New Jersey

Student Learning Assessment—Science

(NJSLA–S)

TECHNICAL REPORT

Grades 5, 8, and 11

2025

April 2026

PTM XXXX.XX

State of New Jersey
Department of Education

Copyright © 2025 by New Jersey Department of Education
All rights reserved

State Board of Education

Kathy Goldenberg .... Burlington
President

Nedd James Johnson .... Salem
Vice President

Arcelio Aponte .... Middlesex

Mary G. Bennett .... Essex

Mary Beth Berry .... Hunterdon

Elaine Bobrove .... Camden

Ronald K. Butcher .... Gloucester

Jack Fornaro .... Warren

Claudine Keenan .... Atlantic

Jeanette Peña .... Hudson

Joseph Ricca Jr. .... Morris

Ahmed Shehata .... Union

Mr. Kevin Dehmer, Commissioner of Education

Secretary, State Board of Education

It is a policy of the New Jersey State Board of Education and the State Department of Education that no person, on the basis of race, creed, national origin, age, sex, handicap, or marital status, shall be subjected to discrimination in employment or be excluded from or denied benefits in any activity, program, or service for which the department has responsibility. The department will comply with all state and federal laws and regulations concerning nondiscrimination.

Table of Contents

Tables and Figures

PART 1: INTRODUCTION

This Technical Report provides information about the technical characteristics of the 2025 administration of the New Jersey Student Learning Assessment–Science (NJSLA–S) to fifth-, eighth-, and eleventh-grade students. The NJSLA–S is administered under the direction of the New Jersey Department of Education (NJDOE). This report provides extensive detail about the development and operation of the NJSLA–S and is intended for use by those who evaluate tests, interpret scores, or use test results for making educational decisions. The documentation in this report is based on the measurement procedures stated in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), hereafter referred to as the "*Standards*."

The NJSLA−S is an integrated program of testing, accountability, and curricular and instructional support. The test itself is one part of a complex network intended to help schools focus their energies on improving student learning. As such, it can only be evaluated properly within this full context. Detailed descriptions of the NJSLA–S 2025 test development, administration, scoring, and reporting are provided in Parts 2, 3, 4, and 10 of this document. Psychometric discussions of item and test statistics, equating and scaling, reliability, and validity can be found in Parts 6, 7, 8, and 9.

Data for the analyses presented in this Technical Report were collected from the NJSLA–S spring administration from April 28, 2025, through May 23, 2025.

- The standard setting discussed in Part 5 of this report is based on a standard-setting study conducted in 2019. More details about the 2019 standard-setting study can be found in the *2019 NJSLA–S Technical Report* (NJDOE, 2019).
- Analyses in Parts 6 (Item and Test Statistics) and 8 (Reliability) of this report are based on test results from the entire state population of fifth-, eighth-, and eleventh-grade students.

1.1 Purpose of the Assessment

The 1965 Elementary and Secondary Education Act (ESEA), as reauthorized by the 2015 Every Student Succeeds Act (ESSA, 2015), contained requirements for each state to assess science at least once during grades 3–5, grades 6–9, and grades 10−12. The NJSLA–S measures student proficiency annually in grades 5, 8, and 11 with regard to the New Jersey Student Learning Standards for Science, adopted in 2014 for implementation by the start of the 2016–17 school year for grades 6–12 and by the start of the 2017–18 school year for grades K–5. These science standards are based upon the *Framework for K–12 Science Education* (National Research Council [NRC], 2012), which identifies the science knowledge and skills that all K–12 students should know, and the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), developed collaboratively by stakeholders across 25 states. The emphasis in instruction and assessment is on learning and understanding core principles and theories.

The New Jersey Student Learning Assessments are part of an ongoing system of activities that provide evidence related to student learning. The data from the NJSLA–S, students' daily interactions with teachers, and their performance on teacher- and district-developed assessments combine to provide a complete picture of student achievement in science. Schools and local education agencies (LEAs) should use the results to identify strengths and weaknesses in their educational programs. The results may also be used, along with other indicators of student progress, to identify those students who may need instructional support to address any identified knowledge or skill gaps.

1.2 Description of the Assessment

The NJSLA–S assesses students in grades 5, 8, and 11 on their understanding and explanations of scientific phenomena and scenarios. The 2018–19 school year marked the first administration of the NJSLA–S, the spring 2019 operational administration was the assessment's baseline year, and 2025 was the fifth year of administration. The assessment was not administered in 2020 and 2021 due to the COVID pandemic.

The NJSLA–S comprises two parts—the performance-based assessment (PBA) and the machine-scorable assessment (MSA). The PBA contains one open-ended, constructed-response item and between two and four technology-enhanced (TE) items. The MSA contains a mixture of TE items and multiple-choice items.

Furthermore, the tests cover a range of materials. To accomplish the necessary scope, each test item requires students to address multiple underlying variables, with items representing an interaction of Disciplinary Core Ideas (DCIs)—within the domains of Physical, Life, and Earth and Space Science; Science and Engineering Practices (SEPs)—Investigating, Sensemaking, or Critiquing; and Crosscutting Concepts (CCC). Every test item counts toward the students' proficiency in exactly one reported domain and one reported practice. (Each item is also aligned to a CCC. The CCCs and the knowledge, skills, and abilities [KSAs] associated with them contribute to the overall scale score. However, there is no specific reported CCC proficiency indicator for the NJSLA−S.)

1.2.1 Content Domains and Scientific Practices

The NJSLA−S is a unidimensional test designed to assess the New Jersey Student Learning Standards for Science (NJSLS–S). The robust standards have been subdivided into six distinct subcategories for test construction and reporting purposes. The six foundational subcategories are equally divided between three science content domain categories (Earth and Space, Life, and Physical) and three scientific practice categories (Sensemaking, Critiquing, and Investigating).

Science content domains. Disciplinary Core Ideas can be classified into three major science content domains: Earth and Space Science, Life Science, and Physical Science. The NJSLA–S is designed to measure student proficiency in each of the three science content domains. The test development processes focus on balancing each science content domain equally. Furthermore, within each content domain, each DCI is balanced. (See the *Framework* for further information.)

1. *Earth and Space Science.* The *Framework* (NRC, 2012) states that “Earth and space sciences (ESS) investigate processes that operate on Earth and also address its place in the solar system” (p. 169). Table 1.2.1 shows the three ESS DCIs and the topics delineated within each.

Table 1.2.1: Earth and Space Science DCIs

DCI Topic Description
ESS1: Earth’s Place in the Universe
ESS1.A: The universe and its stars
ESS1.B: Earth and the solar system
ESS1.C: The history of planet Earth
ESS2: Earth’s Systems
ESS2.A: Earth materials and systems
ESS2.B: Plate tectonics and large-scale system interactions
ESS2.C: The roles of water in Earth’s surface processes
ESS2.D: Weather and climate
ESS2.E: Biogeology
ESS3: Earth and Human Activity
ESS3.A: Natural resources
ESS3.B: Natural hazards
ESS3.C: Human impacts on Earth systems

2. *Life Science.* The *Framework* (NRC, 2012) for the life sciences (LS) “focus[es] on patterns, processes, and relationships of living organisms” (p. 139). Table 1.2.2 presents the four LS DCIs and their underlying topics.

Table 1.2.2: Life Science DCIs

DCI Topic Description
LS1: From Molecules to Organisms: Structures and Processes
LS1.A: Structure and function
LS1.B: Growth and development of organisms
LS1.C: Organization for matter and energy flow in organisms
LS1.D: Information processing
LS2: Ecosystems: Interactions, Energy, and Dynamics
LS2.A: Interdependent relationships in ecosystems
LS2.B: Cycles of matter and energy transfer in ecosystems
LS2.C: Ecosystem dynamics, functioning, and resilience
LS2.D: Social interactions and group behavior
LS3: Heredity: Inheritance and Variation of Traits
LS3.A: Inheritance of traits
LS3.B: Variation of traits
LS4 Biological Evolution: Unity and Diversity
LS4.A: Evidence of common ancestry and diversity
LS4.B: Natural selection
LS4.C: Adaptation
LS4.D: Biodiversity and humans

3. *Physical Science.* According to the *Framework* (NRC, 2012), the goal of learning physical science (PS) “is to help students see that there are mechanisms of cause and effect in all systems and processes that can be understood through a common set of physical chemical principles” (p. 103). Table 1.2.3 illustrates the four PS DCIs along with the associated detailed topics for each.

Table 1.2.3: Physical Science DCIs

DCI Topic Description
PS1: Matter and Its Interactions
Structure and matter
Chemical reactions
PS2: Motion and Stability: Force and Interactions
Force and motion
Types of interactions
Stability and instability in physical systems
PS3: Energy
Definitions of energy
Conservation of energy and energy transfer
Relationship between energy and forces
Energy in chemical processes and everyday life
PS4: Waves and Their Applications in Technologies for Information Transfer
Wave properties
Electromagnetic radiation
Information technologies and instrumentation

Scientific Practices. The *Framework* (2012) contains eight different Science and Engineering Practices (SEPs). One of the goals of the SEPs is to help "students understand how scientific knowledge develops; such direct involvement gives them an appreciation of the wide range of approaches that are used to investigate, model, and explain the world" (p. 42). Within the context of the NJSLA−S, the SEPs are consolidated into three categories of scientific practices: Investigating, Sensemaking, and Critiquing. Table 1.2.4, adapted from the work of McNeill, et al. (2015), shows how the eight *Framework* SEPs were consolidated for the purposes of the NJSLA−S.

Table 1.2.4: SEP Consolidation

SEP	Grouping
Asking questions and defining problems (AQDP)	Investigating
Planning and carrying out investigations (PACI)	Investigating
Using mathematics and computational thinking (UMCT)	Investigating
Analyzing and interpreting data (AID)	Sensemaking
Constructing explanations and designing solutions (CEDS)	Sensemaking
Developing and using models (DUM)	Sensemaking
Engaging in argument from evidence (EAE)	Critiquing
Obtaining, evaluating, and communicating information (OECI)	Critiquing

1. *Investigating.* Investigating Practices (McNeill et al., 2015) involve asking questions, conducting investigations, and using mathematical skills to probe naturally occurring phenomena. Table 1.2.5 delineates the *Framework* definition of each of the Investigating Practices.

Table 1.2.5: Investigating Practices

SEP	NRC Framework
Asking questions and defining problems (AQDP)	Students at any grade level should be able to ask questions of each other about the texts they read, the features of the phenomena they observe, and the conclusions they draw from their models or scientific investigations. For engineering, they should ask questions to define the problem to be solved and to elicit ideas that lead to the constraints and specifications for its solution. (p. 56)
Planning and carrying out investigations (PACI)	Students should have opportunities to plan and carry out several different kinds of investigations during their K−12 years. At all levels, they should engage in investigations that range from those structured by the teacher—in order to expose an issue or question that they would be unlikely to explore on their own (e.g., measuring specific properties of materials)—to those that emerge from students' own questions. (p. 61)
Using mathematics and computational thinking (UMCT)	Although there are differences in how mathematics and computational thinking are applied in science and in engineering, mathematics often brings these two fields together by enabling engineers to apply the mathematical form of scientific theories and by enabling scientists to use powerful information technologies designed by engineers. Both kinds of professionals can thereby accomplish investigations and analyses and build complex models, which might otherwise be out of the question. (p. 65)

2. *Sensemaking.* Sensemaking Practices (McNeill et al., 2015) are conceptualized as analyzing the data that are produced from an investigation and developing models and explanations that can explain naturally occurring phenomena. Table 1.2.6 illustrates the *Framework* definition of each of the Sensemaking Practices.

Table 1.2.6: Sensemaking Practices

SEP	NRC Framework
Developing and using models (DUM)	Modeling can begin in the earliest grades, with students' models progressing from concrete "pictures" and/or physical scale models (e.g., a toy car) to more abstract representations of relevant relationships in later grades, such as a diagram representing forces on a particular object in a system. (p. 58)
Analyzing and interpreting data (AID)	Once collected, data must be presented in a form that can reveal any patterns and relationships and that allows results to be communicated to others. Because raw data as such have little meaning, a major practice of scientists is to organize and interpret data through tabulating, graphing, or statistical analysis. Such analysis can bring out the meaning of data—and their relevance—so that they may be used as evidence. (p. 61)
Constructing explanations and designing solutions (CEDS)	Asking students to demonstrate their own understandings of the implications of a scientific idea by developing their own explanations of phenomena, whether based on observations they have made or models they have developed, engages them in an essential part of the process by which conceptual change can occur. (p. 68)

3. *Critiquing.* Critiquing Practices (McNeill et al., 2015) are conceptualized as the ability of students to evaluate information, engage in argument, and communicate whether the models, explanations, or interpretations are adequate representations of naturally occurring phenomena. Table 1.2.7 shows the *Framework* definition of each of the Critiquing Practices.

Table 1.2.7: Critiquing Practices

SEP	NRC Framework
Engaging in argument from evidence (EAE)	The study of science and engineering should produce a sense of the process of argument necessary for advancing and defending a new idea or an explanation of a phenomenon and the norms for conducting such arguments. In that spirit, students should argue for the explanations they construct, defend their interpretations of the associated data, and advocate for the designs they propose. (p. 73)
Obtaining, evaluating, and communicating information (OECI)	Any education in science and engineering needs to develop students' ability to read and produce domain-specific text. As such, every science or engineering lesson is in part a language lesson, particularly reading and producing the genres of texts that are intrinsic to science and engineering. (p. 76)

1.2.2 Crosscutting Concepts

The *Framework* (2012) contains seven different Crosscutting Concepts (CCCs). They were selected to help "students with an organizational framework for connecting knowledge from the various disciplines into a coherent and scientifically based view of the world" (p. 83). Due to reporting constraints, the CCCs are the lowest priority of the three dimensions described in the *Framework*. However, because each item is aligned to a CCC, the CCCs and the knowledge, skills, and abilities associated with them are still being assessed by the NJSLA−S and contribute to the overall NJSLA−S scale score. Table 1.2.8 shows the CCCs being measured by the NJSLA−S.

Table 1.2.8: Crosscutting Concepts

CCC	NRC Framework (p. 84)
Patterns	Observed patterns of forms and events guide organization and classification, and they prompt questions about relationships and the factors that influence them.
Cause and Effect	Events have causes, sometimes simple, sometimes multifaceted. A major activity of science is investigating and explaining causal relationships and the mechanisms by which they are mediated. Such mechanisms can then be tested across given contexts and used to predict and explain events in new contexts.
Scale, Proportion, and Quantity	In considering phenomena, it is critical to recognize what is relevant at different measures of size, time, and energy and to recognize how changes in scale, proportion, or quantity affect a system's structure or performance.
Systems and System Models	Defining the system under study—specifying its boundaries and making explicit a model of that system—provides tools for understanding and testing ideas that are applicable throughout science and engineering.
Energy and Matter	Tracking fluxes of energy and matter into, out of, and within systems helps one understand the systems' possibilities and limitations.
Structure and Function	The way in which an object or living thing is shaped and its substructure determine many of its properties and functions.
Stability and Change	For natural and built systems alike, conditions of stability and determinants of rates of change or evolution of a system are critical elements of study.

1.2.3 Types of Scores

Student proficiency on the NJSLA–S is described using scale scores and proficiency levels. Each grade level has its own grade-specific scale that represents a composite score of student proficiency on the three NJSLS–S dimensions (DCIs, SEPs, and CCCs). Student proficiency is classified into four grade-specific proficiency levels based on the NJSLA–S Performance-Level Descriptors (PLDs). Both the scale scores and the performance levels are described below.

- Scale Scores. The NJSLA–S reports scale scores to indicate a student's proficiency. A scale score is a conversion of the raw score (the total number of points a student earned on the test), using a predetermined mathematical algorithm, to permit legitimate and meaningful comparisons. As such, scale scores provide the best generalized information about overall proficiency. The total scores in science are reported as scale scores with a range of 100 to 300.
- Proficiency Levels. One of the primary purposes of the NJSLA–S is to identify areas of curricular strength and weakness by examining the extent to which students meet the established proficiency expectations in science. Based on test results, a student's performance is categorized as being at one of four proficiency levels, each of which is defined by a student's scale score and is used to report overall student proficiency on the NJSLA–S. Grade-appropriate Performance-Level Descriptors (PLDs) translate these proficiency levels into words. They describe the KSAs students should have at proficiency levels 1 through 4. Each proficiency level is associated with a range of scale scores, as indicated in Table 1.2.9:

Table 1.2.9: NJSLA–S Scale Score Ranges by Proficiency Level

Grade	Level 1	Level 2	Level 3	Level 4
5	100–149	150–199	200–242	243–300
8	100–149	150–199	200–230	231–300
11	100–157	158–199	200–249	250–300

Students performing at Level 3 and Level 4 are considered proficient and above; they demonstrate appropriate or exemplary understanding of the DCIs and SEPs. Students performing at Level 1 and Level 2 are considered below the state minimum proficiency level. They demonstrate minimal or partial understanding of the DCIs and SEPs. Students at this proficiency level may need additional instructional support, which could be individual or programmatic intervention.

Student proficiency is also classified as "Below," "Near/Met," or "Above" expectations in each of the three content domains (Earth and Space, Life, and Physical Science) and the three scientific practices (Investigating, Sensemaking, Critiquing). These subscore proficiency classifications are primarily meant to provide teachers, schools, and administrators with feedback as to the specific KSAs that their students displayed on the NJSLA–S. Individual students and their parents and teachers receive student-level data on these subscores.

1.3 Organizational Support

The New Jersey Department of Education's Office of Assessments coordinates the development and implementation of the NJSLA–S. In addition to planning, scheduling, and directing all NJSLA–S activities, the staff is extensively involved in numerous test-design, item and statistical review, security, quality assurance, and analytical procedures. Measurement Incorporated (MI), the primary contractor for the NJSLA–S at grades 5, 8, and 11, is responsible for all aspects of the testing program, including activities such as program management, development of tests, publishing documents for test administration, handscoring constructed-response items, and psychometric support (including standard setting). Pearson, the subcontractor for the 2025 NJSLA–S, provides item banking, test registration, administration, digital delivery, and reporting. MI and Pearson work closely together under the direction of the Office of Assessments to ensure ancillary materials and administrative procedures closely match those of the NJSLA–Math and NJSLA–ELA assessments.

PART 2: TEST DEVELOPMENT

The NJSLA−S is aligned to the New Jersey Student Learning Standards for Science (NJSLS−S), adopted in 2014, which in turn are based upon the National Research Council's *Framework for K–12 Science Education* and the Next Generation Science Standards (NGSS).

The Test Design and Development chapter within the *Standards* (2014) outlines a series of five primary phases of the test development process: (1) test specifications; (2) item development and review; (3) assembling and evaluating test forms; (4) development of procedures and materials for test administration and scoring; and (5) test revisions (p. 83). The following sections in Part 2 detail the NJSLA–S test specifications, item development processes, and both the test construction processes and their results in 2025. The development of procedures and materials for test administration and scoring is covered in Parts 2 and 3.

2.1 Test Specifications

According to the *Standards,* "[t]he term *test specifications* is sometimes limited to description of the content and format of the test. In the *Standards*, test specifications are defined more broadly to also include documentation of the purpose and intended uses of the test, as well as detailed decisions about content, format, test length, psychometric characteristics of the items and test, delivery mode, administration, scoring, and score reporting" (p. 76).

The NJSLA−S was developed to measure the knowledge, skills, and abilities (KSAs) identified in the NJSLS−S in grades 5, 8, and 11. The test is designed to provide reporting information for student proficiency levels at the holistic level and each of the three science content domains (Earth and Space, Life, and Physical) and the three scientific practices (Investigating, Sensemaking, and Critiquing). The test specifications call for a balanced test design that prioritizes each science content domain through each DCI, each scientific practice each SEP, and all seven CCCs. (Please refer to Section 1.2 of this document for an explanation of the DCIs, SEPs, and CCCs.) The detailed information recommended in the *Standards* is presented in the sections that follow.

2.1.1 Test Blueprints

Table 2.1.1 depicts the test blueprint—the numbers of items comprising each part of the test—for all grades. Note that each multiple-choice (MC) item is worth one point; each technology-enhanced (TE) item is worth one point; and each constructed-response (CR) item is worth three or four points. Each CR item is scored using an item-specific rubric. The table summarizes the number of items on the operational NJSLA−S for each of the six reporting categories, as well as for both the performance-based assessment (PBA) and machine-scorable assessment (MSA) components. An explanation of the PBA and MSA components is provided in the following section.

Table 2.1.1: Test Blueprints

Domain	Practice	Grade 5		Grade 8		Grade 11	
		PBA	MSA	PBA	MSA	PBA	MSA
Physical Science	Investigating *AQDP, PACI, UMCT*	1–2	3–5	1–2	4–7	1–2	4–8
	Sensemaking *DUM, AID, CEDS*	1–2	3–5	1–2	4–7	1–2	4–8
	Critiquing *EAE, OECI*	1–2	3–5	1–2	4–7	1–2	4–8
	Total Item Counts	3–5	11–13	3–5	14–18	3–5	15–21
Life Science	Investigating *AQDP, PACI, UMCT*	1–2	3–5	1–2	4–7	1–2	4–8
	Sensemaking *DUM, AID, CEDS*	1–2	3–5	1–2	4–7	1–2	4–8
	Critiquing *EAE, OECI*	1–2	3–5	1–2	4–7	1–2	4–8
	Total Item Counts	3–5	11–13	3–5	14–18	3–5	15–21
Earth and Space Science	Investigating *AQDP, PACI, UMCT*	1–2	3–5	1–2	4–7	1–2	4–8
	Sensemaking *DUM, AID, CEDS*	1–2	3–5	1–2	4–7	1–2	4–8
	Critiquing *EAE, OECI*	1–2	3–5	1–2	4–7	1–2	4–8
	Total Item Counts	3–5	11–13	3–5	14–18	3–5	15–21

2.1.2 Unit Design

The NJSLA–S consists of four units—three operational units and one field test unit. The units are numbered 1–4, and the field test unit placement varies from year to year. Each unit contains an MSA and a PBA component; a balance of Earth and Space, Life, and Physical Science items; a balance of Investigating, Sensemaking, and Critiquing practice items; a prescribed proportion of MC, TE, and CR item types; and psychometric constraints that are discussed in Section 2.4 of this Technical Report.

Each MSA and PBA component of a unit is linked to naturally occurring phenomena that provide the impetus for scenarios. The students are provided with the scenario and are subsequently presented with two to five items that measure their mastery of the NJSLS–S. All items attached to a phenomenon-based scenario are independent—that is, for example, if a PBA section contains four total items, a student's response to one of the four items will not impact that student's ability to correctly answer any of the other three. Figure 2.1.1 illustrates the composition of a sample grade 5 unit.

MSA: 4 stimuli, 3 items each				**PBA: 1 stimulus, 5 items**
Stim. 1 3 TE 3 points	Stim. 2 2 TE, 1 MC 3 points	Stim. 3 2 TE, 1 MC 3 points	Stim. 4 2 TE, 1 MC 3 points	• 4 one-point TEs • 1 four-point CR
Total # items, MSA: 12 Total points, MSA: 12				Total # items, PBA: 5 Total points, PBA: 8
Total # items, Unit: 17 Total points, Unit: 20				

Figure 2.1.1. Sample Grade 5 Unit

The MSA component of the NJSLA–S is defined as the portion of the assessment that is scored by a computer. Each cluster of MSA items contains a context-dependent stimulus that presents the students with a naturally occurring phenomenon. Depending on the grade level, each unit contains anywhere from four to seven stimuli, and each stimulus is associated with three to six items. MSA items can be either multiple-choice (MC) or technology-enhanced (TE) items, but within each unit no more than 50% of the MSA items can be MC items.

The PBA component of the NJSLA–S is defined as that portion of the test which requires students to display KSAs to a greater degree of cognitive depth, the degree to which the students displayed depth of knowledge and expertise; it is based on more complex phenomena than the MSA section. The PBA components (one per unit) contain one stimulus, each of which can accommodate two to four TE items and one constructed-response (CR) item. Following the 2024 test administration, NJDOE expanded the point range for the PBA section from requiring 7 or 8 total points to requiring 5, 6, 7, or 8 total points, with 3 or 4 of those points coming from the CR item.

2.1.3 Item Types

Table 2.1.2 describes each NJSLA–S item type. Three main types of items comprise the NJSLA–S: multiple-choice (MC), technology-enhanced (TE), and constructed-response (CR).

- MC items all have a key (A, B, C, or D) associated with them, and students are asked to select the best of the four options. MC items are scored dichotomously, 0/1.
- TE items require students to interact with more complex methods of answering the items. Examples of TE item interactions include drop-down choice; hot spot; text entry; drag and drop; multiple selection; and ordering. (A complete list of TE item interactions is presented in Table 2.1.2.) TE items are scored dichotomously, 0/1.
- CR items are open-ended questions designed to elicit a student response to a range of KSAs that are challenging to measure with traditional MC or TE items. All CR items are rubric-dependent and scored by a human reader.

Table 2.1.2: NJSLA–S Item Types

Item Type	Description
MC: Multiple-Choice	Select one response from four possible options (A, B, C, D).
TE: Multiple Selection	Select two or more answer options.
TE: Drop-Down Choice	Select from a drop-down menu embedded in the prompt.
TE: Ordering	Drag text or image-based options into a particular order.
TE: Drag and Drop	Place one or more text or graphic choices into blank spots within a sentence, table, or diagram.
TE: Matching in a Table	Check a box in the table to match the row to the column.
TE: Text Entry	Type a brief constrained response to the question.
TE: Bar Graph	Drag each bar to the correct length on the graph.
TE: Hot Spot	Select one or more regions on a graphic or image to identify an answer.
TE: Hot Text	Select one or more sentences within a paragraph of text.
CR: Constructed-Response	Type an extended open-ended response to the prompt.

2.2 Item Development Processes

NJSLA–S item development was conducted by MI and Pearson with oversight from NJDOE staff and the New Jersey Science Advisory Committee (NJSAC). The item development process is rigorous and involves item writers, content specialists, editors, graphic artists, programmers, scoring experts, and psychometricians. The resulting products are phenomenon-based scenarios (PBSs) and items that are aligned to the NJSLS–S and the NJSLA–S reporting categories. The PBSs and their items were all housed in Pearson's Assessment Banking for Building and Interoperability (ABBI) item banking system. ABBI is specifically designed to handle next-generation online, interactive, and accessible content. The steps in the item development process are detailed in the sections below. It warrants emphasis that between the NJSAC and the New Jersey Bias and Sensitivity Committee (NJBSC), New Jersey educators and administrators were intimately and actively involved in the item development process; each item that appears on the NJSLA–S was reviewed and approved multiple times. The principles of universal design were incorporated into the development of NJSLA–S phenomenon-based stimuli and their items. There are seven elements of assessments designed to meet the expectations of universal design (Thompson et al., 2002). The seven elements are listed below. All seven elements are incorporated into each step within the item writing process; however, there are specific steps where elements are emphasized and reviewed more extensively by experts.

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, nonbiased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

2.2.1 Item Writing

The item development process begins with the training of item writers on the specifications of NJSLA–S item development. Per the principles of universal design, item writers are trained on how to write PBSs and items that clearly communicate the task at hand for the students while also carefully maintaining alignment to the construct the NJSLA–S is intending to measure.

Once the item writers start item development, they initially identify naturally occurring phenomena that are pertinent for assessing the NJSLS–S. Next, the item writers research and develop a scenario that contains specific examples of how a phenomenon manifests itself in nature (priority is given to scenarios that are specifically relevant to New Jersey, such as native species of plants and animals, weather patterns, geological features, etc.).

Item writers then begin writing clusters of items related to the phenomenon-based scenario. Each item is aligned to a single DCI and its corresponding scientific content domain, an SEP and its corresponding scientific practice, and a CCC. To measure as many KSAs as possible with a single item cluster, item writers are instructed to vary the SEPs and CCCs within each cluster of items. An item type is typically assigned according to the item type's effectiveness and efficiency in measuring the targeted KSAs. To best align the test to the NJSLA–S blueprint, item writers are instructed that no more than 50% of the items in each cluster should be MC items.

Once a phenomenon-based scenario has a diverse cluster of four to ten items, it enters the item writing peer review process. Two different item writers review the scientific justification for the phenomenon and scenario, the alignment of the items to the NJSLS–S, the readability and appropriateness of the content, and any other conceptual understandings inherent to either the scenario or item cluster. The item writers functioning as peer reviewers iteratively rework the scenario with the original item writer until they all reach agreement.

2.2.2 Content Specialist Review

Up to three content specialists review each PBS. The first content specialist review focuses on reviewing references for the stimuli and research articles used with clusters and evaluating the science, scope, and structure of the PBS. If major revisions are needed, then the PBS is sent back to the initial item writer; if the revisions are minor, then the PBS is moved onto the second stage of the content specialist review process.

The second content specialist review focuses on universal design element 2: precisely defined constructs. The content specialist ensures the correct alignment of the PBS and all its associated items to:

- NJSLS–S
- DCI
- SEP
- CCC
- Content Domain Reporting Category
- Scientific Practices Reporting Category

If revisions are suggested, then the first content specialist and the second content specialist discuss the revisions with the item writer. If all parties agree, then the PBS is revised. If resolution is needed, then a third content specialist settles any disputes.

As a final step in the content specialist review process, the third content specialist is also charged with verifying that all the science in the PBS is accurate, that each item is answerable based on the information presented in the PBS, that all answer keys are correct, and that the alignment is in accordance with the NJSLS–S. During this step, universal design elements 5 and 6 are thoroughly reviewed to confirm that the PBS and its items have clear student instructions, that its readability is appropriate, and that it strictly adheres to the New Jersey Science Style Guidelines. Upon the final content review, the PBS is sent to editorial for its review.

2.2.3 Editorial Review

Two editors review each PBS. Their focus is on verifying that universal design elements 5, 6, and 7 are respected. The editors are charged with verifying the readability of the PBS (i.e., the PBS is easy to read and not unnecessarily complex) and checking for grammatical, spelling, and careless errors in the text. They also review each graphic or table for legibility (e.g., graphics have proper legends). Other editorial tasks include ensuring the direction lines and other components within the PBS all adhere to the NJSLA–S New Jersey Style Guidelines. Once the PBS has passed both editorial reviews, it is ready for review by the New Jersey Science Advisory Committee (NJSAC).

2.2.4 NJ Science Advisory Committee Content Review

All items on the NJSLA–S are reviewed by the New Jersey educators who compose the New Jersey Science Advisory Committee (NJSAC). In 2025, the NJSAC comprised a diverse group of New Jersey science educators representing 19 of the 21 New Jersey counties. The districts that the NJSAC members represent and the counties they come from are presented in Appendix B.

The NJSAC is the final authority on the second universal design principle: precisely defined constructs. For the 2025 administration, committee members ensured that each item was aligned to the vision set forth in the NJSLS–S, which includes properly aligning each item to a DCI, SEP, and CCC and confirming that the PBS's content was accurate. They also reviewed the PBS and its items in accordance with universal design principles 5 and 6 by confirming that the items had grade-appropriate vocabulary, that the reading level was appropriate, and that item instructions were simple and clear.

The NJSAC took an active role in editing the content of the items during their item reviews. They collectively interacted with each other, NJDOE, and the content specialists to make suggestions and offer solutions to improve the quality of item development and the NJSLA–S test. The NJSAC item reviews were held both in person at locations approved by NJDOE and in secure online platforms. The PBSs and items were all reviewed in ABBI.

2.2.5 Bias and Sensitivity Committee Review

If an item passes the NJSAC's content review, it proceeds to a review by the New Jersey Bias and Sensitivity Committee (NJBSC). This step in the item development processes is where extra emphasis is placed on universal design elements 1, 3, and 4. The NJBSC's review is designed to provide opportunities for all students to show what they know. They ensure that each item is free from bias and meets the industry guidelines for fairness and sensitivity (ETS, 2015). As described in *Standard* 3.3 (AERA, APA, NCME, 2014), this step helps guard against the introduction of construct-irrelevant language, images, or situations that might either offend or be more familiar to one group of New Jersey students than another.

Of the ten NJBSC members, seven taught special education status students, six specialized in teaching students designated as English learners, and three were bilingual. On average, they had over 12 years of teaching experience. As with NJSAC content reviews, the NJBSC reviews were conducted in person and in ABBI; the NJBSC actively worked with each other, NJDOE, and the content specialists to limit test bias. The NJBSC's district and county representation is presented in Appendix B.

2.2.6 Field Test

Once an item has passed both reviews from the NJSAC and the NJBSC, it is eligible for placement onto one of that year's field test units. The purpose of field testing is to gather data to evaluate whether an item is performing as intended. The field test items are placed into different field test units. Each grade has at least 10 field test units, and there may be as many as 18 field test units. The units are placed into the operational test form in designated positions that rotate from year to year. Each unit is reviewed by content specialists and by NJDOE to ensure that none of the field test items cue answers to the operational test items. The field test units are spiraled at the student level, which ensures that the students who take any of the field test units are a demographically representative sample of New Jersey students. A minimum of 3,500 students respond to each NJSLA–S field test item so that the samples are large enough that the resulting item statistics that are presented at the NJSLA–S Statistical Reviews are stable.

2.2.7 Statistical Review

The NJSAC reviews a battery of statistics for all field test items at the NJSLA–S statistical review. In collaboration with MI's content staff, MI's psychometric staff leads the statistical review and either trains or retrains all NJSAC members on how to interpret the item statistics so that they can make effective evaluative judgments as to the usefulness of the item. Each committee member is presented with the NJSLA–S Statistical Review Reference Sheet that provides them with quick access to definitions of the statistics and the optimal range of values. The NJSAC decides whether the item should be "Accepted," "Rejected," or "Revised and Re-Field Tested." MI's lead content specialists and an NJSAC committee member simultaneously log the decisions made by the committee, including whether an item is to be revised and how to best improve the item.

MI's staff emphasizes to the NJSAC that feedback from statistical review is used to refine future item development in an effort to constantly improve the quality of NJSLA–S stimuli and items. The NJSLA–S Statistical Review Reference Sheet given to panelists is presented in Appendix C.

2.2.8 Second Bias and Sensitivity Review

As a crucial part of statistical review, the NJBSC reviews all items flagged for potential bias against certain groups of New Jersey students. Groups of students include Male/Female, White/Black, White/Hispanic, and White/Asian. The NJBSC members are trained by MI staff prior to reviewing the items on how to interpret the statistics they will see, which include differential item functioning (DIF) statistics and the percentage of each group of students that selected each answer option. DIF is described in Section 2.3.1.1.

2.2.9 Ready for Operational Testing

Once an item has passed both statistical review and the second bias and sensitivity review, it is then eligible to be placed onto an operational test form, and its status in ABBI is updated accordingly.

2.3 Test Construction Process

The NJSLA–S test construction process ensures that the operational test forms balance the specifications set forth in the test blueprint, along with other psychometric constraints. Each form is built to measure students across the whole spectrum of ability levels and to foster valid interpretations of test scores in adherence to the standards for test design and development put forth in the *Standards* (AERA, APA, NCME, 2014). The steps and constraints associated with constructing the NJSLA–S operational tests are detailed in the following sections. An evaluation of the results of the test construction process is presented in Section 2.4.

2.3.1 Test Construction—First Draft

The first step in the NJSLA–S test construction process involves MI's content staff manually selecting approved items that best match the NJSLA–S test blueprint and statistical constraints. The process of selecting items is contingent upon the state of the item bank at each grade level. If specific content constraints are challenging to fulfill given the types of items present within the item bank, then those content constraints are given priority in the initial selection of items. Next, items are selected iteratively based on which content constraints need to be fulfilled while simultaneously balancing the various statistical constraints. Detailed descriptions of the statistical constraints are presented in Section 2.3.1.1.

2.3.1.1 Test construction statistical constraints

To ensure that the NJSLA–S operational test form is reliable and fosters valid interpretations, the following statistical constraints are used by MI's content staff during the test construction process. The primary goal is to balance the content and statistical constraints for the test as a whole; when possible, each unit is designed based on the same statistical constraints. Table 2.3.1 provides a summary of the NJSLA–S test construction constraints.

Item difficulty. Each test form is constructed to a specific difficulty level. The most important decision made from the NJSLA–S is at the Level 2/3 cut score because it is the place on the scale associated with whether students are classified as proficient. To maximize the reliability of those decisions, the average item difficulty parameter of the test form should be as close to the Level 2/3 cut score as possible.

Item discrimination. Item discrimination refers to the ability of the item to discriminate between students with different abilities. A poorly discriminating item could indicate ineffective measurement of the NJSLA–S scale and reduce test form reliability. Under classical test theory (CTT), item discrimination is measured via the item-total correlation, which can range from −1.0 to 1.0; items with item-total correlations that are below 0.2 are only selected for placement on the operational test form if no other viable options are available. Items with negative discrimination are not selected.

IRT model fit. The NJSLA–S uses an item response theory (IRT) model called the partial credit model (PCM; Masters, 1982) to estimate student ability levels. The PCM makes certain assumptions that, if violated, could impact the validity of interpretations made from NJSLA–S test scores. Statistical constraints based on PCM model-fit statistics include infit, outfit, Rasch discrimination, and lower asymptote, which are discussed in detail in Section 6.2.2 of this report. During test construction, the mean item infit, outfit, and Rasch discrimination statistics are all constrained to be as close to 1.0 as possible. If an individual item has an infit or outfit statistic outside of the acceptable range of 0.7 to 1.3 or a Rasch discrimination statistic outside of the acceptable range of 0.5 to 1.5, it is only used if no other viable options are available. The lower asymptote statistic is constrained to be as close to zero as possible; any item whose lower asymptote is greater than 0.1 is flagged. Items with lower asymptote flags are used sparingly to meet content and test specifications.

Time on items. The NJSLA–S is not designed to be a speeded test; consequently, almost all students should be able to finish it within the allotted time. For form review purposes at the test construction stage, the time on items represents the sum of the median time estimates for all items on the test form (i.e., three operational test units). The constraint on the total time is set to be less than the total operational test time minus 30 minutes. See Table 6.1.7 for the time allotted for each unit by grade.

Differential Item Functioning. Differential Item Functioning (DIF) exists when different groups of students have different probabilities of getting an item correct after controlling for their ability levels. NJSLA–S comparison groups include Male/Female, White/Black, White/Hispanic, and White/Asian. If any item favors one group over another based on the ETS Mantel-Haenszel (Dorans & Holland, 1993; Zieky, 1993) and Penfield (2007) DIF classification methods, that item is classified as demonstrating either “B” or “C” level DIF. All items classified as either “B” or “C” are reviewed by the New Jersey Bias and Sensitivity Committee during the statistical review process. If they deem an item biased, then it is ineligible for placement on the operational NJSLA–S regardless of DIF classification. A small number of “B” items can be used to maintain the test blueprint, whereas “C” items are not used on the operational NJSLA–S.

Table 2.3.1: Summary of NJSLA–S Test Construction Statistical Constraints

Statistical Constraint	Description
Item Difficulty	Average Rasch B is as close as possible to the Level 2/3 theta cut score.
Item Discrimination	Item-total correlations are greater than 0.2.
IRT Model Fit	• Item Infit and Outfit statistics range from 0.7 to 1.3 and average 1.0. • Item Discrimination statistics range from 0.5 to 1.5 and average 1.0. • Item Lower Asymptote statistics < 0.1 and average as close to 0.0 as possible.
Time on Items	Total median time on operational items < (total operational test time–30 minutes).
DIF	• "B" items are only used if necessary. • "C" items are not used.

2.3.2 Test Construction Content and Psychometric Review

After MI's content staff finishes the first draft of the operational test forms, content specialists at each grade level check the forms to ensure that no items cue each other or have content that is too similar. The content and psychometric review is an iterative process between content specialists and psychometricians. If, during the review, psychometricians identify items that better meet the statistical constraints and other psychometric properties of the NJSLA–S, the candidate items are replaced. The content and psychometric review then resumes until the test matches the content and statistical criteria of the NJSLA–S. It should be noted that certain candidate items have only undergone field testing at this stage, which means the item statistics are based on smaller field test samples. However, the psychometric analyses presented in Part 6 of this report rely on the full operational test data from the current year.

2.3.3 Test Construction NJDOE Review

All NJSLA–S test forms are reviewed and approved by NJDOE. Once content and psychometrics have agreed upon the operational test forms, they are sent to NJDOE for approval. After NJDOE approves the test forms, they are released for final editorial review and publishing.

2.4 2025 NJSLA–S Test Construction

Overall, the test construction process achieved forms that matched the balance required by the test blueprint in Section 2.1.1. All grade levels had PBA sections to address each of the three content domains. Each PBA section was comprised of three to five items, including a constructed-response item, and was each worth 5 to 8 points for each of the three content domains. Moreover, Table 2.4.1 shows that at each grade level, the science content domains were sufficiently balanced across the scientific practices reporting categories. The largest offset among content domains was in the Critiquing category of the grade 8 test, where there were 6 points aligned to Life Science content domain; 12 points aligned to the Earth Science domain; and 7 points aligned to the Physical Science domain.

Table 2.4.1: Points Available by Domain and Practice

Grade	Practice	Earth	Life	Physical
5	Critiquing	7	8	5
	Investigating	5	7	7
	Sensemaking	8	4	9
8	Critiquing	12	6	7
	Investigating	7	7	8
	Sensemaking	5	11	9
11	Critiquing	9	8	9
	Investigating	9	11	6
	Sensemaking	9	8	9

2.4.1 Grade 5 Test Construction

For grade 5, out of 60 total score points, the three content domains were well-balanced, ranging from 19 to 21 points each, as illustrated in Table 2.4.2. Each content domain had one PBA section devoted to it. The scientific practices were also well-balanced for grade 5, ranging from 19 to 21 out of 60 points being allocated to the Critiquing, Investigating, and Sensemaking reporting categories, respectively. Other content considerations that were met included: MC items only made up 12 points of the total test score (less than 50%), each unit contained a CR item, and all eight SEPs and all seven CCCs were represented on the test. All 11 of the major DCI clusters were represented by multiple points. Table 2.4.2 details the item count and point totals for each of the six reporting categories. Tables 2.4.3 through 2.4.5 show the distributions of DCIs, SEPs, and CCCs.

Table 2.4.2: 2025 NJSLA–S Grade 5 Item Counts and Point Totals by Reporting Category

Domains/Practices	Item Counts				Points
	MC	TE	CR	Total	
Earth and Space	4	13	1	18	20
Life	5	10	1	16	19
Physical	3	15	1	19	21
Total–Domains	12	38	3	53	60
Critiquing	3	10	2	15	20
Investigating	5	14	0	19	19
Sensemaking	4	14	1	19	21
Total–Practices	12	38	3	53	60

Table 2.4.3: 2025 NJSLA–S Grade 5 DCIs

DCI	Item Counts	Points
ESS1	3	3
ESS2	8	8
ESS3	7	9
LS1	2	2
LS2	7	10
LS3	4	4
LS4	3	3
PS1	3	3
PS2	6	6
PS3	3	3
PS4	7	9

Table 2.4.4: 2025 NJSLA–S Grade 5 SEPs

SEP	Item Counts	Points
AID	5	5
AQDP	9	9
CEDS	6	6
DUM	8	10
EAE	10	13
OECI	5	7
PACI	3	3
UMCT	7	7

Table 2.4.5: 2025 NJSLA–S Grade 5 CCCs

CCC	Item Counts	Points
C & E	15	18
E & M	3	5
Patterns	19	19
S & SM	7	9
S, P, & Q	5	5
SC	1	1
SF	3	3

The statistical constraints for the 2025 grade 5 NJSLA–S operational test form were met. Two items had item-total correlations below the .20 threshold, indicating that these were low-discriminating items (see Section 6.1). The two items had observed item-total correlation values of .11 and .18. However, as shown in Table 2.4.6, the mean item-total correlation was higher than the mean target for the form (.35). Additionally, each of the model-fit statistics averaged close to their target values. The median test time of 87.158 minutes was well below the 105-minute threshold, and out of 53 DIF classifications for each of the four group comparisons (i.e., Male/Female, White/Black, White/Hispanic, and White/Asian), there were zero values categorized as “C” and only one value categorized as “B” for all the DIF analyses performed. All “B” DIF items were approved for operational test use by the NJBSC as described in Section 2.3.1.1. Tables 2.4.6 and 2.4.7 summarize the test construction and DIF statistics.

Table 2.4.6: 2025 NJSLA–S Grade 5 Test Construction Statistics

Statistic	Average	Target	Flags
Rasch B	0.884	0.904	N/A
Item-Total Correlation	0.409	> 0.35	2
Infit	1.011	1.00	1
Outfit	1.052	1.00	8
PCM Discrimination	0.967	1.00	3
Lower Asymptote	0.033	0.00	4
Median Time (min)	87.158	< 105	N/A

Table 2.4.7: 2025 NJSLA–S Grade 5 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	52	1	0
White/Black	53	0	0
White/Hispanic	53	0	0
White/Asian	53	0	0

2.4.2 Grade 8 Test Construction

At grade 8, the content domains were well-balanced. The three content domains each accounted for 24 of the total 72 score points, as illustrated in Table 2.4.8. Each content domain was represented in a PBA section. The scientific practices were less balanced than content domains, with Investigating 22 points, Critiquing 25 points, and Sensemaking 25 points. Other content considerations that were met included the following: MC items only accounted for 16 points (less than 50%) of the total test score; each unit contained a CR item, and all eight SEPs and all seven CCCs were represented by multiple points on the test. Similarly, all 11 major DCI clusters were represented by multiple points on the test. Table 2.4.8 details the item count and point totals for each of the six reporting categories; Tables 2.4.9 through 2.4.11 show the distributions of DCIs, SEPs, and CCCs for grade 8.

Table 2.4.8: 2025 NJSLA–S Grade 8 Item Counts and Point Totals by Reporting Category

Domains/Practices	Item Counts MC	TE	CR	Total	Points
Earth and Space	5	15	1	21	24
Life	7	13	1	21	24
Physical	4	16	1	21	24
Total–Domains	16	44	3	63	72
Critiquing	3	14	2	19	25
Investigating	11	11	0	22	22
Sensemaking	2	19	1	22	25
Total–Practices	16	44	3	63	72

Table 2.4.9: 2025 NJSLA–S Grade 8 DCIs

DCI	Item Counts	Points
ESS1	7	7
ESS2	8	11
ESS3	6	6
LS1	3	3
LS2	10	10
LS3	1	1
LS4	7	10
PS1	4	4
PS2	2	2
PS3	8	11
PS4	7	7

Table 2.4.10: 2025 NJSLA–S Grade 8 SEPs

SEP	Item Count	Points
AID	6	6
AQDP	14	14
CEDS	8	11
DUM	8	8
EAE	13	13
OECI	6	12
PACI	6	6
UMCT	2	2

Table 2.4.11: 2025 NJSLA–S Grade 8 CCCs

CCC	Item Counts	Points
C & E	23	26
E & M	6	6
Patterns	10	10
S & SM	11	14
S, P & Q	5	5
SC	3	6
SF	5	5

The statistical constraints for the 2025 grade 8 NJSLA–S operational test form were met. Two grade 8 items were flagged for having item-total correlations below the .20 threshold, indicating a low discriminating item (see Section 6.1). However, both items had observed item-total correlation values above .15, and as shown in Table 2.4.12, the average item-total correlation was close to the mean target value for the form (.35). The infit, outfit, and PCM discrimination model-fit statistics each averaged close to their ideal values of 1.00. The median test time of 78.250 minutes was well below the 105-minute threshold, and out of 63 DIF classifications for each of the four group comparisons (i.e., Male/Female, White/Black, White/Hispanic, and White/Asian), there were zero values categorized as "C" and one value categorized as "B." All "B" DIF items were approved for operational test use by the NJBSC as described in Section 2.3.1.1. Tables 2.4.12 and 2.4.13 summarize the test construction and DIF statistics.

Table 2.4.12: 2025 NJSLA–S Grade 8 Test Construction Statistics

Statistic	Average	Target	Flags
Rasch B	0.342	0.416	N/A
Item-Total Correlation	0.380	> 0.35	2
Infit	1.019	1.00	0
Outfit	1.055	1.00	3
PCM Discrimination	0.956	1.00	3
Lower Asymptote	0.027	0.00	4
Median Time (min)	78.250	< 105	N/A

Table 2.4.13: 2025 NJSLA–S Grade 8 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	62	1	0
White/Black	63	0	0
White/Hispanic	63	0	0
White/Asian	63	0	0

2.4.3 Grade 11 Test Construction

The grade 11 content domains were well-balanced. Out of 78 total score points, the three content domains ranged from 24 to 27 points each. Each content domain had one PBA section.

The scientific practices were also perfectly balanced, with each representing 26 points out of the total test score points of 78. Other content considerations that were met included the following: MC items only made up 26 points (less than 50%) of the total test score; each unit contained a CR item, and all eight SEPs and all eleven DCIs were represented by multiple points on the test. Each of the seven CCCs was adequately represented. Table 2.4.14 details the item count and point totals for each of the six reporting categories; Tables 2.4.15 through 2.4.17 show the distributions of DCIs, SEPs, and CCCs for grade 11.

Table 2.4.14: 2025 NJSLA–S Grade 11 Item Counts and Point Totals by Reporting Category

	Item Counts				
Domains/Practices	**MC**	**TE**	**CR**	**Total**	**Points**
Earth and Space	9	15	1	25	27
Life	10	14	1	25	27
Physical	7	14	1	22	24
Total–Domains	26	43	3	72	78
Critiquing	3	16	2	22	26
Investigating	13	10	1	24	26
Sensemaking	10	17	0	26	26
Total–Practices	26	43	3	72	78

Table 2.4.15: 2025 NJSLA–S Grade 11 DCIs

DCI	Item Counts	Points
ESS1	10	10
ESS2	6	8
ESS3	9	9
LS1	2	2
LS2	11	13
LS3	4	4
LS4	8	8
PS1	5	5
PS2	10	12
PS3	3	3
PS4	4	4

Table 2.4.16: 2025 NJSLA–S Grade 11 SEPs

SEP	Item Counts	Points
AID	7	7
AQDP	10	10
CEDS	10	10
DUM	10	10
EAE	11	11
OECI	10	14
PACI	7	9
UMCT	7	7

Table 2.4.17: 2025 NJSLA–S Grade 11 CCCs

CCC	Item Counts	Points
C & E	24	26
E & M	4	4
Patterns	8	10
S & SM	12	12
S, P & Q	10	10
SC	9	9
SF	5	7

The 2025 grade 11 NJSLA–S operational test form construction included three items flagged for outfit. Nevertheless, the flagged items had outfit values near the target thresholds. Additionally, the values of outfit on the flagged items indicated that their inclusion would not distort or degrade the measures (Engelhard & Wang, 2021; Linacre, 2002; 2016), and the average values of infit and outfit were close to their ideal values of 1.00. As shown in Table 2.4.18, the average item-total correlation was close to the mean target for the form (.35). The mean and median test time was 82.967 minutes, which was well below the 150-minute target. Of 72 DIF classifications for each of the four group comparisons (i.e., Male/Female, White/Black, White/Hispanic, and White/Asian), there were zero values categorized as “C” and five values categorized as “B.” All “B” DIF items were approved for operational test use by the NJBSC as described in Section 2.3.1.1. Tables 2.4.18 and 2.4.19 summarize the test construction and DIF statistics for grade 11.

Table 2.4.18: 2025 NJSLA–S Grade 11 Test Construction Statistics

Statistic	Average	Target	Flags
Rasch B	0.447	0.475	N/A
Item-Total Correlation	0.401	> 0.35	0
Infit	1.011	1.00	0
Outfit	1.037	1.00	3
PCM Discrimination	0.972	1.00	0
Lower Asymptote	0.025	0.00	0
Median Time (min)	82.967	< 150	N/A

Table 2.4.19: 2025 NJSLA–S Grade 11 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	67	5	0
White/Black	72	0	0
White/Hispanic	72	0	0
White/Asian	72	0	0

2.5 2025 NJSLA–S State of the Item Bank

Upon the completion of the 2025 test construction process, MI's psychometricians analyzed the item bank and facilitated a discussion of the results with content specialists and NJDOE staff. The goal of the discussion was to guide future item development so that it could support valid test score interpretations. The item bank analysis looked at how many items were developed, how many survived the field test and statistical review processes, and how many items were available for creating the 2025 NJSLA–S. Item counts were disaggregated by item type, content domain, scientific practice, DCI, SEP, and CCC. Content areas where the bank had been severely depleted were discussed to determine why they had been problematic and how the next round of item development could improve upon the results.

PART 3: TEST ADMINISTRATION

Standard 6.1 (AERA, NCME, APA, 2014) requires that "[t]est administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer" (p. 114). The test developer is responsible for providing "appropriate training, documentation, and oversight so that the individuals who administer or score the test(s) are proficient in the appropriate test administration or scoring procedures and understand the importance of adhering to the directions provided by the test developer" (p. 114). The following sections detail the myriad processes, procedures, and trainings that were undertaken to properly administer the NJSLA–S.

3.1 District Test Coordinator Training

District Test Coordinators (DTCs) were trained in proper test administration procedures during the annual NJSLA District Test Coordinator Training. In turn, they were "responsible for ensuring that all district and school personnel involved in the administration of New Jersey state assessment programs have been trained" (see Figure 3.1.1; NJDOE, 2025). Information about the NJSLA–S administration is in the *Test Coordinator Manual* (*TCM*). That information is not fully replicated here, but the following elements are specific topics that the DTCs were trained on and are also of importance to this Technical Report:

- Scheduling and testing site requirements
- NJSLA–S participation requirements
- Accessibility features and accommodations available for use on the NJSLA–S
- Materials and tools that would be shipped to schools prior to administration
- Student registration and placement procedures
- Protocols for securely handling materials
- Post-testing responsibilities
- Links and contact information related to the NJSLA–S

The NJSLA *TCM* can be read in full on the NJSLA–S website under Educator Resources/Test Content & Other Information/2025 Test Administrator and Coordinator Manuals.

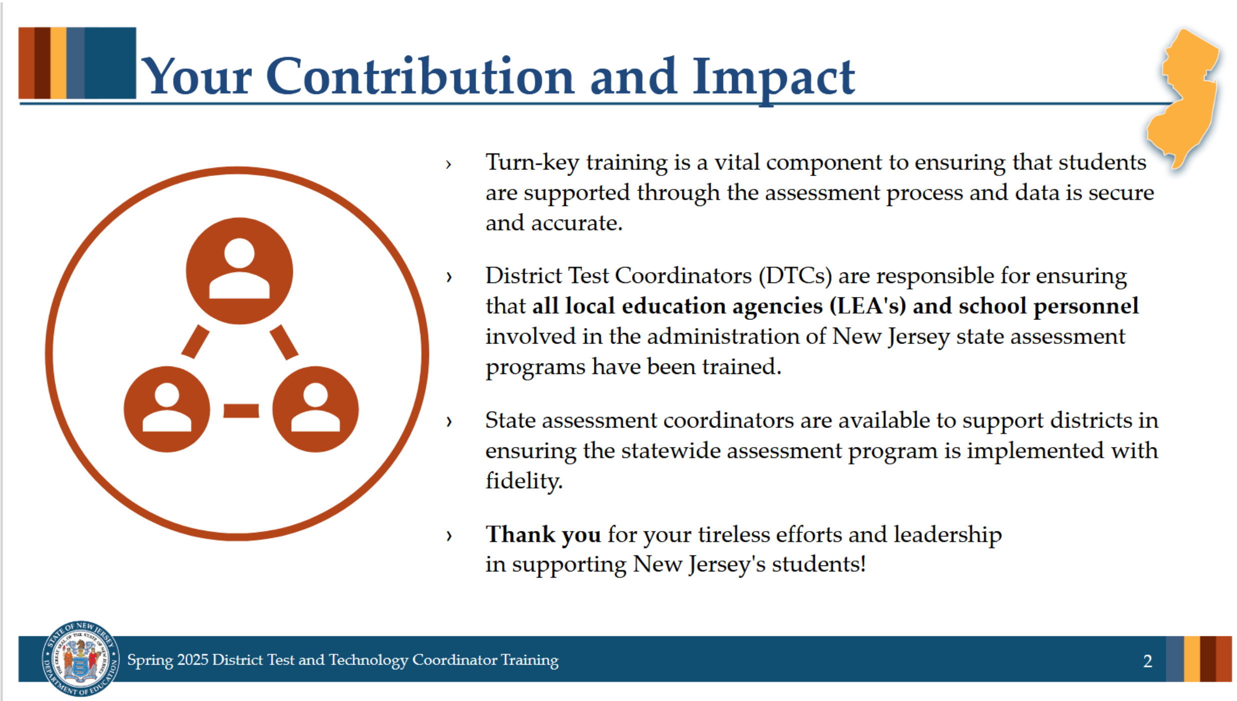


Figure 3.1.1. Slide two from the 2025 DTC

Table 3.1.1 shows the NJSLA–S 2025 testing window dates as well as testing time. Testing times do not include the extra time needed for administrative tasks such as logging students into their testing sessions or reading them directions.

Table 3.1.1: NJSLA–S 2025 Grades 5, 8, and 11 Testing Window

Grade	CBT	PBT	Testing Time
5	4/28/25–5/23/25	4/28/25–5/23/25	45 minutes per unit
8	4/28/25–5/23/25	4/28/25–5/23/25	45 minutes per unit
11	4/28/25–5/23/25	4/28/25–5/23/25	60 minutes per unit

3.2 Test Security and Administration Procedures

This section provides information regarding the NJSLA–S test administration procedures. Descriptions of both the computer-based test (CBT) and paper-based test (PBT) procedures are detailed below. For a complete description of all test administration activities, refer to the NJSLA *TCM*.

3.2.1 Computer-Based Testing

The NJSLA–S CBT forms are delivered via Pearson's test delivery system, TestNav. TestNav is a secure browser that restricts students' actions so that they are unable to access or interact with other applications that are outside of the online test materials. Likewise, the student login process is secure; for every test session, Test Administrators (TAs) provide students with testing tickets that include their unique login and password information. If a student needs to exit the test prior to its completion, the TAs can, to ensure test security, lock a test section for the student to access when they return.

Each School Test Coordinator (STC) is provided with a checklist of tasks that they are required to complete during CBT (see Table 3.2.1). The STCs and TA use PearsonAccess^{next} (PAN) to manage each test session; they can monitor the progress of each of their students and lock and unlock units. PAN is a next-generation web-based platform that allows end-to-end monitoring of test administrations for the TAs. Students are only assigned one unit at a time in a prescribed order. STCs and TAs are also charged with assisting with technical issues if they arise. The *TCM* provides them with a list of typical CBT issues and gives procedures for addressing them. The District Test Coordinator (DTC) and STC are strongly advised to monitor testing and ensure security procedures. Furthermore, they must ensure that TAs provide students with the correct accommodations and accessibility features. After the completion of each unit, STCs collect test materials from the TAs, which include scratch paper, accommodated test materials, and paper copies of the periodic table. Finally, at the end of each day, all NJSLA–S materials must be returned to a secure storage area. Table 3.2.1 shows the checklist of CBT-related tasks that the STCs are charged with completing. For a complete discussion of these procedures, please refer to the *TCM*.

Table 3.2.1: CBT School Test Coordinator Checklist

Tasks	*TCM* Section(s)
Ensure that TAs have a computer or device that meets the PAN system requirements available.	Section 3.4.3
Distribute test materials to TAs.	Section 3.6.3
Manage test sessions in PAN.	Section 4.1.1
Monitor each testing room to ensure that test administration and security protocols are followed, and that required administration information is being documented and collected. Be available during testing to answer questions from TAs.	Section 4.1.3
Investigate all testing irregularities and security breaches, and follow New Jersey policy for reporting these incidents.	Section 2.2
Ensure that TAs provide applicable students with their approved testing accommodations and pre-identified accessibility features.	Section 6.2
Schedule and supervise make-up testing.	Sections 2.4.2 and 4.1.4
Create make-up test sessions in PAN.	Section 4.1.4
Respond to all technology-related issues.	Section 4.1.1
Collect materials from TAs.	Section 4.1.4
Ensure that all units are locked after testing on each testing day.	Section 4.1.4

3.2.2 Paper-Based Testing

The following section describes the responsibilities of the DTC and STC during PBT administration. As in the CBT administration, the DTC and STC are required to complete a checklist of tasks (see Table 3.2.2). The tasks are similar to those on the CBT checklist, except that they are specific to the PBT administration. For instance, the PBT checklist requires STCs to follow protocols for damaged test materials such as test booklets or answer documents. For a complete discussion of these procedures, please refer to the *TCM*.

Table 3.2.2: PBT School Test Coordinator Checklist

Tasks	*TCM* Section(s)
Distribute test materials to TAs.	Section 3.6.3
Monitor each testing room to ensure that test administration and security protocols are followed, and that required administration information is being documented and collected. Be available during testing to answer questions from TAs.	Section 4.1.3
Investigate all testing irregularities and security breaches, and follow New Jersey policy for reporting these incidents.	Section 2.2
Ensure that TAs provide applicable students with their approved testing accommodations and pre-identified accessibility features.	Section 3.7
Schedule and supervise make-up testing.	Sections 2.4.2 and 4.1.4
Follow the protocol for contaminated or damaged test materials, and refer to New Jersey policy for reporting these incidents.	Section 2.1.5
Collect materials from TAs, and ensure that all test booklets and answer documents have a student name or student ID label.	Section 4.1.4

3.3 Test Irregularities and Breaches

If test security is compromised, the validity of the inferences made from test scores can be affected. Thus, any action that compromises test security is prohibited. These actions are classified as testing irregularities or security breaches. A more complete discussion of test irregularities and breaches can be found in the NJSLA *TCM*.

Examples of test irregularities and breaches include, but are not limited to:

- Test Administration Irregularities
  - Student reviewing or working on the wrong unit of the test; if the student completes the wrong unit of a test, the DTC must immediately contact the appropriate State Assessment Program Coordinator for directions.
- Electronic Devices Irregularities
  - Using a cell phone or other prohibited electronic device (e.g., smartphone, iPod®, smartwatch, personal scanner, eReader) while secure test materials are still distributed, while students are testing, after a student turns in his or her test materials, or during a break.
    - Exception: Test Coordinators, Technology Coordinators, Test Administrators, and proctors are permitted to use cell phones in the testing environment only in cases of emergencies or when timely administration assistance is needed. Districts may set additional restrictions on allowable devices as needed.

        - Exception: Certain electronic devices may be allowed for medical or audiological purposes during testing. For specific information, refer to the *NJSLA & NJGPA Accessibility Features and Accommodations Manual* at the New Jersey Assessments Resource Center under Educator Resources > Test Administration Resources > Accessibility Features and Accommodations Resources > Manuals > NJSLA & NJGPA Accessibility Features and Accommodations, 13th Edition.[1]
- Test Supervision Irregularities
    - Coaching students during testing, including giving students verbal or nonverbal cues, hints, suggestions, or paraphrasing or defining any part of the test
    - Engaging in activities (e.g., grading papers, reading a book, newspaper, or magazine) that prevent proper student supervision at all times while secure test materials are still distributed or while students are testing
    - Leaving students unattended without a Test Administrator for any period of time while secure test materials are still distributed or while students are testing. (Proctors must be supervised by a Test Administrator at all times.)
    - Deviating from testing time procedures
    - Allowing cheating of any kind
    - Providing unauthorized persons with access to secure materials
    - Unlocking a test in PAN during non-testing times without NJDOE approval
    - Failing to provide a student with a documented accommodation or providing a student with an accommodation that is not documented and therefore is not appropriate
    - Allowing students to test before or after the test administration window without NJDOE approval
- Test Materials Irregularities and Breaches
    - Losing a student testing ticket
    - Losing a student test booklet or answer document
    - Losing tactile graphics booklets
    - Leaving test materials unattended or failing to keep test materials secure at all times
    - Reading or viewing tests before, during, or after testing
        - Exception: Administration of a Human Reader/Signer accessibility feature or accommodation which requires a Test Administrator to access the tests
    - Copying or reproducing (e.g., taking a picture of) any part of the test or any secure test materials or online test forms
    - Revealing or discussing test items with anyone, including students and school staff, through verbal exchange, email, social media, or any other form of communication
    - Removing secure test materials from the school building or removing them from locked storage for any purpose other than administering the test

[1] https://nj.portal.cambiumast.com/index.html is the current location for such material.

- Testing Environment Irregularities
    - Failing to follow administration directions exactly as specified in the *Test Administrator Manual* (*TAM*) (An electronic version of the manual can be viewed at the NJ Assessments Resource Center on the NJSLA–S website under Educator Resources/Test Content & Other Information/2025 Test Administrator and Coordinator Manuals.)
    - Displaying any resource (e.g., poster, model, display, teaching aid) that defines, explains, or illustrates terminology or concepts, or otherwise provides unauthorized assistance during testing
    - Allowing preventable disruptions such as talking, making noises, or excessive student movement around the classroom
    - Allowing unauthorized visitors in the testing environment
        - Unauthorized Visitors: Visitors, including parents/guardians, school board members, reporters, and school staff not authorized to serve as Test Administrators or proctors, are prohibited from entering the testing environment.
        - Authorized Visitors: Observation visits by the principal, monitors from the NJDOE Office of Assessment, monitors from the district, and NJDOE-authorized observers are allowed as long as these individuals do not disturb the testing process.

