New Jersey
Student Learning Assessment—Science
(NJSLA–S)

TECHNICAL REPORT
Grades 5, 8, and 11
2024

April 2025
PTM XXXX.XX

State of New Jersey
Department of Education

Table of Contents

Tables and Figures

PART 1: INTRODUCTION

This Technical Report provides information about the technical characteristics of the 2024 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 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*."

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 2024 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 24, 2024, through May 24, 2024.

- 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 National Research Council's (NRC; 2012) *Framework for K–12 Science Education*, 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 2024 was the fourth 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 items (TEI). The MSA contains a mixture of TEI and multiple-choice items.

Furthermore, the tests cover a range of material. 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, and the CCC concepts 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 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
PS1.A: Structure and matter
PS1.B: Chemical reactions
PS1.C: Nuclear Processes
PS2: Motion and Stability: Force and Interactions
PS2.A: Force and motion
PS2.B: Types of interactions
PS2.C: Stability and instability in physical systems
PS3: Energy
PS3.A: Definitions of energy
PS3.B: Conservation of energy and energy transfer
PS3.C: Relationship between energy and forces
PS3.D: Energy in chemical processes and everyday life
PS4: Waves and Their Applications in Technologies for Information Transfer
PS4.A: Wave properties
PS4.B: Electromagnetic radiation
PS4.C: Information technologies and instrumentation

Scientific practices. The *Framework* (2012) contains eight different Scientific 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 is 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 CCC concepts 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 Proficiency-Level Descriptors (PLDs). Both the scale score 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, they 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 used to report overall student proficiency on the NJSLA–S. Grade-appropriate Proficiency-Level Descriptors (PLDs) translate these proficiency levels into words. They describe the KSAs students should have at each proficiency level, Level 1 through Level 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 sub-contractor for 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 2024. 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 and each DCI, each scientific practice, and each SEP, as well as 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; each constructed-response (CR) item is worth three or four points. Each constructed-response 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 a machine-scorable (MSA) and a performance-based (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 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

Machine-scorable assessment (MSA). 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.

Performance-based assessment (PBA). 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 student 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. Since 2023, NJDOE has required the PBA section to contain 7 to 8 total points, with three or four 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. 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 are 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, non-biased 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 scientific content domain and DCI, a scientific practice and SEP, 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 to use no more than 50% MC items in each cluster of items. All items are also aligned to one of Webb's (1997; 2002) Depth-of-Knowledge (DOK) classifications.

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 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 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 2024, the NJSAC comprised a diverse group of New Jersey science educators representing 20 of the 21 New Jersey counties. The districts each NJSAC member represents 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 2024 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 thirteen NJBSC members, eleven taught special education status students, seven specialized in teaching students designated as English learners, and five were bilingual. Collectively, they had over 120 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 it was 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 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 4,000 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. MI's psychometric staff leads the statistical review and either trains or re-trains 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 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 psychometric 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 reduces test form reliability. Under classical test theory, 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 .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 and only used if necessary.

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. The content and psychometric review then resumes until the test matches the content and statistical constraints of the NJSLA–S.

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 2024 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 7- or 8-point PBA sections representing 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 8 points aligned to Life Science content domain; 12 points aligned to the Earth Science domain; and 10 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	3
	Investigating	6	6	8
	Sensemaking	8	6	8
8	Critiquing	12	8	10
	Investigating	6	6	8
	Sensemaking	7	9	6
11	Critiquing	9	8	11
	Investigating	9	10	9
	Sensemaking	7	6	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 less balanced than content domains for grade 5, with only 18 out of 60 points being allocated to the Critiquing reporting category. Despite being less than ideal, the 18 points were still enough to produce reliable measures of student Investigating abilities. Other content considerations that were met included: MC items only made up 14 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 by multiple points 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: 2024 NJSLA–S Grade 5 Item Counts and Point Totals by Reporting Category

Domains/Practices	Item Counts				Points
	MC	TE	CR	Total	
Earth and Space	5	12	1	18	21
Life	5	11	1	17	20
Physical	4	12	1	17	19
Total–Domains	14	35	3	52	60
Critiquing	1	9	2	12	18
Investigating	9	11	0	20	20
Sensemaking	4	15	1	20	22
Total–Practices	14	35	3	52	60

Table 2.4.3: 2024 NJSLA–S Grade 5 DCIs

DCI	Item Counts	Points
ESS1	3	3
ESS2	9	12
ESS3	6	6
LS1	2	2
LS2	9	12
LS3	3	3
LS4	3	3
PS1	3	3
PS2	1	1
PS3	7	9
PS4	6	6

