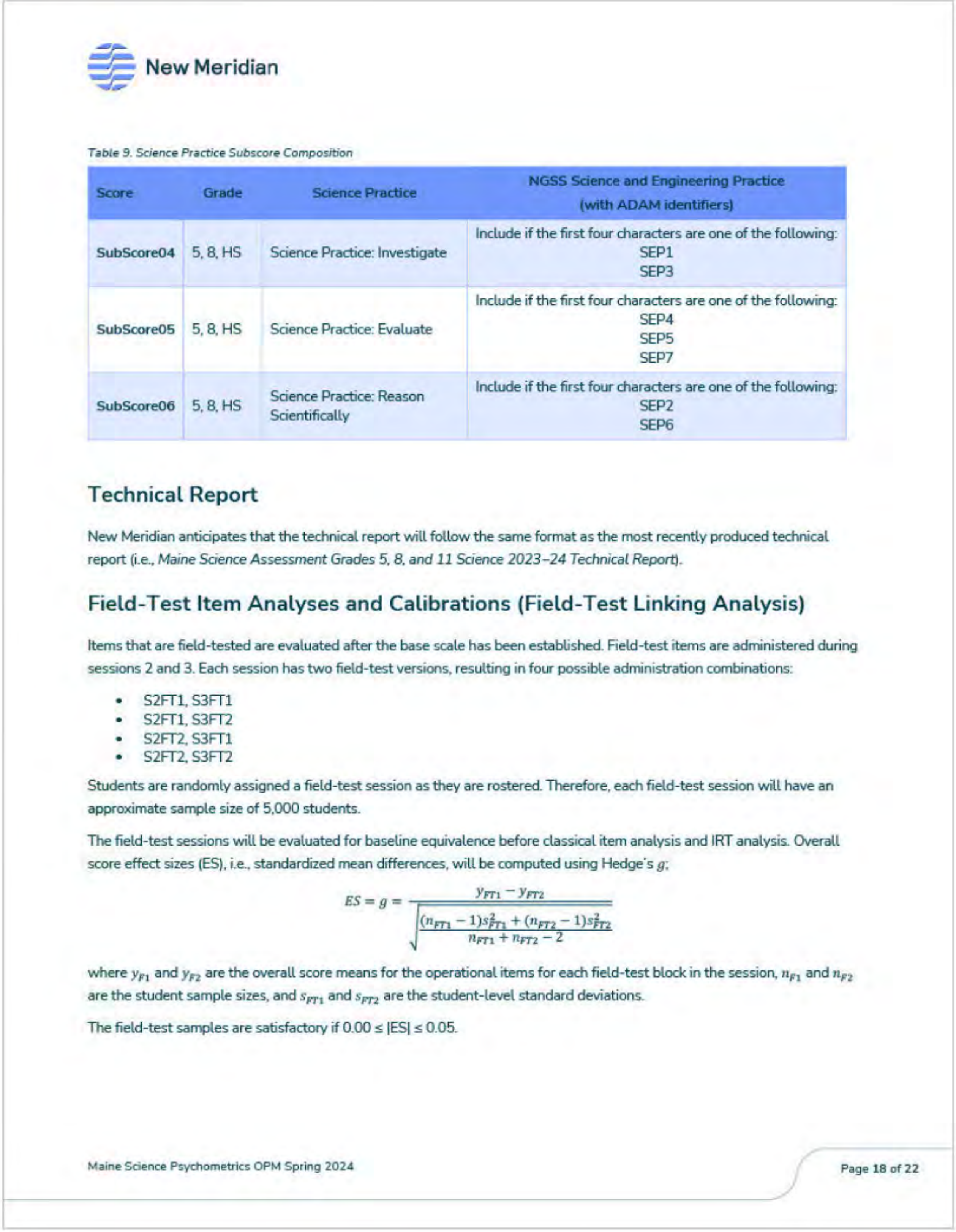
New Meridian

Score	Grade	Topic / Discipline	NGSS Standards (with ADAM identifiers)
SubScore02	8	Life Science	SCI.6-8.MS-LS1-1 SCI.6-8.MS-LS1-2 SCI.6-8.MS-LS1-3 SCI.6-8.MS-LS1-4 SCI.6-8.MS-LS1-5 SCI.6-8.MS-LS1-6 SCI.6-8.MS-LS1-7 SCI.6-8.MS-LS1-8 SCI.6-8.MS-LS2-1 SCI.6-8.MS-LS2-2 SCI.6-8.MS-LS2-3 SCI.6-8.MS-LS2-4 SCI.6-8.MS-LS2-6 SCI.6-8.MS-LS3-1 SCI.6-8.MS-LS3-2 SCI.6-8.MS-LS4-1 SCI.6-8.MS-LS4-2 SCI.6-8.MS-LS4-3 SCI.6-8.MS-LS4-4 SCI.6-8.MS-LS4-5 SCI.6-8.MS-LS4-6
SubScore02	HS	Life Science	SCI.9-12.HS-LS1-1 SCI.9-12.HS-LS1-2 SCI.9-12.HS-LS1-3 SCI.9-12.HS-LS1-4 SCI.9-12.HS-LS1-5 SCI.9-12.HS-LS1-6 SCI.9-12.HS-LS1-7 SCI.9-12.HS-LS2-1 SCI.9-12.HS-LS2-2 SCI.9-12.HS-LS2-3 SCI.9-12.HS-LS2-4 SCI.9-12.HS-LS2-5 SCI.9-12.HS-LS2-6 SCI.9-12.HS-LS2-7 SCI.9-12.HS-LS2-8 SCI.9-12.HS-LS3-1 SCI.9-12.HS-LS3-2 SCI.9-12.HS-LS3-3 SCI.9-12.HS-LS4-1 SCI.9-12.HS-LS4-2 SCI.9-12.HS-LS4-3 SCI.9-12.HS-LS4-4 SCI.9-12.HS-LS4-5 SCI.9-12.HS-LS4-6



New Meridian

Score	Grade	Topic / Discipline	NGSS Standards (with ADAM identifiers)
SubScore03	5	Earth's Systems and Space Systems: Stars and the Solar System	SCI.5.5-PS2-1 SCI.5.5-ESS1-1 SCI.5.5-ESS1-2 SCI.5.5-ESS2-1 SCI.5.5-ESS2-2 SCI.5.5-ESS3-1
SubScore03	8	Earth and Space Science	SCI.6-8.MS-ESS1-1 SCI.6-8.MS-ESS1-2 SCI.6-8.MS-ESS1-3 SCI.6-8.MS-ESS1-4 SCI.6-8.MS-ESS2-1 SCI.6-8.MS-ESS2-2 SCI.6-8.MS-ESS2-3 SCI.6-8.MS-ESS2-4 SCI.6-8.MS-ESS2-5 SCI.6-8.MS-ESS2-6 SCI.6-8.MS-ESS3-1 SCI.6-8.MS-ESS3-2 SCI.6-8.MS-ESS3-3 SCI.6-8.MS-ESS3-4 SCI.6-8.MS-ESS3-5
SubScore03	HS	Earth and Space Science	SCI.9-12.HS-ESS1-1 SCI.9-12.HS-ESS1-2 SCI.9-12.HS-ESS1-3 SCI.9-12.HS-ESS1-4 SCI.9-12.HS-ESS1-5 SCI.9-12.HS-ESS1-6 SCI.9-12.HS-ESS2-1 SCI.9-12.HS-ESS2-2 SCI.9-12.HS-ESS2-3 SCI.9-12.HS-ESS2-4 SCI.9-12.HS-ESS2-5 SCI.9-12.HS-ESS2-6 SCI.9-12.HS-ESS2-7 SCI.9-12.HS-ESS3-1 SCI.9-12.HS-ESS3-2 SCI.9-12.HS-ESS3-3 SCI.9-12.HS-ESS3-4 SCI.9-12.HS-ESS3-5 SCI.9-12.HS-ESS3-6



New Meridian

Table 9. Science Practice Subscore Composition

Score	Grade	Science Practice	NGSS Science and Engineering Practice (with ADAM identifiers)
SubScore04	5, 8, HS	Science Practice: Investigate	Include if the first four characters are one of the following: SEP1 SEP3
SubScore05	5, 8, HS	Science Practice: Evaluate	Include if the first four characters are one of the following: SEP4 SEP5 SEP7
SubScore06	5, 8, HS	Science Practice: Reason Scientifically	Include if the first four characters are one of the following: SEP2 SEP6

Technical Report

New Meridian anticipates that the technical report will follow the same format as the most recently produced technical report (i.e., *Maine Science Assessment Grades 5, 8, and 11 Science 2023–24 Technical Report*).

Field-Test Item Analyses and Calibrations (Field-Test Linking Analysis)

Items that are field-tested are evaluated after the base scale has been established. Field-test items are administered during sessions 2 and 3. Each session has two field-test versions, resulting in four possible administration combinations:

- S2FT1, S3FT1
- S2FT1, S3FT2
- S2FT2, S3FT1
- S2FT2, S3FT2

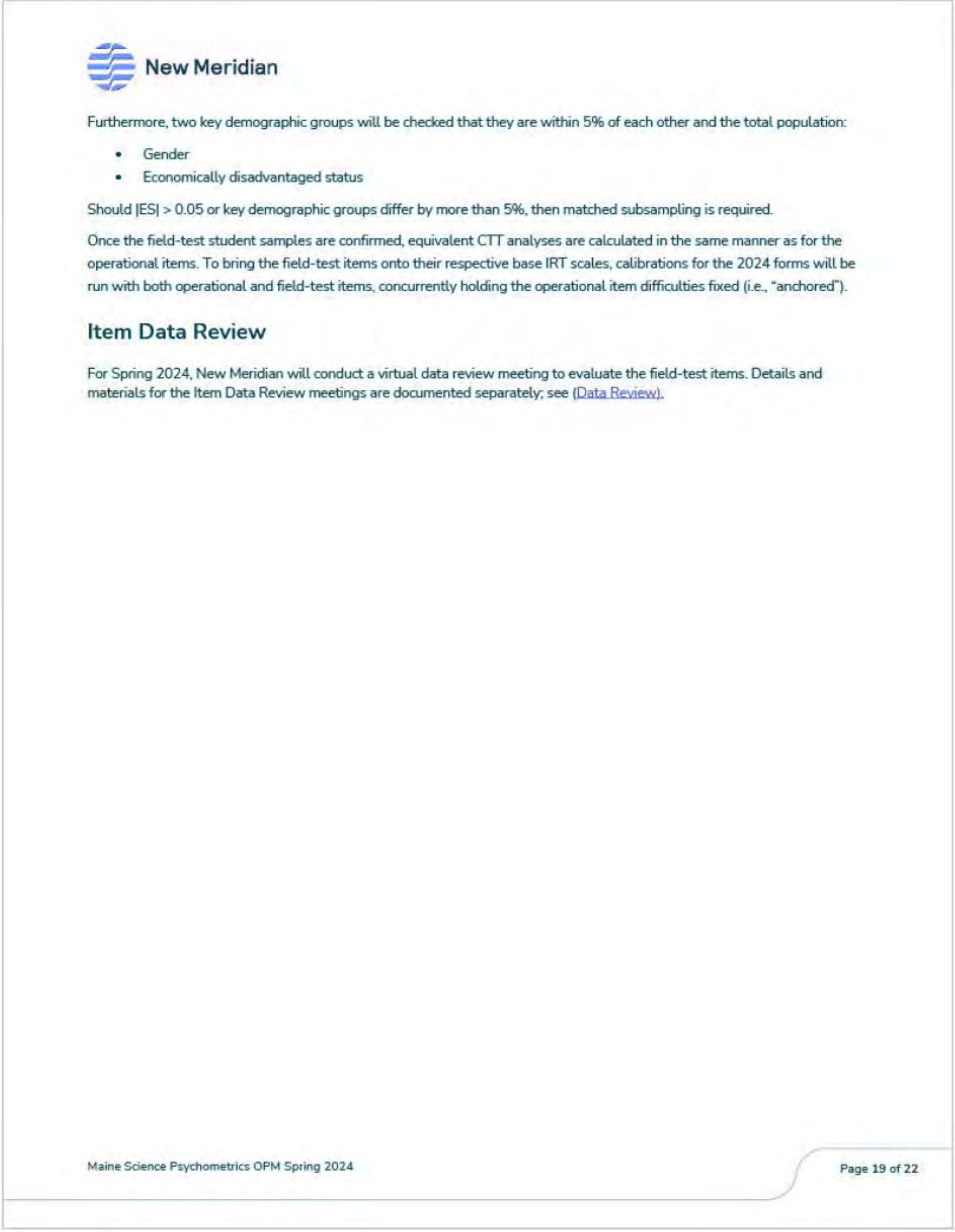
Students are randomly assigned a field-test session as they are rostered. Therefore, each field-test session will have an approximate sample size of 5,000 students.

The field-test sessions will be evaluated for baseline equivalence before classical item analysis and IRT analysis. Overall score effect sizes (ES), i.e., standardized mean differences, will be computed using Hedge's g;

$$ES = g = \frac{y_{FT1} - y_{FT2}}{\sqrt{\frac{(n_{FT1} - 1)s_{FT1}^2 + (n_{FT2} - 1)s_{FT2}^2}{n_{FT1} + n_{FT2} - 2}}}$$

where y_{F1} and y_{F2} are the overall score means for the operational items for each field-test block in the session, n_{F1} and n_{F2} are the student sample sizes, and s_{FT1} and s_{FT2} are the student-level standard deviations.

The field-test samples are satisfactory if 0.00 ≤ |ES| ≤ 0.05.



Furthermore, two key demographic groups will be checked that they are within 5% of each other and the total population:

- Gender
- Economically disadvantaged status

Should |ES| > 0.05 or key demographic groups differ by more than 5%, then matched subsampling is required.

Once the field-test student samples are confirmed, equivalent CTT analyses are calculated in the same manner as for the operational items. To bring the field-test items onto their respective base IRT scales, calibrations for the 2024 forms will be run with both operational and field-test items, concurrently holding the operational item difficulties fixed (i.e., "anchored").

Item Data Review

For Spring 2024, New Meridian will conduct a virtual data review meeting to evaluate the field-test items. Details and materials for the Item Data Review meetings are documented separately; see (Data Review).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Dorans, N. & Holland, P. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Golia, S. (2012). Differential item functioning classification for polytomously scored items. *Electronic Journal of Applied Statistical Analysis*, *5*, 367–373.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika*, *28*(2), 211–218.

Holland, P. & Thayer, D. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Lemke, E., & Wiersma, W. (1976). *Principles of psychological measurement*. Chicago, Ill: McNally.

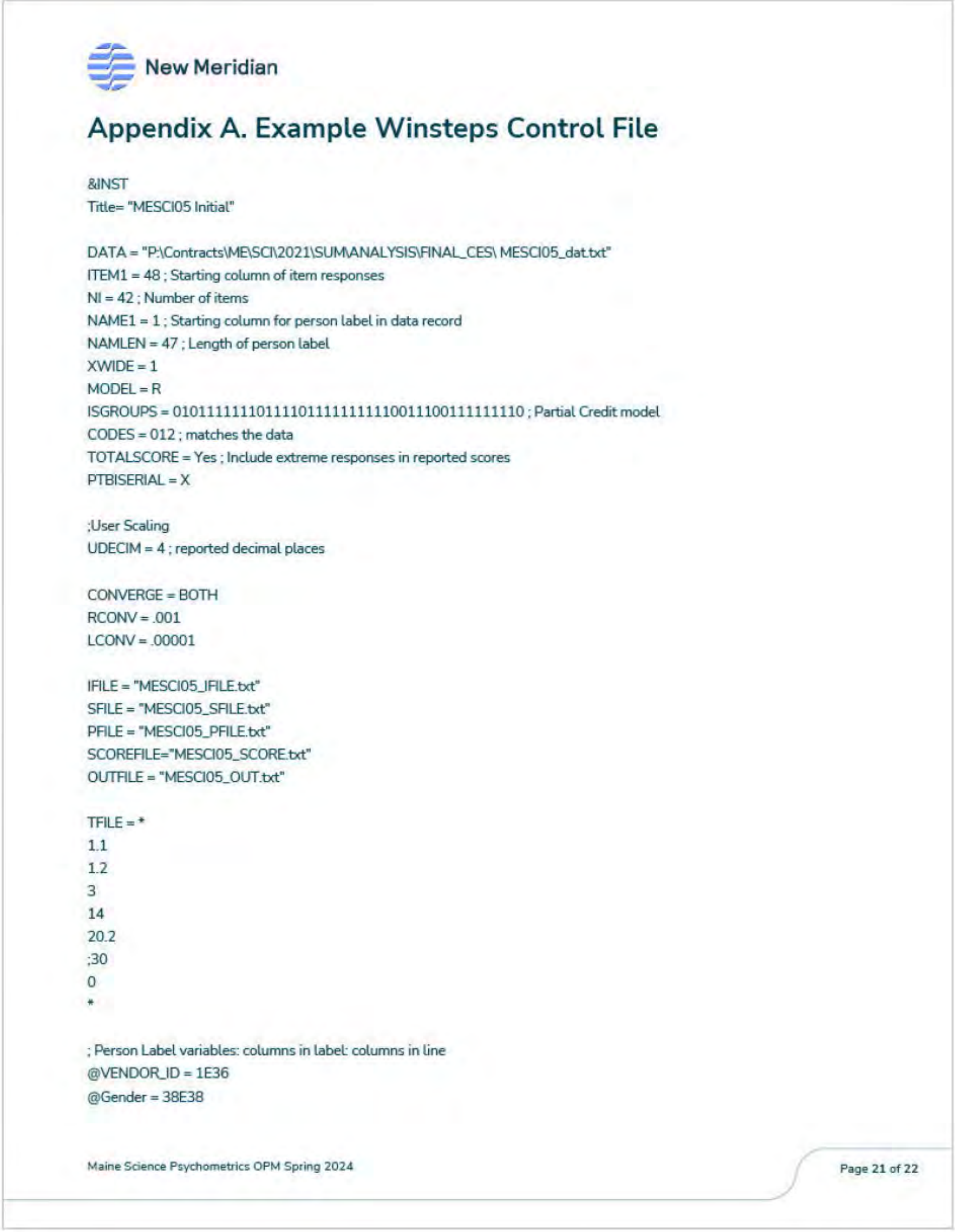
Linacre, J.M. (2023). Winsteps® (Version 5.6.1.0) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690–700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

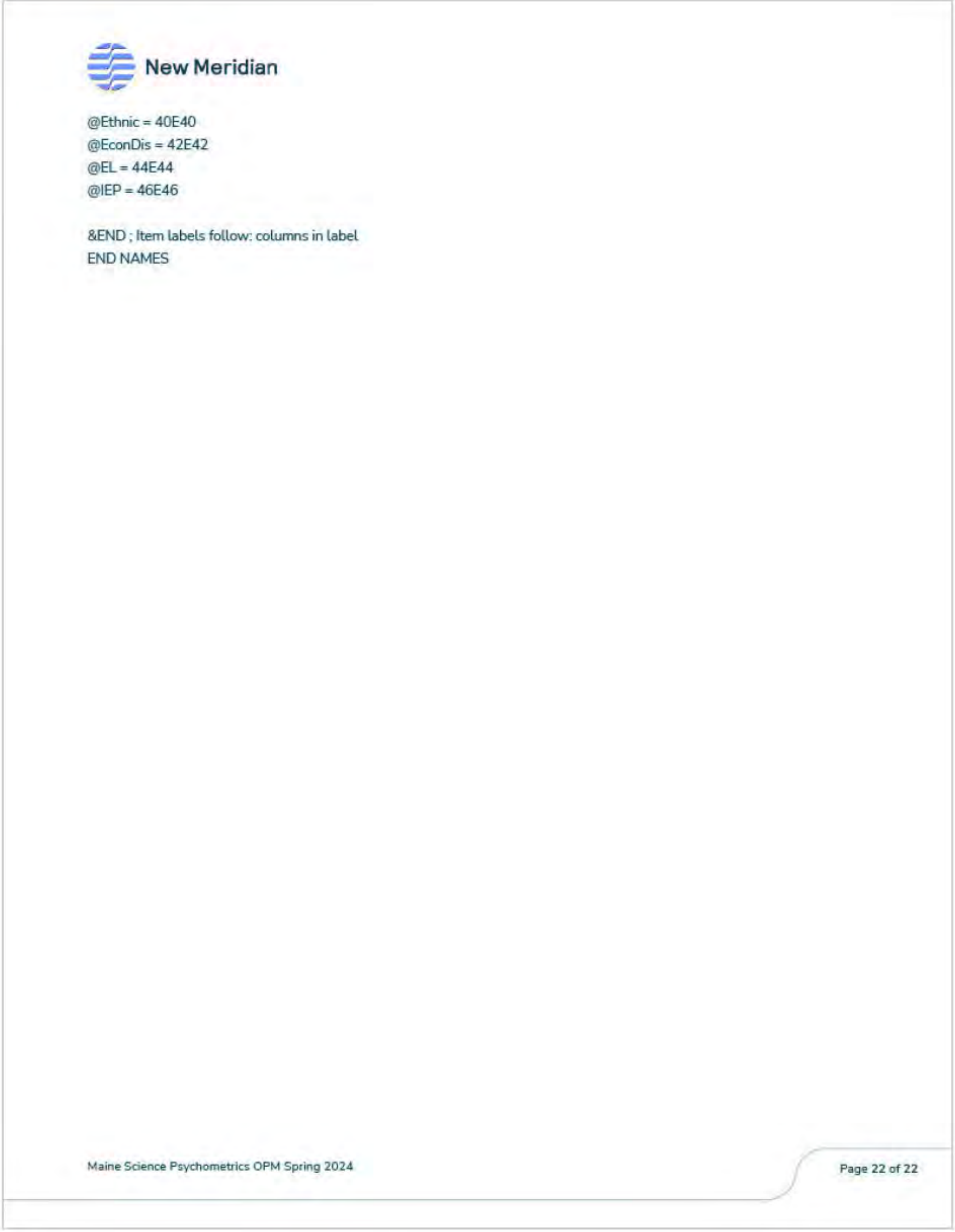
Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*, 149–174.

Rasch, G. (1960). *Probablistic models for same intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.



Appendix A. Example Winsteps Control File

```
&INST
Title= "MESCI05 Initial"

DATA = "P:\Contracts\ME\SCI\2021\SUM\ANALYSIS\FINAL_CES\ MESCI05_dat.txt"
ITEM1 = 48 ; Starting column of item responses
NI = 42 ; Number of items
NAME1 = 1 ; Starting column for person label in data record
NAMLEN = 47 ; Length of person label
XWIDE = 1
MODEL = R
ISGROUPS = 010111111101111011111111110011100111111110 ; Partial Credit model
CODES = 012 ; matches the data
TOTALSCORE = Yes ; Include extreme responses in reported scores
PTBISERIAL = X

;User Scaling
UDECIM = 4 ; reported decimal places

CONVERGE = BOTH
RCONV = .001
LCONV = .00001

IFILE = "MESCI05_IFILE.txt"
SFILE = "MESCI05_SFILE.txt"
PFILE = "MESCI05_PFILE.txt"
SCOREFILE="MESCI05_SCORE.txt"
OUTFILE = "MESCI05_OUT.txt"

TFILE = *
1.1
1.2
3
14
20.2
;30
0
*

; Person Label variables: columns in label: columns in line
@VENDOR_ID = 1E36
@Gender = 38E38
```



```
@Ethnic = 40E40
@EconDis = 42E42
@EL = 44E44
@IEP = 46E46

&END ; Item labels follow: columns in label
END NAMES
```

Appendix J. Classical Item Analysis

Table 60. Grade 5 Item-Level Classical Test Theory Statistics

Item UIN	Type	p-value	Item Mean	Item-Total Correlation	Proportion Omit
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Item UIN	Type	p-value	Item Mean	Item-Total Correlation	Proportion Omit
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 61. Grade 8 Item-Level Classical Test Theory Statistics

Item UIN	Type	p-value	Item Mean	Item-Total Correlation	Proportion Omit
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Item UIN	Type	p-value	Item Mean	Item-Total Correlation	Proportion Omit
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 62. High School Item-Level Classical Test Theory Statistics

Item UIN	Type	p-value	Item Mean	Item-Total Correlation	Proportion Omit
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Item UIN	Type	p-value	Item Mean	Item-Total Correlation	Proportion Omit
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Appendix K. Differential Item Functioning Results

Table 63. Grade 5 Items Classified with DIF by Focal Group

Reference ID	Form Name	C-DIF	Female	Asian	Black	Hisp.	Multi-Race	ML	SWD	Econ. Dis.
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 64. Grade 8 Items Classified with DIF by Focal Group

Reference ID	Form Name	C-DIF	Female	Asian	Black	Hisp.	Multi-Race	ML	SWD	Econ. Dis.
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 65. High School Items Classified with DIF by Focal Group

Reference ID	Form Name	C-DIF	Female	Asian	Black	Hisp.	Multi-Race	ML	SWD	Econ. Dis.
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Reference ID	Form Name	C-DIF	Female	Asian	Black	Hisp.	Multi-Race	ML	SWD	Econ. Dis.
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Appendix L. Item Response Theory Parameters

Table 66. Grade 5 IRT Parameters

Item UIN	b	s.e.[8]	d1	d2	d3
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[8] Standard error.

Item UIN	b	s.e.[8]	d1	d2	d3
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 67. Grade 8 IRT Parameters

Item UIN	b	s.e.	d1	d2	d3
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Item UIN	b	s.e.	d1	d2	d3
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 68. High School IRT Parameters

Item UIN	b	s.e.	d1	d2	d3
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Item UIN	b	s.e.	d1	d2	d3
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Appendix M. Assessment IRT Curves