Protocols are established to report and document any testing irregularity or security breach. All Test Administrators are trained to ensure the proper protocols are implemented. First, both the School and District Test Coordinators must be immediately notified. The DTC is then charged with immediately contacting their NJSLS–S State Contact. The DTC may require the STC to complete the New Jersey Testing Irregularity or Security Breach Form available at the New Jersey Assessments Resource Center under Educator Resources > Test Administration Resources > Forms > NJSLA/NJGPA Testing Irregularity and Security Breach Form to properly document the event.[2] Finally, more information or investigation may be requested by either the DTC or the NJSLS–S State Contact.

3.4 Test Accessibility Features and Accommodations

Standard 3.9 states that "[t]est developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs" (p. 67). Federal and state regulations require that all students—including those classified as English learners (EL) and those with disabilities—be included in the statewide assessment program and assessed annually. The Every Student Succeeds Act of 2015 (ESSA) mandates that all states must test science one time each in three different grade bands: 3–5, 6–8, and 9–12. The *NJSLA Test Coordinator Manual* states:

[2] https://nj.portal.cambiumast.com/index.html is the current location for such material.

> Students who are full-time home-schooled or full-time at a private or parochial school are not eligible to take any statewide assessment. Students with disabilities who attend an approved private school for the disabled and whose tuition is not the financial responsibility of the district are also not eligible to take any statewide assessment. (p. 8)

To ensure that the diverse population of students taking the NJSLA–S is tested under appropriate conditions and to adhere to the principles of universal design (Thompson et al., 2002), NJDOE has adopted test accommodations and accessibility features that may be used when testing special populations of students. The content of the test remains the same, but administration procedures, setting, and answer modes may be adapted. Students requiring accommodations may be tested in a separate location from general education students.

The *NJSLA and NJGPA Accessibility Features and Accommodations Manual (AF&A Manual)* is available online at the New Jersey Assessments Resource Center under Educator Resources > Test Administration Resources > Accessibility Features and Accommodations (AF&A) Resources > Manuals > NJSLA & NJGPA Accessibility Features and Accommodations, 13th Edition.[3] It contains detailed information about each accessibility feature and accommodation. Schools must refer to the *AF&A Manual* for full information about identifying and administering accessibility features and accommodations.

3.4.1 Accessibility Features

The purpose of accessibility features is to ensure that a diverse population of students is being tested fairly and that construct-irrelevant factors are not unduly impacting their test scores. According to the NJSLA and NJGPA *AF&A Manual* (2025), accessibility features are defined as "tools or preferences that are either built into the testing platform or provided externally by TAs" (p. 1, p. 4). All students have access to accessibility features. However, for some accessibility features to be available for students during testing, an administrator must have identified the student as needing the accessibility feature prior to testing. It is essential that students using accessibility features get to practice with them prior to operational testing. Thus, NJSLA–S practice tests that contain the accessibility features are available throughout the year at the NJSLA–S website.

3.4.1.1 Text-to-Speech

The most used NJSLA–S accessibility feature is Text-to-Speech (TTS). Prior to testing, an administrator activates the TTS accessibility feature for individual students. When the selected student gets placed into a testing session, their form automatically defaults to the designated TTS form. During testing, the student can select the TTS player, and the test will be read aloud to them via the TTS software embedded within TestNav. Students using the TTS accessibility feature must be wearing headphones. The items on the TTS form all contain the same phenomenon-based scenarios, item stems, and response options as are presented to the students taking the traditional CBT form. All final TTS forms are verified by NJDOE to ensure that the TTS functionality is working correctly.

[3] https://nj.portal.cambiumast.com/index.html is the current location for such material.

3.4.2 Accommodations

The role of accommodations is to minimize the impact of a student's disabilities or English language proficiency level on his or her assessment performance. The NJSLA and NJGPA *AF&A Manual* (2025) defines an accommodation as "adjustments to the testing conditions, test format, or test administration that provide equitable access during assessments for students with disabilities and students who are multilingual learners (MLs)" (p. 14). Accommodations are only available to students who have an Individualized Education Program (IEP), a Section 504 plan, or an English learner (EL) plan.

Different accommodations are necessary depending on whether the test was administered using a CBT or PBT format. Per NJDOE policy, all students who received PBT versions of the NJSLA–S had appropriate accommodations. A comprehensive explanation of each NJSLA–S accommodation is presented in the NJSLA and NJGPA *AF&A Manual*. The NJSLA–S CBT accommodations include:

- Assistive Technology–screen reader
- Assistive Technology–non-screen reader
- American Sign Language (ASL)
- Text-to-Speech (TTS)
- Human Reader
- Spanish
- Spanish TTS
- Spanish Human Reader

PBT accommodations are received as kits, and they include:

- Braille
- Large Print
- Read-Aloud
- Spanish
- Spanish Large Print
- Spanish Read-Aloud
- Tactile Graphics

3.4.2.1 Accommodated test form development

The *Standards* (AERA, APA, NCME, 2014) state that "an appropriate accommodation is one that responds to specific individual characteristics but does so in a way that does not change the construct the test is measuring or the meaning of the scores" (p. 67). Each of the accommodated test forms requires specific processes to ensure they are addressing the needs of their intended users. After NJDOE approval, the accommodated test forms were sent to various subcontractors so that they could adapt the items to Spanish, braille, and American Sign Language (ASL). The adaptation processes for those forms are presented in Sections 3.4.2.1.1 through 3.4.2.1.3. The paper-based test (PBT) form adaptation process is presented in Section 3.4.2.1.4. Following adaptation, NJDOE verifies each accommodated test form.

*3.4.2.1.1 Spanish*The Teneo Linguistics Company (TLC) made all Spanish accommodations. TLC received the NJDOE-approved tests and created the translations within ABBI. Once the items were translated, a committee of New Jersey Spanish teachers reviewed the items online, with TLC representatives in attendance. Edits were made during the review, and then the final versions of the online forms were verified by NJDOE. The translation that was created for the online version was then used to create the paper version of the Spanish tests.

*3.4.2.1.2 Braille*All braille accommodations were created by the National Braille Press (NBP). NBP received the downloaded paper versions of the operational test forms. NBP provided MI with feedback about any items that were unable to be brailled. Once the tests were brailled, external reviewers received the draft braille versions and reviewed for any issues a student might have taking the braille tests. For the 2025 NJSLA–S, all items were able to be brailled.

*3.4.2.1.3 American Sign Language*All ASL accommodations were created by the ADS Group in Plymouth, MN. They provided ASL video production with two ASL content specialist translators and one ASL proofer. Their video production engineer provided studio editing. Additionally, they provided proofing/QC services as well as closed captioning. Once NJDOE approved the operational test forms, the ADS group created the videos in American Sign Language for each item. These items were verified by external expert reviewers under the guidance of MI.

*3.4.2.1.4 Paper-Based Test*The conversion of the NJSLA–S CBT into PBT form was undertaken by MI's Editorial Department. Most PBT items were the same as their CBT counterparts. However, some aspects needed adaptation. The following bullets represent the major changes that took place with the stimuli and items during the adaptation processes:

- All artwork was converted from color to grayscale.
- Video items were converted to still images. This was accomplished by MI's editorial staff working in conjunction with content specialists to select specific frames from the video that effectively conveyed its essence. In some cases, the captured images were redrawn to ensure that no essential information was lost in the adaptation process.
- TE items were converted to PBT format via multiple methods depending on the TE item type.

3.4.2.2 Accommodated test form equivalenceOccasionally, during the accommodated test form conversion process, an item is deemed unable to be accommodated. This can occur for a multitude of reasons—some items do not translate well from English to Spanish, while others are challenging to braille, for example. The procedures for calculating the separate scale score tables, if needed, are detailed in Part 7: Equating and Scaling. In 2025, all items were deemed adequately accommodated by external reviewers, content specialists, and NJDOE.

PART 4: SCORING

It is the responsibility of the test developer to establish scoring procedures (AERA, APA, NCME, 2014). Standard 6.8 states that "[a] scoring protocol should be established, which may be as simple as an answer key for multiple-choice questions" (p. 118). For constructed-response items, the procedures outlined by the *Standards* require that test developers provide "scoring training materials, scoring rubrics, and examples of test takers' responses at each score level" (p. 118). The procedures for both the machine-scoring and handscoring of NJSLA–S student responses are described in the following sections.

4.1 Machine-Scored Items

All multiple-choice (MC) and technology-enhanced (TE) items are machine-scored. Each item has a key (correct answer) associated with it, which has been supplied and verified by content specialists and approved by NJDOE prior to test administration. All student responses are machine-scored based on these prior approved keys. Prior to the administration, Pearson's Customer Data Quality (CDQ) team creates multiple sets of mock test responses for each test form. These responses are scored and processed just as the real tests will be during the administration. The CDQ team verifies that the student responses were accurately captured from the test and that they were scored accurately. Verification steps include comparing responses to the possible ranges of responses to the item, comparing raw overall scores and subscores for entire tests to the maximum values, validating ID unique item numbers (UINs) against the test map, and flagging inconsistent student records for investigation. After the administration, the same checks are made on the data files containing real student tests before they are transferred to MI for psychometric analysis and the adjudication process.

4.1.1 Adjudication

Adjudication involves the careful review of all student responses to an item, ensuring that its key was applied correctly and that no possible correct answer has been overlooked in the many prior key checks. All machine-scored items are adjudicated by MI's psychometric department. During adjudication, the psychometric team analyzes the student response patterns for each item. The response patterns are simple for items with limited possible options; for instance, an MC item only has five possible student responses (A, B, C, D, or blank). However, some TE items can have hundreds of different student responses. The student response data are used to produce one file for each operational item. This file contains each unique response option, the point-value associated with it (i.e., 0, 1, or 2), the total number and the percentage of students selecting each response, and the item-total correlation associated with each response option that was selected more than 100 times. The item means and item-total correlations are also calculated at the item level, and items are flagged for aberrant behavior across all these metrics. Details of the flagging criteria are presented in Part 6 of this document. Upon completion, the files are securely transferred to each grade level's lead content specialists for review.

The role of the content specialists during the adjudication process is to use the information housed in the adjudication files to identify any possible miskeys. They are instructed to first check items that were flagged for having low item means and item-total correlations because those statistics could indicate that the item is not performing as intended. Next, they look at combinations of student responses that are keyed as receiving 0 points but have item-total correlations above 0. That combination of response-level data could also be an indication of a possible student response that deserves credit for a correct response, but that has been keyed as incorrect. Finally, through a sorting process, the content specialists can relatively quickly review all other combinations of student responses. If there are any miskeys, key changes are submitted to NJDOE, and upon approval, Pearson incorporates them into the scoring algorithm. These steps are essential to ensure both the reliability of student test scores and their valid interpretations.

4.2 Handscored Items

All NJSLA–S CR items are scored by human scorers according to the procedures outlined in the sections that follow.

4.2.1 Selecting Handscoring Staff

MI's recruiting team first recruits qualified scorers who have experience scoring New Jersey Science assessments. To supplement this core pool, MI's recruiting team contacts other scorers in MI's database who have experience in successfully scoring other large-scale assessments. Returning staff are selected based on experience and performance, as well as attendance, punctuality, and cooperation with work procedures and MI policies. MI maintains evaluations and performance data for all staff who work on each scoring project in order to determine employment eligibility for future projects. For new scorers, the recruiting team reviews applications—including prospective scorers' resumes, references, proof of degree, and recognition of scorer requirements—before offering employment. All MI scorers have a minimum of a four-year college degree, and many are current or former educators.

In selecting Team Leaders, MI management staff and Scoring Directors review the files of all returning staff. They look for people who are experienced Team Leaders with a record of good performance on the NJSLA–S and consider scorers who have been recommended for promotion to the Team Leader position.

MI requires that all handscoring staff (Scoring Directors, Team Leaders, scorers, and clerical staff) sign a confidentiality/nondisclosure agreement before receiving training or accessing secure project materials. The employment agreement indicates that participants may not reveal information about the test, the scoring criteria, or the scoring methods to any person.

4.2.2 Operational Range Finding

Range-finding meetings are conducted to establish "true" scores from a representative sample of papers (i.e., responses). One hundred sample papers per task are chosen from the available field test papers. At the beginning of the range-finding meeting, the scoring rubrics of the items are discussed and refined by the committee. The sample responses brought to the range-finding meetings are selected from a broad range of New Jersey local education agencies (LEAs) in order to ensure that the sample is representative of overall student performance. To maximize the probability that papers eligible for the highest score points are included in the sample, special efforts are made by MI management and scoring staff to include high-performing responses. The range-finding committees consist of NJDOE content specialists, New Jersey teacher representatives, and MI management personnel, as well as the Scoring Director responsible for each content area when possible. The final decision on the "true" scores is made by the New Jersey teacher representatives in consultation with the other committee members.

4.2.3 Field Test Range Finding

Prior to field test scoring, content committees consisting of NJDOE personnel, New Jersey teacher representatives, and MI leadership personnel meet virtually to determine "true" scores for 30 selected papers representing each of the score points for each item to be tested. Field test scoring guides and training sets are developed using the papers scored at the range finding. Time is spent determining whether any changes need to be made to the scoring rubrics associated with the items being reviewed before any field test scoring takes place. As with operational range finding, the final decision on the "true" scores is made by the New Jersey teacher representatives with input from the other committee members.

4.2.4 Developing Scoring Guides

After the range-finding meetings, training materials are developed consisting of an anchor set (examples of responses for each score point) and training/qualifying sets (practice papers) for each task using the responses scored at range-finding. Anchor sets usually consist of two or more annotated examples of each score point, arranged in score point order. To maximize consistency, the same anchor sets are used each year for items administered in multiple administrations. Anchor sets include annotations that explain how the scoring criteria are applied to each response's specific features and why the response merits a particular score. These annotations connect to highlighted sections of the student response in training lessons, drawing scorers' attention to the critical training pieces to elucidate the precise scoring rationale and to help scorers define the lines between score points. Training/qualifying sets consist of clearly anchored papers in random score point order. These sets are constructed using responses from the operational range finding, with the scores assigned by the range-finding committee for each response.

4.2.5 Team Leader Training and Duties

After the anchor, training, and qualifying papers have been identified and finalized, the Scoring Director conducts Team Leader training for each task. This process typically takes up to four days depending on the content. Procedures are similar to those for training scorers (described in more detail below) but are more comprehensive, dealing with identification of nonscorable responses, unusual approaches to a prompt, alert situation responses (e.g., child-in-danger), and other duties performed only by leadership. Team Leaders assist in training scorers by serving as a resource when scorers are training.

During scoring, Team Leaders respond to questions, read behind scorers' scored responses, and counsel scorers who are having difficulty with the criteria. Team Leaders also monitor the scoring patterns of each scorer throughout the project, conduct retraining as necessary through responses to scorer questions and reading behind scorers, perform second readings, and maintain a professional working environment.

4.2.6 Scorer Training and Qualifying

All scorers are trained using the rubrics, anchor papers, training papers, and qualifying papers selected during the range-finding meetings and approved by NJDOE. MI's Virtual Scoring Center™ (VSC™) includes an online training interface that presents rubrics, anchor sets, and training/qualifying sets. VSC™ is used for all training and qualifying, whether site-based or remote. VSC™ provides for effortless and timely communication with scoring leadership throughout training and allows scorers to efficiently navigate the training materials.

Recruited staff must maintain rigorous adherence to established training methodologies to ensure the quality and credibility of the scoring. MI enforces strict attendance during training. Scorers are trained as a group to maintain consistency and are trained on all relevant training materials. Scorers have access to all training materials during live scoring, which include annotated versions of the training set indicating the successful parts of each response. The same training protocol is followed for both site-based and remote scorers.

After scorers have signed contracts and nondisclosure forms and have been provided with an introduction to the project, training begins. Scorer training and Team Leader training follow the same format. Scorers and Team Leaders are introduced to the constructed-response task and the anchor set. This process includes modeling how to identify the essential information in anchor responses to establish a consistent scoring vocabulary. Any nuances in interpreting and applying the scoring rubric are also highlighted at this stage.

Scoring personnel log in to Measurement Incorporated Remote Access (MIRA) to review the rubric and anchor responses. MIRA, which includes all online training modules, is the portal to the VSC™ interface and is the data repository of all scoring reports that are used for scorer monitoring. Here, Team Leaders and scorers assign scores to a practice/qualifying set of responses. They are reminded to compare each practice response to comparable anchor responses to ensure accuracy and consistency in scoring the practice responses. MI trains scoring personnel to reference those student responses as representative of the rubric. The rubric is a tool, but the anchor responses represent how the rubric is applied. After Team Leaders and scorers score practice responses, they are provided with the correct scores. The same process is followed for all subsequent practice/qualifying sets.

Scorers must demonstrate their ability to score accurately by attaining 70% perfect agreement and 100% adjacent agreement (within one point) percentage on two of the qualifying sets before they read packets of operational student responses. Any scorer unable to meet the standards set by NJDOE is dismissed.

Training is carefully orchestrated so that scorers understand how to apply the rubric in scoring the papers, learn how to reference the scoring guide, develop the flexibility needed to deal with a variety of responses, and retain the consistency needed to score all papers accurately. In addition to completing all of the initial training and qualifying, scorers are trained in the use of the VSC™ handscoring system, “flagging” of unusual responses for Team Leader review, and other procedures necessary for the conduct of a smooth project.

Levels of staffing for scoring the 2025 NJSLA–S are presented in Table 4.2.1. Specifically, Table 4.2.1 shows the number of scorers, Team Leaders, and Scoring Directors at each grade level who participated in scoring.

Table 4.2.1: Scoring Personnel by Grade

Grade	Scorers	Team Leaders	Scoring Director
5	54	3	1
8	257	3	1
11	88	3	1

4.2.7 Monitoring Scorer Performance

In addition to thorough and consistent training, reliable scoring depends upon careful evaluation of scorer performance to support a continuous loop of feedback among the scorers, Team Leaders, Scoring Directors, and Scoring Management. Scoring Directors offer direct leadership and guidance to Team Leaders as they monitor individual scorer performance. Scoring Directors also furnish scorers with general guidance and clarify appropriate application of the training materials, while Team Leaders provide direct supervision, which allows for a higher degree of scrutiny of scorer performance, individual attention, and opportunities for immediate intervention or correction if required.

Real-time reports that provide both daily and cumulative (project-to-date) data are used to monitor and evaluate scoring performance. Scoring Monitors and Scoring Directors review these reports daily. As they review these data, they can identify any issues evident in scores being generated and address them with Team Leaders and individual scorers when necessary. These reports are described in more detail below.

The quality of MI’s handscoring program is maintained through ongoing monitoring by experienced scoring leadership. Scoring Directors and Team Leaders are skilled in detecting scoring trends and remediating any issues that arise. Scorers who are unable to meet accuracy and productivity standards after feedback and retraining will not be allowed to continue scoring. When this occurs, MI can reset any scores assigned by a dismissed scorer and have the responses immediately rescored.

MI’s handscoring process incorporates ongoing checks for and controls against scorer error. Specifically, MI implements the following quality-assurance procedures:

- Validity checks. MI’s VSC™ scoring system randomly seeds validity responses among operational responses during scoring. A small set of validity responses are selected and approved by Scoring Monitors and Scoring Directors. The “true” scores for these responses are entered into a validity database. Validity responses are indistinguishable from operational responses. Scorer accuracy and drift are evaluated using validity results. The validity responses are dispersed evenly across all of an item's score point levels, and they are selected based on how well they represent typical examples of each score point. Members of the scoring team who score the items, also known as readers or "raters," are encouraged to send responses that are difficult to score to their Team Leader; thus, those types of papers are not selected as validity responses.

- Blind double reads. For each item, a minimum of 10% of responses are randomly selected to receive blind double reads. Scorer agreement is used to evaluate the reliability of scoring across all scorers.
- Daily systematic review of handscoring reports. Scoring Directors monitor and evaluate scorers' performance daily using an array of handscoring reports, described below. MI provides any retraining necessary to ensure scorer accuracy. Retraining strategies are implemented under the direction of the Scoring Monitors in conjunction with Scoring Directors and Team Leaders.
- Targeted read-behinds. Team Leaders conduct targeted read-behinds for scorers who have been identified, based on validity performance, or based on other performance data, as targets for close monitoring. When conducting targeted read-behinds, Team Leaders pay careful attention to the particular score points with which individual scorers have difficulty. This information is obtained by reviewing the results of validity and score point distribution reports. Team Leaders provide feedback by discussing incorrectly scored responses with the individual scorer and continue to monitor to ensure the scorer has understood and applied the feedback appropriately.
- Score verifications. MI implements a series of automated score verifications to ensure the accuracy of scores. For example, when a scorer assigns a condition code of "blank" to a response, a blank check is then conducted, which resets the score when the condition code of "blank" is assigned to a response that has one or more characters in the response string (e.g., a response comprising spaces or tabs). In this case, only after three independent scorers have assigned a condition code of "blank" to a response that appears blank but includes characters in the response string is the score recorded. A similar check is run when a score or condition code other than "blank" is assigned to a response that includes no characters in the response string. Automatic resetting of double-scored responses occurs when two scorers assign nonadjacent scores, mismatched condition codes, or a combination of a condition code and a numeric score, thus providing an additional score verification. In addition to automatically resetting and rescoring these responses, the scorer information (for instance, if a number of resets from one reader ID shows a pattern of scoring high or low over a certain period of time) is captured in a report and reviewed by Scoring Directors, as one of many tools used to determine retraining needs.

VSC™ provides an appropriate infrastructure for facilitating extensive quality-assurance procedures. Through VSC™, handscoring leadership can review scorer performance, conduct read-behinds, provide feedback and respond to questions, deliver retraining and/or recalibration responses on demand and at regularly scheduled intervals, and prevent scorers from scoring additional live responses if they require additional monitoring.

Scorers are dismissed when, in the opinion of the appropriate Scoring Monitor and/or Scoring Director, they have been counseled, retrained, and given a reasonable opportunity to improve and continue to perform below an acceptable standard for accuracy or production. In the case of the former, all scores assigned by a scorer during a given timeframe can be identified and reset, and the responses can be released back into the scoring pool for immediate rescoring.

4.2.8 Automatic Rescores

As shown in Section 8.5, the raters are not in perfect agreement 100% of the time. Thus, to ensure that no student is unjustly penalized because a rater may have been a little too stringent, rescoring is conducted automatically for any student who scores one raw score point below the proficient cut score. MI Scoring Directors and/or Team Leaders review student responses to constructed-response items and verify the original scores or make changes where warranted. A score is never lowered during the automatic rescoring process even if it was deemed to be too high. LEAs do not need to request rescoring. Table 4.2.2 provides automatic rescoring results of the number of responses for all three grade levels. All open-ended/constructed-response item types were scored by a single rater. The first column represents the number of responses eligible for auto rescore; in other words, it represents the number of responses from students who were at the overall cutoff point. The second column is the number of responses with a changed score.

Table 4.2.2: Automatic Rescore Results

Grade	Number of Responses Eligible for Automatic Rescore	Number of Response Score Changes	Percentage Changed (of Those Eligible)
5	4,315	274	6.35
8	3,869	423	10.93
11	3,166	248	7.83

4.3 Quality Control

To confirm that the processing of student tests and test scores was done correctly, MI conducted quality control checks in addition to supporting NJDOE doing its own quality control checks. To produce the score reports, Pearson started with a large data file called the Student Record File (SRF) that includes all the information that will be shown in the reports. MI began by verifying the information in this file.

MI checked the student demographic information against what the districts had entered into the test registration system. MI also verified the results of the machine-scored items. This was straightforward for multiple-choice items but was more complicated for technology-enhanced items that could have multiple correct answers. For open-ended items, MI compared the scores to its handscoring system.

After Pearson produced the final score reports, MI used the previously verified SRF to verify the data shown on a large sample of these reports.

4.3.1 QC Sample

For NJDOE's part of QC, MI selected a sample of several hundred student tests for them to manually review. NJDOE staff compared the scores for these students to the final SRF and score reports produced by Pearson. The sample included all test forms, as well as students with a wide variety of values for demographic variables such as gender, ethnicity/race, English learner status, and disability status. These selected students were provided specifically for validation of ISRs and Rosters. After individual test scores were verified, NJDOE used them to calculate aggregate figures, such as average scale scores, which are shown in data files and score reports. The following section details the processes NJDOE used to ensure that student test scores were accurate.

4.3.2 Key Information Sheets

To help organize NJDOE's QC process for the student-level data, MI produced a Key Information Sheet (KIS) for each student's test. The KIS is a spreadsheet that is used to keep track of all the information from a test, as a helpful aid for the QC process. The KIS is pre-populated with student information from a test (such as name and accommodations), the key for each machine-scored item, and a spot to record the points earned for each item.

First, MI verified the student information on the KIS against the student information in the test registration system. Next, MI scored the student's responses to each selected-response item against the key and recorded the score on the KIS. For open-ended items, MI exported scores from the handscoring system to record on the spreadsheet. Formulas in the KIS automatically tallied the student's overall points, as well as the points in each domain and practice. Any discrepancies between these totals and the preliminary data file from Pearson required scrutiny of the points earned for each item. The KIS helped to narrow down the problem to a particular domain, practice, and unit. MI provided NJDOE with the verified KIS on July 7, 2025.

After external review and NJDOE's approval of the scale score tables, Pearson imported the scale score tables and produced a final set of data files and score reports. NJDOE staff used the KISs prepared by MI to determine each student's scale score, overall proficiency level, and proficiency level for each subscore. Then NJDOE compared this information to student-level score reports. This stage provided NJDOE with confidence that each piece of student-level information on these reports was accurately derived from the original sources of test data.

4.3.3 Aggregate Data

Certain numbers shown on the Individual Student Report and School Student Roster are aggregated figures, such as averages of scores at the school, district, or state level; and the percentages of students achieving each overall proficiency level. In addition, other score reports and the summary data file only show aggregated data. NJDOE verified all of these values by calculating the same figures from the raw data in the SRF. This step was not necessarily limited to the schools in the QC sample.

PART 5: STANDARD SETTING

Cizek and Bunch (2007) define standard setting as "the process of establishing one or more cut scores on examinations" (p. 5). Cut scores divide a distribution of test scores into two or more categories. The purpose of conducting a standard setting is to assist the users of test scores in making valid interpretations. *Standard* 5.21 states that "[w]hen proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly" (p. 107). The *2019 NJSLA–S Technical Report* details the processes, procedures, and analyses used to accomplish the 2019 NJSLA–S Standard Setting. The executive summary from the 2019 NJSLA–S Standard-Setting Report is presented in Appendix D of this report.

The 2019 NJSLA–S Standard Setting was externally reviewed by NJTAC member Stephen Koffler (Koffler, 2019). He evaluated the process based on the *Standards* (2014) and the framework established by Kane (2001). Koffler focused on three major sources of validity evidence: procedural, internal, and external. Overall, he concluded that "the NJSLA–S Standard-Setting Study was sound, followed best practice and met the professional standards for performing a Standard-Setting Study and recommending valid and defensible cut scores" (p. iv).

PART 6: ITEM AND TEST STATISTICS

Standard 5.0 states that "[t]est scores should be derived in a way that supports the interpretations of test scores for proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed uses" (p. 102). The NJSLA–S was designed to support inferences based on the classification of students into four proficiency levels, as has been described throughout this Technical Report. The interpretations of the proficiency-level classifications are dependent upon the test performing as intended. As was described in Section 2.3, the NJSLA–S was constructed using a combination of classical test theory (CTT) statistics, item response theory (IRT) statistics, and the content constraints. The following sections detail how well the 2025 NJSLA–S performed based on those CTT and IRT statistics, along with other criteria. Detailed test maps containing item metadata, various statistics, and Range PLD alignment are presented in Appendix F of this report. The final section in this part presents disaggregated descriptive statistics of scale scores and subscore proficiency classifications.

The data for these and all subsequent analyses were verified by Pearson's Customer Data Quality (CDQ) team prior to delivery to MI. Responses from students who did not attempt to take the test or who had their test scores voided were removed from the data prior to analyses. NJDOE requires a student to attempt at least one item in at least two different operational test units to obtain a scale score. Student responses were voided for cheating, security breaches, or other reasons.

6.1 Classical Test Theory Statistics

For each administration, a set of statistics based on CTT was generated and reviewed for item calibrations and scaling. The statistics can be grouped into measures of four psychometric concepts:

- Item Difficulty
- Item Discrimination
- Speededness
- Differential Item Functioning

These statistics were calculated for every operational item; each statistic provides some key information about the quality of each item from an empirical perspective. If any of the four statistics suggested that an item was negatively impacting the reliability or validity of test score interpretations, a recommendation was made to NJDOE to remove the item from operational use. Descriptions of each type of item statistic appear in the following sections.

6.1.1 Item Difficulty and Discrimination Descriptive Statistics

Monitoring item difficulty is essential for ensuring that the test is reliable and will foster valid test score interpretations. If items are too challenging or too easy for a population of test takers, then the reliability and validity of test score interpretations will suffer. In CTT, the difficulty of a dichotomous item is assessed via the *p*-value, which is defined as the proportion of students who answered an item correctly. *P*-values can range from 0.00 to 1.00; an item with a high *p*-value is easier to answer correctly, whereas an item with a low *p*-value is more challenging. Dichotomous items with *p*-values either below .25 or above .90 were flagged for review during the adjudication process described in Section 4.1.1. For polytomous items (i.e., CR items), item difficulty is expressed as an adjusted mean, which is the proportion of points earned relative to the maximum possible points for that item. The adjusted mean is analogous to a *p*-value. Polytomous items are then flagged if their adjusted mean falls outside of the .25 to .90 range. It should be noted that the flagging criteria only provide a general guideline, and some items that have *p*-values outside of the .25 to .90 range can still be useful operationally.

Item discrimination is also important to monitor. If items are unable to discriminate between students with different ability levels, then both the reliability and the validity of test score interpretations can suffer. In CTT, the item discrimination is expressed as the correlation between item scores and the total score of the remaining items on the test, the latter being a proxy for overall student ability. The item-total correlation (denoted by *rpb* in this Technical Report) can range from –1.00 to 1.00. Items with discrimination values below 0.2 are flagged for review during the adjudication process. Items with item-total correlations that are below zero (i.e., negative) are considered for removal from the test because they could harm both the reliability and the validity of test score interpretations.

Tables 6.1.1, 6.1.3, and 6.1.5 summarize the difficulties of NJSLA–S items in terms of *p--values* for grades 5, 8, and 11, respectively; Tables 6.1.2, 6.1.4, and 6.1.6 show the item discrimination, measured by the item-total correlation (*rpb*), summaries for grades 5, 8, and 11, respectively. Several intervals of item difficulties and item discriminations were created for computing the frequency distributions. In these tables, the descriptive statistics and frequency distributions for each item type (i.e., TE, MC, or CR) are disaggregated by content domain and scientific practice.

The overall patterns in item difficulty and discrimination suggest that the NJSLA–S assessments are effectively measuring student performance in New Jersey. Across all grade levels, TE items were generally more challenging and more discriminating than MC items. CR items, as anticipated, demonstrated the highest levels of discrimination among the three item types.

For each grade, the CR item in Physical Science tended to be more challenging than those in Earth and Space Science or Life Science. At grade 5, MC items showed item difficulties ranging from .25 to .75, reflecting a moderate level of challenge. TE items at this grade were notably more difficult, with 11 (29%) items falling below a *p*-value of .25 and 26 items (68%) between .25 and .50. Grade 8 followed a similar pattern, with most MC items again falling within the .25 to .75 range. However, 19 (44%) of grade 8 TE items were below .25, indicating a substantial portion of more challenging items, while the remainder fell within the average difficulty range. At grade 11, most MC items also had *p*-values between .25 and .75, and 9 (21%) of TE items were below .25, reinforcing the presence of more challenging TE items at this grade level.

Item-total correlations, denoted by *rpb*, for CR items were consistently strong across grades, with average values of .61, .63, and .63 for grades 5, 8, and 11, respectively, indicating strong discrimination. The distributions of item-total correlations for MC and TE items further support their effectiveness in distinguishing between high- and low-performing students. At grade 5, only one TE item had a correlation below .20, while five MC and 20 TE items exceeded .40. In grade 8, five MC and two TE items fell below .20, whereas four MC and 19 TE items surpassed .40. Grade 11 showed similar trends, with only two MC and one TE item below .20 and eight MC and 31 TE items above .40. These results underscore the productive role of TE and CR items in enhancing the discriminatory power of the assessments.

Table 6.1.1: Grade 5 Item Difficulty (*p*-value) Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Difficulty (*p*-value)					Descriptive Statistics		
			[0, .25]	[.25, .5]	[.5, .75]	[.75, .9]	[.9, 1]	Average	S.D.	Median
MC	NJSLA–S	12	0	9	3	0	0	.48	.11	.45
	Earth and Space	4	0	2	2	0	0	.51	.13	.51
	Life	5	0	4	1	0	0	.47	.14	.45
	Physical	3	0	3	0	0	0	.43	.04	.43
	Critiquing	3	0	3	0	0	0	.47	.03	.47
	Investigating	5	0	4	1	0	0	.45	.05	.45
	Sensemaking	4	0	2	2	0	0	.51	.20	.51
TE	NJSLA–S	38	11	26	1	0	0	.33	.11	.33
	Earth and Space	13	5	7	1	0	0	.34	.14	.37
	Life	10	2	8	0	0	0	.33	.11	.30
	Physical	15	4	11	0	0	0	.32	.09	.30
	Critiquing	10	4	6	0	0	0	.31	.11	.27
	Investigating	14	5	9	0	0	0	.32	.10	.33
	Sensemaking	14	2	11	1	0	0	.36	.12	.35
CR	NJSLA–S	3	0	2	1	0	0	.43	.16	.40
	Earth and Space	1	0	0	1	0	0	.59	N/A	.59
	Life	1	0	1	0	0	0	.40	N/A	.40
	Physical	1	0	1	0	0	0	.28	N/A	.28
	Critiquing	2	0	1	1	0	0	.50	.13	.50
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	1	0	1	0	0	0	.28	N/A	.28

Table 6.1.2: Grade 5 Item Discrimination Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Discrimination (*rpb*) [0, .2]	[.2, .3]	[.3, .4]	[.4, .5]	[.5, 1]	Descriptive Statistics Average	S.D.	Median
MC	NJSLA–S	12	0	1	6	5	0	.37	.07	.36
	Earth and Space	4	0	1	2	1	0	.36	.10	.36
	Life	5	0	0	3	2	0	.37	.05	.35
	Physical	3	0	0	1	2	0	.39	.07	.41
	Critiquing	3	0	0	2	1	0	.38	.09	.35
	Investigating	5	0	1	0	4	0	.39	.09	.42
	Sensemaking	4	0	0	4	0	0	.35	.03	.36
TE	NJSLA–S	38	1	7	10	11	9	.41	.11	.42
	Earth and Space	13	0	3	4	4	2	.40	.11	.38
	Life	10	0	2	2	3	3	.42	.13	.45
	Physical	15	1	2	4	4	4	.41	.12	.45
	Critiquing	10	1	5	2	0	2	.32	.12	.27
	Investigating	14	0	1	3	4	6	.46	.11	.45
	Sensemaking	14	0	1	5	7	1	.42	.07	.44
CR	NJSLA–S	3	0	0	0	0	3	.61	.11	.57
	Earth and Space	1	0	0	0	0	1	.53	N/A	.53
	Life	1	0	0	0	0	1	.74	N/A	.74
	Physical	1	0	0	0	0	1	.57	N/A	.57
	Critiquing	2	0	0	0	0	2	.64	.15	.64
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	1	0	0	0	0	1	.57	N/A	.57

Table 6.1.3: Grade 8 Item Difficulty (*p*-value) Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Difficulty (*p*-value)					Descriptive Statistics		
			[0, .25]	[.25, .5]	[.5, .75]	[.75, .9]	[.9, 1]	Average	S.D.	Median
MC	NJSLA–S	16	2	10	4	0	0	.40	.13	.40
	Earth and Space	5	0	4	1	0	0	.43	.09	.41
	Life	7	2	4	1	0	0	.35	.14	.37
	Physical	4	0	2	2	0	0	.46	.14	.43
	Critiquing	3	0	2	1	0	0	.39	.10	.35
	Investigating	11	2	7	2	0	0	.41	.13	.41
	Sensemaking	2	0	1	1	0	0	.41	.21	.41
TE	NJSLA–S	44	19	21	4	0	0	.31	.12	.27
	Earth and Space	15	8	6	1	0	0	.29	.12	.24
	Life	13	7	6	0	0	0	.29	.10	.25
	Physical	16	4	9	3	0	0	.35	.14	.33
	Critiquing	14	8	3	3	0	0	.33	.15	.24
	Investigating	11	4	7	0	0	0	.28	.07	.26
	Sensemaking	19	7	11	1	0	0	.32	.12	.28
CR	NJSLA–S	3	0	3	0	0	0	.37	.08	.38
	Earth and Space	1	0	1	0	0	0	.38	N/A	.38
	Life	1	0	1	0	0	0	.44	N/A	.44
	Physical	1	0	1	0	0	0	.28	N/A	.28
	Critiquing	2	0	2	0	0	0	.41	.04	.41
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	1	0	1	0	0	0	.28	N/A	.28

Table 6.1.4: Grade 8 Item Discrimination Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Discrimination (*rpb*) [0, .2]	[.2, .3]	[.3, .4]	[.4, .5]	[.5, 1]	Descriptive Statistics Average	S.D.	Median
MC	NJSLA–S	16	5	5	2	2	2	.30	.13	.27
	Earth and Space	5	1	1	1	1	1	.34	.12	.32
	Life	7	3	2	0	1	1	.28	.16	.23
	Physical	4	1	2	1	0	0	.29	.07	.29
	Critiquing	3	1	2	0	0	0	.25	.05	.26
	Investigating	11	3	3	1	2	2	.32	.14	.28
	Sensemaking	2	1	0	1	0	0	.25	.10	.25
TE	NJSLA–S	44	2	11	12	15	4	.37	.10	.38
	Earth and Space	15	0	6	2	6	1	.37	.10	.38
	Life	13	1	3	5	2	2	.36	.10	.35
	Physical	16	1	2	5	7	1	.39	.10	.40
	Critiquing	14	0	4	4	6	0	.36	.08	.36
	Investigating	11	1	2	4	4	0	.36	.09	.39
	Sensemaking	19	1	5	4	5	4	.39	.12	.38
CR	NJSLA–S	3	0	0	0	0	3	.63	.03	.63
	Earth and Space	1	0	0	0	0	1	.59	N/A	.59
	Life	1	0	0	0	0	1	.66	N/A	.66
	Physical	1	0	0	0	0	1	.63	N/A	.63
	Critiquing	2	0	0	0	0	2	.62	.05	.62
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	1	0	0	0	0	1	.63	N/A	.63

Table 6.1.5: Grade 11 Item Difficulty (*p*-value) Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Difficulty (*p*-value)					Descriptive Statistics		
			[0, .25]	[.25, .5]	[.5, .75]	[.75, .9]	[.9, 1]	Average	S.D.	Median
MC	NJSLA–S	26	3	16	7	0	0	.42	.13	.41
	Earth and Space	9	0	6	3	0	0	.47	.12	.45
	Life	10	2	6	2	0	0	.39	.14	.38
	Physical	7	1	4	2	0	0	.40	.12	.41
	Critiquing	3	2	1	0	0	0	.30	.13	.23
	Investigating	13	1	8	4	0	0	.43	.14	.44
	Sensemaking	10	0	7	3	0	0	.45	.12	.40
TE	NJSLA–S	43	9	28	6	0	0	.35	.12	.34
	Earth and Space	15	3	10	2	0	0	.34	.10	.34
	Life	14	2	9	3	0	0	.37	.14	.36
	Physical	14	4	9	1	0	0	.34	.11	.34
	Critiquing	16	3	10	3	0	0	.38	.13	.35
	Investigating	10	1	8	1	0	0	.36	.10	.37
	Sensemaking	17	5	10	2	0	0	.33	.12	.32
CR	NJSLA–S	3	0	1	2	0	0	.47	.16	.51
	Earth and Space	1	0	0	1	0	0	.60	N/A	.60
	Life	1	0	0	1	0	0	.51	N/A	.51
	Physical	1	0	1	0	0	0	.29	N/A	.29
	Critiquing	2	0	1	1	0	0	.44	.22	.44
	Investigating	1	0	0	1	0	0	.51	N/A	.51
	Sensemaking	0	0	0	0	0	0	N/A	N/A	N/A

Table 6.1.6: Grade 11 Item Discrimination Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Discrimination (*rpb*) [0, .2]	[.2, .3]	[.3, .4]	[.4, .5]	[.5, 1]	Descriptive Statistics Average	S.D.	Median
MC	NJSLA–S	26	2	9	7	8	0	.33	.10	.32
	Earth and Space	9	0	4	3	2	0	.33	.08	.35
	Life	10	2	4	1	3	0	.30	.12	.27
	Physical	7	0	1	3	3	0	.37	.07	.38
	Critiquing	3	0	1	1	1	0	.33	.09	.33
	Investigating	13	1	3	5	4	0	.35	.10	.38
	Sensemaking	10	1	5	1	3	0	.32	.10	.28
TE	NJSLA–S	43	1	4	7	21	10	.43	.10	.43
	Earth and Space	15	0	2	4	6	3	.41	.09	.42
	Life	14	0	0	1	9	4	.47	.07	.44
	Physical	14	1	2	2	6	3	.41	.11	.45
	Critiquing	16	0	3	2	7	4	.43	.10	.43
	Investigating	10	0	0	2	6	2	.44	.06	.43
	Sensemaking	17	1	1	3	8	4	.42	.11	.42
CR	NJSLA–S	3	0	0	0	0	3	.63	.02	.63
	Earth and Space	1	0	0	0	0	1	.63	N/A	.63
	Life	1	0	0	0	0	1	.65	N/A	.65
	Physical	1	0	0	0	0	1	.61	N/A	.61
	Critiquing	2	0	0	0	0	2	.62	.02	.62
	Investigating	1	0	0	0	0	1	.65	N/A	.65
	Sensemaking	0	0	0	0	0	0	N/A	N/A	N/A

6.1.2 Speededness

The consequence(s) of time limits on examinees' scores is called speededness (Swineford, 1949). A traditional measure of speededness is the number of items that are not attempted by students. Logically, in each separately timed subsection of a test, it can be assumed that a student may have run out of time if the student did not attempt the last item. The percentage of students omitting an item provides information about speededness, although it must be kept in mind that students can omit an item for reasons other than speededness (for example, choosing not to put effort into answering a constructed-response item). Thus, if the percentage of omits is low, that implies that there is little speededness. Conversely, if the percentage of omits is high, speededness, as well as other factors, may be the cause.

The NJSLA–S was not designed to be a speeded test, but rather a power test. That is, all students are expected to have ample time to finish all items and prompts. NJSLA–S assessments were administered during a testing window with a specified amount of time per unit by grade. Students were assumed to have enough time to complete the test. The numbers of items and item types composing each operational test unit for each grade level, along with the testing time, are detailed in Table 6.1.7. Additionally, Table 6.1.8 presents the percentage of students omitting the last TE item in each test section. Overall, the small percentages of students shown in the table indicated that each grade-level test did not show speededness.

Table 6.1.7: Operational Testing Schedule—Items and Time Allocations

Grade	Unit	Items	Time in Minutes
5	1	3 MC, 14 TE, 1 CR	45
	2	5 MC, 11 TE, 1 CR	45
	3	4 MC, 13 TE, 1 CR	45
8	1	5 MC, 15 TE, 1 CR	45
	2	3 MC, 17 TE, 1 CR	45
	3	8 MC, 12 TE, 1 CR	45
11	1	9 MC, 14 TE, 1 CR	60
	2	7 MC, 16 TE, 1 CR	60
	3	10 MC, 13 TE, 1 CR	60

Table 6.1.8: Percentage of Students Omitting the Last TE Item in Each Operational Unit

Grade	Unit	Location	% Student
5	1	17	8.49
	2	15	3.43
	3	16	2.13
8	1	21	3.86
	2	20	1.53
	3	20	1.29
11	1	23	1.36
	2	23	0.62
	3	22	1.80

6.1.3 Operational DIF Analysis

The *Standards* define Differential Item Functioning (DIF) as "when different groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item" (p. 16). If items perform differently for subgroups of students after controlling ability, the test might disadvantage some groups of students over others.

By convention, the two groups of test takers involved in DIF analyses are referred to as the focal and reference groups. Different methods are used for DIF detection depending on whether the item is dichotomous or polytomous. For dichotomous items, DIF was identified using the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959). It is considered effective and efficient (Clauser & Mazor, 1998; Hills, 1989). For the NJSLA–S, under the MH procedure, a statistical significance test (MH Chi-square test) of DIF and an evaluation of effect sizes in DIF measures (MH D-DIF statistic) were performed in conjunction with the ETS DIF classification system (Dorans & Holland, 1993). The letters A, B, and C are used to denote DIF categories in the ETS DIF classification system with A-level indicating a negligible degree of DIF, B-level indicating slight to moderate DIF, and C-level indicating large DIF. Items classified as C-level DIF require a careful review for possible biases. For polytomous items, DIF was identified using the Liu-Agresti (LA) procedure (Liu & Agresti, 1996; Penfield & Algina, 2003, 2006). The LA estimator of the cumulative common odds ratio for DIF detection is a generalization of the MH procedure. This allows the ETS DIF categorization system to be applied to DIF studies of polytomous items. Table 6.1.9 exhibits the DIF evaluation criteria for dichotomous and polytomous items. The effect size in DIF measures under the MH procedure is denoted by MH D-DIF; that under the LA procedure is denoted by Log(LA).

Table 6.1.9: Differential Item Functioning Evaluation Criteria

DIF Category	Dichotomous Items	Polytomous Items
A (Negligible)	Nonsignificant MH Chi-square test ($p \geq .05$) or $\lvert$MH D-DIF$\rvert < 1.0$	Nonsignificant LA Chi-square test ($p \geq .05$) or $\lvert$ Log(LA) $\rvert < 0.43$
B (Slight to moderate)	Significant MH Chi-square test ($p < .05$) and $1.0 \leq \lvert$MH D-DIF$\rvert < 1.5$	Significant LA Chi-square test ($p < .05$) and $0.43 \leq \lvert$ Log(LA) $\rvert < 0.63$
C (Moderate to high)	Significant MH Chi-square test ($p < .05$) and $\lvert$MH D-DIF$\rvert \geq 1.5$	Significant LA Chi-square test ($p < .05$) and $\lvert$ Log(LA) $\rvert \geq 0.63$

*Log indicates the logarithm function.

NJSLA–S DIF analyses for the field test items only focused on four major comparisons of students: Male/Female, White/Black, White/Hispanic, and White/Asian. For the operational assessment, four other comparisons were made: non-multilingual learner (ML–No)/ multilingual learner (ML–Yes), students with disabilities (SWD–Yes)/ students without disabilities (SWD–No), Not economically disadvantaged (EconDis–No)/economically disadvantaged (EconDis–Yes), and TTS/CBT test takers due to the large numbers of students taking the TTS forms. The traditional CBT test takers were the reference group, whereas the TTS test takers were the focal group.

Tables 6.1.10 through 6.1.12 present the DIF classifications for all eight comparison groups across grades 5, 8, and 11 in the operational assessments. Overall, the DIF analysis yielded encouraging results, with only a small number of items flagged as “C” in the ML–No versus ML–Yes comparison. Specifically, this included one MC item and one TE item at grade 11, as well as two CR items at grade 11. No “C” classifications were observed in any other comparison group across the three grade levels.

In addition, the presence of “B” classifications was minimal across all grades, item types, and comparison groups, further supporting the stability of the assessment design. Notably, every item received an “A” classification in the CBT versus TTS DIF analysis for grades 5, 8, and 11, indicating no evidence of differential functioning in that context. The few flagged items will be revisited once additional test data become available, ensuring continued scrutiny and refinement of the assessment instruments.

Table 6.1.10: Grade 5 DIF Classification by Item Type

Grade	Group	Item Type	A	B	C
5	Male/Female	MC	12	0	0
	Male/Female	TE	37	1	0
	Male/Female	CR	3	0	0
	Male/Female	Total	52	1	0
5	White/Black	MC	12	0	0
	White/Black	TE	38	0	0
	White/Black	CR	3	0	0
	White/Black	Total	53	0	0
5	White/Hispanic	MC	12	0	0
	White/Hispanic	TE	38	0	0
	White/Hispanic	CR	3	0	0
	White/Hispanic	Total	53	0	0
5	White/Asian	MC	12	0	0
	White/Asian	TE	38	0	0
	White/Asian	CR	3	0	0
	White/Asian	Total	53	0	0
5	ML–No/ML–Yes	MC	12	0	0
	ML–No/ML–Yes	TE	37	1	0
	ML–No/ML–Yes	CR	3	0	0
	ML–No/ML–Yes	Total	52	1	0
5	SWD–No/SWD–Yes	MC	12	0	0
	SWD–No/SWD–Yes	TE	38	0	0
	SWD–No/SWD–Yes	CR	3	0	0
	SWD–No/SWD–Yes	Total	53	0	0
5	EconDis–No/EconDis–Yes	MC	12	0	0
	EconDis–No/EconDis–Yes	TE	38	0	0
	EconDis–No/EconDis–Yes	CR	3	0	0
	EconDis–No/EconDis–Yes	Total	53	0	0
5	CBT/TTS	MC	12	0	0
	CBT/TTS	TE	38	0	0
	CBT/TTS	CR	3	0	0
	CBT/TTS	Total	53	0	0

Table 6.1.11: Grade 8 DIF Classification by Item Type

Grade	Group	Item Type	A	B	C
8	Male/Female	MC	16	0	0
	Male/Female	TE	43	1	0
	Male/Female	CR	3	0	0
	Male/Female	Total	62	1	0
8	White/Black	MC	16	0	0
	White/Black	TE	44	0	0
	White/Black	CR	3	0	0
	White/Black	Total	63	0	0
8	White/Hispanic	MC	16	0	0
	White/Hispanic	TE	44	0	0
	White/Hispanic	CR	3	0	0
	White/Hispanic	Total	63	0	0
8	White/Asian	MC	16	0	0
	White/Asian	TE	44	0	0
	White/Asian	CR	3	0	0
	White/Asian	Total	63	0	0
8	ML–No/ML–Yes	MC	16	0	0
	ML–No/ML–Yes	TE	43	1	0
	ML–No/ML–Yes	CR	2	1	0
	ML–No/ML–Yes	Total	61	2	0
8	SWD–No/SWD–Yes	MC	16	0	0
	SWD–No/SWD–Yes	TE	44	0	0
	SWD–No/SWD–Yes	CR	3	0	0
	SWD–No/SWD–Yes	Total	63	0	0
8	EconDis–No/EconDis–Yes	MC	16	0	0
	EconDis–No/EconDis–Yes	TE	44	0	0
	EconDis–No/EconDis–Yes	CR	3	0	0
	EconDis–No/EconDis–Yes	Total	63	0	0
8	CBT/TTS	MC	16	0	0
	CBT/TTS	TE	44	0	0
	CBT/TTS	CR	3	0	0
	CBT/TTS	Total	63	0	0

Table 6.1.12: Grade 11 DIF Classification by Item Type

Grade	Group	Item Type	A	B	C
11	Male/Female	MC	25	1	0
	Male/Female	TE	41	2	0
	Male/Female	CR	2	1	0
	Male/Female	Total	68	4	0
11	White/Black	MC	26	0	0
	White/Black	TE	43	0	0
	White/Black	CR	3	0	0
	White/Black	Total	72	0	0
11	White/Hispanic	MC	26	0	0
	White/Hispanic	TE	43	0	0
	White/Hispanic	CR	3	0	0
	White/Hispanic	Total	72	0	0
11	White/Asian	MC	25	1	0
	White/Asian	TE	43	0	0
	White/Asian	CR	3	0	0
	White/Asian	Total	71	1	0
11	ML–No/ML–Yes	MC	22	3	1
	ML–No/ML–Yes	TE	39	3	1
	ML–No/ML–Yes	CR	1	0	2
	ML–No/ML–Yes	Total	62	6	4
11	SWD–No/SWD–Yes	MC	26	0	0
	SWD–No/SWD–Yes	TE	43	0	0
	SWD–No/SWD–Yes	CR	3	0	0
	SWD–No/SWD–Yes	Total	72	0	0
11	EconDis–No/EconDis–Yes	MC	26	0	0
	EconDis–No/EconDis–Yes	TE	43	0	0
	EconDis–No/EconDis–Yes	CR	3	0	0
	EconDis–No/EconDis–Yes	Total	72	0	0
11	CBT/TTS	MC	26	0	0
	CBT/TTS	TE	43	0	0
	CBT/TTS	CR	3	0	0
	CBT/TTS	Total	72	0	0

6.2 Item Response Theory

The grade-specific NJSLA–S student ability estimates and subsequent scale scores are calibrated via IRT statistical processes. Section 6.2 of this report explains how IRT is used in the context of the NJSLA–S. First, the concept of IRT is explained. Then, the specific IRT model used for the NJSLA–S is described in conjunction with the assumptions underlying the model. The remainder of Section 6.2 presents evaluations of how well the assumptions of IRT are met.

IRT is conceptualized as a family of mathematical models that provide an equation for the relationship between the probability of a student response to a test item and student latent ability on the construct of interest (Hambleton & Swaminathan, 1985). While latent traits (e.g., anxiety, intelligence, or mastery of the NJSLA–S) are not directly observable, student responses to items are directly observable. Within the context of the NJSLA–S, the latent trait theoretically being measured by the items is student understanding of the New Jersey science curriculum: the NJSLA–S. The directly observable behaviors resulting from that latent trait are the responses of students to those items.

IRT addresses many of the limitations of CTT, such as sample and test dependency, and can improve both the construction and uses of tests (Hambleton & van der Linden, 1982). Hence, IRT can enhance the validity of the inferences made from test scores. Under IRT, item parameters (e.g., item difficulty) are independent of the students who took the test. Similarly, student ability estimates are independent of the test items. Moreover, the test information function (TIF; see Section 8.2 for a more detailed explanation) allows for test construction to be targeted to specific places on the student ability spectrum where decisions are most important to maximize the test's ability to reliably classify examinees. The increased power of IRT in comparison to CTT requires that certain assumptions be met. When the assumptions of IRT are met, data can be used for psychometric analyses such as equating.

Comprising dichotomous and polytomous items with varying score points (e.g., 0–4-point CR items), the NJSLA–S was constructed to meet the assumptions of a specific IRT model: the Rasch-based partial credit model (PCM). The Rasch family of IRT models is a special case of IRT models. That is, Rasch models assume that items discriminate equally and that guessing on items is minimal (Hambleton & Swaminathan, 1985). The PCM is a flexible Rasch-based model that can be used with both dichotomous and polytomous item response data (Ostini & Nering, 2010).

For each polytomous item, there are some ordered levels of performance, and an associated number of steps required to move from one level to the next. Statistically, under the PCM, the probability, $\Pi_{ij}(x)$, of student *j* obtaining an item score *x* with *x* = 0, 1, . . ., *m* on polytomous item *i* can be written as follows:

$$\pi_{ij}\left(x;\theta_j;\ \beta_i,\tau_{in}\right) = \frac{exp\left[\sum_{n=0}^{x}\theta_j-(\beta_i+\tau_{in})\right]}{\sum_{k=0}^{m} exp\left[\sum_{n=0}^{k}\theta_j-(\beta_i+\tau_{in})\right]}, \qquad \text{Equation 6.1}$$

where θ_j is the student proficiency score, β_i denotes the difficulty or location parameter for item *i* and τ_{in} with *n* = 0, 1, . . ., *m* denotes the threshold or step parameters. For model identification, it is defined that $\tau_{io} = 0$, $\sum_{n=0}^{m} \tau_{in} = 0$ and $exp\left[\sum_{n=0}^{0} \theta_j - (\beta_i + \tau_{in})\right] = 1$. Accordingly, the predicated probability of a correct response (i.e., item score *x* = 1) to a dichotomous item is given by the following:

$$P(x_{ij} = 1; \theta_j, \beta_i) = \frac{exp(\theta_j - \beta_i)}{1 + exp(\theta_j - \beta_i)}, \qquad \text{Equation 6.2}$$

where $P(x_{ij} = 1; \theta_j, \beta_i)$ is the probability of student *j* with a proficiency score θ_j to obtain a correct response to item *i* and β_i denotes the item difficulty parameter for item *i*.

Assessing the IRT model fit (i.e., how well the NJSLA–S data meet the assumptions of the PCM) is imperative before using the PCM to analyze NJSLA–S data. If the NJSLA–S data do not meet the assumptions made by the PCM, then any results obtained by using the PCM would be suspect. However, if the NJSLA–S data meet the assumptions required by the PCM, then equating (as presented in Part 7 of this Technical Report) can be performed under the PCM to place item parameter and ability estimates on a common scale. This allows meaningful, grade-specific comparisons across forms and is important for ensuring the equivalence of the test across years.

The main assumptions of the PCM as they apply to the NJSLA–S are that the test is unidimensional, the items discriminate relatively equally, guessing on items is minimal, each individual item is independent of the others, and the resulting item parameter estimates are invariant regardless of who answered the items. Each of these five IRT assumptions will be explained in detail in the sections below as they relate to the PCM. Also, the PCM item category characteristic functions are graphically presented to show the relationships between student ability estimates and the probability of achieving a specific score point on the 0–3-point or 0–4-point CR items. Overall, the results of the 2025 NJSLA–S indicate that the assumptions of the PCM were adequately met.

6.2.1 Unidimensionality

Unidimensionality was checked via multiple methods. First, the intercorrelations among the subscores were evaluated. High correlations would indicate strong linear relationships among the subscore variables, providing evidence of unidimensionality. Second, the eigenvalues of the principal components analysis (PCA) were evaluated. A dominant first eigenvalue, in comparison to the other eigenvalues, is evidence of unidimensionality. Overall, there is ample evidence that the NJSLA–S is a unidimensional test and that the PCM assumption of unidimensionality has been met.

6.2.1.1 Intercorrelations

Tables 6.2.1 and 6.2.2 show the Pearson product-moment correlations among the domains and practices, respectively. High correlations would be evidence of a unidimensional test. Generally, more items in a cluster (i.e., a domain or a practice) will lead to a higher correlation between that cluster and the total test score.

At each grade level, all domain and practice subscores correlated with the total NJSLA–S test score at .93 or above. The lowest correlation between clusters was .78. The intercorrelations among subscores indicate that the NJSLA–S is a unidimensional test.

Table 6.2.1: Correlation Matrix for Domains

Grade	Domain	NJSLA–S	Earth and Space	Life	Physical
5	Earth and Space	.93	1.00	-	-
	Life	.94	.81	1.00	-
	Physical	.93	.79	.81	1.00
8	Earth and Space	.93	1.00	-	-
	Life	.92	.79	1.00	-
	Physical	.93	.80	.78	1.00
11	Earth and Space	.95	1.00	-	-
	Life	.95	.85	1.00	-
	Physical	.93	.83	.83	1.00

Table 6.2.2: Correlation Matrix for Practices

Grade	Practice	NJSLA–S	Critiquing	Investigation	Sensemaking
5	Critiquing	.93	1.00	-	-
	Investigating	.94	.80	1.00	-
	Sensemaking	.93	.80	.82	1.00
8	Critiquing	.93	1.00	-	-
	Investigating	.91	.76	1.00	-
	Sensemaking	.94	.81	.79	1.00
11	Critiquing	.95	1.00	-	-
	Investigating	.95	.85	1.00	-
	Sensemaking	.94	.84	.84	1.00

6.2.1.2. Principal Component AnalysisPrincipal Component Analysis (PCA) is a data reduction technique that attempts to account for the variance in measures by converting them into uncorrelated principal components (Brown, 2006). The resulting principal components can be ordered according to the eigenvalues (i.e., the magnitudes of variance accounted for) from the largest to the smallest. The first principal component accounts for as much measured variance as possible, and each succeeding factor does the same until there are as many principal components as original variables (Gorsuch, 1983). Then, a scree plot displays the eigenvalues on the Y-axis and the number (i.e., the order) of principal components on the X-axis. Gorsuch (1983) noted that this method of interpretation works well when sample sizes are large and the factors are well-defined. The scree plots are interpreted by finding the place on the plot where the slope leveled off. The principal components to the left of that point on the plot are deemed practically significant.

Figures 6.2.1 through 6.2.3 show the scree plots for grades 5, 8, and 11, respectively. As exhibited in these plots, the second most prominent eigenvalue for each grade level is below 2, whereas the most prominent eigenvalues approximately range from 11 to 14. At each grade, the scree plot shows that only one major dimension is contributing to the variability in student responses to items. The results of each grade's PCA provide further evidence of the unidimensionality of the NJSLA–S.

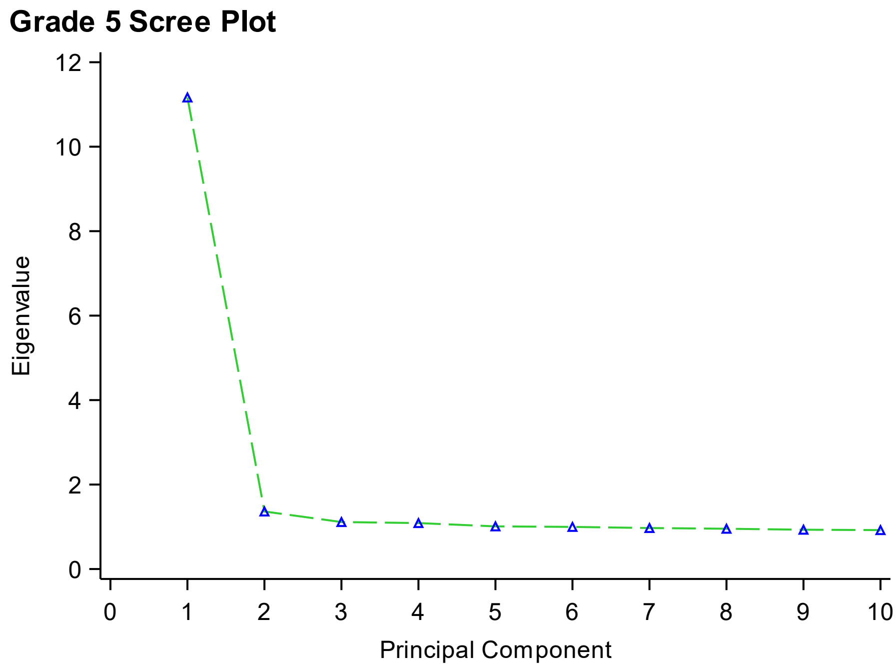


Figure 6.2.1. Grade 5 Scree Plot

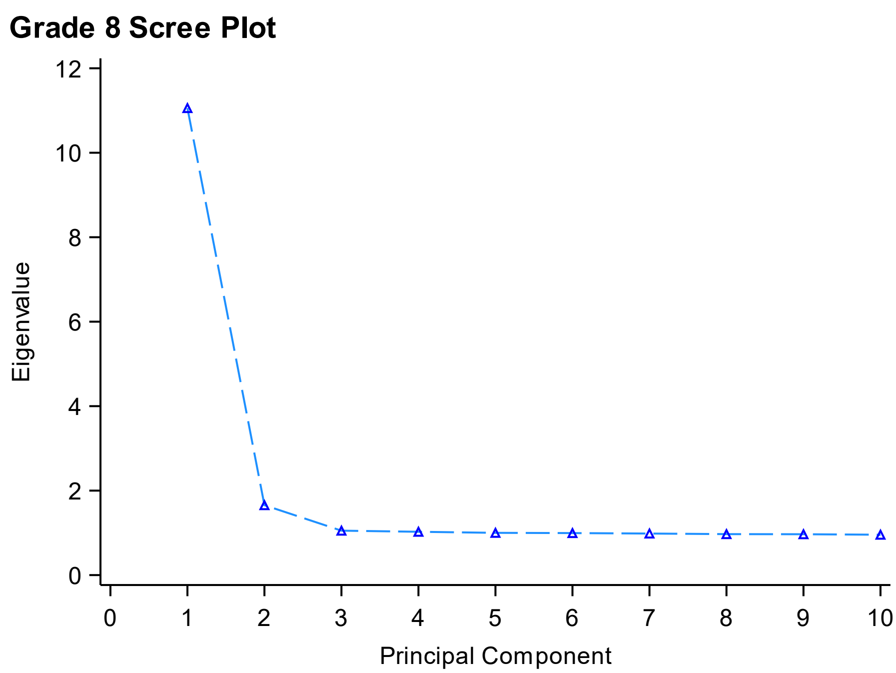


Figure 6.2.2. Grade 8 Scree Plot

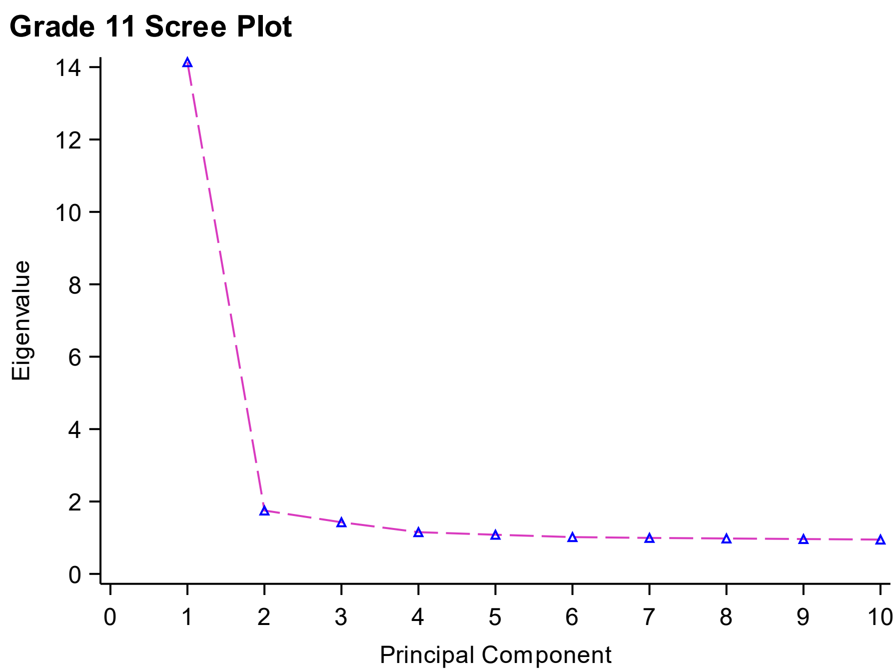


Figure 6.2.3. Grade 11 Scree Plot

6.2.2 Partial Credit Model-Fit Statistics

Hambleton, Swaminathan, and Rogers (1991) noted that "[a] poorly fitting IRT model will not yield invariant item and ability parameters" (p. 53), which diminishes the beneficial properties inherent to IRT. The PCM model fit was assessed at the item level via Rasch-based item infit and outfit, discrimination, and guessing statistics. At the person level, model fit was evaluated using Rasch-based person infit and outfit statistics. These statistics were calculated using the 2025 NJSLA–S test data via Winsteps 3.92.1 (Linacre, 2016); Appendix H of this Technical Report provides the resulting item fit statistics. Overall, there is ample evidence that items fit the assumptions of the PCM for all grades. In particular, the grades 8 and 11 item performance were remarkable, with a small (or zero) percentage of items flagged for each of the four model-fit categories.