Table 2.4.4: 2024 NJSLA–S Grade 5 SEPs

SEP	Item Counts	Points
AQDP	11	11
PACI	6	6
UMCT	3	3
DUM	7	7
AID	3	3
CEDS	10	12
EAE	11	17
OECI	1	1

Table 2.4.5: 2024 NJSLA–S Grade 5 CCCs

CCC	Item Counts	Points
C & E	13	16
E & M	2	2
Patterns	15	15
S & SM	13	18
S, P & Q	2	2
SC	2	2
SF	5	5

The statistical constraints for the 2024 Grade 5 NJSLA–S operational test form were met. One item had item-total correlations below the .20 threshold indicating a low discriminating item (see Section 6.1). 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 87.826 minutes was well below the 105-minute threshold, and out of 52 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: 2024 NJSLA–S Grade 5 Test Construction Statistics

Statistic	Average	Target	Flags
Rasch B	0.698	0.904	N/A
Item-Total Correlation	0.400	> 0.35	1
Infit	1.066	1.00	1
Outfit	1.155	1.00	13
PCM Discrimination	0.863	1.00	8
Lower Asymptote	0.051	0.00	8
Median Time (min)	87.826	< 105	N/A

Table 2.4.7: 2024 NJSLA–S Grade 5 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	51	1	0
White/Black	51	1	0
White/Hispanic	52	0	0
White/Asian	51	1	0

2.4.2 Grade 8 Test Construction

At grade 8, the content domains were well balanced. Out of 72 total score points, the three content domains ranged from 23 to 25 points each, as illustrated in Table 2.4.8. Each content domain had one PBA section devoted to it. The scientific practices were less balanced than content domains, with Investigating 20 points, Critiquing 22 points, and Sensemaking 30 points. Other content considerations that were met included: MC items only made up 14 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: 2024 NJSLA–S Grade 8 Item Counts and Point Totals by Reporting Category

Domains/Practices	Item Counts				Points
	MC	TE	CR	Total	
Earth and Space	4	17	1	22	25
Life	6	15	1	22	24
Physical	4	15	1	20	23
Total–Domains	14	47	3	64	72
Critiquing	5	14	3	22	30
Investigating	7	13	0	20	20
Sensemaking	2	20	0	22	22
Total–Practices	14	47	3	64	72

Table 2.4.9: 2024 NJSLA–S Grade 8 DCIs

DCI	Item Counts	Points
ESS1	5	5
ESS2	8	11
ESS3	9	9
LS1	3	3
LS2	7	9
LS3	6	6
LS4	5	5
PS1	2	2
PS2	2	2
PS3	8	11
PS4	9	9

Table 2.4.10: 2024 NJSLA–S Grade 8 SEPs

SEP	Item Count	Points
AQDP	12	12
PACI	4	4
UMCT	4	4
DUM	6	6
AID	9	9
CEDS	7	7
EAE	12	15
OECI	10	15

Table 2.4.11: 2024 NJSLA–S Grade 8 CCCs

CCC	Item Counts	Points
C & E	23	26
E & M	5	5
Patterns	17	17
S & SM	6	6
S, P & Q	4	4
SC	4	9
SF	5	5

The statistical constraints for the 2024 Grade 8 NJSLA–S operational test form were met. Five grade 8 items were flagged for having item-total correlations below the .20 threshold indicating a low discriminating item (see Section 6.1), including two with the lowest value around .11. However, the other three had values above .18, 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 66.938 was well below the 105-minute threshold, and out of 64 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 three 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 2.4.12 and 2.4.13 summarize the test construction and DIF statistics.

Table 2.4.12: 2024 NJSLA–S Grade 8 Test Construction Statistics

Statistic	Average	Target	Flags
Rasch B	0.302	0.416	N/A
Item-Total Correlation	0.363	> 0.35	5
Infit	1.032	1.00	0
Outfit	1.078	1.00	8
PCM Discrimination	0.946	1.00	4
Lower Asymptote	0.030	0.00	6
Median Time (min)	66.938	< 105	N/A

Table 2.4.13: 2024 NJSLA–S Grade 8 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	64	0	0
White/Black	62	2	0
White/Hispanic	63	1	0
White/Asian	64	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 29 points each. Each content domain had one PBA section. The scientific practices were also less balanced than content domains, with Critiquing 28 points, Investigating 28 points, and Sensemaking 22 points. Other content considerations that were met included: MC items only made up 29 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 were 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: 2024 NJSLA–S Grade 11 Item Counts and Point Totals by Reporting Category