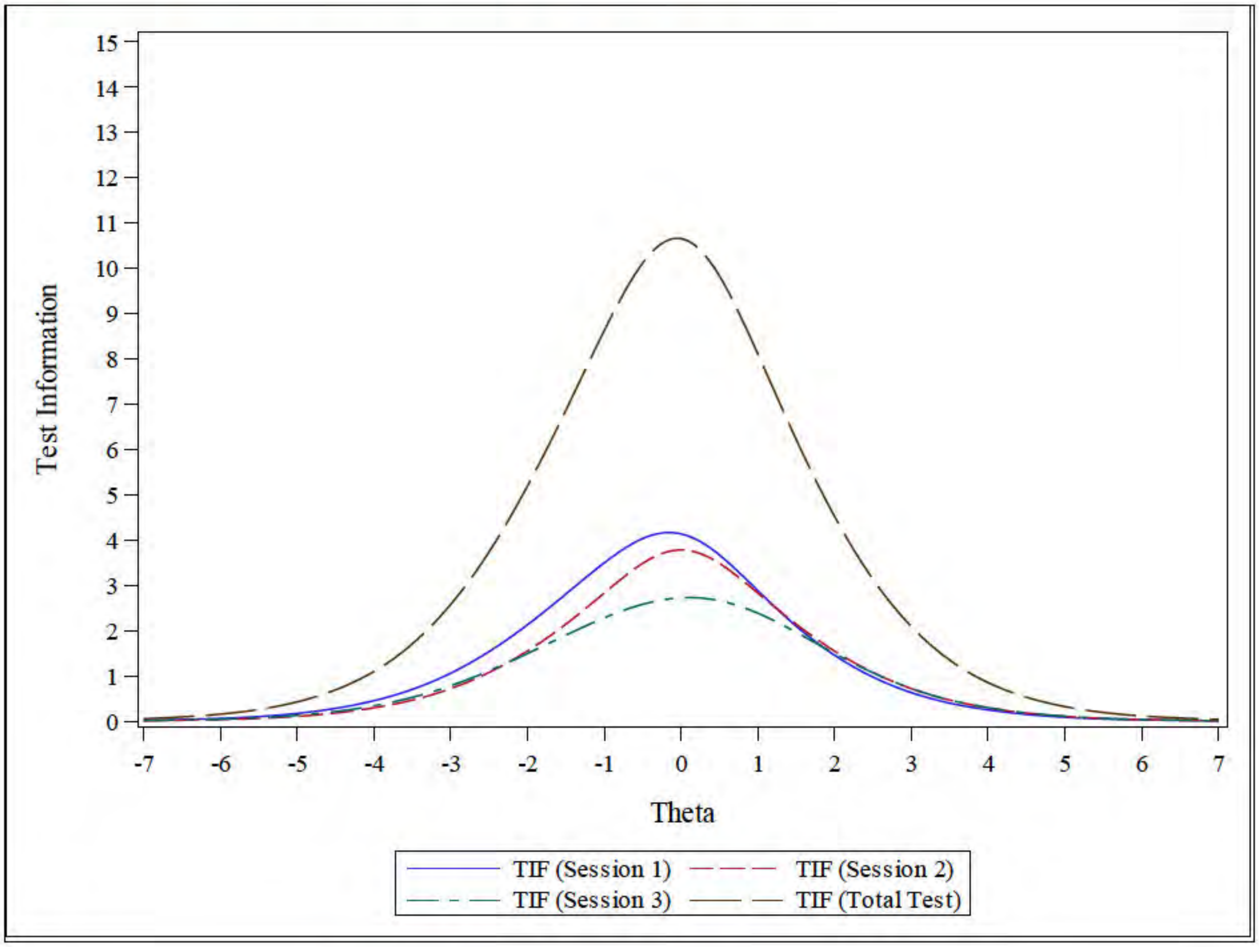


Figure 8. Grade 5 Test Information Function for 2024 Form

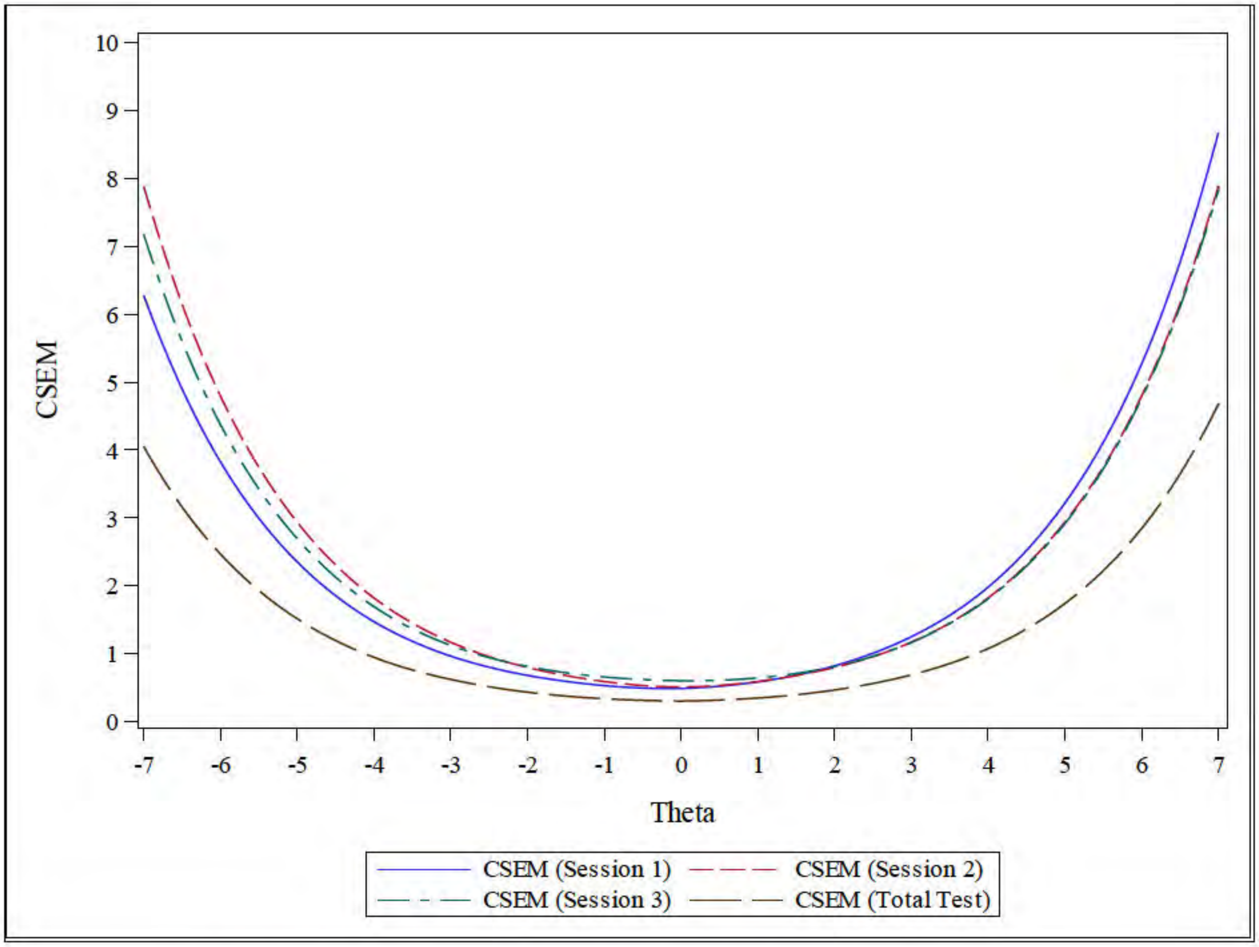


Figure 9. Grade 5 Conditional Standard Error of Measurement for 2024 Form

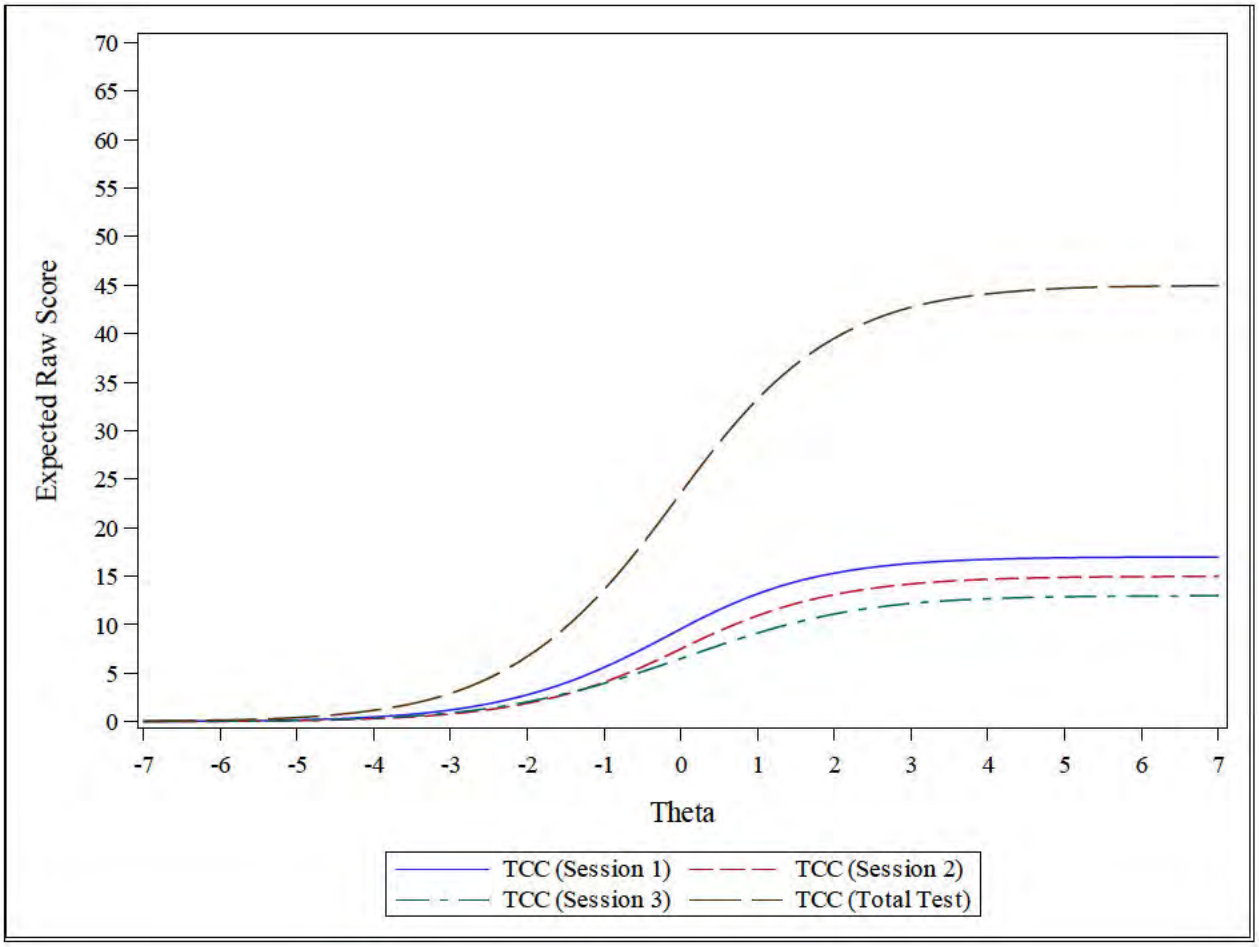


Figure 10. Grade 5 Test Characteristic Curve for 2024 Form

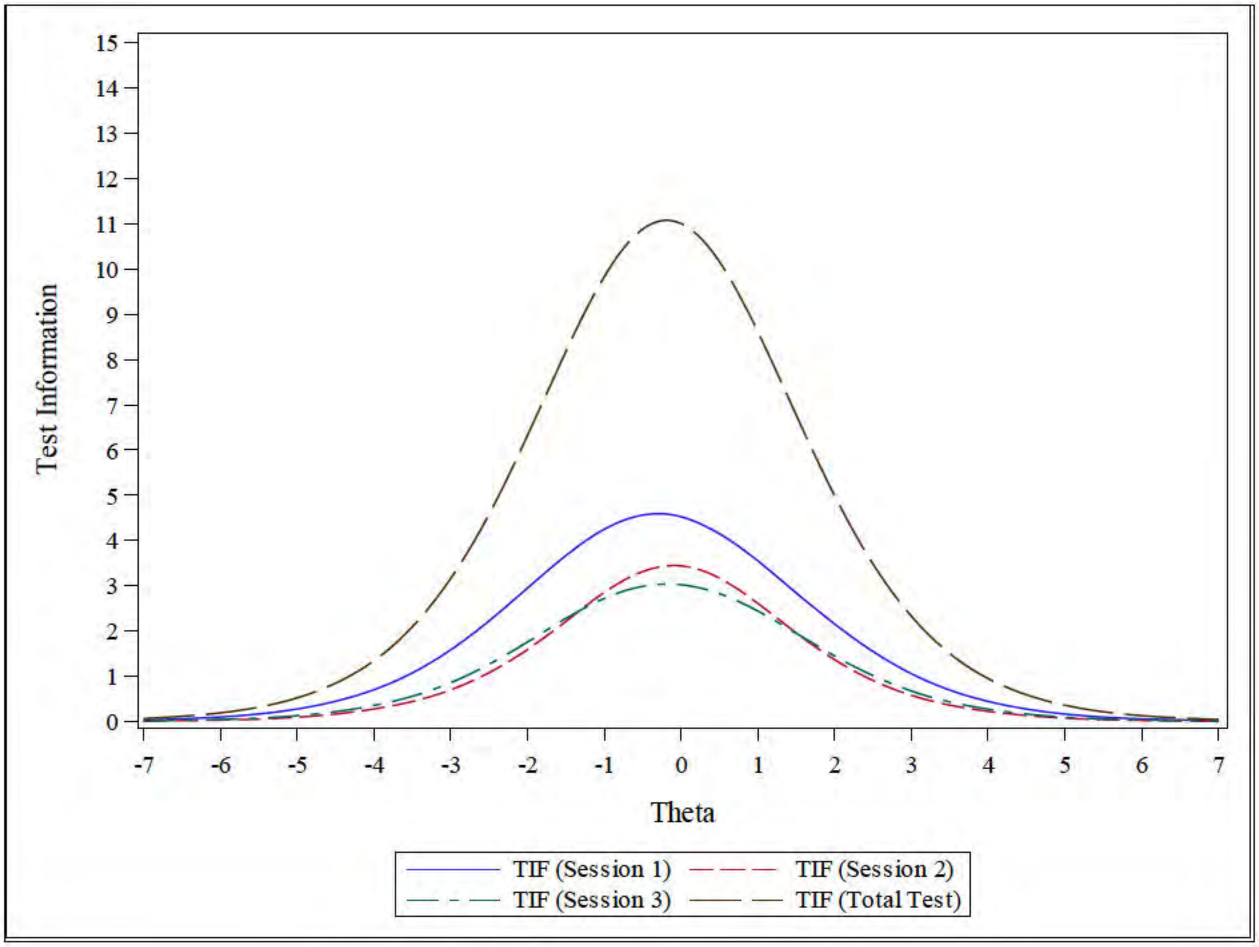


Figure 11. Grade 8 Test Information Function for 2024 Form

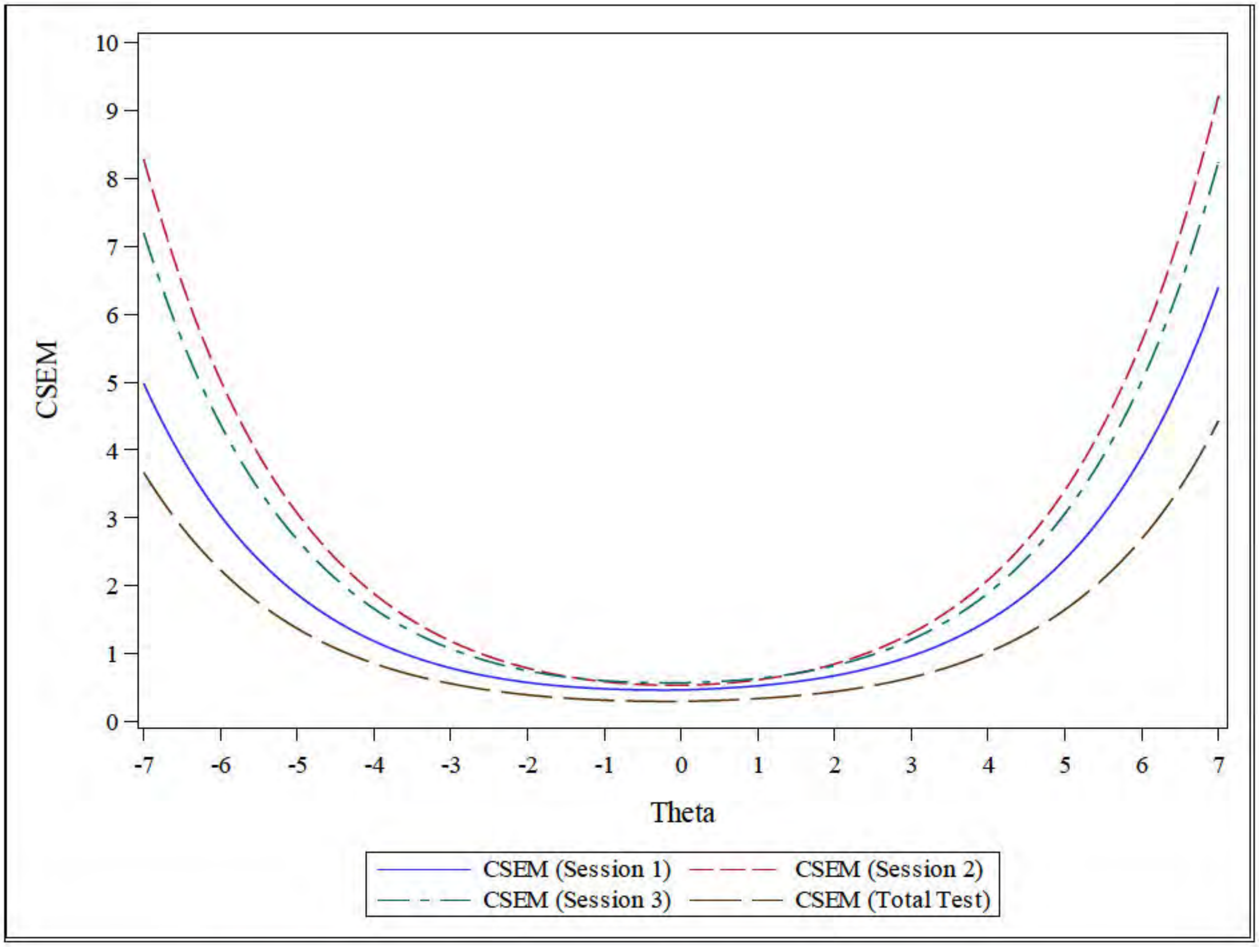


Figure 12. Grade 8 Conditional Standard Error of Measurement for 2024 Form

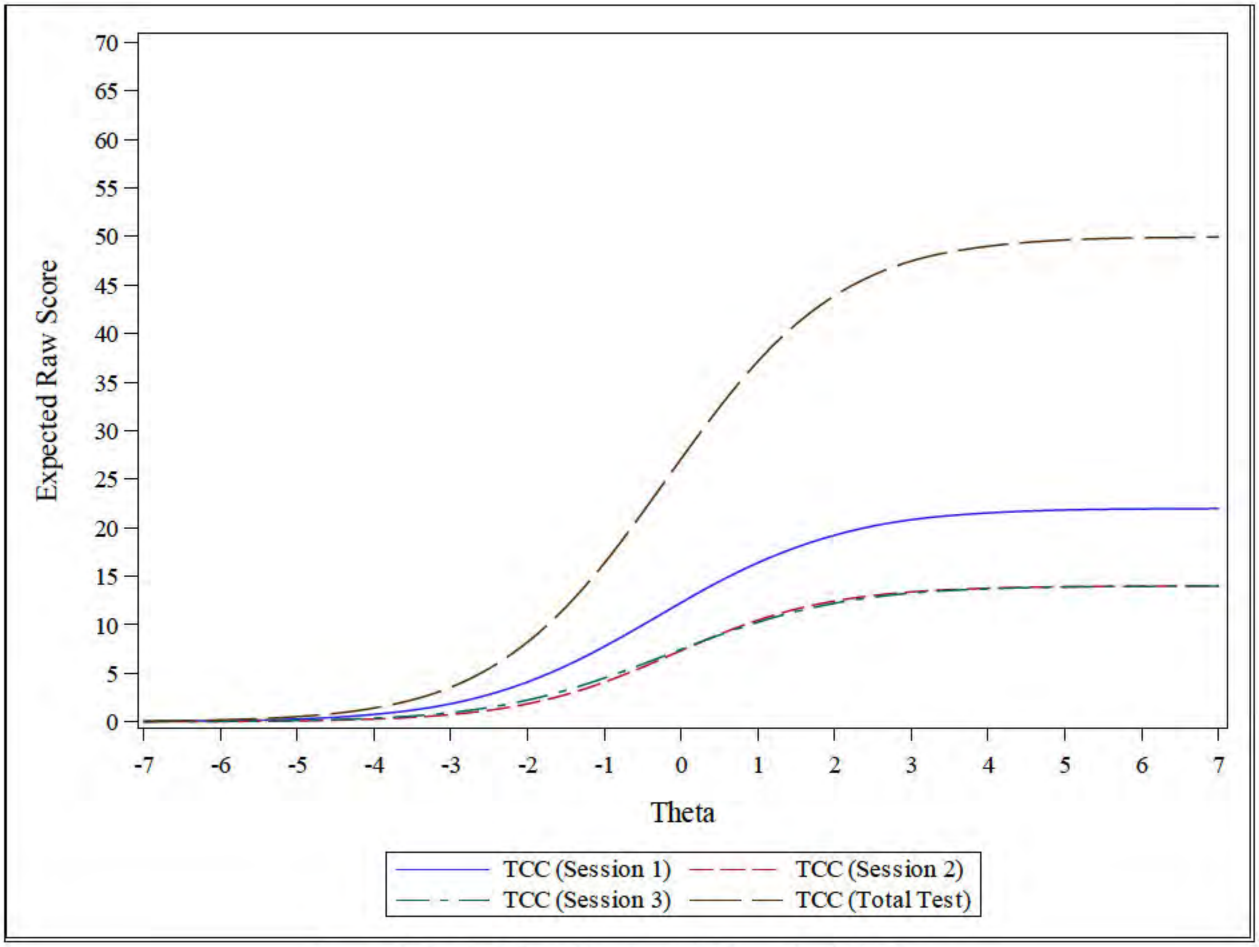


Figure 13. Grade 8 Test Characteristic Curve for 2024 Form

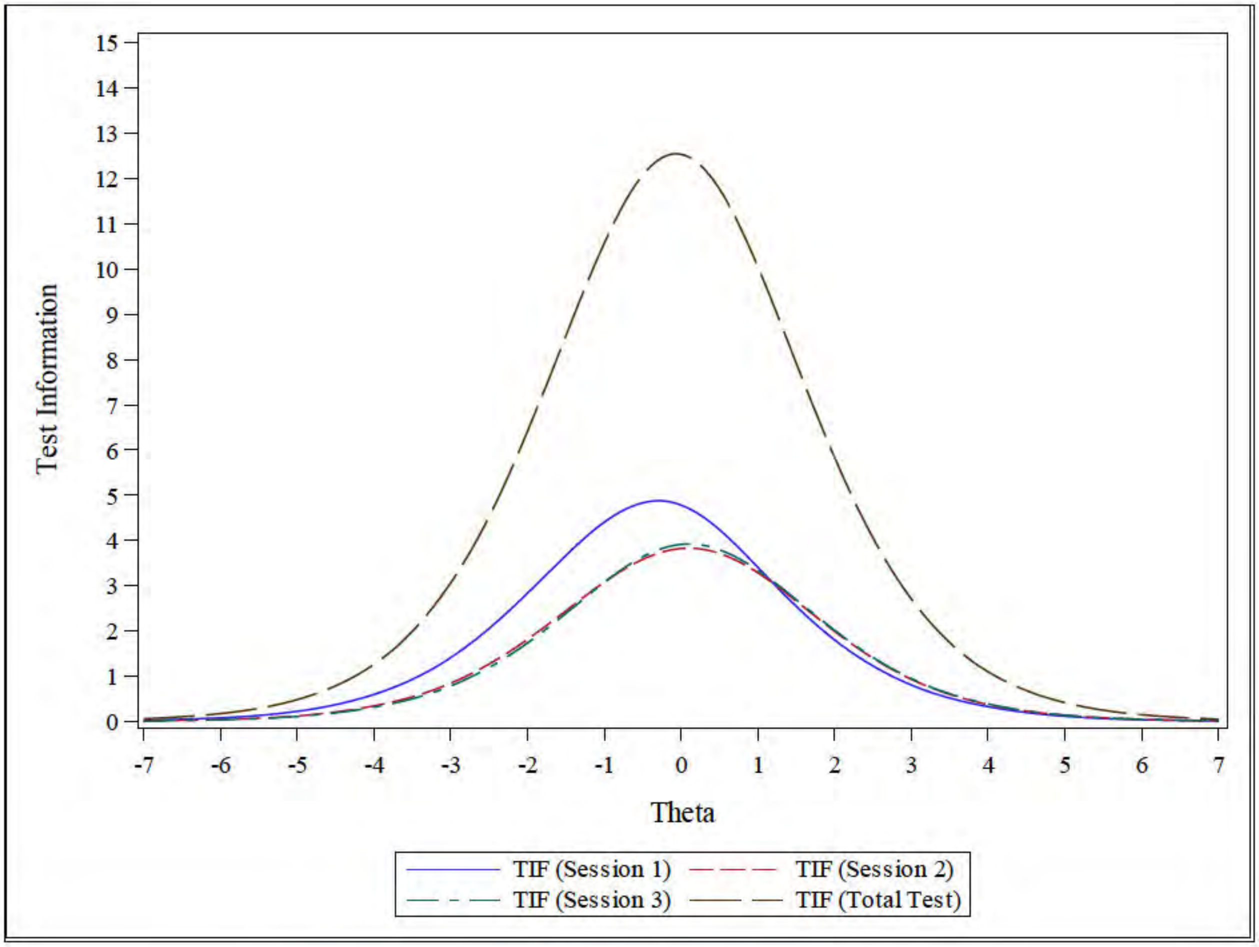


Figure 14. High School Test Information Function for 2024 Form

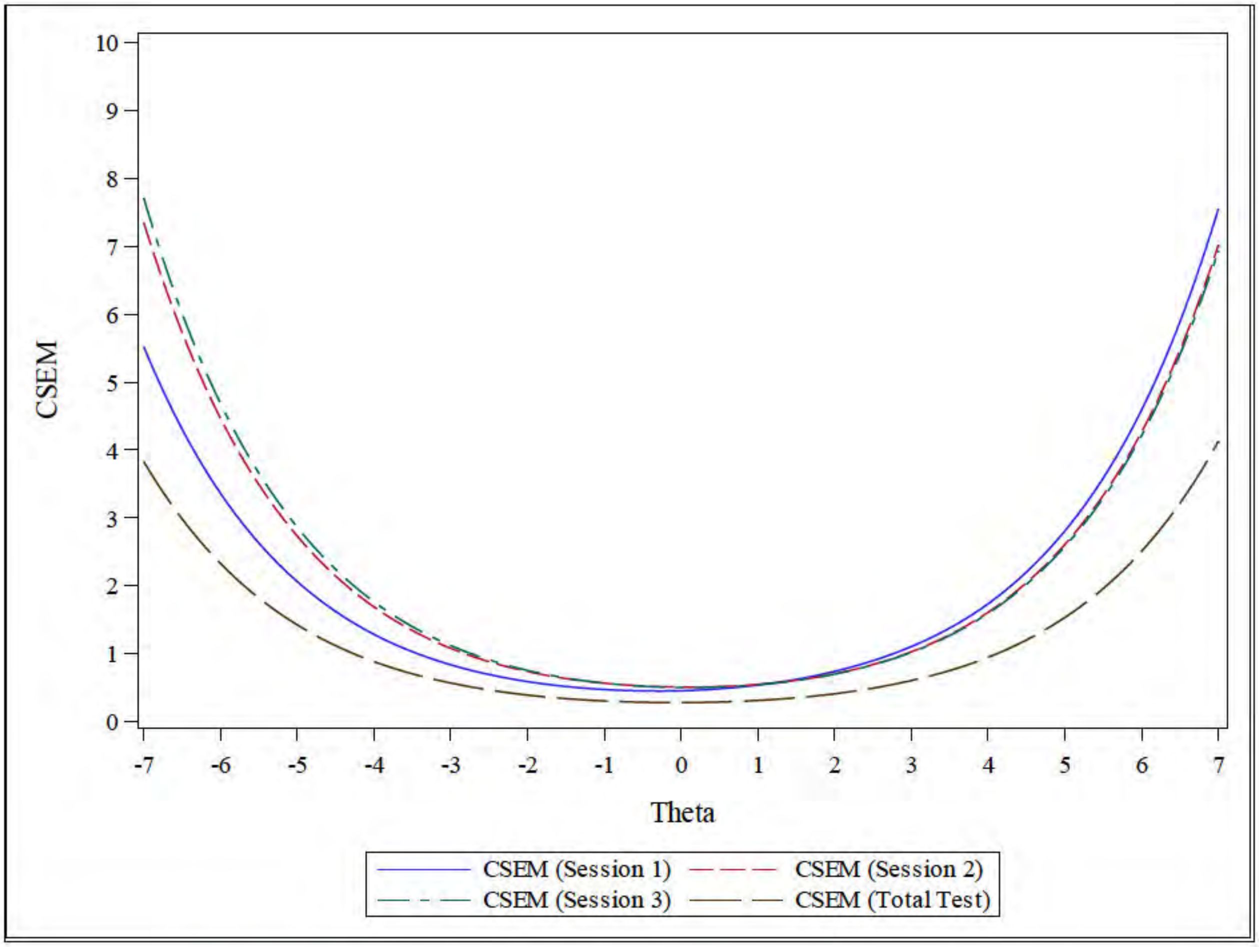


Figure 15. High School Conditional Standard Error of Measurement for 2024 Form

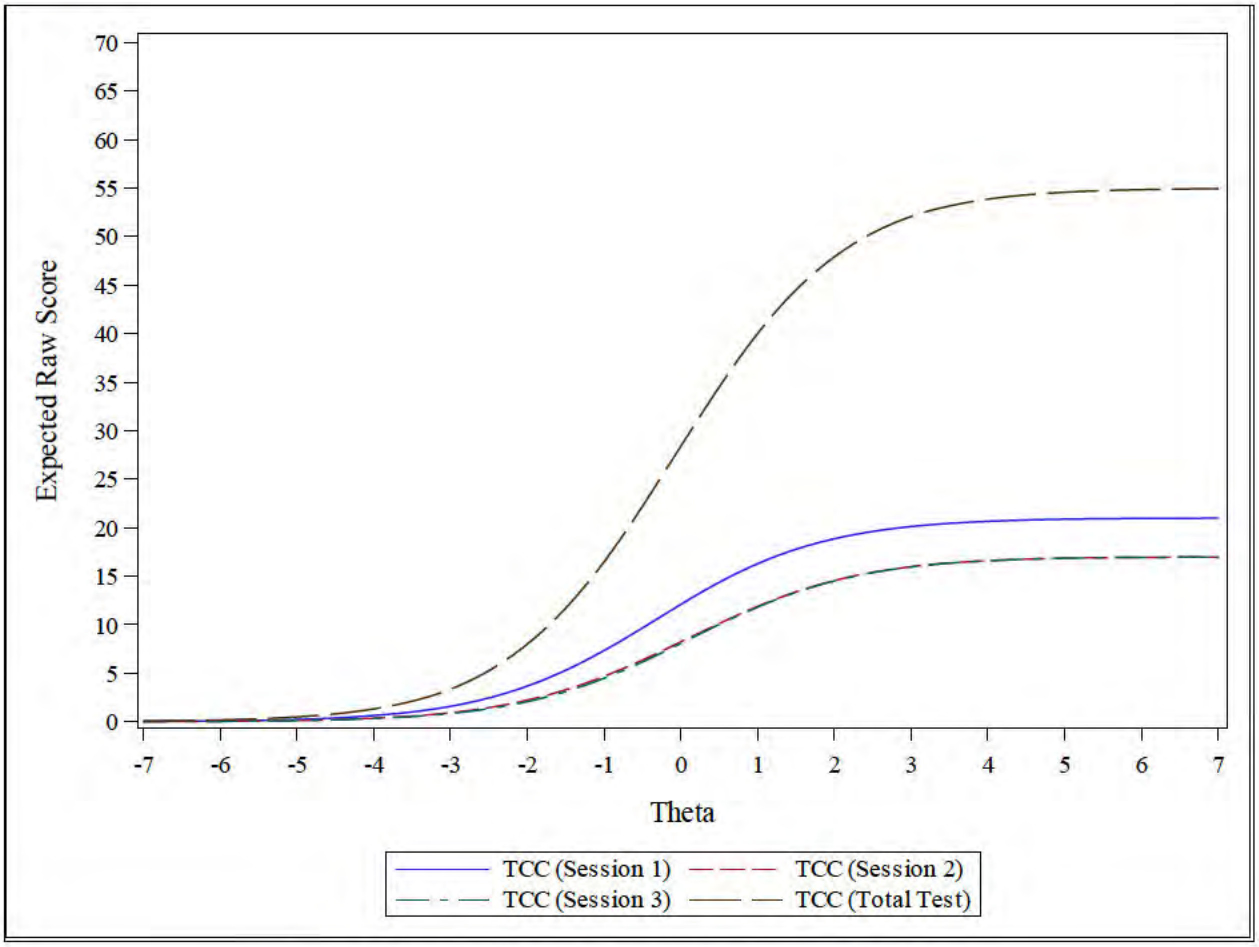


Figure 16. High School Test Characteristic Curve for 2024 Form

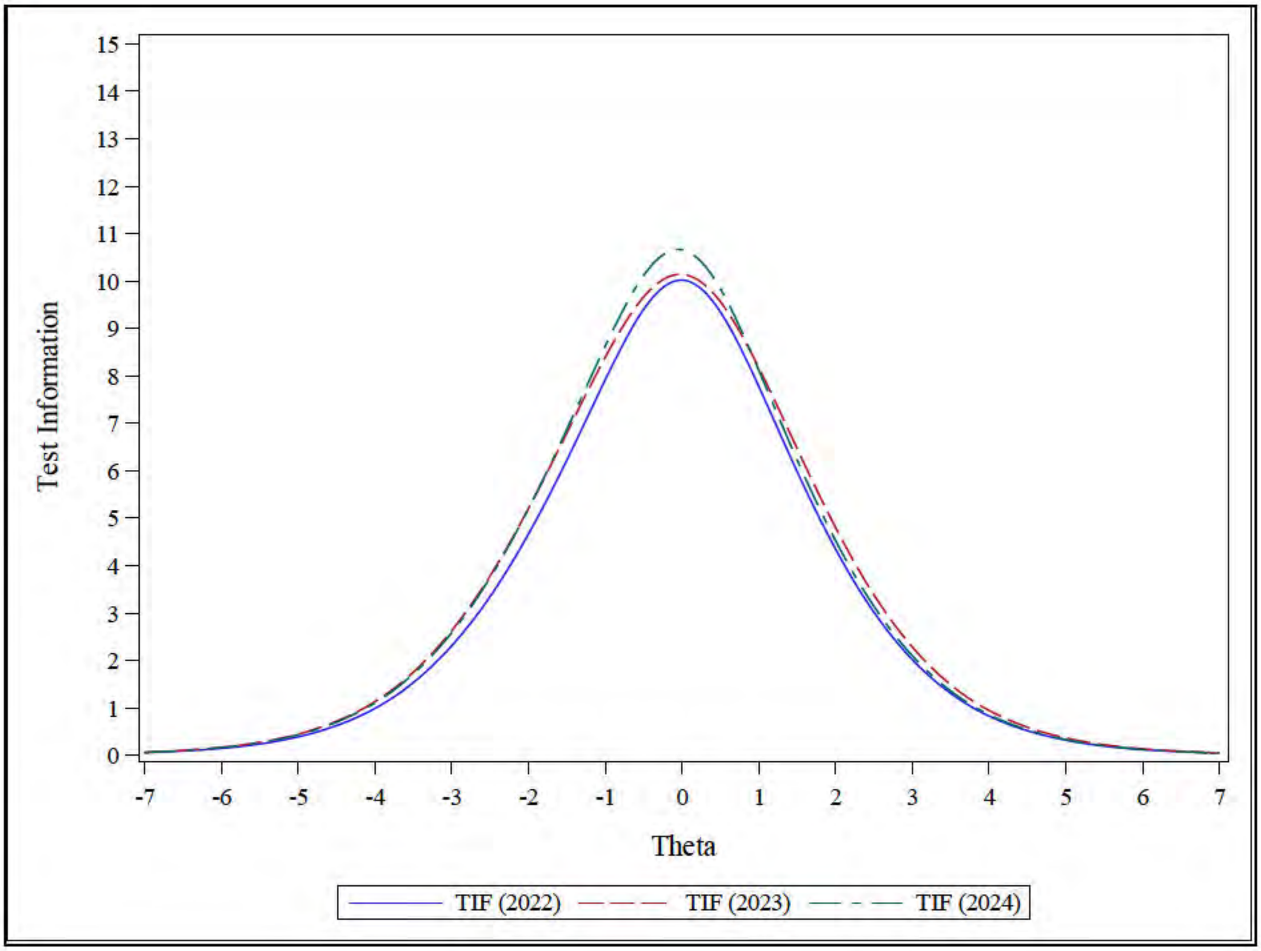


Figure 17. Grade 5 Test Information Functions 2022-2024

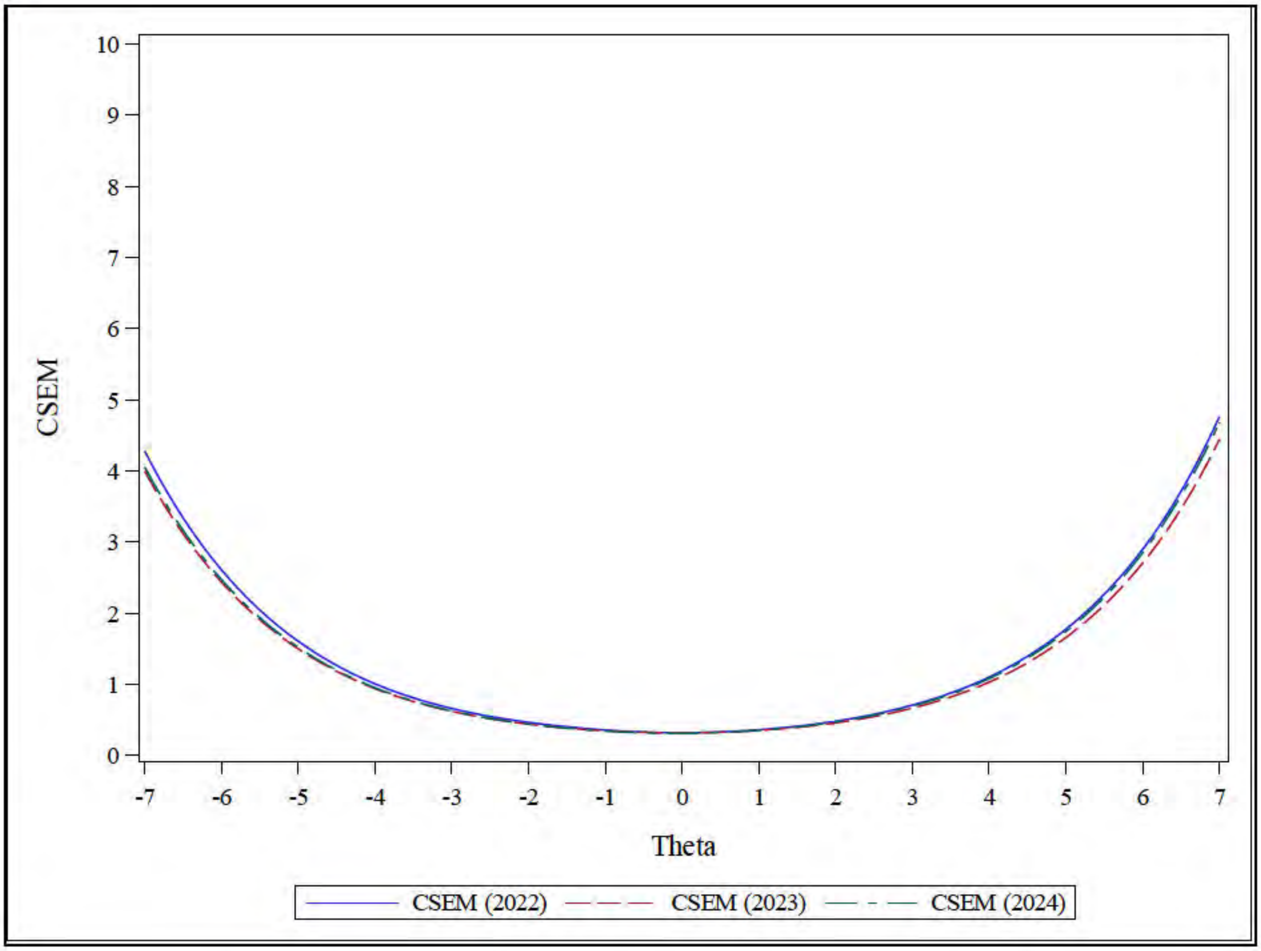


Figure 18. Grade 5 Conditional Standard Errors of Measurement 2022-2024