6.2.2.1 Item infit and outfitRasch infit and outfit statistics range from zero to infinity, with 1.0 representing ideal model fit (i.e., no misfit). For the NJSLA–S, items were flagged for having infit or outfit statistics outside of the 0.7 to 1.3 range (Wright and Linacre, 1994). Infit statistics are influenced by unexpected responses from students on items that have item difficulties near their ability level (Wright and Masters, 1982). Conversely, outfit statistics are heavily influenced by unexpected student responses to items that are either relatively easy or relatively hard for the student based on the student's ability level.

Table 6.2.3 summarizes the item infit and outfit statistics across all grade levels, offering a reassuring view of item performance. No items were flagged for problematic infit at any grade, which is particularly encouraging given that infit statistics are more critical to the validity of score interpretations. However, there were several items flagged based on outfit statistics: three at grade 5, four at grade 8, and two at grade 11. Items flagged for outfit are less concerning for the psychometric integrity of a test. Generally, the overall pattern of infit and outfit results supports the conclusion that the assumptions underlying the PCM have been reasonably satisfied.

Table 6.2.3: Summary of Item Infit and Outfit Statistics

Grade	Fit Statistic	Average	Min	Max	Outside 0.7 to 1.3	% Flagged
5	Infit	1.00	0.75	1.22	0 out of 53	0.0
	Outfit	1.01	0.64	1.36	3 out of 53	5.6
8	Infit	0.99	0.74	1.18	0 out of 63	0.0
	Outfit	1.01	0.66	1.71	4 out of 63	6.3
11	Infit	0.99	0.80	1.23	0 out of 72	0.0
	Outfit	1.01	0.74	1.74	2 out of 72	2.8

6.2.2.2 Rasch discriminationThe PCM assumes that all items discriminate equally. Practically, items never discriminate equally, but if they are within reasonable thresholds then the assumption will be met. PCM does not model item discrimination, nor does it adjust item difficulty or person-ability estimates based on item discrimination. However, Winsteps provides an index of item discrimination that approximates the discrimination parameter from the 2PL model (Linacre, 2016). Rasch discrimination statistics are centered at 1.0, which indicates that the item is discriminating exactly as expected by the PCM. Items are flagged when their Rasch discrimination statistics fall outside of the range of 0.5 to 1.5.

Table 6.2.4 presents a summary of Rasch discrimination statistics across grades 5, 8, and 11. Most items fell within the acceptable range of 0.5 to 1.5, with only a small number flagged for values outside this threshold: three items at grade 5 (5.7%), three at grade 8 (4.8%), and one at grade 11 (1.4%). Notably, the flagged items had discrimination values that were close to the threshold boundaries. Generally, these results offer solid evidence that the constant discrimination assumption of the PCM has not been violated in a way that will impact the parameter estimates for the NJSLA–S.

Table 6.2.4: Summary of Rasch Discrimination Statistics

Grade	Average	Min	Max	Outside 0.5 to 1.5	% Flagged
5	1.01	0.44	1.59	3 out of 53	5.7
8	1.02	0.41	1.56	3 out of 63	4.8
11	1.01	0.47	1.42	1 out of 72	1.4

6.2.2.3 Rasch lower asymptoteThe PCM assumes that there is minimal guessing on the test items. Practically, however, students guess, and sometimes they guess correctly. Thus, as with the assumption of equal discrimination, the assumption of minimal guessing is met if item guessing statistics remain within a reasonable threshold. The PCM models guessing as misfit (i.e., infit and outfit) and does not adjust item difficulty or person-ability estimates based on guessing. However, Winsteps provides an index that approximates a guessing parameter in the form of lower asymptote statistics (Linacre, 2016). Rasch lower asymptote statistics are ideally zero, which indicates that an item is displaying little to no guessing. Items are flagged when their lower asymptote statistics fall outside of the range of 0.0 to 0.1.

Table 6.2.5 presents a summary of the lower asymptote statistics for grades 5, 8, and 11. Only a small number of items exceeded the .1 threshold: three (5.8%) in grade 5, three (4.8%) in grade 8, and five (6.9%) in grade 11. These flagged items tended to show relatively low item-total correlations, with values clustering around .25. None of the items flagged for their lower asymptote values were flagged for infit or outfit concerns. Overall, the lower asymptote results support that minimal guessing occurred on the NJSLA–S.

Table 6.2.5: Summary of Rasch Lower Asymptote Statistics

Grade	Average	Min	Max	Greater Than 0.1	% Flagged
5	0.02	0.00	0.17	3 out of 53	5.8
8	0.02	0.00	0.16	3 out of 63	4.8
11	0.02	0.00	0.19	5 out of 72	6.9

6.2.2.4 Rasch person infit and outfitPCM person-fit statistics are useful for evaluating whether student response patterns are reasonable. Reasonableness includes not only response patterns that are improbable, but those that are too probable. Multiple factors can cause distortions in the expected patterns of test scores, including:

- Carelessness—examinees miss items that they should have answered correctly.
- Cheating—examinees receive information to correctly answer items that they would normally miss.
- Guessing—examinees correctly answer items without knowing the correct answer.
- Creative responses—examinees misinterpret the item.
- Test administration errors.

Two PCM person-fit statistics were used: infit and outfit. Person infit is more influenced by responses to items that are targeted at the person's ability level; outfit is more influenced by responses to items that are relatively easy or hard for a student (Wright & Masters, 1982). Ideally, both statistics would be close to 1.0. For the NJSLA–S, values larger than 1.3 indicate model underfit, while values smaller than .7 indicate model overfit.

Tables 6.2.6 and 6.2.7 show, respectively, the person infit and outfit descriptive statistics by demographic variables. For NJSLA–S, person-fit statistics were evaluated based on the following demographics: gender, ethnicity, multilingual learner (ML) status, economically disadvantaged

(EconDis) status, and students with disabilities (SWD) status. Tables 6.2.8 and 6.2.9 break down, respectively, the person infit and outfit descriptive statistics by test forms including CBT, PBT, TTS, Spanish, Spanish TTS, and Human Reader forms. Figures 6.2.4 through 6.2.6 exhibit grade-level distributions of both the person infit and outfit statistics for all students.

At the overall level across all combinations of grade and demographic variables, as shown in Table 6.2.6, 4.80% of grade 5 students, 3.93% of grade 8 students, and 1.54% of grade 11 students were flagged for person infit statistics. Asian students tended to have the highest percentage of students flagged for person infit among the ethnic groups investigated at grades 5 and 8. As shown in Table 6.2.8, the grade 5 Human Reader (HR) forms flagged 5.65% of students for person infit. However, it should be noted that there were only 177 students taking the grade 5 HR form.

Overall, there were relatively more students flagged for aberrant person outfit statistics than for person infit statistics. As shown in Table 6.2.7, the multilingual learners and students with disabilities tended to have a higher percentage of students that were flagged for person outfit statistics in comparison to other demographic groups at their grade level. Within those groups, the students that were flagged also tended to be lower performing. Additionally, they were more likely to have taken accommodated forms. These forms had higher percentages of students flagged for person outfit than did the CBT forms.

As stated earlier, aberrant person outfit statistics are less of a threat to the validity of the test score inferences than are aberrant person infit statistics. Based on the large percentages of person outfit flags, it is likely that these students tended to be lower performing, but there were some items that they were able to unexpectedly answer correctly. Moreover, because the flagged students tended to be lower performing, it is unlikely that the misfit was having any meaningful impact on the reliability of the student proficiency classification. However, a deeper investigation into the person outfit statistics for multilingual learners, students with disabilities, and the accommodated forms was conducted to ensure there were no concerns. The results of this investigation are summarized in Section 6.2.2.5.

Table 6.2.6: Summary of Person Infit Statistics by Demographic Group

Grade	Group	N Students	Mean Scale Score	Person Infit Average	Person Infit Min	Person Infit Max	N Flagged	% Flagged	Flagged Mean Scale Score
5	NJSLA–S	96,215	173.11	1.00	0.59	2.99	4,620	4.80	180.57
	Male	48,685	174.45	1.00	0.59	2.84	2,276	4.67	185.94
	Female	47,516	171.73	1.00	0.61	2.99	2,344	4.93	175.35
	Am. Indian	192	181.20	0.97	0.71	1.39	2	1.04	199.50
	Asian	10,550	206.59	1.00	0.63	2.53	535	5.07	204.20
	Black	13,303	152.21	1.00	0.62	2.34	581	4.37	169.75
	Hispanic	32,784	155.97	1.00	0.61	2.99	1,564	4.77	166.23
	Pacific Islander	135	174.56	0.98	0.66	1.77	7	5.19	190.71
	White	35,920	185.57	0.99	0.59	2.63	1,789	4.98	189.23
	ML–Yes	10,424	133.48	1.00	0.61	2.26	366	3.51	146.10
	ML–No	85,788	177.93	1.00	0.59	2.99	4,254	4.96	183.53
	EconDis–Yes	41,269	153.27	1.00	0.61	2.99	1,895	4.59	164.26
	EconDis–No	54,944	188.01	1.00	0.59	2.84	2,725	4.96	191.90
	SWD–Yes	21,597	152.52	0.99	0.61	2.99	795	3.68	168.24
	SWD–No	74,618	179.07	1.00	0.59	2.63	3,825	5.13	183.13
8	NJSLA–S	99,219	165.59	0.99	0.56	2.33	3,901	3.93	176.41
	Male	50,914	165.41	0.99	0.58	2.33	1,964	3.86	180.22
	Female	48,275	165.76	0.98	0.56	2.29	1,936	4.01	172.52
	Am. Indian	148	162.69	0.98	0.68	1.46	6	4.05	187.33
	Asian	10,781	195.27	0.99	0.60	2.29	509	4.72	202.99
	Black	13,888	149.49	0.99	0.61	2.16	460	3.31	161.35
	Hispanic	34,174	151.75	0.99	0.56	2.16	1,285	3.76	162.79
	Pacific Islander	188	171.37	0.97	0.70	1.56	6	3.19	180.00
	White	37,138	175.01	0.98	0.58	2.33	1,520	4.09	182.90
	ML–Yes	9,176	133.90	1.01	0.64	2.16	271	2.95	149.59
	ML–No	90,039	168.81	0.98	0.56	2.33	3,629	4.03	178.41
	EconDis–Yes	41,311	150.65	0.99	0.56	2.16	1,510	3.66	162.15
	EconDis–No	57,904	176.24	0.98	0.58	2.33	2,390	4.13	185.42

Grade	Group	N Students	Mean Scale Score	Person Infit Average	Person Infit Min	Person Infit Max	N Flagged	% Flagged	Flagged Mean Scale Score
	SWD–Yes	20,582	148.75	1.00	0.58	2.33	649	3.15	167.50
	SWD–No	78,637	169.99	0.98	0.56	2.29	3,252	4.14	178.19
11	NJSLA–S	97,156	173.27	0.99	0.64	2.25	1,498	1.54	182.89
	Male	49,581	174.40	0.99	0.64	2.25	855	1.72	192.64
	Female	47,441	172.04	0.98	0.67	1.97	639	1.35	169.78
	Am. Indian	173	171.12	0.99	0.79	1.41	4	2.31	156.25
	Asian	10,409	215.95	0.99	0.69	1.82	160	1.54	209.80
	Black	13,787	151.32	0.99	0.68	1.98	236	1.71	171.17
	Hispanic	32,455	154.50	0.99	0.64	2.04	580	1.79	171.61
	Pacific Islander	195	173.15	0.99	0.77	1.45	4	2.05	200.00
	White	37,716	184.89	0.98	0.67	2.25	482	1.28	191.88
	ML–Yes	7,617	128.56	1.01	0.72	1.84	168	2.21	179.87
	ML–No	89,530	177.08	0.99	0.64	2.25	1,330	1.49	183.28
	EconDis–Yes	37,793	153.55	0.99	0.67	1.98	648	1.71	171.27
	EconDis–No	59,354	185.84	0.98	0.64	2.25	850	1.43	191.75
	SWD–Yes	19,765	152.87	0.99	0.68	2.25	261	1.32	176.70
	SWD–No	77,391	178.49	0.99	0.64	1.98	1,237	1.60	184.20

Table 6.2.7: Summary of Person Outfit Statistics by Demographic Group

Grade	Group	N Students	Mean Scale Score	Person Infit			N Flagged	% Flagged	Flagged Mean Scale Score
				Average	Min	Max			
5	NJSLA–S	96,215	173.11	1.01	0.25	3.40	8,539	8.87	136.67
	Male	48,685	174.45	1.02	0.26	3.40	4,453	9.15	138.64
	Female	47,516	171.73	1.01	0.25	3.13	4,084	8.59	134.52
	Am. Indian	192	181.20	0.99	0.61	1.62	9	4.69	134.33
	Asian	10,550	206.59	1.00	0.30	2.83	578	5.48	202.49
	Black	13,303	152.21	1.03	0.33	3.40	1,786	13.43	122.39
	Hispanic	32,784	155.97	1.03	0.25	3.13	3,952	12.05	124.41
	Pacific Islander	135	174.56	1.01	0.64	2.86	10	7.41	138.00
	White	35,920	185.57	1.00	0.28	2.54	1,974	5.50	153.00
	ML–Yes	10,424	133.48	1.08	0.25	3.40	2,104	20.18	117.42
	ML–No	85,788	177.93	1.00	0.28	2.98	6,435	7.50	142.96
	EconDis–Yes	41,269	153.27	1.03	0.25	3.40	5,184	12.56	122.75
	EconDis–No	54,944	188.01	1.00	0.26	2.95	3,355	6.11	158.17
	SWD–Yes	21,597	152.52	1.03	0.26	2.98	3,028	14.02	121.28
	SWD–No	74,618	179.07	1.01	0.25	3.40	5,511	7.39	145.12
8	NJSLA–S	99,219	165.59	1.01	0.33	3.49	7,950	8.01	123.20
	Male	50,914	165.41	1.02	0.33	3.49	4,649	9.13	122.88
	Female	48,275	165.76	1.00	0.41	3.28	3,298	6.83	123.66
	Am. Indian	148	162.69	1.03	0.69	1.74	16	10.81	123.75
	Asian	10,781	195.27	0.98	0.51	2.19	277	2.57	147.51
	Black	13,888	149.49	1.04	0.33	3.24	1,711	12.32	120.24
	Hispanic	34,174	151.75	1.03	0.34	3.28	3,978	11.64	121.09
	Pacific Islander	188	171.37	0.99	0.56	1.85	13	6.91	120.92
	White	37,138	175.01	0.99	0.44	3.49	1,772	4.77	126.61
	ML–Yes	9,176	133.90	1.10	0.34	3.49	2,003	21.83	118.67
	ML–No	90,039	168.81	1.00	0.33	3.28	5,947	6.60	124.73
	EconDis–Yes	41,311	150.65	1.04	0.33	3.49	4,916	11.90	120.91
	EconDis–No	57,904	176.24	1.00	0.35	2.41	3,034	5.24	126.92
	SWD–Yes	20,582	148.75	1.07	0.33	2.92	3,156	15.33	119.37
	SWD–No	78,637	169.99	1.00	0.34	3.49	4,794	6.10	125.73

Grade	Group	N Students	Mean Scale Score	Person Infit Average	Person Infit Min	Person Infit Max	N Flagged	% Flagged	Flagged Mean Scale Score
11	NJSLA–S	97,156	173.27	1.01	0.35	4.35	5,954	6.13	129.54
	Male	49,581	174.40	1.02	0.46	3.21	3,485	7.03	131.43
	Female	47,441	172.04	1.01	0.35	4.35	2,464	5.19	126.77
	Am. Indian	173	171.12	0.98	0.66	1.69	8	4.62	121.50
	Asian	10,409	215.95	1.00	0.54	4.35	364	3.50	222.40
	Black	13,787	151.32	1.03	0.52	2.39	1,212	8.79	114.64
	Hispanic	32,455	154.50	1.03	0.35	2.41	2,747	8.46	117.17
	Pacific Islander	195	173.15	1.03	0.71	1.98	13	6.67	121.46
	White	37,716	184.89	0.99	0.46	3.21	1,486	3.94	140.13
	ML–Yes	7,617	128.56	1.10	0.49	2.35	1,279	16.79	113.65
	ML–No	89,530	177.08	1.01	0.35	4.35	4,673	5.22	133.91
	EconDis–Yes	37,793	153.55	1.03	0.35	4.35	3,227	8.54	115.67
	EconDis–No	59,354	185.84	1.00	0.46	3.21	2,725	4.59	145.99
	SWD–Yes	19,765	152.87	1.04	0.51	3.21	1,921	9.72	115.77
	SWD–No	77,391	178.49	1.01	0.35	4.35	4,033	5.21	136.10

Table 6.2.8: Summary of Person Infit Statistics by Form

Grade	Form	N Students	Mean Scale Score	Person Infit Average	Person Infit Min	Person Infit Max	N Flagged	% Flagged	Flagged Mean Scale Score
5	CBT	68,817	180.60	1.00	0.59	2.84	3,455	5.02	184.93
	PBT	72	156.79	1.01	0.73	1.42	2	2.78	163.00
	TTS	24,261	156.86	0.99	0.61	2.99	1,045	4.31	170.24
	SP	1,660	133.35	1.01	0.66	2.02	70	4.22	138.60
	SP TTS	1,175	131.65	1.01	0.69	1.91	37	3.15	151.19
	HR	177	150.10	1.01	0.67	2.26	10	5.65	158.30
8	CBT	77,160	170.03	0.98	0.56	2.33	3,153	4.09	179.64
	PBT	40	148.23	1.05	0.81	1.26	0	0.00	N/A
	TTS	18,877	152.59	0.99	0.60	2.29	640	3.39	165.12
	SP	2,063	134.65	1.02	0.67	1.96	80	3.88	150.99
	SP TTS	999	135.03	1.00	0.69	1.75	25	2.50	142.16
	HR	55	136.45	1.01	0.69	1.58	2	3.64	146.50
11	CBT	85,163	176.73	0.99	0.64	2.25	1,302	1.53	185.10
	PBT	49	158.69	1.01	0.76	1.54	1	2.04	134.00
	TTS	9,288	152.96	0.99	0.69	1.82	129	1.39	163.07
	SP	2,231	133.86	1.01	0.73	1.81	59	2.64	178.32
	SP TTS	289	132.47	1.01	0.75	1.84	7	2.42	182.86
	HR	92	141.36	0.97	0.78	1.30	0	0.00	N/A

Note. CBT: Computer-Based Test; PBT: Paper-Based Test; TTS: Text-to-Speech; SP: Spanish; SP TTS: Spanish Text-to-Speech; HR: Human Reader

Table 6.2.9: Summary of Person Outfit Statistics by Form

Grade	Form	N Students	Mean Scale Score	Person Outfit Average	Person Outfit Min	Person Outfit Max	N Flagged	% Flagged	Flagged Mean Scale Score
5	CBT	68,817	180.60	1.00	0.27	3.13	4,819	7.00	146.50
	PBT	72	156.79	1.09	0.56	1.73	14	19.44	133.71
	TTS	24,261	156.86	1.03	0.25	3.40	3,117	12.85	125.07
	SP	1,660	133.35	1.08	0.44	2.65	327	19.70	117.61
	SP TTS	1,175	131.65	1.09	0.35	2.58	225	19.15	116.76
	HR	177	150.10	1.03	0.44	1.92	30	16.95	125.53
8	CBT	77,160	170.03	1.00	0.33	3.28	4,836	6.27	125.58
	PBT	40	148.23	1.17	0.87	1.83	9	22.50	119.67
	TTS	18,877	152.59	1.05	0.35	2.92	2,456	13.01	119.64
	SP	2,063	134.65	1.10	0.35	3.49	442	21.43	119.00
	SP TTS	999	135.03	1.08	0.34	2.25	188	18.82	119.13
	HR	55	136.45	1.12	0.66	2.02	14	25.45	119.07
11	CBT	85,163	176.73	1.01	0.35	4.35	4,723	5.55	133.10
	PBT	49	158.69	1.01	0.68	1.27	1	2.04	117.00
	TTS	9,288	152.96	1.04	0.48	2.76	892	9.60	115.61
	SP	2,231	133.86	1.07	0.49	1.91	284	12.73	116.49
	SP TTS	289	132.47	1.08	0.72	1.73	36	12.46	122.69
	HR	92	141.36	1.04	0.63	2.41	8	8.70	109.00

Note. CBT: Computer-Based Test; PBT: Paper-Based Test; TTS: Text-to-Speech; SP: Spanish; SP TTS: Spanish Text-to-Speech; HR: Human Reader

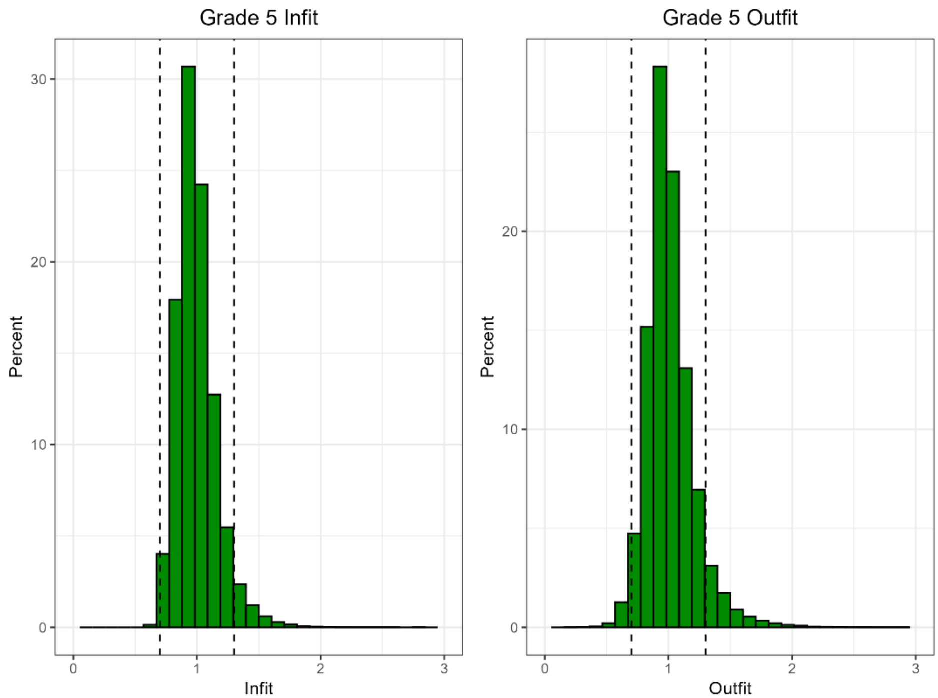


Figure 6.2.4. Grade 5 Person Infit and Outfit Distributions

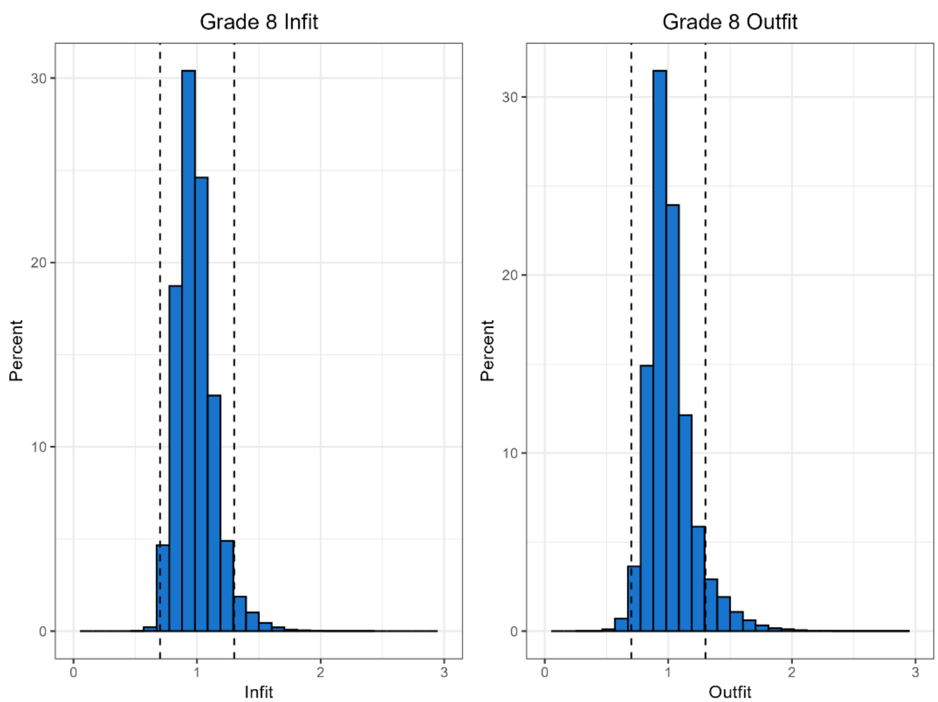


Figure 6.2.5. Grade 8 Person Infit and Outfit Distributions

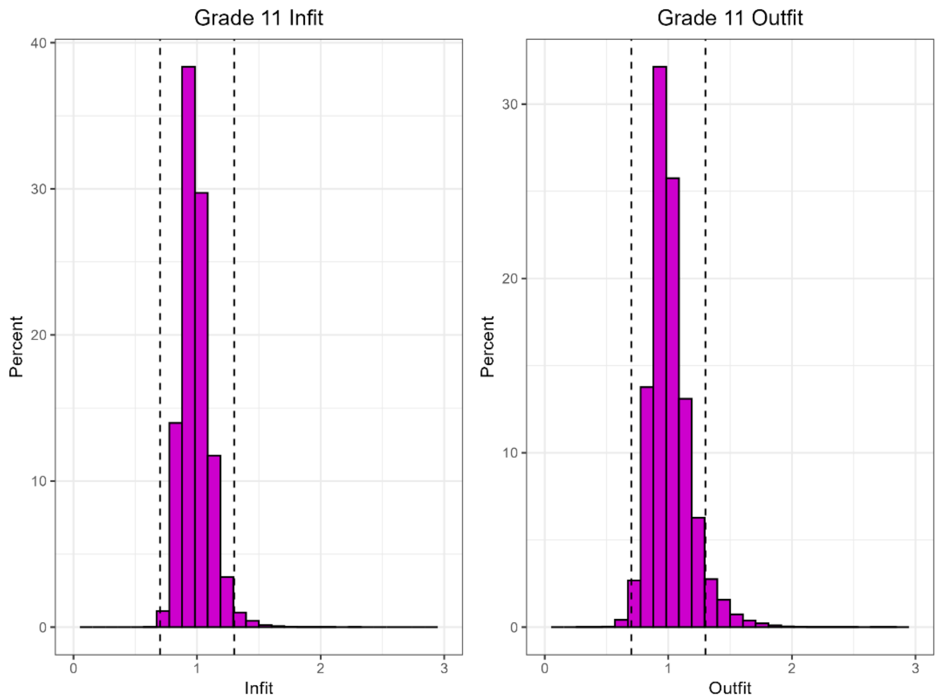


Figure 6.2.6. Grade 11 Person Infit and Outfit Distributions

6.2.2.5 Further investigation of person outfit statisticsWithin the context of the Rasch model, person infit and outfit statistics are based on model residuals and are interpreted in terms of fit and misfit. It is essential that the model-fit residuals are neither larger (indicating underfit) nor smaller (indicating overfit) than expected. In brief, overfit implies that the item scores were too predictable while underfit implies item scores were too unpredictable. Underfit is more problematic than overfit because an excess of expected scores (i.e., overfit) does not necessarily constitute evidence of model–data misfit within the measurement framework (Engelhard & Wind, 2018; Linacre, 2016). However, an excess of unexpected scores (i.e., underfit) could suggest there are some unaccounted-for factors affecting item scores. Therefore, this investigation focused on further examining the underfitting students in the student cohorts that showed a high frequency of outfit flags. Table 6.2.10 presents the distributions of total fit and total misfit, which further breaks down into overfit and underfit by student cohort for each grade.

Table 6.2.10: Distribution of Person Fit for Grades 5, 8, and 11

Grade	Group	ACC (%)	EconDis (%)	ML (%)	SWD (%)
5	Total	27,345 (100%)	41,269 (100%)	10,424 (100%)	21,597 (100%)
	Fit	23,632 (86.4%)	36,085 (87.4%)	8,320 (79.8%)	18,569 (86.0%)
	Total Misfit	3,713 (13.6%)	5,184 (12.6%)	2,104 (20.2%)	3,028 (14.0%)
	Overfit	838 (3.1%)	1,171 (2.8%)	387 (3.7%)	722 (3.3%)
	Underfit	2,875 (10.5%)	4,013 (9.7%)	1,717 (16.5%)	2,306 (10.7%)
8	Total	22,034 (100%)	41,311 (100%)	9,176 (100%)	20,582 (100%)
	Fit	18,925 (85.9%)	36,395 (88.1%)	7,173 (78.2%)	17,426 (84.7%)
	Total Misfit	3,109 (14.1%)	4,916 (11.9%)	2,003 (21.8%)	3,156 (15.3%)
	Overfit	336 (1.5%)	626 (1.5%)	183 (2.0%)	301 (1.5%)
	Underfit	2,773 (12.6%)	4,290 (10.4%)	1,820 (19.8%)	2,855 (13.9%)
11	Total	11,949 (100%)	37,793 (100%)	7,617 (100%)	19,765 (100%)
	Fit	10,728 (89.8%)	34,566 (91.5%)	6,338 (83.2%)	17,844 (90.3%)
	Total Misfit	1,221 (10.2%)	3,227 (8.5%)	1,279 (16.8%)	1,921 (9.7%)
	Overfit	88 (0.7%)	308 (0.8%)	61 (0.8%)	166 (0.8%)
	Underfit	1,133 (9.5%)	2,919 (7.7%)	1,218 (16.0%)	1,755 (8.9%)

Note. ACC = Students taking an accommodated test form; EconDis = Students with economically disadvantaged status; ML = Multilingual learners; SWD = Students with disabilities

While the PCM provides summary statistics to identify overall person misfit (i.e., infit and outfit), other methods are required to identify which items contributed to the unexpected student performance. Therefore, this investigation used the item difficulty parameters to identify which items showed unexpected performance for the group of students showing underfit. Specifically, the delta plot method (Angoff & Ford, 1973; discussed in Section 7.1) was employed. To conduct the delta method, the *p*-values of fitting students within a cohort were compared to the *p*-values of underfitting students within a cohort. This comparison was made between the students exhibiting no misfit and the underfitting students for each cohort (i.e., multilingual learners, students with economically disadvantaged status, students with disabilities, and students taking an accommodated test form) for all three grades. The observed *p*-values for the students exhibiting no misfit and students exhibiting underfit are provided in Appendix N of this report.

Table 6.2.11 exhibits the items flagged resulting from a single round of delta analysis. There were three, three, and five items flagged for grades 5, 8, and 11, respectively. For an item flagged for a student cohort, the associated *p*-values for the group of students showing no misfit (P_FT) and the *p*-values for the underfit subgroup (P_UF) are presented in Table 6.2.11. While there was some overlap in items flagged across cohorts, not all items were flagged for all cohorts. When an item was not flagged for a specific cohort, no P_FT or P_UF values were provided. As shown in Table 6.2.11, when flagged TE and MC items had Rasch B values larger than 1.2, the *p*-values associated with the underfit subgroup were higher than those associated with the fit subgroup. This pattern was observed in all three grades. In contrast, the underfit subgroup had lower *p*-values than the fit subgroup on the flagged TE, MC, and CR items with smaller Rasch B values.

Based on the study results, the higher rate of outfit flags seen in these student cohorts likely stems from the fact that these, on average, lower performing (see Table 6.2.7) students unexpectedly answered the challenging items correctly and did not perform as expected on the easier items. As previously mentioned, aberrant person infit statistics pose a greater threat to the validity of inferences than aberrant person outfit statistics. Furthermore, considering that the students flagged for outfit statistics were predominantly low performing, it is improbable that the presence of misfit significantly impacted the reliability of student proficiency classification.

Table 6.2.11: *P*-values of Items Flagged by Delta Method

	Flagged Items					ACC		EconDis		ML		SWD	
Grade	**Item UIN**	**DCI**	**SEP**	**Item Type**	**Rasch B**	**P_FT**	**P_UF**	**P_FT**	**P_UF**	**P_FT**	**P_UF**	**P_FT**	**P_UF**
5	2205B000_07	LS	CRI	CR	0.793	-	-	.30	.05	-	-	-	-
	2305B011_03	ESS	CRI	TE	2.540	.08	.28	.08	.27	.03	.22	.07	.29
	2405B008_05	ESS	SEN	TE	2.028	.14	.30	.13	.30	.10	.30	.14	.29
8	2308B018_09	LS	CRI	CR	-0.294	-	-	.36	.03	.20	.02	-	-
	2308M006_03	LS	INV	MC	1.250	.16	.46	.13	.45	.13	.44	.16	.47
	2408B005_04	PS	SEN	CR	0.545	-	-	.22	.01	-	-	-	-
11	2011M059_08	LS	INV	MC	2.033	.12	.45	.12	.43	.12	.44	.13	.44
	2111B002_11	LS	INV	CR	-0.096	.37	.05	.42	.05	.21	.02	.39	.06
	2311B004_04	PS	SEN	TE	1.859	.13	.29	.14	.29	.11	.32	.15	.31
	2311B014_04	ESS	CRI	CR	-0.483	-	-	-	-	.30	.06	-	-
	2311M058_05	ESS	CRI	TE	1.519	-	-	.16	.28	-	-	.17	.30

Note. ACC = Students taking an accommodated test form; EconDis = Students with economically disadvantaged status; ML = Multilingual learners; SWD = Students with disabilities; P_FT = Average *p*-value of students showing no misfit; P_UF = Average *p*-value of students showing underfit.

6.2.3 Local Independence

The PCM assumes that student responses to items are independent of responses to other items. In other words, student performance on one item does not affect performance on the other items on the test. If the assumption of local independence is violated, then the validity of inferences made from test scores could be threatened, the reliability of the assessment could be overestimated, and item-total correlations could be inflated. The assumption of local independence was tested via calculations of Yen's Q3 (Yen, 1984), which is an item residual correlation. The item residual (d_i) for item *i* at a student ability estimate $\hat{\theta}_j$ is defined as follows:

$$d_i = X_i - E(X_i; \theta_j), \qquad \text{Equation 6.3}$$

where X_i is an observed item score and $E(X_i; \theta_j)$ is the conditional expected item score under the IRT model of interest. The Q3 statistics for items *i* and *k* ($i \neq k$) are then computed as the Pearson correlation of d_i and d_k over all test takers.

Table 6.2.12 presents Yen's Q3 statistics for the NJSLA–S test across grade levels, evaluating all pairwise item combinations. A combination was flagged if its Q3 values exceeded .2 or fell below –.2, following the criteria suggested by Chen and Thissen (1997). At each grade level, the results support the assumption of local independence, as no pairwise item combination exhibited Q3 values outside the acceptable threshold.

Table 6.2.12: Summary of Yen's Q3 Statistics

Grade	Average	Min	Max	Outside –.2 to .2	% Flagged
5	–0.02	–0.11	0.18	0 out of 1,378	0.00
8	–0.01	–0.10	0.10	0 out of 1,953	0.00
11	–0.01	–0.09	0.11	0 out of 2,556	0.00

6.2.4 Item Characteristic Curves–CR Items

Under IRT, the item characteristic curves (ICC; the item categorical response functions) for a CR item show the relationship between latent student ability (theta) and the probability of achieving a specific score point on that item. The ICCs for each of the handscored, constructed-response, 0–3- or 0–4-point items are presented in Figures 6.2.7 through 6.2.15. The vertical dashed lines represent, from left to right, the Level 1/2, 2/3, and 3/4 cut scores on the theta scale. In addition, Table 6.2.13 shows the percentages of students receiving each score point for all the CR items by grade level.

Table 6.2.13: Constructed-Response Point Distribution Percentages

Grade	Item	%0	%1	%2	%3	%4
5	CR Item 1	35.48	18.59	8.98	22.70	14.25
	CR Item 2	19.55	18.47	26.52	35.45	N/A
	CR Item 3	47.47	27.90	17.43	7.20	N/A
8	CR Item 1	24.74	24.73	27.80	17.36	5.36
	CR Item 2	23.90	23.67	20.34	16.52	15.57
	CR Item 3	43.17	21.37	18.80	12.75	3.90
11	CR Item 1	29.39	16.43	26.34	27.85	N/A
	CR Item 2	47.70	27.68	15.92	8.69	N/A
	CR Item 3	19.06	17.03	28.06	35.84	N/A

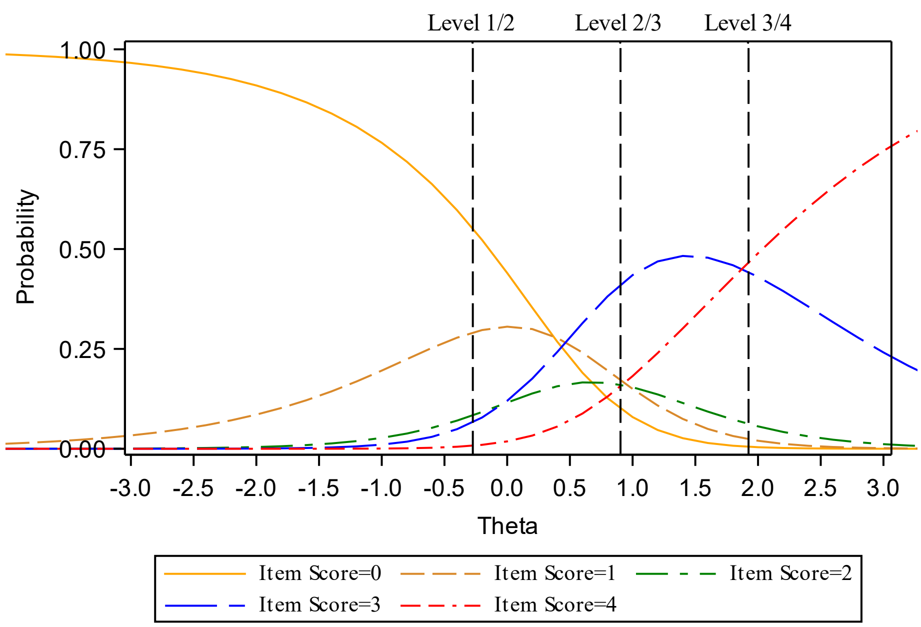


Figure 6.2.7. ICC Plot for Grade 5 Constructed-Response Item 1

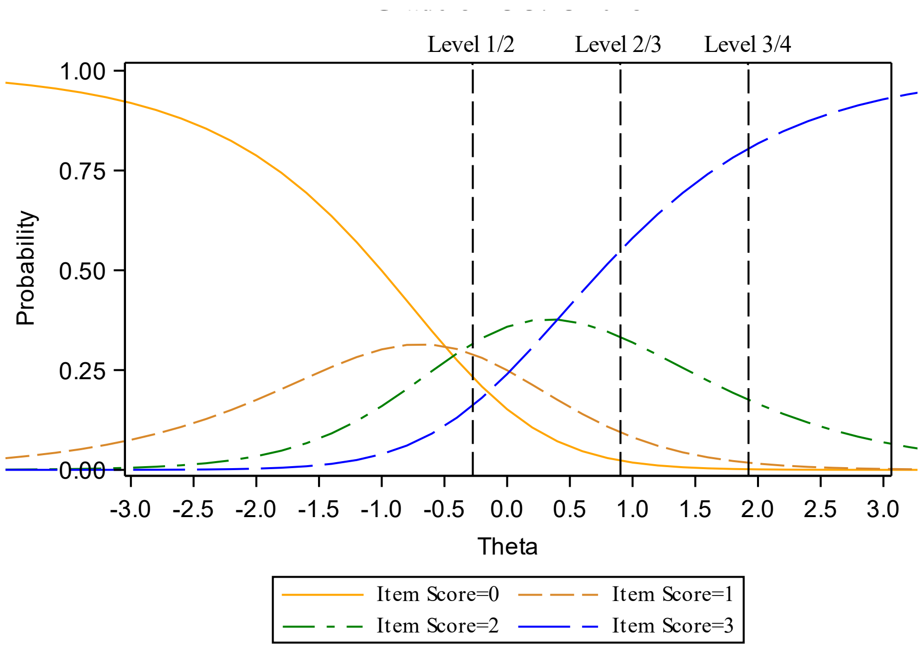


Figure 6.2.8. ICC Plot for Grade 5 Constructed-Response Item 2

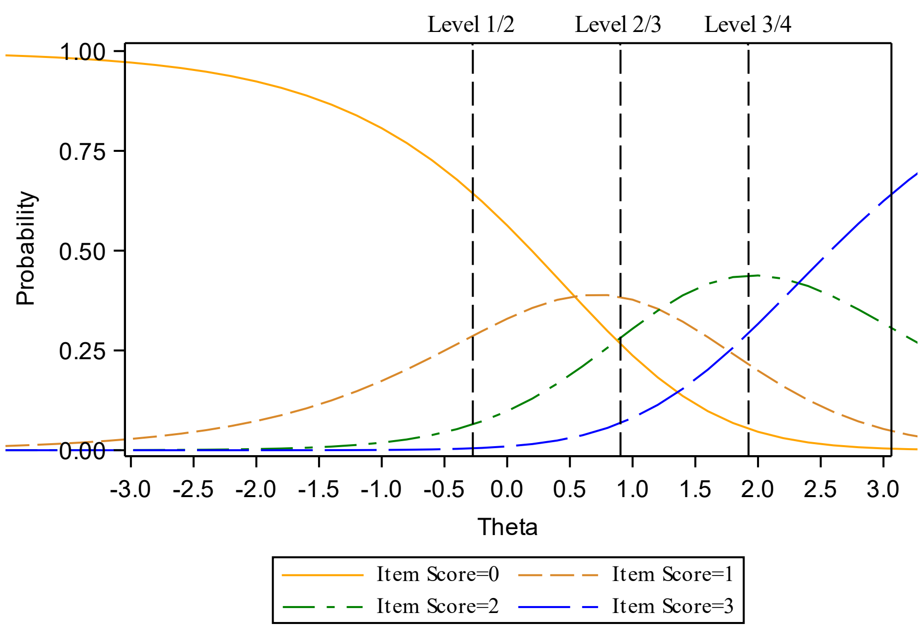


Figure 6.2.9. ICC Plot for Grade 5 Constructed-Response Item 3

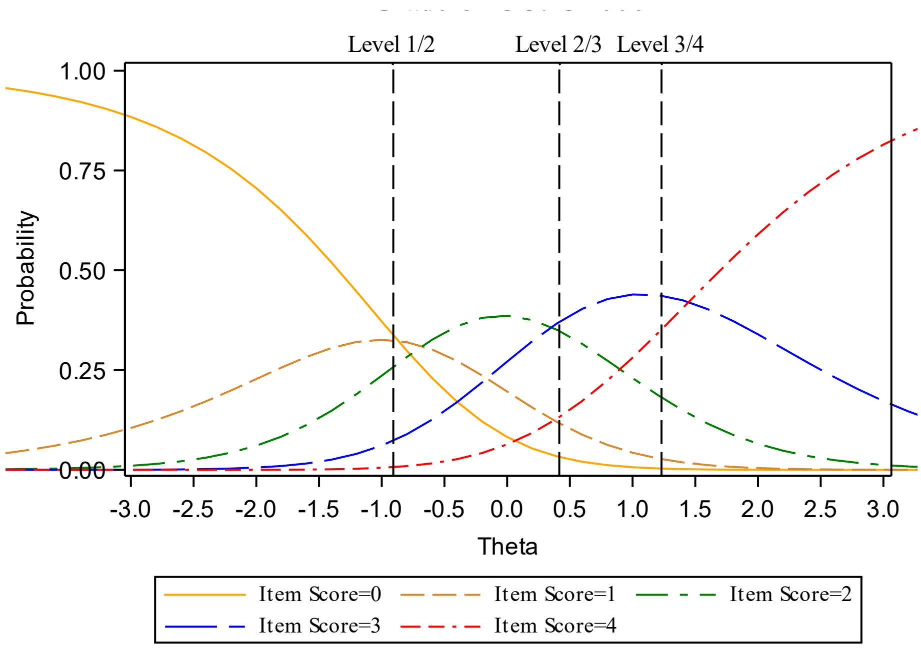


Figure 6.2.10. ICC Plot for Grade 8 Constructed-Response Item 1

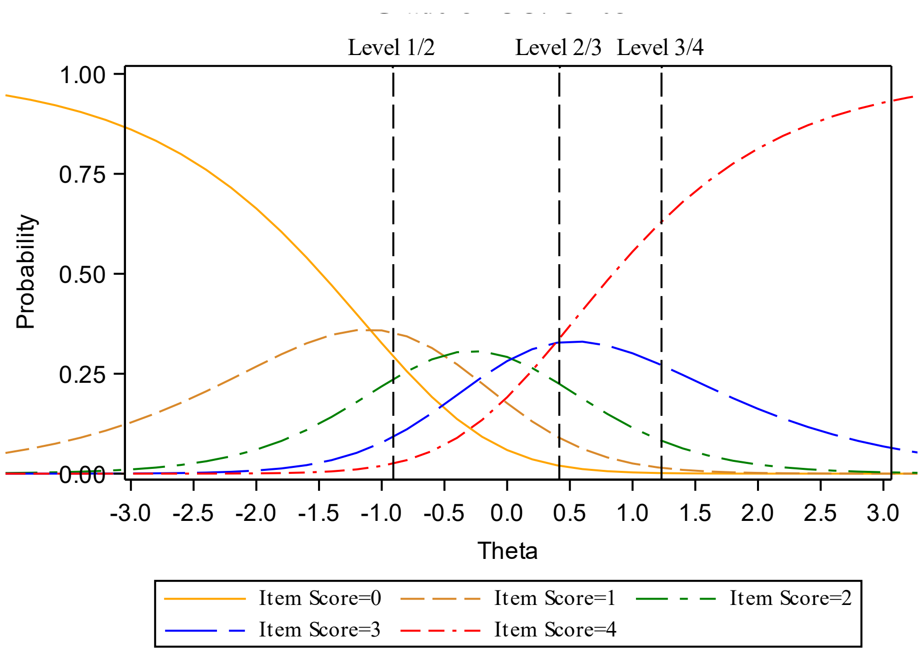


Figure 6.2.11. ICC Plot for Grade 8 Constructed-Response Item 2

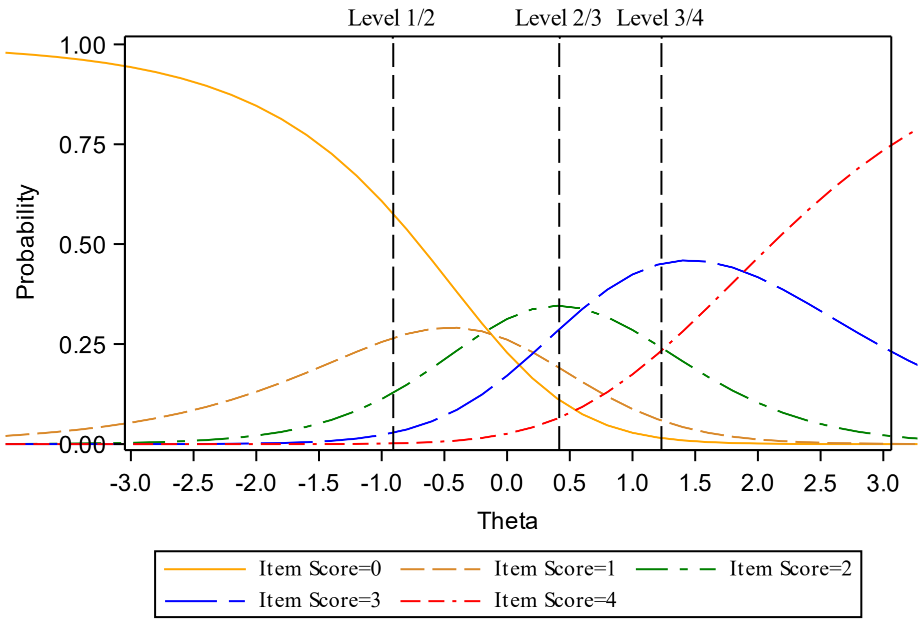


Figure 6.2.12. ICC Plot for Grade 8 Constructed-Response Item 3

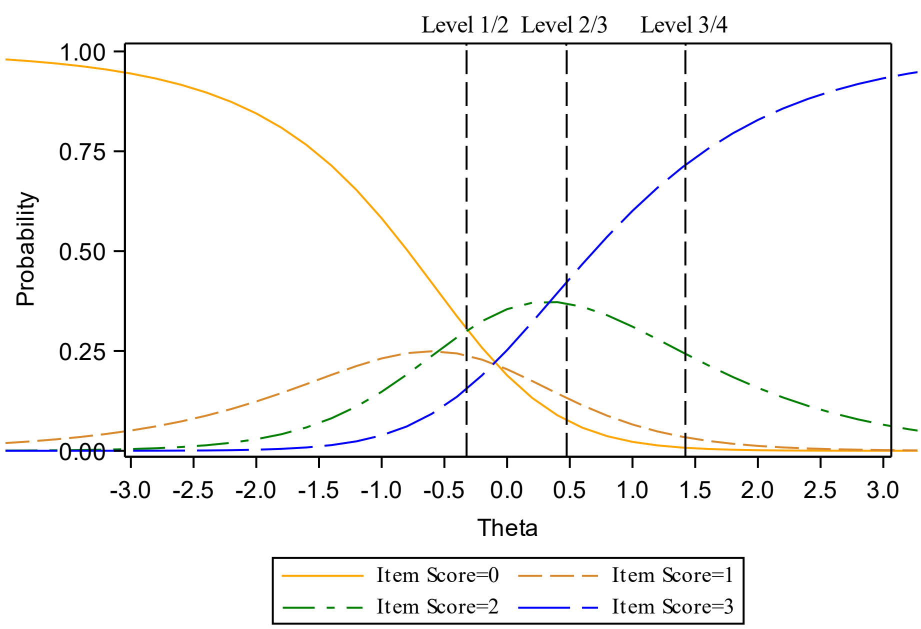


Figure 6.2.13. ICC Plot for Grade 11 Constructed-Response Item 1

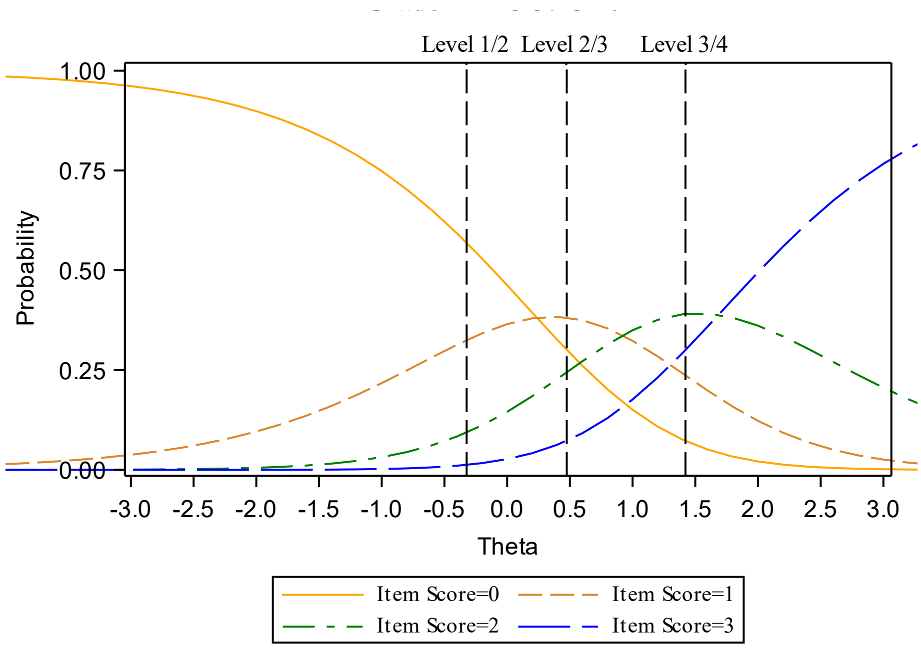


Figure 6.2.14. ICC Plot for Grade 11 Constructed-Response Item 2

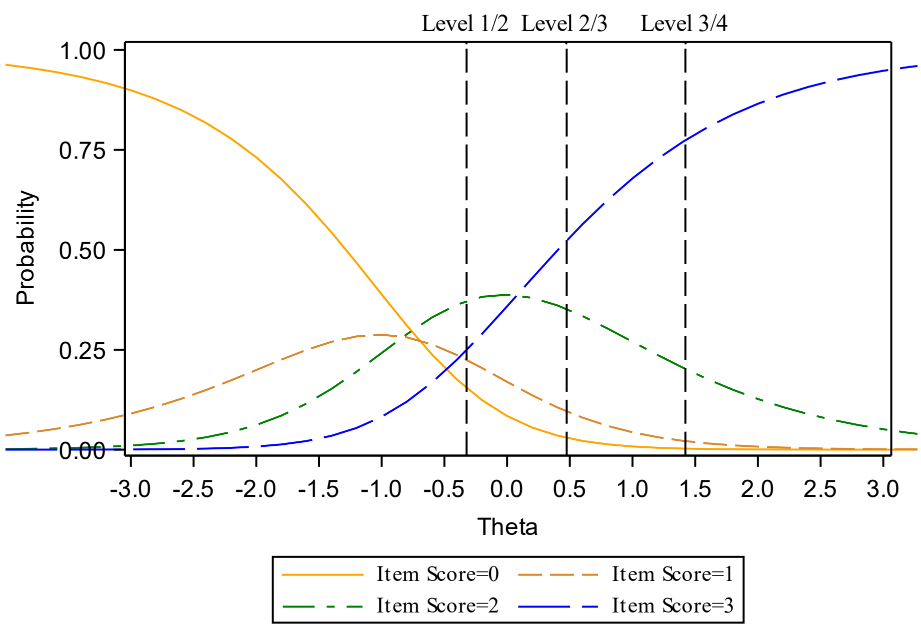


Figure 6.2.15. ICC Plot for Grade 11 Constructed-Response Item 3

6.3 Student Test Performance

Descriptive statistics for scale scores and proficiency-level distributions by form are presented in the following sections. For all the forms, scale scores have a range of 100 to 300. The Level 2/3 cut score is 200 at each grade level. Students who score at Level 3 or above are deemed proficient according to the results of the 2019 NJSLA–S Standard Setting. The Level 1/2 and 3/4 cut score ranges are more complex, and details regarding them can be found in

Section 7.1 of this Technical Report. It should be noted that no scale score comparisons should be made across grade levels.

6.3.1 Scale Score Distribution by Form

Descriptive statistics for scale scores and percentage distributions of students' proficiency levels by form are summarized by grade in Table 6.3.1. The forms include CBT, PBT, TTS, SP, SP TTS, and HR.

Table 6.3.1: Descriptive Statistics of Students' Test Proficiency by Form

Grade	Form	N Students	Average	S.D.	Min	Max	%L1	%L2	%L3	%L4
5	CBT	68,817	180.60	43.45	100	300	22.96	41.73	25.81	9.50
	PBT	72	156.79	35.44	103	254	45.83	37.50	15.28	1.39
	TTS	24,261	156.86	42.55	100	300	46.93	34.29	14.32	4.46
	SP	1,660	133.35	26.21	100	243	71.69	25.78	2.47	0.06
	SP TTS	1,175	131.65	25.82	100	250	74.81	22.81	2.30	0.09
	HR	177	150.10	37.48	100	279	55.37	29.94	11.86	2.82
8	CBT	77,160	170.03	34.84	100	300	28.12	50.10	16.12	5.66
	PBT	40	148.23	29.43	100	216	55.00	37.50	7.50	0.00
	TTS	18,877	152.59	33.32	100	300	50.09	39.14	8.31	2.46
	SP	2,063	134.65	22.01	100	219	74.65	24.43	0.92	0.00
	SP TTS	999	135.03	20.93	100	211	74.27	25.33	0.40	0.00
	HR	55	136.45	21.82	100	207	70.91	27.27	1.82	0.00
11	CBT	85,163	176.73	51.96	100	300	37.45	28.93	22.62	11.00
	PBT	49	158.69	43.78	100	267	55.10	26.53	14.29	4.08
	TTS	9,288	152.96	46.92	100	300	58.37	23.60	13.15	4.89
	SP	2,231	133.86	29.75	100	275	77.14	18.92	3.68	0.27
	SP TTS	289	132.47	26.39	100	232	80.97	17.30	1.73	0.00
	HR	92	141.36	34.31	100	229	67.39	23.91	8.70	0.00

Note. CBT: Computer-Based Test; PBT: Paper-Based Test; TTS: Text-to-Speech; SP: Spanish; SP TTS: Spanish Text-to-Speech; HR: Human Reader

6.3.2 Scale Score Distributions by Demographic Group

Descriptive statistics of scale scores and percentage distributions of students' test performance by demographic groups can be found on the New Jersey Statewide Assessment Reports web page. Scale score cumulative frequency distributions are attached as Appendix G of this Technical Report.

6.3.3 Subscore Proficiency Classification

There are no scale scores for the various subscores. As is explained in Section 7.3, student proficiency on the subscore categories was classified into three levels: Below, Near/Met, and Above Expectations. Appendix K of this report presents the percentages of students who were placed in the three subscore proficiency classifications. The data are disaggregated by form type, gender, ethnicity, and other demographic variables for the content domains and practices at each grade level.

At grade 5, the percentages of students classified as Below Expectation ranged from 47.04% for Earth and Space Science to 53.47% for Physical Science, and the percentages of students classified as Below Expectations for three practices (i.e., Critiquing, Investigating, and Sensemaking) hovered around 50%. At grade 8, all the contents and practices had Below Expectation percentages within the 60% to 70% range except for Earth and Space Science (58.44%). At grade 11, the Below Expectation percentages were below 50% for all the contents and practices except Earth and Space Science (50.62%) and Sensemaking (53.02%). Overall, across content domains and practices at each grade level, there was a noticeable difference between the percentage of students classified as Below Expectation within the demographic groups. The exception to this was gender, where the percentage of students classified as Below Expectation was similar for both male and female students across grades.

PART 7: EQUATING AND SCALING

Standard 5.12 states that "A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternative forms of a test may be used interchangeably" (p. 105). Equating is the process that allows for the interchangeability of test scores from year to year and within year test forms (Kolen & Brennan, 2004).

7.1 Summary of Equating and Scaling Procedures

The NJSLA–S uses an internal anchor item equating design, in which an anchor item set is a subset of the operational items, as well as the PCM (discussed in Section 6.2 of this report) for maintaining the scale. In addition to students who took the regular NJSLA–S test, the equating samples include students who took the accommodated forms, as New Jersey Department of Education policy requires the same score tables be used for all accommodated test forms. The equating samples are demographically representative of the population of NJSLA–S test takers in 2025 in terms of demographic distributions of gender, ethnicity, and socioeconomic status. Before the equating samples were created, additional analyses were conducted to guarantee the appropriateness of items for use in generating student test scores. A preliminary item analysis was conducted on multiple-choice items to validate the keys. Item scores of all the multiple-choice items were determined to have been correct.

For the NJSLA–S, equating and item calibrations include three phases of psychometric analyses. A free (unconstrained) calibration was first conducted under the PCM using the equating sample for each grade. The free calibration run converged successfully for the 2025 NJSLA–S equating sample at each grade. The free-run item parameter estimates were employed for the second phase of equating analyses: anchor item stability evaluations.

The NJSLA–S assessment uses two methods for anchor item stability evaluation. The first is the displacement evaluation method, which investigates the deviation (i.e., displacement) in the IRT item difficulty parameter estimates (i.e., Rasch B) of the anchored parameter values in comparison to the free-run parameter estimates. An item is flagged using the 0.3-logit absolute difference criterion (Miller et al., 2004). The second anchor item stability evaluation method is the delta plot method (Angoff & Ford, 1973), which compares the item means (using *p*-values for dichotomous items and adjusted mean scores for polytomous items) obtained from the current year equating samples to those obtained from previous test administration(s). The item means are converted to Delta values, which in turn are used to compute a best-fitted line. For all the anchor items under investigation, their perpendicular distances (PD) in Delta scores to the best-fitted line are computed and evaluated. An item is flagged if its resulting PD is more than two standard deviations away from the best-fitted line.

For the NJSLA–S, among the items flagged by both methods in a round of the anchor item evaluation process, the one with the largest absolute displacement value is dropped from the anchor set. The anchor item evaluation analysis is iteratively conducted until either no items are flagged by both evaluation methods or the number of dropped items reaches 20% of the original set of anchor items. For the 2025 NJSLA–S equating, for both grades 5 and 11, zero anchor items were flagged and dropped from the respective anchor item sets. At grade 8, the anchor evaluation processes were terminated at round two and yielded one (4.76%) dropped item.

After the completion of anchor evaluations, Winsteps 3.92.1 was used to calibrate the Rasch values (i.e., the B-parameter estimates and the step parameter estimates) of all operational items to the base theta scale (i.e., the base scale). This was done by constraining the retained anchor items to their Rasch values from the previous administrations or item bank that were already calibrated to the base scale. The results of the fixed Winsteps calibration run are used to develop the raw-to-theta-to-scale conversion tables for scoring. The development of scaling constants (i.e., intercept and slope) for converting theta scores to NJSLA–S scale scores is discussed in the following paragraph.

The NJSLA–S was scaled via a linear transformation that converted the IRT student ability estimates into scale scores. New Jersey has historically used a 100–300 scale for statewide assessments; in the past, with only three proficiency levels, scale scores of 200 and 250 represented proficient and advanced proficient performance, respectively. The NJSLA–S scaling procedure maintained the 100–300 scale; however, the scaling was slightly more complex due to the introduction of a third cut score (i.e., four proficiency levels). Policy decisions based on minimizing the number of students receiving the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) necessitated that, at grades 5 and 8, the Level 1/2 and the Level 2/3 cut scores be anchored during the linear transformation and at grade 11, the Level 2/3 and Level 3/4 cut scores be anchored. The linear transformation is described in detail below.