Domains/Practices	Item Counts				Points
	MC	TE	CR	Total	
Earth and Space	9	12	1	22	25
Life	11	10	1	22	24
Physical	9	16	1	26	29
Total–Domains	29	38	3	70	78
Critiquing	6	14	2	22	28
Investigating	14	11	1	26	28
Sensemaking	9	13	0	22	22
Total–Practices	29	38	3	70	78

Table 2.4.15: 2024 NJSLA–S Grade 11 DCIs

DCI	Item Counts	Points
ESS1	7	7
ESS2	8	8
ESS3	7	10
LS1	4	4
LS2	10	12
LS3	3	3
LS4	6	6
PS1	6	6
PS2	4	4
PS3	12	15
PS4	3	3

Table 2.4.16: 2024 NJSLA–S Grade 11 SEPs

SEP	Item Counts	Points
AQDP	8	8
PACI	11	13
UMCT	7	7
DUM	7	7
AID	7	7
CEDS	8	8
EAE	17	23
OECI	5	5

Table 2.4.17: 2024 NJSLA–S Grade 11 CCCs

CCC	Item Counts	Points
C & E	22	27
E & M	7	10
Patterns	5	5
S & SM	14	14
S, P & Q	8	8
SC	10	10
SF	4	4

The 2024 Grade 11 NJSLA–S operational test form construction saw five items flagged for outfit, one item flagged for both infit and outfit, and one item flagged for PCM discrimination. Nevertheless, the flagged items had an outfit, infit, or PCM discrimination 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 value of infit or outfit were close to their ideal values of 1.00. Two grade 11 items were flagged for having item-total correlations below the .20 threshold indicating a low discriminating item (see Section 6.1), with values of .18 and .12. However, 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 68.886 minutes, which was well below the 150-minute target. Of 70 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 three 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: 2024 NJSLA–S Grade 11 Test Construction Statistics

Statistic	Average	Target	Flags
Rasch B	0.536	0.475	N/A
Item-Total Correlation	.389	> 0.35	2
Infit	1.024	1.00	1
Outfit	1.061	1.00	6
PCM Discrimination	0.947	1.00	1
Lower Asymptote	0.033	0.00	0
Median Time (min)	68.886	< 150	N/A

Table 2.4.19: 2024 NJSLA–S Grade 11 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	68	2	0
White/Black	69	1	0
White/Hispanic	69	1	0
White/Asian	70	0	0

2.5 2024 NJSLA–S State of the Item Bank

Upon the completion of the 2024 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 on 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, 2024). 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 at the NJSLA–S website under Documents and Downloads.