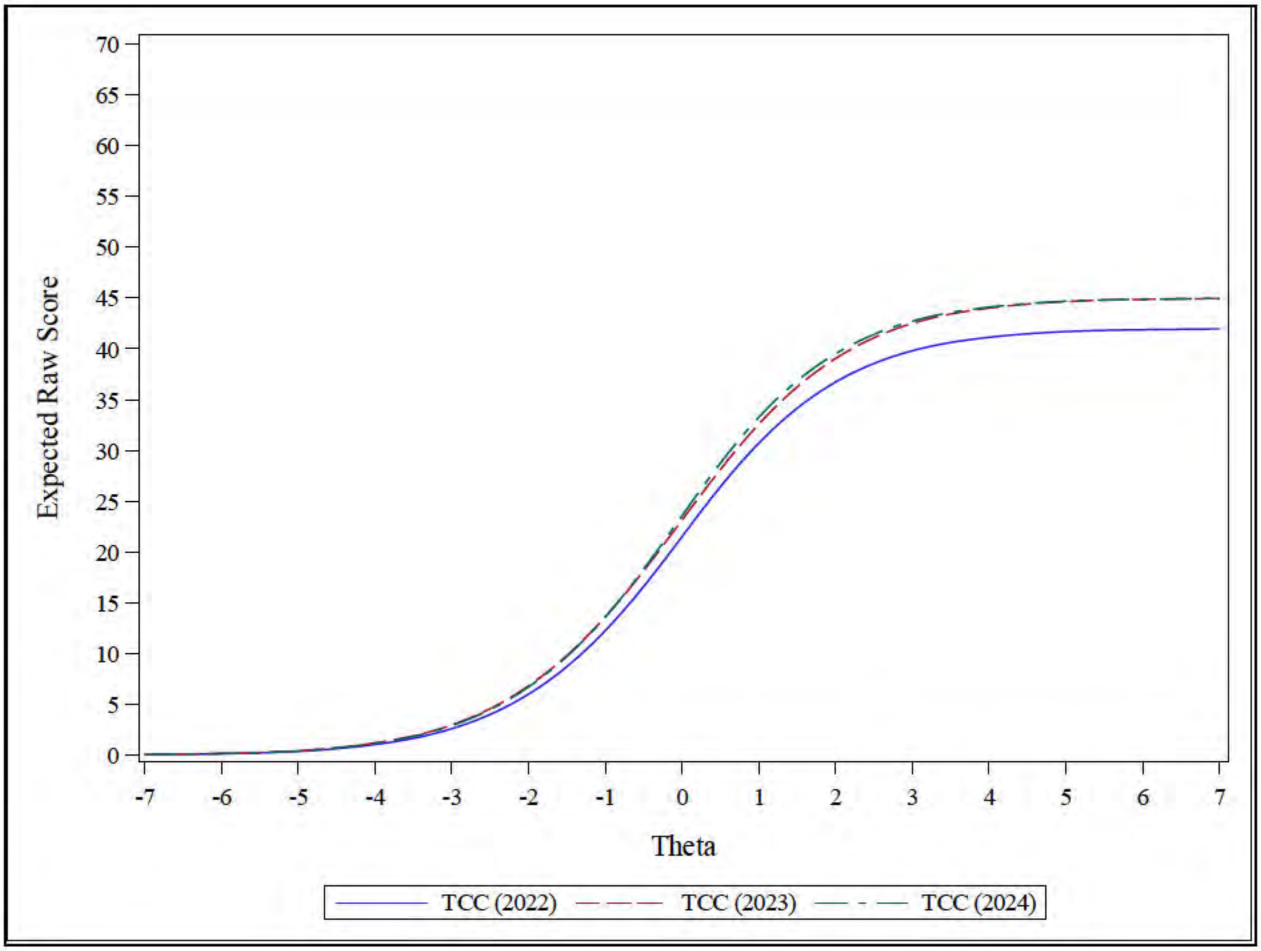


Figure 19. Grade 5 Test Characteristic Curves 2022-2024

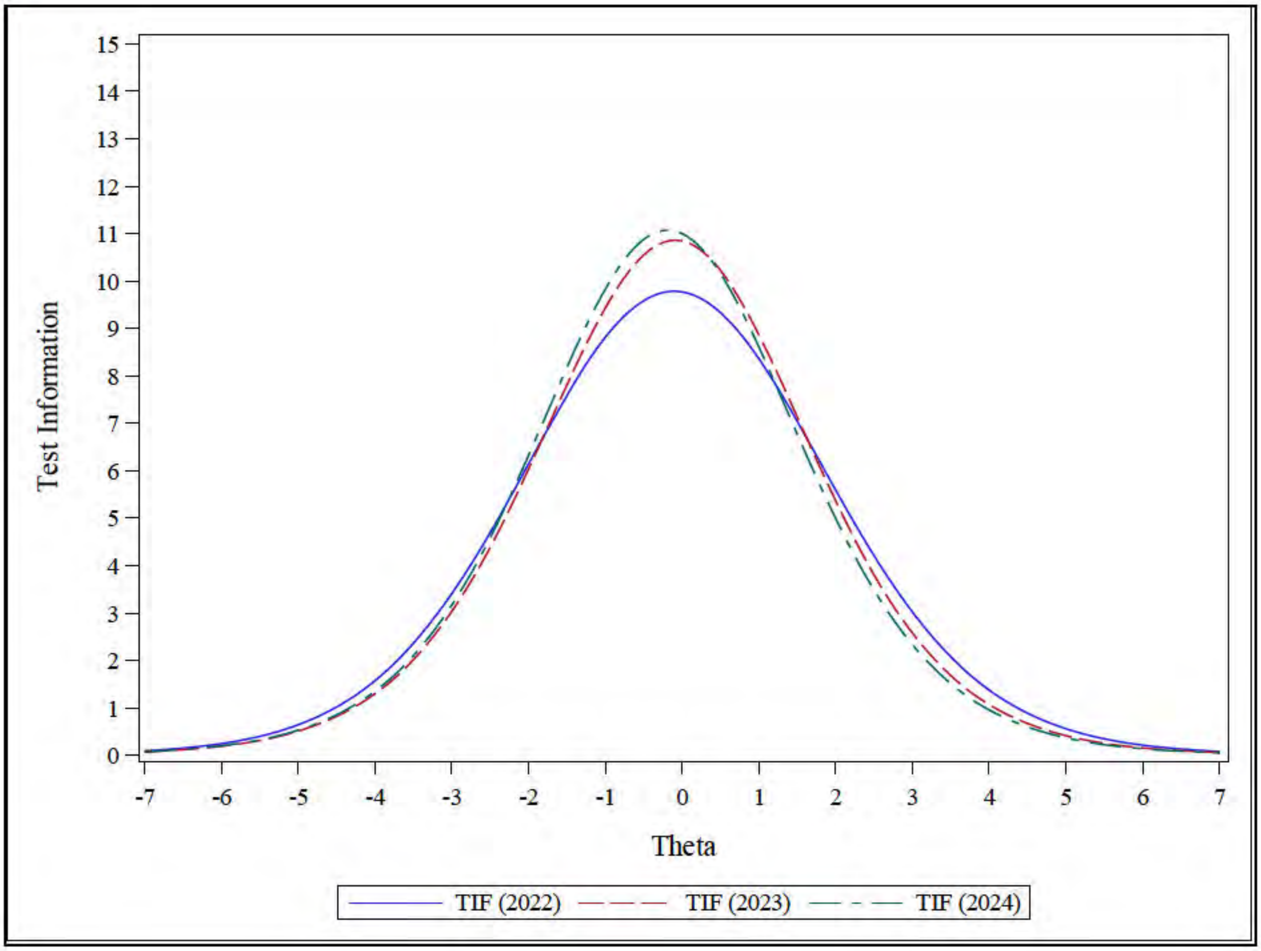


Figure 20. Grade 8 Test Information Functions 2022-2024

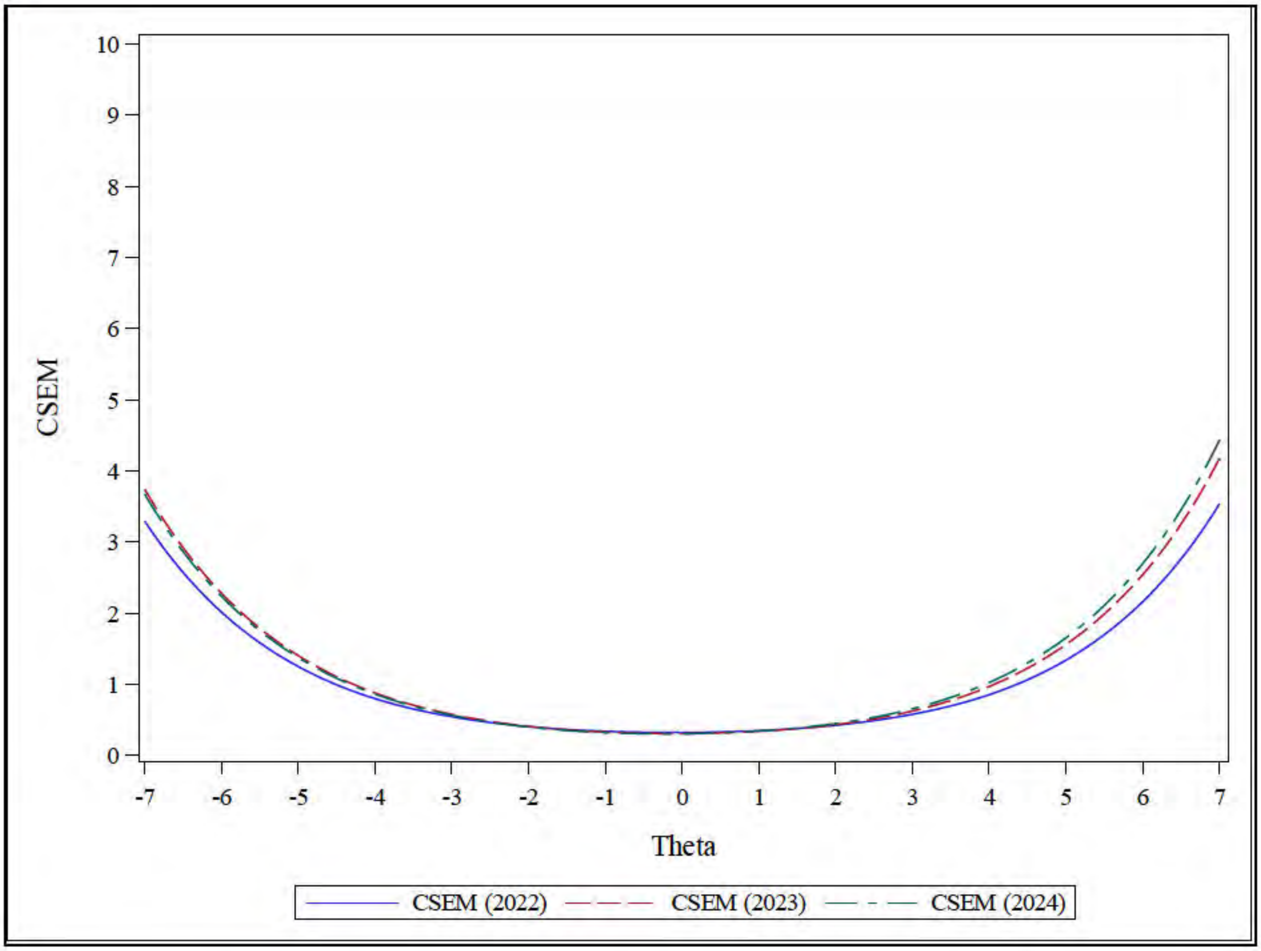


Figure 21. Grade 8 Conditional Standard Errors of Measurement 2022-2024

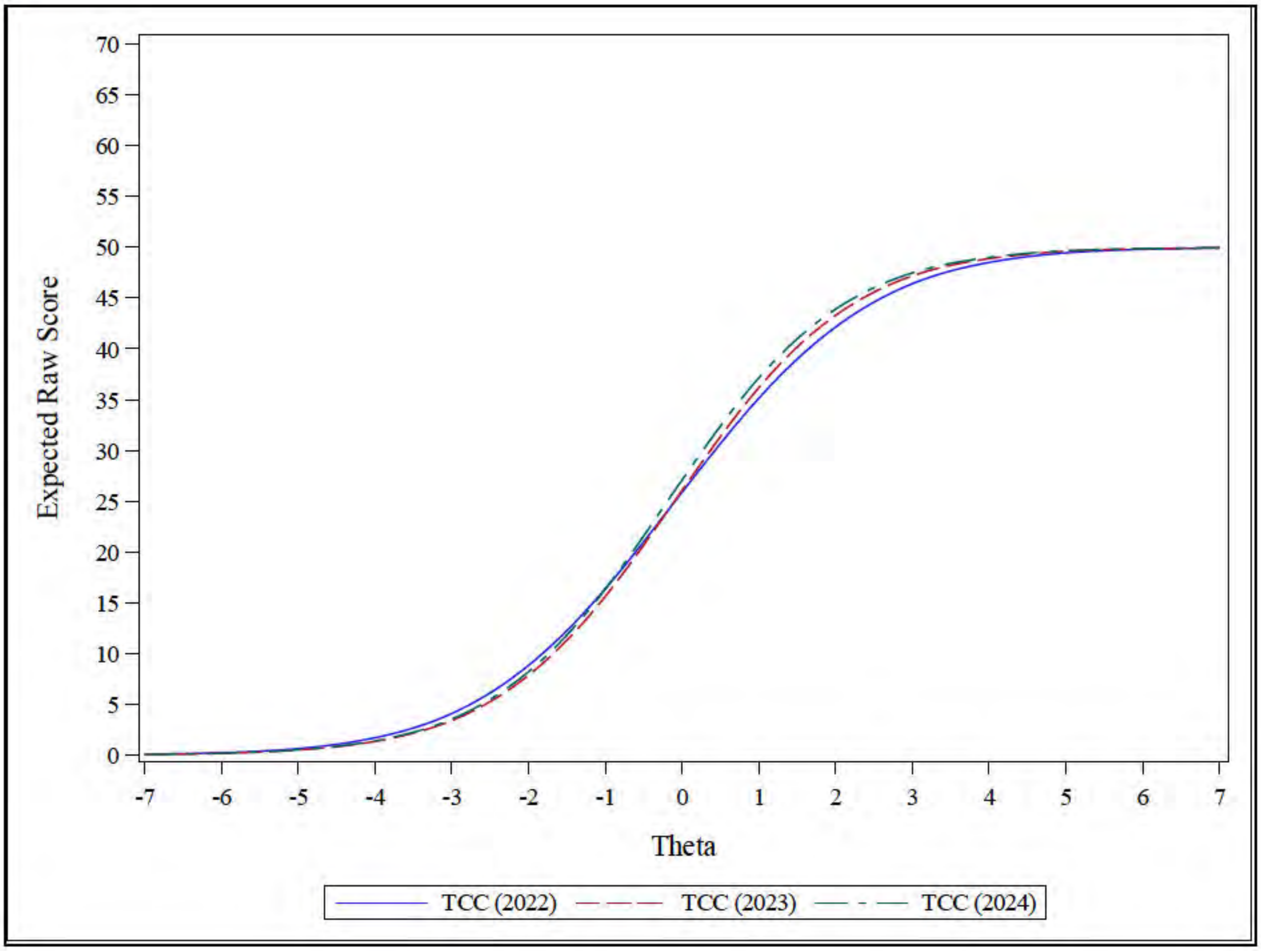


Figure 22. Grade 8 Test Characteristic Curves 2022-2024

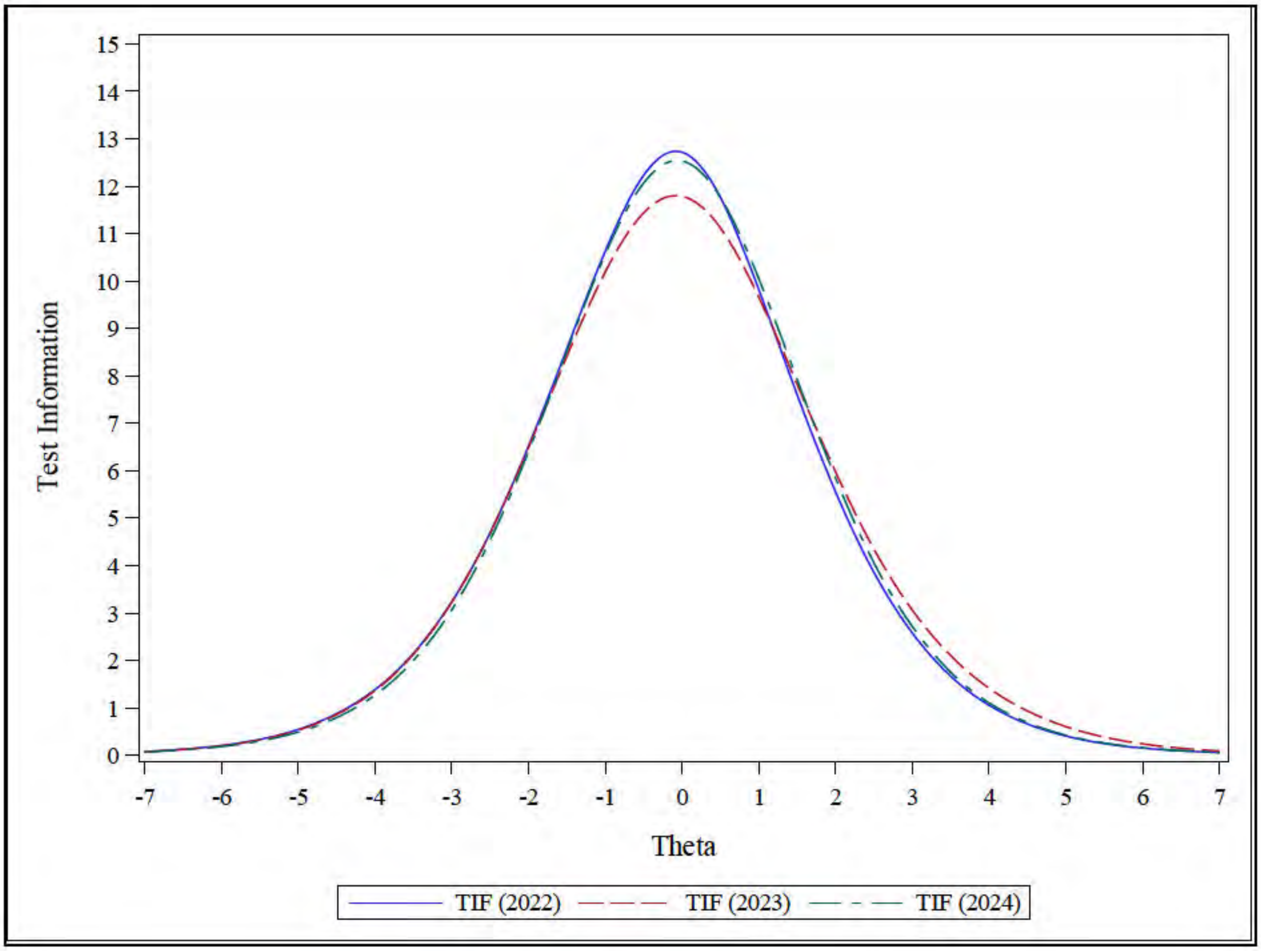


Figure 23. High School Test Information Functions 2022-2024

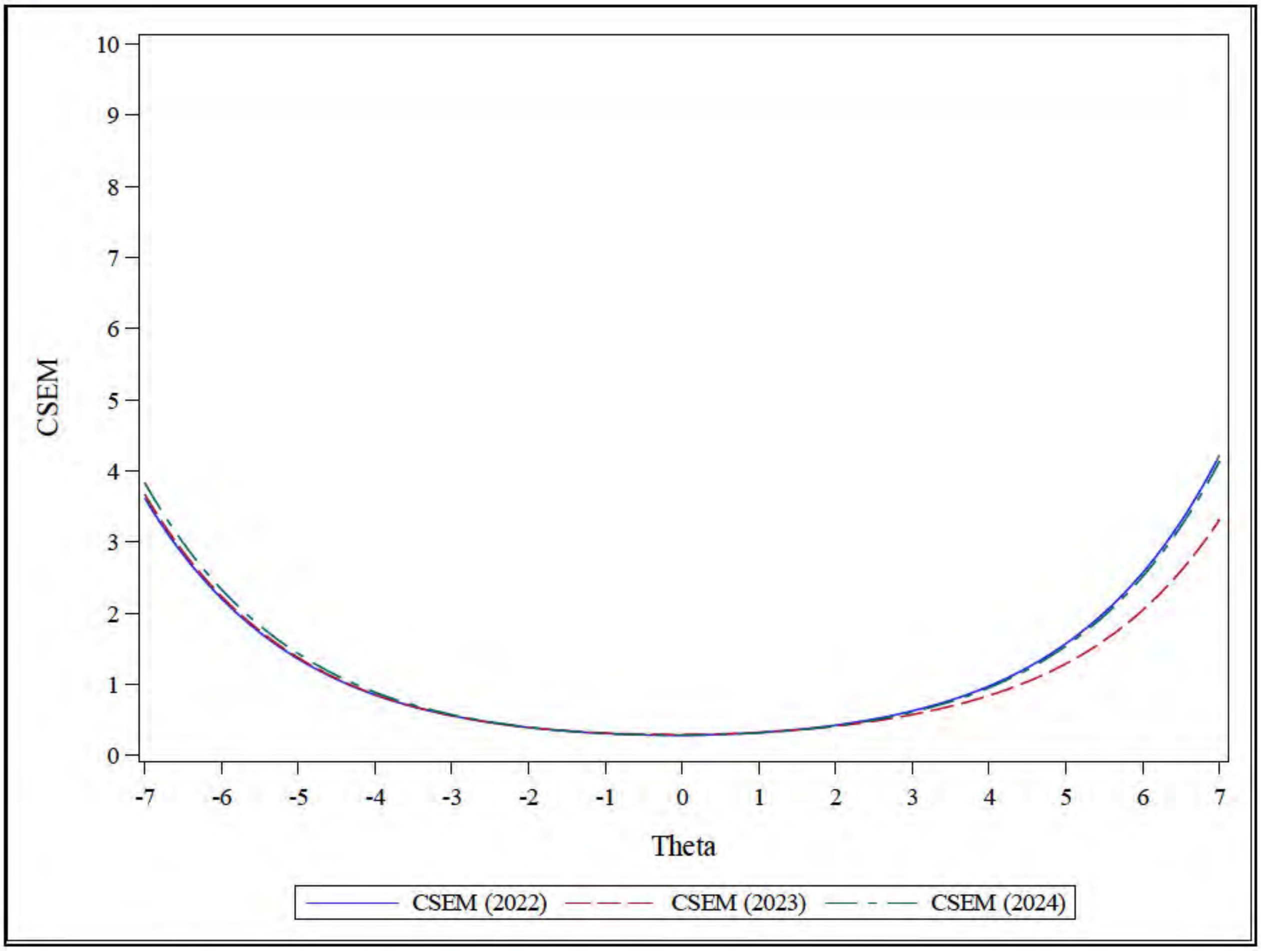


Figure 24. High School Conditional Standard Errors of Measurement 2022-2024

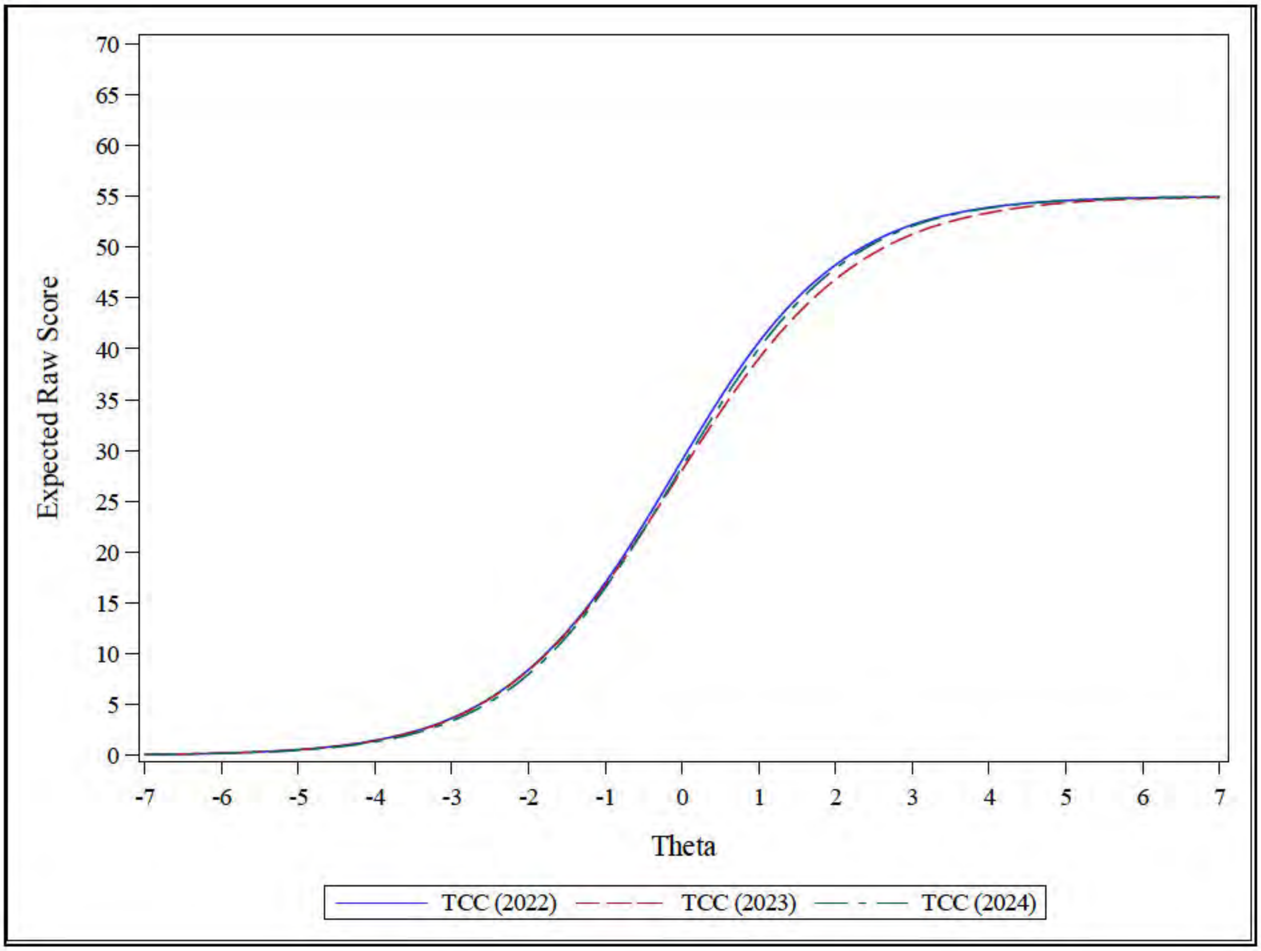


Figure 25. High School Test Characteristic Curves 2022-2024

Appendix N. Raw to Scaled score Tables

Table 69. Grade 5 Raw-to-Scaled Score Look-Up Table

Raw Score	Scaled Score	Scaled Score CSEM
0	6	7
1	6	7
2	11	5
3	15	4
4	17	4
5	19	3
6	21	3
7	23	3
8	24	3
9	25	2
10	26	2
11	27	2
12	28	2
13	29	2
14	30	2
15	31	2
16	32	2
17	33	2
18	33	2
19	34	2
20	35	2
21	36	2
22	36	2
23	37	2
24	38	2
25	39	2
26	39	2
27	40	2
28	41	2
29	42	2
30	42	2
31	43	2
32	44	2
33	45	2
34	46	2
35	47	2
36	48	2
37	49	2
38	50	3
39	52	3
40	53	3
41	55	3
42	58	4
43	61	5
44	67	7
45	80	13