At all grades, a scale score of 200 still represents the proficient cut point (i.e., Level 2/3 cut). Students who score below 200 are placed in either Level 1 or Level 2. They are classified as below proficient and display minimal or partial understanding of the NJSLS–S. Students who score 200 or above are classified as either Level 3 or Level 4. Their performance is deemed proficient, and it represents an appropriate or exemplary understanding of the NJSLS–S.

The scale score ranges are reflected in Table 7.1.1 below. The scale scores representing the cut score differentiating Level 1 from Level 2 and differentiating Level 3 from Level 4 vary depending on each grade. At grades 5 and 8 the Level 1/2 cut score was anchored at a scale score of 150, whereas at grade 11 the scale score cut was 158. The Level 3/4 cut score was anchored at 250 for grade 11, while it was 243 for grade 5 and 231 for grade 8.

Table 7.1.1: Scale Score Ranges for Proficiency Levels by Grade

Grade	Level 1	Level 2	Level 3	Level 4
5	100–149	150–199	200–242	243–300
8	100–149	150–199	200–230	231–300
11	100–157	158–199	200–249	250–300

To produce the scale score ranges above, linear transformations were applied to theta (θ) estimates and scale scores. The following formula, adapted from Kolen and Brennan (2004, p. 337), was used to obtain the slopes and intercepts for the transformation functions:

$$sc(y) = \left[\frac{sc(y_2)-sc(y_1)}{\theta_2-\theta_1}\right] y + \left\{sc(y_1) - \left[\frac{sc(y_2)-sc(y_1)}{\theta_2-\theta_1}\right]\theta_1\right\}, \qquad \text{Equation 7.1}$$

where θ_1 and θ_2 are student ability estimates that correspond to the approved cut score points, and $sc(y_1)$ and $sc(y_2)$ are scale score points corresponding to θ_1 and θ_2, respectively. The resulting slopes and intercepts of the linear transformations at each grade level are shown in Table 7.1.2.

Table 7.1.2: Slope and Intercept of Theta-to-Scale Score Transformations and Proficiency-Level Cut Scores by Grade

Grade	Level 1/2 Cut		Level 2/3 Cut		Level 3/4 Cut		Slope	Intercept
	Theta	Scale Score	Theta	Scale Score	Theta	Scale Score		
5	–0.2739	150	0.9035	200	1.9243	243	42.4639	161.6317
8	–0.9077	150	0.4156	200	1.2306	231	37.7800	184.2960
11	–0.3230	158	0.4751	200	1.4217	250	52.8189	174.9036

The following sections specify how these slopes and intercepts were used to generate the scale scores at each grade level. The complete raw-to-scale score conversion tables can be found in Appendix I.

7.1.1 Rounding Rules

NJDOE policy requires that scaled scores below 100 are rounded up to 100, and that scaled scores above 300 are rounded down to 300. Additional rules of adjustments to scale score tables required for scaling are as follows:

- If a raw score maps to an unrounded scale score that is greater than 199.499 and less than or equal to 200.000, it will serve as the proficient (Level 2/3) cut score. Otherwise, the highest raw score that maps to a scale score less than or equal to 199.499 will serve as the cut score. The selected cut score will be assigned a value of exactly 200.
- If a raw score maps to an unrounded scale score that is greater than 249.499 and less than or equal to 250.000 for Level 4 at grade 11, it will serve as the advanced (Level 3/4) cut score. Otherwise, the highest raw score that maps to a scale score less than or equal to 249.499 will serve as the cut score. The selected cut score will be assigned a value of exactly 250. The same rounding procedures apply to the cut scores for Levels 1/2 and 2/3 for grade 11 as well as Levels 1/2, 2/3, and 3/4 for grades 5 and 8.
- If two unrounded scale scores fall in the range of greater than 199.499 and less than 200.000 (i.e., the Level 2/3 cut scores for all the grades), the lower of these two scores would become the cut score. The same rounding procedures apply to the cut scores of Levels 2 and 4 for all the grades.

- When the implementation of the above rounding rules results in two raw scores mapping to an equivalent rounded scale score, the scale score associated with the higher of the two raw scores will be adjusted upward by one (1) scale score.

7.2 Accommodative Form Equivalence

NJDOE has traditionally used the same score tables for their accommodative forms as for their traditional operational test forms, a decision that is predicated on several assumptions. These were checked for all accommodative forms by either content experts versed in universal design or, in the case of the braille and Spanish forms, external reviewers.

First, it must be assumed that the latent trait measured by the accommodative forms is the same as the latent trait measured by the regular operational test forms. Given that the same items measuring the same skills and abilities were used across the tests, it seems reasonable to assume that changes to item format or item presentation would not greatly change the overall latent trait or construct measured by each assessment form. Moreover, all items were written based on the principles of universal design as described in Section 3.4.

A second assumption is that item parameters across the test forms within each content cluster are identical. This, of course, is a potentially tenuous assumption considering the different item formats across the test forms. However, NJDOE's policy requiring that the same score tables be used for all accommodative test forms rendered this assumption necessary. Thus, all the accommodative forms for the NJSLA–S were assumed to be equivalent. If an operational item is unable to be properly adapted to a specific accommodative form, then the assumption of equivalence is violated, and a special equating is required. For any given administration year, if this assumption is violated for any accommodative form(s), the special equating procedure described in the following section is conducted. For the 2025 NJSLA–S administration, no special equating was needed.

7.2.1 Special Equating

In the event of errors during the test construction process that led to the removal of item(s) from the test, special equating is conducted to recalculate score tables so that the students who received those forms were placed onto a scale equivalent to that underlying the other CBT forms. The following steps are taken to ensure the special equating and CBT forms are on the same scale.

1. Anchored item calibration. The inequivalent items are removed prior to the special equating calibrations, and the item parameters and steps of the accommodated test items are fixed with the estimates resulting from the corresponding regular test items.
2. Theta to the scale score metric transformation. Because the theta values obtained from the anchored calibration and those obtained from the regular test score calibration are on the same metric, the transformation functions applied to the regular test scores could likewise be applied to the accommodated test scores.

3. Raw-to-scale score tables for each special equating. The rounding rules described in Section 7.1.1 are applied to the transformed scale scores, resulting in a separate raw-to-scale score table for each special equating that could be interpreted the same as the other operational forms.

7.3 Subscore Proficiency Levels

The NJSLA–S assessment reports student proficiency in three Disciplinary Core Ideas (DCI) including Earth and Space Science, Life Science, and Physical Science. The NJSLA–S also reports proficiency in three Science and Engineering Practices (SEP) including Investigating, Sensemaking, and Critiquing. In each DCI and SEP, subscore proficiency is classified as "Below," (Level 1) "Near/Met," (Level 2), or "Above" (Level 3) expectations. The subscores for these six reporting categories are described in Part 1 of this Technical Report. This section details the processes used to create the NJSLA–S subscore proficiency-level classifications.

For a given DCI or SEP at a grade level, the process for classifying NJSLA–S subscore proficiency first involved creating a subscore table. The subscore table was generated through a Winsteps fixed-parameter calibration run with the item parameter estimates of each item associated with the given DCI or SEP held constant (i.e., anchored) at the values obtained from operational item calibration results. A subscore table consisted of raw subscores, their associated thetas (θ), and the conditional standard errors of measurement (CSEM). The subscore proficiency-level classifications were based on the extent to which the subscore theta values within the subscore score tables were above or below the overall scale's Level 3 (proficient) theta cut score (denoted by θ^*) after taking the CSEM into account. Based on the subscore table, the CSEM associated with θ^*, denoted by CSEM^*, was estimated for subscore proficiency classification analyses. The "1.5 standard error rule" (Smarter Balanced Assessment Consortium, 2018) was then used to generate the subscore proficiency-level classifications as follows:

- A raw subscore is classified as "Above" if its associated θ is at or above (θ^* + 1.5 CSEM^*) units.
- A raw subscore is classified as "Below" if its associated θ is below (θ^* – 1.5 CSEM^*) units.
- A raw subscore is classified as "Near/Met" if its associated θ does not meet the definition of Above or Below.

The subscore score tables for each combination of grade and reporting category are presented in Appendix J of this report.

PART 8: RELIABILITY

Test reliability refers to the consistency of test scores. Ultimately, valid interpretations of test scores are dependent upon those scores being reliable. *Standard 2.0* states that "[a]ppropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (p. 42). Examples of appropriate evidence include reliability coefficients, conditional standard errors of measurement (CSEM), test information functions (TIF), and decision consistency measures, amongst others. The following sections detail evidence supporting the reliability of the NJSLA–S test scores and subscores.

8.1 CTT Reliability Estimates

This section describes the CTT reliability estimates calculated for the NJSLA–S. Section 8.1.1 describes the concept of reliability in the CTT framework, and Section 8.1.2 displays the reliability analysis results based on CTT.

8.1.1 Reliability and Measurement Error

Under the assumptions of CTT any observed measurement—such as a test score, X—is defined as a composite of true score, T, and its associated error:

$$X = T + \text{error} \qquad \text{Equation 8.1}$$

Errors in measurement can result from any of a multitude of factors, including environmental factors (e.g., testing conditions) and examinee factors (e.g., fatigue, stress). CTT provides a means for this quantification of examinee inconsistency (i.e., measurement error). Student test scores are reliable when measurement error is minimized. Increasing reliability by minimizing measurement error is an important goal in the construction of any test.

Estimating the size of the measurement error associated with the true score is the key to estimating reliability. The definitions or assumptions in CTT lead to several important properties. For example, it can be demonstrated that observed score variance (σ_X^2) equals the sum of true score variance (σ_T^2) and error variance (σ_e^2) or mathematically:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \qquad \text{Equation 8.2}$$

The relationships among the variance terms (i.e., $\sigma_X^2, \sigma_T^2, \sigma_e^2$) are critical to a more thorough understanding of important CTT concepts, including reliability and the standard error of measurement. Under CTT, reliability ($\rho_{X_1X_2}$) is defined as the correlation between observed scores (X_1, X_2) on parallel forms, which is equal to true score variance (σ_T^2) divided by observed score variance (σ_X^2):

$$\rho_{X_1X_2} = \frac{\sigma_T^2}{\sigma_X^2} \qquad \text{Equation 8.3}$$

With just a few algebraic steps, the CTT definition of the standard error of measurement (SEM, σ_e) can be shown as:

$$\sigma_e = \sigma_X\sqrt{1 - \rho_{X_1X_2}} \qquad \text{Equation 8.4}$$

Although the concepts of reliability and SEM are relatively straightforward, issues underlying the estimation of reliability are not. Reliability can be estimated via the correlation of scores on parallel forms or from test-retest data, or it can be estimated from a single test administration using any one of a variety of techniques (e.g., Brown, 1910; Cronbach, 1951; Kuder & Richardson, 1937).

For NJSLA–S, consistency of individual student performance was estimated using Cronbach's (1951) coefficient alpha. Coefficient alpha is conceptualized as the proportion of total raw score variance that may be attributed to a student's true score variance. Ideally, more score variance should be attributable to true test scores than to measurement errors. Coefficient alpha was estimated using the following formula:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right], \qquad \text{Equation 8.5}$$

where *n* is the number of items on the test, σ_i^2 is the item score variance of item *i*, and σ_X^2 is the variance of the observed total test score. Accordingly, SEMs were estimated and calculated using the following formula:

$$SEM = S_X\sqrt{1 - \alpha}, \qquad \text{Equation 8.6}$$

where S_X is the standard deviation of observed total scores. For the NJSLA–S assessments, separate analyses were performed for each grade level. Scores from all item types were used in the computations.

8.1.2 Raw Score Internal Consistency

In order to accommodate the state's diverse testing population, the NJSLA–S was delivered in multiple formats. The most used forms were the traditional online (CBT), the Text-to-Speech (TTS), the Spanish (SP), the paper-based test (PBT), and the Human Reader (HR). Reliability measures decrease when the students taking a given test form are more homogeneous in their test performance.

Table 8.1.1 displays the coefficient alpha and SEM for each form by grade. Overall, the reliability coefficients at each grade level indicate that students' raw scores were reliable for the NJSLA–S. The results at grade 5 stand out as particularly exceptional given that the grade 5 test is shorter than either the grade 8 or 11 tests. The most likely reason for the better results at grade 5, despite it being a shorter test, is that the grade 5 item difficulties were closer to the ability levels of the grade 5 students, thereby increasing the variance among test scores. The grade 5 reliability coefficients ranged from .77 for the SP and SP TTS forms to .91 for the CBT form. At grade 8, reliability ranged from .73 for the SP TTS form to

.91 for the CBT form. The grade 11 alpha coefficients ranged from .77 for the SP TTS form to .94 for the CBT form. At each grade level, the SP and SP TTS form test takers had lower average raw scores on the test compared to students taking other forms. This would result in less variance in test scores for these groups of test takers, which may explain the lower reliability estimates for these forms compared to the other test forms.

Table 8.1.1: Coefficient Alpha and SEM by Form

Grade	Form	N Students	Average Raw Score	S.D.	Alpha	SEM
5	CBT	68,817	24.57	12.31	.91	3.59
	PBT	72	17.86	9.78	.87	3.48
	TTS	24,261	18.13	11.59	.91	3.39
	SP	1,660	11.74	6.32	.77	3.03
	SP TTS	1,175	11.37	6.22	.77	3.00
	HR	177	16.18	10.13	.89	3.30
8	CBT	77,160	26.17	13.13	.91	3.87
	PBT	40	18.23	10.29	.87	3.67
	TTS	18,877	19.92	11.89	.90	3.67
	SP	2,063	13.58	6.71	.76	3.31
	SP TTS	999	13.62	6.30	.73	3.30
	HR	55	14.02	6.81	.75	3.37
11	CBT	85,163	31.36	15.75	.94	3.95
	PBT	49	25.76	13.27	.91	3.87
	TTS	9,288	24.26	14.01	.93	3.82
	SP	2,231	18.53	8.51	.82	3.61
	SP TTS	289	18.07	7.44	.77	3.57
	HR	92	20.66	9.99	.86	3.69

Note. CBT: Computer-Based Test; PBT: Paper-Based Test; TTS: Text-to-Speech; SP: Spanish; SP TTS: Spanish Text-to-Speech; HR: Human Reader

Table 8.1.2 presents coefficient alpha and standard error of measurement (SEM) values for raw scores across six reporting categories by grade level. Consistent with findings from Traub and Rowley (1991), longer assessments generally produce higher reliability estimates than shorter ones. This pattern is evident in the Critiquing category at grade 5, which contained fewer items and consequently showed lower reliability. Within the content domain, the Life Science subscore exhibited relatively low reliability (.75) at grade 8, whereas reliability levels were more consistent across grades 5 and 11. In the practice, Critiquing at grade 5 had the lowest subscore reliability (.72), reflecting its limited item count. In contrast, at grades 8 and 11, Critiquing demonstrated reliability estimates comparable to the other practice categories. Taken together, the reliability estimates for both overall test scores and subscores across grade levels demonstrate a consistently strong and acceptable level of measurement precision.

Table 8.1.2: Coefficient Alpha and SEM by Reporting Category

Grade	Reporting Category	Item Counts				Max Points	Alpha	SEM
		Total	MC	TE	CR			
5	NJSLA–S	53	12	38	3	60	.92	3.54
	Earth and Space	18	4	13	1	20	.78	2.02
	Life	16	5	10	1	19	.78	2.13
	Physical	19	3	15	1	21	.81	1.99
	Critiquing	15	3	10	2	20	.72	2.32
	Investigating	19	5	14	0	19	.84	1.81
	Sensemaking	19	4	14	1	21	.81	1.99
8	NJSLA–S	63	16	44	3	72	.91	3.84
	Earth and Space	21	5	15	1	24	.79	2.16
	Life	21	7	13	1	24	.75	2.28
	Physical	21	4	16	1	24	.80	2.21
	Critiquing	19	3	14	2	25	.77	2.41
	Investigating	22	11	11	0	22	.76	2.01
	Sensemaking	22	2	19	1	25	.81	2.20
11	NJSLA–S	72	26	43	3	78	.94	3.94
	Earth and Space	25	9	15	1	27	.83	2.35
	Life	25	10	14	1	27	.84	2.33
	Physical	22	7	14	1	24	.83	2.13
	Critiquing	21	3	16	2	25	.83	2.26
	Investigating	24	13	10	1	26	.83	2.34
	Sensemaking	27	10	17	0	27	.83	2.23

Table 8.1.3 shows the coefficient alpha and SEM by demographic group. These calculations are based on the entire test. In general, the coefficient alphas are consistently high among the various demographic groups. At grade 5, the lowest value was .80, for multilingual learner (ML-Yes) students, which is still very strong. At grade 8, the coefficient alphas hovered close to .90 except for the multilingual learners (α_{ML-Yes} = .77). The pattern for grade 11 was the same as for grade 8. The coefficient alpha values for all groups at grade 11 were above .90 except for the multilingual learners (α_{ML-Yes} = .83).

Table 8.1.3: Coefficient Alpha and SEM by Demographic Group

Grade	Group	N Students	Average Raw Score	S.D.	Alpha	SEM
5	NJSLA–S	96,215	22.54	12.46	.92	3.54
	Male	48,685	22.96	12.89	.92	3.53
	Female	47,516	22.11	11.99	.91	3.56
	Am. Indian	192	24.65	12.51	.92	3.57
	Asian	10,550	32.01	12.53	.92	3.56
	Black	13,303	16.75	10.08	.89	3.36
	Hispanic	32,784	17.75	10.49	.89	3.41
	Pacific Islander	135	23.05	12.18	.91	3.59
	White	35,920	25.97	11.92	.91	3.61
	ML–Yes	10,424	11.82	6.70	.80	3.03
	ML–No	85,788	23.84	12.37	.92	3.58
	EconDis–Yes	41,269	16.99	9.98	.89	3.38
	EconDis–No	54,944	26.71	12.51	.92	3.60
	SWD–Yes	21,597	16.97	11.18	.91	3.31
	SWD–No	74,611	24.15	12.34	.92	3.59
8	NJSLA–S	99,219	24.58	13.13	.91	3.84
	Male	50,914	24.61	13.60	.92	3.82
	Female	48,275	24.54	12.60	.91	3.85
	Am. Indian	148	23.55	12.68	.91	3.80
	Asian	10,781	35.96	13.46	.91	3.95
	Black	13,888	18.66	10.35	.88	3.63
	Hispanic	34,174	19.42	10.64	.88	3.66
	Pacific Islander	188	26.67	13.12	.91	3.87
	White	37,138	27.97	12.65	.90	3.91
	ML–Yes	9,176	13.40	6.87	.77	3.26
	ML–No	90,039	25.72	13.08	.91	3.86
	EconDis–Yes	41,311	19.01	10.34	.88	3.64
	EconDis–No	57,904	28.55	13.46	.92	3.91
	SWD–Yes	20,582	18.57	11.31	.90	3.59
	SWD–No	78,633	26.15	13.11	.91	3.87
11	NJSLA–S	97,156	30.32	15.70	.94	3.94
	Male	49,581	30.70	16.50	.94	3.91
	Female	47,441	29.91	14.80	.93	3.96
	Am. Indian	173	29.72	15.65	.94	3.94
	Asian	10,409	43.32	15.88	.94	3.93
	Black	13,787	23.71	12.76	.91	3.84
	Hispanic	32,455	24.65	13.15	.91	3.86
	Pacific Islander	195	30.28	16.52	.94	3.94
	White	37,716	33.80	15.29	.93	3.97
	ML–Yes	7,617	17.06	8.42	.83	3.52

Grade	Group	N Students	Average Raw Score	S.D.	Alpha	SEM
	ML–No	89,530	31.45	15.66	.94	3.95
	EconDis–Yes	37,793	24.36	13.04	.91	3.86
	EconDis–No	59,354	34.12	16.06	.94	3.96
	SWD–Yes	19,765	24.24	14.19	.93	3.80
	SWD–No	77,381	31.88	15.69	.94	3.96

Table 8.1.4 displays coefficient alpha and SEM by the three main item types: multiple-choice (MC), technology-enhanced (TE), and constructed-response (CR). Those item types are more thoroughly described in Part 2 of this Technical Report. As would be expected, as the number of points associated with a specific item type increase, so does the corresponding coefficient alpha. More than half of the points available on each test were associated with TE item types; thus, it is not surprising that at each grade level, the TE items displayed alphas close to .90. The alphas associated with each grade level's CR items were above .70, which is relatively strong given the limited number of points associated with them.

Table 8.1.4: Coefficient Alpha and SEM by Item Type

Grade	Item Type	Item Counts	Points	Average Raw Score	S.D.	Alpha	SEM
5	MC	12	12	5.70	2.75	.68	1.56
	TE	38	38	12.60	7.82	.89	2.55
	CR	3	10	4.24	2.89	.70	1.57
8	MC	16	16	6.46	3.01	.64	1.80
	TE	44	44	13.69	8.22	.89	2.73
	CR	3	12	4.43	3.07	.74	1.58
11	MC	26	26	10.99	4.88	.78	2.28
	TE	43	43	15.14	9.22	.91	2.72
	CR	3	9	4.19	2.69	.75	1.34

8.2 IRT Reliability

The reliability of the scale scores from the PCM (discussed in Section 6.2) was assessed in multiple ways. Test information functions (TIFs), conditional standard error measurements (CSEMs), and person-fit statistics were evaluated at each grade level. Overall, all results support that the 2025 NJSLA–S scores were reliable from the perspective of IRT and the PCM.

8.2.1 Test Information Functions

In IRT, the reliability of an assessment is conceptualized via the test information function (TIF, Hambleton & Swaminathan, 1985). Unlike coefficient alpha, the TIF is not uniform across the entire range of test scores. Instead, the TIF can assess test reliability across the full range of scores. This is particularly important to a criterion-referenced test such as the NJSLA–S because it allows for the reliability of the assessment to be evaluated at the most important decision points (i.e., the Level 2/3 cut scores).

Psychometrically, under the IRT assumption of local independence, the TIF for a test is the summation of all the item information functions (IIF; Lord & Novick, 1968; Hambleton, 1989) as follows:

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta) \qquad \text{Equation 8.7}$$

where $I(\theta)$ is the amount of test information at an ability level of θ, $I_i(\theta)$ is the amount of information for item *i* at an ability level of θ, and *N* is the number of items on a test. It should be noted that the mathematical definition of the amount of item information depends on the IRT model employed. Under the PCM where the responses to item *i* are scored as the integers 0, 1, . . ., m_i, the item information in item *i* is given by (Donoghue, 1994):

$$I_i(\theta) = \sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_i} k P_{ik}(\theta)\right)^2 \qquad \text{Equation 8.8}$$

where $P_{ik}(\theta)$ is the probability that an examinee of a given ability level θ will obtain a score of *k* on item *i*. With a few algebraic steps, the item information for a dichotomous item under the Rasch model is given by the following:

$$I_i(\theta) = P_i(\theta)\left(1 - P_i(\theta)\right) \qquad \text{Equation 8.9}$$

Figures 8.2.1 to 8.2.3 illustrate, respectively, the TIFs for grades 5, 8, and 11 at person-ability estimates ranging from –6 to +6. Within each figure, there are three vertical dash lines representing the test proficiency cut scores. More information at a specific ability level implies less measurement error. Ideally, the Level 2/3 cut score would occur at the peak of the information function where the most information and the least measurement error occur. Given the importance of making decisions at the Level 1/2 and 3/4 cut scores, the graph would also maintain ample information at those places along the scale.

The TIFs at each grade level were assessed primarily by whether they peaked close to the Level 2/3 cut score, and whether there was a precipitous drop in information at the Level 1/2 and 3/4 cut scores. For all grades, the TIFs peaked close to the Level 2/3 cut score but had drops in information at the Level 1/2 and Level 3/4 cut scores. Overall, the TIFs provide ample evidence that student ability estimates are reliable at the most important decision points. Nonetheless, all grades would benefit from more information around the Level 1/2 and Level 3/4 cut scores on future tests.

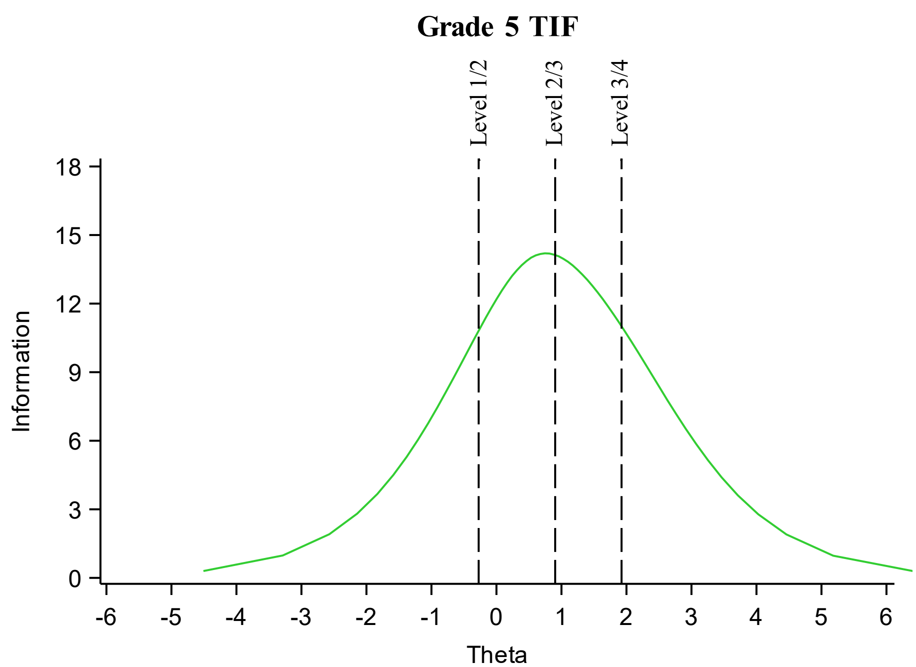


Figure 8.2.1. Grade 5 Test Information Function

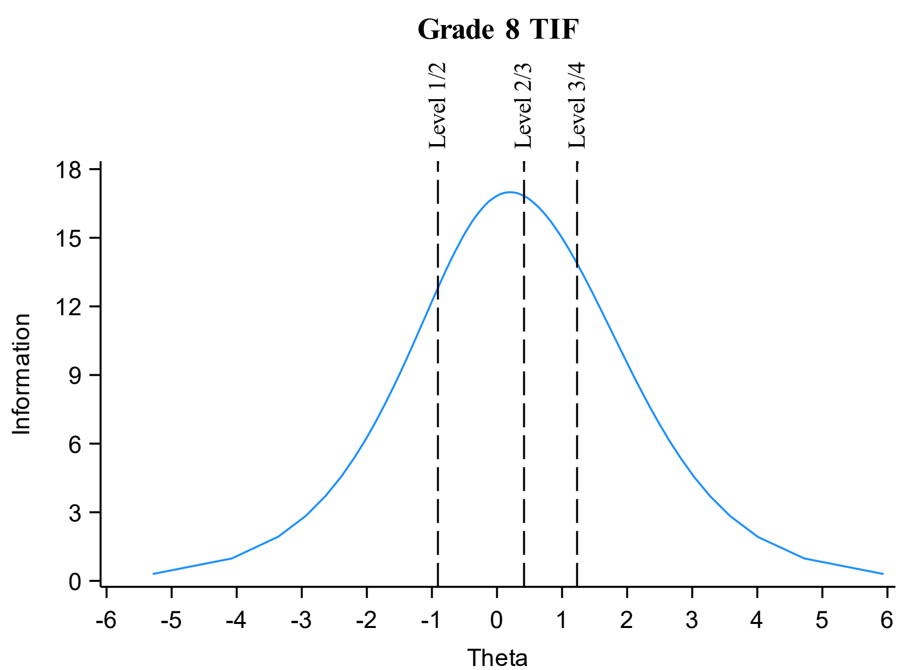


Figure 8.2.2. Grade 8 Test Information Function

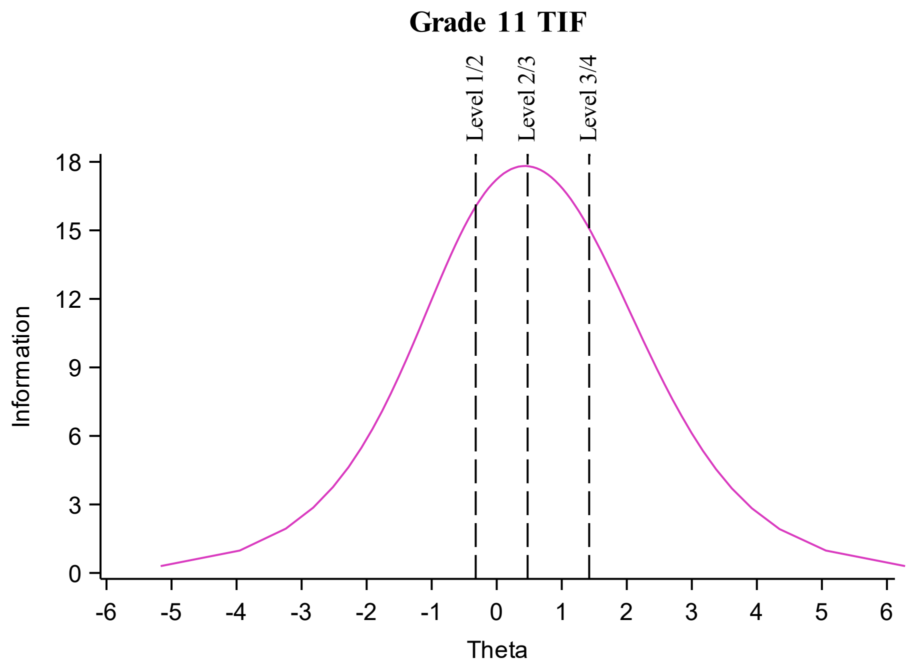


Figure 8.2.3. Grade 11 Test Information Function

8.2.2 Conditional Standard Error of Measurement

Under IRT, the conditional standard error of measurement (CSEM) of an examinee's estimated ability plays an important role in psychometric analyses. Mathematically, CSEMs are inversely related to the TIF, $I(\theta)$, and given by the following:

$$CSEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad \text{Equation 8.10}$$

where $I(\theta)$ indicates the amount of test information (TIF) at an ability level of θ. TIF was discussed in Section 8.2.1 of this report. Given that the CSEMs are the inverse of the TIF, their interpretations are similar. If the amount of information at a given level of θ is large and hence the corresponding CSEM is small, it means an examinee whose true ability is at that level can be estimated with precision. That is, the estimates will be reasonably close to the true value. It should be noted that the TIF and CSEM do not depend on the distribution of examinees over the ability scale.

Figures 8.2.4 through 8.2.6 illustrate, respectively, the CSEMs for grades 5, 8, and 11 at ability estimates ranging from –6 to +6. As shown in these figures, the CSEMs around the three cut scores are around .25 for each grade level, indicating ability scores around the three cut scores are estimated with precision.

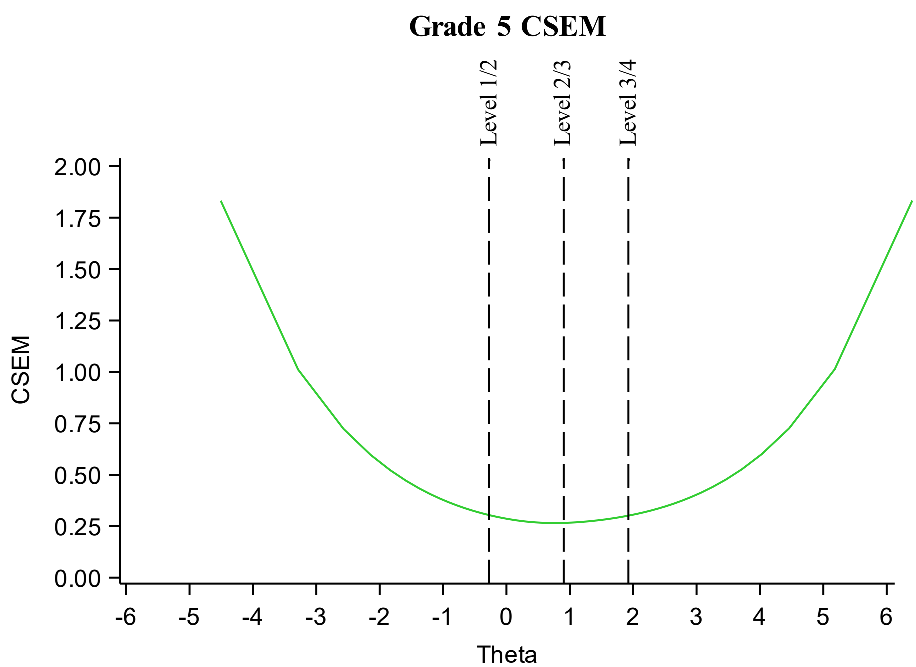


Figure 8.2.4. Grade 5 Conditional Standard Error of Measurement

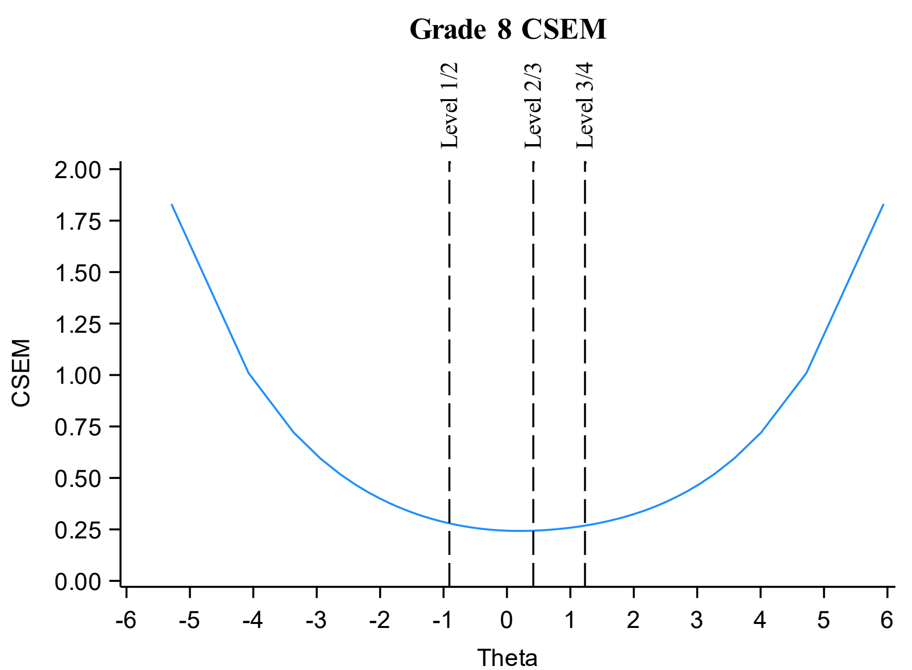


Figure 8.2.5. Grade 8 Conditional Standard Error of Measurement

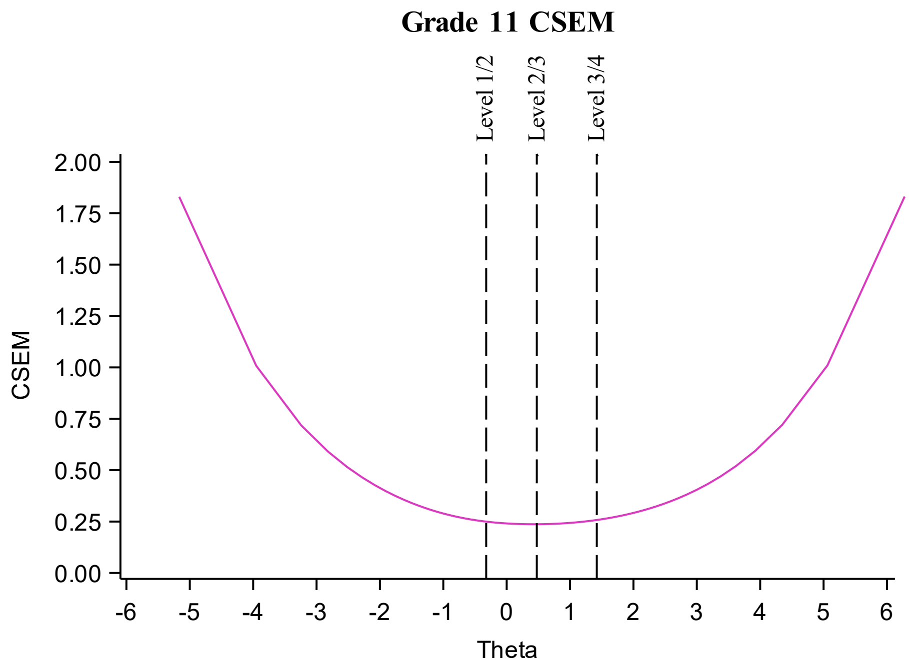


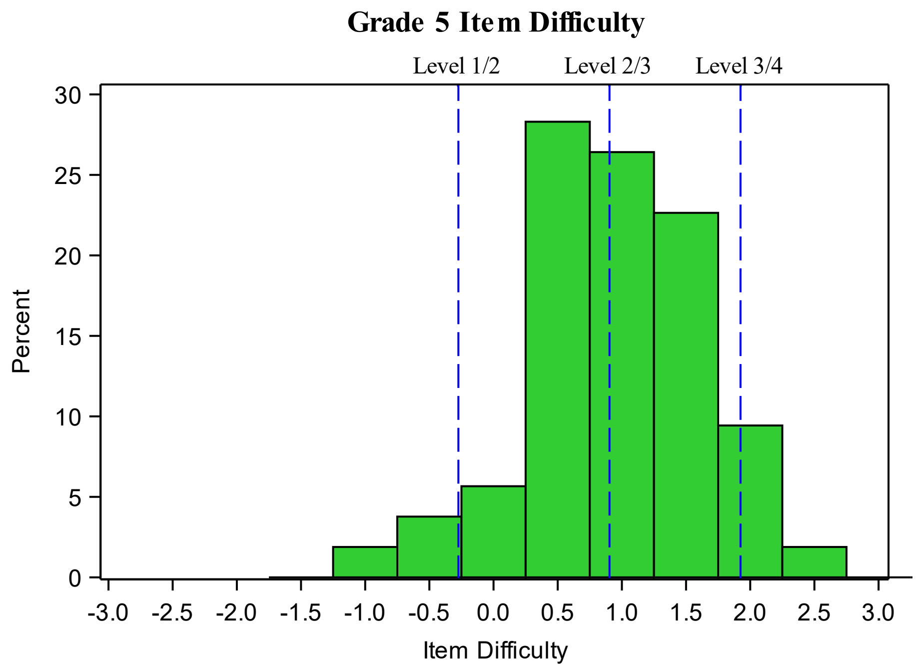
Figure 8.2.6. Grade 11 Conditional Standard Error of Measurement

8.2.3 Item Maps

An item map exhibits the distribution of person-ability estimates, and the distribution of item difficulty parameter estimates along the latent scale (i.e., theta). Item maps are useful to compare the range and positions of the item difficulty distribution to those of the person-ability measure distribution. Items that are targeted to the ability levels of the students taking the test will result in more reliable measures of student ability.

Figures 8.2.7 through 8.2.9 display the 2025 item maps for grades 5, 8, and 11, each comprising two panels: the upper panel illustrates the distribution of item difficulty estimates, while the lower panel depicts the distribution of student ability (theta). Across all grade levels, the NJSLA–S items were well aligned with the corresponding student ability distributions, indicating appropriate targeting of item difficulty.

At grade 5, item difficulty estimates clustered between the Level 1/2 and Level 2/3 cut scores, while student ability peaked around the Level 1/2 threshold. Few students performed above the Level 3/4 cut score, and correspondingly, there were limited items positioned at that higher end of the scale. In grade 8, the item difficulty distribution showed a relative scarcity of easier items, creating a slight mismatch with the student ability distribution, which included more students at the lower end. For grade 11, item difficulty estimates peaked near the Level 2/3 cut score and included a notable number of more challenging items, whereas student ability was concentrated closer to the Level 1/2 cut score. These patterns suggest that while item targeting was generally effective, some refinements may enhance alignment at specific grade levels.



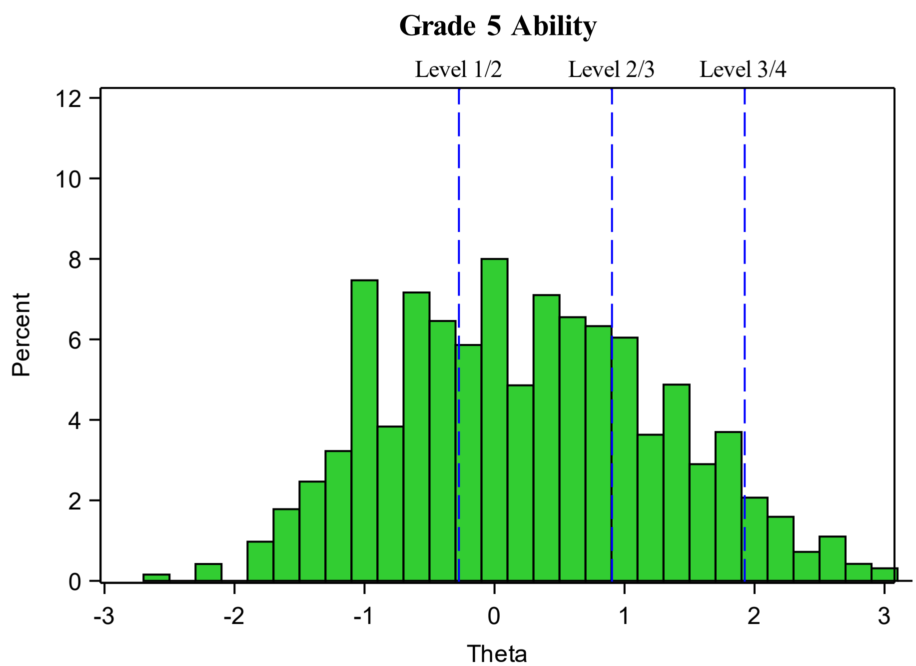


Figure 8.2.7. Grade 5 Item Difficulty and Student Ability Distributions

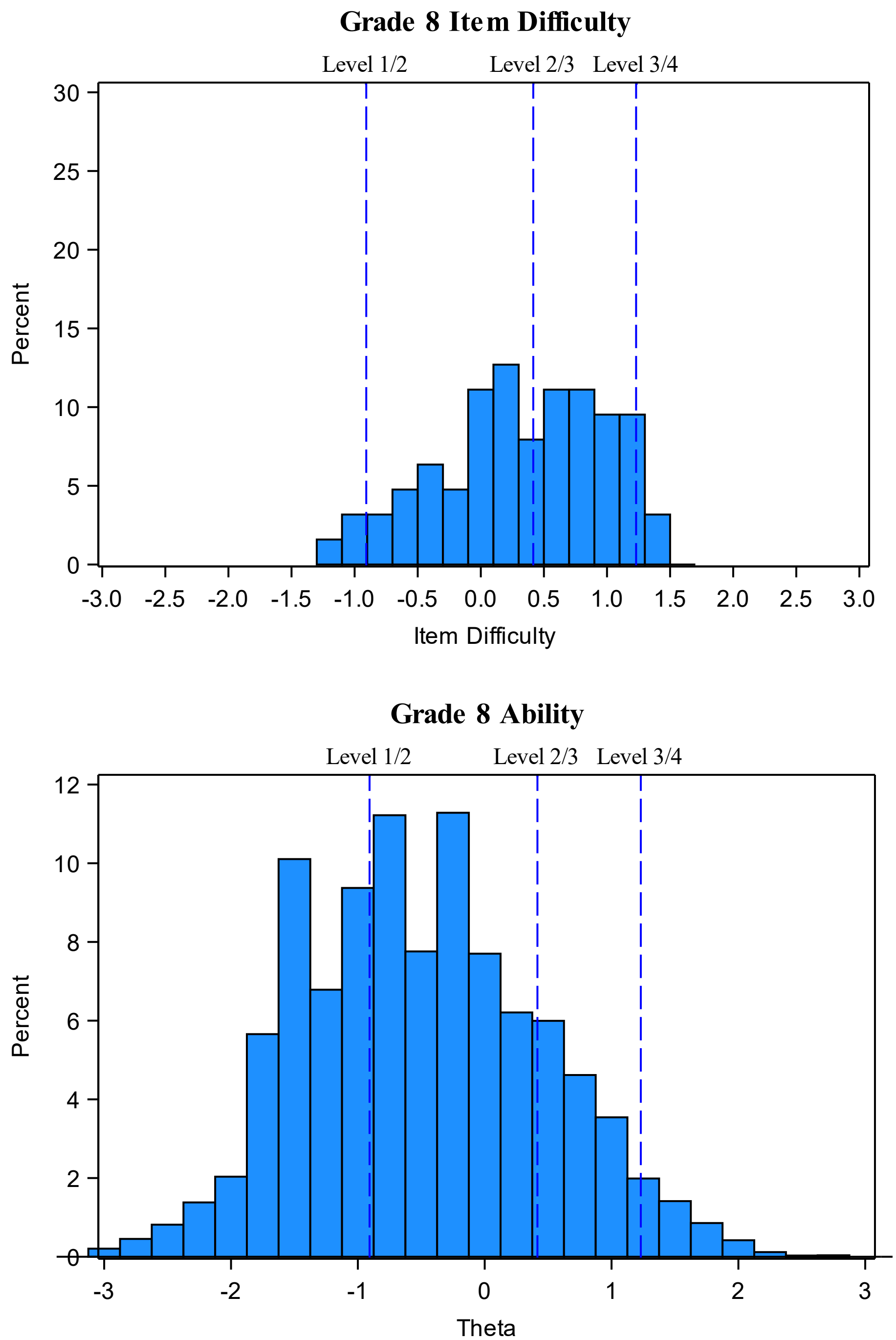


Figure 8.2.8. Grade 8 Item Difficulty and Student Ability Distributions

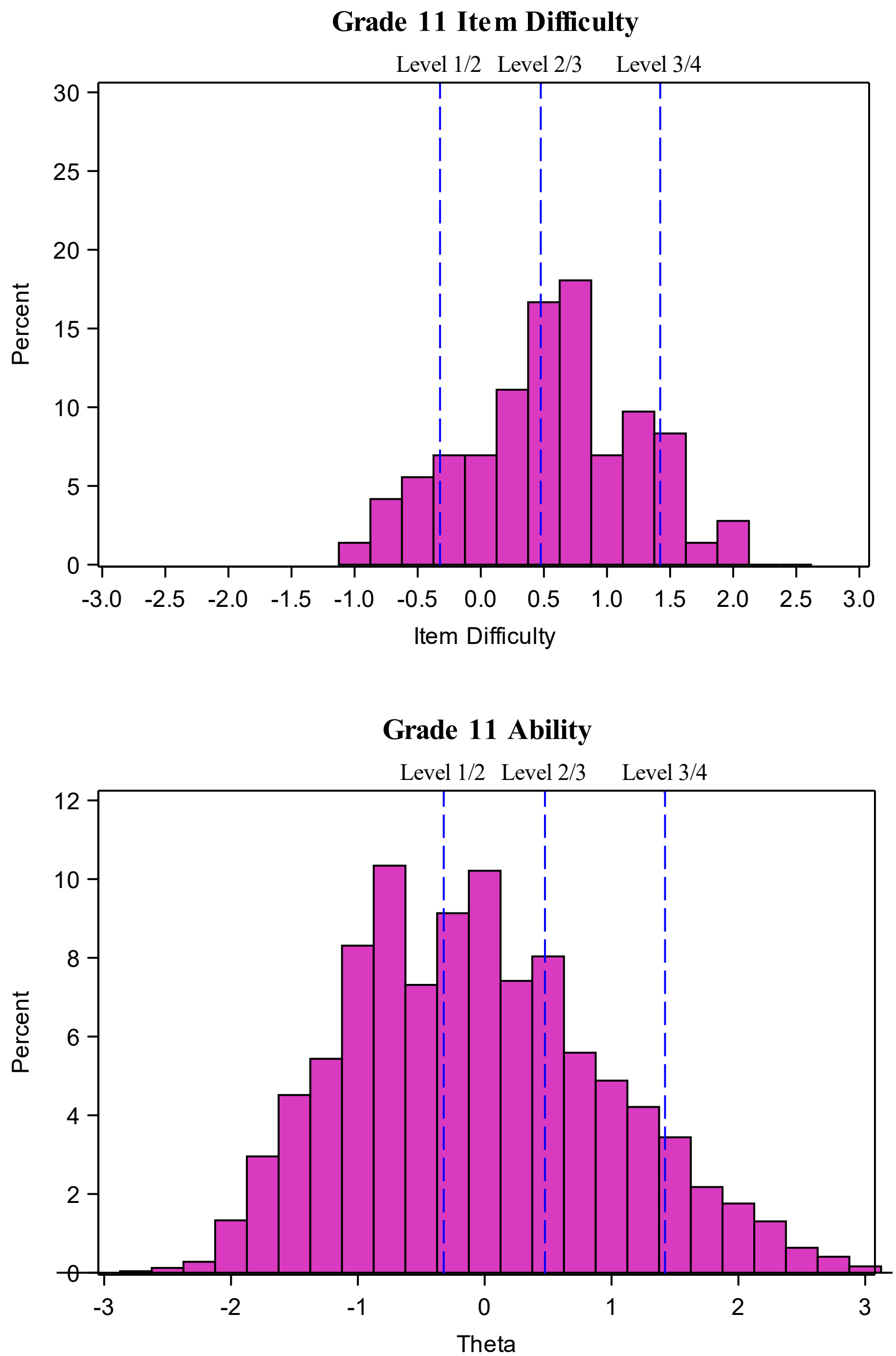


Figure 8.2.9. Grade 11 Item Difficulty and Student Ability Distributions

8.3 Reliability of Proficiency Classifications

The reliability of the proficiency-level classifications was evaluated via two methods. First, error bands were placed around each cut score using the CSEM. Next, the BB-CLASS (Brennan, 2004) program was used to calculate proficiency-level classification consistency indices. The results of both methods indicate that the 2025 NJSLA–S proficiency-level classifications were reliable.

8.3.1 Conditional Standard Error of Measurement at Each Cut Score

As discussed in Section 8.2.2, the conditional standard error of measurements (CSEM) can be computed and evaluated along the theta (θ) scale. Also, the CSEM can be converted and placed on reported scales as needed and appropriate.

The 2025 NJSLA–S cut scores and the corresponding CSEM on the NJSLA–S scales are summarized in Table 8.3.1, and the CSEM tables for all raw and scale scores are presented in Appendix I of this report. The values in Table 8.3.1 have been placed on the same scale as the scale score. At each grade, the cut score with the least amount of error is the level 2/3 cut score (200). At grades 5 and 8, the Level 1/2 cut score's CSEM was slightly higher than that at the Level 3/4. At grade 11, the CSEM at the Level 1/2 cut is approximately the same as the CSEM for the Level 3/4 cut score. Table 8.3.1 also presents error bands that were placed around each of the cut scores to create upper and lower boundaries. The upper and lower bounds were defined by multiplying the cut score's CSEM by two and either adding it to or subtracting it from the cut score. Overlap between the upper or lower bounds of a cut score and one of the other cut scores may indicate reliability issues among the proficiency-level classifications. Overlap between the upper or lower bounds of a cut score and an adjacent cut score may indicate reliability concerns in the proficiency classifications. For 2025, the confidence bounds around adjacent cut scores do not consistently remain distinct across grades. In grades 5 and 8, the Level 1/2 and Level 2/3 bounds are clearly separated, but the Level 2/3 and Level 3/4 bounds overlap. In grade 11, both adjacent pairs of cut scores (Level 1/2 vs. 2/3 and Level 2/3 vs. 3/4) show overlapping confidence bounds, indicating greater uncertainty in the classification boundaries at this grade. The reliability of classification is investigated further with classification consistency indices discussed in Section 8.3.2.

Table 8.3.1: Cut Scores with Conditional Standard Error of Measurement

Grade	Level	Scale Score Cuts	CSEM	Lower Bound	Upper Bound
5	Level 1/2	150	13.2	123.7	176.3
	Level 2/3	200	11.3	177.4	222.6
	Level 3/4	243	12.7	217.5	268.5
8	Level 1/2	150	10.7	128.7	171.3
	Level 2/3	200	9.2	181.6	218.4
	Level 3/4	231	10.0	210.9	251.1
11	Level 1/2	158	13.3	131.4	184.6
	Level 2/3	200	12.5	175.0	225.0
	Level 3/4	250	13.5	223.0	277.0

8.3.2 Classification Consistency Indices

A classification consistency index can be regarded as the percentage of examinees that would hypothetically be assigned to the same proficiency level if the same test was administered a second time or an equivalent test was administered under the same conditions. Cohen's Kappa (Cohen, 1960, 1968) is a statistic that is often used to assess classification consistency. Coefficient Kappa (K) is given by:

$$K = \frac{P_O - P_C}{1 - P_C}, \quad \text{Equation 8.11}$$

where P_o is the probability of a consistent classification and P_c is the probability of a consistent classification by chance. For the NJSLA–S, the classification consistency index for proficiency classifications is an estimate of how reliably the test classifies students into the proficiency categories (i.e., Levels 1–4).

Table 8.3.2 displays the results from BB-CLASS (Brennan, 2004) using the Livingston and Lewis (1995) consistency results. At each grade level, the classification consistency rates (P_o) ranged from .73 to .76. Thus, if the NJSLA–S had been administered a second time, approximately 75% of the students would have been classified at the exact same proficiency level. The most important decision is at the Level 2/3 cut score (200) because it demarcates the point along the scale where students are deemed proficient or not. The decision consistency at the Level 2/3 cut score or above was remarkable at .90 to .91, indicating 90% to 91% probability of being correctly classified as Level 3 or above. The results suggest that the overall NJSLA–S proficiency classification was consistent across grades.

Table 8.3.2: Proficiency-Level Classification Consistency

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut	Kappa	P_0	P_0 for Level 3 or above
5	150	200	243	.61	.73	.90
8	150	200	231	.62	.76	.91
11	158	200	250	.64	.75	.91

8.4 Reliability of Subscore Proficiency Classifications

As discussed in Section 7.3, in each content domain and practice, subscore proficiency is classified as "Below," (Level 1) "Near/Met," (Level 2), or "Above" (Level 3) expectations. The methodology used to create the subscore proficiency-level classifications was dependent on the CSEMs in the raw-to-theta subscore tables. Subscores associated with large CSEMs would indicate unreliable subscore proficiency-level classifications. The complete raw-to-theta subscore tables are presented in Appendix J of this report.

Table 8.4.1 shows that the CSEMs associated with the subscore proficiency cut scores on the theta scale for each content domain and practice by grade were relatively small, indicating reliable subscore classifications. As presented in Table 8.4.1, the classification consistency rates (P_o) were above .69, given the short tests for content domains or practices at each grade level.

Table 8.4.1: Subscore Proficiency Classification Consistency and Conditional Standard Error of Measurement

Grade	Domain/Practice	Kappa	P_o	Level	Subscore Cuts Raw	Subscore Cuts Theta	Subscore Cuts CSEM
5	Earth and Space	.49	.69	Near/Met	8	0.236	0.482
				Above	14	1.697	0.530
	Life	.50	.70	Near/Met	7	0.324	0.461
				Above	13	1.624	0.511
	Physical	.54	.73	Near/Met	7	0.374	0.470
				Above	13	1.592	0.457
	Critiquing	.44	.67	Near/Met	8	0.352	0.442
				Above	14	1.626	0.516
	Investigating	.58	.75	Near/Met	6	0.205	0.509
				Above	13	1.872	0.512
	Sensemaking	.53	.72	Near/Met	8	0.363	0.474
				Above	14	1.651	0.479
8	Earth and Space	.53	.75	Near/Met	9	–0.174	0.435
				Above	16	1.132	0.451
	Life	.52	.75	Near/Met	9	–0.209	0.426
				Above	16	1.059	0.449
	Physical	.54	.76	Near/Met	10	–0.160	0.425
				Above	17	1.136	0.465
	Critiquing	.50	.75	Near/Met	11	–0.083	0.408
				Above	18	1.179	0.464
	Investigating	.54	.77	Near/Met	9	–0.075	0.454
				Above	15	1.164	0.477
	Sensemaking	.57	.79	Near/Met	10	–0.024	0.415
				Above	17	1.191	0.441
11	Earth and Space	.56	.73	Near/Met	11	–0.059	0.407
				Above	18	1.114	0.428
	Life	.58	.74	Near/Met	10	–0.086	0.409
				Above	18	1.263	0.434
	Physical	.55	.73	Near/Met	8	–0.049	0.449
				Above	15	1.232	0.433
	Critiquing	.55	.72	Near/Met	9	–0.096	0.435
				Above	16	1.170	0.434
	Investigating	.57	.74	Near/Met	10	–0.111	0.407
				Above	18	1.265	0.447
	Sensemaking	.57	.75	Near/Met	10	0.011	0.418
				Above	17	1.179	0.417

8.5 Rater Reliability

For constructed-response (CR) items, raters used item-specific scoring rubrics with a score range of 0 to 3 or 0 to 4, depending on the CR item. There were no half points assigned for any of the CR items. Only 10% of the constructed-response items were read by a second independent rater; the purpose of the second read was to investigate the consistency between raters. If the second read score was nonadjacent, then the scores for the response were erased and the paper was rescored. Thus, all scores in the 10% of second reads were either perfect or adjacent agreement.

Table 8.5.1 shows, at the item level, the percentages of constructed-response items scored with exact or adjacent agreement and weighted Kappa. Weighted Kappa is a variation of Cohen's Kappa designed for ordinal variables. As shown in Table 8.5.1, the exact agreement rates ranged from 71.6% to 77.2% for grade 5, from 63.0% to 73.0% for grade 8, and from 70.2% to 80.7% for grade 11. While there was only one CR item that showed weighted Kappa above .90 at each grade level, the other CR items at all grade levels had weighted Kappa estimates above .84. Overall, these results support that rater agreement on the NJSLA–S CR items was acceptable.

Table 8.5.1: Inter-rater Agreement Rate of Constructed-Response Items

Grade	Item	% Raters in Exact Agreement	% Raters in Adjacent Agreement	Weighted Kappa
5	CR 1	77.10	22.90	0.95
	CR 2	77.20	22.80	0.89
	CR 3	71.60	28.40	0.84
8	CR 1	64.10	35.90	0.86
	CR 2	73.00	27.00	0.92
	CR 3	63.00	37.00	0.87
11	CR 1	70.20	29.80	0.88
	CR 2	73.10	26.90	0.86
	CR 3	80.70	19.30	0.91

PART 9: VALIDITY

The *Standards* state that "[v]alidity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (AERA, APA, NCME, p. 14). If there is ample evidence to support reasonable interpretations and test uses, then they are considered to possess high validity (Kane, 2013). Conversely, interpretations and test uses that lack evidence possess low validity. Conceptually, Kane (2006) labeled the process of evaluating that evidence as validation. Test validation is an ongoing, ever-evolving process that extends through the duration of an assessment program. Every component within this Technical Report, from test development to score reporting, is evidence both for and against the valid interpretation and uses of test scores.

The *Standards* categorize validity evidence into five sections:

- evidence based on test content.
- evidence based on response processes.
- evidence based on internal structure.
- evidence based on relation to other variables.
- evidence based on the consequences of testing.

The following sections detail what evidence exists both for and against those five categories of validity evidence. Overall, the evidence suggests that the NJSLA–S fosters valid interpretations and uses of test scores as they pertain to the overall proficiency-level classifications of students.

9.1 Evidence Based on Test Content

Validity evidence based on test content refers to the relevance of the content of the test to the construct that the test is purporting to measure. *Standard* 1.11 states that:

> [w]hen the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. (AERA, APA, NCME, p. 26)

The content-related evidence of validity includes the extent to which the test items represent the specified content domains and cognitive dimensions. Adequacy of the content representation of the NJSLA–S is critical because the tests must provide an indication of student progress toward achieving the KSAs identified in the NJSLS–S, and the tests must fulfill the requirements under ESSA (2015).

Adequate representation of the content domains defined in the NJSLS–S is assured by using a test blueprint and a responsible test construction process as was described in Part 2. The NJSLS–S is taken into consideration in the writing of all NJSLA–S items. In accordance with the test blueprint, the test construction process attempts to balance the six reporting categories and to ensure that the NJSLA–S contains an adequate representation of each content domain and scientific practice. Furthermore, all DCIs, SEPs, and CCCs are represented on the test. Section 2.4 provides a summary of test construction in comparison to the goals established in the test blueprint.

The test content was well-balanced at the content domain level (i.e., Earth and Space, Life, and Physical Science). At each grade level, the content domains were all within five points of being perfectly balanced. The scientific practices (i.e., Investigating, Sensemaking, and Critiquing) were less balanced for grades 8 and 11 and within four points of being perfectly balanced for grade 5. At grade 8, the Critiquing practice was overrepresented, accounting for 10 more points than the Investigating practice and 8 more points than the Sensemaking practice. At grade 11, the Sensemaking practice was underrepresented, accounting for 6 fewer points than both Critiquing and Investigating. At a more granular level all DCIs, SEPs, and CCCs were represented on each grade level's test. The relative balance of the DCIs, SEPs, and CCCs was less impressive, with many categories being either over- or underrepresented. Overall, the content domains and the range of DCIs, SEPs, and CCCs provide evidence that the test is adequately measuring the KSAs defined by the NJSLS–S. However, the relative lack of balance in the scientific practices and individual DCIs, SEPs, and CCCs provides evidence that the scale may be overrepresented by certain components within the NJSLS–S, which could affect interpretations of test scores at both the overall and subscore levels.

9.1.1 Alignment Study

In August of 2022, NJDOE commissioned an independent evaluation of the alignment quality of the NJSLA–S administered at grades 5, 8, and 11. Evidence of alignment quality is critical to validity evaluation for standards-based assessments (Forte, 2017; Webb, 1997, 1999). Such evidence must draw upon an examination of how a test has been designed and developed, as well as instances of the test itself (Forte, 2013). As is the case for all validity evidence, evidence of alignment quality is necessary to support the interpretation and use of test scores. A well-aligned test is one that elicits a sample of student performance that is adequate to support inferences about student achievement in relation to the standards-based domains on which the test is based. To address the unique aspects of the three-dimensional nature of the NJSLS–S and the NJSLA–S items and test forms, the following alignment questions guided the evaluation: (1) To what extent do the blueprints support the consistent creation of test forms that reflect the standards and the score scale? (2) To what extent do the Performance-Level Descriptors (PLDs) reflect meaningful and appropriate score interpretations across the full range of the score scale? (3) To what extent does the set of phenomena, tasks, and items reflect the blueprints and provide performance opportunities across the full range of the score scale?

The results of the study found that the blueprint development was well-documented across all three grades (5, 8, and 11) and included a clear description of the review and revision process by stakeholders. Each blueprint met the criteria of strong evidence of alignment for Domain Concurrence, Balance of Representation, and Phenomena Design. The PLDs for all three grade levels were determined to have strong evidence of alignment with the NJSLS–S and were found to describe increasingly sophisticated and reasonable levels of performance for the concepts defined in the standards. All three test forms met the criteria for strong evidence of alignment with the intended DCI and were judged as strongly representing the multidimensionality of the standards with 100% of items aligning to the additional dimensions of the standards (SEP and CCC). Finally, all test forms met expectations for Domain Concurrence, Range of Knowledge, and Balance of Representation. Further, panelists evaluated the items on the form as being cognitively challenging, though panelists noted that, while a range of cognitive challenge levels is present within the form, items tend to skew toward the higher levels of cognitive challenge, with less representation at the lower levels. The results of this alignment evaluation will be used to inform future item and assessment development activities. The executive summary of the alignment evaluation study is included in Appendix L.

9.2 Evidence Based on Response Processes

Standard 1.12 states that "[i]f the rationale for a test score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided" (AERA, APA, NCME, p. 26). Evidence based on response processes is complementary to evidence based on test content; it can come from several sources including response times, eye tracking, think-aloud protocols, interviews, and/or focus groups. This complementary evidence is different from content evidence because its source is not content experts or teachers but rather the actual student test takers. Padilla and Benitez (2014) noted that "validation studies aimed at obtaining evidence from response processes are scant" (p. 139). The NJSLA–S evidence based on judgment from the NJSAC, content specialists, and a cognitive lab study is described below.

The alignment of each item to the Range PLDs provides limited evidence of the cognitive processes theoretically being assessed by the NJSLA–S. As described in the 2*019 NJSLA–S Technical Report* (NJDOE, 2019), the Range PLDs were created in a collaborative effort by NJDOE, the NJSAC, content specialists, and psychometricians; they are based upon the NJSLS–S content standards. Note that the Range PLDs were not finalized until well after the completion of the item development process for the 2019 NJSLA–S.