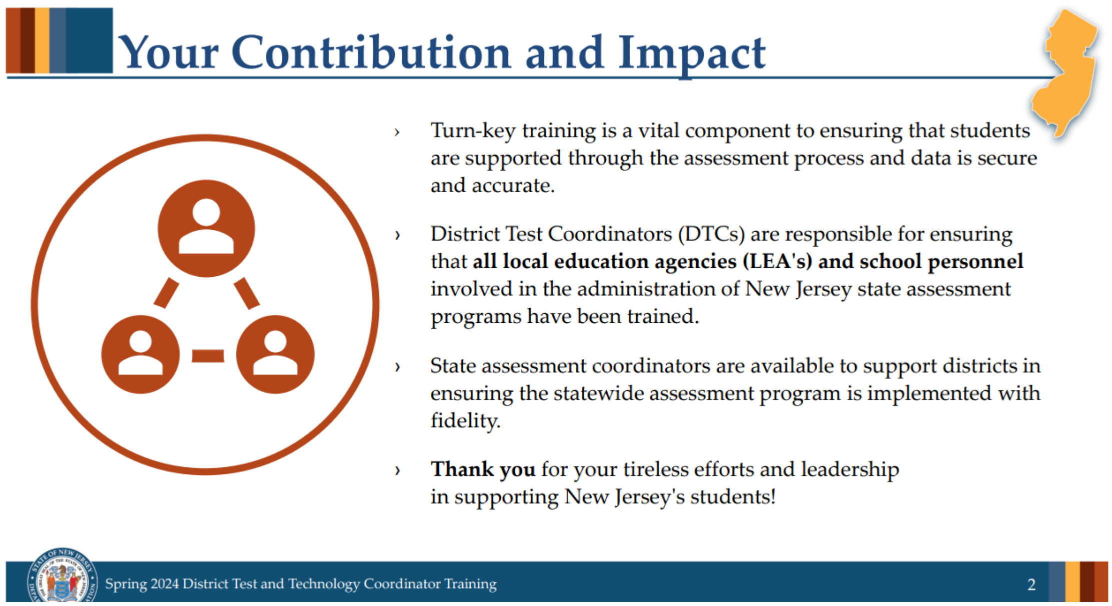


Figure 3.1.1. Slide two from the 2024 DTC

Table 3.1.1 shows the NJSLA–S 2024 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 2024 Grades 5, 8, and 11 Testing Window

Grade	CBT	PBT	Testing Time
5	4/29/24–5/24/24	4/29/24–5/24/24	45 minutes per unit
8	4/29/24–5/24/24	4/29/24–5/24/24	45 minutes per unit
11	4/29/24–5/24/24	4/29/24–5/24/24	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 tablet available.	Section 3.5
Distribute test materials to TAs.	Section 3.9
Manage test sessions in PAN.	Section 4.1.2
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 4.1.4
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 4.1.4
Schedule and supervise make-up testing.	Sections 2.4.2 and 4.1.5
Create make-up test sessions in PAN.	Section 4.1.5
Respond to all technology-related issues.	Section 4.1.3
Collect materials from TAs.	Section 4.1.5
Ensure that all units are locked after testing on each testing day.	Section 4.1.2

3.2.2 Paper-Based Testing

The following section describes the responsibilities of the DTC and STC during PBT administration. Like 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 3.10
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 4.2.2
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 4.2.2
Schedule and supervise make-up testing.	Sections 2.4.2 and 4.2.4
Follow the protocol for contaminated or damaged test materials, and refer to New Jersey policy for reporting these incidents.	Section 4.2.3
Collect materials from TAs, and ensure that all test booklets and answer documents have a student name or student ID label.	Section 4.2.4

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.
- 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

- 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, located under Educator Resources > Test Administration Resources > Test Administrator Manuals as well as on the NJSLA–S website.)
    - 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. 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:

> 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. 13)

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, 12th Edition. 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* (2024) accessibility features are defined as "tools or preferences that are either built into the testing platform or provided externally by Test Administrators" (p. 54). 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.

3.4.2 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* (2024) defines an accommodation as "an assessment practice or procedure that changes the presentation, response, setting, and/or time and scheduling of assessments" (p. 64). 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 Text-to-Speech
- 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 2023 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 equivalence. Occasionally, 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 2023, 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 5 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 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 our 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 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 non-scorable 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 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 the 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 our 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 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 the 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 2024 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	73	3	1
8	78	3	1
11	96	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. Readers 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 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 our 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 reviews student responses to constructed response items and verifies the original scores or makes 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 for all three grade levels. All open-ended/constructed-response item types were scored by a single rater.

Table 4.2.2: Automatic Rescore Results

Grade	Eligible for Automatic Rescore	Number of Changes	Percentage Changed (of those Eligible)
5	4,303	217	5.04
8	3,139	315	10.0
11	3,209	289	9.0

4.3 Quality Control

To confirm that the processing of student tests and test scores was done correctly, Measurement Incorporated 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 on 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 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 August 7, 2023.

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 2024 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 were 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 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 item 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 item mean. The polytomous item flagging criteria involves converting the item mean to a proportion by dividing it by the maximum points possible on the item (i.e., making it an adjusted item mean or a *p-value*). Polytomous items are then flagged if their converted *p-value* falls outside of the .25 to .90 range. It should be noted that the flagging criteria only provide a general guideline, and some productive items have *p-values* outside of the .25 to .90 range. 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.

For NJSLA–S items, Tables 6.1.1, 6.1.3, and 6.1.5 summarize the item difficulties 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 (*rpb*) summaries for grades 5, 8, and 11, respectively. Several intervals of item difficulties or 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.

Overall, the average item difficulties and discriminations appear to be productive for measuring students in New Jersey. At each grade level, the average TE items tended to be slightly more challenging and more discriminating than MC items. The CR items were, as expected, more discriminating than the MC and TE items.