Table 70. Grade 8 Raw-to-Scaled Score Look-Up Table

Raw Score	Scaled Score	Scaled Score CSEM
0	3	7
1	3	7
2	3	7
3	3	7
4	3	7
5	7	6
6	10	6
7	13	6
8	15	5
9	18	5
10	20	5
11	22	5
12	23	5
13	25	4
14	27	4
15	28	4
16	30	4
17	31	4
18	33	4
19	34	4
20	36	4
21	37	4
22	39	4
23	40	4
24	41	4
25	43	4
26	44	4
27	45	4
28	47	4
29	48	4
30	50	4
31	51	4
32	52	4
33	54	4
34	56	4
35	57	4
36	59	4
37	60	5
38	62	5
39	64	5
40	66	5
41	68	5
42	71	5
43	73	6
44	76	6
45	80	7
46	84	7
47	89	8
48	90	8
49	90	8
50	90	8

Table 71. High School Raw-to-Scaled Score Look-Up Table

Raw Score	Scaled Score	Scaled Score CSEM
0	5	8
1	5	8
2	11	5
3	14	4
4	17	4
5	19	3
6	21	3
7	22	3
8	23	3
9	25	3
10	26	2
11	27	2
12	28	2
13	29	2
14	29	2
15	30	2
16	31	2
17	32	2
18	33	2
19	33	2
20	34	2
21	35	2
22	36	2
23	36	2
24	37	2
25	38	2
26	38	2
27	39	2
28	40	2
29	40	2
30	41	2
31	42	2
32	43	2
33	43	2
34	44	2
35	45	2
36	46	2
37	47	2
38	47	2
39	48	2
40	49	2
41	50	2
42	51	2
43	52	2
44	53	3
45	55	3
46	56	3
47	57	3
48	59	3
49	60	3
50	62	4
51	65	4
52	68	5
53	72	6
54	78	8
55	90	14

Appendix O. Score Distributions

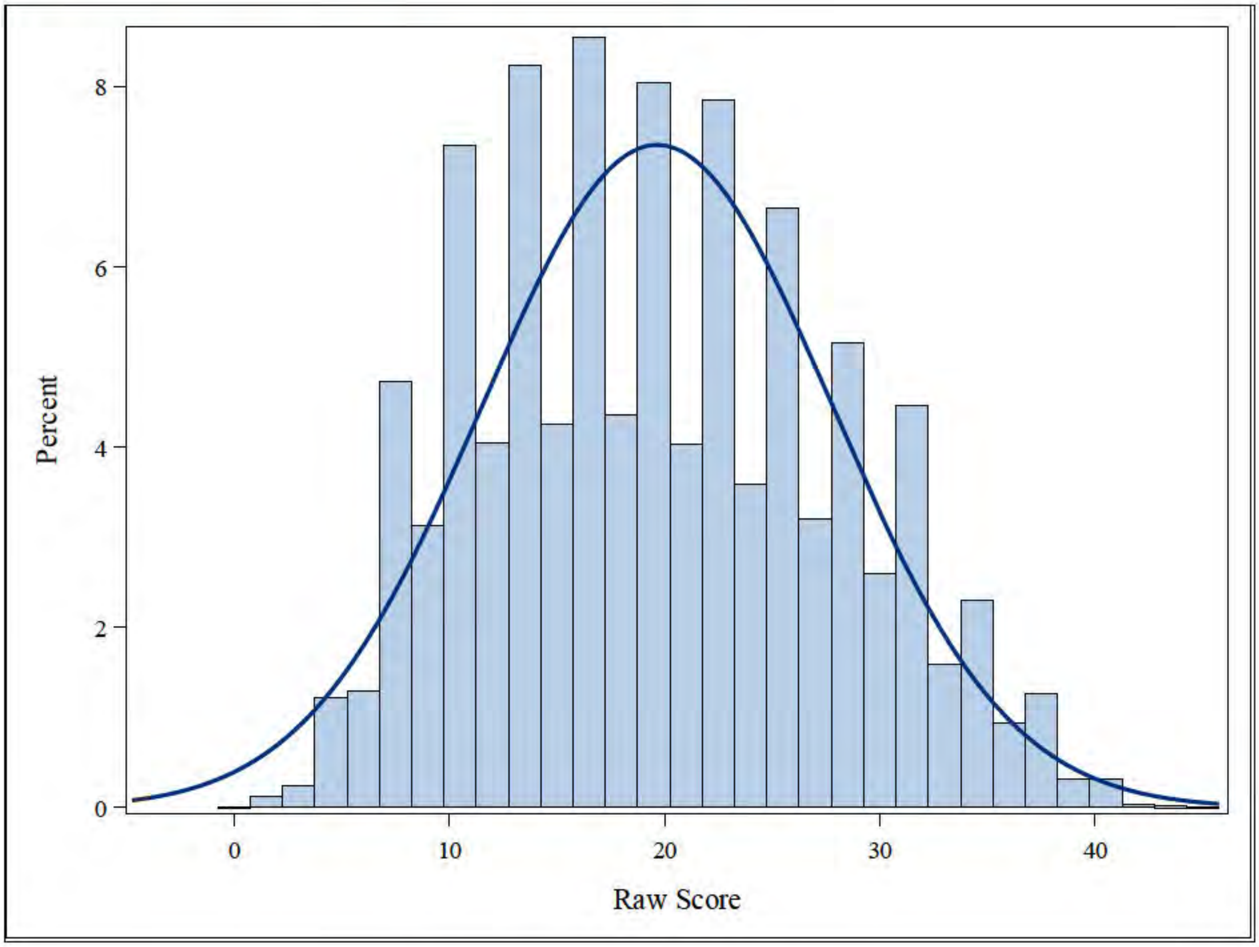


Figure 26. Grade 5 Raw Score Distribution

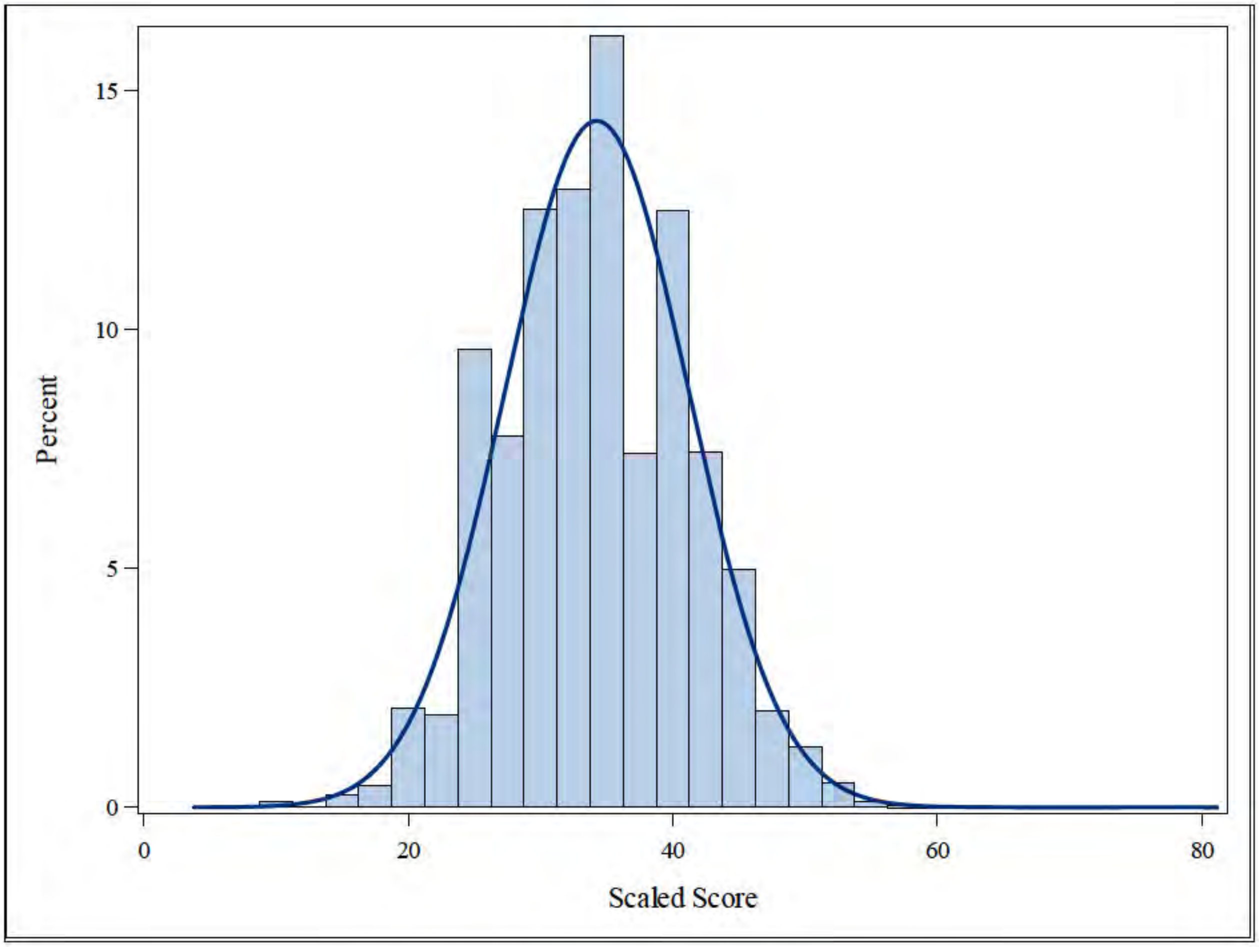


Figure 27. Grade 5 Scaled Score Distribution

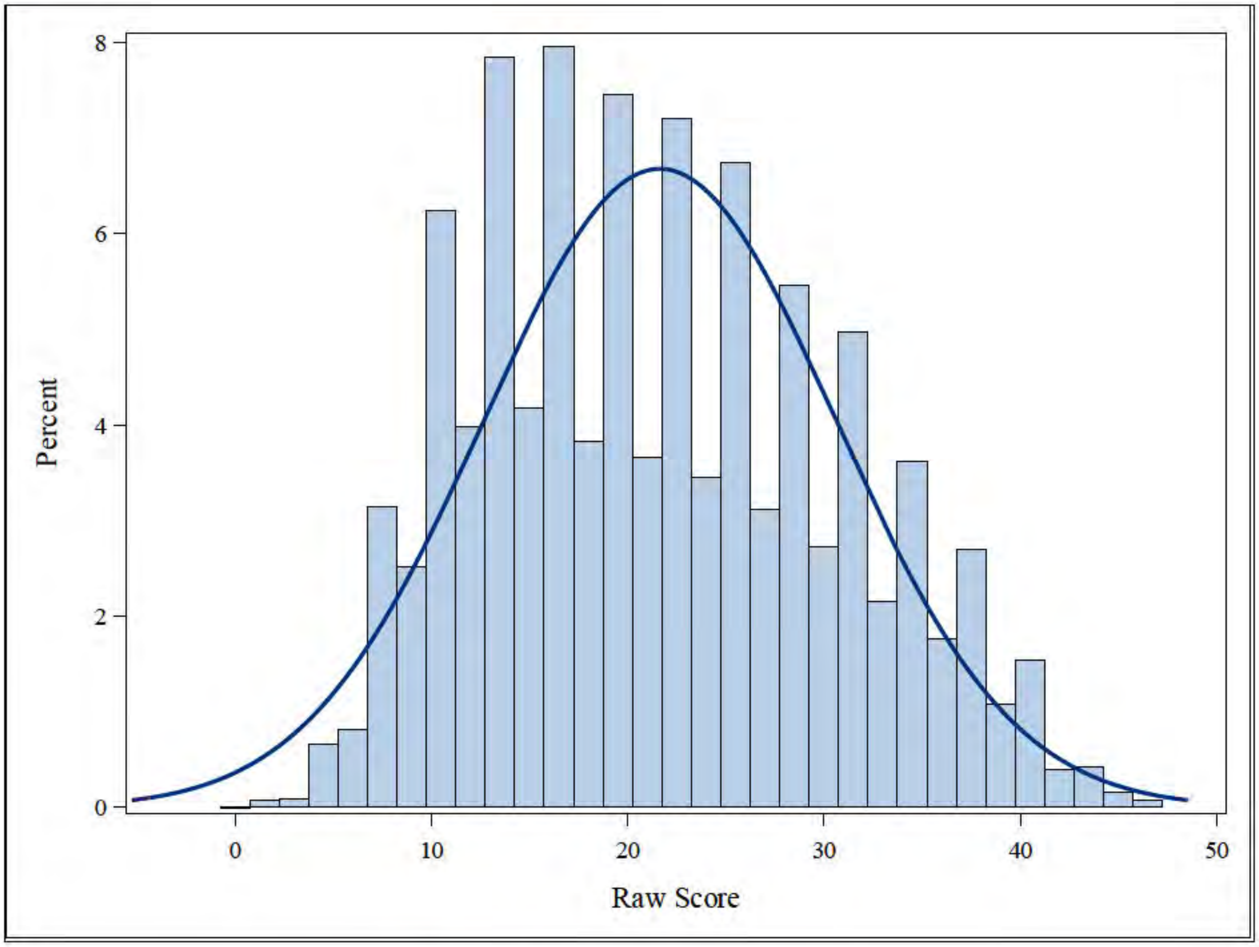


Figure 28. Grade 8 Raw Score Distribution

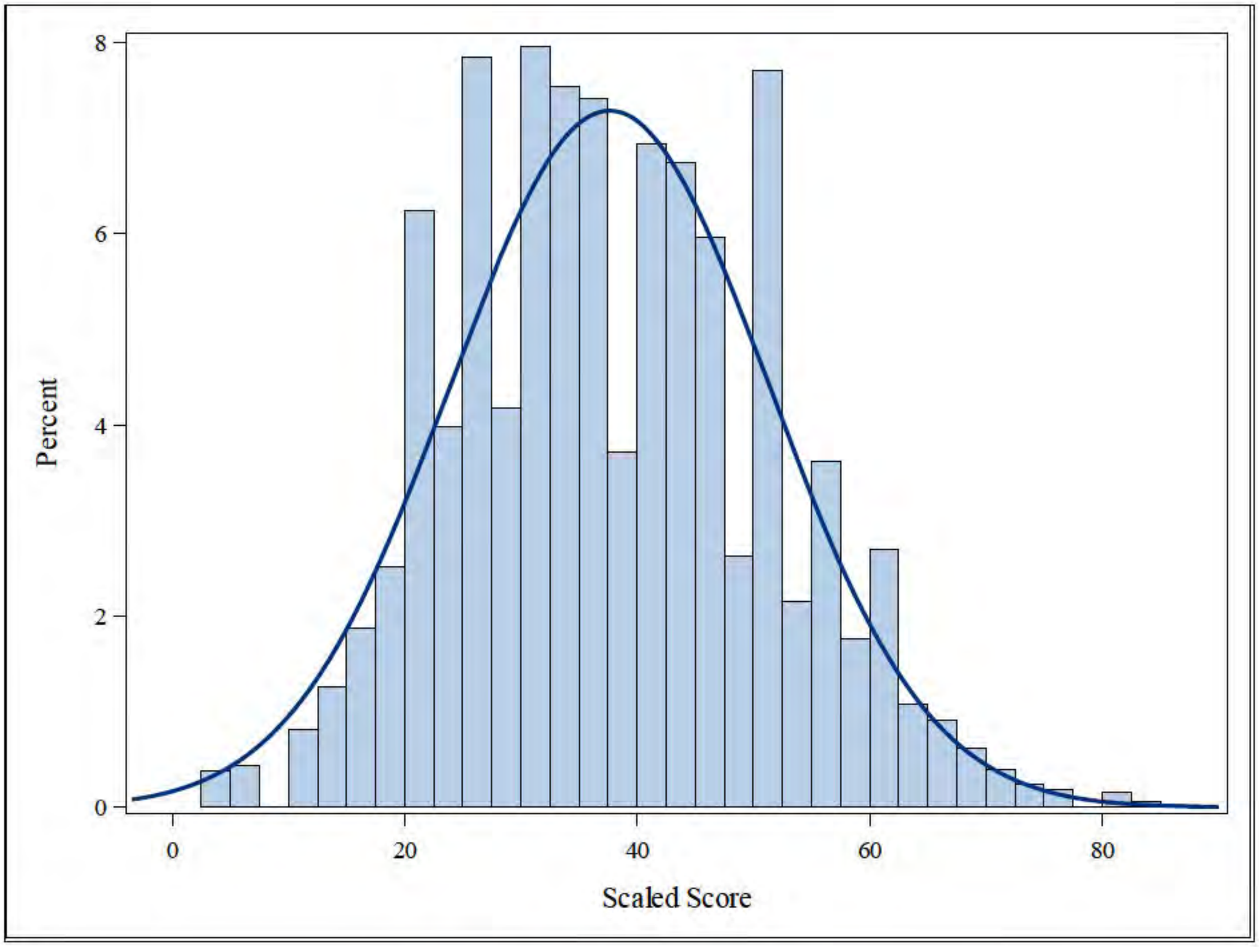


Figure 29. Grade 8 Scaled Score Distribution

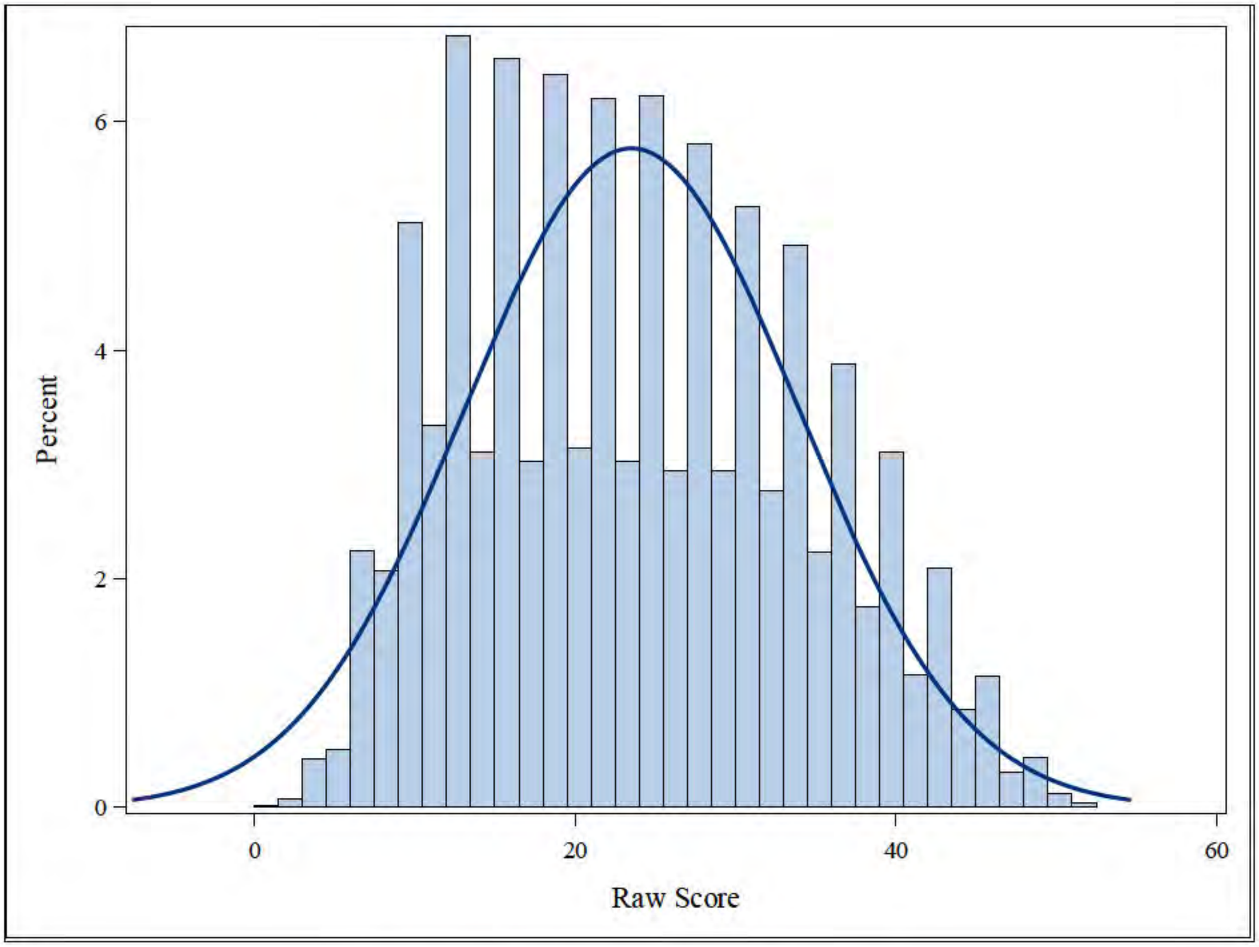


Figure 30. High School Raw Score Distribution

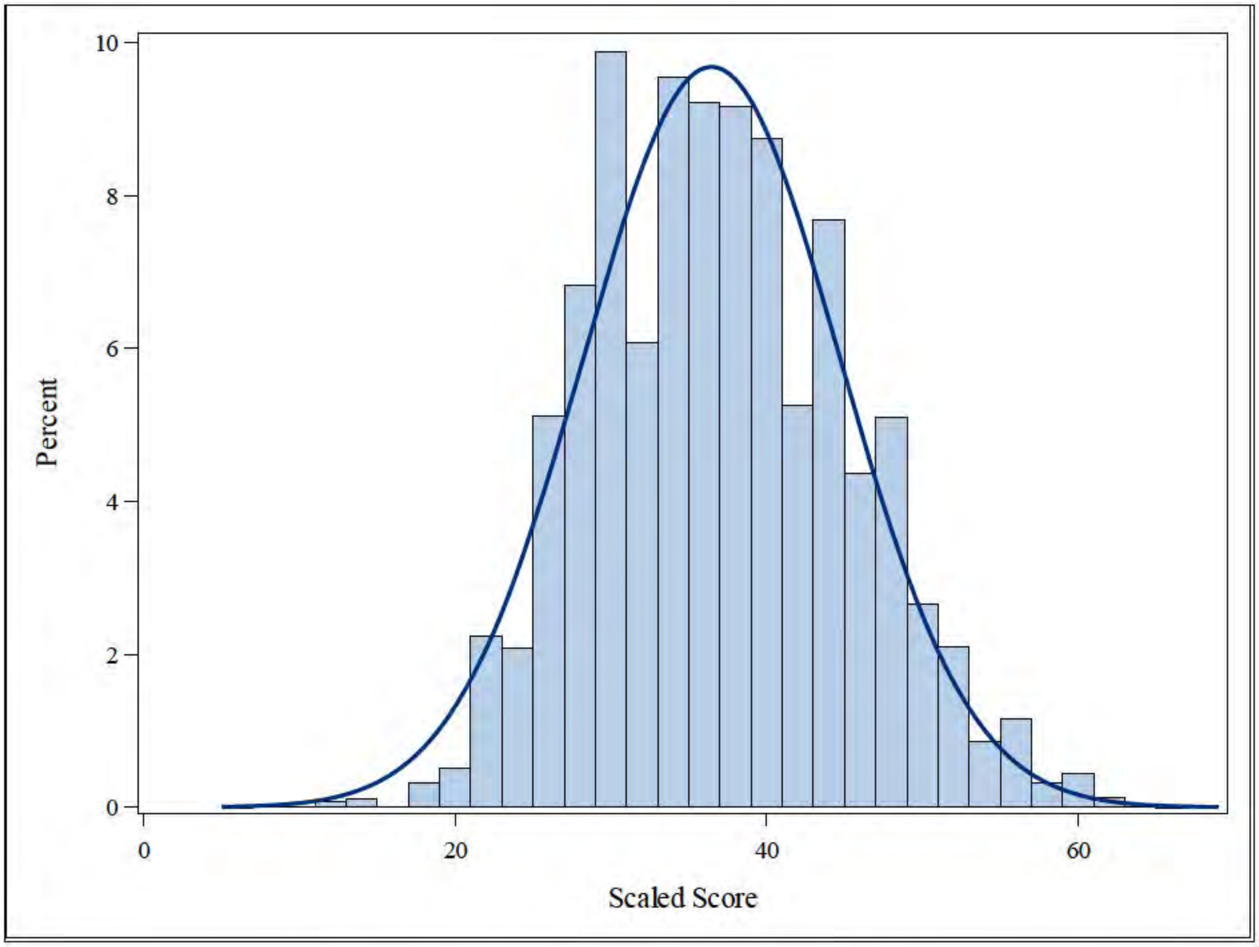


Figure 31. High School Scaled Score Distribution

Appendix P. Reliability[9]

Table 72. Grade 5 Subgroup Reliabilities

Group	N	Maximum	Mean	Std. Dev.	Alpha	SEM
All Students	12,176	45	19.65	8.13	0.84	3.20
Male	6,303	45	19.58	8.33	0.85	3.21
Female	5,873	42	19.73	7.91	0.83	3.20
Hispanic or Latino	416	41	18.22	8.33	0.85	3.18
American Indian or Alaskan Native	103	36	17.98	6.65	0.81	2.83
Asian	168	39	20.69	8.06	0.85	3.10
Black or African American	599	41	13.89	6.78	0.79	3.10
Native Hawaiian or Other Pacific Islander	14	—	—	—	—	—
White	10,411	43	20.03	8.07	0.84	3.21
Two or More races	463	45	19.87	8.27	0.85	3.17
English Speaking Students	11,462	45	20.01	8.06	0.84	3.20
Currently receiving EL services	528	33	11.63	5.47	0.70	2.96
Former EL student - monitoring year 1	152	35	19.36	6.50	0.75	3.16
Former EL student - monitoring year 2	17	—	—	—	—	—
Former EL student - monitoring year 3	17	—	—	—	—	—
IEP: All Other Students	9,479	45	21.15	7.85	0.82	3.22
Students with an IEP	2,697	41	14.39	6.81	0.79	3.05
SES: All Other Students	7,331	45	21.48	8.06	0.83	3.23
Economically Disadvantaged Students	4,845	42	16.89	7.44	0.82	3.14
Migrant: All Other Students	12,169	45	19.66	8.13	0.84	3.21
Migrant Students	7	—	—	—	—	—
Plan 504	672	42	20.59	7.86	0.83	3.19
Plan 504: All Other Students	11,504	45	19.60	8.15	0.84	3.21

[9] Data is not presented for groups with student counts below 25.