The Range PLDs are the theoretical cognitive structure underlying all current NJSLA–S item and test development. They contain detailed descriptions of the knowledge, skills, and abilities (KSAs) that a student needs to display to be classified at a given performance level. Each item on the NJSLA–S was aligned to two Range PLDs: one based on the DCI and one based on the SEP. Those alignments were verified by the NJSAC. The alignment of each item to the Range PLDs offers a theoretical link from the NJSLA–S's underlying cognitive structure to the student responses, which provides limited validity evidence based on response processes. The detailed test maps presented in Appendix F display the Range PLD alignment for each item.

Table 9.2.1 shows the distributions of the performance levels associated with each item by grade level and by DCI and SEP. The DCI distribution of items at grade 5 and the DCI and SEP distributions at grade 11 clustered at Levels 1 and 2, tapering off at Level 3. The grade 5 SEP distribution was clustered at Levels 2 and 3, as was the grade 8 DCI distribution. The grade 8 SEP distribution was more heavily centered at Level 2. These distributions largely correspond to the item difficulty distributions illustrated in Figures 8.2.4 through 8.2.6.

Table 9.2.1: Range PLD Alignment by DCI, SEP, and Grade Level

Grade	Domain/Practice	Level 1	Level 2	Level 3	Level 4
5	DCI	19	22	8	1
	SEP	12	22	16	0
8	DCI	6	27	22	4
	SEP	10	33	10	6
11	DCI	19	29	13	7
	SEP	18	34	15	1

9.2.1 Cognitive Lab Study

To evaluate the degree to which the items and tasks on the NJSLA–S in grades 5, 8, and 11 elicit the intended response processes as represented in the NJSLS for Science, cognitive interviews with students were conducted. The purpose of this study was to gather evidence of the response process. Messick (1995) argued that the substantive validity of test scores relates to the theoretical underpinnings of the construct that is meant to be measured. In the case of statewide, standards-based academic assessments, the construct that is meant to be measured derives from the set of standards in each content area and grade level. The validity evidence necessary to support score interpretation and use includes evidence regarding the alignment of test tasks to the standards in terms of breadth and depth (Webb, 1997; Forte, 2013, 2017) as well as consideration of whether the test tasks elicit the intended cognitive processes as students generate responses to the tasks (Thelk et al., 2006). Items must be developed to elicit those cognitive processes and examined to determine whether, in practice, students' cognitive processing is influenced by variables other than the ones test designers are interested in measuring, which introduces construct-irrelevant variance (Thelk et al., 2009). Two evaluation questions guided the evaluation: (1) To what extent do the tasks on the NJSLA–S tap the intended cognitive processes as represented in the NJSLS for Science? and (2) How do students interact with the task types within the NJSLA–S?

To answer the evaluation questions, cognitive laboratories (often referred to as cog labs) were conducted with 12 students in each grade level across two New Jersey districts (one urban and one suburban) in November 2022. The cog labs used a think-aloud protocol, in which each student outlined their thinking as they worked to answer each item. The study included both a concurrent account of problem-solving and a retrospective cognitive interview. Because the study was conducted in the fall, an off-grade approach was used to ensure that participating students had the opportunity to learn the assessed standards. Students in grades 6, 9, and 12 participated in the study as they received instruction on the assessed standards during the previous school year. Two evaluators observed each student, audio-recorded the session, and independently coded their observations using a standard protocol. The executive summary of the cog-labs study is included in Appendix M.

9.3 Evidence Based on Internal Structure

According to the *Standards,* "[a]nalyses of the internal structure of a test can indicate 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, p. 16). The NJSLA–S was constructed as a unidimensional test. However, it also assesses student performance in several content clusters. It is important to study the pattern of relationships among the content clusters and testing methods. Therefore, this section addresses evidence based on responses and internal structure. Overall, the evidence supports the notion that the internal structure of the NJSLA–S is unidimensional and that its items are measuring the same construct.

9.3.1 Intercorrelations

One method for studying patterns of relationships to provide evidence supporting the inferences made from test scores is to evaluate the correlations between the total test score and its subscores. If the subscores are highly correlated, then that provides evidence that the test is unidimensional. Section 6.2.1.1 of this document summarizes correlation coefficients among test content domains and clusters by grade level. The intercorrelations of the NJSLA–S provide clear evidence that the NJSLA–S is unidimensional. Among the content domain subscores at all grade levels, the lowest correlation was .79 at grade 5 between Physical Science and Earth and Space Science. Among the scientific practices, the intercorrelation ranged from .76 to .95.

9.3.2 Other Internal Structure Evidence

Further evidence of the internal structure of the NJSLA–S was also presented via a principal component analysis (PCA). The PCA results are presented in Section 6.2.1.2. These scree plots show further evidence that the variability in the NJSLA–S test scores is due to a single dimension. No secondary factors at any grade level practically contributed to explaining the variation in the overall NJSLA–S test scores, while subtest scores could convey pedagogical information on a specific content domain or scientific practice.

Part 8 of this Technical Report provides ample evidence to support NJSLA–S reliability. Reliability is the extent to which items within a test measure aspects of a singular construct. Internal consistency reliability (for which the evidence presented in Part 8) is one measure of reliability. The grade-level internal consistency reliability coefficients presented in Section 8.1 were strong, ranging from .91 to .94 for the CBT form. At the subscore level, the reliability coefficients were relatively impressive, with the lowest estimates of .72 for both the grade 5 Critiquing subscore.

9.4 Evidence Based on Relationships to Other Variables

Evidence based on relationships to other variables refers to the degree to which test scores are related to external measures in theoretically expected ways (AERA, APA, & NCME, 2014). Such evidence may be obtained by examining relationships among assessments that measure related constructs, assessments that measure different constructs, or other outcomes that the test is intended to predict. In 2025, NJDOE conducted an internal validity study examining the relationships between NJSLA–S scores and other New Jersey large-scale statewide assessment scores (i.e., NJSLA–ELA, NJSLA–Math, and NJGPA). Across grades 5, 8, and 11, NJSLA–S scores demonstrated moderate to strong positive correlations with ELA and mathematics scores, consistent with theoretical expectations given the shared literacy, reasoning, and analytical demands of science assessments. At the same time, the magnitude and pattern of these relationships, particularly differences across ELA claims and mathematics courses, indicate that the scientific KSAs assessed by the NJSLA–S represent a related but distinct construct within the statewide assessment system.

The results at grade 5 are presented in Table 9.4.1. The ELA assessment comprises two major claims: Reading Complex Text and Writing. The scale scores for these two major claims were included in the correlation matrix. At grade 5, science scores showed correlations of .80 and.82 with ELA (ELA 5) and mathematics (Math 5), respectively. The correlation between science and ELA Reading Complex Text claim was .82, and the correlation between science and ELA Writing claim was .67.

Table 9.4.1: Grade 5 Intercorrelations by Content Area

Content Areas	N	Correlation
SCI 5 - ELA 5	94,898	.80
SCI 5 - ELA 5 Reading	94,898	.82
SCI 5 - ELA 5 Writing	94,898	.67
SCI 5 - MAT 5	96,101	.82
MAT 5 - ELA 5	94,847	.77
MAT 5 - ELA 5 Reading	94,847	.76
MAT 5 - E5 LA Writing	94,847	.68

The results at grade 8 are displayed in Table 9.4.2. The only difference between the grade 8 and grade 5 intercorrelation matrices relates to the mathematics scale scores. Depending on course enrollment, grade 8 students may have taken one of four mathematics assessments: Math 8 (MAT 8), Algebra I (ALG 1), Algebra II (ALG 2), or Geometry (GEO). Accordingly, the grade 8 intercorrelation matrix includes four separate mathematics scale scores rather than a single math score. Because students cannot receive scale scores on more than one mathematics assessment, correlations among the mathematics tests were not included in Table 9.4.2.

The correlation between grade 8 science (SCI 8) and ELA 8 was .78. Correlations between science and the various mathematics assessments ranged from .54 to .76. This range of variability across mathematics assessments is likely attributable to differences in course enrollment, with higher- and lower-achieving students taking different mathematics assessments. Such sorting can reduce scale score variance within each mathematics test, which in turn attenuates the observed correlations. Accordingly, the magnitude of the correlations between science and the mathematics assessments appears reasonable, given that mathematics achievement is more homogeneous within each subgroup than it would be if all students across ability levels took the same assessment.

Table 9.4.2: Grade 8 Intercorrelations by Content Area

Content Areas	N	Correlation
SCI 8 - ELA 8	97,795	.78
SCI 8 - ELA 8 Reading	97,795	.80
SCI 8 - ELA 8 Writing	97,795	.66
SCI 8 - MAT 8	65,892	.71
SCI 8 - ALG I	27,802	.76
SCI 8 - GEO 1	4,679	.70
SCI 8 - ALG 2	538	.54
MAT 8 - ELA 8	64,721	.65
MAT 8 - ELA 8 Reading	64,721	.65
MAT 8 - ELA 8 Writing	64,721	.55
ALG 1 - ELA 8	27,676	.62
ALG 1 - ELA 8 Reading	27,676	.63
ALG1 - ELA 8 Writing	27,676	.49
GEO 1 - ELA 8	4,674	.50
GEO 1 - ELA 8 Reading	4,674	.50
GEO 1 - ELA 8 Writing	4,674	.39
ALG 2 - ELA 8	538	.29
ALG 2 - ELA 8 Reading	538	.32
ALG 2 - ELA 8 Writing	538	.20

The results at grade 11 are presented in Table 9.4.3. The correlation between science (SCI 11) and ELA GPA was .70, while the correlation between science and mathematics was .78. These results are consistent with the relationships between science, mathematics, and ELA observed at other grade levels and align with theoretical expectations. Correlations between science and ELA Reading GPA and ELA Writing GPA were .74 and .58, respectively, reflecting patterns similar to those observed in grades 5 and 8.

Table 9.4.3: Grade 11 Intercorrelations by Content Area

Content Areas	N	Correlations
SCI 11 - ELA GPA	95,785	.70
SCI 11 - ELA Reading GPA	95,785	.74
SCI 11 - ELA Writing GPA	95,785	.58
SCI 11 - MAT GPA	96,054	.78
MAT GPA - ELA GPA	95,635	.70
MAT GPA - ELA Reading GPA	95,635	.73
MAT GPA - ELA Writing GPA	95,635	.59

9.5 Evidence Based on the Consequences of Testing

Standard 1.25 states that "[w]hen unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct" (p. 30). Lane and Stone (2002, p. 24) list the types of evidence that can be collected to evaluate the consequences of a large-scale statewide accountability assessment program.

- Student, teacher, and administrator motivation and effort
- Curriculum and instructional content and strategies
- Content and format of classroom assessments
- Improved learning for all students
- Professional development support
- Use and nature of test preparation activities
- Student, teacher, administrator, and public awareness and beliefs about the assessment and criteria for judging performance and the use of assessment results

No NJSLA–S validity evidence based on the consequences of testing currently exists. Future NJSLA–S validity studies, including evidence based on consequences, are detailed in Section 9.7.3.

9.6 Other Validity Evidence

Each part within this Technical Report contributes evidence relevant to validity. The following is a summary of evidence within each part:

Part 1: Introduction—This part describes the purpose of the assessment, including:

- intended inferences and uses of test scores
- the relationship between the NJSLS–S and NJSLA–S

Part 2: Test Development—This part describes the processes used to design and develop the NJSLA–S, including:

- the steps taken to link test development to the intended inferences and uses of the NJSLA–S
- the training and QC procedures implemented in the item development process
- the use of NJDOE, the NJSAC, and the Sensitivity committee to ensure the work of item writers and content specialists was aligned to the NJSLS–S
- the statistical review of each item after being field tested
- the steps taken to ensure the test construction process matched the NJSLA–S blueprint and statistical constraints

Part 3: Test Administration—This part describes the care that was taken to implement standardized test administration procedures, including:

- documents produced to communicate NJSLA–S test administration procedures for all versions of the test
- steps taken to ensure testing materials were handled using safe and secure procedures
- accommodations and accessibility features that were used during the test administration to provide all NJSLA–S test takers with equal opportunities on the test

Part 4: Scoring—This part describes the procedures that were implemented to verify the accuracy of scoring student responses, including:

- confirming all computer-scored answer keys for both MC and TE item types
- development of unique scoring guides for each CR item
- selecting and training the scorers, Team Leaders, and Scoring Directors charged with handscoring the CR items
- monitoring handscorers to verify they are implementing the scoring rubric accurately
- verifying that student raw scores and subscores were calculated accurately

Part 5: Standard Setting—This part and the *2019 NJSLA–S Technical Report* describe the methods that were undertaken to set the NJSLA–S performance standards, including:

- approval of all NJSLA–S Standard-Setting methods by the NJTAC
- development of performance-level descriptors
- selection of a representative group of New Jersey educators to serve as standard-setting panelists
- evaluation of the standard-setting meeting by the standard-setting panelists
- external review of the standard-setting meeting by an NJTAC member
- documentation of all results in the NJSLA–S Standard-Setting Report

Part 6: Item and Test Statistics—This part describes the battery of statistics that was used to evaluate the NJSLA–S at both the test and item levels, including:

- summaries of item performance across grade level, content domain, scientific practice, and item type to verify that the items were appropriate
- measures of test speededness to assess whether students could finish the test in the allotted time
- confirming the test items were not disadvantaging large subgroups of students via DIF statistics
- descriptive statistics of raw and scale scores by test form and subgroups of students to evaluate how appropriate the test was for portions of the population
- evaluating the IRT assumptions of the PCM to ensure it was appropriate for modeling student ability estimates

- evaluating IRT person-fit statistics by subgroups of students

Part 7: Equating and Scaling—This part describes the methods used to ensure all students at a given grade level received scale scores that were comparable, including:

- documenting the equating and scaling procedures
- descriptions of the special equating(s)

Part 8: Reliability—This part describes the reliability statistics that were calculated to verify the consistency of the NJSLA–S test scores, including:

- verifying the reliability at the total score, form, subscore, item type, and subgroup levels
- evaluating graphic displays of IRT reliability such as TIFs and CSEMs
- assessing the consistency of student performance-level classifications
- assessing rater agreement rates for the handscoring of all CR items

9.7 Summary

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 13). Making an integrated evaluative judgment with such a diverse assortment of evidence is challenging, given that the validity process is ongoing and exists throughout the duration of the testing program. Overall, there is ample evidence that the NJSLA–S fosters valid inferences and uses. However, the NJSLA–S validity argument requires continuing attention, and an iterative process of identifying its weakest components, making modifications, and then reevaluating their effectiveness is needed. As Cronbach (1980) said, "the job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it" (p. 103). The following sections set forth the pros and cons of the NJSLA–S validity evidence by the primary inferences and uses of the test.

9.7.1 Student Performance-Level Classifications: Overall Scale Score

The most important inferences made from the NJSLA–S involve the student performance-level classifications. Students are classified in Levels 1 through 4; students at or above Level 3 are deemed proficient. All interpretations based on NJSLA–S performance-level classifications should be validated for evaluating student performance as it pertains to the KSAs defined in the NJSLS–S.

Overwhelming validity evidence in support of the proposed performance-level classification interpretations has been presented throughout this document and within the validity section. The NJSLA–S was developed and constructed by well-trained experts with assistance from NJDOE and the NJSAC to specifically measure the wide range of KSAs defined in the NJSLS–S. It was administered under strict standardized processes and procedures. The accuracy of the scoring of all NJSLA–S items was verified. The performance-level classifications were determined at standard setting using methodology that was reviewed and approved by the NJTAC. After the test administration the items were statistically evaluated to ensure they met the assumptions of the proposed IRT model. Finally, both the overall scale and the performance-level classifications were verified as being internally consistent.

There are some areas in which the validity evidence in support of the performance-level classification inferences could be improved. The validity section on consequences also has no evidence, which is somewhat expected due to the challenge of integrating consequential validity evidence into a coherent validity argument (Cizek, 2016), as well as to the fact that it is hard to identify the long-term consequences of any testing program after its first year of operational use. Also, the Reporting PLDs would be more useful in providing guidance to test score users if they contained both performance level- and grade-specific KSAs. The current versions are generic for each performance level and do not differentiate among grade-level skills.

Overall, the evidence in favor of the valid interpretations of performance-level classification outweighs the areas in which evidence is lacking or nonexistent. As a standards-based assessment, the content validity evidence linking the test scores and interpretations to the NJSLS–S and the test blueprint are of chief importance (Sireci et al., 2008). Studying the issues noted above would enhance the validity evidence.

9.7.2 Student Performance-Level Classifications: Domains and Practices Subscores

Inferences and uses of subscores are of secondary importance to the overall scale score and performance-level classifications. Student subscores are used to classify their performance as Below Expectations, Near/Met Expectations, or Above Expectations. Students do not receive either a raw score or a scale score in any of the subscore categories. The validity evidence pertaining to interpretations based on NJSLA–S subscore performance-level classifications is limited, and caution in using the subscores should be emphasized.

Some validity evidence in support of the interpretations of subscores is presented throughout this document. Much of the validity evidence supporting the overall scale score—for instance, the test administration and scoring procedures—also contributes to subscore validity evidence. Aside from that, item development, test construction, and PLD creation were all undertaken with the explicit goal of being able to report student performance in the six subscore categories. The subscore performance-level procedures were approved by the NJTAC, and each subscore raw-to-theta score table was independently calibrated and verified by two MI psychometricians. Psychometrically, the subscores displayed adequate reliability coefficients and CSEMs.

Finally, the connection of the NJSLA–S subscores to the NJSLS–S is unclear. The NJSLS–S emphasizes the SEPs, DCIs, and CCCs, whereas the NJSLA–S reports subscore categories back to students, teachers, and administrators in categories that are clusters of SEPs and DCIs. One of the stated goals of the NJSLA–S is to provide feedback to schools on their overall performance in the six subscore categories, but it is not clear how to use or interpret that information within the framework of the NJSLS–S. Constructing links between the NJSLS–S and the reporting categories of the NJSLA–S would improve the ability of teachers, schools, and administrators to use and interpret the information in the subscores.

Overall, the intended inferences being made from the NJSLA–S subscores lack enough validity evidence that any interpretations and uses should be made with caution. NJDOE has sagaciously emphasized caution both in their communications with local education agencies (LEAs) and in the Score Interpretation Guide. Future studies of response processes and factor structures, as well as links from the NJSLS–S to the NJSLA–S reporting categories, could provide insights into how to best interpret and use the subscores; as previously noted in Section 2.4, ongoing, iterative improvements to item development and test construction might alleviate the lack of balance between individual scientific practices and the three content domains.

9.7.3 Future NJSLA–S Validity Studies

As was noted earlier, Kane (2006) labeled the process of evaluating validity evidence as validation, and he conceptualized that process as ongoing, ever evolving, and extending through the duration of an assessment program. NJDOE is committed to addressing the limitations within the NJSLA–S validity evidence and iteratively enhancing the validity of the inferences made from its test scores. One future study is planned, and some details are provided in Section 9.7.3.1.

9.7.3.1 Consequences of the NJSLA–S

Two of the goals of the NJSLA–S are to influence adoption of the NJSLS–S curriculum and to inform instruction, which will in turn improve the educational opportunities for New Jersey students. As described in Section 9.5: Evidence Based on the Consequences of Testing, Lane and Stone (2002) list many possible studies of the consequences of testing programs. They generally involve evaluating whether the testing program is having its intended effect and/or whether it is having unintended consequences. Sources of the data come from students, teachers, administrators, and parents. The future study would likely follow recommendations from Lane and Stone to evaluate the consequences of the NJSLA–S as NJDOE is committed to evaluating the effects of the NJSLA–S.

PART 10: REPORTING

Standard 6.10 states that "[w]hen test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience" (p. 119). The NJSLA–S score reports were designed to effectively communicate test scores while avoiding possible misinterpretations or overinterpretation of the figures. This means that the score reports only show scale scores and performance levels rather than raw scores. This section briefly describes the five different reports that were produced for the NJSLA–S. An example of each of the five reports is explained and presented below. More comprehensive descriptions of each component within the reports can be found in the NJSLA–S Score Interpretation Guide (SIG) at the NJSLA–S website under NJSLA–Science Guides. Two versions of the SIG are publicly available. One version is targeted to educators, administrators, and other district personnel who need to understand the score reports. The other version is targeted to parents and focuses on the Individual Student Reports.

10.1 Individual Student Report

The Individual Student Report (ISR) is a two-sided document intended for use by students, parents, teachers, and other school personnel who have to know a student's strengths and weaknesses in science. It shows the student scale score; the Reporting PLD associated with the student's performance; data for comparison across the state, district, and school; subscore performance levels; and descriptions of the Near/Met Expectations performance level for each subscore. Figures 10.1.1 and 10.1.2 show examples of the front and back of an ISR. A complete list of Reporting PLDs can be found in Appendix E.

FIRSTNAME004 C. LASTNAME004
Spring 2025 Grade: 11
SID: **1200001006** DOB: **04/25/2006**
Local Student Identification: **674316**
NJ SAMPLE DISTRICT NAME
NJ SAMPLE SCHOOL NAME

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

New Jersey Student Learning Assessment - Science (NJSLA-S)
Individual Student Report

This report shows how FIRSTNAME004 performed on the high school science assessment.
This assessment is just one measure of how well your child is performing academically. The results from this assessment should be used in combination with other indicators of achievement in drawing conclusions about your student's performance in science.

Visit the NJ Parent Portal at nj-results.pearsonaccessnext.com and use this code to access your student's results online.

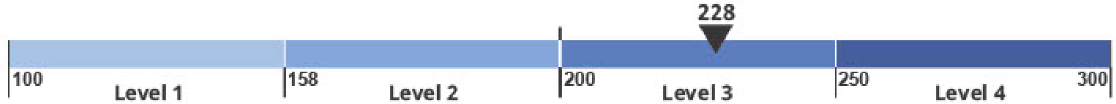
Rz8Nppysqw7T

How did **FIRSTNAME004** perform on the NJSLA-S?

Your student's score: **228**
Performance: **Level 3**

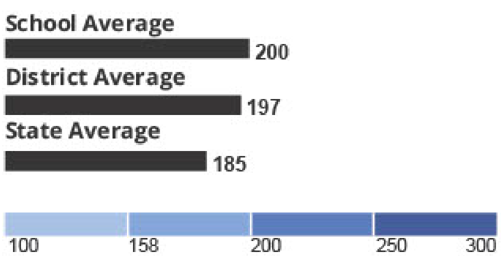
Proficient

Level 4 (250 – 300) Advanced Proficiency
Level 3 (200 – 249) Proficient
Level 2 (158 – 199) Near Proficiency
Level 1 (100 – 157) Below Proficient

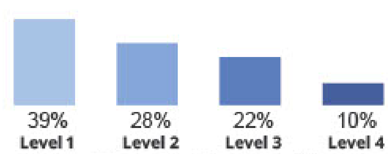


FIRSTNAME004's score on the NJSLA-S indicates that your student is at Level 3.
Students who are at Level 3 demonstrated appropriate grade-level understanding of the New Jersey Student Learning Standards-Science (NJSLS-S) by comprehending information from a variety of sources (e.g., text, charts, graphs, tables) and applying the knowledge gained from scientific investigations to develop accurate explanations and models of observed phenomena. The students often chose and used the appropriate tools to make observations and to gather, classify, and present data. The students used both essential and non-essential information to recognize patterns and relationships between data and designed systems. The students were able to use information to make real-world connections and predictions.

11



How Students Statewide Performed



Percentage of students at each performance level

See page 2 of this report for specific information on your student's performance using the science domains and practices.

Page 1 of 2

Figure 10.1.1. Sample Individual Student Report–Page 1

FIRSTNAME004 C. LASTNAME004

How did your student perform using the domains and practices?

The domains are the content components related to specific disciplines of science.

Earth & Space Science

Your student's performance is **Below Expectations**.

A student designated as Near/Met Expectations demonstrates knowledge of the processes that operate on and within the Earth and also its place in the solar system and galaxy.

Life Science

Your student's performance is **Near/Met Expectations**.

A student designated as Near/Met Expectations demonstrates knowledge of patterns, processes, and relationships of living organisms.

Physical Science

Your student's performance is **Below Expectations**.

A student designated as Near/Met Expectations demonstrates knowledge of the mechanisms of cause and effect in all systems and processes that can be understood through a common set of physical and chemical processes.

The practices are methods by which scientists investigate and build models and theories about the world.

Investigating Practices

Your student's performance is **Above Expectations**.

A student designated as Near/Met Expectations asks questions, plans and carries out investigations based on observations of phenomena, and organizes the data effectively.

Sensemaking Practices

Your student's performance is **Below Expectations**.

A student designated as Near/Met Expectations recognizes patterns and relationships in data to develop explanations or models of the phenomena.

Critiquing Practices

Your student's performance is **Near/Met Expectations.**

A student designated as Near/Met Expectations evaluates and creates arguments regarding different explanations and claims to convey a deeper understanding of the natural world.

LEGEND

Below Expectations | **Near/Met Expectations** | **Above Expectations**

How will my student's school use the test results?
Results from the test give your student's teacher information about their academic performance. The results also give your school and school district important information to make improvements to the education program.

Learn more about the New Jersey Student Learning Assessment — Science
For more information about the assessment, sample questions, practice tests, and the Score Interpretation Guide (SIG) for this report please visit www.measinc.com/nj/science.

Learn More about the New Jersey Learning Standards
Explore your school website, or ask your principal, for information on your school's annual assessment schedule; the curriculum chosen by your district to give students more hands-on learning experiences that meet state standards; and to learn more about how test results contribute to school improvements. You can also learn more about New Jersey's K-12 standards at https://www.nj.gov/education/standards/science/Index.shtml.

Page 2 of 2

Figure 10.1.2. Sample Individual Student Report–Page 2

10.2 Student Label

The NJSLA–S Student Label is designed to be applied to hard-copy files of student records to provide teachers and other school staff with a quick reference to a student's overall performance level. Each label contains student name and identifiers, school information, performance level, and scale score. Figure 10.2.1 depicts a sample Student Label.

LASTNAME, FIRSTNAME M.

NJSLA-Science SPRING 2024 Grade 5

SID: **0123456789** DOB: **02/02/2013**

Local ID: **012345678901234567890123456789** SE: **IEP** ML: **Y**

District: 999999 SAMPLE DISTRICT NAME

School: 999 SAMPLE SCHOOL NAME

Below Proficient **Scale Score: 139**

Figure 10.2.1. Sample Student Label

10.3 Student Roster

The NJSLA–S Student Roster displays student test results within a school. It is used by teachers, other school faculty, and administrators. The Student Roster shows state, district, and school-level average scale scores and the percentages of students with subscore performance that was Near/Met or Above Expectations. At the student level, the roster shows student names, Special Education status, English learner status, scale score, overall performance level, and the performance levels for the subscores. Figure 10.3.1 shows a sample Student Roster.

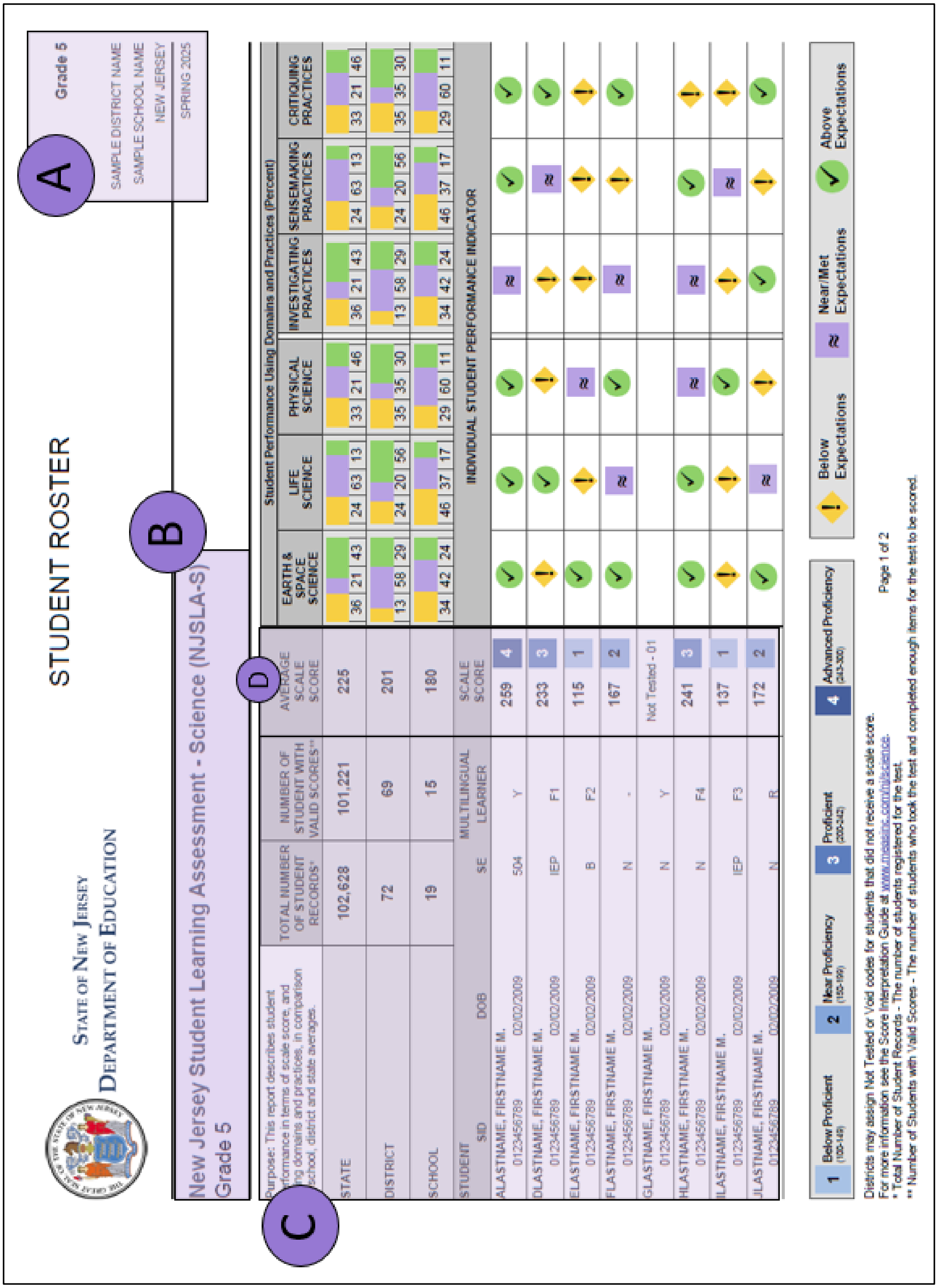


Figure 10.3.1. Sample Student Roster

10.4 School Summary and District Summary of Schools

The NJSLA–S School Summary and District Summary of Schools reports display aggregate student performance at the state, district, and school levels. The School Summary shows only one school while the District Summary of Schools shows all the schools in a district. Other aggregations include gender, ethnicity/race, disability status, and English learner status. Aggregate student performance is illustrated by the percentages of students with each subscore performance level. Figure 10.4.1 displays an example of the School Summary report. Figure 10.4.2 displays an example of the District Summary of Schools report.

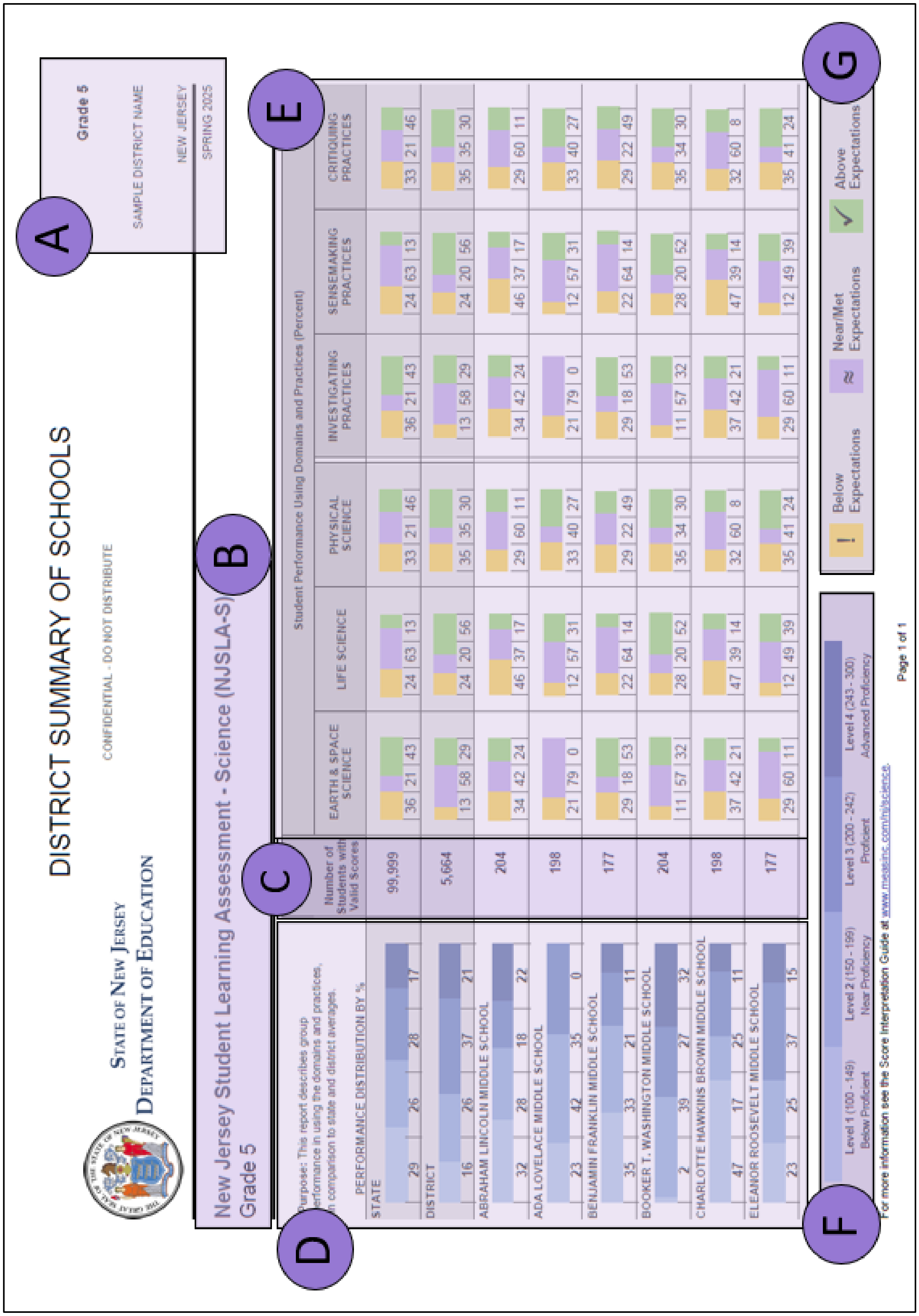
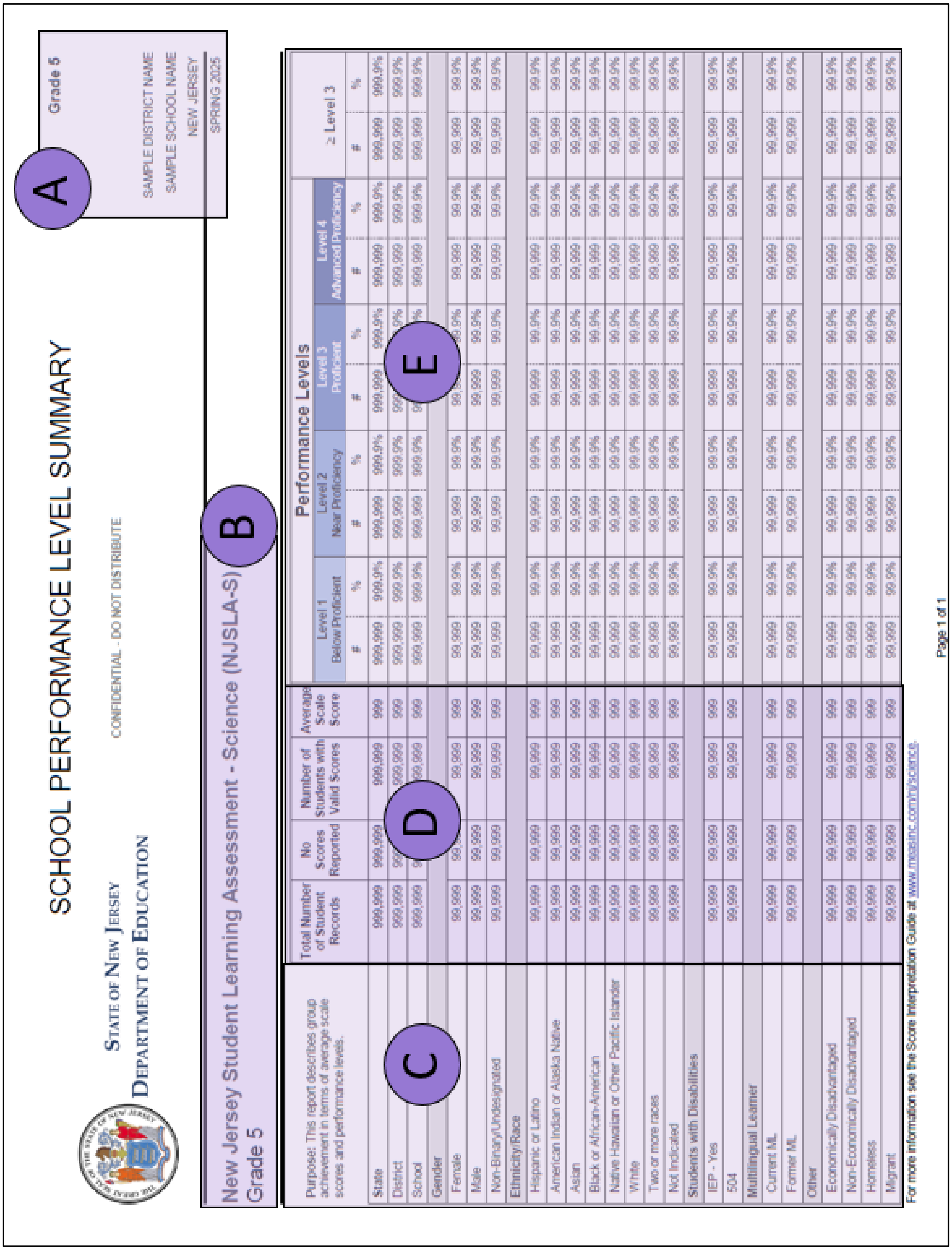


Figure 10.4.1. Sample District Performance-Level Summary Report–Domains and Practices

10.5 School and District Performance-Level Summary Reports

The NJSLA–S School and District Performance-Level Summary reports display aggregate student performance for the state and district. The School Performance-Level Summary also shows student performance at the school level. Other aggregations for the district or school include gender, ethnicity/race, disability status, and English learner status. Aggregate student performance is illustrated by the average scale score and the percentages of students in each performance-level classification. Figures 10.5.1 and 10.5.2 display examples of the School and District Performance-Level Summary reports.



SCHOOL PERFORMANCE LEVEL SUMMARY

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

CONFIDENTIAL - DO NOT DISTRIBUTE

Grade 5

SAMPLE DISTRICT NAME
SAMPLE SCHOOL NAME
NEW JERSEY
SPRING 2025

New Jersey Student Learning Assessment - Science (NJSLA-S)
Grade 5

Purpose: This report describes group achievement in terms of average scale scores and performance levels.	Total Number of Student Records	No Scores Reported	Number of Students with Valid Scores	Average Scale Score	Performance Levels									
					Level 1 Below Proficient		Level 2 Near Proficiency		Level 3 Proficient		Level 4 Advanced Proficiency		≥ Level 3	
					#	%	#	%	#	%	#	%	#	%
State	999,999	999,999	999,999	999	999,999	999.9%	999,999	999.9%	999,999	999.9%	999,999	999.9%	999,999	999.9%
District	999,999	[illegible]	999,999	999	999,999	999.9%	999,999	999.9%	[illegible]	[illegible]	999,999	999.9%	999,999	999.9%
School	999,999	[illegible]	[illegible]	999	999,999	999.9%	999,999	999.9%	[illegible]	[illegible]	999,999	999.9%	999,999	999.9%
Gender														
Female	99,999	[illegible]	99,999	999	99,999	99.9%	99,999	99.9%	[illegible]	[illegible]	99,999	99.9%	99,999	99.9%
Male	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Non-Binary/Undesignated	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Ethnicity/Race														
Hispanic or Latino	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
American Indian or Alaska Native	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Asian	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Black or African-American	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Native Hawaiian or Other Pacific Islander	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
White	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Two or more races	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Not Indicated	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Students with Disabilities														
IEP - Yes	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
504	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Multilingual Learner														
Current ML	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Former ML	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Other														
Economically Disadvantaged	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Non-Economically Disadvantaged	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Homeless	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%
Migrant	99,999	99,999	99,999	999	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%	99,999	99.9%

For more information see the Score Interpretation Guide at www.measinc.com/nj/science.

Page 1 of 1

Figure 10.5.1. Sample School Performance-Level Summary Report

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95–106.

Brennan, R. L. (2004). Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (version 1). *CASMA Research Report 9*. Iowa City, IA.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Cizek, G.J., & Bunch, M.B. (2007). *Standard Setting*. Sage Publications.

Cizek, G.J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy, & Practice, 23*(2), 212–225.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31–44.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

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

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement over a Decade, 5*, 99–108.

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). Lawrence Erlbaum.

Educational Testing Service. (2015). *ETS guidelines for fair tests and communications.* Educational Test Service.

Engelhard, G. & Wang, J. (2021). *Rasch model for solving measurement problems: Invariant measurement in the social sciences*. Sage Publications, Inc.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

Forte, E. (2013). *Re-conceptualizing alignment in the evidence-centered design context* [Paper presentation]. The Annual Meeting of the American Educational Research Association, San Francisco, CA, United States.

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems* [White paper]. Council of Chief State School Officers.

Gorsuch, R. L. (1983). *Factor Analysis*. Lawrence Erlbaum Associates.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). American Council on Education and Macmillan.

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

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

Hambleton, R. K., & van der Linden (1982). Advanced in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*, 373–378.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*, 5–1 1.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–88). Erlbaum

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement (*pp. 17–64). American Council on Education/Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Koffler, S. (2019). *NJSLA–S Cut Score Evaluation*. New Jersey Technical Advisory Committee.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practice.* Springer.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*, 23–30.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J.M. (2016). A *User's Guide to Winsteps MINISTEP Rasch-Model Computer Programs*.

Liu, I-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223–1234.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test* scores. Addison-Welsley.

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

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

McNeill, K.L., Katsh-Singer, R., & Pelletier, P. (2015). Assessing science practices: Moving your class along a continuum. *Science Scope, 39*, 21–28

Messick S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Miller, G., Rotou, O., & Twing, J. (2004). Evaluation of the 0.3 logits screening criterion in common item equating. *Journal of Applied Measurement, 5*, 172–177.

National Research Council (2012). *A framework for K–12 science education: Practices, s and core ideas.* National Academies Press.

New Jersey Department of Education (2019). *New Jersey Student Learning Assessment–Science (NJSLA–S): Technical report grades 5, 8, and 11, 2019.* https://measinc-nj-science.com/sites/default/files/2022-03/2019%20NJSLA-S%20Technical%20Report-Accessible.pdf

New Jersey Department of Education (2025). *Spring 2025 District Test and District Technology Coordinator Training: New Jersey Student Learning Assessments (NJSLA) and New Jersey Graduation Proficiency Assessment (NJGPA)* [Training materials]. https://nj.mypearsonsupport.com/resources/test-administration-resource/Spring%202025_NJSLA_NJGPA_DTC_Training_Part1_01062025.pdf

NGSS Lead States (2013). *Next generation science standards: For states, by states*. The National Academies Press.

Ostini, R., & Nering, M. L. (2010). New perspectives and applications. In M. L. Nering & R. Ostini (Ed.), *Handbook of Polytomous Item Response Models* (pp. 3–20). Routledge.

Padilla, J. L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136–144.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*, 335–355.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 353–370.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295–312.

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., Shea, C. L., Han, K., Deng, N., Delton, J., & Hambleton, R. K. (2008). *Massachusetts Adult Proficiency Tests Technical Manual.* Massachusetts Department of Elementary and Secondary Education.

Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced Assessment Consortium: 2017–18 summative technical report*. Retrieved from https://portal.smarterbalanced.org/library/en/2017-18-summative-assessment-technical-report.pdf/

Swineford, F. (1949). *Law School Admission Test–WLS*. Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1949.tb00915.x

Thelk, A.D., Hoole, E.R., & Lottridge, S.M. (2006). What are you thinking? Postsecondary student think-alouds of scientific and quantitative reasoning tasks. *The Journal of General Education, 55*(1), 17–39.

Thelk, A., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). *Examining inferences about test-taking motivation: The Student Opinion Scale (SOS)* [Paper presentation]. Annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT, United States.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (Synthesis Report 44). National Center on Educational Outcomes.

Traub, R. E., & Rowley, G. L. (1991). NCME instructional module: Understanding reliability. *Educational Measurement: Issues and Practice, 10*(1), 37–45.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research monograph No. 6). Council of Chief State School Officers.

Webb, N.L. (1999). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Council of Chief State School Officers.

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

Wright B.D., & Linacre J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright B.D., & Masters G.N. (1982). *Rating Scale Analysis.* MESA Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Erlbaum.

APPENDIX A: Glossary of Abbreviations

Table A.1: Glossary of NJSLA–S Abbreviations

Abbreviation	Definition
ABBI	Assessment Banking for Building Interoperability
AERA	American Educational Research Association
AF&A	Accessibility Features and Accommodations
APA	American Psychological Association
ASL	American Sign Language
CBT	Computer-Based Test
CCC	Crosscutting Concept
CR	Constructed-Response
CSEM	Conditional Standard Error of Measurement
CTT	Classical Test Theory
DCI	Disciplinary Core Idea
DIF	Differential Item Functioning
DTC	District Test Coordinator
EconDis	Economically Disadvantaged
EL	English Learner
ESEA	Elementary and Secondary Education Act
ESSA	Every Student Succeeds Act
ICC	Item Characteristic Curve
IIF	Item Information Function
IRT	Item Response Theory
ISR	Individual Student Report
KIS	Key Information Sheet
KSA	Knowledge, Skills, and Abilities
LEA	Local Education Agency
MC	Multiple-Choice
MH	Mantel-Haenszel
MI	Measurement Incorporated
MSA	Machine-Scorable Assessment
NBP	National Braille Press
NCME	National Council on Measurement in Education
NJBCT	New Jersey Biology Competency Test
NJBSC	New Jersey Bias and Sensitivity Committee
NJDOE	New Jersey Department of Education
NJSAC	New Jersey Science Advisory Committee
NJTAC	New Jersey Technical Advisory Committee
NJSLA–S	New Jersey Student Learning Assessment–Science
NJSLS–S	New Jersey Student Learning Standards for Science
NRC	National Research Council

Abbreviation	Definition
OIB	Ordered Item Booklet
OPLS	Online Performance-Level Setting
PAN	PearsonAccess^{next}
PBA	Performance-Based Assessment
PBS	Phenomenon-Based Scenario
PBT	Paper-Based Test
PCA	Principal Components Analysis
PCM	Partial Credit Model
PIA	Preliminary Item Analysis
PLD	Performance-Level Descriptor
rpb	Item-Total Correlation
SEM	Standard Error of Measurement
SEP	Science and Engineering Practice
SIG	Score Interpretation Guide
SRF	Summative Record File
STC	School Test Coordinator
SWD	Students with Disabilities
TA	Test Administrator
TCM	*Test Coordinator Manual*
TE	Technology-enhanced
TIF	Test Information Function
TLC	Teneo Linguistics Company
TTS	Text-to-Speech

APPENDIX B: New Jersey Science Advisory and Bias and Sensitivity Committees–District and County Representation

Table B.1: Grade 5 NJSAC District and County Representation

Number	District	County
1	Saddle River School District	Bergen
2	Northern Burlington County Regional Schools District	Burlington
3	Lindenwold Public Schools	Camden
4	Avalon School District	Cape May
5	Orange School District	Essex
6	Swedesboro-Woolwich School District	Gloucester
7	Readington Township School District	Hunterdon
8	Alexandria Township School District	Hunterdon
9	West-Windsor-Plainsboro Regional School District	Mercer
10	Metuchen Public School District	Middlesex
11	Freehold Township Public Schools	Monmouth
12	Rumson School District	Monmouth
13	Roselle Public Schools	Union

Table B.2: Grade 8 NJSAC District and County Representation

Number	District	County
1	Oakland K-8 Public Schools	Bergen
2	Lyndhurst Public Schools	Bergen
3	Leonia Public School District	Bergen
4	Cinnaminson Township Public Schools	Burlington
6	West Orange Public Schools	Essex
7	Greenwich Township School District	Gloucester
8	West Windsor-Plainsboro Regional Schools	Mercer
9	East Brunswick Township	Middlesex
10	Rumson Borough Elementary School District	Monmouth
12	Matawan Aberdeen Regional School District	Monmouth
13	Mount Olive Township Schools	Morris
15	Passaic City Schools	Passaic
16	Somerville Public Schools	Somerset
17	Plainfield Public School District	Union
18	Roselle Park Public Schools	Union
19	Roselle Public Schools	Union

Table B.3: Grade 11 NJSAC District and County Representation

Number	District	County
1	Atlantic County Vocational School District	Atlantic
2	Park Ridge School District	Bergen
3	Wood-Ridge School District	Bergen
5	Woodbury City Public Schools	Gloucester
6	Jersey City Public Schools	Hudson
7	Delaware Valley Regional High School District	Hunterdon
8	East Windsor Regional School District	Mercer
9	Ewing Township School District	Mercer
10	Piscataway Township School District	Middlesex
11	West Windsor-Plainsboro Regional School District	Middlesex
12	Mount Olive Township School District	Morris
13	Point Pleasant Borough School District	Ocean
14	Passaic City Schools	Passaic
15	Patterson Public Schools	Passaic
16	Pennsville School District	Salem
17	Somerset County Vocational & Technical School District	Somerset
18	Lenape Valley Regional School District	Sussex
19	Elizabeth Public Schools	Union
20	Union County Vocational-Technical School District	Union

Table B.4: NJBSC District and County Representation

Number	District	County
1	Oakland Public School District	Bergen
2	Pemberton Township School District	Burlington
3	Cherry Hill Public Schools	Camden
4	Newark Public Schools	Essex
5	Middlesex Borough School District	Middlesex
6	Freehold Township School District	Monmouth
7	Long Branch Public School District	Monmouth
8	Morris County Vocational School District	Morris
9	Clifton Public Schools	Passaic
10	Paterson Public Schools	Passaic

APPENDIX C: Statistical Review Reference Sheet

P-VALUE (All items: 1-point to 4-point items)	• A measure of item difficulty based on classical test theory • Proportion correct; the proportion of students who answered a dichotomous (1-point) item correctly • Percentage of maximum score point; item mean divided by the highest attainable score point for a polytomous (2- or 4-point) item • *P*-values can range from 0 to 1.
**FLAGGED IF:*	*P < .25 (too hard)* *P > .90 (too easy)*

RASCH VALUE (All items: 1-point to 4-point items)	• A measure of item difficulty based on item response theory, with values generally ranging from –3 to +3. Higher values indicate greater difficulty (reverse of *p*-value). • Items with Rasch values targeted at cut scores for performance categories are especially useful for measurement.

SCORE POINT DISTRIBUTION (2-/4-point items)	• Percentage of responses at each score point • If any score point has fewer than 10% of responses (2-point item) or 5% of responses (4-point item), the score point is not measuring relevant ability effectively.
**FLAGGED IF:*	*Response percentage < 10% at any score point (2-point items)* *Response percentage < 5% at any score point (4-point items)*

ITEM-TOTAL CORRELATION (All items: 1-point to 4--point items)	• A measure of the degree to which an item discriminates between those students who know the material (using total test score as a proxy for that knowledge) and those who do not • *rpb*: correlation between an item and the total test score • *rpb* can range from –1 to +1.
**FLAGGED IF:*	*rpb < .20 (All items, 1-point to 4-point items)*

DIF CATEGORY (All items: 1-point to 4-point items)	• A statistical procedure for detecting potential item bias • Differential item functioning (DIF) categorization looks at the extent to which an item performs differently across different groups—Male/Female, White/Black, White/Hispanic, and White/Asian—controlling for the groups' ability (using total test score as a proxy). • Each item is classified as A, B, or C: ➢ A: Item displays negligible DIF; does not need review for bias. ➢ B: Item displays moderate DIF; needs review for bias. ➢ C: Item displays severe DIF; needs *careful* review for bias.
**FLAGGED IF:*	*DIF CATEGORY = B or C*

APPENDIX D: 2019 NJSLA–S Standard Setting: Executive Summary

Appendix D contains the executive summary from the standard-setting report submitted by MI in 2019. The standard-setting study is summarized in greater detail in the *2019 NJSLA–S Technical Report* (NJDOE, 2019).

The New Jersey Student Learning Assessment–Science (NJSLA–S) is the assessment battery New Jersey uses to satisfy reporting requirements for the Every Student Succeeds Act (ESSA; P.L. 115–94) for science in grades 5, 8, and 11.

The New Jersey Department of Education (NJDOE) conducted standard setting for science tests in grades 5, 8, and 11 during the week of July 23–25, 2019. Educators from throughout the state of New Jersey participated in this three-day meeting. Staff of Measurement Incorporated (MI), the contractor, and Pearson Education, its subcontractor, facilitated the meeting.

The main goals of the meeting were to

1. allow workshop participants (panelists) to gain an understanding of the test contents and performance-level descriptors (PLDs),
2. learn a standard-setting procedure known as the Bookmark procedure, and
3. have panelists recommended cut scores for each test that differentiate Level 1 from Level 2, Level 2 from Level 3, and Level 3 from Level 4 performance (i.e., three cut scores to yield four performance levels).

These recommendations are designed to help inform the New Jersey State Board of Education (Board) as it completes its task of establishing performance standards for these assessments.

From July 23 through July 25, 2019, MI/Pearson staff met with representatives of NJDOE and 39 educator-panelists from around the state to recommend performance standards on the three tests.

Process and Procedures

The panelists, nominated by district superintendents, were chosen specifically to represent the demographics and geographic distribution of educators throughout the state. A profile of the 39 panelists is provided in the *2019 NJSLA–S Technical Report* (NJDOE, 2019). Panelists spent the entire first day examining the tests and PLDs under the direction of NJDOE and MI staff. On the second day, following an introduction to the Bookmark standard-setting procedure, the panelists separated into their respective grade-level groups, each led by two facilitators (one psychometrician and one content specialist) from MI/Pearson. Panelists in all groups received a thorough orientation to the standard-setting software and practice exercises to prepare them for their standard-setting task. MI staff provided additional information to panelists as they proceeded through three rounds of recommending cut scores, discussing decisions, and settling on final recommendations.

In accordance with a plan previously approved by NJDOE, MI employed the Bookmark procedure. This procedure is the most widely used standard-setting procedure for statewide assessments and is thoroughly documented in the approved plan and elsewhere (cf. Cizek & Bunch, 2007). In this procedure, panelists review all test items in a specially formatted test booklet (ordered item booklet, or OIB) that places the easiest item on page one, the most difficult item on the final page, and all items in between ordered by difficulty, based on actual student responses. Using threshold PLDs developed previously by NJDOE (with the assistance of New Jersey educators), panelists place a bookmark at the point in the test booklet where they believe the probability of a student at the threshold of Level 2, Level 3, or Level 4 would begin to have less than a two-thirds chance of answering correctly. These page numbers are then mathematically translated into raw cut scores. The average (median) of the panelists' bookmarked pages becomes the group bookmark, and the associated raw score becomes the cut score for that level for that grade for that round. The procedure is more fully described in Chapter 1 of the report. All reviews were completed within software created by MI and used previously for several other successful standard-setting activities.

Panelists considered each test in three online rounds. During Round 1, each panelist placed three bookmarks, one for Level 2, one for Level 3, and one for Level 4. MI staff analyzed the data for Round 1 and led discussions of the results: difficulties encountered, dispersion of bookmarks for each level, reasons for those dispersions, rationales for individual bookmark placements, and differences in interpretation of the PLDs.

After discussion of Round 1 results, panelists then started Round 2, repeating the process of placing bookmarks as they had in Round 1. After Round 2, MI staff again analyzed the data and presented results to the panelists, along with score distributions showing percentages of students who would be classified at each level on the basis of the Round 2 cut scores (impact data).

After discussion of Round 2 results and impact data, panelists once again placed three bookmarks in Round 3. These bookmarks defined the final cut scores (averaged over all panelists in a given group) to be forwarded to NJDOE. Facilitators then presented Round 3 results to panelists and gave them an opportunity to evaluate the process and outcomes. One panelist in grade 11 had to leave after Round 2.

Results

Final recommended performance standards are reported in Table ES-1. The cut scores include both the raw score associated with the median bookmark and that score expressed in terms of a percentage of the total points possible. The final column in Table ES-1 shows the total number of points possible for each test. There were no cross-grade discussions of cut scores.

Table ES.1: Final Recommendations from Standard-Setting Panelists

Grade	Level	Total Points	Raw Cut Score	Cut Score % Correct
Grade 5	Level 2	60	25	42%
Grade 5	Level 3	60	39	65%
Grade 5	Level 4	60	49	82%
Grade 8	Level 2	70	20	29%
Grade 8	Level 3	70	40	57%
Grade 8	Level 4	70	52	74%
Grade 11	Level 2	78	31	40%
Grade 11	Level 3	78	45	58%
Grade 11	Level 4	78	60	77%

The impact of these cut scores on New Jersey students is summarized in Figure ES-1. Overall, 26.3% of grade 5 students, 17.6% of grade 8 students, and 26.5% of grade 11 students scored at or above Level 3. The numbers of students upon which these percentages are based are not the entire population. By prior agreement between NJDOE and MI, the available data was analyzed during the week prior to standard setting: 64,419 fifth graders, 88,295 eighth graders, and 76,001 eleventh graders. It should be noted that special care was taken to make sure these data were representative of the entire state. Thus, when all of the data are analyzed, it is possible that the percentages in each category could change slightly.

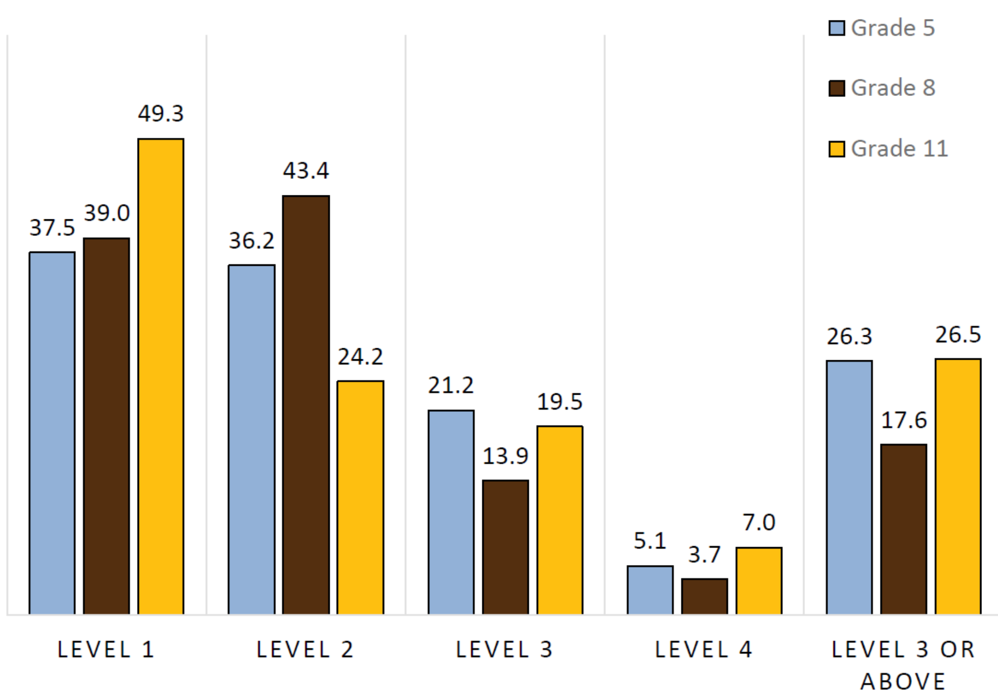


Figure ES.1. Percentages of Students Classified at Each Level after Round 3

Impact of impact data. From Round 2 to Round 3, there was some movement (in both directions) in cut scores. In grade 5, the Level 2 cut score actually went up by 1 raw score point. At grade 8, the Level 2 cut score went down by 7 raw score points (a difference of two pages in the OIB), but the cut scores for Levels 3 and 4 did not change. One grade 8 panelist commented on the back of the evaluation form that anticipated pressure from local school administrators may have caused some panelists to lower their cut scores for Level 2. Yet, there was no change in the Level 3 or Level 4 cut scores for grade 8. At grade 11, the Level 2 and Level 3 raw cut scores went down by 4 and 2 points, respectively; the Level 4 cut score was unchanged from Round 2 to Round 3.

Evaluation of process and outcomes. The panelists were given an opportunity after presentation of Round 3 results to evaluate the entire process and outcomes. Those conducting the standard setting were especially interested in knowing how reasonable they found the final cut scores to be. Their responses to key statements on the evaluation form are summarized in Table ES-2.

Table ES.2: Responses to Key Evaluation Questions

[Responses: Grade 5–14; Grade 8–12; Grade 11–12]

Statement	% Strongly Disagree			% Disagree			% Uncertain			% Agree			% Strongly Agree		
	5	8	11	5	8	11	5	8	11	5	8	11	5	8	11
The process was fair.	0	0	0	0	0	0	0	0	8	7	17	33	93	83	58
The process was orderly.	0	0	0	0	0	0	0	0	0	7	17	33	93	83	67
My group's final cut score for Level 2 is reasonable.	0	0	0	0	0	0	0	8	0	14	8	50	86	83	50
My group's final cut score for Level 3 is reasonable.	0	0	0	0	0	0	0	0	0	14	17	25	86	83	75
My group's final cut score for Level 4 is reasonable.	0	0	0	0	0	0	0	0	0	21	8	25	79	92	75

These last three statements had a follow-up direction: If you disagree, should it have been higher or lower? Circle one.

Panelists were also encouraged to enter comments on the back of the form, particularly if they disagreed with the reasonableness of any of the cut scores. The open-ended responses to the reasonableness items are summarized in Table ES-3.

Table ES.3: Summary of Reasonableness Ratings and Comments

Statement	Grade 5	Grade 8	Grade 11
My group's final cut score for Level 2 is reasonable.	No objections; no recommended changes	No objections; one suggestion that impacts data skew Round 3 cuts	No objections; no recommended changes
My group's final cut score for Level 3 is reasonable.	No objections; no recommended changes	No objections; no recommended changes	No objections; no recommended changes
My group's final cut score for Level 4 is reasonable.	No objections; one recommendation to raise cut by 1	No objections; no recommended changes	No objections; no recommended changes

Summary and Recommendations

The standard setting for NJSLA–S was conducted in strict accordance with the approved plan. Panelists understood the process well, as indicated by their responses to the evaluation form. The standard-setting process for NJSLA–S was sound, both in conception and execution, representative of the highest standards in contemporary educational measurement, and representative of standards operating among state assessment programs nationwide. The cut scores produced after three rounds of test review reflect well the PLDs panelists used to complete the standard-setting task. It is proposed that the cut score recommendations presented here be given strong consideration for approval.

APPENDIX E: NJSLA–S Performance-Level Descriptors

E.1 Policy PLDs

NJSLA–S Policy-Level Performance-Level Descriptors

Level 1	Level 2	Level 3	Level 4
Level 1 students demonstrate minimal understanding of the disciplinary concepts and have difficulty applying the scientific practices. They may have significant difficulty engaging in public discussion on scientific topics and discerning valid and reliable scientific technological information related to their everyday lives even with focused effort achieving minimal success.	Level 2 students demonstrate partial understanding of the disciplinary concepts and performance with the scientific practices. They may have difficulty engaging in public discussion on scientific topics and discerning valid and reliable scientific technological information related to their everyday lives without the focused effort needed to achieve some success.	Level 3 students demonstrate appropriate grade-level understanding of the disciplinary concepts and performance with the scientific practices. They can likely engage in public discussion on scientific topics and discern valid and reliable scientific technological information related to their everyday lives with some success.	Level 4 students demonstrate a deep understanding of the disciplinary concepts and superior performance with the scientific practices. They can likely engage in public discussions on scientific topics and discern valid and reliable scientific and technological information related to their everyday lives with a high degree of success.