At grade 5, most MC and TE items had item difficulties between .25 and .75, indicating an average item difficulty level on the scale. The grade 5 CR items in Earth and Space Science and Life science tended to be slightly more challenging than the CR item in Physical Science. At grade 8, most MC items had item difficulties between .25 and .75, indicating an average item difficulty level for MC items. While 14 (30%) grade 8 TE items had *p-values* below .25, the remaining TE items had item difficulty between .25 and .75. The grade 8 CR item in Physical Science tended to be more challenging than those in Life Science and Earth and Space Science. At grade 11, most MC items had *p-values* between .25 and .75, and 12 (32%) TE items had *p-values* below .25. indicating that there are several TE items on the harder end of the scale. Additionally for Grade 11, the CR item in Physical Science tended to be more challenging than the CR items in Earth and Space Science and Life Science.

The average item-total correlations (*rpb*) for CR items were .62, .63, and .61 for grades 5, 8, and 11, respectively, indicating good item discrimination. Also, the frequency distributions of item-total correlations for MC and TE items appear to be productive for discriminating between high- and low-achieving students. At grade 5, only one MC item and one TE item had item-total correlations below .20, while five MC and 21 TE items had item-total correlations above .40. At grade 8, three MC items and three TE items had item-total correlations below .20, while two MC and 22 TE items had discriminations above .40. At grade 11, two MC items and one TE items had item-total correlations below .20, while 10 MC and 20 TE items had discriminations above .40.

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	14	1	8	5	0	0	.45	.13	.45
	Earth and Space	5	1	2	2	0	0	.42	.15	.44
	Life	5	0	3	2	0	0	.47	.14	.44
	Physical	4	0	3	1	0	0	.47	.10	.46
	Critiquing	1	0	0	1	0	0	.61	N/A	N/A
	Investigating	9	1	6	2	0	0	.42	.09	.44
	Sensemaking	4	0	2	2	0	0	.48	.19	.47
TE	NJSLA–S	35	4	28	3	0	0	.36	.11	.35
	Earth and Space	12	1	10	1	0	0	.35	.10	.35
	Life	11	1	10	0	0	0	.35	.10	.36
	Physical	12	2	8	2	0	0	.39	.12	.39
	Critiquing	9	2	7	0	0	0	.29	.07	.28
	Investigating	11	2	8	1	0	0	.35	.12	.30
	Sensemaking	15	0	13	2	0	0	.42	.09	.43
CR	NJSLA–S	3	0	3	0	0	0	.37	.07	.35
	Earth and Space	1	0	1	0	0	0	.31	N/A	N/A
	Life	1	0	1	0	0	0	.35	N/A	N/A
	Physical	1	0	1	0	0	0	.45	N/A	N/A
	Critiquing	2	0	2	0	0	0	.33	.03	.33
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	1	0	1	0	0	0	.45	N/A	N/A

Table 6.1.2: Grade 5 Item Discrimination Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Discrimination (*rpb*)					Descriptive Statistics		
			[0, .2)	[.2, .3)	[.3, .4)	[.4, .5)	[.5, 1]	Average	S.D.	Median
MC	NJSLA–S	14	1	4	4	5	0	.35	.10	.39
	Earth and Space	5	1	1	2	1	0	.32	.13	.36
	Life	5	0	2	2	1	0	.35	.08	.40
	Physical	4	0	1	0	3	0	.39	.11	.43
	Critiquing	1	0	1	0	0	0	.24	N/A	N/A
	Investigating	9	1	2	1	5	0	.36	.12	.41
	Sensemaking	4	0	1	3	0	0	.35	.06	.38
TE	NJSLA–S	35	1	7	6	9	12	.41	.13	.43
	Earth and Space	12	0	3	5	2	2	.36	.10	.33
	Life	11	1	2	1	4	3	.41	.14	.40
	Physical	12	0	2	0	3	7	.46	.12	.52
	Critiquing	9	1	3	2	1	2	.33	.14	.30
	Investigating	11	0	2	0	5	4	.45	.12	.46
	Sensemaking	15	0	2	4	3	6	.43	.11	.43
CR	NJSLA–S	3	0	0	0	1	2	.62	.14	.64
	Earth and Space	1	0	0	0	0	1	.64	N/A	N/A
	Life	1	0	0	0	0	1	.74	N/A	N/A
	Physical	1	0	0	0	1	0	.47	N/A	N/A
	Critiquing	2	0	0	0	0	2	.69	.07	.69
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	1	0	0	0	1	0	.47	N/A	N/A