Table 73. Grade 8 Subgroup Reliabilities

Group	N	Maximum	Mean	Std. Dev	Alpha	SEM
All Students	12,210	47	21.63	8.97	0.89	3.02
Male	6,365	47	21.78	9.29	0.89	3.05
Female	5,845	46	21.47	8.60	0.88	2.97
Hispanic or Latino	409	45	20.22	9.15	0.89	3.03
American Indian or Alaskan Native	94	39	18.56	8.34	0.86	3.16
Asian	159	42	22.56	9.23	0.88	3.26
Black or African American	607	41	15.74	7.35	0.83	3.05
Native Hawaiian or Other Pacific Islander	12	—	—	—	—	—
White	10,500	47	22.05	8.91	0.89	3.00
Two or More races	425	44	21.48	9.02	0.89	3.01
English Speaking Students	11,560	47	22.03	8.92	0.89	3.00
Currently receiving EL services	489	39	12.92	5.25	0.71	2.82
Former EL student - monitoring year 1	21	—	—	—	—	—
Former EL student - monitoring year 2	29	33	18.34	5.33	0.62	3.48
Former EL student - monitoring year 3	111	37	18.92	7.61	0.83	3.14
IEP: All Other Students	9,836	47	23.06	8.73	0.88	3.02
Students with an IEP	2,374	45	15.69	7.34	0.85	2.87
SES: All Other Students	7,561	47	23.42	8.96	0.89	3.03
Economically Disadvantaged Students	4,649	46	18.72	8.18	0.87	2.97
Migrant: All Other Students	12,199	47	21.63	8.97	0.89	3.02
Migrant Students	11	—	—	—	—	—
Plan 504	911	46	22.10	8.84	0.90	2.86
Plan 504: All Other Students	11,299	47	21.59	8.98	0.89	3.03

Table 74. High School Subgroup Reliabilities

Group	N	Maximum	Mean	Std. Dev.	Alpha	SEM
All Students	11,595	52	23.48	10.40	0.92	2.94
Male	5,983	52	23.22	10.95	0.92	3.01
Female	5,612	50	23.76	9.78	0.92	2.85
Hispanic or Latino	401	48	21.06	9.95	0.92	2.79
American Indian or Alaskan Native	76	41	19.91	8.64	0.91	2.68
Asian	220	50	25.67	10.91	0.93	2.94
Black or African American	540	52	15.69	8.66	0.89	2.91
Native Hawaiian or Other Pacific Islander	15	—	—	—	—	—
White	9,952	52	24.00	10.32	0.92	2.93
Two or More races	390	50	22.96	10.28	0.92	2.96
English Speaking Students	11,102	52	23.93	10.31	0.92	2.92
Currently receiving EL services	429	35	12.52	5.67	0.82	2.47
Former EL student - monitoring year 1	36	41	17.97	8.34	0.89	2.84
Former EL student - monitoring year 2	14	—	—	—	—	—
Former EL student - monitoring year 3	14	—	—	—	—	—
IEP: All Other Students	9,806	52	24.76	10.23	0.92	2.93
Students with an IEP	1,789	50	16.50	8.33	0.90	2.72
SES: All Other Students	8,086	52	25.00	10.50	0.92	2.97
Economically Disadvantaged Students	3,509	50	19.99	9.27	0.91	2.83
Migrant: All Other Students	11,586	52	23.48	10.40	0.92	2.94
Migrant Students	9	—	—	—	—	—
Plan 504	1,012	52	23.89	10.30	0.93	2.79
Plan 504: All Other Students	10,583	52	23.44	10.41	0.92	2.96

Appendix Q. Student Questionnaires

Grade 5

(Document begins on next page.)

STUDENT QUESTIONNAIRE SESSION 4

This session has fifteen multiple-choice questions.

Choose the best answer for each multiple-choice question. During only this session, you may ask for help.

Background

1. Which of the following best describes how you rate yourself as a student overall?

Ⓐ very good

Ⓑ good

Ⓒ fair

Ⓓ poor

2. During the school day, how much time do you spend using a computer?

Ⓐ none

Ⓑ less than one hour

Ⓒ one to two hours

Ⓓ more than two hours

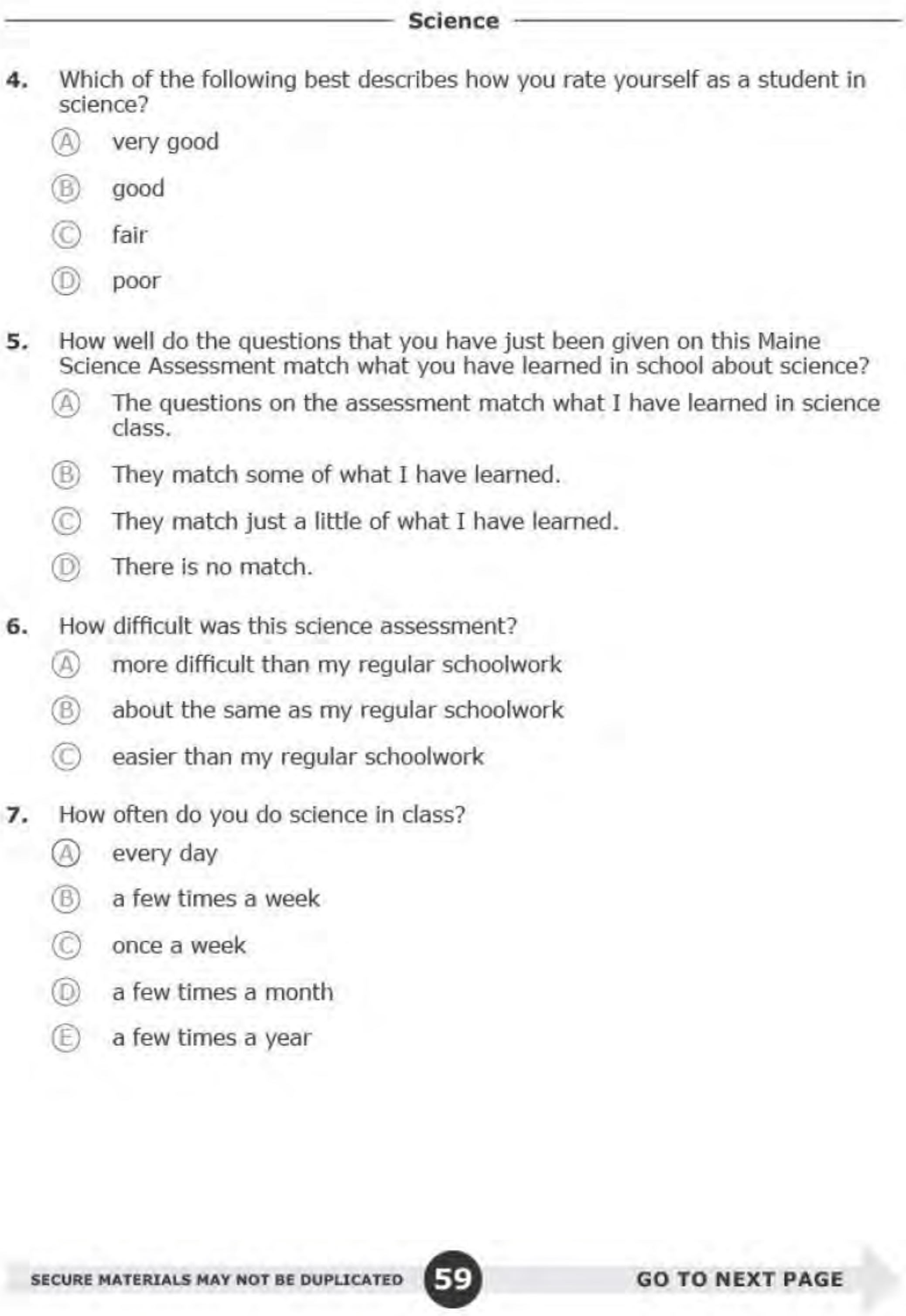
3. How much homework do you do on school nights?

Ⓐ none

Ⓑ less than one hour

Ⓒ one to two hours

Ⓓ more than two hours



Science

4. Which of the following best describes how you rate yourself as a student in science?

Ⓐ very good

Ⓑ good

Ⓒ fair

Ⓓ poor

5. How well do the questions that you have just been given on this Maine Science Assessment match what you have learned in school about science?

Ⓐ The questions on the assessment match what I have learned in science class.

Ⓑ They match some of what I have learned.

Ⓒ They match just a little of what I have learned.

Ⓓ There is no match.

6. How difficult was this science assessment?

Ⓐ more difficult than my regular schoolwork

Ⓑ about the same as my regular schoolwork

Ⓒ easier than my regular schoolwork

7. How often do you do science in class?

Ⓐ every day

Ⓑ a few times a week

Ⓒ once a week

Ⓓ a few times a month

Ⓔ a few times a year

SECURE MATERIALS MAY NOT BE DUPLICATED 59 GO TO NEXT PAGE

8. Which statement best describes how you learn science?

Ⓐ I read a textbook and answer questions, and/or take notes and do assignments.

Ⓑ I work in groups to design and conduct experiments.

Ⓒ I watch videos and answer questions.

Ⓓ I do a combination of A, B and C, mostly A.

Ⓔ I do a combination of A, B and C, mostly B.

Ⓕ I do a combination of A, B and C, mostly C.

9. How often do you make observations and collect data in science class?

Ⓐ a few times a week

Ⓑ a few times a month

Ⓒ once a month

Ⓓ never or almost never

10. How often do you go outside for lessons during the school day?

Ⓐ a few times a week

Ⓑ a few times a month

Ⓒ once a month

Ⓓ never or almost never

11. Do you think you would like to have a job that is related to SCIENCE when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

12. Do you think you would like to have a job that is related to MATH when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

13. Do you think you would like to have a job that is related to TECHNOLOGY when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

14. Do you think you would like to have a job that is related to ENGINEERING when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

15. Do you think you would like to have a job that is related to COMPUTER SCIENCE when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

STOP

Grade 8

(Document begins on next page.)

STUDENT QUESTIONNAIRE SESSION 4

This session has fifteen multiple-choice questions.

Choose the best answer for each multiple-choice question. During only this session, you may ask for help.

Background

1. Which of the following best describes how you rate yourself as a student overall?

Ⓐ very good

Ⓑ good

Ⓒ fair

Ⓓ poor

2. During a typical school day, how much time do you spend using a computer?

Ⓐ none

Ⓑ less than one hour

Ⓒ one to two hours

Ⓓ more than two hours

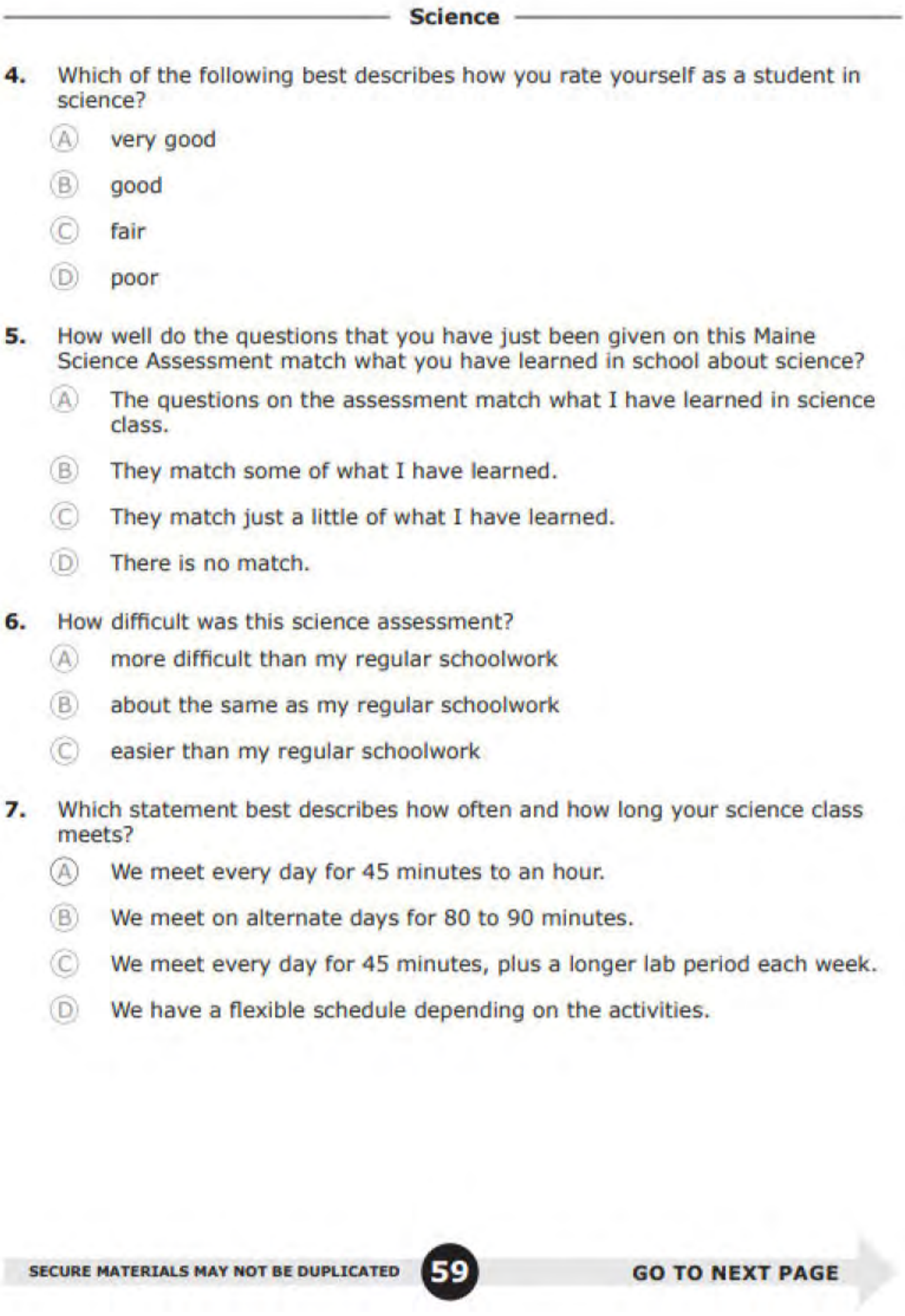
3. How much homework do you do on a typical school night?

Ⓐ none

Ⓑ less than one hour

Ⓒ one to two hours

Ⓓ more than two hours



4. Which of the following best describes how you rate yourself as a student in science?

   Ⓐ very good

   Ⓑ good

   Ⓒ fair

   Ⓓ poor

5. How well do the questions that you have just been given on this Maine Science Assessment match what you have learned in school about science?

   Ⓐ The questions on the assessment match what I have learned in science class.

   Ⓑ They match some of what I have learned.

   Ⓒ They match just a little of what I have learned.

   Ⓓ There is no match.

6. How difficult was this science assessment?

   Ⓐ more difficult than my regular schoolwork

   Ⓑ about the same as my regular schoolwork

   Ⓒ easier than my regular schoolwork

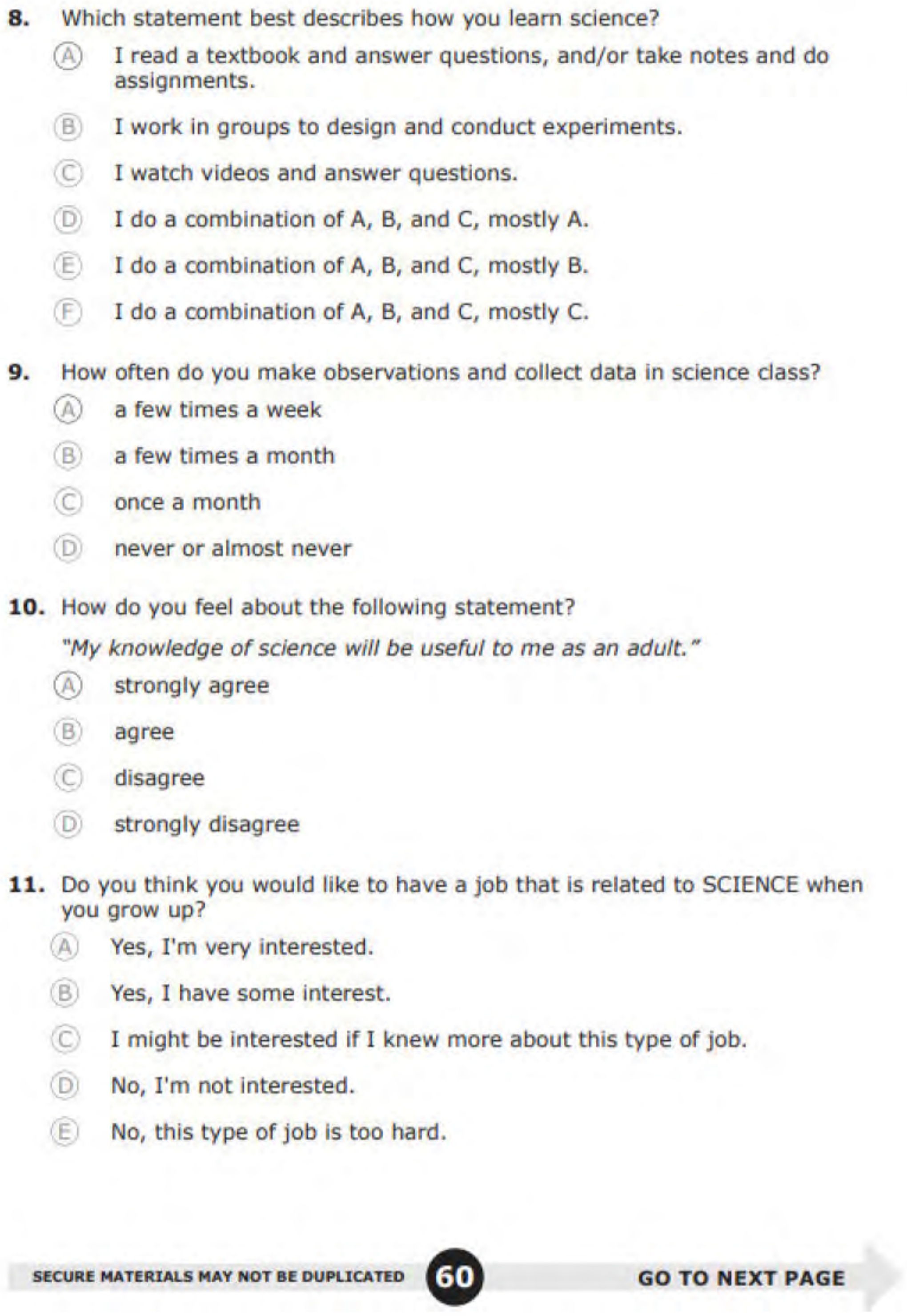
7. Which statement best describes how often and how long your science class meets?

   Ⓐ We meet every day for 45 minutes to an hour.

   Ⓑ We meet on alternate days for 80 to 90 minutes.

   Ⓒ We meet every day for 45 minutes, plus a longer lab period each week.

   Ⓓ We have a flexible schedule depending on the activities.



8. Which statement best describes how you learn science?

Ⓐ I read a textbook and answer questions, and/or take notes and do assignments.

Ⓑ I work in groups to design and conduct experiments.

Ⓒ I watch videos and answer questions.

Ⓓ I do a combination of A, B, and C, mostly A.

Ⓔ I do a combination of A, B, and C, mostly B.

Ⓕ I do a combination of A, B, and C, mostly C.

9. How often do you make observations and collect data in science class?

Ⓐ a few times a week

Ⓑ a few times a month

Ⓒ once a month

Ⓓ never or almost never

10. How do you feel about the following statement?

"My knowledge of science will be useful to me as an adult."

Ⓐ strongly agree

Ⓑ agree

Ⓒ disagree

Ⓓ strongly disagree

11. Do you think you would like to have a job that is related to SCIENCE when you grow up?

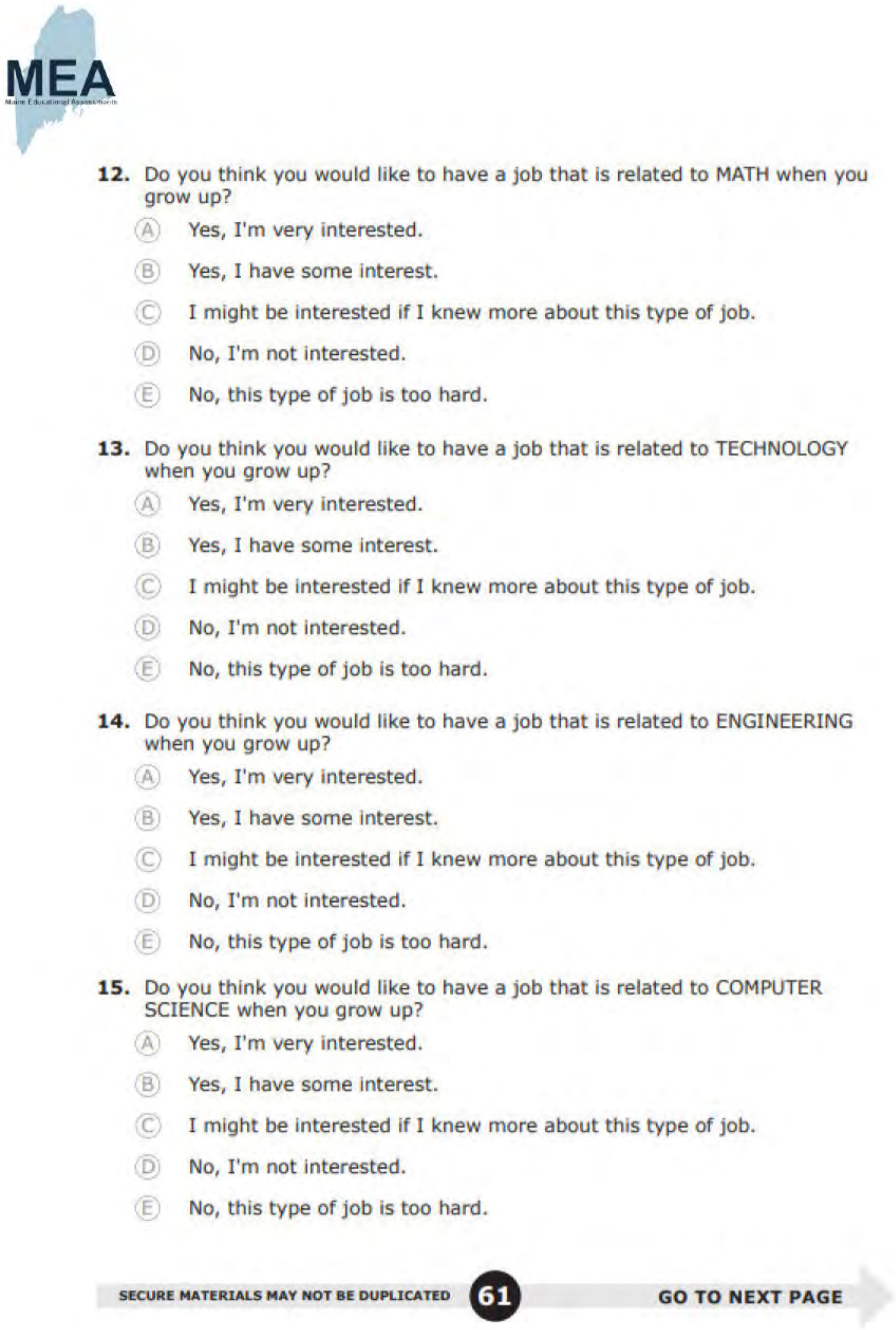
Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.



MEA

Maine Educational Assessments

12. Do you think you would like to have a job that is related to MATH when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

13. Do you think you would like to have a job that is related to TECHNOLOGY when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

14. Do you think you would like to have a job that is related to ENGINEERING when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

15. Do you think you would like to have a job that is related to COMPUTER SCIENCE when you grow up?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

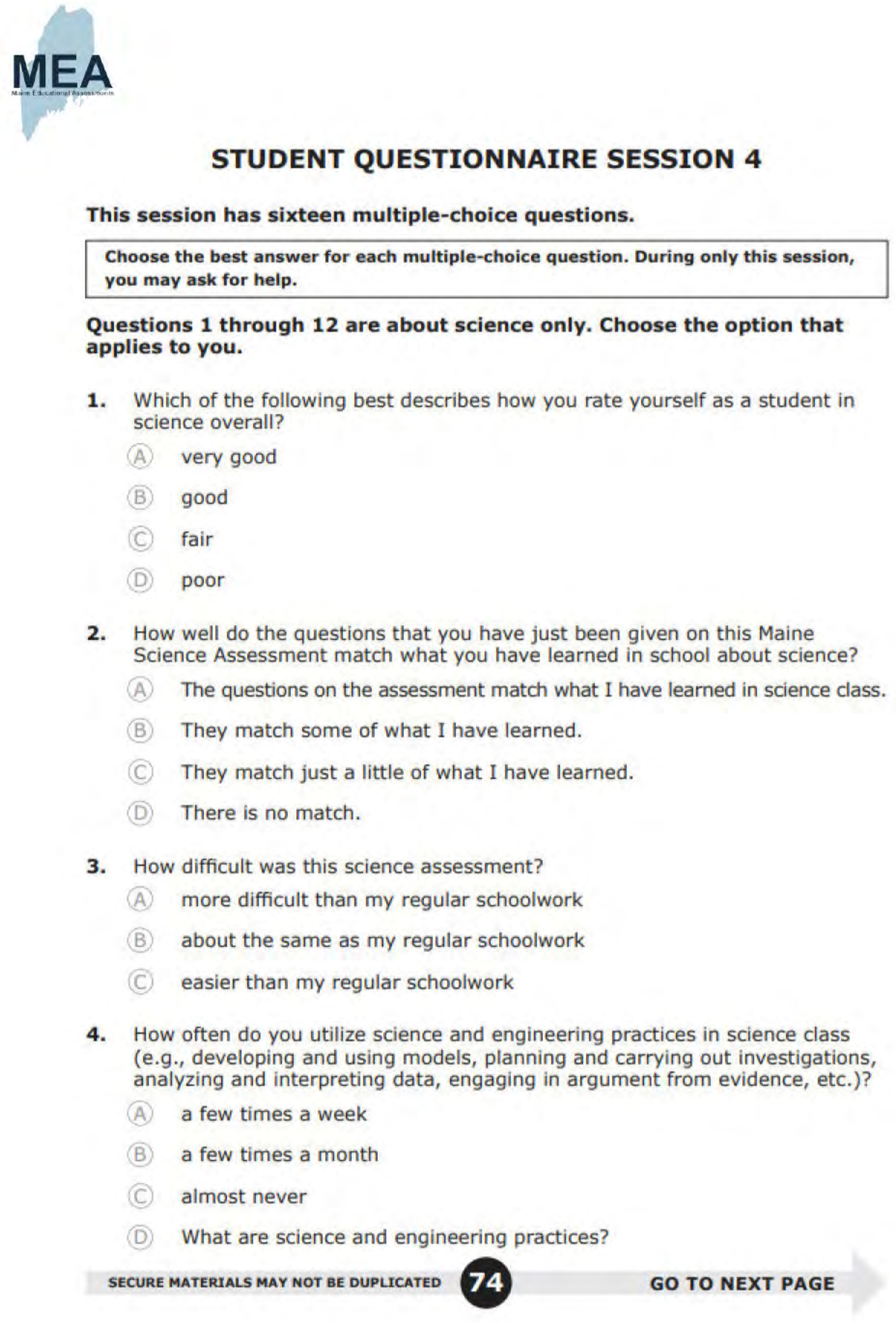
Ⓒ I might be interested if I knew more about this type of job.

Ⓓ No, I'm not interested.

Ⓔ No, this type of job is too hard.

High School

(Document begins on next page.)



STUDENT QUESTIONNAIRE SESSION 4

This session has sixteen multiple-choice questions.

Choose the best answer for each multiple-choice question. During only this session, you may ask for help.

Questions 1 through 12 are about science only. Choose the option that applies to you.

1. Which of the following best describes how you rate yourself as a student in science overall?

Ⓐ very good

Ⓑ good

Ⓒ fair

Ⓓ poor

2. How well do the questions that you have just been given on this Maine Science Assessment match what you have learned in school about science?

Ⓐ The questions on the assessment match what I have learned in science class.

Ⓑ They match some of what I have learned.

Ⓒ They match just a little of what I have learned.

Ⓓ There is no match.

3. How difficult was this science assessment?

Ⓐ more difficult than my regular schoolwork

Ⓑ about the same as my regular schoolwork

Ⓒ easier than my regular schoolwork

4. How often do you utilize science and engineering practices in science class (e.g., developing and using models, planning and carrying out investigations, analyzing and interpreting data, engaging in argument from evidence, etc.)?

Ⓐ a few times a week

Ⓑ a few times a month

Ⓒ almost never

Ⓓ What are science and engineering practices?

5. Select how often you have the opportunity to engage in each of the science and engineering practices in your science courses.

	Regularly	Occasionally	Seldom or Never
a Asking questions and defining problems	Ⓐ	Ⓑ	Ⓒ
b Developing and using models	Ⓐ	Ⓑ	Ⓒ
c Planning and carrying out investigations	Ⓐ	Ⓑ	Ⓒ
d Analyzing and interpreting data	Ⓐ	Ⓑ	Ⓒ
e Using mathematics and computational thinking	Ⓐ	Ⓑ	Ⓒ
f Constructing explanations and designing solutions	Ⓐ	Ⓑ	Ⓒ
g Engaging in argument from evidence	Ⓐ	Ⓑ	Ⓒ
h Obtaining, evaluating and communicating information	Ⓐ	Ⓑ	Ⓒ

6. How do you feel about the following statement?

"My knowledge of science will be useful to me as an adult."

Ⓐ strongly agree

Ⓑ agree

Ⓒ disagree

Ⓓ strongly disagree

7. Do you think you would like to have a career that is related to SCIENCE?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of career.

Ⓓ No, I'm not interested.

Ⓔ No, this type of career is too hard.

8. Do you think you would like to have a career that is related to MATH?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of career.

Ⓓ No, I'm not interested.

Ⓔ No, this type of career is too hard.