E.2 Threshold PLDs

E.2.1 Grade 5 Threshold PLDs

The Threshold Performance-Level Descriptors (PLDs) define the minimum knowledge, skills, and practices that students must display for each Disciplinary Core Idea and Science and Engineering Practice to reach a certain performance level. They expand upon the brief overall PLDs included in the Score Interpretation Guide.

Grade 5 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS1: Matter and Its Interactions	• that matter is made of particles that can be identified by their properties and that weight does not change during visible physical changes • that the properties of substances may change when combined, but the total weight will stay the same	• that matter is made of particles with unique, measurable properties that are conserved when changing state • that a change to a substance(s) may or may not result in one or more new substances, but the total weight will remain the same	• of distinguishing properties of matter and the relationship between visible and non-visible matter • that the outcome of the combination of one or more substances is predictable based on the properties of the substances
PS2: Motion and Stability: Forces and Interactions	• that objects are acted upon by forces that can cause predictable patterns of motion • that the size of a force, the properties of objects, and the position of the objects relative to one another have an effect on their interaction	• that an object's motion is a product of the net force acting on the object and can therefore cause predictable motion • of how certain relationships among the interactions between objects are interconnected and can explain how the objects ultimately affect each other	• of the relationship between net force and motion of an object in predicting future movement • that the relationships between the interactions and the properties of objects are dependent upon systems in which the objects exist

Grade 5 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS3: Energy	• that differences in the movement of energy can cause objects to move at different speeds • that energy in various forms can be transferred from place to place • that energy is transferred when objects collide • that energy can be converted into forms for practical use	• that energy can move from place to place in different forms with varying levels of magnitude • that effects of transferred energy are observable • of the relationship between the transfer of energy and the change in motion when objects collide • that there is a relationship between energy and its conversion for practical uses	• that predictions can be made regarding the interactions of objects based on the amount of energy the objects possess • of the transformation from one type of energy to other type(s) of energy • that when objects collide, there are predictable outcomes • that stored energy is converted energy from the Sun

Grade 5 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS4: Waves and Their Applications in Technologies for Information Transfer	• that there are similarities and differences in the patterns of waves • that in order for an object to be seen, light must reflect off the object • that information can be transmitted over long distances using communication methods/devices	• that the characteristics of a wave determine the net motion of the wave • that there exists a relationship among the path of light, light reflection, and the visibility of objects • of how different communication methods/devices operate	• of how changing the amount of energy can change the characteristics of a wave • that a change in the path of light or light reflection will cause a change in the visibility of an object • of the advantages of different communication methods/devices and how those devices transmit digitized information over long distances

Grade 5 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS1: From Molecules to Organisms: Structures and Processes	• of the internal or external structures of plants or animals and their functions • that animals or plants reproduce and have life cycles • that both animals and plants take in materials to survive • that animals have sense receptors that they use to guide their actions	• of internal and external structures of plants and animals and how their functions support survival, growth, behavior, or reproduction • that animals and plants reproduce for continued existence and have life cycles that are unique but have some similarities • of the relationship between plants and animals and the materials they take in for specific various functions • that an animal's brain processes information received from specialized sense receptors that they use to guide their actions	• of the variation and function of internal and external structures across the plant and animal kingdoms • of the relationships among the components of life cycles • that animals and plants acquire energy from different sources but use the energy for similar functions • that animals respond to environmental changes using sensory information and stored memories

Grade 5 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS2: Ecosystems: Interactions, Energy, and Dynamics	• that in a food web, all organisms have a role OR • of the requirements of a healthy ecosystem • that materials cycle through an environment • that organisms respond to changes in their environment • that living in groups helps animals	• that organisms have different roles in a food web, with a focus on the cycling of materials • that the health and stability of an ecosystem depend on the overall biodiversity and the availability of resources • of how materials cycle through multiple components of an environment • of organisms responding to changes in their environment • that living in specialized groups helps animals, depending on the situation	• that the materials that animals consume can be traced through multiple levels of the food web back to plants • that the balance of the flow of matter can be disrupted by changes in the ecosystem • of the impact of change on the cycling of matter in a system • of how changes in an environment affect multiple organisms • that the dynamics of a group can change over time
LS3: Heredity: Inheritance and Variation of Traits	• that traits and characteristics are based on both inheritance and environmental factors • that organisms have variations in traits	• that while there are similarities in traits between siblings, they each have characteristics that are influenced by the environment • that some traits are inherited in a predictable way while others may be influenced by the environment	• that environmental factors affect traits or functions • that patterns in traits are expressed over multiple generations • that traits, whether inherited or influenced by the environment, have some similarities and some differences

Grade 5 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS4: Biological Evolution: Unity and Diversity	• that fossils are evidence of plant and animal life long ago • that variations among organisms help them survive and reproduce • that some organisms can survive in a particular environment while others cannot • that plants and animals are affected by change in their habitat	• that fossils are evidence of varying environments • that certain characteristics are advantageous to the survival of a species • that an environment must meet the needs of an organism for survival • that plants and animals may adapt to changes in their environment	• that fossils are evidence of changing environments over time • that specific variation in a characteristic can influence an organism's survival • that changes in an environment affect an organism's ability to survive • that the effects of habitat change may cause adaptation to occur

Grade 5 Threshold Performance-Level Descriptors (Earth and Space Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
ESS1: Earth's Place in the Universe	• that the Sun is an object in the sky and gives off light • that Earth is a rotating body in relative position to the Sun • that Earth's rotation affects day and night • that there are observable patterns in moon phases, shadows, and star patterns • that patterns of rock formations can contain fossils and can change due to Earth forces	• that distance affects relative size • of changes in patterns (daylight hours, shadow length, stars, moon phases) that can be observed during day and night as Earth rotates and orbits around the Sun • that fossil records can help identify rock layer formations because of changes caused by natural processes	• that relative distance affects brightness • that Earth's orbit and rotation at different times of day and year, together with the orbit of the Moon and position of the Sun, create patterns that affect how humans view objects from Earth • that a geological history can be determined by examining rock layers and fossil records
ESS2: Earth's Systems	• that Earth's four major systems can interact with each other and that components of the systems can change • that maps can be used to locate Earth's features and processes • that Earth has oceans and areas of freshwater • that weather conditions in different areas change over time • that organisms affect the environment	• of how specific processes change components of Earth's four major systems and, in turn, have an effect on the systems themselves • that maps can be used to determine patterns of Earth's features and processes • of the distribution of water on Earth and its availability and accessibility • that patterns of weather form the basis of climate data • of how organisms affect the environment	• of patterns of processes affecting Earth's four major systems and how changes in those processes will likely affect the components of those systems • that the locations of Earth's features are related to geologic changes • that the water cycle affects the distribution of water on Earth • that climatic patterns can be used to predict future weather conditions of an area • that behavior of organisms in an environment can help predict changes to the physical characteristics of that environment

Grade 5 Threshold Performance-Level Descriptors (Earth and Space Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
ESS3: Earth and Human Activity	• that humans use both renewable and nonrenewable resources for fuel and energy and that such use can affect the environment • that humans can identify different types of natural hazards • that humans have different effects on the environment or its resources	• that using fuel from natural sources can be positive and negative in multiple ways • that Earth's processes create unavoidable hazards and that humans have an important role in designing solutions to reduce negative impact • that individuals and communities can protect and reduce the negative effects that human activities can have on the environment	• that humans have to make informed decisions about which natural resources to use by analyzing their risks and benefits • that there are benefits and risks to human-created solutions designed to lessen the impact of natural hazards • that humans have to make informed decisions based on the positive and negative effects of their activities in an effort to protect Earth

Grade 5 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
Asking Questions (for Science) and Defining Problems (for Engineering) (AQDP): A practice of science is to ask and refine questions that lead to descriptions and explanations of how the natural and designed world works and which can be empirically tested.	• identify or ask relevant questions that are testable and that can show cause-and-effect relationships in the natural or designed world	• identify or ask relevant questions that can be investigated • describe problems that can be solved • predict reasonable outcomes • clarify and redesign a solution to a problem	• generate questions based on investigations incorporating variables to determine patterns while defining and solving a design problem
Developing and Using Models (DUM): A practice of both science and engineering is to use and construct models as helpful tools for representing ideas and explanations. These tools include diagrams, drawings, physical replicas, mathematical representations, analogies, and computer simulations.	• describe or use a model to show the relationship among components in a phenomenon	• develop or refine a model to minimize limitations, or test cause-and-effect relationships	• evaluate and revise or develop models to show relationships in cause-and-effect systems
Planning and Carrying Out Investigations (PACI): Scientists and engineers plan and carry out investigations in the field or laboratory, working collaboratively as well as individually. Their investigations are systematic and require clarifying what counts as data and identifying variables or parameters.	• plan an investigation and collect observational data using appropriate methods or tools that help identify outcomes from changing a variable	• plan or conduct an investigation by evaluating appropriate methods or tools for collecting data while making predictions about a fair test in which variables are controlled	• plan and conduct multiple trials of an investigation to produce data that can be compared to make predictions, to serve as evidence for an explanation of a phenomenon, or to test a design solution

Grade 5 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
Analyzing and Interpreting Data (AID): Scientific investigations produce data that must be analyzed in order to derive meaning. Because data patterns and trends are not always obvious, scientists use a range of tools—including tabulation, graphical interpretation, visualization, and statistical analysis—to identify the significant features and patterns in the data. Scientists identify sources of error in the investigations and calculate the degree of certainty in the results. Modern technology makes the collection of large data sets much easier, providing secondary sources for analysis.	• organize relevant data to identify similarities or differences and describe how the data can be interpreted to make sense of phenomena	• analyze and represent relevant data describing how the data can be interpreted to make sense of phenomena	• evaluate and analyze data to refine a problem statement or make sense of phenomena
Using Mathematics and Computational Thinking (UMCT): In both science and engineering, mathematics and computation are fundamental tools for representing physical variables and their relationships. They are used for a range of tasks, such as constructing simulations; statistically analyzing data; and recognizing, expressing, and applying quantitative relationships.	• identify ways to organize or analyze qualitative or quantitative data	• collect and organize data to reveal patterns, determine whether qualitative or quantitative data would be more appropriate	• organize complex data sets of qualitative or quantitative data, as determined to be appropriate, for determining relationships and patterns, creating algorithms, or utilizing mathematical representations to support conclusions

Grade 5 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
Constructing Explanations (for Science) and Designing Solutions (for Engineering) (CEDS): The products of science are explanations and the products of engineering are solutions.	• identify evidence or scientific ideas that support relationships to create solutions to a problem	• construct an explanation using evidence which utilizes scientific ideas to solve problems	• using evidence, evaluate and refine explanations of relationships among variables in determining the strengths and weaknesses of a design
Engaging in Argument from Evidence (EAE): Argumentation is the process by which explanations and solutions are reached.	• identify evidence or compare facts in a claim	• distinguish among facts to construct, support, or evaluate a claim	• make or evaluate a claim using multiple sets of data
Obtaining, Evaluating, and Communicating Information (OECI): Scientists and engineers must be able to communicate clearly and persuasively the ideas and methods they generate. Critiquing and communicating ideas individually and in groups is a critical professional activity.	• compare and summarize information to communicate basic scientific explanations of a phenomenon	• compare and combine information from various sources to communicate scientific explanations in various media	• evaluate scientific information to describe evidence and support future investigations

E.2.2 Grade 8 Threshold PLDs

The Threshold Performance-Level Descriptors (PLDs) define the minimum knowledge, skills, and practices that students must display for each Disciplinary Core Idea and Science and Engineering Practice to reach a certain performance level. They expand upon the brief overall PLDs included in the Score Interpretation Guide.

Grade 8 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS1: Matter and Its Interactions	• that everything is made from atoms and that the states of matter have some unique characteristics • that temperature and/or pressure have an effect on changes of state • that chemical reactions create new substances while the mass does not change, and energy is involved	• that substances are made from one or more types of atoms and that the particles in the states of matter have unique characteristics • that atoms are regrouped and conserved during chemical processes, and energy is either released or stored	• that substances can be made from two to thousands of atoms that can be combined in a variety of ways • that the same numbers of atoms are regrouped into different molecules to create new substances with different properties, and therefore, the mass does not change
PS2: Motion and Stability: Forces and Interactions	• that the movement of an object is the sum of its forces • that forces among objects are either attractive or repulsive and are dependent upon the distance between the objects	• that in every interaction, there is a pair of forces acting on the two interacting objects and that the size of the forces on the first object equals the size of the forces on the second object • that the size of the electromagnetic force depends upon the magnitudes of the charges, currents, or magnetic strengths due to the fields created	• of the effect of balanced versus unbalanced forces on the motion of objects • that there is a relationship among forces, the fields created, and the magnitudes of the charges, currents, or magnetic strengths involved and among the distance between interacting objects and the masses of the interacting objects

Grade 8 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS3: Energy	• to identify kinetic energy, potential energy, temperature, and heat • that if there is a change in motion energy, it is due to energy being transferred in or out of the system • to identify that, during a collision, energy is transferred, and both objects exert a force • to identify reactants needed to make food in plants and the products of cellular respiration	• of the proportional relationships that define kinetic and potential energy and the relationship between temperature and energy • of the relationship between energy and motion and how the amount of energy needed to cause changes is related to the properties of the substance • by describing the interaction between two objects in terms of force and energy transfer • to describe in general the processes of photosynthesis and cellular respiration including their reactants and products	• to explain the relationship among the variables for kinetic and potential energy and explain how temperature is affected by composition, state, and energy of the particles in the system • to explain the flow of energy in a system, the relationship between the properties of a substance, and the energy needed to change the temperature or motion of the particles • to explain why objects exert a force on each other and that energy is transferred during an interaction • to explain the relationship between photosynthesis and cellular respiration and predict effects of a change to the system
PS4: Waves and Their Applications in Technologies for Information Transfer	• to identify properties of a simple wave • to identify the effect on a beam of light as it crosses between media and when it interacts with an object • to identify methods and their characteristics for transmitting information	• to describe the properties of a simple wave and how it moves • to describe the effect on light as it crosses between media, the path it follows, and its interaction with objects • by describing how digitized signals are a more reliable way to encode and transmit information than analog signals	• to explain the relationship between the properties of a wave and the requirement of a medium for transmission • by explaining how the properties of an object affect how light interacts with it and that the wave model of light is useful for explaining certain properties of light • to explain why digitized signals are a more reliable way to encode and transmit information than analog signals

Grade 8 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS1: From Molecules to Organisms: Structures and Processes	• that cells contain special structures which may be specific to the type of cell in a living unicellular or multicellular organism • of why genetic material is transferred differently in asexual reproduction and sexual reproduction, of how animal behaviors aid in reproduction for both the animal and/or some plants, and discuss genetic factors and local conditions that can affect growth of an organism • that matter and energy cycle through plants, creating sugars which can be broken down or rearranged to release the energy • that sense receptors can send various signals to the brain	• that cells are the smallest unit of life, that living organisms can consist of one or more cells, and that multicellular organisms often contain specialized systems working together, and discuss the functions of special structures within cells • of characteristics, specialized features, and animal behaviors that increase the reproduction chance for both animals and plants, and explain how growth is affected by both genetic and environmental factors • of the process of photosynthesis for the creation of food and of the fact that to use that food, it needs to be broken down through another series of chemical reactions • that nerves transmit sense receptor inputs to be processed in the brain, resulting in memories or responses	• of how parts of a cell function together in a manner similar to how systems interact in multicellular organisms • of characteristics, specialized features, and animal behaviors that increase the reproduction chance for both animals and plants and explain how growth is affected by both genetic and environmental factors • of the relationship between photosynthesis and cellular respiration and of how an organism obtains energy to sustain life • of the different ways a sense receptor reacts to inputs and of the process by which the signal is processed

Grade 8 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS2: Ecosystems: Interactions, Energy, and Dynamics	• that organisms are dependent on resources for which they may need to compete • that matter and/or energy are cycled through a food web of an ecosystem • that there are physical and biological components of ecosystems, that changes to those will cause disruption, and that biodiversity is related to species representation and can be used to determine overall health of an ecosystem • that changes in biodiversity have an impact on humans	• of how growth and survival of an organism are dependent on access to limited resources and interactions with other organisms • of how matter and energy transfer between trophic levels • of the dynamic nature of ecosystems and of how biodiversity is used as a measure of an ecosystem's health • of how changing biodiversity can affect humans and the services humans rely on	• of an organism's reliance on the environment and of how populations are limited by access to resources, predatory interactions, and competition • of how a food web can model mechanisms for the cycling of matter, including the role of decomposers, which in turn account for the conservation of energy • of the relationship between biodiversity and ecosystem health, and of the predicted outcomes of disturbances to an ecosystem • of why changes in biodiversity affect humans
LS3: Heredity: Inheritance and Variation of Traits	• that genes are located on inherited chromosomes and that the gene may be slightly different from the parent's • that in sexual reproduction, each parent contributes half of the genetic material and that mutations that occur can be beneficial, harmful, or neutral	• that genes control production of proteins and that mutations cause genetic variation • about genetic contributions during sexual reproduction and the general effects that mutations cause	• of how genes control protein production and of what effect mutations could have on this process • of why individuals have two of each chromosome and how mutations may result in structural and functional changes

Grade 8 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS4: Biological Evolution: Unity and Diversity	• that fossils can show the evolutionary progression of organisms living today, that organisms may be artificially selected for reproduction based on desired traits, and that while embryos across species may have similarities as they develop, the organisms with more advantageous traits are more likely to survive • that environmental conditions will drive trait commonality in species	• of the uses for the fossil record and of embryological development, including similarities not evident in the fully formed anatomy, where certain traits, whether natural or artificially selected, will provide advantages for survival • of how environmental conditions can change a species over generations and of how distributions of traits reflect adaptation by natural selection	• of evolutionary history based on anatomical similarities and to predict predominance of certain traits in a population • to predict trait distribution in a species based on changing environmental conditions

Grade 8 Threshold Performance-Level Descriptors (Earth and Space Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
ESS1: Earth's Place in the Universe	• that the celestial bodies have observable patterns and that we exist in a galaxy called the Milky Way • that gravity acts on objects, that there are eclipses, and that Earth's tilt causes seasons • that fossils are used to date rock layers and that tectonic processes change Earth	• to predict the observed motion of the Sun, Moon, and stars • that gravity is an attractive force, that alignment of the Earth-Moon-Sun causes solar and lunar eclipses, and that changes in seasons are due to intensity of sunlight • that Earth's history can be determined from rock layers and that tectonic processes create and destroy Earth materials	• to explain the predictable observed patterns of the Sun, Moon, and stars • to predict eclipses and seasonal changes based on data or models • that rock layers and fossils only provide relative dates and that the sea floor has different ages
ESS2: Earth's Systems	• of where Earth's energy comes from and that Earth processes vary in timeframe and size • that Earth's plates move in different ways • that water cycles in Earth's spheres and affects weather patterns, that ocean water density varies, and that moving water affects landforms • that both living and nonliving factors influence complex weather patterns	• that energy and matter have caused, and continue to cause, changes on Earth • that rocks and fossils help determine how Earth's plates have moved • of the way that water cycles, of the factors that affect the movement of water in Earth's spheres, of the causes of ocean density differences, and of the way that moving water affects landforms • of how weather patterns are influenced by living and nonliving factors that vary with location and of how the ocean is a major driving factor	• of the interaction between Earth's processes driven by differing energy sources to explain Earth's history or predict future geological events • to predict effects of plate movement on Earth's landscape • to predict weather patterns that are the result of the cycling of water and of impacts of density on ocean currents • to predict the effect living and nonliving factors, including the ocean, have on weather and climate

Grade 8 Threshold Performance-Level Descriptors (Earth and Space Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
ESS3: Earth and Human Activity	• that resources are not evenly distributed • that natural hazards can be mapped • that human populations may negatively impact resources and that human activity has both positive and negative impacts on different organisms • of climate science and of the fact that human activities have an effect on global temperatures	• that there are renewable and nonrenewable resources • that mapping hazards can help understand geological forces • on how humans have altered the biosphere and that humans are making technological gains to minimize negative impacts • of how human activities affect temperatures and that climate science may help lead to decisions to benefit life on Earth	• of the relationship of past geological processes and the distribution of resources • to predict future hazards based on historical occurrences • to predict whether human activities would be positive or negative and to evaluate solutions based on the rate of resource consumption • to predict when human activities will have significant impacts on the Earth's climate

Grade 8 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
Analyzing and Interpreting Data (AID): Scientific investigations produce data that must be analyzed in order to derive meaning. Because data patterns and trends are not always obvious, scientists use a range of tools—including tabulation, graphical interpretation, visualization, and statistical analysis—to identify the significant features and patterns in the data. Scientists identify sources of error in the investigations and calculate the degree of certainty in the results. Modern technology makes the collection of large data sets much easier, providing secondary sources for analysis.	• identify and/or interpret data, graphical displays, and/or concepts of statistics and/or their limitations to provide evidence for phenomena	• analyze, interpret, and/or use simple data sets and/or concepts of statistics to identify relationships and/or define operational ranges for objects, processes, and/or systems	• analyze and interpret complex or multiple data sets and/or construct graphical displays to identify and/or explain relationships, limitations of data, when to use concepts of statistics, and/or to justify operational ranges for objects, processes, and/or systems
Asking Questions (for Science) and Defining Problems (for Engineering) (AQDP): A practice of science is to ask and refine questions that lead to descriptions and explanations of how the natural and designed world works and which can be empirically tested.	• identify questions that arise from observations and models in order to clarify information and/or arguments, refine models, and/or determine relationships	• ask testable questions that arise from observations of phenomena, models, and/or unexpected results in order to clarify information, evidence, arguments, and/or design problems that can be solved through development of objects/tools, processes, and/or systems	• analyze and/or evaluate testable questions that arise from observations of phenomena, models, and/or unexpected results in order to clarify information, evidence, arguments, and/or design problems that can be solved through development of objects/tools, processes, and/or systems

Grade 8 SEP Threshold Performance-Level Descriptors
Students should be able to:

SEP	Level 2	Level 3	Level 4
Constructing Explanations (for Science) and Designing Solutions (for Engineering) (CEDS): The products of science are explanations, and the products of engineering are solutions.	• identify or revise an explanation and/or design project based on models or representations, or by applying scientific reasoning and/or evidence	• construct, revise, and/or use an explanation based on models or representations, or by applying scientific reasoning and/or evidence, or by undertaking a design project to construct and/or implement a solution	• analyze, construct, and/or elaborate on an explanation based on models or representations by applying scientific reasoning and/or evidence, or by evaluating a design project to construct and/or implement solutions and/or optimize performance
Developing and Using Models (DUM): A practice of both science and engineering is to use and construct models as helpful tools for representing ideas and explanations. These tools include diagrams, drawings, physical replicas, mathematical representations, analogies, and computer simulations.	• use a simple model to show relationships, make predictions, or generate data and/or describe its limitations	• develop and/or revise a simple model to show relationships, make predictions, or generate data and/or evaluate its limitations	• develop, revise, and/or evaluate a complex model to show relationships, make predictions, or generate data and/or evaluate its merits and limitations
Engaging in Argument from Evidence (EAE): Argumentation is the process by which explanations and solutions are reached.	• identify evidence in arguments to support or refute explanations, • provide critiques of procedures or models, and/or • identify competing design solutions	• identify and/or compare multiple pieces of evidence in arguments, • provide critiques about explanations or questions, and/or • write arguments that support or refute the advertised performance of a device, process, or system	• critique arguments, procedures, or models; • construct and/or use written arguments to support or refute explanations, models, and/or solutions; or • analyze empirical evidence to support written arguments

Grade 8 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
Obtaining, Evaluating, and Communicating Information (OEIC): Scientists and engineers must be able to communicate clearly and persuasively the ideas and methods they generate. Critiquing and communicating ideas individually and in groups is a critical professional activity.	• read and use information from multiple simple scientific sources to describe patterns, clarify claims, and/or assess accuracy	• integrate information from multiple, complex, qualitative sources to clarify claims, assess accuracy, and evaluate conclusions	• integrate information from multiple, complex, quantitative sources to describe patterns, clarify claims, assess accuracy, and evaluate conclusions
Planning and Carrying Out Investigations (PACI): Scientists and engineers plan and carry out investigations in the field or laboratory, working collaboratively as well as individually. Their investigations are systematic and require clarifying what counts as data and identifying variables or parameters.	• plan and/or conduct an investigation that includes the identification of appropriate tools and methods for collecting data in order to provide evidence or test a design solution	• plan an investigation that includes the identification of variables and/or controls, or indicates how much data is sufficient to serve as evidence necessary to test a design solution, or evaluate an experimental design	• plan and refine an investigation that includes the identification of variables and controls, tools, how data will be collected, and how much data is sufficient to serve as evidence necessary to test a design solution, or revise an experimental design
Using Mathematics and Computational Thinking (UMCT): In both science and engineering, mathematics and computation are fundamental tools for representing physical variables and their relationships. They are used for a range of tasks such as constructing simulations; statistically analyzing data; and recognizing, expressing, and applying quantitative relationships.	• identify qualitative and quantitative data and when the use of digital tools is warranted, • select appropriate mathematical representations, and • use algorithms to solve problems and/or address engineering questions	• decide whether to use qualitative or quantitative data, • use digital tools to analyze large data sets, • use mathematical representations, and • explain and/or evaluate algorithms or mathematical concepts for solving problems and/or addressing engineering questions	• explain when to use qualitative or quantitative data, • evaluate digital tools, • explain mathematical representations, and/or • create algorithms to solve problems and/or address engineering questions

E.2.3 Grade 11 Threshold PLDs

The Threshold Performance-Level Descriptors (PLDs) define the minimum knowledge, skills, and practices that students must display for each Disciplinary Core Idea and Science and Engineering Practice to reach a certain performance level. They expand upon the brief overall PLDs included in the Score Interpretation Guide.

Grade 11 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS1: Matter and Its Interactions	• of subatomic particles, their interactions, and the involvement of energy in these interactions • of an understanding of how collisions between molecules affect reaction rates • that some reactions are reversible • that atoms are conserved during reactions • that nuclear processes involve energy	• of atomic properties and patterns through the use of the periodic table, as well as different types of particle interactions and the energy involved • of the factors that affect reaction rates and equilibrium systems • of the energy involved in the rearranging of atoms and molecules • of the different types of reactions and how to make predictions about them • that energy and matter are conserved in nuclear processes	• of varying atomic structures • of how the periodic table models the patterns of the properties and electron structure of the elements • of how particle interactions affect bulk properties of substances • of how collisions lead to changes in the sum of all the bond energies • of how atom conservation and chemical properties can be used to make predictions on chemical reactions • of multiple nuclear processes

Grade 11 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS2: Motion and Stability: Forces and Interactions	• of quantified acceleration and momentum • of types of fields and attractive/repulsive forces of gravitational and/or electric fields • that electrical energy can be stored or transmitted	• (quantified knowledge) of factors that affect Newton's second law, single object momentum systems, and conservation of momentum • of how interactions happen at a distance due to fields • of electrical interactions at the atomic level • of the difference between magnetic and electric fields OR • (quantified knowledge) of Coulomb's law and Newton's universal law of gravitation • of how electrical energy can be stored in a battery or transmitted by electric currents	• (quantified knowledge) of outside interactions that affect the momentum and acceleration of a single- or multiple-object system • of how to predict changes in electrical and gravitational forces • of how to describe fields as force and energy fields and predict the effect of electrical and/or magnetic fields due to interactions between the two fields
PS3: Energy	• of how different types of energy can be transferred • of systems in which energy is conserved and how the availability of energy restricts what is possible in a closed system • of the nature of the relationship between two objects interacting in a field using the energy prospective • of how energy can be converted to different forms	• of how energy manifests itself at the microscopic and macroscopic scale and how energy transfers in a system • (quantified knowledge) of how energy transfers in and out of a system OR • of possible and impossible events based on energy availability, and defined stable states • of how the distance between two objects affects the energy of a field • of how energy can be converted to less useful forms • of how solar energy can be captured and used for other processes, such as photosynthesis	• of the amount of various types of energy in a given situation and how microscopic changes affect macroscopic manifestations of energy • of how to evaluate physical changes in a system using the conservation of energy • of how to predict changes in energy in a field based on the position and nature of objects • of the importance of energy conservation and efficiency

Grade 11 Threshold Performance-Level Descriptors (Physical Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
PS4: Waves and Their Applications in Technologies for Information Transfer	• of how a wave travels through a medium, including understanding of examples of digitized information, and qualitative understanding of superposition principle • of the wave and particle models of electromagnetic radiation, the absorption of electromagnetic radiation, and the relationship between frequency and energy of light • of everyday experiences that involve waves and how wave signals are produced, transmitted, and captured	• (quantified knowledge) of the relationship among frequency, wavelength, and speed in a real-world phenomenon OR • of the advantages and disadvantages of digitizing information • of the effect of absorption of electromagnetic waves, features of electromagnetic radiation that can be explained by either the wave or particle model, and the nature of photoelectric materials • of technologies used to produce, transmit, and/or capture signals and technologies used to store and interpret information	• of waves in various media and how combining waves of different frequencies can make a wide variety of patterns and thereby encode and transmit information • of the difference between the wave- and particle-like behavior of electromagnetic radiation and how either the wave or particle model can be used to explain how an electron is emitted and how it can damage living cells • of how technology can be used to store and/or interpret information

Grade 11 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS1: From Molecules to Organisms: Structures and Processes	• of how multicellular organisms utilize feedback mechanisms and have specialized cells that are organized and function according to the proteins coded by the DNA • of the role of cellular division (mitosis) in creating genetically identical cells that differentiate into complex multicellular organisms • of photosynthesis and cellular respiration as the chemical processes of life that produce or utilize carbon-based molecules that are recombined into different products of living systems	• of how positive and negative feedback mechanisms are beneficial to multicellular organisms, which have systems of specialized cells that perform essential life functions expressed through proteins coded for by genes • of how mitosis and differentiation produce and maintain complex organisms from a single cell • of the chemistry behind photosynthesis, how cellular respiration uses energy to maintain the organism, and how the products of these processes are used to build larger molecules	• of how changing genes (mutation) can lead to functional changes of a protein and how positive and/or negative feedback helps maintain the equilibrium of an organism • of how genetic material from two variants of each chromosome pair is maintained as a single cell (fertilized egg) grows to a multicellular organism • of the interdependence of photosynthesis and cellular respiration and their role in the growth and maintenance of living systems

Grade 11 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS2: Ecosystems: Interactions, Energy, and Dynamics	• of both living and nonliving factors that contribute to the carrying capacity of the ecosystem • of how food webs often have photosynthetic producers at the lowest level, how a small amount of matter and energy will transfer upward in the food web reducing the amount of organisms that can exist at higher levels, and how this relates to the carbon cycle • of how ecosystems have interactions that keep the population numbers stable, and ecosystems are resilient to modest changes, but humans can disrupt ecosystems and species survival • of how group behavior has evolved to increase individual and group survival	• of how carrying capacity is affected by challenges and/or availability of resources • of how photosynthesis and cellular respiration are connected and use carbon in maintaining life processes, that the matter and energy of a food web are used and restructured by the organisms in the food web, and that a small amount is used by the next levels of the food web • of complex ecosystem interactions and their effects on population size, including biological and physical disturbances, extreme fluctuations, and the ways human activity can have an effect on an ecosystem • of how group behaviors can increase the chances of survival for individuals and their genetic relatives	• of how carrying capacity affects the population size of a given species within an ecosystem • of how carbon and matter are used in the maintenance of life processes (including photosynthesis and both anaerobic and aerobic respiration) through the food web, including how carbon cycles through Earth's spheres • of how changes to populations and environments caused by human interactions and other physical events within ecosystems can result in changes that affect both the organisms and the environment • of how changes to the group or conditions can affect the survival of individuals and their genetic relatives

Grade 11 Threshold Performance-Level Descriptors (Life Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
LS3: Heredity: Inheritance and Variation of Traits	• of how all cells have the same DNA containing genes that are the organisms' characteristics, but not all DNA codes for protein • of the processes within meiosis, errors that can occur during DNA replication, and mutations due to environmental factors that can create genetic diversity, which may be passed to future generations	• that chromosomes contain genes that code for proteins and regions that do not code for proteins, and that different cells express different genes • that while the process of DNA replication is tightly regulated and highly accurate, errors still occur, and combined with mutations due to environmental factors, DNA replication can create genetic diversity that may affect survivability and the transmission of traits to future generations	• of the mechanisms of gene regulation and different possible functions of segments of non-protein coding DNA • of the mechanisms within meiosis that create genetic diversity, as well as the effects of environmental factors on DNA replication and the impact of the changes to DNA on genetic diversity within populations
LS4: Biological Evolution: Unity and Diversity	• of the different types of evidence of evolution • of how natural selection allows inheritable advantageous traits to become more common if they increase chances of survival within populations • that natural selection selects for inheritable traits that provide a survival advantage for a particular environment • that changes to the environment may cause the selection of different traits leading to changes in the population known as adaptation • that the frequency of traits depends on natural selection forces that can change with a changing environment • of how biodiversity increases or decreases and how humans need resources and biodiversity, but are having adverse effects on biodiversity	• of how different sources of evidence for evolution can support each other • of how gene expression and genetic variation in the individual lead to differences in performance of the individuals in a population, and how positively selected traits are more common in a population because they increase survival • that evolution occurs when there is genetic variation, competition, and selective reproduction of organisms with desirable genetic traits • that organisms with desirable traits will become more common, but as the environment changes, different traits may provide the selective advantages • that some populations may increase while others may go extinct • of specific results of human activities that affect the environment and biodiversity and reasons why preservation of biodiversity is desirable	• of how DNA sequences, amino acid sequences, and anatomical and embryological evidence support that evolution has occurred • of how natural selection leads to different levels of performance of the individual • that factors affecting natural selection work together creating changes in the diversity within populations and ecosystems • that changing environments cause changes in selection pressures that result in further adaptation or extinction • of ways that humans can maintain or increase biodiversity while meeting the needs of humanity and why this is beneficial to life on Earth

Grade 11 Threshold Performance-Level Descriptors (Earth and Space Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
ESS1: Earth's Place in the Universe	• of the big bang, which allowed for the creation of galaxies and stars, where many elements are created • of identifying properties of orbits, factors that affect the orbit, and how the orbit affects the stellar body • of plate tectonics and erosion, which cause the destruction of early rock records on Earth, and that we have to rely on other objects in the solar system for information on Earth's formation	• that light spectra emitted from a star can give information about its life cycle, composition, and distance • of features of motion of orbital objects, what changes that motion, and the effects of changing the motion of the stellar body • of the fact that while there is a range in the age of the rocks on Earth, the early rock history has been destroyed, and we rely on studying other stellar bodies to explain how Earth formed	• of the life cycle of stars and explain how the characteristics of a star can support the big bang theory • of the laws explaining motions of orbiting objects, their changes, and the changes to the stellar bodies as a result of those changes • of why different areas of Earth have rocks of different ages and the processes that are erasing the early rock history
ESS2: Earth's Systems	• of how Earth has a series of interacting dynamic systems • that Earth's surface is in motion, and that motion can create physical features on Earth's surface • of the properties of water that are essential to Earth's dynamics • of Earth's atmosphere and how it undergoes temperature changes • that dynamic and delicate feedbacks between Earth's systems and biosphere exist	• of methods of investigation of Earth's dynamic systems and how the data can be used to describe the effects of these systems • that Earth's surface is in motion due to convection, creating physical features that have changed throughout history • of how the properties of water are essential to Earth's processes • of how Earth's atmosphere undergoes short-term and long-term temperature changes at the global scale due to changes in the biosphere, including human activities • of how dynamic and delicate feedback between Earth's systems and biosphere causes a continual co-evolution of Earth's surface and the life that exists on it	• of Earth's dynamic systems in explaining the effects of these systems and the development of the currently accepted model of the structure of the planet • of the theory of plate tectonics allowing for the prediction of future plate movements and interpretations of Earth's geologic history • of how the properties of water can be used to explain Earth's processes • of why Earth's atmosphere undergoes short-term and long-term temperature changes at the global scale • of how positive and/or negative feedbacks between the biosphere and other Earth systems cause a continual co-evolution of Earth's surface and the life that exists on it

Grade 11 Threshold Performance-Level Descriptors (Earth and Space Science)

Students should be able to demonstrate knowledge:

DCI	Level 2	Level 3	Level 4
ESS3: Earth and Human Activity	• that new technologies have associated costs, risks, and benefits • that natural hazards have shaped human history • that human activities can have both positive and negative impacts on biodiversity • of humans' abilities to use technology to model, predict, and manage current and future impacts	• that new technologies have associated costs, risks, and benefits at the economic, social, environmental, and/or geopolitical level • of how natural hazards and geological events have shaped human history through changes in the human population including through migration at the local, regional, and/or global scale • that human impacts on biodiversity can be mitigated by the development of new technologies and/or responsible resource management • of technologies that allow modeling, predicting, and managing of current and future impacts on oceans, the atmosphere, and the biosphere	• of new technologies in order to explain their associated costs, risks, and benefits at the economic, social, environmental, and/or geopolitical level • of how natural hazards affect human population and migration at the local, regional, and global scale • of new technologies and responsible resource management to predict their effects on biodiversity • to explain how humans' abilities to model, predict, and manage current and future impacts have increased alongside the magnitudes of human impacts

Grade 11 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
(Investigating) Asking Questions (for Science) and Defining Problems (for Engineering) (AQDP): A practice of science is to ask and refine questions that lead to descriptions and explanations of how the natural and designed worlds work and which can be empirically tested. Engineering questions clarify problems to determine criteria for successful solutions and identify constraints to solve problems about the designed world. Both scientists and engineers also ask questions to clarify ideas. Asking questions and defining problems in 9–12 progresses to formulating, refining, and evaluating empirically testable questions and design problems using models and simulations.	• ask relevant questions or define problems in different contexts, based on unexpected results, independent and dependent variables, models, theories, etc.	• ask relevant and testable questions that arise from careful observation of phenomena, unexpected results, or models or theories for the purpose of determining relationships, providing an explanation, or clarifying and refining a design	• analyze, evaluate, and/or revise questions that arise from careful observation of phenomena, unexpected results, or models or theories for the purpose of determining relationships, providing an explanation, or clarifying and refining a design
(Sensemaking) Developing and Using Models (DUM): A practice of both science and engineering is to use and construct models as helpful tools for representing ideas and explanations. These tools include diagrams, drawings, physical replicas, mathematical representations, analogies, and computer simulations. Modeling in 9–12 progresses to using, synthesizing, and developing models to predict and show relationships among variables between systems and their components in the natural and designed worlds.	• use a model to generate data that test the model's reliability and/or evaluates its merits and limitations	• develop simple models and revise different types of models that test and/or predict relationships among systems/ phenomena based on the models' merits and limitations	• develop or revise complex models that test and/or predict relationships/ phenomena based on the models' merits and limitations

Grade 11 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
(Investigating) Planning and Carrying Out Investigations (PACI): Scientists and engineers plan and carry out investigations in the field or laboratory, working collaboratively as well as individually. Their investigations are systematic and require clarifying what counts as data and identifying variables or parameters. Planning and carrying out investigations in 9–12 progresses to include investigations that provide evidence for and test conceptual, mathematical, physical, and empirical models.	• identify ways to conduct an investigation (including making a directional hypothesis) or test a design solution through manipulating variables or acquiring data	• plan and/or conduct an investigation (including making a directional hypothesis) or test a design solution through manipulating variables or acquiring data	• revise and/or evaluate an investigation in which an independent variable is manipulated or an unsatisfactory performance is found
(Sensemaking) Analyzing and Interpreting Data (AID): Scientific investigations produce data that must be analyzed in order to derive meaning. Because data patterns and trends are not always obvious, scientists use a range of tools—including tabulation, graphical interpretation, visualization, and statistical analysis—to identify the significant features and patterns in the data. Scientists identify sources of error in the investigations and calculate the degree of certainty in the results. Modern technology makes the collection of large data sets much easier, providing secondary sources for analysis. Analyzing data in 9–12 progresses to introducing more detailed statistical analysis, the comparison of data sets for consistency, and the use of models to generate and analyze data.	• identify the appropriate statistics and/or data, and/or their limitations, when providing evidence for claims, design solutions, or solving problems	• apply and/or analyze data and statistics to identify or solve scientific and engineering problems, or to make scientific claims	• evaluate the use of data and statistics and/or their limitations to solve problems, make claims, or design solutions

Grade 11 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
(Investigating) Using Mathematics and Computational Thinking (UMCT): In both science and engineering, mathematics and computation are fundamental tools for representing physical variables and their relationships. They are used for a range of tasks such as constructing simulations; statistically analyzing data; and recognizing, expressing, and applying quantitative relationships. Mathematical and computational thinking in 9–12 progresses to using algebraic thinking and analysis, a range of linear and nonlinear functions including trigonometric functions, exponentials and logarithms, and computational tools for statistical analysis to analyze, represent, and model data. Simple computational simulations are created and used based on mathematical models of basic assumptions.	• apply/use mathematical concepts to describe conclusions that may require deciding when to use qualitative versus quantitative data	• apply/use mathematical computational representations to see if a model is viable, or decide if qualitative or quantitative data meet criteria for success	• through the use of evaluation of mathematical computations, create a model or justify the choice of qualitative versus quantitative data
(Sensemaking) Constructing Explanations (for science) and Designing Solutions (for engineering) (CEDS): The products of science are explanations, and the products of engineering are solutions. Constructing explanations and designing solutions in 9–12 progresses to explanations and designs that are supported by multiple and independent student-generated sources of evidence consistent with scientific ideas, principles, and theories.	• identify and describe appropriate data and/or evidence for supporting claims, solving problems, constructing explanations, or designing solutions	• make or revise claims, explanations, or solutions by applying appropriate data and/or evidence	• evaluate, design, or construct claims, explanations, or solutions by applying appropriate data, evidence, and/or scientific theories and laws

Grade 11 SEP Threshold Performance-Level Descriptors

Students should be able to:

SEP	Level 2	Level 3	Level 4
(Critiquing) Engaging in Argument from Evidence (EAE): Argumentation is the process by which explanations and solutions are reached. Engaging in argument from evidence in 9–12 progresses to using appropriate and sufficient evidence and scientific reasoning to defend and critique claims and explanations about the natural and designed worlds. Arguments may also come from current scientific or historical episodes in science.	• identify and/or describe the main points of an argument or claim that is based on scientific evidence	• evaluate and/or defend a claim or argument—or choose between competing arguments—related to currently accepted explanations or solutions	• construct and/or critique an argument or claim by using scientific evidence
(Critiquing) Obtaining, Evaluating, and Communicating Information (OECI): Scientists and engineers must be able to communicate clearly and persuasively the ideas and methods they generate. Critiquing and communicating ideas individually and in groups is a critical professional activity. Obtaining, evaluating, and communicating information in 9–12 progresses to evaluating the validity and reliability of the claims, methods, and designs.	• read and compare sources of information to describe patterns in evidence and/or evidence for solving problems or answering scientific questions	• integrate information from multiple sources to gather valid and reliable evidence for solving problems or answering scientific questions	• evaluate information from multiple sources and determine the usefulness of evidence, ensuring it is valid and reliable, for solving problems or answering scientific questions

E.3 Reporting PLDs

E.3.1 Reporting PLDs–Level 1

Students who are at Level 1 demonstrated a minimal understanding of the New Jersey Student Learning Standards for Science (NJSLS–S) by misinterpreting information from a variety of sources (e.g., text, charts, graphs, tables) and inconsistently applying the knowledge gained from scientific investigations to develop incorrect explanations or models of observed phenomena. The students had difficulty choosing and using, even with significant scaffolding, the appropriate tools to make observations and to gather, classify, and present data. The students struggled to use essential information to recognize patterns and relationships between data and designed systems. The students seldom used information to make real-world connections or predictions.

E.3.2 Reporting PLDs–Level 2

Students who are at Level 2 demonstrated a limited grade-level understanding of the New Jersey Student Learning Standards for Science (NJSLS–S) by partially interpreting information from a variety of sources (e.g., text, charts, graphs, tables) and inconsistently applying the knowledge gained from scientific investigations to develop incomplete explanations or models of observed phenomena. The students had some difficulty choosing and using the appropriate tools to make observations and to gather, classify, and present data. The students may be able to use essential information to recognize patterns and relationships between data and designed systems. The students inconsistently used information to make real-world connections and predictions.

E.3.3 Reporting PLDs–Level 3

Students who are at Level 3 demonstrated appropriate grade-level understanding of the New Jersey Student Learning Standards for Science (NJSLS–S) by comprehending information from a variety of sources (e.g., text, charts, graphs, tables) and applying the knowledge gained from scientific investigations to develop accurate explanations and models of observed phenomena. The students often chose and used the appropriate tools to make observations and to gather, classify, and present data. The students used both essential and nonessential information to recognize patterns and relationships between data and designed systems. The students were able to use information to make real-world connections and predictions.

E.3.4 Reporting PLDs–Level 4

Students who are at Level 4 demonstrated advanced understanding of the New Jersey Student Learning Standards for Science (NJSLS–S) by integrating information from a variety of sources (e.g., text, charts, graphs, tables) and analyzing the knowledge gained from scientific investigations to develop sophisticated explanations and models of observed phenomena. The students consistently chose and used the appropriate tools to make observations and to gather, classify, and present relevant data. The students considered both essential and nonessential information to explain patterns and relationships between data and designed systems. The students regularly used information and provided supporting explanations in making real-world connections and predictions.

APPENDIX F: Detailed Test Maps

Table F.1: Grade 5 Test Map–Metadata and Item Statistics

Grade 5											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
1905B001_04B	1	TE	AID	ESS1.B	PAT	Earth and Space	Sensemaking	–0.655	.67	.44	85
1905B001_08B	1	TE	UMCT	ESS1.B	PAT	Earth and Space	Investigating	0.831	.38	.41	142
1905B001_11	1	MC	AQDP	ESS1.B	C and E	Earth and Space	Investigating	0.139	.52	.23	115
1905M069_01	1	TE	UMCT	PS1.B	S, P, and Q	Physical	Investigating	0.863	.38	.54	147
1905M069_02	1	TE	AID	PS1.B	PAT	Physical	Sensemaking	1.131	.30	.54	49
1905M069_05	1	TE	UMCT	PS1.B	PAT	Physical	Investigating	1.569	.21	.52	75
2005M015_02	1	MC	AID	ESS3.B	C and E	Earth and Space	Sensemaking	–0.415	.67	.36	100
2005M015_04	1	MC	CEDS	ESS3.B	C and E	Earth and Space	Sensemaking	1.009	.35	.37	65
2005M015_05	1	TE	CEDS	ESS3.B	SF	Earth and Space	Sensemaking	0.934	.34	.33	66
2005M015_07	1	TE	UMCT	ESS3.B	PAT	Earth and Space	Investigating	0.696	.39	.57	84
2005M015_10	1	TE	DUM	ESS3.B	S & SM	Earth and Space	Sensemaking	0.875	.37	.38	51
2105M003_01	1	TE	AQDP	LS3.A	PAT	Life	Investigating	1.157	.32	.46	150
2105M003_04	1	MC	EAE	LS3.A	PAT	Life	Critiquing	0.520	.44	.35	60
2105M003_07	1	MC	CEDS	LS3.A	PAT	Life	Sensemaking	–0.756	.69	.35	46
2105M003_09	1	TE	EAE	LS3.A	PAT	Life	Critiquing	1.557	.26	.52	79
2205B000_02	1	TE	AQDP	LS2.C	C and E	Life	Investigating	2.044	.22	.40	120
2205B000_03	1	MC	PACI	LS2.C	PAT	Life	Investigating	0.409	.46	.42	61
2205B000_04	1	TE	PACI	LS4.C	C and E	Life	Investigating	0.409	.49	.62	48
2205B000_07	4	CR	EAE	LS2.C	C and E	Life	Critiquing	0.793	.40	.74	424
2205B000_10	1	MC	AID	LS2.C	PAT	Life	Sensemaking	1.114	.32	.30	75
2205B007_02	1	TE	DUM	PS3.D	SF	Physical	Sensemaking	1.398	.28	.45	117
2205B007_03	3	CR	OECI	ESS3.C	S & SM	Earth and Space	Critiquing	–0.153	.59	.53	379
2205B007_04	1	TE	AQDP	PS3.D	C and E	Physical	Investigating	0.571	.43	.45	63
2205B007_05	1	TE	UMCT	PS3.D	S & SM	Physical	Investigating	1.895	.21	.32	129
2205M007_01	1	TE	AQDP	ESS2.B	PAT	Earth and Space	Investigating	0.902	.37	.59	155

Grade 5											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
2205M007_02	1	TE	EAE	ESS2.B	PAT	Earth and Space	Critiquing	1.688	.24	.27	133
2205M007_04	1	MC	OECI	ESS2.B	PAT	Earth and Space	Critiquing	0.243	.50	.49	73
2205M007_05	1	TE	CEDS	ESS2.B	PAT	Earth and Space	Sensemaking	0.473	.45	.35	81
2205M024_03	1	TE	DUM	PS2.B	SF	Physical	Sensemaking	0.547	.44	.37	125
2205M024_04	1	TE	EAE	PS2.B	S & SM	Physical	Critiquing	1.371	.29	.19	94
2205M024_05	1	TE	PACI	PS2.B	S, P, and Q	Physical	Investigating	1.111	.33	.55	55
2305B004_01	1	MC	AQDP	PS4.A	C and E	Physical	Investigating	0.589	.43	.44	42
2305B004_03	1	TE	OECI	PS4.A	S, P, and Q	Physical	Critiquing	0.509	.45	.35	109
2305B004_04	1	TE	EAE	PS4.A	PAT	Physical	Critiquing	1.647	.24	.27	47
2305B004_09	3	CR	DUM	PS4.A	E&M	Physical	Sensemaking	1.358	.28	.57	430
2305B006_01	1	TE	DUM	PS4.C	S & SM	Physical	Sensemaking	0.508	.43	.47	99
2305B006_03	1	MC	AQDP	PS4.C	C and E	Physical	Investigating	0.812	.40	.41	61
2305B006_04	1	TE	EAE	PS4.A	C and E	Physical	Critiquing	0.875	.36	.24	61
2305B011_01	1	TE	AQDP	LS2.C	C and E	Life	Investigating	1.491	.27	.33	80
2305B011_02	1	TE	EAE	LS1.D	C and E	Life	Critiquing	0.570	.43	.23	43
2305B011_03	1	TE	EAE	ESS2.B	S & SM	Earth and Space	Critiquing	2.540	.13	.27	70
2305B011_04	1	TE	CEDS	ESS3.C	C and E	Earth and Space	Sensemaking	0.593	.43	.48	46
2305B011_06	1	TE	DUM	LS1.D	C and E	Life	Sensemaking	0.853	.38	.46	58
2405B002_01	1	TE	AQDP	LS4.D	C and E	Life	Investigating	1.730	.23	.23	142
2405B002_02	1	MC	UMCT	LS4.D	PAT	Life	Investigating	0.500	.45	.43	74
2405B002_03	1	TE	OECI	LS2.B	E&M	Life	Critiquing	0.296	.49	.54	77
2405B002_04	1	TE	CEDS	LS2.C	SC	Life	Sensemaking	1.560	.26	.43	63
2405B008_03	1	TE	UMCT	ESS2.D	S, P, and Q	Earth and Space	Investigating	2.085	.18	.42	146
2405B008_05	1	TE	AID	ESS2.D	PAT	Earth and Space	Sensemaking	2.028	.19	.28	135
2405B008_06	1	TE	EAE	ESS2.D	PAT	Earth and Space	Critiquing	1.659	.24	.34	69
2405M008_02	1	TE	DUM	PS2.B	E&M	Physical	Sensemaking	1.905	.21	.49	86
2405M008_04	1	MC	OECI	PS2.B	S & SM	Physical	Critiquing	0.379	.47	.32	47
2405M008_05	1	TE	DUM	PS2.B	S, P, and Q	Physical	Sensemaking	1.268	.30	.36	53

Table F.2: Grade 8 Test Map–Metadata and Item Statistics

Grade 8											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
2008M025_01	1	TE	AID	PS2.B	E&M	Physical	Sensemaking	0.086	.38	.48	90
2008M025_02	1	MC	AQDP	PS3.A	S, P, and Q	Physical	Investigating	0.145	.36	.37	82
2008M025_08	1	MC	EAE	PS2.A	E&M	Physical	Critiquing	0.216	.35	.19	61
2108M000_02	1	TE	AQDP	ESS3.C	C and E	Earth and Space	Investigating	0.213	.35	.45	55
2108M000_03	1	TE	AID	ESS3.D	C and E	Earth and Space	Sensemaking	0.698	.23	.49	41
2108M000_06	1	TE	EAE	ESS3.D	SC	Earth and Space	Critiquing	0.893	.22	.22	45
2308B003_01	1	TE	DUM	ESS2.A	S & SM	Earth and Space	Sensemaking	0.044	.46	.56	103
2308B003_03	1	TE	AQDP	ESS2.B	S & SM	Earth and Space	Investigating	0.397	.32	.42	97
2308B003_05	4	CR	OECI	ESS2.A	SC	Earth and Space	Critiquing	0.064	.38	.59	285
2308B003_07	1	TE	PACI	ESS2.B	S, P, and Q	Earth and Space	Investigating	0.937	.26	.43	55
2308B003_11	1	TE	EAE	ESS3.B	S & SM	Earth and Space	Critiquing	–0.704	.59	.45	70
2308B006_01	1	TE	CEDS	LS4.C	C and E	Life	Sensemaking	1.133	.19	.29	65
2308B006_07	1	TE	CEDS	LS2.B	C and E	Life	Sensemaking	1.027	.23	.18	48
2308B006_08	1	TE	AID	LS2.A	PAT	Life	Sensemaking	0.837	.23	.33	37
2308B011_01	1	TE	AQDP	PS1.A	S, P, and Q	Physical	Investigating	0.990	.22	.19	83
2308B011_06	1	MC	AQDP	ESS2.C	S & SM	Earth and Space	Investigating	0.046	.38	.40	142
2308B011_07	1	TE	OECI	PS1.A	S & SM	Physical	Critiquing	–0.625	.52	.47	83
2308B011_08	1	TE	CEDS	PS1.A	S & SM	Physical	Sensemaking	–0.155	.42	.32	73
2308B018_01	1	TE	DUM	LS2.A	C and E	Life	Sensemaking	–0.498	.50	.52	79
2308B018_03	1	MC	AQDP	LS2.A	C and E	Life	Investigating	–0.572	.51	.52	67
2308B018_04	1	TE	PACI	LS2.A	C and E	Life	Investigating	1.066	.21	.39	50
2308B018_07	1	TE	EAE	LS4.C	C and E	Life	Critiquing	0.817	.25	.42	52
2308B018_09	4	CR	OECI	LS4.B	C and E	Life	Critiquing	–0.294	.44	.66	333
2308M003_01	1	TE	CEDS	LS4.C	SF	Life	Sensemaking	0.216	.35	.31	53
2308M003_04	1	MC	AQDP	LS1.B	C and E	Life	Investigating	0.095	.37	.23	39
2308M003_07	1	MC	UMCT	LS3.B	C and E	Life	Investigating	1.240	.18	.49	43
2308M006_03	1	MC	PACI	LS2.C	S, P, and Q	Life	Investigating	1.250	.18	.12	43

Grade 8											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
2308M006_05	1	TE	DUM	LS2.C	C and E	Life	Sensemaking	0.655	.27	.41	67
2308M006_06	1	TE	EAE	ESS3.C	PAT	Earth and Space	Critiquing	1.184	.19	.33	58
2308M006_08	1	TE	EAE	LS2.C	C and E	Life	Critiquing	–0.233	.44	.35	79
2308M006_09	1	TE	DUM	ESS3.C	C and E	Earth and Space	Sensemaking	0.713	.26	.28	95
2308M010_01	1	TE	EAE	ESS1.C	S & SM	Earth and Space	Critiquing	0.923	.23	.30	51
2308M010_04	1	TE	EAE	ESS1.C	PAT	Earth and Space	Critiquing	0.862	.24	.38	48
2308M010_05	1	TE	DUM	ESS1.C	PAT	Earth and Space	Sensemaking	1.200	.19	.29	58
2308M010_07	1	MC	PACI	ESS1.C	PAT	Earth and Space	Investigating	–0.066	.41	.20	68
2308M020_05	1	TE	EAE	PS4.B	E&M	Physical	Critiquing	0.212	.39	.42	117
2308M020_09	1	TE	UMCT	PS3.B	PAT	Physical	Investigating	0.410	.32	.35	64
2308M020_11	1	MC	AQDP	PS4.A	C and E	Physical	Investigating	–1.025	.64	.28	44
2308M027_01	1	TE	AQDP	PS4.B	C and E	Physical	Investigating	0.189	.36	.34	30
2308M027_02	1	TE	PACI	PS4.B	C and E	Physical	Investigating	0.592	.24	.27	41
2308M027_03	1	TE	EAE	PS4.B	E&M	Physical	Critiquing	0.778	.22	.42	37
2308M027_05	1	TE	DUM	PS4.B	PAT	Physical	Sensemaking	0.340	.29	.59	46
2308M043_01	1	MC	AID	LS4.C	S, P, and Q	Life	Sensemaking	0.679	.27	.18	73
2308M043_03	1	TE	DUM	LS1.D	SF	Life	Sensemaking	0.961	.22	.26	65
2308M043_07	1	MC	AQDP	LS2.C	C and E	Life	Investigating	–0.336	.46	.24	46
2308M049_01	1	TE	DUM	LS4.B	C and E	Life	Sensemaking	0.605	.24	.53	55
2308M049_03	1	TE	AID	PS4.B	SF	Physical	Sensemaking	0.295	.29	.47	54
2308M049_05	1	TE	AID	LS2.A	E&M	Life	Sensemaking	0.634	.28	.38	70
2308M049_07	1	MC	AQDP	LS4.B	SF	Life	Investigating	–0.391	.48	.19	36
2408B005_01	1	TE	AQDP	PS3.B	S & SM	Physical	Investigating	0.782	.25	.39	48
2408B005_02	1	TE	CEDS	PS3.A	E&M	Physical	Sensemaking	–1.188	.64	.49	70
2408B005_03	1	TE	OECI	PS3.B	C and E	Physical	Critiquing	–1.022	.60	.44	35
2408B005_04	4	CR	CEDS	PS3.B	S & SM	Physical	Sensemaking	0.545	.28	.63	484
2408B005_05	1	TE	EAE	PS3.B	C and E	Physical	Critiquing	0.324	.33	.35	48
2408M003_01	1	MC	AQDP	ESS2.D	SC	Earth and Space	Investigating	–0.373	.47	.51	64

Grade 8											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
2408M003_02	1	TE	EAE	ESS2.D	PAT	Earth and Space	Critiquing	1.308	.18	.28	47
2408M003_03	1	MC	OECI	ESS2.D	C and E	Earth and Space	Critiquing	0.401	.32	.26	59
2408M039_02	1	TE	PACI	ESS1.A	PAT	Earth and Space	Investigating	0.038	.39	.46	72
2408M039_03	1	TE	EAE	ESS1.A	S & SM	Earth and Space	Critiquing	1.203	.19	.26	43
2408M039_08	1	MC	CEDS	ESS1.A	PAT	Earth and Space	Sensemaking	–0.800	.56	.32	46
2408M047_03	1	TE	AQDP	PS1.B	SF	Physical	Investigating	1.401	.16	.25	74
2408M047_04	1	TE	CEDS	LS1.C	C and E	Life	Sensemaking	0.128	.37	.30	127
2408M047_06	1	MC	OECI	PS3.D	S & SM	Physical	Critiquing	–0.525	.50	.30	88

Table F.3: Grade 11 Test Map–Metadata and Item Statistics

Grade 11											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
2011M028_01	1	MC	AQDP	LS4.D	C and E	Life	Investigating	–0.682	.64	.26	90
2011M028_03	1	TE	CEDS	ESS3.C	S & SM	Earth and Space	Sensemaking	0.237	.44	.40	75
2011M028_04	1	MC	EAE	LS4.D	S & SM	Life	Critiquing	1.348	.22	.24	84
2011M028_05	1	TE	PACI	ESS3.C	C and E	Earth and Space	Investigating	0.025	.51	.39	39
2011M031_03	1	TE	CEDS	ESS1.C	SC	Earth and Space	Sensemaking	1.447	.20	.32	62
2011M031_04	1	TE	EAE	ESS2.B	SC	Earth and Space	Critiquing	0.798	.31	.24	41
2011M031_05	1	MC	DUM	ESS1.C	S & SM	Earth and Space	Sensemaking	0.756	.28	.23	42
2011M031_06	1	MC	AQDP	ESS2.B	S & SM	Earth and Space	Investigating	0.533	.38	.26	27
2011M059_01	1	TE	AID	LS4.C	SC	Life	Sensemaking	0.858	.32	.57	145
2011M059_02	1	MC	CEDS	LS2.C	C and E	Life	Sensemaking	–0.025	.49	.29	61
2011M059_05	1	MC	AID	LS2.B	E&M	Life	Sensemaking	0.617	.36	.24	57
2011M059_08	1	MC	AQDP	LS2.C	C and E	Life	Investigating	2.033	.15	.11	25
2111B002_01	1	TE	EAE	LS4.C	C and E	Life	Critiquing	0.732	.36	.62	75
2111B002_03	1	TE	AID	LS2.A	SC	Life	Sensemaking	0.278	.46	.41	82
2111B002_07	1	TE	AQDP	LS2.A	SC	Life	Investigating	0.760	.38	.40	36
2111B002_10	1	TE	CEDS	LS2.A	C and E	Life	Sensemaking	1.177	.27	.47	85
2111B002_11	3	CR	PACI	LS2.C	C and E	Life	Investigating	–0.096	.51	.65	266
2111M020_01	1	MC	UMCT	PS2.A	S, P, and Q	Physical	Investigating	–0.593	.58	.43	59
2111M020_03	1	TE	CEDS	PS2.A	SF	Physical	Sensemaking	1.345	.22	.36	61
2111M020_05	1	TE	EAE	PS2.A	PAT	Physical	Critiquing	–0.057	.48	.51	43
2211M000_02	1	TE	EAE	LS1.A	PAT	Life	Critiquing	1.605	.20	.47	44
2211M000_03	1	MC	DUM	LS4.C	PAT	Life	Sensemaking	0.466	.39	.47	65
2211M000_06	1	TE	EAE	LS1.A	SF	Life	Critiquing	0.664	.35	.56	31
2211M000_08	1	MC	PACI	LS4.C	C and E	Life	Investigating	0.955	.30	.32	21
2211M000_09	1	TE	UMCT	LS4.C	PAT	Life	Investigating	1.996	.15	.43	61
2211M013_02	1	MC	OECI	ESS3.D	S, P, and Q	Earth and Space	Critiquing	0.177	.45	.42	59
2211M013_03	1	TE	DUM	ESS3.C	C and E	Earth and Space	Sensemaking	0.748	.34	.56	100

Grade 11											
UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	Rasch B	*p*-value	*rpb*	Median Time
2211M013_05	1	TE	PACI	ESS3.C	C and E	Earth and Space	Investigating	0.542	.38	.43	46
2211M016_01	1	TE	DUM	PS3.A	E&M	Physical	Sensemaking	0.548	.34	.48	76
2211M016_02	1	MC	UMCT	PS3.B	S & SM	Physical	Investigating	–0.239	.50	.38	56
2211M016_06	1	TE	OECI	ESS3.C	C and E	Earth and Space	Critiquing	0.711	.34	.37	35
2211M016_08	1	TE	CEDS	PS3.D	E&M	Physical	Sensemaking	1.537	.19	.47	57
2211M031_01	1	TE	AQDP	LS2.D	C and E	Life	Investigating	0.577	.37	.38	108
2211M031_03	1	TE	EAE	LS2.D	S, P, and Q	Life	Critiquing	1.085	.28	.43	72
2211M031_04	1	TE	CEDS	LS2.A	PAT	Life	Sensemaking	–0.337	.55	.51	82
2211M031_05	1	TE	OECI	LS2.D	C and E	Life	Critiquing	–0.350	.56	.42	55
2311B004_01	1	TE	UMCT	PS2.A	S, P, and Q	Physical	Investigating	0.289	.43	.56	34
2311B004_04	1	TE	DUM	PS1.B	E&M	Physical	Sensemaking	1.859	.17	.18	27
2311B004_05	3	CR	OECI	PS2.A	SF	Physical	Critiquing	0.948	.29	.61	194
2311B014_01	1	MC	AQDP	ESS2.D	C and E	Earth and Space	Investigating	0.378	.41	.38	42
2311B014_02	1	TE	DUM	ESS2.D	SC	Earth and Space	Sensemaking	0.861	.32	.42	137
2311B014_04	3	CR	OECI	ESS2.D	PAT	Earth and Space	Critiquing	–0.483	.60	.63	421
2311M011_03	1	TE	EAE	PS2.B	C and E	Physical	Critiquing	0.990	.33	.47	107
2311M011_05	1	MC	DUM	ESS1.B	S & SM	Earth and Space	Sensemaking	–0.699	.64	.28	54
2311M011_07	1	MC	AQDP	ESS1.B	C and E	Earth and Space	Investigating	–0.355	.59	.40	32
2311M015_01	1	TE	DUM	PS1.B	S, P, and Q	Physical	Sensemaking	–0.511	.59	.42	88
2311M015_03	1	TE	OECI	PS1.B	SF	Physical	Critiquing	0.249	.43	.47	47
2311M015_05	1	MC	UMCT	PS1.B	S, P, and Q	Physical	Investigating	1.142	.27	.43	75
2311M015_06	1	MC	EAE	PS1.A	SF	Physical	Critiquing	1.421	.23	.33	67
2311M019_01	1	TE	AID	PS2.A	C and E	Physical	Sensemaking	0.813	.33	.42	66
2311M019_02	1	MC	UMCT	PS2.A	C and E	Physical	Investigating	0.206	.44	.46	61
2311M019_03	1	TE	EAE	PS2.A	C and E	Physical	Critiquing	0.686	.35	.27	72
2311M019_06	1	MC	CEDS	PS2.B	S, P, and Q	Physical	Sensemaking	0.517	.38	.32	37
2311M038_01	1	MC	UMCT	PS4.A	S, P, and Q	Physical	Investigating	0.396	.41	.26	48
2311M038_03	1	TE	OECI	PS4.A	C and E	Physical	Critiquing	0.775	.33	.52	35