9. Do you think you would like to have a career that is related to TECHNOLOGY?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of career.

Ⓓ No, I'm not interested.

Ⓔ No, this type of career is too hard.

10. Do you think you would like to have a career that is related to ENGINEERING?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of career.

Ⓓ No, I'm not interested.

Ⓔ No, this type of career is too hard.

11. Do you think you would like to have a career that is related to COMPUTER SCIENCE?

Ⓐ Yes, I'm very interested.

Ⓑ Yes, I have some interest.

Ⓒ I might be interested if I knew more about this type of career.

Ⓓ No, I'm not interested.

Ⓔ No, this type of career is too hard.

For questions 12 and 13, choose all of the options that apply to you.

12. Select the Science courses you have taken before 9th grade, or have taken or currently are taking during high school (grades 9–12), and identify if they were/are Honors or AP courses.

	Before 9th Grade	9th-12th Grade	AP/Honors
a Environmental, Earth, or Space Science	Ⓐ	Ⓑ	Ⓒ
b Biology	Ⓐ	Ⓑ	Ⓒ
c Chemistry	Ⓐ	Ⓑ	Ⓒ
d Physics	Ⓐ	Ⓑ	Ⓒ
e Other science course	Ⓐ	Ⓑ	Ⓒ

13. Select the average grade for all courses you have already taken in each subject.

	A	B	C	D	E/F
a Mathematics	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ
b English and Language Arts	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ
c Science	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ

For questions 14–16, choose only one answer.

14. Do you plan to enroll in an educational program the year after high school graduation?

Ⓐ Yes, a four-year college.

Ⓑ Yes, a two-year community/vocational or technical school.

Ⓒ Yes, through military enlistment.

Ⓓ Yes, but I am not sure yet.

Ⓔ No, I am not planning to attend an educational program the year after my high school graduation.

15. What is the highest level of education you plan to complete beyond high school?

Ⓐ specialized training or certificate program

Ⓑ two-year associate of arts or science degree (AA, AAS, or AS)

Ⓒ bachelor's degree (BA or BS)

Ⓓ graduate degree (MA, MBA, MS, PhD, JD, MD, or DVM)

Ⓔ other

Ⓕ undecided

16. How important to you is your score on this science assessment you just completed?

Ⓐ extremely important

Ⓑ important

Ⓒ somewhat important

Ⓓ not very important

STOP

Appendix R. Questionnaire Data

Difficulty of the Assessment

All students were asked how difficult the content of the assessment was compared to their classroom instruction.

Question: How difficult was this science test?

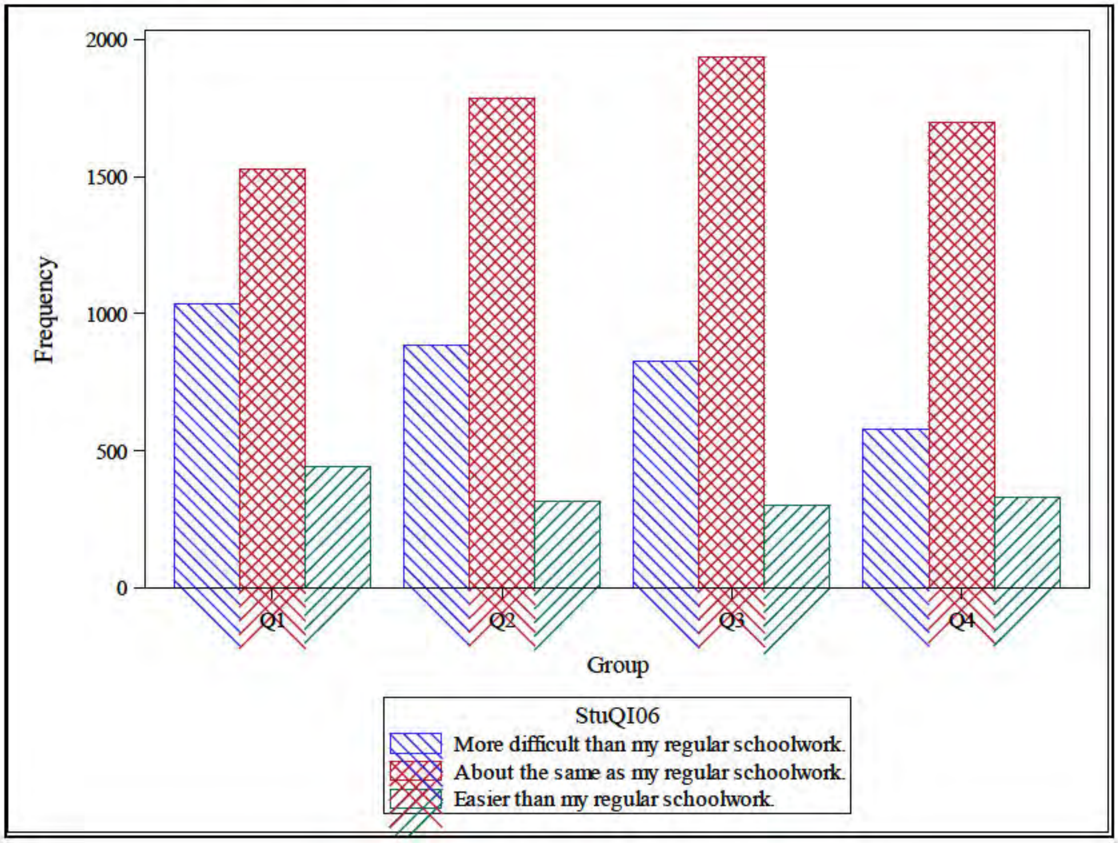


Figure 32. Grade 5 Difficulty of the Assessment

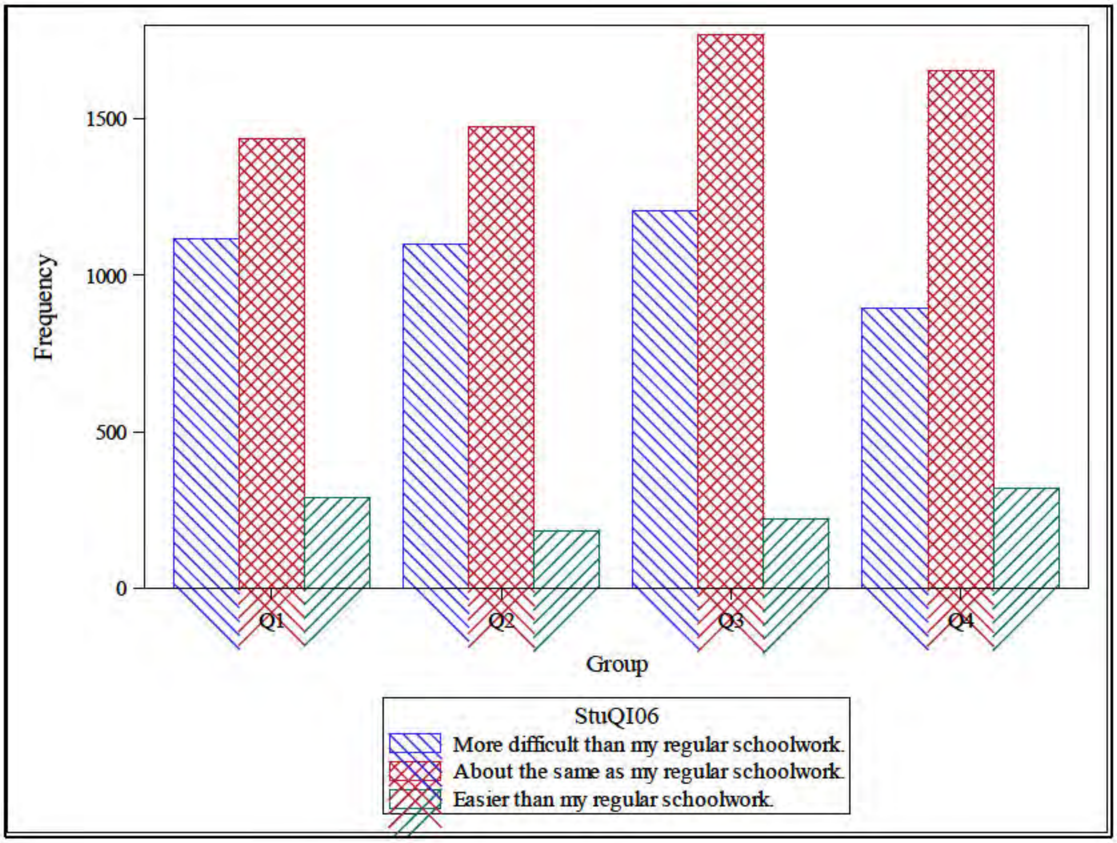


Figure 33. Grade 8 Difficulty of the Assessment

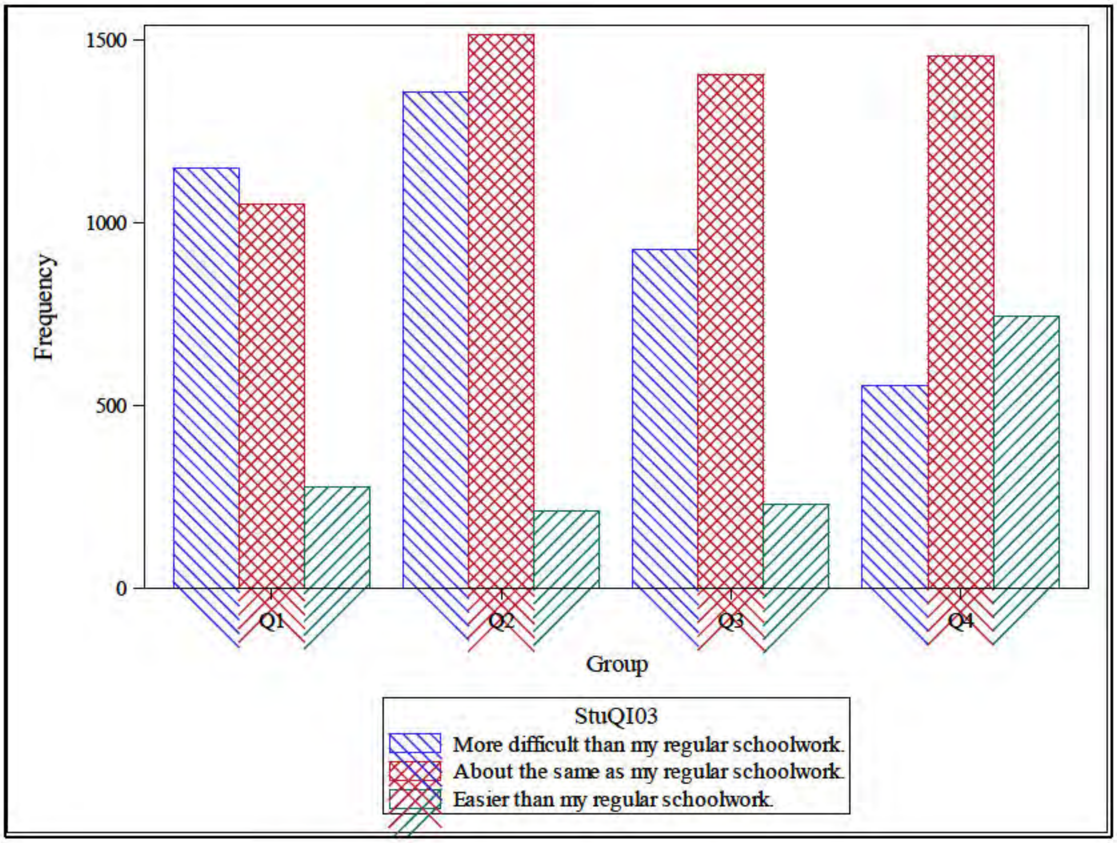


Figure 34. High School Difficulty of the Assessment

For all three grades, most students indicated that the Maine Science Assessment was about the same difficulty as their regular schoolwork. In general, the lower performing students indicated that the Maine Science Assessment was more difficult than their regular schoolwork while higher performing students were more likely to indicate the assessment was easier than their regular schoolwork.

Frequency of Science Class

Students in grades 5 and 8 were asked how frequently they received science instruction. Note that the wording of this question varied by grade.

Grade 5 Question: How often do you do science in class?

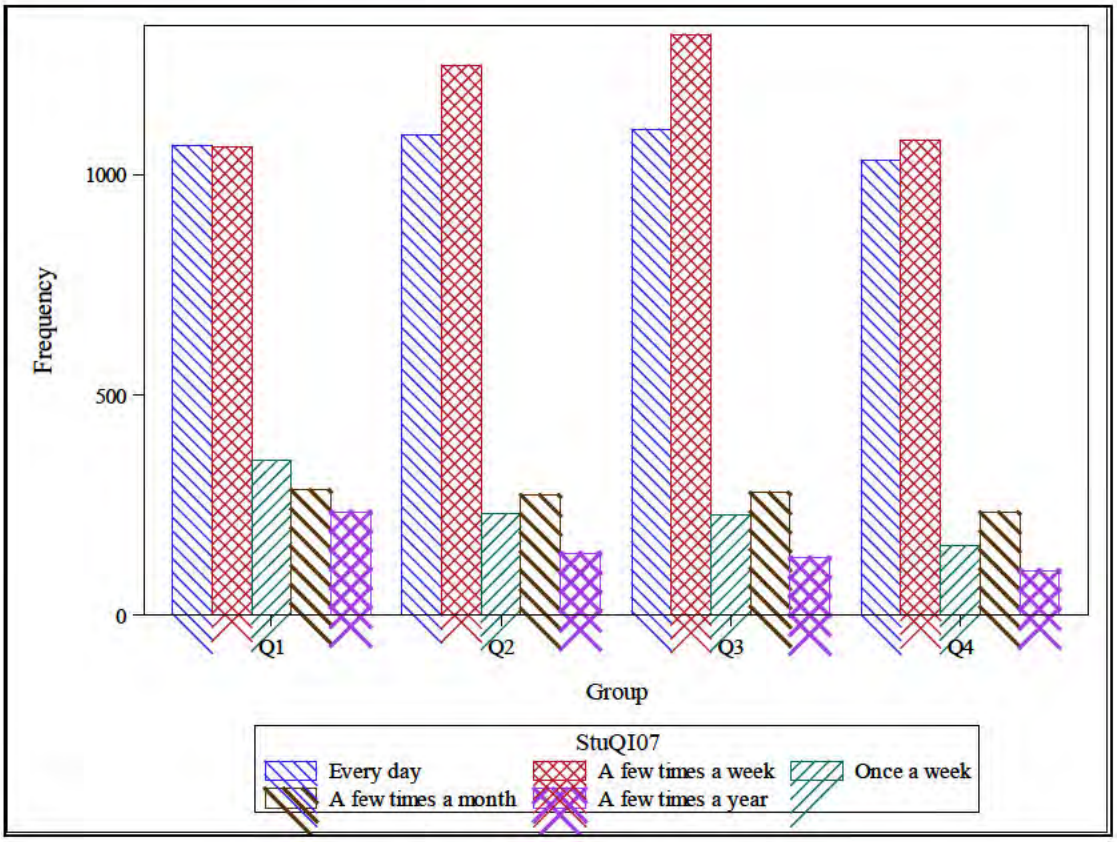


Figure 35. Grade 5 Frequency of Science Class

Grade 8 Question: Which statement best describes how often and how long your science class meets?

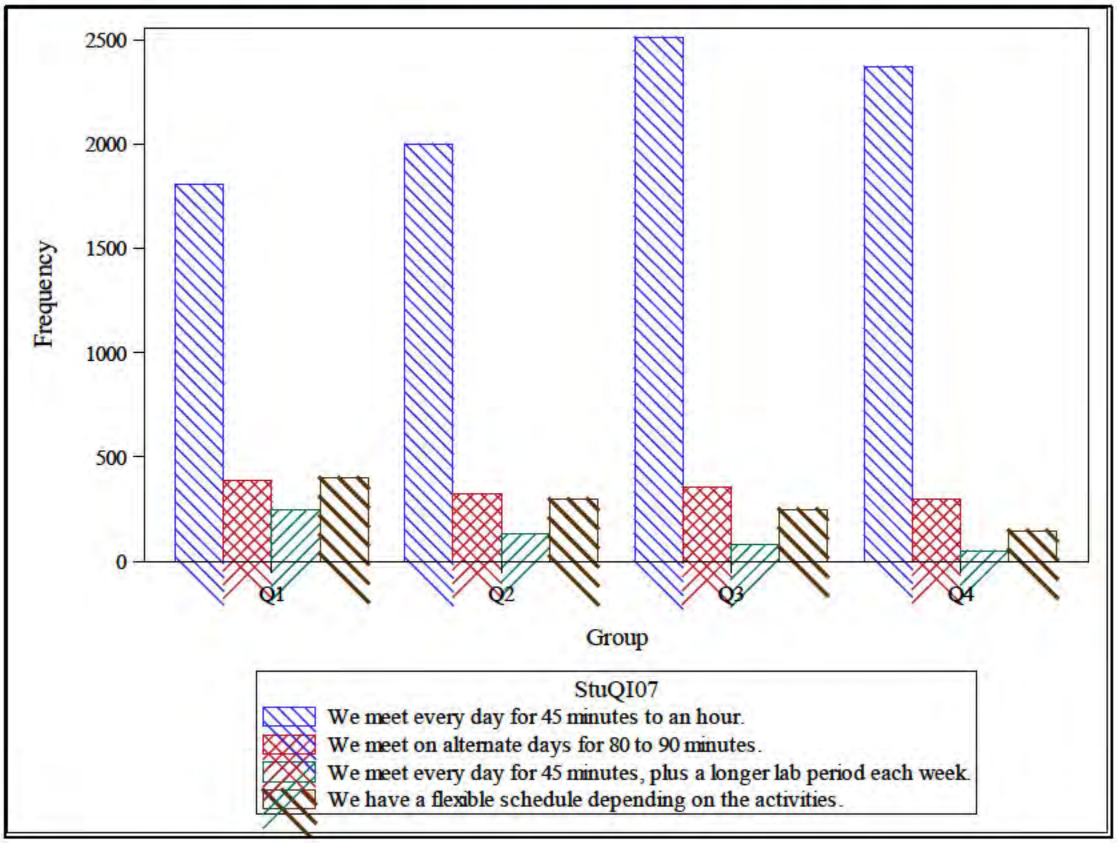


Figure 36. Grade 8 Frequency of Science Class

Utilization of SEPs

High school students were asked how frequently engineering practices were incorporated into their instruction. Figure 27 provides the result by performance quartile.

Question: How often do you utilize science and engineering practices in science class?

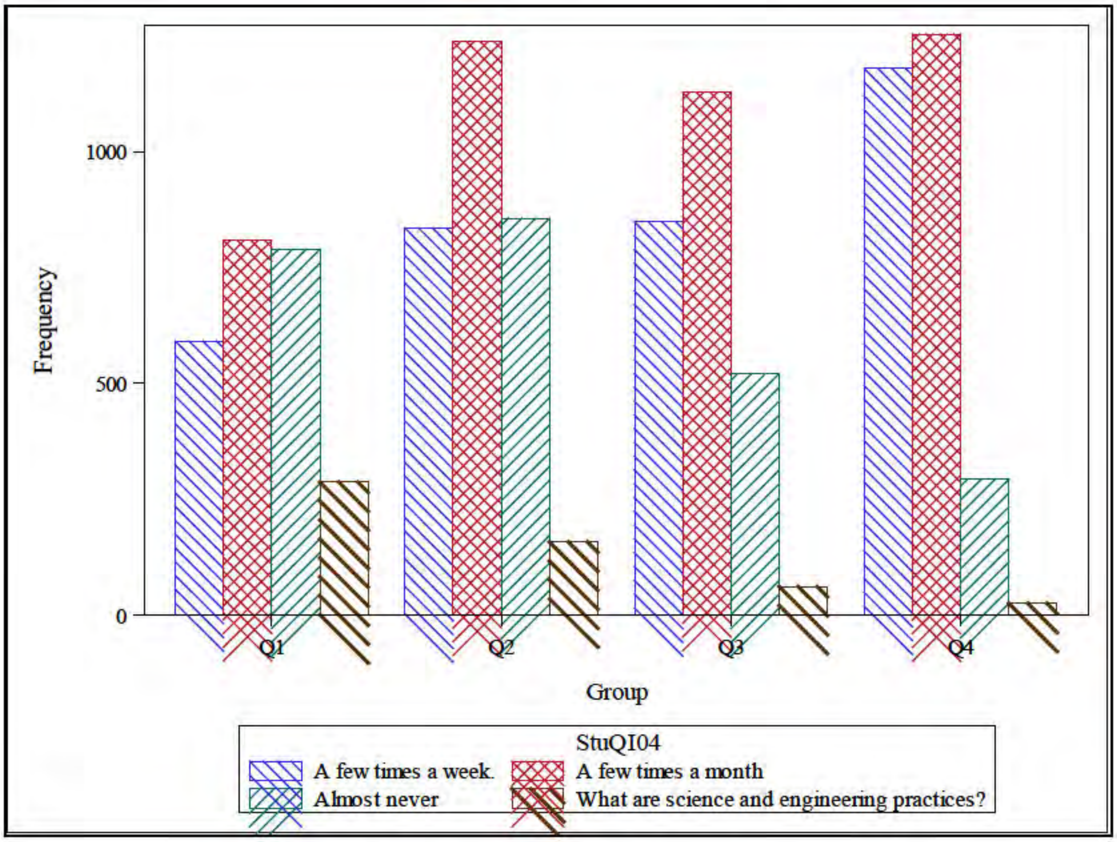


Figure 37. High School How Science and Engineering Are Utilized

How Science Is Learned

Students in grades 5 and 8 were asked how they learn science. Figure 28 and Figure 29 provide the results by performance quartile.

Question: Which statement best describes how you learn science?

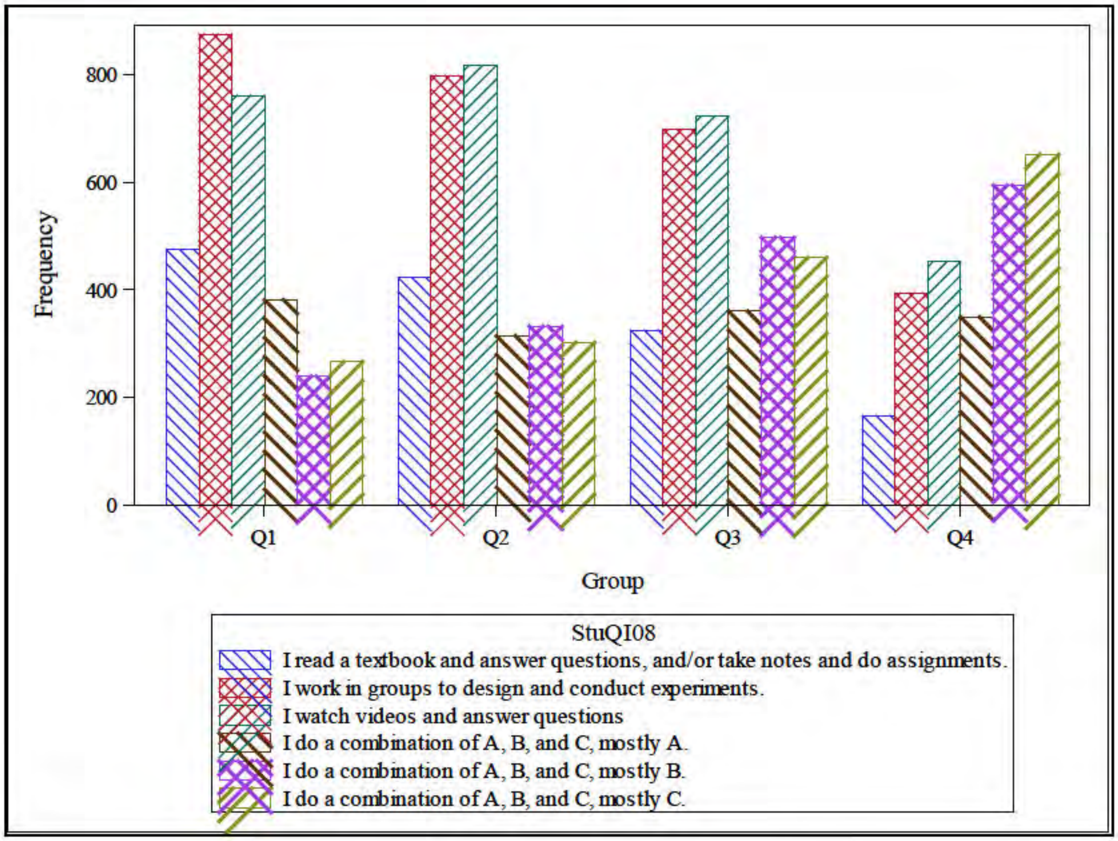


Figure 38. Grade 5 How Science Is Learned

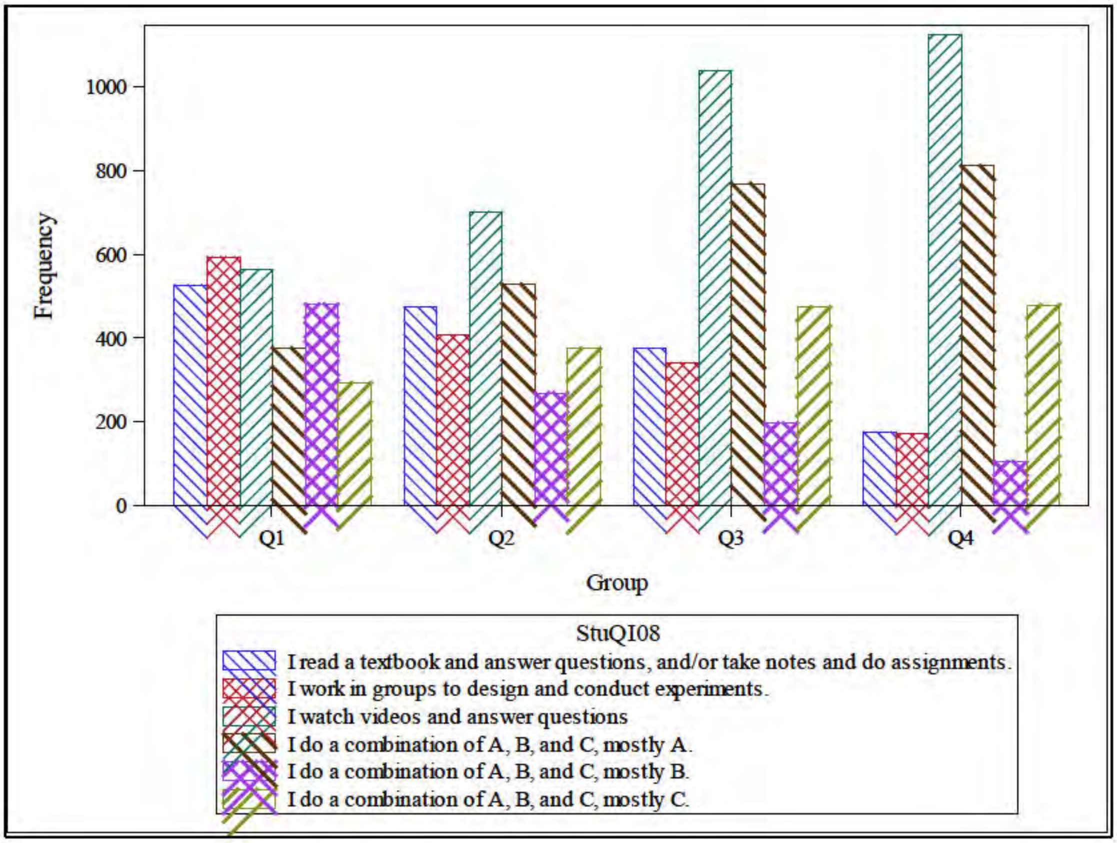


Figure 39. Grade 8 How Science Is Learned

Score Importance

High school students were asked how important the Maine Science Assessment score was to them. Figure 30 provides the result by performance quartile.

How important to you is your score on this science test you just completed?

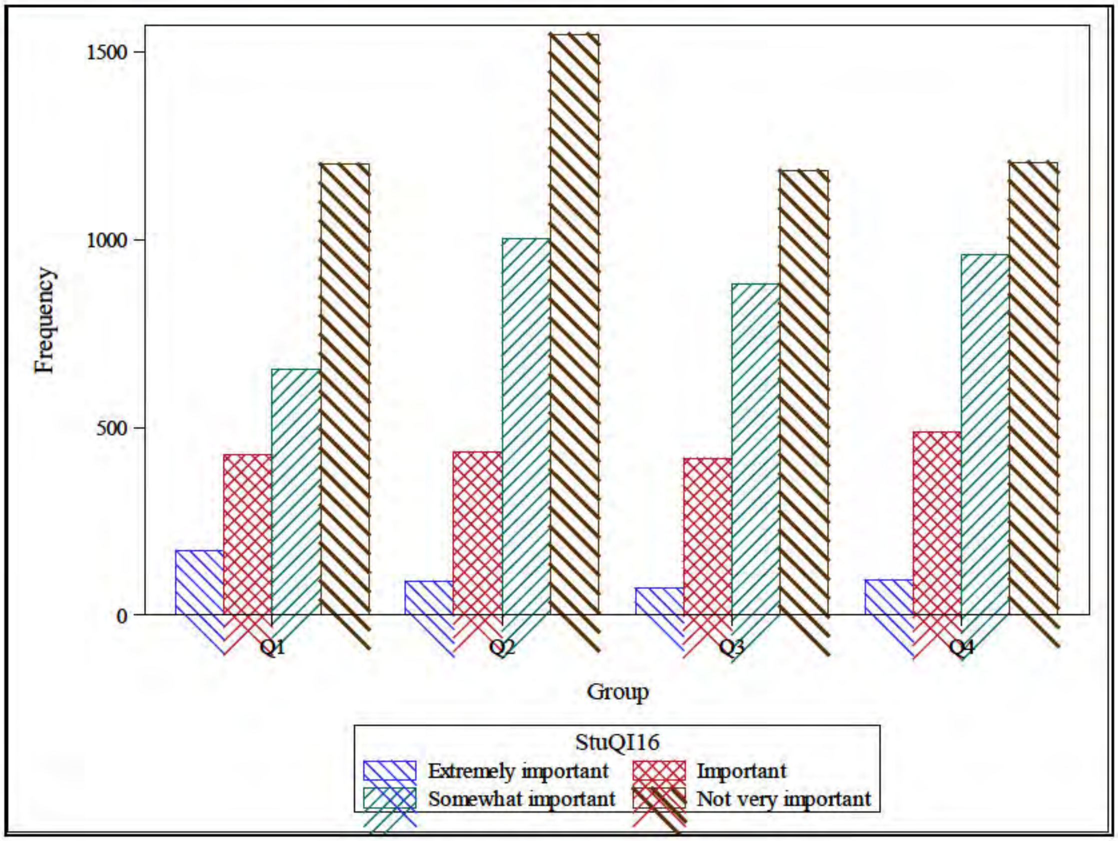


Figure 40. High School Score Importance

Appendix S. Contributing State A Item Development

Item Development Processes from Science Exchange Contributing State A

Contributing state A contributed items to New Meridian's Science Exchange in all science disciplines for Grades 3–8 and life science items for high school. The following process summary is from contributing state A's technical report. The stimuli for the contributing state A assessment are anchored on a scientific phenomenon described by text, images, tables, graphs, models, and graphic organizers created by the contributing state's vendor. Phenomena and bundles were chosen to represent the breadth of assessable science content. As part of the item development plan, all performance expectations were aligned to at least one standalone item or to an item in an item set or task. After studying the science standards of contributing state A, the content lead generated lists of bundled and associated phenomena for item sets and tasks. When identifying a phenomenon, the content lead considered the following:

- The emphasis of each performance expectation as described in the clarification statements for each performance expectation
- Whether a proposed phenomenon was rich enough to support the required number of items, including overage
- Whether the phenomenon fit with the "PE bundles" developed earlier to provide meaningful, three-dimensional assessment of performance expectations

Phenomena were chosen to represent the breadth of content described by contributing state A's state standards. The process of determining phenomena and associated bundles was iterative and included the identification of phenomena that could be assessed with a particular bundle as well as the understanding of the need to assess as many PEs as possible in the field test.

Sets were purposefully designated as item sets or tasks, and the designation of the set (whether item or task) influenced the selection of phenomena. The tasks were based on stimuli that allowed students to delve deeply into a topic and were made up of items that built upon each other and often led to a culminating extended-response (ER) item. The items in a task could require a specific order, and information in one item could be used to build upon in subsequent items.

Outline and Stimuli Development for Contributing State A

Contributing state A's vendor used both experienced internal and external science assessment editors to develop the phenomena-based stimuli for item sets and tasks. Before the editors began the process, the vendor's content lead trained them on the process of conducting an effective internet search for science articles on contributing state A DOE's objectives, including training in universal design and bias and sensitivity issues. To support the outline development process, writers were given the state standards for contributing state A. They were also provided specific item set or task templates that described the PE bundle to be written to in addition to the point value, item types, and dimensional alignment of each of the items in the set and whether the dimensions of the bundled PEs could be mixed or matched. The outline contained space for writers to enter the primary sources they used in researching their phenomenon and in writing their stimulus, space for the writers to include a draft of the stimulus and its supporting data, and space to describe each item and its metadata. Writers submitted their item outlines to the editors, who finalized the item set and task outlines before submitting them to the content lead and manager for senior review. After this review, the outlines were submitted to the DOE for Contributing State A.

Item Writing and Review Process for Contributing State A

The vendor for Contributing State A employed a cadre of item writers for the grades 3–8 assessment. All writers were approved by the DOE of Contributing State A before they engaged in any item development activities. As the first step in the item writing process, the vendor content lead provided a webinar training to all writers.

In the training, writers were provided context for the assessment, including DOE expectations, the science standards of Contributing State A, and a review of best practices for item development. The item writers were provided the approved item topics and drafts of the stimuli and item outlines that provided explanations of the phenomena underlying the tasks and item sets. Item writers were also provided with alignment to the Science and Engineering Practices, Crosscutting Concepts, and Disciplinary Core Ideas of the science standards for Contributing State A and guidance on how each item set or task should be developed.

The use of item sets and task overviews allowed the vendor to provide direction for the items developed during the development cycle. For standalone development, item writers were provided with assignments that indicated the number of items to write to each performance expectation as well as the specific dimensions to align to for each item. The item writing assignments for each set or task also specified the set type, the item types and number of items to be written, and potential item stems to be used for each item. Significant attention was devoted to understanding how to write TE items and scoring guides for CR and ER items.

Although all the writers were science writers with experience in writing three-dimensional items, the vendor gave instructions in basic assessment item writing principles. Writers were instructed to make certain that the vocabulary and context of the items were grade level appropriate, to ensure that the distractors were incorrect but plausible, and to avoid cueing and outliers in the items.

The vendor hosted an online training for writers that included information regarding universal design and bias/sensitivity. A variety of items were presented and reviewed using universal design and bias/sensitivity lenses. The vendor provided training and feedback to the writers throughout the development cycle as the DOE of Contributing State A and their vendor gained a clearer understanding of how the stimuli, items, and sets worked together.

The vendor provided additional training to a subset of editors outlining the specific responsibilities for those who served as editors for the grades 3–8 assessment. Items went through two rounds of content editing that examined characteristics of items, including alignment to the dimensions of the performance expectations of the state standards for Contributing State A, content accuracy, cognitive complexity, and quality of distractors. Items then went through one round of proofreading, which focused on grammar, usage, and consistent style of graphics, and a final round of review before being submitted to the DOE of Contributing State A for their first round of review.

Content and Bias Review for Items from Contributing State A

After the completion of item development, the vendor coordinated face-to-face content and bias review meetings. The meetings were led by facilitators from the DOE of Contributing State A and from their vendor. Participants included current classroom teachers, retired teachers, content specialists, and school administrators. For the content and bias review meeting, participants completed nondisclosure agreements as part of the activities. The recruitment process, conducted by DOE staff of Contributing State A, also included participants from regions across the state. Participants represent the population of students served in Contributing State A—including special education, multilingual learners, and students with disabilities—as well as the diverse geographic and demographic composition of the state. Because the content and bias review meeting took place over five days, committee members could not participate every day of the meeting. As a result, the vendor and the DOE separated the meeting into two parts and had participants attend the meeting in the first half of the week or the second half of the week. Consequently, most of the individual participants did not review or discuss every item, although every item was reviewed by a committee. Table 70 and Table 71 provide the demographic characteristics of the review committees.

Table 75. Educator Representation in the 2018–2019 Content & Bias Reviews for HS Biology

	Characteristic	**Number of Participants**
Role	Classroom Teacher	9
	Content/Curriculum Specialist	0
	School Administrator	0
	Other Staff	2
	ML Teacher	0
	Special Education Teacher	0
	Special Ed Teacher—Gifted	0
	Visually Impaired or Hearing-Impaired Teacher	1
Race	Black or African American	2
	Asian	0
	Hispanic/Latino	1
	White	7
Gender	Male	3
	Female	7
Total Participants		**10**

Note: As teachers may fulfill multiple roles, representation of roles exceeds the number of total participants.

Table 76. Educator Representation in the 2017–2018 Content & Bias Reviews for Grade 5 and Middle School

	Characteristic	**Number of Participants by Grade Level**				
		03	**04**	**05**	**06–08**	
					Group A	**Group B**
Role	Classroom Teacher	7	8	4	10	10
	Content/Curriculum Specialist	0	1	1	1	1
	Instructional Lead	0	0	0	1	0
	School Administrator	0	1	0	0	0
	Other Staff	2	0	0	1	0
	ML Teacher	1	1	1	0	0
	Language Immersion Teacher	0	0	0	0	1
	Special Education Teacher	1	0	0	1	0
	Special Education Teacher–Gifted	1	0	0	0	0
	Visually Impaired or Hearing-Impaired Teacher	0	1	0	1	1
Race	Black or African American	1	2	3	2	3
	Asian	0	1	1	0	0
	Hispanic/Latino	0	1	0	1	0
	White	7	6	2	7	8
Gender	Male	2	2	1	2	1
	Female	7	8	6	9	10
Total Participants		**9**	**10**	**7**	**11**	**11**

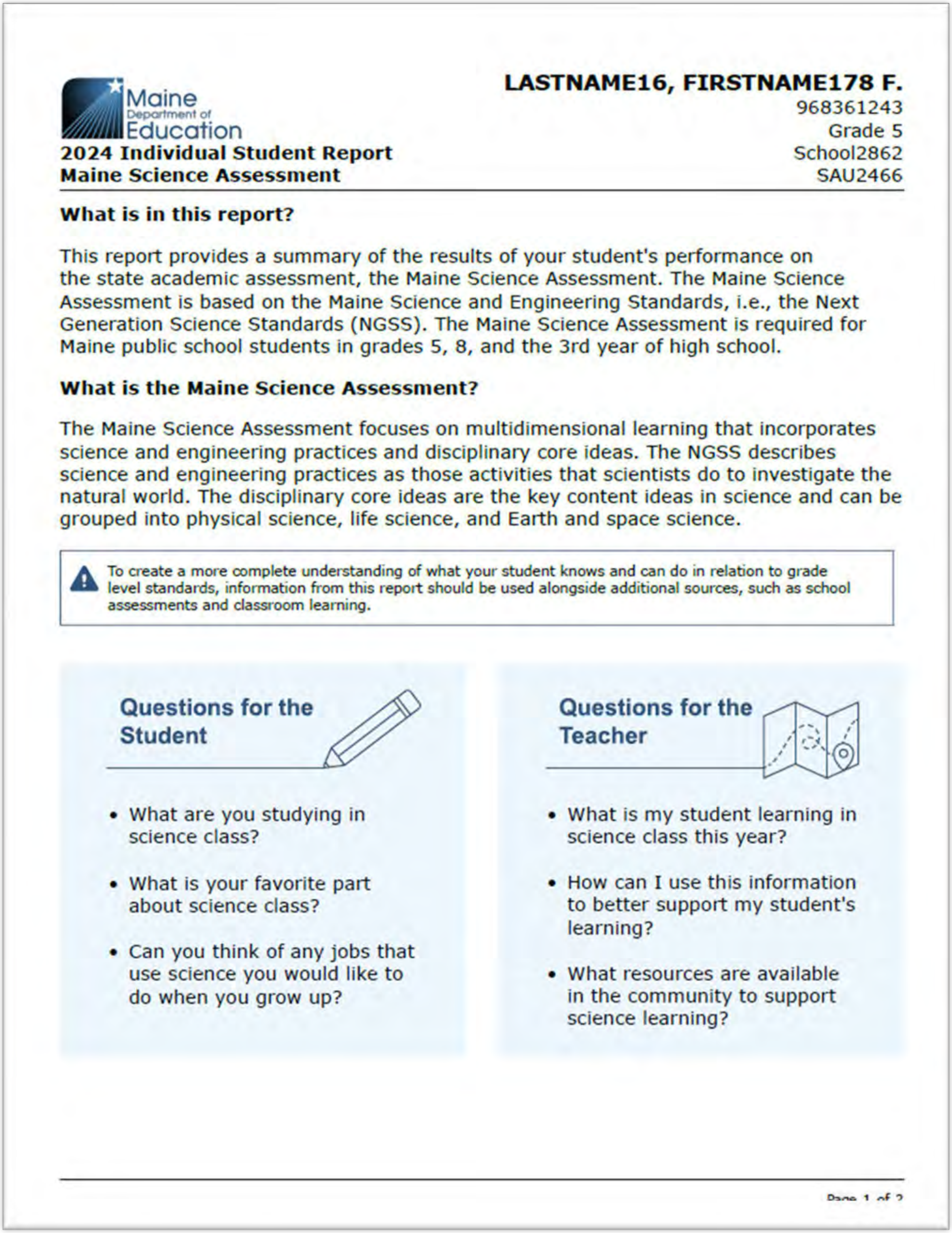
The committee members individually reviewed PE, SEP, DCI, and CCC alignment for each item and recorded the degree of alignment for each dimension and overall alignment on a worksheet on a scale of 0 (not aligned) to 3 (well-aligned) as they referred to Contributing State A's state standards. An item was considered to have a high degree of alignment if it aligned to the bullet listed in the PE. An item was considered to have a lower degree of alignment if it aligned to another bullet listed in the learning progression for that SEP or CCC. Committee members also recorded whether the science for each item was accurate and whether each item was free of bias. Areas of concern included opportunity and access, portrayal of groups represented, protecting privacy, and avoiding offensive content.

Appendix T. Score Reports

(Documents begin on the next page.)

Individual Student Reports (ISRs) – PDFs

(Documents begin on the next page.)



Maine Department of Education

2024 Individual Student Report
Maine Science Assessment

LASTNAME16, FIRSTNAME178 F.
968361243
Grade 5
School2862
SAU2466

What is in this report?

This report provides a summary of the results of your student's performance on the state academic assessment, the Maine Science Assessment. The Maine Science Assessment is based on the Maine Science and Engineering Standards, i.e., the Next Generation Science Standards (NGSS). The Maine Science Assessment is required for Maine public school students in grades 5, 8, and the 3rd year of high school.

What is the Maine Science Assessment?

The Maine Science Assessment focuses on multidimensional learning that incorporates science and engineering practices and disciplinary core ideas. The NGSS describes science and engineering practices as those activities that scientists do to investigate the natural world. The disciplinary core ideas are the key content ideas in science and can be grouped into physical science, life science, and Earth and space science.

⚠ To create a more complete understanding of what your student knows and can do in relation to grade level standards, information from this report should be used alongside additional sources, such as school assessments and classroom learning.

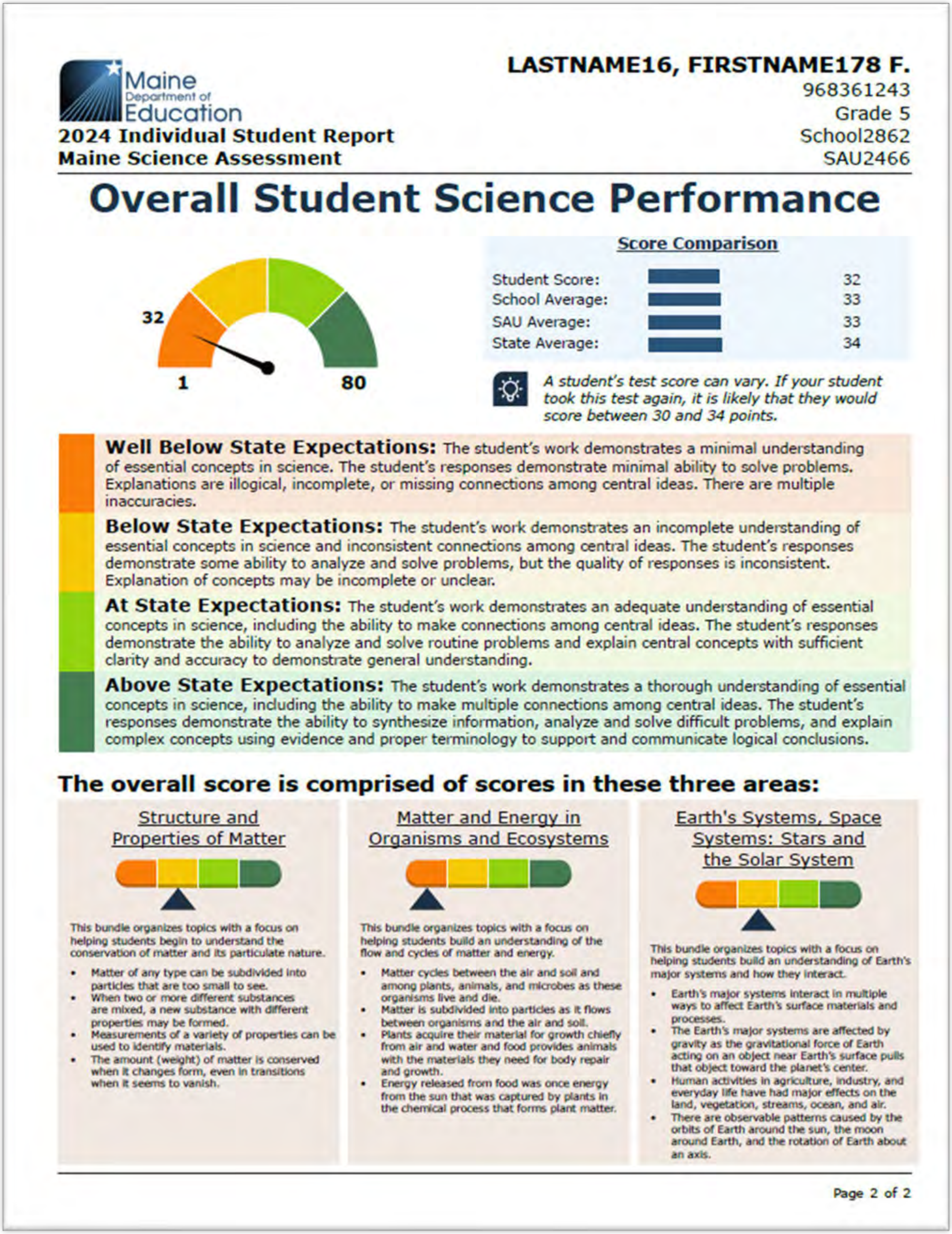
Questions for the Student

- What are you studying in science class?
- What is your favorite part about science class?
- Can you think of any jobs that use science you would like to do when you grow up?

Questions for the Teacher

- What is my student learning in science class this year?
- How can I use this information to better support my student's learning?
- What resources are available in the community to support science learning?

Page 1 of 2



Maine Department of Education

2024 Individual Student Report
Maine Science Assessment

LASTNAME16, FIRSTNAME178 F.
968361243
Grade 5
School2862
SAU2466

Overall Student Science Performance

Score Comparison

Student Score:	32
School Average:	33
SAU Average:	33
State Average:	34

A student's test score can vary. If your student took this test again, it is likely that they would score between 30 and 34 points.

Well Below State Expectations: The student's work demonstrates a minimal understanding of essential concepts in science. The student's responses demonstrate minimal ability to solve problems. Explanations are illogical, incomplete, or missing connections among central ideas. There are multiple inaccuracies.

Below State Expectations: The student's work demonstrates an incomplete understanding of essential concepts in science and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems, but the quality of responses is inconsistent. Explanation of concepts may be incomplete or unclear.

At State Expectations: The student's work demonstrates an adequate understanding of essential concepts in science, including the ability to make connections among central ideas. The student's responses demonstrate the ability to analyze and solve routine problems and explain central concepts with sufficient clarity and accuracy to demonstrate general understanding.

Above State Expectations: The student's work demonstrates a thorough understanding of essential concepts in science, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and explain complex concepts using evidence and proper terminology to support and communicate logical conclusions.

The overall score is comprised of scores in these three areas:

Structure and Properties of Matter

This bundle organizes topics with a focus on helping students begin to understand the conservation of matter and its particulate nature.

- Matter of any type can be subdivided into particles that are too small to see.
- When two or more different substances are mixed, a new substance with different properties may be formed.
- Measurements of a variety of properties can be used to identify materials.
- The amount (weight) of matter is conserved when it changes form, even in transitions when it seems to vanish.

Matter and Energy in Organisms and Ecosystems

This bundle organizes topics with a focus on helping students build an understanding of the flow and cycles of matter and energy.

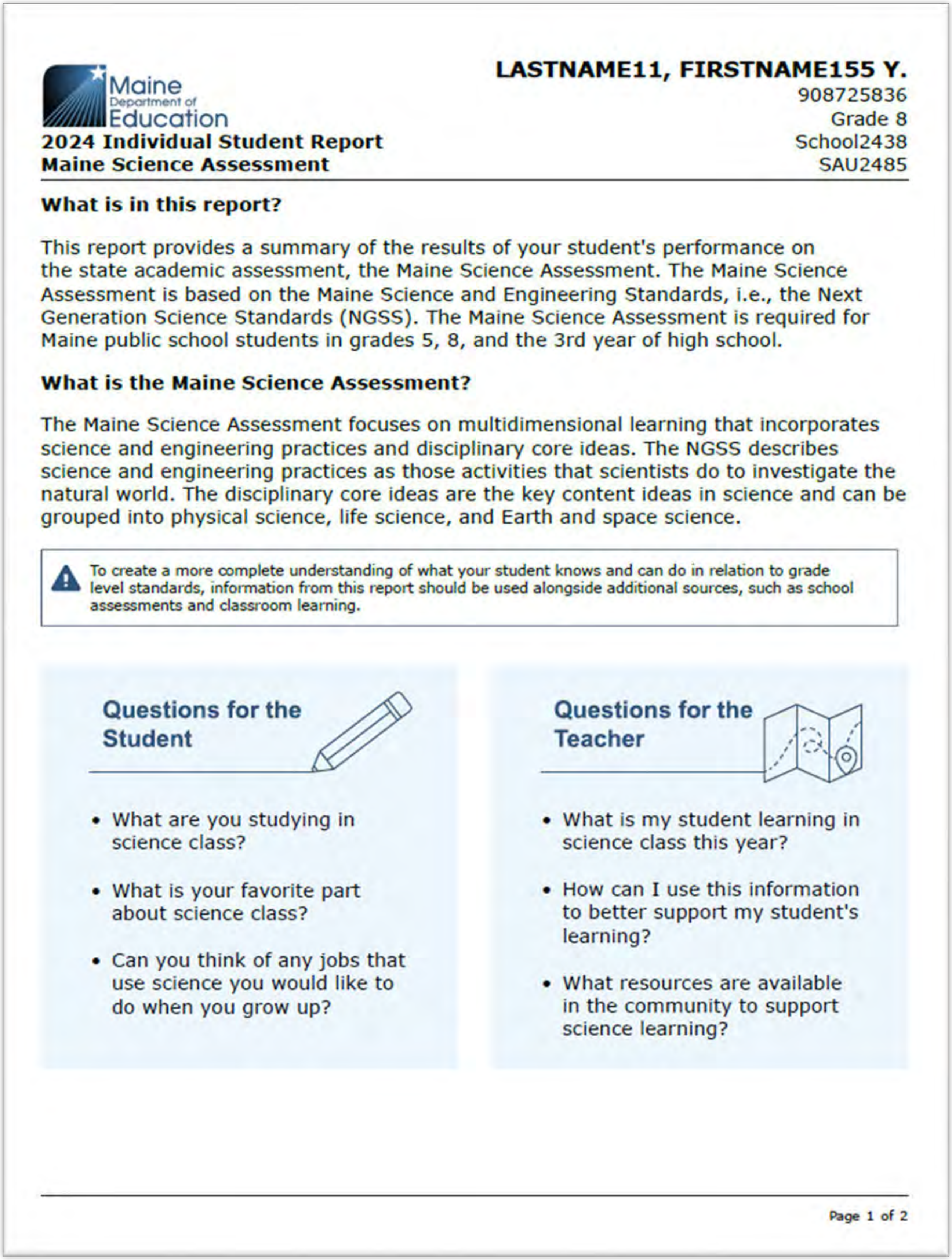
- Matter cycles between the air and soil and among plants, animals, and microbes as these organisms live and die.
- Matter is subdivided into particles as it flows between organisms and the air and soil.
- Plants acquire their material for growth chiefly from air and water and food provides animals with the materials they need for body repair and growth.
- Energy released from food was once energy from the sun that was captured by plants in the chemical process that forms plant matter.

Earth's Systems, Space Systems: Stars and the Solar System

This bundle organizes topics with a focus on helping students build an understanding of Earth's major systems and how they interact.

- Earth's major systems interact in multiple ways to affect Earth's surface materials and processes.
- The Earth's major systems are affected by gravity as the gravitational force of Earth acting on an object near Earth's surface pulls that object toward the planet's center.
- Human activities in agriculture, industry, and everyday life have had major effects on the land, vegetation, streams, ocean, and air.
- There are observable patterns caused by the orbits of Earth around the sun, the moon around Earth, and the rotation of Earth about an axis.

Page 2 of 2



Maine Department of Education

LASTNAME11, FIRSTNAME155 Y.
908725836
Grade 8
School2438
SAU2485

2024 Individual Student Report
Maine Science Assessment

What is in this report?

This report provides a summary of the results of your student's performance on the state academic assessment, the Maine Science Assessment. The Maine Science Assessment is based on the Maine Science and Engineering Standards, i.e., the Next Generation Science Standards (NGSS). The Maine Science Assessment is required for Maine public school students in grades 5, 8, and the 3rd year of high school.

What is the Maine Science Assessment?

The Maine Science Assessment focuses on multidimensional learning that incorporates science and engineering practices and disciplinary core ideas. The NGSS describes science and engineering practices as those activities that scientists do to investigate the natural world. The disciplinary core ideas are the key content ideas in science and can be grouped into physical science, life science, and Earth and space science.

⚠ To create a more complete understanding of what your student knows and can do in relation to grade level standards, information from this report should be used alongside additional sources, such as school assessments and classroom learning.

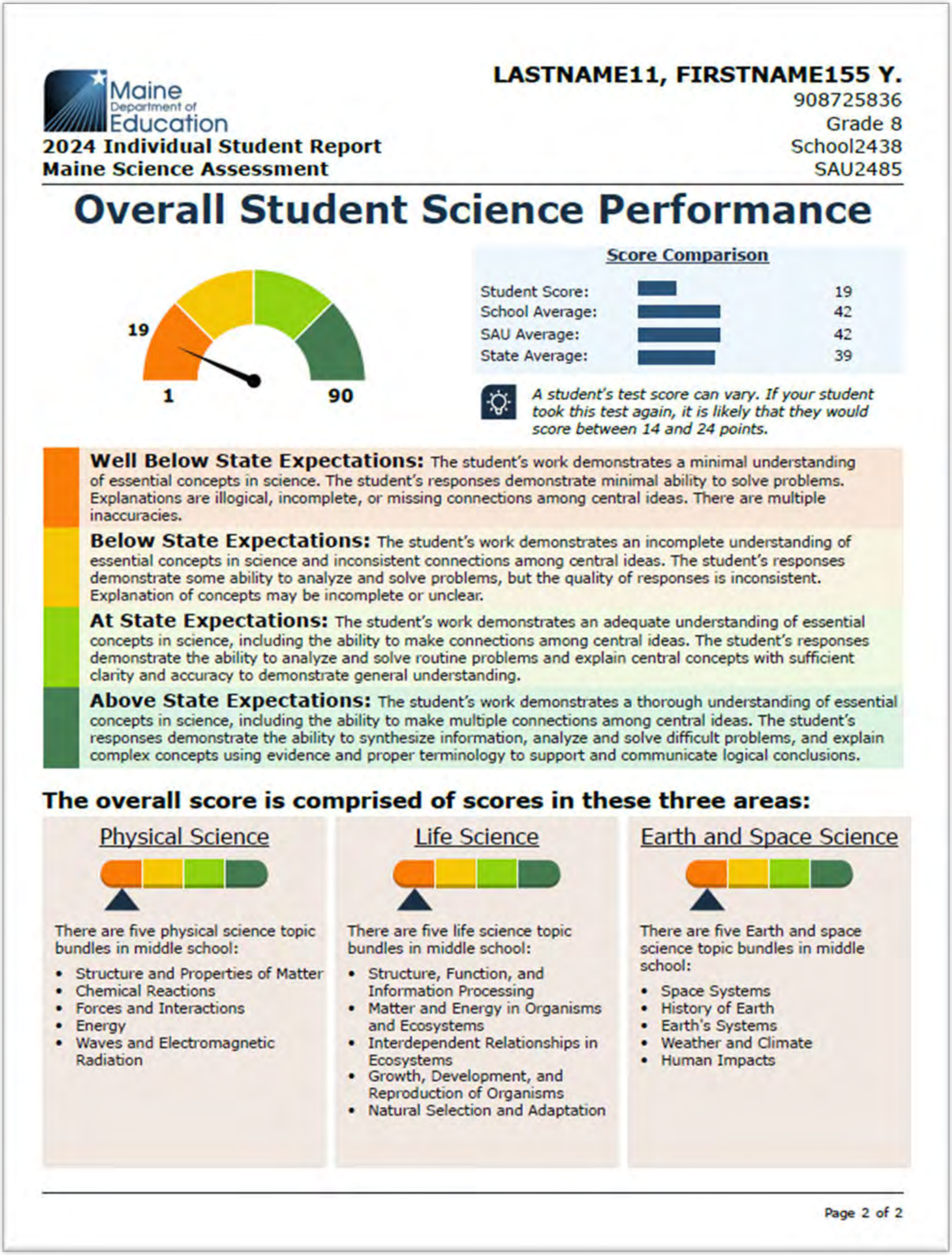
Questions for the Student

- What are you studying in science class?
- What is your favorite part about science class?
- Can you think of any jobs that use science you would like to do when you grow up?

Questions for the Teacher

- What is my student learning in science class this year?
- How can I use this information to better support my student's learning?
- What resources are available in the community to support science learning?

Page 1 of 2



Maine Department of Education

2024 Individual Student Report
Maine Science Assessment

LASTNAME11, FIRSTNAME155 Y.
908725836
Grade 8
School2438
SAU2485

Overall Student Science Performance

Score Comparison

Student Score:		19
School Average:		42
SAU Average:		42
State Average:		39

A student's test score can vary. If your student took this test again, it is likely that they would score between 14 and 24 points.

Well Below State Expectations: The student's work demonstrates a minimal understanding of essential concepts in science. The student's responses demonstrate minimal ability to solve problems. Explanations are illogical, incomplete, or missing connections among central ideas. There are multiple inaccuracies.

Below State Expectations: The student's work demonstrates an incomplete understanding of essential concepts in science and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems, but the quality of responses is inconsistent. Explanation of concepts may be incomplete or unclear.

At State Expectations: The student's work demonstrates an adequate understanding of essential concepts in science, including the ability to make connections among central ideas. The student's responses demonstrate the ability to analyze and solve routine problems and explain central concepts with sufficient clarity and accuracy to demonstrate general understanding.

Above State Expectations: The student's work demonstrates a thorough understanding of essential concepts in science, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and explain complex concepts using evidence and proper terminology to support and communicate logical conclusions.