Grade 11											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p*-value**	***rpb***	**Median Time**
2311M038_04	1	TE	CEDS	PS4.A	C and E	Physical	Sensemaking	1.293	.25	.30	24
2311M053_01	1	MC	AQDP	LS4.B	S & SM	Life	Investigating	0.146	.46	.45	92
2311M053_03	1	MC	AID	LS3.B	SC	Life	Sensemaking	–0.195	.52	.45	56
2311M053_04	1	TE	PACI	LS3.B	C and E	Life	Investigating	1.189	.26	.44	58
2311M053_05	1	MC	AID	LS3.B	S, P, and Q	Life	Sensemaking	0.558	.37	.19	70
2311M053_06	1	TE	EAE	LS3.B	SC	Life	Critiquing	–0.902	.66	.41	56
2311M058_01	1	MC	AQDP	ESS1.B	S & SM	Earth and Space	Investigating	–0.024	.49	.35	33
2311M058_04	1	TE	DUM	PS4.B	S & SM	Physical	Sensemaking	0.755	.34	.27	29
2311M058_05	1	TE	OECI	ESS1.B	S & SM	Earth and Space	Critiquing	1.519	.21	.29	32
2411M017_01	1	MC	AID	ESS2.A	S, P, and Q	Earth and Space	Sensemaking	–0.650	.62	.44	53
2411M017_02	1	TE	PACI	ESS3.A	C and E	Earth and Space	Investigating	1.031	.29	.47	60
2411M017_04	1	TE	PACI	ESS3.A	C and E	Earth and Space	Investigating	0.337	.42	.51	62
2411M017_06	1	TE	OECI	ESS3.D	S & SM	Earth and Space	Critiquing	1.387	.23	.43	58
2411M030_02	1	TE	AQDP	ESS1.C	PAT	Earth and Space	Investigating	0.616	.36	.42	29
2411M030_03	1	MC	DUM	ESS1.C	SC	Earth and Space	Sensemaking	0.416	.40	.26	66
2411M030_07	1	TE	CEDS	ESS1.C	S & SM	Earth and Space	Sensemaking	1.212	.26	.58	36
2411M030_09	1	TE	OECI	ESS1.C	PAT	Earth and Space	Critiquing	–0.395	.57	.35	42

APPENDIX G: Scale Score Cumulative Frequency Distributions

Table G.1: Grade 5 Scale Score Cumulative Frequency Distribution

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
0	100	7	0.01	0.01	0.01	0.00	0.02	0.01	0.01
1	100	41	0.04	0.03	0.06	0.02	0.08	0.06	0.03
2	100	190	0.20	0.17	0.22	0.05	0.32	0.33	0.09
3	100	590	0.61	0.56	0.66	0.20	1.18	0.92	0.28
4	100	1,520	1.58	1.46	1.70	0.39	3.07	2.40	0.69
5	100	3,222	3.35	3.09	3.60	0.98	6.32	5.10	1.47
6	103	5,579	5.80	5.38	6.21	1.57	10.59	8.91	2.66
7	110	8,661	9.00	8.42	9.57	2.36	15.76	13.89	4.27
8	117	12,149	12.63	12.07	13.17	3.42	21.90	19.30	6.21
9	123	5,798	16.42	15.78	17.05	4.59	28.05	24.86	8.40
10	128	19,462	20.23	19.55	20.89	5.83	34.16	30.29	10.75
11	133	22,951	23.85	23.21	24.48	6.94	39.36	35.55	13.10
12	138	26,310	27.35	26.90	27.78	8.22	44.34	40.45	15.42
13	143	29,419	30.58	30.33	30.82	9.72	48.87	44.73	17.80
14	150	32,481	33.76	33.76	33.76	11.14	52.99	48.68	20.48
15	151	35,270	36.66	36.83	36.50	12.52	56.35	52.25	23.08
16	155	38,082	39.58	39.97	39.21	13.99	59.75	55.77	25.73
17	158	40,810	42.42	43.06	41.79	15.72	62.72	59.09	28.42
18	162	43,255	44.96	45.78	44.16	17.23	65.41	61.91	30.94
19	165	45,726	47.52	48.51	46.57	18.86	68.13	64.47	33.75
20	168	48,104	50.00	51.17	48.86	20.90	70.38	66.97	36.46
21	172	50,369	52.35	53.60	51.13	22.80	72.59	69.21	39.11
22	175	52,634	54.70	56.05	53.39	24.74	74.67	71.44	41.83
23	178	54,931	57.09	58.55	55.67	26.57	76.70	73.65	44.68
24	181	57,155	59.40	60.98	57.87	28.66	78.58	75.64	47.50
25	184	59,371	61.71	63.36	60.10	31.03	80.31	77.67	50.21
26	187	61,425	63.84	65.60	62.13	33.39	81.95	79.43	52.78
27	190	63,416	65.91	67.72	64.15	35.50	83.51	81.15	55.36
28	193	65,477	68.05	69.97	66.19	37.90	85.03	82.77	58.08
29	196	67,249	69.89	71.84	68.00	39.94	86.43	84.11	60.45
30	200	69,465	72.20	74.08	70.36	42.64	88.04	85.89	63.38
31	202	71,475	74.29	76.16	72.47	45.55	89.36	87.31	66.10
32	205	73,400	76.29	78.14	74.49	48.01	90.66	88.61	68.84
33	208	75,241	78.20	80.08	76.37	50.77	91.78	89.86	71.39
34	211	77,025	80.06	81.94	78.22	53.65	92.69	91.08	73.84
35	214	78,711	81.81	83.63	80.03	56.41	93.76	92.19	76.11
36	218	80,343	83.50	85.30	81.76	59.31	94.61	93.11	78.44
37	221	81,904	85.13	86.91	83.39	62.34	95.33	93.97	80.65
38	224	83,371	86.65	88.38	84.97	65.13	95.93	94.85	82.70
39	228	84,819	88.16	89.82	86.54	68.19	96.56	95.62	84.75
40	231	86,142	89.53	91.09	88.02	70.91	97.15	96.30	86.58
41	234	87,414	90.85	92.29	89.46	73.73	97.65	96.90	88.35

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
42	238	88,589	92.07	93.39	90.79	76.56	98.11	97.44	89.97
43	243	89,676	93.20	94.43	92.01	79.27	98.47	97.85	91.53
44	246	90,709	94.28	95.38	93.20	82.01	98.76	98.28	92.91
45	250	91,652	95.26	96.21	94.32	84.40	99.06	98.62	94.25
46	254	92,468	96.11	96.95	95.28	86.80	99.21	98.93	95.36
47	258	93,173	96.84	97.56	96.13	88.83	99.41	99.22	96.29
48	263	93,861	97.55	98.09	97.03	90.98	99.64	99.42	97.19
49	268	94,450	98.17	98.57	97.77	93.09	99.75	99.58	97.93
50	273	94,913	98.65	98.98	98.32	94.66	99.84	99.70	98.51
51	279	95,316	99.07	99.30	98.84	96.24	99.92	99.79	99.00
52	285	95,614	99.38	99.54	99.21	97.40	99.96	99.87	99.33
53	292	95,861	99.63	99.73	99.54	98.39	99.98	99.93	99.62
54	300	96,024	99.80	99.86	99.74	99.06	99.99	99.97	99.81
55	300	96,113	99.89	99.93	99.86	99.47	100.00	99.99	99.91
56	300	96,175	99.96	99.97	99.94	99.79	100.00	100.00	99.96
57	300	96,198	99.98	99.99	99.98	99.90	100.00	100.00	99.99
58	300	96,209	99.99	99.99	100.00	99.94	100.00	100.00	100.00
59	300	96,215	100.00	100.00	100.00	100.00	100.00	100.00	100.00
60	300	96,215	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table G.2: Grade 8 Scale Score Cumulative Frequency Distribution

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
0	100	5	0.01	0.00	0.01	0.00	0.01	0.01	0.00
1	100	29	0.03	0.01	0.05	0.00	0.08	0.03	0.02
2	100	92	0.09	0.06	0.12	0.01	0.22	0.13	0.04
3	100	296	0.30	0.24	0.35	0.04	0.66	0.39	0.16
4	100	745	0.75	0.63	0.86	0.13	1.60	1.05	0.36
5	100	1,553	1.57	1.28	1.83	0.34	2.95	2.29	0.77
6	102	2,921	2.94	2.44	3.42	0.70	5.34	4.45	1.38
7	108	4,938	4.98	4.18	5.73	1.13	9.14	7.42	2.40
8	114	7,470	7.53	6.39	8.60	1.73	13.29	11.33	3.73
9	119	10,545	10.63	9.24	11.95	2.39	18.26	16.14	5.36
10	124	13,776	13.88	12.27	15.41	3.19	23.42	21.04	7.20
11	128	17,213	17.35	15.56	19.04	4.26	28.44	26.18	9.31
12	132	20,561	20.72	18.86	22.49	5.12	33.85	30.91	11.46
13	136	24,005	24.19	22.33	25.97	6.08	38.70	35.92	13.82
14	139	27,287	27.50	25.67	29.24	7.14	43.35	40.52	16.15
15	143	30,475	30.71	29.03	32.32	8.27	47.70	44.82	18.64
16	146	33,516	33.78	32.20	35.29	9.61	51.74	48.66	21.17
17	150	36,576	36.86	35.32	38.34	11.00	55.49	52.46	23.89
18	152	39,458	39.77	38.48	40.99	12.14	58.83	56.02	26.58
19	155	42,243	42.58	41.41	43.69	13.64	62.13	59.18	29.25
20	158	45,036	45.39	44.46	46.28	15.21	65.14	62.33	32.05
21	160	47,697	48.07	47.24	48.86	16.84	67.84	65.25	34.80
22	163	50,359	50.76	50.18	51.31	18.45	70.33	68.09	37.74
23	165	52,943	53.36	53.09	53.62	20.21	72.78	70.72	40.62
24	168	55,387	55.82	55.75	55.90	21.96	74.91	73.17	43.39
25	170	57,730	58.18	58.28	58.10	23.65	77.28	75.24	46.23
26	173	60,082	60.55	60.87	60.26	25.74	79.30	77.40	49.03
27	175	62,339	62.83	63.34	62.35	27.72	81.24	79.41	51.75
28	177	64,497	65.00	65.71	64.35	29.89	82.85	81.22	54.44
29	180	66,570	67.09	67.86	66.38	32.08	84.45	82.93	57.03
30	182	68,589	69.13	69.94	68.37	34.30	85.78	84.46	59.69
31	184	70,577	71.13	72.01	70.30	36.60	87.18	85.97	62.25
32	187	72,461	73.03	74.02	72.10	38.89	88.50	87.32	64.74
33	189	74,204	74.79	75.89	73.75	41.37	89.73	88.51	67.00
34	191	75,956	76.55	77.73	75.44	44.24	90.78	89.58	69.31
35	193	77,660	78.27	79.46	77.15	46.66	91.86	90.67	71.60
36	195	79,195	79.82	81.06	78.64	49.05	92.73	91.63	73.64
37	198	80,357	80.99	82.28	79.77	51.07	93.33	92.36	75.15
38	200	82,162	82.81	84.09	81.60	54.31	94.20	93.39	77.59
39	202	83,699	84.36	85.56	83.22	57.35	94.95	94.15	79.69
40	204	85,035	85.70	86.88	84.59	59.86	95.68	94.84	81.52
41	207	86,299	86.98	88.11	85.91	62.55	96.24	95.48	83.22
42	209	87,555	88.24	89.34	87.20	65.39	96.75	96.03	84.95

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
43	211	88,729	89.43	90.52	88.39	67.64	97.27	96.53	86.65
44	214	89,869	90.58	91.63	89.58	70.23	97.67	97.01	88.25
45	216	90,876	91.59	92.58	90.65	72.89	98.01	97.39	89.60
46	219	91,835	92.56	93.47	91.70	75.29	98.38	97.77	90.88
47	221	92,741	93.47	94.31	92.67	77.71	98.63	98.12	92.09
48	224	93,596	94.33	95.14	93.57	79.98	98.93	98.39	93.28
49	226	94,388	95.13	95.78	94.51	82.27	99.12	98.64	94.34
50	231	95,134	95.88	96.46	95.34	84.70	99.29	98.91	95.25
51	232	95,783	96.54	97.06	96.05	86.58	99.47	99.14	96.08
52	234	96,359	97.12	97.58	96.68	88.51	99.57	99.30	96.81
53	237	96,886	97.65	98.06	97.26	90.33	99.68	99.46	97.43
54	240	97,375	98.14	98.48	97.82	92.21	99.77	99.58	98.01
55	243	97,760	98.53	98.79	98.28	93.62	99.84	99.71	98.44
56	246	98,113	98.89	99.12	98.67	94.95	99.86	99.77	98.88
57	250	98,412	99.19	99.38	99.00	96.20	99.91	99.84	99.21
58	253	98,609	99.39	99.55	99.23	97.02	99.94	99.89	99.42
59	257	98,774	99.55	99.67	99.44	97.75	99.97	99.92	99.59
60	261	98,913	99.69	99.79	99.59	98.44	99.98	99.95	99.72
61	265	99,024	99.80	99.88	99.73	99.02	99.99	99.97	99.81
62	269	99,094	99.87	99.92	99.83	99.32	99.99	99.99	99.89
63	274	99,141	99.92	99.95	99.89	99.56	99.99	99.99	99.94
64	279	99,171	99.95	99.97	99.93	99.76	99.99	99.99	99.96
65	285	99,195	99.98	99.99	99.96	99.88	100.00	99.99	99.98
66	292	99,208	99.99	99.99	99.98	99.96	100.00	100.00	99.99
67	300	99,216	100.00	100.00	100.00	99.99	100.00	100.00	99.99
68	300	99,217	100.00	100.00	100.00	100.00	100.00	100.00	99.99
69	300	99,219	100.00	100.00	100.00	100.00	100.00	100.00	100.00
70	100	99,219	100.00	100.00	100.00	100.00	100.00	100.00	100.00
71	100	99,219	100.00	100.00	100.00	100.00	100.00	100.00	100.00
72	100	99,219	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table G.3: Grade 11 Scale Score Cumulative Frequency Distribution

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
0	100	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	100	5	0.01	0.00	0.01	0.00	0.01	0.01	0.00
2	100	28	0.03	0.02	0.04	0.00	0.06	0.04	0.02
3	100	65	0.07	0.04	0.09	0.01	0.15	0.08	0.04
4	100	183	0.19	0.13	0.24	0.03	0.33	0.26	0.12
5	100	452	0.47	0.36	0.57	0.07	0.80	0.70	0.26
6	100	917	0.94	0.77	1.11	0.19	1.62	1.41	0.52
7	100	1,74	1.79	1.52	2.05	0.34	3.18	2.70	0.94
8	100	2,940	3.03	2.61	3.43	0.59	5.36	4.48	1.66
9	100	4,603	4.74	4.12	5.33	1.06	8.33	6.88	2.66
10	100	6,672	6.87	6.06	7.65	1.51	12.00	10.08	3.86
11	101	8,976	9.24	8.26	10.19	2.01	15.75	13.67	5.23
12	107	11,547	11.89	10.65	13.08	2.55	19.94	17.56	6.88
13	112	14,241	14.66	13.29	15.99	3.30	24.02	21.67	8.63
14	117	17,027	17.53	16.02	18.99	4.21	28.35	25.75	10.50
15	121	19,671	20.25	18.59	21.87	5.02	32.27	29.59	12.40
16	125	22,291	22.94	21.32	24.53	6.06	35.96	33.32	14.36
17	130	24,840	25.57	23.95	27.15	7.04	39.57	36.88	16.30
18	134	27,355	28.16	26.72	29.57	8.12	42.93	40.18	18.41
19	138	29,856	30.73	29.45	32.00	9.10	46.32	43.58	20.46
20	141	32,309	33.25	32.08	34.42	9.95	49.41	46.66	22.78
21	145	34,723	35.74	34.72	36.76	11.01	52.33	49.60	25.11
22	149	37,039	38.12	37.29	38.97	12.21	55.04	52.55	27.21
23	152	39,393	40.55	39.96	41.16	13.48	58.11	55.33	29.44
24	158	41,741	42.96	42.59	43.37	14.64	60.92	58.04	31.80
25	159	43,916	45.20	44.98	45.46	15.91	63.34	60.55	33.98
26	162	46,106	47.46	47.37	47.59	17.31	65.51	63.03	36.35
27	165	48,241	49.65	49.73	49.63	18.72	67.71	65.33	38.66
28	168	50,306	51.78	52.07	51.54	20.36	69.80	67.46	40.93
29	171	52,282	53.81	54.30	53.40	21.65	71.64	69.52	43.24
30	175	54,256	55.84	56.47	55.30	23.23	73.59	71.48	45.50
31	178	56,268	57.92	58.73	57.19	25.06	75.55	73.36	47.85
32	181	58,134	59.84	60.83	58.93	26.73	77.43	75.07	50.02
33	184	59,994	61.75	62.92	60.68	28.40	79.04	76.86	52.22
34	187	61,829	63.64	64.93	62.44	30.22	80.39	78.46	54.55
35	190	63,589	65.45	66.91	64.10	31.84	81.93	79.98	56.72
36	193	65,316	67.23	68.87	65.69	33.79	83.20	81.52	58.85
37	196	66,732	68.69	70.41	67.07	35.58	84.35	82.63	60.60
38	200	68,537	70.54	72.35	68.85	37.58	85.63	84.13	62.93
39	202	70,120	72.17	74.05	70.42	39.73	86.73	85.36	64.94
40	204	71,623	73.72	75.71	71.85	41.77	87.87	86.49	66.88
41	207	73,101	75.24	77.25	73.36	44.01	88.95	87.56	68.75
42	210	74,525	76.71	78.73	74.81	46.06	89.82	88.61	70.60

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
43	213	75,964	78.19	80.23	76.27	48.09	90.79	89.58	72.56
44	216	77,241	79.50	81.61	77.52	50.12	91.67	90.44	74.21
45	220	78,516	80.81	82.93	78.83	51.96	92.51	91.22	75.96
46	223	79,773	82.11	84.27	80.07	53.98	93.23	92.09	77.62
47	226	80,968	83.34	85.40	81.39	56.20	93.85	92.84	79.23
48	229	82,119	84.52	86.66	82.50	58.46	94.47	93.55	80.75
49	232	83,244	85.68	87.80	83.68	60.57	95.11	94.16	82.30
50	235	84,323	86.79	88.87	84.82	62.71	95.65	94.81	83.69
51	238	85,400	87.90	89.92	86.00	65.01	96.16	95.42	85.10
52	242	86,382	88.91	90.87	87.06	67.12	96.73	95.86	86.45
53	245	87,324	89.88	91.78	88.08	69.22	97.16	96.29	87.76
54	250	88,197	90.78	92.61	89.04	71.48	97.50	96.72	88.85
55	252	89,095	91.70	93.45	90.04	73.86	97.87	97.13	90.00
56	256	89,909	92.54	94.19	90.98	76.01	98.21	97.54	91.03
57	259	90,657	93.31	94.85	91.85	78.17	98.44	97.85	91.98
58	263	91,448	94.12	95.57	92.75	80.36	98.76	98.14	93.01
59	267	92,201	94.90	96.23	93.63	82.49	99.12	98.47	93.94
60	271	92,766	95.48	96.71	94.32	84.09	99.30	98.71	94.66
61	275	93,385	96.12	97.27	95.01	86.13	99.38	98.93	95.44
62	280	93,931	96.68	97.73	95.68	87.85	99.49	99.12	96.15
63	284	94,469	97.23	98.17	96.35	89.77	99.62	99.30	96.79
64	289	94,922	97.70	98.52	96.92	91.38	99.68	99.49	97.29
65	294	95,332	98.12	98.81	97.47	92.79	99.75	99.58	97.83
66	299	95,733	98.54	99.14	97.95	94.31	99.82	99.71	98.30
67	300	96,054	98.87	99.37	98.38	95.43	99.89	99.80	98.69
68	300	96,347	99.17	99.56	98.79	96.65	99.93	99.86	99.02
69	300	96,563	99.39	99.71	99.08	97.54	99.94	99.90	99.30
70	300	96,741	99.57	99.80	99.36	98.31	99.98	99.93	99.50
71	300	96,897	99.73	99.87	99.60	98.83	99.99	99.95	99.73
72	300	96,998	99.84	99.92	99.75	99.31	100.00	99.97	99.83
73	300	97,074	99.92	99.97	99.87	99.67	100.00	99.99	99.89
74	300	97,113	99.96	99.99	99.93	99.78	100.00	100.00	99.95
75	300	97,139	99.98	99.99	99.97	99.89	100.00	100.00	99.98
76	300	97,153	100.00	100.00	100.00	99.99	100.00	100.00	99.99
77	300	97,155	100.00	100.00	100.00	100.00	100.00	100.00	100.00
78	300	97,156	100.00	100.00	100.00	100.00	100.00	100.00	100.00

APPENDIX H: Item Parameters and Model-Fit Tables

Table H.1: Grade 5 IRT Item Parameters and Fit Statistics

NJSLA–S Grade 5							
Item	Rasch B	Infit	Outfit	*rpb*	Rasch Discrim.	Lower Asym.	*p*-value
1905B001_04B	–0.655	0.92	0.86	.44	1.18	0.00	.67
1905B001_08B	0.831	1.00	0.96	.42	1.03	0.00	.38
1905B001_11	0.139	1.20	1.25	.23	0.44	0.17	.52
1905M069_01	0.863	0.87	0.81	.54	1.31	0.00	.38
1905M069_02	1.131	0.81	0.74	.54	1.37	0.00	.30
1905M069_05	1.569	0.75	0.64	.52	1.38	0.00	.21
2005M015_02	–0.415	0.97	0.96	.36	1.07	0.06	.67
2005M015_04	1.009	1.04	1.08	.37	0.91	0.03	.35
2005M015_05	0.934	1.06	1.06	.34	0.89	0.01	.34
2005M015_07	0.696	0.82	0.77	.57	1.45	0.00	.39
2005M015_10	0.875	1.02	1.01	.38	0.96	0.00	.37
2105M003_01	1.157	0.94	0.93	.46	1.11	0.00	.32
2105M003_04	0.520	1.07	1.09	.35	0.81	0.04	.44
2105M003_07	–0.756	1.00	1.03	.35	1.00	0.00	.69
2105M003_09	1.557	0.86	0.80	.52	1.21	0.00	.26
2205B000_02	2.044	1.10	1.06	.40	0.90	0.02	.22
2205B000_03	0.409	1.00	1.01	.42	1.00	0.02	.46
2205B000_04	0.409	0.78	0.73	.62	1.59	0.00	.49
2205B000_07	0.793	0.88	0.81	.74	1.11	0.00	.40
2205B000_10	1.114	1.09	1.15	.30	0.81	0.03	.32
2205B007_02	1.398	0.95	0.90	.45	1.10	0.00	.28
2205B007_03	–0.153	1.12	1.22	.53	0.82	0.00	.59
2205B007_04	0.571	0.96	0.93	.45	1.12	0.00	.43
2205B007_05	1.895	1.03	1.20	.32	0.93	0.02	.21
2205M007_01	0.902	0.81	0.75	.59	1.43	0.00	.37
2205M007_02	1.688	1.10	1.26	.27	0.82	0.04	.24
2205M007_04	0.243	0.92	0.89	.49	1.24	0.00	.50
2205M007_05	0.473	1.07	1.08	.35	0.82	0.01	.45
2205M024_03	0.547	1.05	1.06	.37	0.87	0.03	.44
2205M024_04	1.371	1.22	1.33	.19	0.62	0.07	.29
2205M024_05	1.111	0.85	0.81	.55	1.29	0.00	.33
2305B004_01	0.589	0.97	0.97	.44	1.07	0.00	.43
2305B004_03	0.509	1.06	1.07	.36	0.84	0.02	.45
2305B004_04	1.647	1.12	1.24	.27	0.81	0.03	.24
2305B004_09	1.358	1.09	1.03	.57	0.94	0.00	.28
2305B006_01	0.508	0.94	0.91	.47	1.17	0.00	.43
2305B006_03	0.812	1.01	1.04	.41	0.96	0.02	.40

NJSLA–S Grade 5							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p*-value**
2305B006_04	0.875	1.16	1.27	.25	0.61	0.08	.36
2305B011_01	1.491	1.05	1.22	.33	0.88	0.03	.27
2305B011_02	0.570	1.21	1.28	.23	0.46	0.12	.43
2305B011_03	2.540	1.03	1.22	.27	0.95	0.01	.13
2305B011_04	0.593	0.93	0.91	.48	1.18	0.00	.43
2305B011_06	0.853	0.95	0.90	.46	1.14	0.00	.38
2405B002_01	1.730	1.13	1.36	.23	0.77	0.04	.23
2405B002_02	0.500	0.98	0.99	.43	1.03	0.01	.45
2405B002_03	0.296	0.86	0.82	.54	1.39	0.00	.49
2405B002_04	1.560	0.94	0.97	.43	1.07	0.00	.26
2405B008_03	2.085	0.95	0.79	.42	1.08	0.00	.18
2405B008_05	2.028	1.05	1.30	.28	0.90	0.02	.19
2405B008_06	1.659	1.04	1.12	.34	0.93	0.02	.24
2405M008_02	1.905	0.88	0.79	.49	1.16	0.00	.21
2405M008_04	0.379	1.12	1.14	.32	0.69	0.11	.47
2405M008_05	1.268	1.04	1.14	.36	0.90	0.03	.30

Table H.2: Grade 8 IRT Item Parameters and Fit Statistics

NJSLA–S Grade 8							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p*-value**
2008M025_01	0.086	0.90	0.85	.48	1.28	0.00	.38
2008M025_02	0.145	1.00	1.02	.37	1.00	0.01	.36
2008M025_08	0.216	1.16	1.25	.19	0.59	0.09	.35
2108M000_02	0.213	0.92	0.90	.45	1.19	0.00	.35
2108M000_03	0.698	0.81	0.71	.49	1.32	0.00	.23
2108M000_06	0.893	1.06	1.09	.22	0.91	0.01	.22
2308B003_01	0.044	0.90	0.88	.56	1.27	0.00	.46
2308B003_03	0.397	0.95	0.90	.43	1.12	0.00	.32
2308B003_05	0.064	0.95	0.95	.59	1.01	0.00	.38
2308B003_07	0.937	1.00	1.03	.43	0.99	0.01	.26
2308B003_11	–0.704	0.90	0.87	.45	1.35	0.00	.59
2308B006_01	1.133	0.98	1.05	.29	1.01	0.00	.19
2308B006_07	1.027	1.18	1.38	.18	0.74	0.05	.23
2308B006_08	0.837	0.97	1.01	.33	1.02	0.00	.23
2308B011_01	0.990	1.12	1.25	.19	0.81	0.04	.22
2308B011_06	0.046	0.97	0.97	.41	1.08	0.00	.38
2308B011_07	–0.625	0.90	0.88	.47	1.35	0.00	.52
2308B011_08	–0.155	1.05	1.07	.32	0.84	0.04	.42
2308B018_01	–0.498	0.85	0.82	.52	1.50	0.00	.50
2308B018_03	–0.572	0.85	0.82	.52	1.50	0.00	.51
2308B018_04	1.066	0.94	0.95	.39	1.06	0.00	.21
2308B018_07	0.817	0.93	0.95	.42	1.09	0.00	.25
2308B018_09	–0.294	0.93	0.93	.66	1.11	0.00	.44
2308M003_01	0.216	1.06	1.07	.31	0.87	0.03	.35
2308M003_04	0.095	1.13	1.20	.23	0.64	0.08	.37
2308M003_07	1.240	0.85	0.78	.50	1.18	0.00	.18
2308M006_03	1.250	1.13	1.71	.12	0.76	0.05	.18
2308M006_05	0.655	0.95	1.02	.41	1.06	0.00	.27
2308M006_06	1.184	0.99	1.03	.33	1.00	0.00	.19
2308M006_08	–0.233	1.02	1.00	.35	0.97	0.00	.44
2308M006_09	0.713	1.06	1.07	.29	0.90	0.02	.26
2308M010_01	0.923	1.04	1.02	.30	0.95	0.01	.23
2308M010_04	0.862	0.97	0.93	.38	1.05	0.00	.24
2308M010_05	1.200	1.02	1.16	.29	0.95	0.01	.19
2308M010_07	–0.066	1.17	1.25	.20	0.49	0.13	.41
2308M020_05	0.212	1.00	1.00	.42	0.99	0.01	.39
2308M020_09	0.410	1.03	1.03	.35	0.95	0.01	.32
2308M020_11	–1.025	1.02	1.05	.28	0.93	0.04	.64
2308M027_01	0.189	1.03	1.02	.34	0.94	0.01	.36

NJSLA–S Grade 8							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p*-value**
2308M027_02	0.592	0.98	1.06	.27	1.01	0.00	.24
2308M027_03	0.778	0.86	0.82	.42	1.21	0.00	.22
2308M027_05	0.340	0.74	0.66	.59	1.56	0.00	.29
2308M043_01	0.679	1.15	1.35	.18	0.70	0.07	.27
2308M043_03	0.961	1.06	1.21	.26	0.89	0.02	.22
2308M043_07	–0.336	1.13	1.17	.24	0.58	0.11	.46
2308M049_01	0.605	0.77	0.70	.53	1.41	0.00	.24
2308M049_03	0.295	0.84	0.84	.47	1.33	0.00	.29
2308M049_05	0.634	0.98	1.04	.38	1.02	0.00	.28
2308M049_07	–0.391	1.18	1.21	.19	0.41	0.16	.48
2408B005_01	0.782	0.96	0.95	.39	1.06	0.00	.25
2408B005_02	–1.188	0.85	0.80	.49	1.41	0.00	.64
2408B005_03	–1.022	0.90	0.88	.44	1.29	0.00	.60
2408B005_04	0.545	0.99	0.91	.63	1.05	0.00	.28
2408B005_05	0.324	1.02	1.00	.35	0.97	0.00	.33
2408M003_01	–0.373	0.87	0.85	.51	1.41	0.00	.47
2408M003_02	1.308	1.04	1.07	.28	0.96	0.01	.18
2408M003_03	0.401	1.10	1.09	.26	0.81	0.03	.32
2408M039_02	0.038	0.92	0.87	.46	1.23	0.00	.39
2408M039_03	1.203	1.04	1.18	.27	0.93	0.01	.19
2408M039_08	–0.800	1.04	1.04	.32	0.87	0.05	.56
2408M047_03	1.401	1.05	1.07	.25	0.95	0.01	.16
2408M047_04	0.128	1.07	1.08	.30	0.83	0.03	.37
2408M047_06	–0.525	1.08	1.08	.30	0.75	0.09	.50

Table H.3: Grade 11 IRT Item Parameters and Fit Statistics

NJSLA–S Grade 11							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p*-value**
2011M028_01	–0.682	1.10	1.17	.26	0.71	0.19	.64
2011M028_03	0.237	1.01	1.01	.40	0.97	0.03	.44
2011M028_04	1.348	1.08	1.28	.24	0.85	0.03	.22
2011M028_05	0.025	1.02	1.02	.39	0.96	0.00	.51
2011M031_03	1.447	0.97	1.04	.32	1.02	0.00	.20
2011M031_04	0.798	1.12	1.22	.24	0.74	0.06	.31
2011M031_05	0.756	1.10	1.13	.23	0.80	0.02	.28
2011M031_06	0.533	1.16	1.22	.26	0.62	0.06	.38
2011M059_01	0.858	0.83	0.74	.57	1.33	0.00	.32
2011M059_02	–0.025	1.13	1.15	.29	0.63	0.12	.49
2011M059_05	0.617	1.17	1.24	.24	0.61	0.08	.36
2011M059_08	2.033	1.19	1.74	.11	0.75	0.04	.15
2111B002_01	0.732	0.80	0.74	.62	1.41	0.00	.36
2111B002_03	0.278	1.01	1.00	.41	0.98	0.00	.46
2111B002_07	0.760	1.08	1.06	.41	0.86	0.03	.38
2111B002_10	1.177	0.95	0.84	.48	1.10	0.00	.27
2111B002_11	–0.096	0.88	0.90	.65	1.10	0.00	.51
2111M020_01	–0.593	0.98	0.96	.43	1.06	0.00	.58
2111M020_03	1.345	0.96	1.08	.36	1.02	0.00	.22
2111M020_05	–0.057	0.88	0.86	.51	1.35	0.00	.48
2211M000_02	1.605	0.90	0.82	.47	1.13	0.00	.20
2211M000_03	0.466	0.94	0.93	.47	1.15	0.00	.39
2211M000_06	0.664	0.84	0.80	.56	1.34	0.00	.35
2211M000_08	0.955	1.08	1.13	.32	0.85	0.03	.30
2211M000_09	1.996	0.92	0.81	.43	1.09	0.00	.15
2211M013_02	0.177	0.99	0.98	.42	1.04	0.00	.45
2211M013_03	0.748	0.84	0.76	.56	1.34	0.00	.34
2211M013_05	0.542	0.98	0.99	.43	1.04	0.00	.38
2211M016_01	0.548	0.90	0.86	.48	1.25	0.00	.34
2211M016_02	–0.239	1.02	1.03	.38	0.92	0.00	.50
2211M016_06	0.711	1.03	1.06	.37	0.92	0.02	.34
2211M016_08	1.537	0.83	0.76	.47	1.20	0.00	.19
2211M031_01	0.577	1.03	1.07	.38	0.92	0.03	.37
2211M031_03	1.085	0.97	0.91	.43	1.07	0.00	.28
2211M031_04	–0.337	0.86	0.86	.51	1.39	0.00	.55
2211M031_05	–0.350	0.97	0.96	.42	1.10	0.00	.56
2311B004_01	0.289	0.84	0.80	.56	1.42	0.00	.43
2311B004_04	1.859	1.15	1.48	.18	0.80	0.04	.17
2311B004_05	0.948	1.04	0.99	.61	1.00	0.00	.29

NJSLA–S Grade 11							
Item	Rasch B	Infit	Outfit	*rpb*	Rasch Discrim.	Lower Asym.	*p*-value
2311B014_01	0.378	1.03	1.03	.38	0.93	0.02	.41
2311B014_02	0.861	0.98	0.99	.42	1.03	0.00	.32
2311B014_04	–0.483	0.84	0.85	.63	1.19	0.00	.60
2311M011_03	0.990	1.00	0.95	.47	1.02	0.00	.33
2311M011_05	–0.699	1.08	1.15	.28	0.78	0.16	.64
2311M011_07	–0.355	0.98	0.96	.40	1.07	0.07	.59
2311M015_01	–0.511	0.96	0.94	.42	1.13	0.00	.59
2311M015_03	0.249	0.93	0.91	.47	1.18	0.00	.43
2311M015_05	1.142	0.95	0.99	.43	1.06	0.00	.27
2311M015_06	1.421	1.04	1.15	.33	0.92	0.02	.23
2311M019_01	0.813	0.98	0.96	.42	1.04	0.00	.33
2311M019_02	0.206	0.94	0.93	.46	1.16	0.00	.44
2311M019_03	0.686	1.15	1.15	.27	0.70	0.05	.35
2311M019_06	0.517	1.10	1.12	.32	0.77	0.05	.38
2311M038_01	0.396	1.15	1.22	.26	0.59	0.12	.41
2311M038_03	0.775	0.88	0.82	.52	1.25	0.00	.33
2311M038_04	1.293	1.08	1.17	.30	0.87	0.03	.25
2311M053_01	0.146	0.96	0.97	.45	1.10	0.02	.46
2311M053_03	–0.195	0.94	0.92	.45	1.18	0.00	.52
2311M053_04	1.189	0.94	0.93	.45	1.08	0.00	.26
2311M053_05	0.558	1.23	1.30	.19	0.47	0.09	.37
2311M053_06	–0.902	0.93	0.91	.41	1.16	0.00	.66
2311M058_01	–0.024	1.06	1.07	.35	0.82	0.07	.49
2311M058_04	0.755	1.14	1.20	.27	0.71	0.06	.34
2311M058_05	1.519	1.05	1.29	.30	0.89	0.03	.21
2411M017_01	–0.650	0.92	0.92	.44	1.20	0.00	.62
2411M017_02	1.031	0.93	0.90	.47	1.12	0.00	.29
2411M017_04	0.337	0.89	0.88	.51	1.27	0.00	.42
2411M017_06	1.387	0.96	0.90	.43	1.06	0.00	.23
2411M030_02	0.616	0.99	0.99	.42	1.02	0.00	.36
2411M030_03	0.416	1.16	1.27	.26	0.57	0.14	.40
2411M030_07	1.212	0.80	0.74	.58	1.29	0.00	.26
2411M030_09	–0.395	1.03	1.08	.35	0.90	0.00	.57

APPENDIX I: Raw Score-to-Scale Score Conversion Tables

Table I.1: Grade 5 Operational Raw-to-Scale Score Conversion

NJSLA–S Grade 5							
Raw Score	Theta	Standard Error	Unrounded Scale Score	Scale Score (SS)	SS CSEM	SS Lower Bound	SS Upper Bound
0	–4.512	1.832	–29.949	100	20	100	120
1	–3.291	1.012	21.879	100	20	100	120
2	–2.575	0.723	52.296	100	20	100	120
3	–2.147	0.597	70.483	100	20	100	120
4	–1.836	0.522	83.659	100	20	100	120
5	–1.591	0.472	94.089	100	20	100	120
6	–1.386	0.435	102.785	103	18	100	121
7	–1.209	0.407	110.289	110	17	100	127
8	–1.053	0.384	116.921	117	16	101	133
9	–0.912	0.366	122.892	123	16	107	139
10	–0.784	0.351	128.349	128	15	113	143
11	–0.665	0.338	133.393	133	14	119	147
12	–0.554	0.328	138.094	138	14	124	152
13	–0.450	0.318	142.514	143	14	129	157
14	–0.352	0.310	146.701	150	13	137	163
15	–0.258	0.303	150.689	151	13	138	164
16	–0.168	0.297	154.502	155	13	142	168
17	–0.082	0.291	158.167	158	12	146	170
18	0.002	0.286	161.704	162	12	150	174
19	0.082	0.282	165.131	165	12	153	177
20	0.161	0.278	168.464	168	12	156	180
21	0.238	0.275	171.717	172	12	160	184
22	0.313	0.273	174.906	175	12	163	187
23	0.386	0.270	178.036	178	11	167	189
24	0.459	0.269	181.118	181	11	170	192
25	0.531	0.267	184.163	184	11	173	195
26	0.602	0.266	187.187	187	11	176	198
27	0.673	0.266	190.189	190	11	179	201
28	0.743	0.265	193.182	193	11	182	204
29	0.813	0.266	196.172	196	11	185	207
30	0.884	0.266	199.170	200	11	189	211
31	0.955	0.267	202.181	202	11	191	213
32	1.026	0.268	205.208	205	11	194	216
33	1.098	0.269	208.266	208	11	197	219
34	1.171	0.271	211.353	211	11	200	222
35	1.245	0.272	214.482	214	12	202	226
36	1.319	0.275	217.654	218	12	206	230

NJSLA–S Grade 5							
Raw Score	**Theta**	**Standard Error**	**Unrounded Scale Score**	**Scale Score (SS)**	**SS CSEM**	**SS Lower Bound**	**SS Upper Bound**
37	1.395	0.277	220.882	221	12	209	233
38	1.473	0.280	224.173	224	12	212	236
39	1.552	0.283	227.532	228	12	216	240
40	1.633	0.286	230.967	231	12	219	243
41	1.716	0.290	234.496	234	12	222	246
42	1.801	0.295	238.126	238	13	225	251
43	1.890	0.300	241.872	243	13	230	256
44	1.981	0.305	245.749	246	13	233	259
45	2.076	0.311	249.778	250	13	237	263
46	2.175	0.318	253.982	254	14	240	268
47	2.279	0.326	258.386	258	14	244	272
48	2.388	0.336	263.031	263	14	249	277
49	2.504	0.346	267.961	268	15	253	283
50	2.628	0.359	273.231	273	15	258	288
51	2.762	0.373	278.913	279	16	263	295
52	2.908	0.391	285.113	285	17	268	300
53	3.069	0.413	291.971	292	18	274	300
54	3.251	0.441	299.695	300	19	281	300
55	3.461	0.477	308.604	300	19	281	300
56	3.712	0.527	319.237	300	19	281	300
57	4.026	0.600	332.600	300	19	281	300
58	4.459	0.726	350.961	300	19	281	300
59	5.178	1.013	381.527	300	19	281	300
60	6.401	1.833	433.443	300	19	281	300

Note. Grade 5 theta-to-scale linear conversion: slope = 42.46393; intercept = 161.6317

Table I.2: Grade 8 Operational Raw-to-Scale Score Conversion

NJSLA–S Grade 8							
Raw Score	Theta	Standard Error	Unrounded Scale Score	Scale Score (SS)	SS CSME	SS Lower Bound	SS Upper Bound
0	–5.289	1.831	–15.523	100	18	100	118
1	–4.072	1.009	30.463	100	18	100	118
2	–3.360	0.720	57.363	100	18	100	118
3	–2.936	0.593	73.389	100	18	100	118
4	–2.629	0.518	84.965	100	18	100	118
5	–2.387	0.468	94.100	100	18	100	118
6	–2.186	0.431	101.698	102	16	100	118
7	–2.013	0.402	108.230	108	15	100	123
8	–1.861	0.379	113.984	114	14	100	128
9	–1.724	0.361	119.148	119	14	105	133
10	–1.600	0.345	123.848	124	13	111	137
11	–1.486	0.332	128.170	128	13	115	141
12	–1.379	0.321	132.186	132	12	120	144
13	–1.280	0.311	135.949	136	12	124	148
14	–1.186	0.302	139.493	139	11	128	150
15	–1.097	0.295	142.851	143	11	132	154
16	–1.012	0.288	146.051	146	11	135	157
17	–0.931	0.282	149.115	150	11	139	161
18	–0.853	0.277	152.062	152	10	142	162
19	–0.778	0.272	154.903	155	10	145	165
20	–0.705	0.268	157.650	158	10	148	168
21	–0.635	0.264	160.321	160	10	150	170
22	–0.566	0.261	162.920	163	10	153	173
23	–0.499	0.258	165.455	165	10	155	175
24	–0.433	0.255	167.937	168	10	158	178
25	–0.369	0.253	170.370	170	10	160	180
26	–0.305	0.251	172.766	173	9	164	182
27	–0.243	0.249	175.123	175	9	166	184
28	–0.181	0.248	177.450	177	9	168	186
29	–0.120	0.246	179.751	180	9	171	189
30	–0.060	0.245	182.033	182	9	173	191
31	0.000	0.244	184.296	184	9	175	193
32	0.060	0.244	186.544	187	9	178	196
33	0.119	0.243	188.784	189	9	180	198
34	0.178	0.243	191.021	191	9	182	200
35	0.237	0.243	193.250	193	9	184	202
36	0.296	0.243	195.486	195	9	186	204
37	0.355	0.244	197.723	198	9	189	207
38	0.415	0.244	199.967	200	9	191	209

NJSLA–S Grade 8							
Raw Score	**Theta**	**Standard Error**	**Unrounded Scale Score**	**Scale Score (SS)**	**SS CSME**	**SS Lower Bound**	**SS Upper Bound**
39	0.475	0.245	202.223	202	9	193	211
40	0.535	0.246	204.493	204	9	195	213
41	0.595	0.247	206.779	207	9	198	216
42	0.656	0.248	209.091	209	9	200	218
43	0.718	0.249	211.422	211	9	202	220
44	0.781	0.251	213.787	214	9	205	223
45	0.844	0.253	216.182	216	10	206	226
46	0.908	0.255	218.612	219	10	209	229
47	0.974	0.257	221.086	221	10	211	231
48	1.041	0.260	223.610	224	10	214	234
49	1.109	0.263	226.183	226	10	216	236
50	1.178	0.266	228.816	231	10	221	241
51	1.250	0.269	231.514	232	10	222	242
52	1.323	0.273	234.287	234	10	224	244
53	1.399	0.277	237.143	237	10	227	247
54	1.477	0.282	240.093	240	11	229	251
55	1.558	0.287	243.150	243	11	232	254
56	1.642	0.293	246.327	246	11	235	257
57	1.730	0.300	249.644	250	11	239	261
58	1.822	0.307	253.116	253	12	241	265
59	1.919	0.315	256.777	257	12	245	269
60	2.021	0.325	260.650	261	12	249	273
61	2.130	0.336	264.775	265	13	252	278
62	2.248	0.349	269.207	269	13	256	282
63	2.375	0.364	274.009	274	14	260	288
64	2.514	0.383	279.271	279	14	265	293
65	2.669	0.405	285.120	285	15	270	300
66	2.844	0.433	291.742	292	16	276	300
67	3.047	0.470	299.419	300	18	282	300
68	3.291	0.520	308.630	300	18	282	300
69	3.599	0.595	320.270	300	18	282	300
70	4.025	0.721	336.357	300	18	282	300
71	4.738	1.010	363.302	300	18	282	300
72	5.956	1.831	409.314	300	18	282	300

Note. Grade 8 theta-to-scale linear conversion: slope = 37.78004; intercept = 184.2960

Table I.3: Grade 11 Operational Raw-to-Scale Score Conversion

NJSLA–S Grade 11							
Raw Score	Theta	Standard Error	Unrounded Scale Score	Scale Score (SS)	SS CSEM	SS Lower Bound	SS Upper Bound
0	–5.171	1.830	–98.234	100	18	100	118
1	–3.956	1.008	–34.048	100	18	100	118
2	–3.247	0.718	3.401	100	18	100	118
3	–2.826	0.591	25.648	100	18	100	118
4	–2.522	0.516	41.678	100	18	100	118
5	–2.283	0.465	54.302	100	18	100	118
6	–2.085	0.427	64.766	100	18	100	118
7	–1.915	0.399	73.750	100	18	100	118
8	–1.765	0.376	81.657	100	18	100	118
9	–1.631	0.357	88.735	100	18	100	118
10	–1.510	0.341	95.168	100	18	100	118
11	–1.398	0.328	101.079	101	17	100	118
12	–1.294	0.317	106.572	107	17	100	124
13	–1.197	0.307	111.706	112	16	100	128
14	–1.105	0.298	116.544	117	16	101	133
15	–1.018	0.291	121.129	121	15	106	136
16	–0.935	0.284	125.497	125	15	110	140
17	–0.856	0.278	129.675	130	15	115	145
18	–0.780	0.273	133.694	134	14	120	148
19	–0.707	0.269	137.566	138	14	124	152
20	–0.636	0.264	141.316	141	14	127	155
21	–0.567	0.261	144.961	145	14	131	159
22	–0.500	0.258	148.505	149	14	135	163
23	–0.434	0.255	151.964	152	13	139	165
24	–0.370	0.252	155.350	158	13	145	171
25	–0.307	0.250	158.672	159	13	146	172
26	–0.246	0.248	161.931	162	13	149	175
27	–0.185	0.246	165.143	165	13	152	178
28	–0.125	0.244	168.307	168	13	155	181
29	–0.066	0.243	171.433	171	13	158	184
30	–0.007	0.241	174.523	175	13	162	188
31	0.051	0.240	177.587	178	13	165	191
32	0.108	0.239	180.629	181	13	168	194
33	0.166	0.239	183.645	184	13	171	197
34	0.222	0.238	186.645	187	13	174	200
35	0.279	0.238	189.635	190	13	177	203
36	0.335	0.237	192.614	193	13	180	206
37	0.392	0.237	195.588	196	13	183	209
38	0.448	0.237	198.556	200	13	187	213

NJSLA–S Grade 11							
Raw Score	**Theta**	**Standard Error**	**Unrounded Scale Score**	**Scale Score (SS)**	**SS CSEM**	**SS Lower Bound**	**SS Upper Bound**
39	0.504	0.237	201.524	202	13	189	215
40	0.560	0.237	204.498	204	13	191	217
41	0.617	0.238	207.477	207	13	194	220
42	0.673	0.238	210.472	210	13	197	223
43	0.730	0.239	213.472	213	13	200	226
44	0.787	0.240	216.493	216	13	203	229
45	0.845	0.240	219.536	220	13	207	233
46	0.903	0.242	222.599	223	13	210	236
47	0.962	0.243	225.694	226	13	213	239
48	1.021	0.244	228.821	229	13	216	242
49	1.081	0.246	231.985	232	13	219	245
50	1.141	0.247	235.191	235	13	222	248
51	1.203	0.249	238.445	238	13	225	251
52	1.266	0.251	241.757	242	13	229	255
53	1.329	0.254	245.121	245	13	232	258
54	1.394	0.256	248.554	250	14	236	264
55	1.461	0.259	252.067	252	14	238	266
56	1.529	0.262	255.659	256	14	242	270
57	1.599	0.266	259.345	259	14	245	273
58	1.671	0.270	263.138	263	14	249	277
59	1.745	0.274	267.046	267	14	253	281
60	1.821	0.279	271.087	271	15	256	286
61	1.900	0.285	275.281	275	15	260	290
62	1.983	0.291	279.644	280	15	265	295
63	2.069	0.297	284.202	284	16	268	300
64	2.160	0.305	288.987	289	16	273	300
65	2.255	0.313	294.026	294	17	277	300
66	2.357	0.323	299.377	299	17	282	300
67	2.465	0.334	305.076	300	18	282	300
68	2.581	0.348	311.214	300	18	282	300
69	2.707	0.363	317.869	300	18	282	300
70	2.845	0.381	325.174	300	18	282	300
71	2.999	0.404	333.302	300	18	282	300
72	3.173	0.432	342.509	300	18	282	300
73	3.376	0.469	353.205	300	18	282	300
74	3.619	0.520	366.045	300	18	282	300
75	3.926	0.594	382.287	300	18	282	300
76	4.352	0.721	404.751	300	18	282	300
77	5.064	1.010	442.400	300	18	282	300
78	6.282	1.831	506.718	300	18	282	300

Note. Grade 11 theta-to-scale linear conversion: slope = 52.81995; intercept = 174.9036

APPENDIX J: Raw Score-to-Theta Subscore Tables

Table J.1: Grade 5 Earth and Space Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.605	1.841	–6.367	–0.843	Below
1	–2.362	1.027	–3.903	–0.821	Below
2	–1.613	0.746	–2.732	–0.494	Below
3	–1.150	0.627	–2.090	–0.209	Below
4	–0.799	0.563	–1.643	0.045	Below
5	–0.505	0.524	–1.291	0.282	Below
6	–0.243	0.501	–0.995	0.509	Below
7	0.001	0.488	–0.731	0.734	Below
8	0.236	0.482	–0.486	0.958	Near/Met
9	0.467	0.480	–0.253	1.186	Near/Met
10	0.698	0.482	–0.025	1.420	Near/Met
11	0.932	0.487	0.201	1.663	Near/Met
12	1.174	0.497	0.429	1.919	Near/Met
13	1.427	0.510	0.661	2.192	Near/Met
14	1.697	0.530	0.902	2.491	Above
15	1.991	0.558	1.155	2.827	Above
16	2.324	0.598	1.426	3.221	Above
17	2.718	0.663	1.724	3.712	Above
18	3.228	0.777	2.063	4.394	Above
19	4.024	1.050	2.450	5.599	Above
20	5.299	1.853	2.520	8.079	Above

Table J.2: Grade 5 Life Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.335	1.852	–6.114	–0.557	Below
1	–2.065	1.045	–3.633	–0.497	Below
2	–1.283	0.765	–2.430	–0.135	Below
3	–0.796	0.640	–1.757	0.164	Below
4	–0.436	0.566	–1.284	0.413	Below
5	–0.145	0.516	–0.918	0.628	Below
6	0.103	0.482	–0.620	0.826	Below
7	0.324	0.461	–0.367	1.016	Near/Met
8	0.531	0.450	–0.144	1.207	Near/Met
9	0.733	0.449	0.059	1.406	Near/Met
10	0.937	0.455	0.254	1.619	Near/Met
11	1.149	0.468	0.447	1.851	Near/Met
12	1.376	0.486	0.647	2.106	Near/Met
13	1.624	0.511	0.858	2.390	Above
14	1.901	0.542	1.087	2.714	Above
15	2.218	0.586	1.339	3.096	Above
16	2.598	0.652	1.620	3.575	Above
17	3.093	0.767	1.942	4.244	Above
18	3.874	1.042	2.311	5.437	Above
19	5.137	1.849	2.364	7.910	Above

Table J.3: Grade 5 Physical Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.216	1.841	–5.978	–0.455	Below
1	–1.972	1.029	–3.515	–0.429	Below
2	–1.220	0.749	–2.343	–0.097	Below
3	–0.753	0.630	–1.697	0.192	Below
4	–0.400	0.562	–1.244	0.443	Below
5	–0.109	0.519	–0.888	0.669	Below
6	0.144	0.490	–0.591	0.879	Below
7	0.374	0.470	–0.331	1.079	Near/Met
8	0.588	0.457	–0.096	1.273	Near/Met
9	0.793	0.448	0.120	1.465	Near/Met
10	0.991	0.444	0.325	1.657	Near/Met
11	1.188	0.444	0.522	1.855	Near/Met
12	1.387	0.448	0.715	2.060	Near/Met
13	1.592	0.457	0.906	2.277	Above
14	1.806	0.471	1.101	2.512	Above
15	2.037	0.491	1.301	2.773	Above
16	2.291	0.520	1.511	3.071	Above
17	2.583	0.563	1.738	3.428	Above
18	2.937	0.631	1.991	3.883	Above
19	3.405	0.750	2.281	4.530	Above
20	4.160	1.030	2.615	5.704	Above
21	5.405	1.842	2.642	8.167	Above

Table J.4: Grade 5 Critiquing Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.268	1.831	–6.013	–0.522	Below
1	–2.050	1.010	–3.566	–0.535	Below
2	–1.333	0.727	–2.423	–0.242	Below
3	–0.895	0.608	–1.807	0.017	Below
4	–0.567	0.542	–1.380	0.246	Below
5	–0.297	0.500	–1.047	0.453	Below
6	–0.061	0.472	–0.769	0.647	Below
7	0.152	0.453	–0.528	0.832	Below
8	0.352	0.442	–0.311	1.015	Near/Met
9	0.545	0.438	–0.112	1.202	Near/Met
10	0.738	0.441	0.076	1.400	Near/Met
11	0.937	0.451	0.260	1.613	Near/Met
12	1.147	0.467	0.447	1.847	Near/Met
13	1.375	0.488	0.642	2.107	Near/Met
14	1.626	0.516	0.853	2.399	Above
15	1.909	0.550	1.085	2.733	Above
16	2.235	0.595	1.343	3.127	Above
17	2.627	0.661	1.635	3.619	Above
18	3.136	0.777	1.971	4.302	Above
19	3.933	1.050	2.358	5.509	Above
20	5.209	1.854	2.429	7.990	Above

Table J.5: Grade 5 Investigating Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.237	1.844	–6.003	–0.471	Below
1	–1.985	1.034	–3.536	–0.434	Below
2	–1.221	0.757	–2.356	–0.086	Below
3	–0.741	0.640	–1.702	0.220	Below
4	–0.374	0.576	–1.238	0.489	Below
5	–0.067	0.535	–0.870	0.736	Below
6	0.205	0.509	–0.559	0.968	Near/Met
7	0.455	0.492	–0.284	1.193	Near/Met
8	0.692	0.482	–0.032	1.415	Near/Met
9	0.921	0.478	0.205	1.638	Near/Met
10	1.149	0.478	0.432	1.867	Near/Met
11	1.380	0.484	0.655	2.106	Near/Met
12	1.619	0.495	0.877	2.361	Near/Met
13	1.872	0.512	1.104	2.641	Above
14	2.148	0.539	1.339	2.957	Above
15	2.460	0.580	1.589	3.330	Above
16	2.832	0.645	1.864	3.799	Above
17	3.318	0.761	2.177	4.460	Above
18	4.089	1.037	2.533	5.645	Above
19	5.346	1.846	2.577	8.115	Above

Table J.6: Grade 5 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.645	1.851	–6.422	–0.868	Below
1	–2.376	1.046	–3.944	–0.807	Below
2	–1.589	0.770	–2.744	–0.434	Below
3	–1.091	0.653	–2.070	–0.112	Below
4	–0.710	0.585	–1.588	0.168	Below
5	–0.394	0.542	–1.206	0.418	Below
6	–0.118	0.511	–0.885	0.648	Below
7	0.132	0.489	–0.602	0.865	Below
8	0.363	0.474	–0.348	1.074	Near/Met
9	0.583	0.464	–0.113	1.279	Near/Met
10	0.795	0.458	0.108	1.483	Near/Met
11	1.004	0.457	0.319	1.689	Near/Met
12	1.214	0.460	0.524	1.903	Near/Met
13	1.428	0.467	0.728	2.128	Near/Met
14	1.651	0.479	0.932	2.371	Above
15	1.890	0.499	1.142	2.638	Above
16	2.152	0.527	1.361	2.943	Above
17	2.451	0.570	1.597	3.306	Above
18	2.812	0.636	1.858	3.766	Above
19	3.288	0.754	2.156	4.419	Above
20	4.049	1.033	2.499	5.598	Above
21	5.299	1.844	2.533	8.065	Above

Table J.7: Grade 8 Earth and Space Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.076	1.838	–6.832	–1.319	Below
1	–2.840	1.022	–4.374	–1.307	Below
2	–2.102	0.739	–3.211	–0.992	Below
3	–1.648	0.619	–2.576	–0.720	Below
4	–1.309	0.551	–2.135	–0.483	Below
5	–1.031	0.507	–1.792	–0.270	Below
6	–0.789	0.478	–1.506	–0.072	Below
7	–0.570	0.458	–1.258	0.117	Below
8	–0.367	0.445	–1.034	0.300	Below
9	–0.174	0.435	–0.827	0.479	Near/Met
10	0.013	0.429	–0.631	0.657	Near/Met
11	0.196	0.426	–0.444	0.835	Near/Met
12	0.377	0.426	–0.262	1.016	Near/Met
13	0.559	0.428	–0.083	1.201	Near/Met
14	0.744	0.432	0.095	1.393	Near/Met
15	0.934	0.440	0.274	1.594	Near/Met
16	1.132	0.451	0.456	1.808	Above
17	1.342	0.466	0.642	2.041	Above
18	1.568	0.487	0.837	2.299	Above
19	1.820	0.517	1.044	2.596	Above
20	2.109	0.561	1.268	2.951	Above
21	2.461	0.629	1.517	3.404	Above
22	2.927	0.748	1.805	4.049	Above
23	3.679	1.028	2.136	5.221	Above
24	4.922	1.841	2.161	7.684	Above

Table J.8: Grade 8 Life Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.085	1.838	–6.842	–1.328	Below
1	–2.849	1.023	–4.383	–1.315	Below
2	–2.109	0.740	–3.219	–1.000	Below
3	–1.656	0.618	–2.583	–0.728	Below
4	–1.318	0.549	–2.141	–0.495	Below
5	–1.042	0.504	–1.798	–0.287	Below
6	–0.805	0.473	–1.514	–0.095	Below
7	–0.592	0.451	–1.269	0.085	Below
8	–0.395	0.436	–1.050	0.259	Below
9	–0.209	0.426	–0.849	0.430	Near/Met
10	–0.030	0.421	–0.661	0.600	Near/Met
11	0.145	0.418	–0.482	0.772	Near/Met
12	0.320	0.419	–0.308	0.948	Near/Met
13	0.497	0.422	–0.136	1.129	Near/Met
14	0.677	0.428	0.035	1.319	Near/Met
15	0.863	0.437	0.208	1.518	Near/Met
16	1.059	0.449	0.386	1.733	Above
17	1.268	0.466	0.570	1.966	Above
18	1.495	0.488	0.763	2.226	Above
19	1.747	0.519	0.969	2.525	Above
20	2.038	0.563	1.194	2.882	Above
21	2.391	0.631	1.445	3.337	Above
22	2.860	0.750	1.735	3.984	Above
23	3.614	1.030	2.069	5.158	Above
24	4.859	1.842	2.097	7.622	Above

Table J.9: Grade 8 Physical Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.353	1.844	–7.120	–1.587	Below
1	–3.101	1.034	–4.652	–1.551	Below
2	–2.340	0.754	–3.471	–1.209	Below
3	–1.866	0.634	–2.817	–0.915	Below
4	–1.510	0.564	–2.356	–0.663	Below
5	–1.218	0.519	–1.996	–0.440	Below
6	–0.966	0.486	–1.695	–0.237	Below
7	–0.742	0.463	–1.435	–0.048	Below
8	–0.536	0.445	–1.204	0.132	Below
9	–0.344	0.433	–0.993	0.306	Below
10	–0.160	0.425	–0.797	0.477	Near/Met
11	0.018	0.420	–0.612	0.648	Near/Met
12	0.194	0.419	–0.435	0.822	Near/Met
13	0.370	0.421	–0.262	1.001	Near/Met
14	0.549	0.426	–0.090	1.189	Near/Met
15	0.734	0.435	0.082	1.387	Near/Met
16	0.929	0.448	0.257	1.600	Near/Met
17	1.136	0.465	0.439	1.833	Above
18	1.363	0.488	0.631	2.094	Above
19	1.616	0.520	0.836	2.395	Above
20	1.908	0.565	1.061	2.755	Above
21	2.264	0.633	1.314	3.214	Above
22	2.737	0.753	1.607	3.866	Above
23	3.497	1.033	1.948	5.046	Above
24	4.747	1.844	1.981	7.512	Above

Table J.10: Grade 8 Critiquing Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.238	1.837	–6.993	–1.483	Below
1	–3.006	1.020	–4.536	–1.475	Below
2	–2.272	0.735	–3.375	–1.169	Below
3	–1.825	0.613	–2.744	–0.906	Below
4	–1.494	0.542	–2.308	–0.681	Below
5	–1.226	0.497	–1.971	–0.481	Below
6	–0.995	0.465	–1.693	–0.297	Below
7	–0.790	0.443	–1.455	–0.124	Below
8	–0.600	0.428	–1.242	0.042	Below
9	–0.422	0.418	–1.048	0.205	Below
10	–0.250	0.411	–0.867	0.366	Below
11	–0.083	0.408	–0.695	0.529	Near/Met
12	0.084	0.408	–0.528	0.695	Near/Met
13	0.251	0.410	–0.365	0.866	Near/Met
14	0.421	0.415	–0.202	1.043	Near/Met
15	0.596	0.423	–0.038	1.230	Near/Met
16	0.779	0.433	0.130	1.428	Near/Met
17	0.972	0.446	0.302	1.641	Near/Met
18	1.179	0.464	0.483	1.874	Above
19	1.404	0.487	0.674	2.135	Above
20	1.656	0.518	0.878	2.434	Above
21	1.947	0.563	1.102	2.791	Above
22	2.301	0.631	1.354	3.248	Above
23	2.770	0.751	1.644	3.896	Above
24	3.526	1.030	1.980	5.071	Above
25	4.773	1.842	2.009	7.536	Above