The overall score is comprised of scores in these three areas:

Physical Science

There are five physical science topic bundles in middle school:

- Structure and Properties of Matter
- Chemical Reactions
- Forces and Interactions
- Energy
- Waves and Electromagnetic Radiation

Life Science

There are five life science topic bundles in middle school:

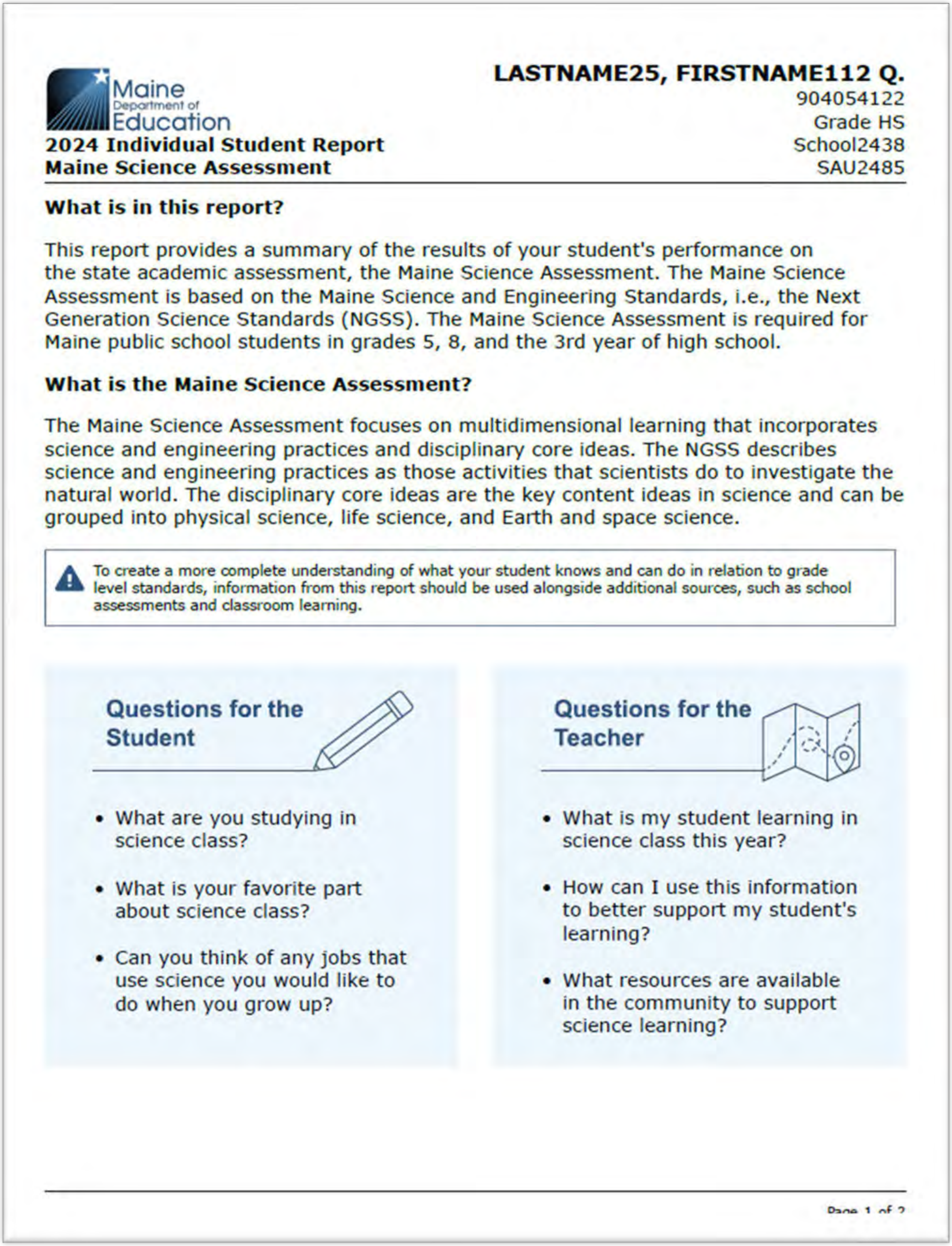
- Structure, Function, and Information Processing
- Matter and Energy in Organisms and Ecosystems
- Interdependent Relationships in Ecosystems
- Growth, Development, and Reproduction of Organisms
- Natural Selection and Adaptation

Earth and Space Science

There are five Earth and space science topic bundles in middle school:

- Space Systems
- History of Earth
- Earth's Systems
- Weather and Climate
- Human Impacts

Page 2 of 2



Maine Department of Education

2024 Individual Student Report
Maine Science Assessment

LASTNAME25, FIRSTNAME112 Q.
904054122
Grade HS
School2438
SAU2485

What is in this report?

This report provides a summary of the results of your student's performance on the state academic assessment, the Maine Science Assessment. The Maine Science Assessment is based on the Maine Science and Engineering Standards, i.e., the Next Generation Science Standards (NGSS). The Maine Science Assessment is required for Maine public school students in grades 5, 8, and the 3rd year of high school.

What is the Maine Science Assessment?

The Maine Science Assessment focuses on multidimensional learning that incorporates science and engineering practices and disciplinary core ideas. The NGSS describes science and engineering practices as those activities that scientists do to investigate the natural world. The disciplinary core ideas are the key content ideas in science and can be grouped into physical science, life science, and Earth and space science.

⚠ To create a more complete understanding of what your student knows and can do in relation to grade level standards, information from this report should be used alongside additional sources, such as school assessments and classroom learning.

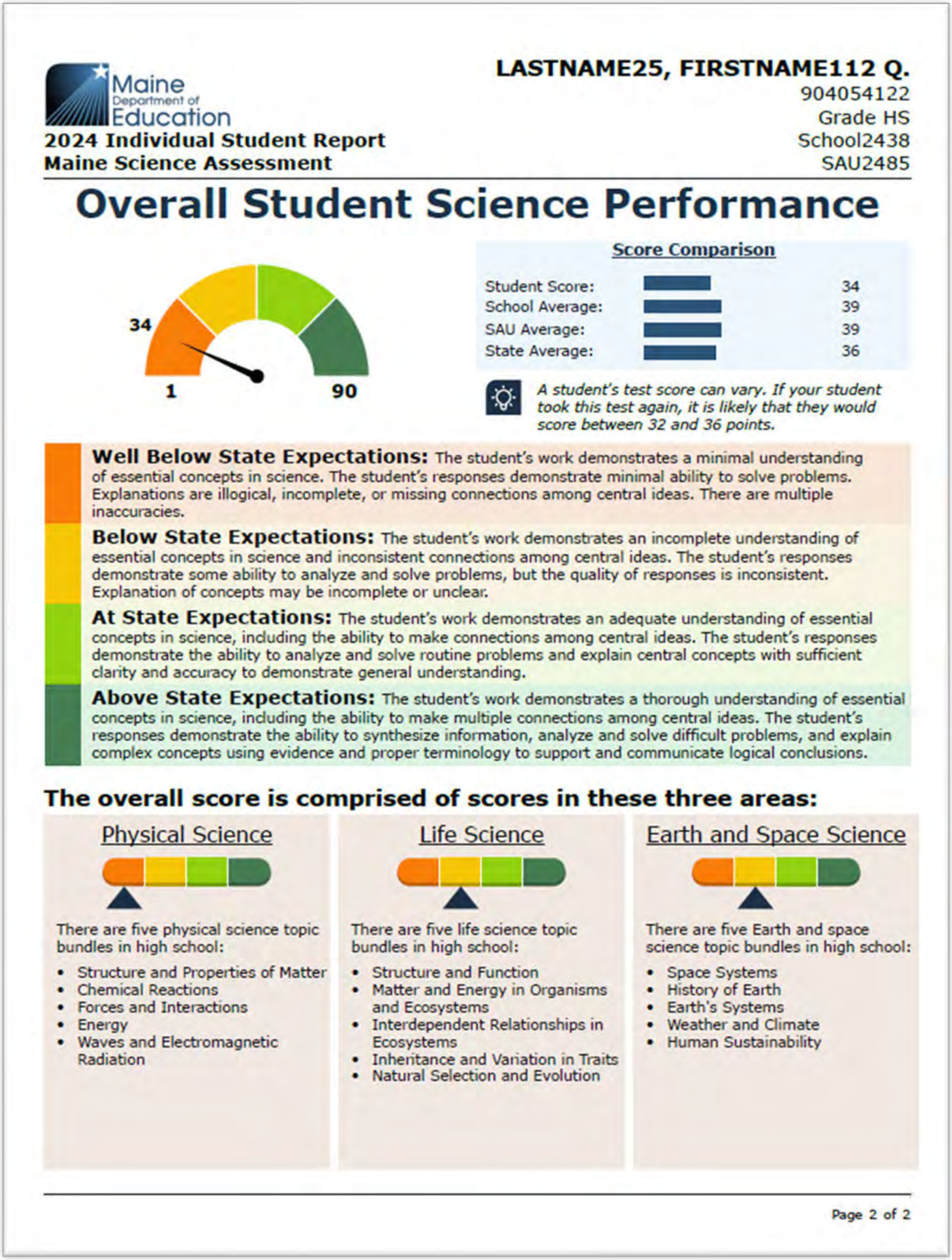
Questions for the Student

- What are you studying in science class?
- What is your favorite part about science class?
- Can you think of any jobs that use science you would like to do when you grow up?

Questions for the Teacher

- What is my student learning in science class this year?
- How can I use this information to better support my student's learning?
- What resources are available in the community to support science learning?

Page 1 of 2



Maine Department of Education

2024 Individual Student Report
Maine Science Assessment

LASTNAME25, FIRSTNAME112 Q.
904054122
Grade HS
School2438
SAU2485

Overall Student Science Performance

34

1

90

Score Comparison

Student Score:	34
School Average:	39
SAU Average:	39
State Average:	36

A student's test score can vary. If your student took this test again, it is likely that they would score between 32 and 36 points.

Well Below State Expectations: The student's work demonstrates a minimal understanding of essential concepts in science. The student's responses demonstrate minimal ability to solve problems. Explanations are illogical, incomplete, or missing connections among central ideas. There are multiple inaccuracies.

Below State Expectations: The student's work demonstrates an incomplete understanding of essential concepts in science and inconsistent connections among central ideas. The student's responses demonstrate some ability to analyze and solve problems, but the quality of responses is inconsistent. Explanation of concepts may be incomplete or unclear.

At State Expectations: The student's work demonstrates an adequate understanding of essential concepts in science, including the ability to make connections among central ideas. The student's responses demonstrate the ability to analyze and solve routine problems and explain central concepts with sufficient clarity and accuracy to demonstrate general understanding.

Above State Expectations: The student's work demonstrates a thorough understanding of essential concepts in science, including the ability to make multiple connections among central ideas. The student's responses demonstrate the ability to synthesize information, analyze and solve difficult problems, and explain complex concepts using evidence and proper terminology to support and communicate logical conclusions.

The overall score is comprised of scores in these three areas:

Physical Science

There are five physical science topic bundles in high school:

- Structure and Properties of Matter
- Chemical Reactions
- Forces and Interactions
- Energy
- Waves and Electromagnetic Radiation

Life Science

There are five life science topic bundles in high school:

- Structure and Function
- Matter and Energy in Organisms and Ecosystems
- Interdependent Relationships in Ecosystems
- Inheritance and Variation in Traits
- Natural Selection and Evolution

Earth and Space Science

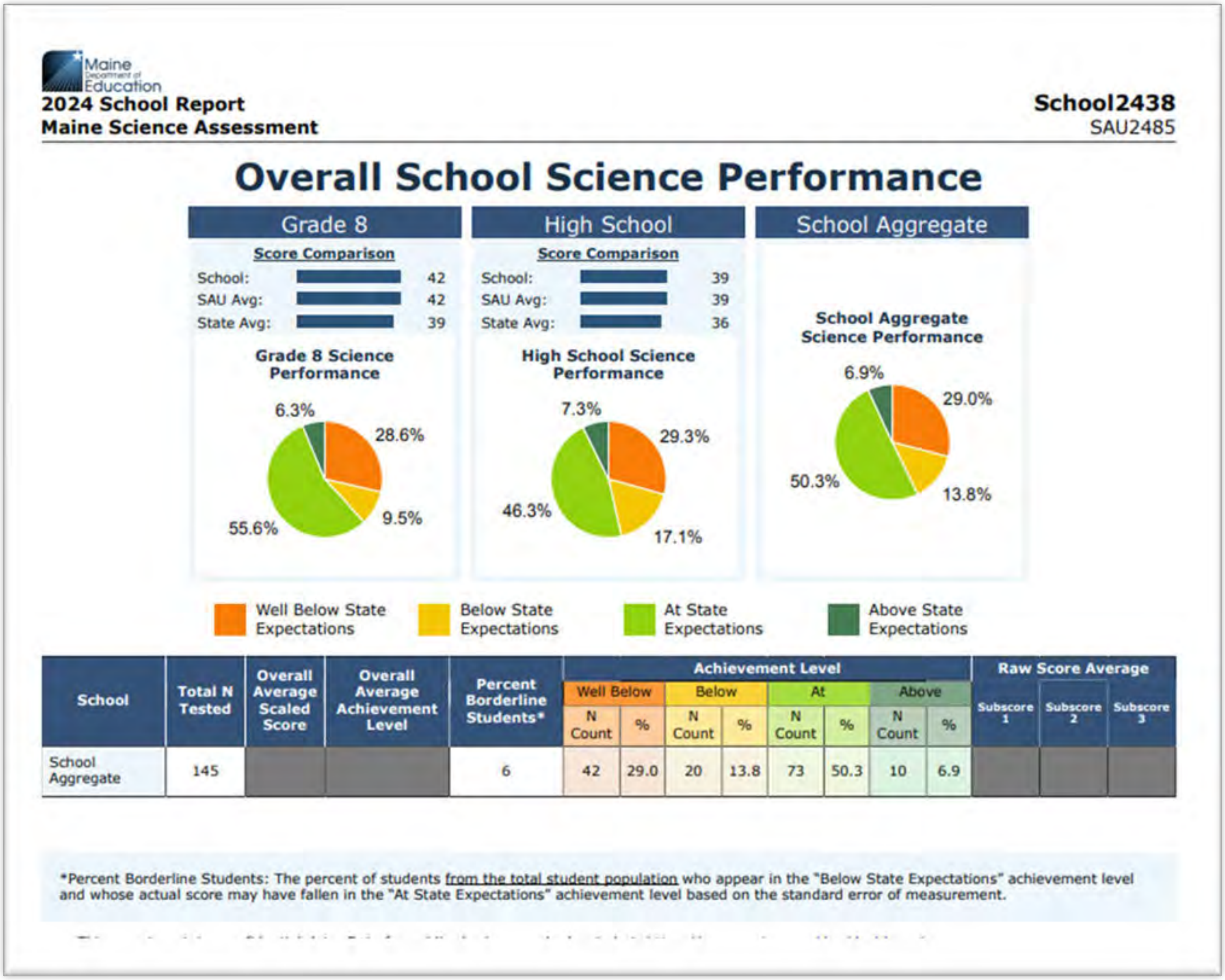
There are five Earth and space science topic bundles in high school:

- Space Systems
- History of Earth
- Earth's Systems
- Weather and Climate
- Human Sustainability

Page 2 of 2

School Summary Report (PDF)

Page 1



Maine Department of Education
2024 School Report
Maine Science Assessment

School2438
SAU2485

Overall School Science Performance

Grade 8		High School		School Aggregate
Score Comparison		**Score Comparison**		
School:	42	School:	39	
SAU Avg:	42	SAU Avg:	39	
State Avg:	39	State Avg:	36	

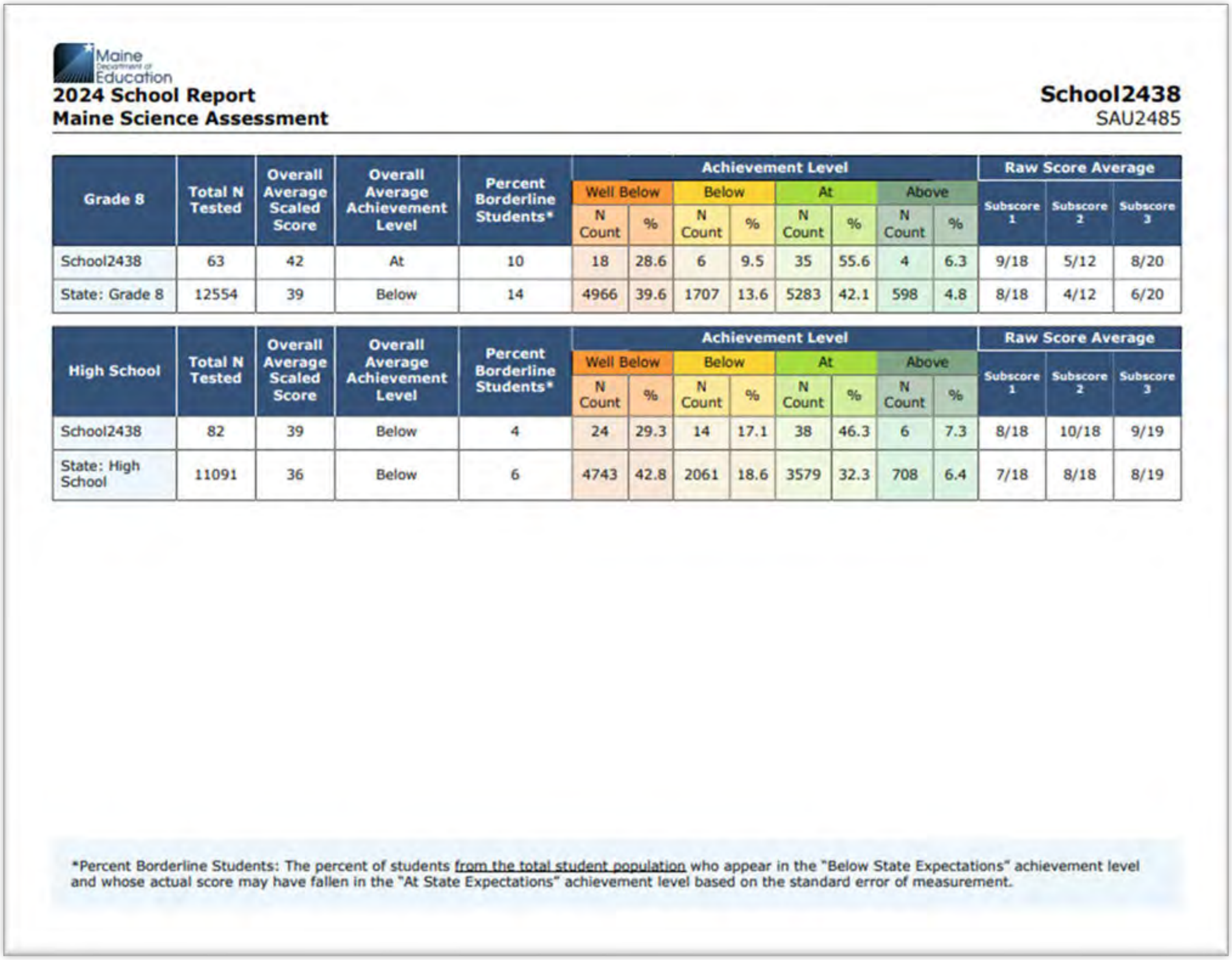
Grade 8 Science Performance
6.3%
28.6%
9.5%
55.6%
High School Science Performance
7.3%
29.3%
17.1%
46.3%
School Aggregate Science Performance
6.9%
29.0%
13.8%
50.3%

🟧 Well Below State Expectations
🟨 Below State Expectations
🟩 At State Expectations
🟩 Above State Expectations

School	Total N Tested	Overall Average Scaled Score	Overall Average Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School Aggregate	145			6	42	29.0	20	13.8	73	50.3	10	6.9			

*Percent Borderline Students: The percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

School Summary Report (PDF) – Page 2



Maine Department of Education

2024 School Report
Maine Science Assessment

School2438
SAU2485

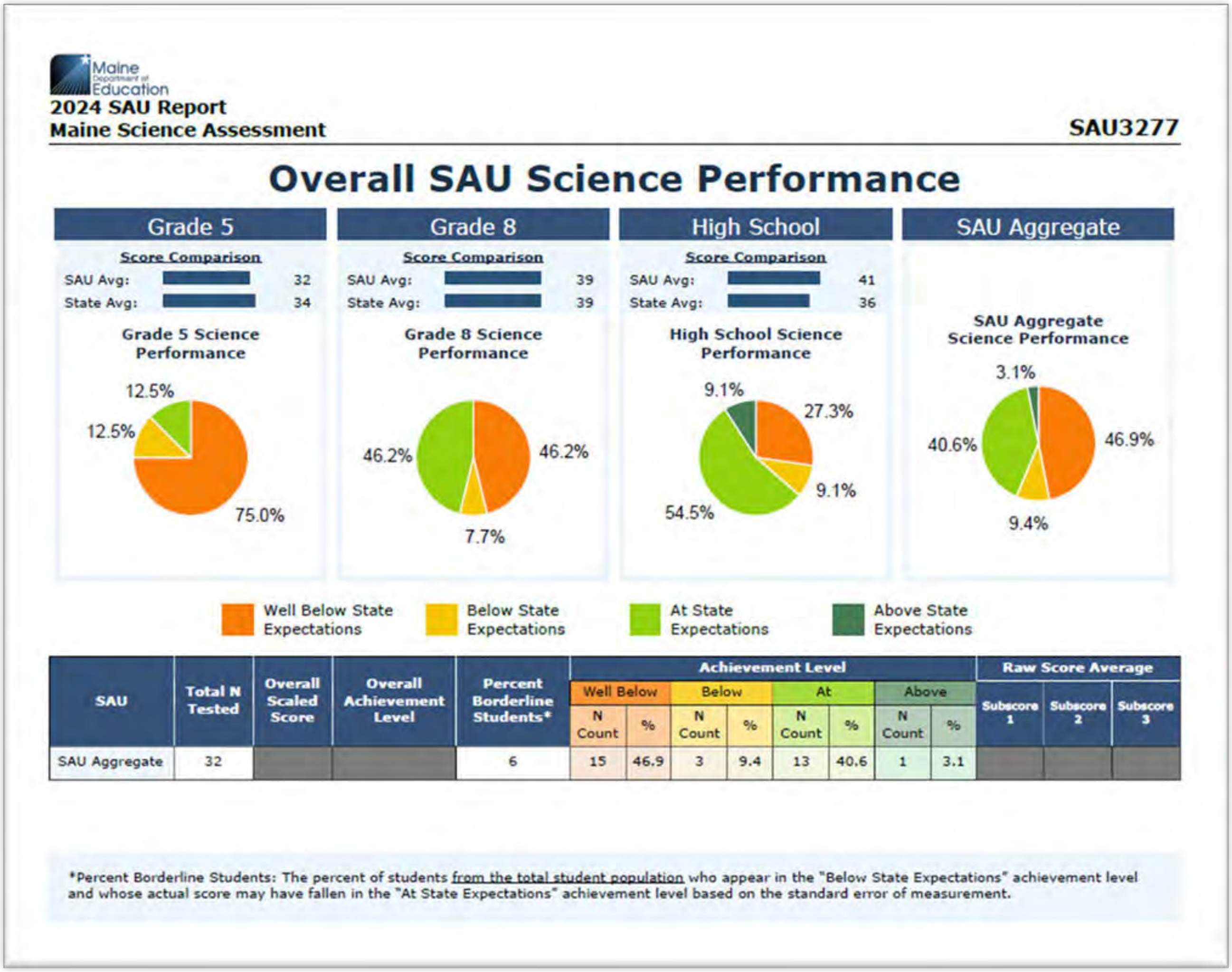
Grade 8	Total N Tested	Overall Average Scaled Score	Overall Average Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School2438	63	42	At	10	18	28.6	6	9.5	35	55.6	4	6.3	9/18	5/12	8/20
State: Grade 8	12554	39	Below	14	4966	39.6	1707	13.6	5283	42.1	598	4.8	8/18	4/12	6/20

High School	Total N Tested	Overall Average Scaled Score	Overall Average Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School2438	82	39	Below	4	24	29.3	14	17.1	38	46.3	6	7.3	8/18	10/18	9/19
State: High School	11091	36	Below	6	4743	42.8	2061	18.6	3579	32.3	708	6.4	7/18	8/18	8/19

*Percent Borderline Students: The percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

SAU Summary Report (PDF)

Page 1



Maine Department of Education
2024 SAU Report
Maine Science Assessment

SAU3277

Overall SAU Science Performance

Grade 5

Score Comparison

SAU Avg:	32
State Avg:	34

Grade 5 Science Performance

Grade 8

Score Comparison

SAU Avg:	39
State Avg:	39

Grade 8 Science Performance

High School

Score Comparison

SAU Avg:	41
State Avg:	36

High School Science Performance

SAU Aggregate

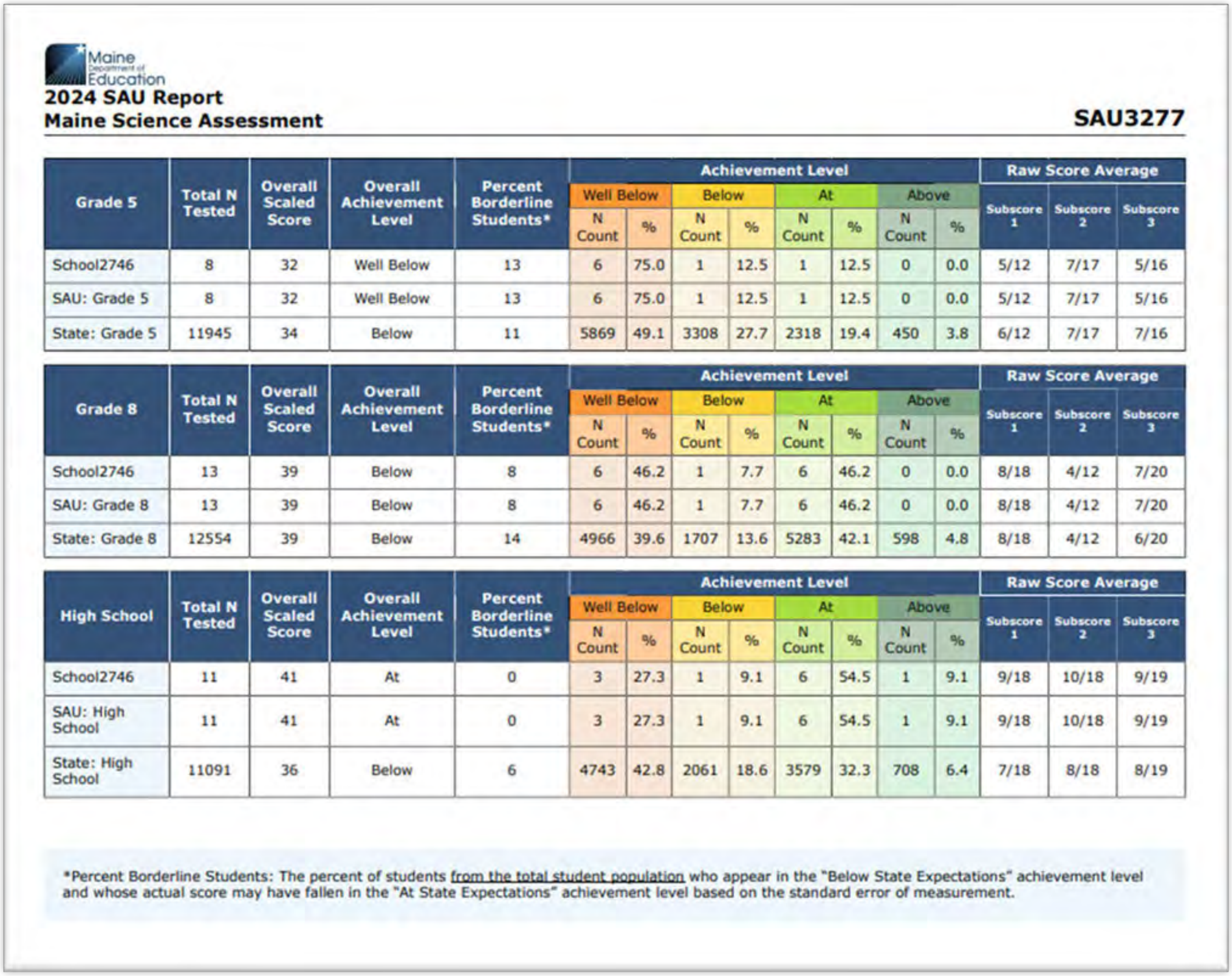
SAU Aggregate Science Performance

🟧 Well Below State Expectations 🟨 Below State Expectations 🟩 At State Expectations 🟩 Above State Expectations

SAU	Total N Tested	Overall Scaled Score	Overall Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
SAU Aggregate	32			6	15	46.9	3	9.4	13	40.6	1	3.1			

*Percent Borderline Students: The percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.

SAU Summary Report (PDF) – Page 2



Maine Department of Education

2024 SAU Report

Maine Science Assessment

SAU3277

Grade 5	Total N Tested	Overall Scaled Score	Overall Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School2746	8	32	Well Below	13	6	75.0	1	12.5	1	12.5	0	0.0	5/12	7/17	5/16
SAU: Grade 5	8	32	Well Below	13	6	75.0	1	12.5	1	12.5	0	0.0	5/12	7/17	5/16
State: Grade 5	11945	34	Below	11	5869	49.1	3308	27.7	2318	19.4	450	3.8	6/12	7/17	7/16

Grade 8	Total N Tested	Overall Scaled Score	Overall Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School2746	13	39	Below	8	6	46.2	1	7.7	6	46.2	0	0.0	8/18	4/12	7/20
SAU: Grade 8	13	39	Below	8	6	46.2	1	7.7	6	46.2	0	0.0	8/18	4/12	7/20
State: Grade 8	12554	39	Below	14	4966	39.6	1707	13.6	5283	42.1	598	4.8	8/18	4/12	6/20

High School	Total N Tested	Overall Scaled Score	Overall Achievement Level	Percent Borderline Students*	Achievement Level								Raw Score Average		
					Well Below		Below		At		Above		Subscore 1	Subscore 2	Subscore 3
					N Count	%	N Count	%	N Count	%	N Count	%			
School2746	11	41	At	0	3	27.3	1	9.1	6	54.5	1	9.1	9/18	10/18	9/19
SAU: High School	11	41	At	0	3	27.3	1	9.1	6	54.5	1	9.1	9/18	10/18	9/19
State: High School	11091	36	Below	6	4743	42.8	2061	18.6	3579	32.3	708	6.4	7/18	8/18	8/19

*Percent Borderline Students: The percent of students from the total student population who appear in the "Below State Expectations" achievement level and whose actual score may have fallen in the "At State Expectations" achievement level based on the standard error of measurement.