Table J.11: Grade 8 Investigating Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.152	1.844	–6.918	–1.386	Below
1	–2.901	1.033	–4.451	–1.351	Below
2	–2.140	0.755	–3.271	–1.008	Below
3	–1.664	0.636	–2.618	–0.709	Below
4	–1.303	0.570	–2.157	–0.449	Below
5	–1.004	0.527	–1.794	–0.214	Below
6	–0.743	0.498	–1.489	0.004	Below
7	–0.506	0.477	–1.222	0.210	Below
8	–0.285	0.463	–0.980	0.410	Below
9	–0.075	0.454	–0.756	0.606	Near/Met
10	0.129	0.449	–0.545	0.802	Near/Met
11	0.329	0.447	–0.342	1.000	Near/Met
12	0.530	0.449	–0.144	1.203	Near/Met
13	0.733	0.454	0.052	1.414	Near/Met
14	0.943	0.463	0.248	1.638	Near/Met
15	1.164	0.477	0.448	1.880	Above
16	1.401	0.498	0.655	2.147	Above
17	1.662	0.527	0.873	2.452	Above
18	1.961	0.569	1.107	2.815	Above
19	2.321	0.636	1.367	3.275	Above
20	2.796	0.754	1.665	3.927	Above
21	3.557	1.032	2.008	5.105	Above
22	4.806	1.843	2.041	7.571	Above

Table J.12: Grade 8 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.144	1.844	–6.909	–1.379	Below
1	–2.895	1.031	–4.442	–1.348	Below
2	–2.139	0.750	–3.264	–1.014	Below
3	–1.672	0.628	–2.614	–0.730	Below
4	–1.323	0.557	–2.159	–0.488	Below
5	–1.040	0.510	–1.805	–0.275	Below
6	–0.797	0.477	–1.513	–0.081	Below
7	–0.581	0.453	–1.261	0.099	Below
8	–0.384	0.436	–1.038	0.270	Below
9	–0.199	0.424	–0.835	0.436	Below
10	–0.024	0.415	–0.647	0.599	Near/Met
11	0.146	0.410	–0.469	0.762	Near/Met
12	0.314	0.408	–0.299	0.926	Near/Met
13	0.481	0.409	–0.133	1.095	Near/Met
14	0.650	0.413	0.030	1.269	Near/Met
15	0.822	0.419	0.194	1.451	Near/Met
16	1.002	0.428	0.359	1.644	Near/Met
17	1.191	0.441	0.529	1.852	Above
18	1.392	0.458	0.705	2.080	Above
19	1.613	0.481	0.891	2.334	Above
20	1.859	0.513	1.090	2.628	Above
21	2.144	0.558	1.307	2.980	Above
22	2.492	0.627	1.552	3.432	Above
23	2.956	0.747	1.836	4.076	Above
24	3.706	1.028	2.164	5.248	Above
25	4.949	1.841	2.188	7.710	Above

Table J.13: Grade 11 Earth and Space Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.196	1.835	–6.949	–1.444	Below
1	–2.968	1.017	–4.494	–1.443	Below
2	–2.240	0.732	–3.338	–1.143	Below
3	–1.798	0.609	–2.712	–0.884	Below
4	–1.471	0.540	–2.280	–0.661	Below
5	–1.205	0.495	–1.947	–0.462	Below
6	–0.975	0.465	–1.672	–0.277	Below
7	–0.769	0.444	–1.435	–0.103	Below
8	–0.578	0.429	–1.222	0.065	Below
9	–0.399	0.419	–1.027	0.229	Below
10	–0.227	0.412	–0.844	0.391	Below
11	–0.059	0.407	–0.670	0.551	Near/Met
12	0.105	0.404	–0.502	0.711	Near/Met
13	0.268	0.404	–0.337	0.873	Near/Met
14	0.431	0.404	–0.176	1.038	Near/Met
15	0.596	0.407	–0.015	1.206	Near/Met
16	0.763	0.412	0.145	1.381	Near/Met
17	0.935	0.419	0.307	1.563	Near/Met
18	1.114	0.428	0.472	1.756	Above
19	1.303	0.441	0.642	1.964	Above
20	1.504	0.458	0.817	2.191	Above
21	1.724	0.481	1.003	2.444	Above
22	1.969	0.512	1.201	2.737	Above
23	2.253	0.557	1.418	3.088	Above
24	2.600	0.626	1.661	3.539	Above
25	3.063	0.746	1.944	4.182	Above
26	3.812	1.027	2.271	5.352	Above
27	5.054	1.840	2.293	7.814	Above

Table J.14: Grade 11 Life Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.105	1.840	–6.865	–1.345	Below
1	–2.865	1.025	–4.402	–1.327	Below
2	–2.121	0.741	–3.233	–1.009	Below
3	–1.667	0.618	–2.594	–0.740	Below
4	–1.331	0.546	–2.150	–0.512	Below
5	–1.060	0.499	–1.808	–0.311	Below
6	–0.828	0.466	–1.527	–0.128	Below
7	–0.622	0.443	–1.286	0.043	Below
8	–0.433	0.427	–1.073	0.208	Below
9	–0.255	0.416	–0.879	0.368	Below
10	–0.086	0.409	–0.699	0.527	Near/Met
11	0.079	0.405	–0.527	0.686	Near/Met
12	0.242	0.403	–0.362	0.846	Near/Met
13	0.405	0.403	–0.200	1.009	Near/Met
14	0.568	0.406	–0.040	1.176	Near/Met
15	0.734	0.410	0.120	1.348	Near/Met
16	0.904	0.415	0.281	1.527	Near/Met
17	1.080	0.423	0.445	1.715	Near/Met
18	1.263	0.434	0.613	1.914	Above
19	1.457	0.447	0.786	2.127	Above
20	1.664	0.464	0.967	2.361	Above
21	1.890	0.488	1.159	2.622	Above
22	2.143	0.519	1.364	2.922	Above
23	2.435	0.564	1.589	3.281	Above
24	2.790	0.633	1.841	3.739	Above
25	3.262	0.752	2.133	4.390	Above
26	4.020	1.032	2.472	5.568	Above
27	5.270	1.843	2.505	8.034	Above

Table J.15: Grade 11 Physical Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	−3.858	1.843	−6.622	−1.093	Below
1	−2.609	1.031	−4.156	−1.063	Below
2	−1.852	0.751	−2.979	−0.725	Below
3	−1.382	0.632	−2.329	−0.434	Below
4	−1.027	0.563	−1.872	−0.182	Below
5	−0.736	0.519	−1.514	0.042	Below
6	−0.484	0.488	−1.215	0.248	Below
7	−0.257	0.465	−0.955	0.440	Below
8	−0.049	0.449	−0.722	0.624	Near/Met
9	0.146	0.437	−0.508	0.801	Near/Met
10	0.333	0.428	−0.309	0.975	Near/Met
11	0.514	0.423	−0.121	1.149	Near/Met
12	0.692	0.421	0.061	1.323	Near/Met
13	0.869	0.422	0.237	1.502	Near/Met
14	1.048	0.426	0.410	1.687	Near/Met
15	1.232	0.433	0.583	1.881	Above
16	1.424	0.444	0.759	2.089	Above
17	1.627	0.459	0.938	2.316	Above
18	1.848	0.481	1.126	2.570	Above
19	2.094	0.512	1.326	2.862	Above
20	2.379	0.557	1.543	3.214	Above
21	2.726	0.626	1.787	3.665	Above
22	3.189	0.746	2.069	4.309	Above
23	3.938	1.028	2.397	5.480	Above
24	5.181	1.841	2.420	7.943	Above

Table J.16: Grade 11 Critiquing Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.964	1.836	–6.718	–1.210	Below
1	–2.734	1.018	–4.261	–1.207	Below
2	–2.004	0.734	–3.105	–0.903	Below
3	–1.558	0.613	–2.478	–0.638	Below
4	–1.224	0.546	–2.044	–0.405	Below
5	–0.950	0.504	–1.707	–0.193	Below
6	–0.710	0.477	–1.425	0.005	Below
7	–0.492	0.458	–1.179	0.195	Below
8	–0.289	0.445	–0.956	0.378	Below
9	–0.096	0.435	–0.748	0.557	Near/Met
10	0.090	0.428	–0.552	0.733	Near/Met
11	0.272	0.424	–0.364	0.907	Near/Met
12	0.450	0.421	–0.182	1.081	Near/Met
13	0.627	0.421	–0.004	1.258	Near/Met
14	0.804	0.423	0.171	1.438	Near/Met
15	0.985	0.427	0.344	1.625	Near/Met
16	1.170	0.434	0.518	1.821	Above
17	1.363	0.445	0.695	2.030	Above
18	1.567	0.461	0.876	2.258	Above
19	1.789	0.483	1.065	2.513	Above
20	2.036	0.513	1.266	2.806	Above
21	2.321	0.558	1.485	3.158	Above
22	2.669	0.626	1.730	3.608	Above
23	3.132	0.746	2.013	4.252	Above
24	3.882	1.027	2.341	5.423	Above
25	5.124	1.841	2.363	7.885	Above

Table J.17: Grade 11 Investigating Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.099	1.838	–6.856	–1.341	Below
1	–2.863	1.022	–4.397	–1.330	Below
2	–2.125	0.738	–3.232	–1.019	Below
3	–1.676	0.614	–2.597	–0.754	Below
4	–1.344	0.543	–2.158	–0.529	Below
5	–1.075	0.496	–1.819	–0.331	Below
6	–0.846	0.464	–1.541	–0.151	Below
7	–0.642	0.441	–1.303	0.019	Below
8	–0.455	0.425	–1.092	0.182	Below
9	–0.279	0.414	–0.900	0.342	Below
10	–0.111	0.407	–0.722	0.500	Near/Met
11	0.054	0.404	–0.552	0.659	Near/Met
12	0.216	0.403	–0.389	0.821	Near/Met
13	0.379	0.405	–0.228	0.986	Near/Met
14	0.544	0.409	–0.068	1.157	Near/Met
15	0.714	0.414	0.092	1.335	Near/Met
16	0.889	0.423	0.255	1.522	Near/Met
17	1.071	0.433	0.421	1.721	Near/Met
18	1.265	0.447	0.594	1.936	Above
19	1.473	0.465	0.775	2.171	Above
20	1.700	0.489	0.966	2.434	Above
21	1.955	0.521	1.173	2.736	Above
22	2.249	0.567	1.399	3.099	Above
23	2.608	0.636	1.654	3.562	Above
24	3.085	0.756	1.951	4.218	Above
25	3.849	1.035	2.296	5.402	Above
26	5.103	1.845	2.335	7.871	Above

Table J.18: Grade 11 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.120	1.841	–6.882	–1.358	Below
1	–2.876	1.028	–4.418	–1.333	Below
2	–2.124	0.748	–3.246	–1.003	Below
3	–1.659	0.627	–2.600	–0.718	Below
4	–1.311	0.558	–2.148	–0.473	Below
5	–1.025	0.513	–1.795	–0.256	Below
6	–0.779	0.481	–1.501	–0.057	Below
7	–0.559	0.458	–1.246	0.129	Below
8	–0.357	0.441	–1.018	0.305	Below
9	–0.168	0.428	–0.810	0.474	Below
10	0.011	0.418	–0.617	0.638	Near/Met
11	0.183	0.412	–0.435	0.800	Near/Met
12	0.350	0.407	–0.261	0.960	Near/Met
13	0.514	0.405	–0.093	1.121	Near/Met
14	0.678	0.404	0.071	1.285	Near/Met
15	0.842	0.406	0.233	1.451	Near/Met
16	1.009	0.410	0.393	1.624	Near/Met
17	1.179	0.417	0.554	1.804	Above
18	1.357	0.426	0.718	1.995	Above
19	1.543	0.438	0.885	2.201	Above
20	1.742	0.455	1.059	2.426	Above
21	1.960	0.478	1.242	2.677	Above
22	2.203	0.510	1.438	2.968	Above
23	2.485	0.555	1.653	3.318	Above
24	2.830	0.624	1.894	3.766	Above
25	3.291	0.745	2.174	4.408	Above
26	4.038	1.026	2.499	5.577	Above
27	5.279	1.840	2.519	8.039	Above

APPENDIX K: Subscore Proficiency Classifications

Table K.1: Grade 5 Content Disaggregated Subscore Proficiency Classifications

		Earth and Space Science			Life Science			Physical Science		
Group	N	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above
All Students	96,215	47.04	39.88	13.08	49.58	33.82	16.60	53.47	32.17	14.37
Male	48,685	46.21	39.07	14.73	49.10	32.98	17.92	52.01	31.97	16.02
Female	47,516	47.89	40.71	11.40	50.08	34.67	15.25	54.96	32.37	12.67
Am. Indian	192	36.46	43.75	19.79	49.48	31.77	18.75	50.00	31.25	18.75
Asian	10,550	19.76	47.27	32.97	21.83	39.40	38.77	25.05	39.73	35.22
Black	13,303	66.80	29.35	3.84	68.90	24.49	6.61	72.28	22.58	5.13
Hispanic	32,784	63.14	31.47	5.39	64.98	27.37	7.64	69.51	24.56	5.93
Pacific Islander	135	41.48	43.70	14.81	48.15	30.37	21.48	54.81	31.11	14.07
White	35,920	33.93	48.92	17.15	37.44	41.31	21.25	41.06	40.26	18.68
ML–Yes	10,424	86.37	12.91	0.72	87.61	11.25	1.14	88.93	10.24	0.83
ML–No	85,788	42.26	43.16	14.58	44.96	36.56	18.48	49.16	34.83	16.01
EconDis–Yes	41,269	65.64	30.10	4.26	67.58	26.06	6.36	72.05	23.11	4.84
EconDis–No	54,944	33.06	47.23	19.71	36.07	39.64	24.29	39.51	38.97	21.52
SWD–Yes	21,597	66.72	27.11	6.17	69.82	21.79	8.39	71.72	20.71	7.58
SWD–No	74,611	41.34	43.58	15.08	43.72	37.30	18.98	48.18	35.48	16.34
CBT	68,817	39.70	44.65	15.64	42.48	37.93	19.58	46.78	36.19	17.02
PBT	72	63.89	33.33	2.78	63.89	29.17	6.94	70.83	23.61	5.56
TTS	24,261	62.66	29.89	7.45	64.96	24.95	10.09	68.01	23.40	8.59
SP	1,660	87.77	11.99	0.24	87.47	11.39	1.14	88.61	11.02	0.36
SP TTS	1,175	90.47	9.36	0.17	89.11	9.87	1.02	89.45	9.96	0.60
Human Reader	177	72.32	24.29	3.39	71.75	21.47	6.78	74.01	20.90	5.08

Table K.2: Grade 5 Practice Disaggregated Subscore Proficiency Classifications

		Critiquing			Investigating			Sensemaking		
Group	**N**	**% Below**	**% Near/Met**	**% Above**	**% Below**	**% Near/Met**	**% Above**	**% Below**	**% Near/Met**	**% Above**
All Students	96,215	49.73	37.47	12.81	48.64	38.07	13.29	51.57	34.80	13.64
Male	48,685	48.94	37.32	13.74	47.87	37.28	14.85	50.25	34.16	15.59
Female	47,516	50.53	37.62	11.84	49.43	38.87	11.70	52.92	35.45	11.63
Am. Indian	192	44.79	38.54	16.67	43.75	37.50	18.75	43.75	36.98	19.27
Asian	10,550	20.99	47.14	31.88	21.35	45.52	33.14	24.27	42.27	33.46
Black	13,303	70.04	25.59	4.37	67.51	27.99	4.50	70.17	25.07	4.76
Hispanic	32,784	65.53	29.07	5.40	64.92	29.60	5.48	67.15	27.15	5.70
Pacific Islander	135	49.63	37.78	12.59	48.15	37.78	14.07	48.15	38.52	13.33
White	35,920	37.16	46.29	16.55	35.73	47.03	17.23	39.39	42.90	17.71
ML–Yes	10,424	87.01	12.26	0.73	87.02	12.20	0.78	88.64	10.65	0.71
ML–No	85,788	45.20	40.53	14.28	43.97	41.22	14.81	47.06	37.73	15.21
EconDis–Yes	41,269	68.26	27.45	4.29	67.08	28.64	4.28	69.74	25.63	4.63
EconDis–No	54,944	35.81	44.99	19.21	34.79	45.16	20.05	37.91	41.68	20.40
SWD–Yes	21,597	69.31	24.45	6.24	68.95	24.60	6.45	69.72	23.05	7.23
SWD–No	74,611	44.05	41.24	14.71	42.76	41.98	15.27	46.31	38.20	15.49
CBT	68,817	42.67	42.16	15.17	41.49	42.64	15.87	44.66	39.17	16.17
PBT	72	66.67	29.17	4.17	65.28	31.94	2.78	68.06	27.78	4.17
TTS	24,261	65.22	27.12	7.65	64.06	28.34	7.60	66.24	25.63	8.12
SP	1,660	85.30	14.10	0.60	87.05	12.47	0.48	91.45	8.49	0.06
SP TTS	1,175	87.06	12.17	0.77	88.77	10.98	0.26	91.23	8.26	0.51
Human Reader	177	73.45	23.73	2.82	71.19	23.73	5.08	71.75	22.03	6.21

Table K.3: Grade 8 Content Disaggregated Subscore Proficiency Classifications

		Earth and Space Science			Life Science			Physical Science		
Group	N	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above
All Students	99,219	58.44	34.20	7.36	61.09	30.71	8.19	60.30	32.03	7.67
Male	50,914	57.66	34.42	7.92	60.65	30.06	9.30	61.26	30.36	8.38
Female	48,275	59.26	33.97	6.77	61.57	31.40	7.03	59.30	33.79	6.91
Am. Indian	148	60.14	30.41	9.46	65.54	27.03	7.43	61.49	33.78	4.73
Asian	10,781	26.48	52.09	21.43	29.70	44.89	25.41	26.21	49.03	24.76
Black	13,888	76.95	20.98	2.07	77.64	20.05	2.30	78.20	19.79	2.02
Hispanic	34,174	74.67	22.78	2.55	76.40	20.90	2.70	76.04	21.47	2.49
Pacific Islander	188	52.66	37.23	10.11	56.91	31.38	11.70	51.06	39.89	9.04
White	37,138	46.63	43.97	9.41	50.71	39.12	10.17	49.84	40.82	9.34
ML–Yes	9,176	92.36	7.35	0.29	93.33	6.40	0.27	92.43	7.31	0.26
ML–No	90,039	54.98	36.94	8.08	57.81	33.19	9.00	57.03	34.55	8.42
EconDis–Yes	41,311	76.09	21.84	2.07	77.33	20.27	2.41	77.33	20.56	2.11
EconDis–No	57,904	45.84	43.02	11.14	49.51	38.17	12.32	48.15	40.22	11.63
SWD–Yes	20,582	76.53	20.32	3.14	77.78	18.26	3.96	79.25	17.71	3.04
SWD–No	78,633	53.70	37.84	8.47	56.72	33.98	9.30	55.34	35.78	8.88
CBT	77,160	53.56	37.93	8.51	56.60	33.97	9.42	55.55	35.57	8.88
PBT	40	75.00	25.00	0.00	62.50	32.50	5.00	87.50	12.50	0.00
TTS	18,877	72.78	23.34	3.88	74.15	21.34	4.50	74.25	21.78	3.97
SP	2,063	91.61	8.29	0.10	92.73	7.03	0.24	92.73	7.13	0.15
SP TTS	999	91.99	7.81	0.20	93.29	6.71	0.00	92.49	7.41	0.10
Human Reader	55	90.91	9.09	0.00	87.27	12.73	0.00	98.18	0.00	1.82

Table K.4: Grade 8 Practice Disaggregated Subscore Proficiency Classifications

		Critiquing			Investigating			Sensemaking		
Group	N	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above
All Students	99,219	61.42	33.40	5.18	65.16	27.17	7.67	66.03	26.53	7.44
Male	50,914	62.51	32.08	5.42	64.18	27.24	8.58	64.90	26.67	8.43
Female	48,275	60.29	34.78	4.92	66.20	27.08	6.72	67.23	26.37	6.39
Am. Indian	148	65.54	29.73	4.73	68.92	26.35	4.73	68.24	27.70	4.05
Asian	10,781	28.87	54.36	16.77	32.57	43.31	24.13	32.90	43.30	23.80
Black	13,888	78.89	19.73	1.38	81.48	16.55	1.97	83.17	15.14	1.68
Hispanic	34,174	77.59	20.79	1.62	80.08	17.47	2.46	80.95	16.62	2.43
Pacific Islander	188	56.38	39.36	4.26	60.11	32.98	6.91	63.30	27.66	9.04
White	37,138	50.35	43.34	6.31	55.50	34.93	9.56	56.28	34.50	9.22
ML–Yes	9,176	94.29	5.57	0.14	94.67	5.07	0.26	94.81	4.90	0.28
ML–No	90,039	58.07	36.23	5.69	62.15	29.42	8.43	63.10	28.73	8.17
EconDis–Yes	41,311	78.76	19.85	1.39	81.10	16.78	2.11	82.24	15.80	1.96
EconDis–No	57,904	49.06	43.06	7.88	53.79	34.57	11.64	54.46	34.18	11.35
SWD–Yes	20,582	79.63	18.36	2.01	81.11	15.69	3.20	82.01	14.54	3.44
SWD–No	78,633	56.66	37.34	6.01	60.99	30.17	8.85	61.85	29.67	8.48
CBT	77,160	56.65	37.35	6.00	60.90	30.26	8.84	61.77	29.62	8.61
PBT	40	77.50	17.50	5.00	87.50	12.50	0.00	85.00	12.50	2.50
TTS	18,877	75.40	21.93	2.66	77.41	18.40	4.18	78.76	17.37	3.88
SP	2,063	94.13	5.82	0.05	96.12	3.83	0.05	93.60	6.11	0.29
SP TTS	999	95.20	4.70	0.10	95.60	4.40	0.00	94.69	5.31	0.00
Human Reader	55	94.55	5.45	0.00	94.55	5.45	0.00	96.36	3.64	0.00

Table K.5: Grade 11 Content Disaggregated Subscore Proficiency Classifications

		Earth and Space Science			Life Science			Physical Science		
Group	N	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above
All Students	97,156	50.62	32.83	16.56	48.00	36.99	15.01	49.99	35.62	14.38
Male	49,581	49.28	31.78	18.94	49.83	34.29	15.88	49.10	34.14	16.75
Female	47,441	52.07	33.89	14.04	46.14	39.79	14.07	50.96	37.15	11.88
Am. Indian	173	51.45	34.68	13.87	46.24	39.88	13.87	54.91	30.64	14.45
Asian	10,409	21.75	38.11	40.14	19.31	42.31	38.38	19.29	39.84	40.87
Black	13,787	68.49	25.34	6.17	64.18	29.65	6.17	66.85	28.13	5.03
Hispanic	32,455	65.53	26.86	7.61	62.35	30.76	6.89	65.06	29.06	5.88
Pacific Islander	195	54.36	28.21	17.44	49.74	32.82	17.44	51.28	33.85	14.87
White	37,716	39.73	39.13	21.14	38.26	43.38	18.36	39.86	42.65	17.49
ML–Yes	7,617	88.05	10.82	1.13	86.28	13.10	0.62	86.25	12.80	0.95
ML–No	89,530	47.43	34.70	17.87	44.74	39.02	16.23	46.90	37.57	15.53
EconDis–Yes	37,793	66.55	26.26	7.18	62.99	30.54	6.47	65.66	28.60	5.74
EconDis–No	59,354	40.46	37.01	22.53	38.46	41.10	20.44	40.01	40.10	19.89
SWD–Yes	19,765	66.91	23.54	9.55	65.31	26.68	8.01	67.10	25.77	7.12
SWD–No	77,381	46.45	35.20	18.35	43.58	39.63	16.79	45.62	38.14	16.24
CBT	85,163	47.79	34.34	17.86	45.01	38.74	16.26	47.18	37.26	15.55
PBT	49	61.22	28.57	10.20	63.27	30.61	6.12	59.18	34.69	6.12
TTS	9,288	66.90	24.12	8.98	65.44	26.85	7.71	65.97	26.50	7.54
SP	2,231	83.73	14.84	1.43	82.97	16.45	0.58	84.18	14.88	0.94
SP TTS	289	86.85	12.80	0.35	85.12	14.88	0.00	87.54	11.42	1.04
Human Reader	92	79.35	18.48	2.17	71.74	28.26	0.00	71.74	27.17	1.09

Table K.6: Grade 11 Practice Disaggregated Subscore Proficiency Classifications

		Critiquing			Investigating			Sensemaking		
Group	N	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above	% Below	% Near/Met	% Above
All Students	97,156	47.18	36.24	16.57	48.11	37.14	14.75	53.02	32.44	14.54
Male	49,581	47.28	34.89	17.83	48.64	34.90	16.46	52.06	30.93	17.00
Female	47,441	47.12	37.66	15.23	47.61	39.45	12.95	54.07	34.00	11.93
Am. Indian	173	45.66	37.57	16.76	50.29	35.84	13.87	53.18	36.42	10.40
Asian	10,409	18.68	40.46	40.86	17.67	42.12	40.22	23.73	38.14	38.13
Black	13,787	64.86	28.45	6.69	66.53	28.35	5.11	68.32	26.26	5.42
Hispanic	32,455	61.93	30.31	7.76	62.90	30.97	6.13	67.49	26.15	6.36
Pacific Islander	195	48.72	30.26	21.03	48.21	35.38	16.41	57.95	26.67	15.38
White	37,716	36.46	42.89	20.66	37.62	44.11	18.28	43.61	38.38	18.01
ML–Yes	7,617	86.37	12.73	0.89	85.78	13.18	1.04	88.34	10.82	0.84
ML–No	89,530	43.84	38.25	17.91	44.90	39.18	15.92	50.01	34.28	15.71
EconDis–Yes	37,793	62.86	29.80	7.34	63.95	30.27	5.77	67.94	25.83	6.23
EconDis–No	59,354	37.19	40.35	22.46	38.02	41.52	20.47	43.52	36.65	19.83
SWD–Yes	19,765	64.37	26.62	9.01	65.96	26.66	7.38	68.71	23.19	8.10
SWD–No	77,381	42.79	38.71	18.51	43.55	39.82	16.63	49.01	34.81	16.19
CBT	85,163	44.24	37.85	17.92	45.17	38.85	15.98	50.30	33.95	15.75
PBT	49	53.06	42.86	4.08	63.27	28.57	8.16	61.22	30.61	8.16
TTS	9,288	64.09	27.22	8.69	65.17	27.40	7.43	68.50	24.07	7.43
SP	2,231	82.12	16.72	1.17	82.21	16.76	1.03	85.88	13.40	0.72
SP TTS	289	87.20	11.76	1.04	86.51	13.49	0.00	87.89	11.42	0.69
Human Reader	92	70.65	26.09	3.26	73.91	26.09	0.00	80.43	19.57	0.00

APPENDIX L: Executive Summary of the NJSLA–S Alignment Evaluation Study

Appendix L contains the executive summary from the alignment evaluation study report submitted by edCount, LLC, in September 2022.

Introduction

The New Jersey Student Learning Assessment–Science (NJSLA–S) assesses students in grades 5, 8, and 11 on their understanding and explanations of scientific phenomena and scenarios. In spring 2019, the NJSLA–S was administered for the first time. Due to the coronavirus pandemic, statewide assessments were cancelled for the 2019–2020 and 2020–2021 school years; thus, the 2022 year marked the second administration of this assessment.

The NJSLA–S is composed of two parts: a performance-based assessment (PBA) and a machine-scorable assessment (MSA). Each item within the NJSLA–S represents an interaction of Disciplinary Core Ideas (DCIs), Science and Engineering Practices (SEPs), and Crosscutting Concepts (CCCs).

The New Jersey Department of Education (NJDOE) commissioned edCount, LLC, (edCount) to conduct an independent evaluation of the alignment quality of the NJSLA–S for grades 5, 8, and 11 with the New Jersey Student Learning Standards for Science (NJSLS–S) in 2022. This report documents the methodology for and results of this independent alignment evaluation. NJDOE intends to use the information gained via this evaluation to inform decisions about future item and assessment development and for federal peer review purposes.

Evaluation Methodology

Evidence of alignment quality is critical to validity evaluation for standards-based assessments (Forte, 2017; Webb, 1997, 1999). Such evidence must draw upon an examination of how a test has been designed and developed, as well as instances of the test itself (Forte, 2013). As is the case for all validity evidence, evidence of alignment quality is necessary to support the interpretation and use of test scores. A well-aligned test is one that elicits a sample of student performance that is adequate to support inferences about student achievement in relation to the standards-based domains on which the test is based.

None of the traditional alignment methods are suited to meet the challenge of evaluating the multidimensional science standards within the NJSLS–S. These methods, such as Webb (1999), involve panelists' ratings of content and cognitive complexity and the analysis of those ratings in relation to overall criteria for Domain Concurrence, Balance of Representation, Range of Knowledge, and Depth of Knowledge (DOK), which will not alone address the needs of the NJSLS–S.

To address the unique aspects of the three-dimensional nature of the NJSLS–S and the NJSLA–S, edCount addresses the following alignment questions.

1. To what extent do the blueprints support the consistent creation of test forms that reflect the standards and the score scale?
2. To what extent do the Performance-Level Descriptors (PLDs) reflect meaningful and appropriate score interpretations across the full range of the score scale?
3. To what extent does the set of phenomena, tasks, and items reflect the blueprints and provide performance opportunities across the full range of the score scale?

This approach evaluates the quality of alignment of the assessment to the multidimensional standards and provides evidence of the extent to which the assessment supports inferences about student achievement in relation to the standards-based domains.

Evaluation Findings and Recommendations

Evaluation Question 1 addresses the blueprints (and not test items). This question focuses on the extent to which the blueprints support the consistent creation of test forms that reflect the NJSLS–S and the score scale. edCount evaluators found that the blueprint development of the NJSLA–S is well-documented across all grades, including a clear description of the review and revision process by stakeholders. Each blueprint also meets the alignment criteria of strong evidence of alignment for Domain Concurrence, Balance of Representation, and Phenomena Design.

edCount commends the NJDOE for the use of the emerging best practice of PLDs as a cognitive complexity framework for forms development and for the NJDOE's plans to include Range PLD expectations within the test blueprints, as well as on the close monitoring of test content, including longitudinal representation of content and item types by form. Another commendable practice the NJDOE used is the thoroughness of the phenomena design guidance provided within the test development documentation.

To further supplement these practices, edCount recommends that the NJDOE consider including guidance on the balance of score points within the test blueprint, in addition to guidance around the number of items by domain. edCount also recommends that the NJDOE consider including guidance on the longitudinal sampling of content, such as the number of forms to be developed before all assessed DCIs have been represented on a form at least once.

Evaluation Question 2 addresses the PLDs. Evaluators found that the PLDs were developed to represent clear and appropriate expectations for performance on the assessments. Evaluators also found documentation that indicated that the item review process included item alignment to PLDs as part of the new item development process. The PLDs for all three grade levels exhibit strong evidence of alignment with the NJSLS–S in terms of PLD-Domain Concurrence. Panelists noted that every reporting category/domain from the standards is fully represented by the PLDs. The PLDs describe increasingly sophisticated and reasonable levels of performance for the concepts defined in the standards. However, the grade 11 panelists noted some uneven progressions between Levels 3 and 4 of the Earth and Space Science portion of the PLDs, due primarily to vagueness in the PLD wording. Overall, the PLDs for all three evaluated forms meet the criteria for “adequately differentiated.”

edCount commends the NJDOE on the development of extremely detailed Range PLDs and on leveraging these PLDs during item development. edCount also commends the NJDOE on the inclusion of New Jersey educators and experts in the assessment field in the PLD development process.

As a result of these findings for Evaluation Question 2, edCount recommends the NJDOE review the grade 11 Earth and Space Science PLDs to ensure that sufficient progression is clarified in the language.

Evaluation Question 3 focuses on the relationship among the dimensions of the standards, phenomena, and items that contribute to students’ scores. For the first part of this question, which examines the development process for the phenomena, tasks, clusters, and items on the test form, edCount evaluators found that the test forms were developed to ensure that they consisted of item clusters tied to phenomena and the overall test forms reflect each respective blueprint. edCount commends the NJDOE on the test form design and the inclusion of clusters of items designed around the same phenomena.

In terms of the extent to which phenomena represent the intended concepts and problems to be solved, panelists found the phenomena from all three test forms are engaging and display strong evidence of alignment. edCount commends the NJDOE on the use of state-specific phenomena and relevant everyday phenomena, which contribute to student engagement. Panelists evaluating the grade 5 form judged the phenomena to be “highly accessible,” though panelists evaluating the grades 8 and 11 forms judged these phenomena to be “somewhat accessible,” citing some distracting or confusing elements, inclusion of content more appropriately suited to a different grade level, or the requirement of skills other than science knowledge; specifically, reading or mathematics skills.

Across the test forms, panelists aligned each item to a DCI, SEP, and CCC, with the option to indicate “no alignment” for each of these dimensions. edCount evaluators used this information to determine the alignment with intended targets. All three forms meet the criteria for strong evidence of alignment with the intended targets, indicating that panelists judged more than 75 percent of the items on the form to align to the intended DCI; panelists also identified 100 percent of items on all three forms as aligning to additional dimensions of the standards (SEP and CCC). edCount commends the NJDOE for the strong representation of multidimensionality within the test forms, reflecting the nature of the NJSLS–S.

All test forms meet expectations for Domain Concurrence, Range of Knowledge, and Balance of Representation. Further, all three test forms display strong evidence of alignment in terms of all three criteria above, with the exception of the grade 5 test form, which meets the criteria for moderate evidence of alignment in terms of Range of Knowledge, given that 27 percent of the Earth and Space Science DCIs are represented on this form. edCount commends the NJDOE for their strong plan for monitoring sampling of all DCIs in these grade levels across forms.

For all three forms, panelists evaluated the items on the form as being cognitively challenging, though panelists noted in all three forms that, while a range of cognitive challenge levels is present within the form, items tend to skew toward the higher levels of cognitive challenge, with less representation at the lower levels.

Mapping items to PLDs is not required in the federal peer review elements but provides critical insight into how well the set of items on which students' scores are based reflects the descriptions of performance and skills within the PLDs. edCount commends the NJDOE on the inclusion of PLD levels within their item and test development processes. For this component of alignment, edCount evaluators examined panelist alignments of items to PLD levels to determine the extent to which the distribution is adequate to support score interpretations for all performance levels. The grade 11 test form shows moderate evidence of alignment in terms of PLD range, with the distribution of items across the PLDs unevenly supporting adequate score interpretations for all performance levels. The grades 5 and 8 test forms show limited evidence of alignment in terms of PLD range; the distribution of items across the PLDs is inadequate to support score interpretations for the lower performance levels for both of these forms. These findings are consistent with panelist comments on the cognitive challenge level of the NJSLA–S forms.

Given the findings for Evaluation Question 3, edCount recommends the NJDOE 1) consider including the intended representation of score point values by reporting category, as well as item numbers, within test development documentation; 2) consider reviewing the performance levels of items on the assessment to ensure that the forms support score interpretations across all four performance levels; and 3) consider reviewing phenomena that panelists identified as not meeting the highest expectations for accessibility.

For the three NJSLA–S forms reviewed, all forms meet expectations across most evaluation criteria addressing both test development and alignment outcomes. Test development activities follow industry practices, and edCount commends the NJDOE for the inclusion of key stakeholders throughout the process. While the alignment outcomes for the test forms reviewed are overwhelmingly positive, edCount encourages the NJDOE to consider the findings and recommendations in their ongoing improvement efforts.

These findings are notable given the depth and breadth of the methodology used, which exceeds the requirements laid out for state assessments through federal peer review. A test form is the product of a complex, multifaceted development process, and the levels of alignment for the NJSLA–S forms are the outcome of a clear and standardized test development process. We commend the NJDOE on the development of test forms that meet the majority of the rigorous expectations of this alignment evaluation.

APPENDIX M: Executive Summary of Evaluation of the Cognitive Process Study

Appendix M contains the executive summary from the cognitive process study report submitted by edCount, LLC, in February 2023.

Introduction

The New Jersey Department of Education (NJDOE) commissioned edCount, LLC, (edCount) to conduct an independent evaluation of the degree to which the items on the New Jersey Student Learning Assessment–Science (NJSLA–S) in grades 5, 8, and 11 elicit the intended response processes as represented in the New Jersey Student Learning Standards for Science (NJSLS–S) in 2023. Critical Element 3.2 of the state assessment peer review guidance requires states to provide evidence that their assessments tap the intended cognitive processes appropriate for each grade level as represented in the state's academic content standards (U.S. Department of Education, 2018). The NJDOE intends to use the results of this evaluation to inform decisions about future item and assessment development and for federal peer review purposes.

To serve this purpose, edCount conducted a cognitive lab study of a strategic sample of grade 5, 8, and 11 NJSLA–S items with a sample of New Jersey students. The goal of the cognitive lab study was to investigate the degree to which:

1. the items elicit the intended construct-relevant response processes appropriate for the grade level;
2. the items include any construct-irrelevant attributes that interfere with students' demonstration of their knowledge and skills; and
3. the items require complex demonstrations or applications of knowledge and skills.

edCount managed all materials and logistics for the cognitive lab study, provided evaluators to conduct the study, conducted all data analyses, and produced a final written report documenting the study and its findings. edCount worked with the NJDOE to confirm all data collection protocols and cognitive lab methodology, and finalize the sample of items to be evaluated, as well as the characteristics of the sample of students. Additionally, edCount presented the proposal for the study to New Jersey's Technical Advisory Committee (TAC). Finally, edCount worked with the selected NJDOE-approved districts on the selection of students for inclusion in the study, as well as the timing of the data collection event.

edCount recommends that the NJDOE also use the results of this evaluation to inform decisions about future item and assessment development.

Evaluation Methodology

Evidence of intended and elicited cognitive processes is critical to a validity evaluation for standards-based assessments. Such evidence must draw upon an examination of how the test was designed and developed and how the items elicit the cognitive processes they intend to measure. As in the case for all validity evidence, evidence of cognitive processes is necessary to support the interpretation and use of test scores, to support inferences of student performance as an accurate reflection of the way the content is assessed.

edCount's approach to evaluating cognitive processes between the aforementioned NJSLA–S and the NJSLS–S encompasses the collection and evaluation of a comprehensive body of evidence. This evidence aligns with the demands of both the federal peer review criteria and the Standards for Educational and Psychological Testing (*The Standards*; AERA, APA, & NCME, 2014).

The methodology for this evaluation was carefully developed based on studies that have utilized cognitive labs to capture a verbal report of a problem solver's account of their own mental processing (Baxter & Glaser, 1998; Ericsson & Simon, 1993) using concurrent and retrospective accounts (Cohen, 1987; Leighton, 2004). Evaluators asked students to describe their thinking concurrently while working through the sample items and had them complete retrospective cognitive interview accounts to clarify the concurrent account and allow additional time for probing questions such as students' level of familiarity with the topics/phenomena included within the item set.

To address the unique aspects of the multidimensional nature of the NJSLA–S and the NJSLS–S, edCount addressed the following questions:

1. To what extent do the items on the NJSLA–S elicit the intended cognitive processes as represented in the NJSLS–S?
    a. To what degree do the items on the NJSLA–S elicit the intended construct-relevant response processes appropriate for the grade level?
    b. To what degree do the items include any construct-irrelevant attributes that interfere with students' demonstration of their knowledge and skills?
2. How do students interact with the item types within the NJSLA–S?
    a. To what degree do students interact with the stimuli and assessment activities as intended?
    b. Do any aspects of the item interfere with students' ability to respond, and if so, how?

Evaluation Findings and Recommendations

Evaluation Question 1: Findings and Recommendations

Overall, students engaged with items from the NJSLA–S in the manner intended. Students made valid attempts at the items, fully understood key information most of the time, and grappled in problem-solving in an appropriate manner. The data results in Chapter 3 of the report prepared by edCount detail how students engaged with the items and the associated SEP reporting categories of Critiquing, Investigating, and Sensemaking. When students had difficulty in engaging with items, it was largely during the problem-solving process, which is an appropriate construct-relevant challenge for students to be facing.

While very few items were found to be problematic for students, edCount recommends the NJDOE review the items some students found to be confusing or unclear to determine whether there is a more direct manner in which to pose the question.

For items containing multiple stimuli, edCount recommends the NJDOE consider scaffolding the stimuli so as not to overwhelm students with information that may not be needed to answer some questions. As the students work through a set of items associated with a phenomenon, additional information can be introduced.

edCount commends the overall construction of the NJSLA–S items, as students were challenged, yet able to apply appropriate reasoning and problem-solving. The vast majority of students made valid attempts on the items (97 percent) and demonstrated full understanding of key information in 89 percent of item attempts. For problem-solving, students executed the problem-solving process without issues in 76 percent of the attempts.

edCount commends the format and diversity of items. Throughout the study, students engaged with the test as intended. In post-cognitive lab interviews, no item was universally liked or disliked by all students. In fact, the questions that students most often mentioned tended to be polarizing, both favorite and least favorite, and reflected a matter of personal preference rather than any issues with item construction.

Evaluation Question 2: Findings and Recommendations

Students found the technological interface of the test to be accessible and faced no trouble in moving through all test items. Many students deftly moved between items based on their comfort level with the assigned items. Few cited any difficulties with the test format or the interface of the various item types. In instances where students discussed these item types in the post-cognitive lab interviews, it was clear that their comments were a matter of preference rather than confusion.

The analysis on item types executed in Chapter 3 showed that there were some differences between different item types. Students performed more poorly on multi-select selected-response items as well as drag-and-drop/order items. While not always the case, some of the challenges noted in these items were related to gaps in content knowledge or academic vocabulary.

Because some item types, such as drag-and-drop and multi-select selected response, resulted in lower response accuracy, edCount recommends a review of these item types across the item bank to determine whether this trend is consistent or an artifact of the items selected for this study.

The test administration platform did not interfere with students' ability to respond on the NJSLA–S. Across all students and items, no events occurred, and there were no comments made about how the technological interface of the test was a challenge to navigate.

edCount commends the thoroughness with which the NJSLA–S items were developed. Fewer than 2 percent of item attempts were marred by any difficulty in understanding the item, and this trend was not concentrated in any area (DCI or SEP). Students reported being appropriately challenged by the items, and comments were generally positive.

REFERENCES

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

Baxter, G. P., & Glaser, R. (1998) Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 17(3), 37–45.

Cohen, A. (1987). Using verbal reports in research on language learning. In: C. Færch & G. Kasper (Eds.), Introspection in second language research (pp. 82–95). Clevedon: Multilingual Matters.

Ericsson, K. A., & Simon H. A. (1993). Protocol analysis: verbal reports as data. Cambridge, MA: MIT Press.

Leighton, J.P. (2021). Rethinking think-alouds: The often-problematic collection of response process data. Applied Measurement in Education, 31(1), 61–74.

APPENDIX N: Observed *p*-values for the Fit and Underfit Subgroups of Students

Table N.1: Grade 5 Fit and Underfit *p*-values

				ACC		ML		EconDis		SWD	
UIN	Domain	Practice	Item Type	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
1905B001_04B	Earth and Space Science	Sensemaking	TE	.592	.251	.484	.225	.605	.262	.581	.232
1905B001_08B	Earth and Space Science	Investigating	TE	.296	.091	.167	.055	.314	.099	.275	.072
1905B001_11	Earth and Space Science	Investigating	MC	.481	.336	.421	.328	.479	.341	.468	.338
1905M069_01	Physical Science	Investigating	TE	.289	.088	.192	.069	.272	.086	.297	.086
1905M069_02	Physical Science	Sensemaking	TE	.216	.077	.118	.041	.205	.068	.221	.070
1905M069_05	Physical Science	Investigating	TE	.145	.077	.068	.049	.119	.058	.156	.072
2005M015_02	Earth and Space Science	Sensemaking	MC	.617	.309	.469	.245	.607	.307	.618	.315
2005M015_04	Earth and Space Science	Sensemaking	MC	.312	.209	.239	.185	.280	.190	.297	.209
2005M015_05	Earth and Space Science	Sensemaking	TE	.296	.166	.195	.126	.283	.148	.289	.140
2005M015_07	Earth and Space Science	Investigating	TE	.286	.099	.139	.064	.287	.100	.265	.099
2005M015_10	Earth and Space Science	Sensemaking	TE	.296	.111	.181	.082	.301	.125	.308	.114
2105M003_01	Life Science	Investigating	TE	.249	.137	.128	.105	.228	.133	.242	.124
2105M003_04	Life Science	Critiquing	MC	.399	.209	.315	.184	.386	.196	.381	.205
2105M003_07	Life Science	Sensemaking	MC	.641	.289	.506	.253	.655	.296	.632	.290
2105M003_09	Life Science	Critiquing	TE	.181	.113	.098	.084	.158	.100	.177	.092
2205B000_02	Life Science	Investigating	TE	.165	.145	.098	.130	.160	.139	.139	.114
2205B000_03	Life Science	Investigating	MC	.399	.235	.313	.193	.388	.219	.390	.227
2205B000_04	Life Science	Investigating	TE	.375	.097	.176	.043	.374	.088	.354	.073
2205B000_07	Life Science	Critiquing	CR	.304	.060	.156	.026	.298	.053	.279	.047
2205B000_10	Life Science	Sensemaking	MC	.265	.211	.225	.193	.277	.206	.255	.203
2205B007_02	Physical Science	Sensemaking	TE	.219	.104	.130	.083	.195	.104	.217	.091
2205B007_03	Earth and Space Science	Critiquing	CR	.523	.151	.421	.137	.522	.147	.518	.144
2205B007_04	Physical Science	Investigating	TE	.353	.111	.222	.068	.353	.106	.326	.093
2205B007_05	Physical Science	Investigating	TE	.160	.253	.101	.222	.140	.231	.157	.250
2205M007_01	Earth and Space Science	Investigating	TE	.271	.089	.122	.048	.254	.078	.257	.076
2205M007_02	Earth and Space Science	Critiquing	TE	.194	.237	.132	.208	.178	.229	.190	.245
2205M007_04	Earth and Space Science	Critiquing	MC	.416	.173	.250	.123	.392	.147	.418	.161
2205M007_05	Earth and Space Science	Sensemaking	TE	.382	.159	.284	.139	.388	.159	.386	.149

UIN	Domain	Practice	Item Type	ACC		ML		EconDis		SWD	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2205M024_03	Physical Science	Sensemaking	TE	.390	.203	.290	.179	.364	.189	.392	.199
2205M024_04	Physical Science	Critiquing	TE	.251	.213	.205	.204	.259	.223	.244	.194
2205M024_05	Physical Science	Investigating	TE	.266	.133	.144	.097	.225	.102	.255	.105
2305B004_01	Physical Science	Investigating	MC	.365	.176	.250	.146	.352	.162	.352	.147
2305B004_03	Physical Science	Critiquing	TE	.395	.178	.320	.156	.397	.170	.372	.163
2305B004_04	Physical Science	Critiquing	TE	.204	.231	.142	.202	.184	.215	.198	.223
2305B004_09	Physical Science	Sensemaking	CR	.210	.070	.124	.039	.204	.066	.200	.057
2305B006_01	Physical Science	Sensemaking	TE	.344	.128	.176	.076	.350	.128	.353	.127
2305B006_03	Physical Science	Investigating	MC	.345	.192	.248	.166	.317	.175	.344	.177
2305B006_04	Physical Science	Critiquing	TE	.330	.265	.278	.241	.319	.261	.335	.277
2305B011_01	Life Science	Investigating	TE	.224	.240	.171	.238	.202	.244	.222	.244
2305B011_02	Life Science	Critiquing	TE	.418	.280	.342	.250	.396	.252	.431	.269
2305B011_03	Earth and Space Science	Critiquing	TE	.079	.276	.031	.217	.077	.273	.075	.288
2305B011_04	Earth and Space Science	Sensemaking	TE	.350	.162	.238	.152	.332	.166	.356	.155
2305B011_06	Life Science	Sensemaking	TE	.315	.104	.158	.060	.302	.097	.319	.104
2405B002_01	Life Science	Investigating	TE	.189	.286	.152	.266	.181	.268	.197	.272
2405B002_02	Life Science	Investigating	MC	.384	.215	.287	.188	.368	.209	.371	.203
2405B002_03	Life Science	Critiquing	TE	.387	.126	.238	.077	.391	.121	.386	.118
2405B002_04	Life Science	Sensemaking	TE	.204	.148	.128	.130	.194	.152	.191	.141
2405B008_03	Earth and Space Science	Investigating	TE	.123	.064	.064	.038	.115	.066	.125	.058
2405B008_05	Earth and Space Science	Sensemaking	TE	.140	.301	.100	.300	.129	.300	.136	.286
2405B008_06	Earth and Space Science	Critiquing	TE	.182	.219	.129	.209	.178	.217	.180	.212
2405M008_02	Physical Science	Sensemaking	TE	.141	.115	.053	.083	.129	.106	.139	.105
2405M008_04	Physical Science	Critiquing	MC	.437	.299	.358	.257	.414	.282	.440	.303
2405M008_05	Physical Science	Sensemaking	TE	.261	.230	.209	.228	.253	.226	.257	.222

Note. P_FT = *p*-values for the group of students not showing underfit; P_UF = *p*-values for the group of students showing underfit.

Table N.2: Grade 8 Fit and Underfit *p*-values

UIN	Domain	Practice	Item Type	ACC		ML		EconDis		SWD	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2008M025_01	Physical Science	Sensemaking	TE	.291	.032	.179	.032	.302	.047	.277	.037
2008M025_02	Physical Science	Investigating	MC	.320	.159	.261	.168	.317	.158	.306	.152
2008M025_08	Physical Science	Critiquing	MC	.326	.250	.260	.250	.320	.267	.325	.260
2108M000_02	Earth and Space Science	Investigating	TE	.280	.094	.179	.093	.274	.097	.279	.096
2108M000_03	Earth and Space Science	Sensemaking	TE	.167	.034	.066	.021	.150	.031	.182	.034
2108M000_06	Earth and Space Science	Critiquing	TE	.187	.132	.134	.126	.185	.139	.186	.133
2308B003_01	Earth and Space Science	Sensemaking	TE	.378	.101	.239	.096	.363	.101	.371	.110
2308B003_03	Earth and Space Science	Investigating	TE	.249	.072	.149	.063	.250	.072	.244	.079
2308B003_05	Earth and Space Science	Critiquing	CR	.330	.041	.248	.047	.323	.048	.318	.041
2308B003_07	Earth and Space Science	Investigating	TE	.184	.139	.106	.138	.168	.152	.193	.146
2308B003_11	Earth and Space Science	Critiquing	TE	.508	.103	.355	.099	.522	.115	.516	.110
2308B006_01	Life Science	Sensemaking	TE	.142	.133	.084	.132	.133	.151	.148	.149
2308B006_07	Life Science	Sensemaking	TE	.201	.222	.174	.210	.192	.215	.216	.222
2308B006_08	Life Science	Sensemaking	TE	.187	.128	.143	.139	.177	.136	.197	.120
2308B011_01	Physical Science	Investigating	TE	.186	.171	.156	.170	.188	.182	.182	.163
2308B011_06	Earth and Space Science	Investigating	MC	.308	.144	.178	.142	.302	.153	.321	.159
2308B011_07	Physical Science	Critiquing	TE	.438	.086	.277	.069	.442	.105	.435	.105
2308B011_08	Physical Science	Sensemaking	TE	.380	.198	.337	.199	.382	.180	.364	.189
2308B018_01	Life Science	Sensemaking	TE	.410	.070	.261	.070	.418	.079	.407	.064
2308B018_03	Life Science	Investigating	MC	.432	.116	.320	.137	.435	.118	.402	.104
2308B018_04	Life Science	Investigating	TE	.153	.092	.085	.106	.140	.106	.147	.093
2308B018_07	Life Science	Critiquing	TE	.198	.127	.105	.115	.189	.127	.191	.129
2308B018_09	Life Science	Critiquing	CR	.353	.033	.197	.021	.363	.032	.343	.030
2308M003_01	Life Science	Sensemaking	TE	.329	.148	.263	.152	.323	.145	.310	.129
2308M003_04	Life Science	Investigating	MC	.356	.264	.322	.265	.345	.252	.351	.242
2308M003_07	Life Science	Investigating	MC	.122	.085	.053	.085	.105	.087	.112	.090
2308M006_03	Life Science	Investigating	MC	.155	.459	.134	.443	.131	.447	.159	.471
2308M006_05	Life Science	Sensemaking	TE	.216	.197	.141	.176	.203	.183	.217	.217
2308M006_06	Earth and Space Science	Critiquing	TE	.141	.123	.079	.127	.127	.137	.136	.123
2308M006_08	Life Science	Critiquing	TE	.379	.088	.254	.080	.372	.097	.389	.088
2308M006_09	Earth and Space Science	Sensemaking	TE	.222	.103	.164	.103	.220	.109	.213	.099

UIN	Domain	Practice	Item Type	ACC		ML		EconDis		SWD	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2308M010_01	Earth and Space Science	Critiquing	TE	.177	.070	.116	.063	.179	.076	.178	.067
2308M010_04	Earth and Space Science	Critiquing	TE	.172	.078	.081	.066	.178	.080	.175	.085
2308M010_05	Earth and Space Science	Sensemaking	TE	.143	.191	.101	.178	.136	.183	.151	.196
2308M010_07	Earth and Space Science	Investigating	MC	.380	.334	.341	.319	.372	.323	.383	.342
2308M020_05	Physical Science	Critiquing	TE	.311	.124	.209	.124	.324	.131	.301	.119
2308M020_09	Physical Science	Investigating	TE	.266	.109	.186	.121	.255	.121	.249	.105
2308M020_11	Physical Science	Investigating	MC	.600	.200	.511	.199	.613	.230	.609	.198
2308M027_01	Physical Science	Investigating	TE	.323	.100	.241	.098	.320	.109	.304	.089
2308M027_02	Physical Science	Investigating	TE	.197	.186	.140	.169	.181	.186	.198	.207
2308M027_03	Physical Science	Critiquing	TE	.159	.087	.089	.091	.153	.100	.155	.080
2308M027_05	Physical Science	Sensemaking	TE	.191	.033	.079	.028	.173	.036	.181	.032
2308M043_01	Life Science	Sensemaking	MC	.249	.266	.218	.283	.235	.279	.247	.249
2308M043_03	Life Science	Sensemaking	TE	.185	.203	.156	.219	.178	.213	.185	.199
2308M043_07	Life Science	Investigating	MC	.428	.308	.373	.293	.415	.296	.445	.324
2308M049_01	Life Science	Sensemaking	TE	.172	.049	.080	.043	.161	.047	.182	.052
2308M049_03	Physical Science	Sensemaking	TE	.227	.117	.160	.127	.212	.121	.226	.108
2308M049_05	Life Science	Sensemaking	TE	.221	.202	.149	.192	.202	.193	.219	.197
2308M049_07	Life Science	Investigating	MC	.465	.290	.442	.305	.463	.293	.469	.283
2408B005_01	Physical Science	Investigating	TE	.186	.093	.095	.098	.180	.108	.180	.093
2408B005_02	Physical Science	Sensemaking	TE	.561	.107	.341	.078	.562	.120	.567	.118
2408B005_03	Physical Science	Critiquing	TE	.543	.124	.469	.145	.555	.141	.508	.117
2408B005_04	Physical Science	Sensemaking	CR	.221	.014	.141	.016	.217	.015	.192	.009
2408B005_05	Physical Science	Critiquing	TE	.275	.088	.202	.101	.269	.095	.274	.085
2408M003_01	Earth and Space Science	Investigating	MC	.413	.132	.326	.163	.388	.134	.392	.120
2408M003_02	Earth and Space Science	Critiquing	TE	.133	.134	.063	.135	.121	.149	.132	.140
2408M003_03	Earth and Space Science	Critiquing	MC	.275	.118	.207	.129	.293	.122	.256	.109
2408M039_02	Earth and Space Science	Investigating	TE	.313	.048	.170	.038	.304	.048	.296	.049
2408M039_03	Earth and Space Science	Critiquing	TE	.149	.203	.085	.197	.135	.209	.146	.203
2408M039_08	Earth and Space Science	Sensemaking	MC	.515	.240	.438	.239	.513	.243	.527	.235
2408M047_03	Physical Science	Investigating	TE	.132	.093	.082	.091	.129	.102	.124	.084
2408M047_04	Life Science	Sensemaking	TE	.331	.135	.235	.134	.324	.132	.343	.136
2408M047_06	Physical Science	Critiquing	MC	.461	.237	.417	.250	.462	.250	.454	.239

Note. P_FT = *p*-values for the group of students not showing underfit; P_UF = *p*-values for the group of students showing underfit.

Table N.3: Grade 11 Fit and Underfit *p*-values

UIN	Domain	Practice	Item Type	ACC		ML		EconDis		SWD	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2011M028_01	Life Science	Investigating	MC	.570	.326	.507	.340	.610	.354	.597	.332
2011M028_03	Earth and Space Science	Sensemaking	TE	.377	.224	.298	.218	.377	.214	.389	.207
2011M028_04	Life Science	Critiquing	MC	.203	.311	.228	.342	.192	.290	.192	.277
2011M028_05	Earth and Space Science	Investigating	TE	.397	.137	.240	.108	.428	.144	.457	.168
2011M031_03	Earth and Space Science	Sensemaking	TE	.146	.200	.094	.172	.149	.183	.163	.187
2011M031_04	Earth and Space Science	Critiquing	TE	.268	.229	.226	.234	.275	.243	.275	.235
2011M031_05	Earth and Space Science	Sensemaking	MC	.266	.199	.244	.200	.253	.176	.257	.193
2011M031_06	Earth and Space Science	Investigating	MC	.338	.266	.327	.285	.341	.260	.328	.244
2011M059_01	Life Science	Sensemaking	TE	.185	.034	.074	.021	.237	.040	.235	.050
2011M059_02	Life Science	Sensemaking	MC	.430	.277	.371	.256	.457	.274	.458	.301
2011M059_05	Life Science	Sensemaking	MC	.310	.278	.318	.312	.340	.317	.315	.307
2011M059_08	Life Science	Investigating	MC	.122	.452	.121	.443	.120	.430	.127	.436
2111B002_01	Life Science	Critiquing	TE	.224	.077	.108	.057	.247	.078	.264	.079
2111B002_03	Life Science	Sensemaking	TE	.367	.102	.271	.090	.409	.119	.379	.104
2111B002_07	Life Science	Investigating	TE	.303	.128	.226	.125	.318	.127	.319	.119
2111B002_10	Life Science	Sensemaking	TE	.178	.058	.080	.051	.198	.059	.204	.065
2111B002_11	Life Science	Investigating	CR	.374	.054	.207	.024	.425	.051	.390	.058
2111M020_01	Physical Science	Investigating	MC	.495	.233	.428	.255	.521	.239	.493	.215
2111M020_03	Physical Science	Sensemaking	TE	.165	.231	.111	.209	.164	.230	.162	.242
2111M020_05	Physical Science	Critiquing	TE	.360	.126	.211	.099	.382	.108	.388	.125
2211M000_02	Life Science	Critiquing	TE	.111	.091	.053	.077	.141	.101	.137	.103
2211M000_03	Life Science	Sensemaking	MC	.298	.192	.263	.220	.339	.217	.317	.193
2211M000_06	Life Science	Critiquing	TE	.236	.095	.131	.103	.283	.109	.271	.108
2211M000_08	Life Science	Investigating	MC	.228	.228	.178	.228	.253	.244	.254	.246
2211M000_09	Life Science	Investigating	TE	.086	.101	.039	.082	.100	.102	.100	.109
2211M013_02	Earth and Space Science	Critiquing	MC	.371	.224	.311	.213	.393	.207	.369	.215
2211M013_03	Earth and Space Science	Sensemaking	TE	.209	.046	.067	.032	.250	.055	.262	.051
2211M013_05	Earth and Space Science	Investigating	TE	.297	.192	.210	.181	.298	.172	.325	.207
2211M016_01	Physical Science	Sensemaking	TE	.239	.070	.139	.067	.255	.080	.268	.079
2211M016_02	Physical Science	Investigating	MC	.417	.205	.338	.213	.437	.206	.427	.201

UIN	Domain	Practice	Item Type	ACC		ML		EconDis		SWD	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2211M016_06	Earth and Space Science	Critiquing	TE	.269	.196	.190	.171	.273	.171	.306	.199
2211M016_08	Physical Science	Sensemaking	TE	.112	.074	.046	.070	.116	.089	.130	.089
2211M031_01	Life Science	Investigating	TE	.302	.208	.254	.216	.318	.208	.327	.219
2211M031_03	Life Science	Critiquing	TE	.200	.071	.101	.056	.208	.078	.214	.081
2211M031_04	Life Science	Sensemaking	TE	.412	.110	.242	.101	.475	.133	.450	.130
2211M031_05	Life Science	Critiquing	TE	.491	.214	.416	.219	.507	.207	.492	.202
2311B004_01	Physical Science	Investigating	TE	.301	.109	.225	.135	.335	.125	.317	.110
2311B004_04	Physical Science	Sensemaking	TE	.135	.295	.111	.315	.135	.288	.155	.307
2311B004_05	Physical Science	Critiquing	CR	.199	.060	.097	.035	.219	.048	.206	.048
2311B014_01	Earth and Space Science	Investigating	MC	.334	.208	.257	.199	.330	.186	.335	.190
2311B014_02	Earth and Space Science	Sensemaking	TE	.243	.164	.174	.155	.253	.162	.270	.158
2311B014_04	Earth and Space Science	Critiquing	CR	.482	.104	.302	.058	.518	.095	.516	.109
2311M011_03	Physical Science	Critiquing	TE	.236	.121	.149	.098	.277	.107	.257	.104
2311M011_05	Earth and Space Science	Sensemaking	MC	.588	.365	.554	.406	.611	.387	.608	.364
2311M011_07	Earth and Space Science	Investigating	MC	.540	.293	.490	.301	.541	.273	.521	.267
2311M015_01	Physical Science	Sensemaking	TE	.493	.228	.407	.219	.520	.226	.510	.213
2311M015_03	Physical Science	Critiquing	TE	.330	.147	.197	.112	.348	.130	.361	.158
2311M015_05	Physical Science	Investigating	MC	.189	.206	.154	.186	.201	.209	.197	.205
2311M015_06	Physical Science	Critiquing	MC	.173	.238	.158	.257	.180	.235	.174	.221
2311M019_01	Physical Science	Sensemaking	TE	.261	.129	.169	.110	.246	.125	.258	.126
2311M019_02	Physical Science	Investigating	MC	.342	.194	.262	.191	.353	.189	.350	.198
2311M019_03	Physical Science	Critiquing	TE	.310	.153	.248	.129	.310	.134	.322	.137
2311M019_06	Physical Science	Sensemaking	MC	.336	.213	.288	.211	.333	.206	.356	.211
2311M038_01	Physical Science	Investigating	MC	.368	.287	.336	.287	.358	.272	.373	.277
2311M038_03	Physical Science	Critiquing	TE	.227	.079	.138	.071	.255	.081	.245	.081
2311M038_04	Physical Science	Sensemaking	TE	.199	.172	.166	.167	.198	.152	.209	.160
2311M053_01	Life Science	Investigating	MC	.362	.274	.269	.226	.379	.251	.400	.268
2311M053_03	Life Science	Sensemaking	MC	.438	.195	.309	.172	.442	.184	.453	.199
2311M053_04	Life Science	Investigating	TE	.177	.116	.095	.125	.168	.127	.194	.126
2311M053_05	Life Science	Sensemaking	MC	.348	.261	.311	.250	.343	.251	.350	.262
2311M053_06	Life Science	Critiquing	TE	.590	.229	.446	.190	.602	.200	.619	.222

UIN	Domain	Practice	Item Type	ACC		ML		EconDis		SWD	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2311M058_01	Earth and Space Science	Investigating	MC	.452	.302	.455	.378	.449	.322	.432	.296
2311M058_04	Physical Science	Sensemaking	TE	.278	.218	.233	.198	.301	.211	.291	.224
2311M058_05	Earth and Space Science	Critiquing	TE	.150	.266	.123	.254	.159	.283	.167	.301
2411M017_01	Earth and Space Science	Sensemaking	MC	.515	.203	.398	.204	.552	.205	.545	.191
2411M017_02	Earth and Space Science	Investigating	TE	.197	.105	.107	.102	.207	.105	.216	.101
2411M017_04	Earth and Space Science	Investigating	TE	.346	.164	.234	.157	.328	.123	.342	.107
2411M017_06	Earth and Space Science	Critiquing	TE	.154	.127	.074	.100	.161	.103	.192	.111
2411M030_02	Earth and Space Science	Investigating	TE	.279	.143	.209	.135	.298	.169	.296	.169
2411M030_03	Earth and Space Science	Sensemaking	MC	.363	.329	.332	.310	.357	.311	.385	.334
2411M030_07	Earth and Space Science	Sensemaking	TE	.162	.106	.068	.072	.172	.100	.187	.109
2411M030_09	Earth and Space Science	Critiquing	TE	.531	.218	.469	.221	.528	.186	.513	.174

Note. P_FT = *p*-values for the group of students not showing underfit; P_UF = *p*-values for the group of students showing underfit.