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	14	1	11	2	0	0	.41	.12	.42
	Earth and Space	4	0	4	0	0	0	.41	.09	.40
	Life	6	0	6	0	0	0	.38	.07	.39
	Physical	4	1	1	2	0	0	.47	.20	.51
	Critiquing	5	0	4	1	0	0	.48	.11	.46
	Investigating	7	1	5	1	0	0	.37	.13	.34
	Sensemaking	2	0	2	0	0	0	.38	.08	.38
TE	NJSLA–S	47	14	29	4	0	0	.32	.10	.29
	Earth and Space	17	5	11	1	0	0	.31	.09	.28
	Life	15	6	8	1	0	0	.29	.11	.27
	Physical	15	3	10	2	0	0	.35	.11	.32
	Critiquing	14	4	8	2	0	0	.34	.12	.33
	Investigating	13	5	7	1	0	0	.30	.11	.29
	Sensemaking	20	5	14	1	0	0	.31	.09	.28
CR	NJSLA–S	3	1	2	0	0	0	.35	.12	.35
	Earth and Space	1	0	1	0	0	0	.38	N/A	N/A
	Life	1	0	1	0	0	0	.45	N/A	N/A
	Physical	1	1	0	0	0	0	.21	N/A	N/A
	Critiquing	3	1	2	0	0	0	.35	.12	.35
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	0	0	0	0	0	0	N/A	N/A	N/A

Table 6.1.4: Grade 8 Item Discrimination Distribution and Summary Statistics

Item Type	Domain/ Practice	Item Counts	Distribution of Item Discrimination (*rpb*)					Descriptive Statistics		
			[0, .2)	[.2, .3)	[.3, .4)	[.4, .5)	[.5, 1]	Average	S.D.	Median
MC	NJSLA–S	14	3	3	6	1	1	.30	.12	.32
	Earth and Space	4	1	0	3	0	0	.28	.08	.32
	Life	6	1	2	2	0	1	.32	.13	.30
	Physical	4	1	1	1	1	0	.28	.13	.28
	Critiquing	5	0	2	3	0	0	.29	.07	.31
	Investigating	7	3	1	1	1	1	.29	.15	.26
	Sensemaking	2	0	0	2	0	0	.34	.00	.34
TE	NJSLA–S	47	3	7	15	15	7	.39	.13	.39
	Earth and Space	17	1	2	6	4	4	.40	.12	.38
	Life	15	2	2	3	6	2	.38	.14	.45
	Physical	15	0	3	6	5	1	.39	.09	.39
	Critiquing	14	0	3	1	8	2	.41	.10	.44
	Investigating	13	1	1	6	3	2	.37	.12	.39
	Sensemaking	20	2	3	8	4	3	.39	.13	.39
CR	NJSLA–S	3	0	0	0	0	3	.63	.32	.62
	Earth and Space	1	0	0	0	0	1	.62	N/A	N/A
	Life	1	0	0	0	0	1	.65	N/A	N/A
	Physical	1	0	0	0	0	1	.68	N/A	N/A
	Critiquing	3	0	0	0	0	3	.63	.05	.62
	Investigating	0	0	0	0	0	0	N/A	N/A	N/A
	Sensemaking	0	0	0	0	0	0	N/A	N/A	N/A

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	29	1	20	8	0	0	.43	.11	.44
	Earth and Space	9	0	7	2	0	0	.43	.10	.42
	Life	11	1	7	3	0	0	.43	.13	.44
	Physical	9	0	6	3	0	0	.44	.10	.47
	Critiquing	6	1	5	0	0	0	.33	.09	.31
	Investigating	14	0	9	5	0	0	.45	.11	.46
	Sensemaking	9	0	6	3	0	0	.46	.09	.45
TE	NJSLA–S	38	12	25	1	0	0	.31	.11	.29
	Earth and Space	12	3	8	1	0	0	.33	.14	.30
	Life	10	4	6	0	0	0	.28	.09	.28
	Physical	16	5	11	0	0	0	.32	.10	.29
	Critiquing	14	3	11	0	0	0	.32	.09	.32
	Investigating	11	3	7	1	0	0	.32	.15	.29
	Sensemaking	13	6	7	0	0	0	.30	.11	.29
CR	NJSLA–S	3	1	1	1	0	0	.40	.19	.49
	Earth and Space	1	0	0	1	0	0	.53	N/A	N/A
	Life	1	0	1	0	0	0	.49	N/A	N/A
	Physical	1	1	0	0	0	0	.19	N/A	N/A
	Critiquing	2	1	0	1	0	0	.36	.24	.36
	Investigating	1	0	1	0	0	0	.49	N/A	N/A
	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*)					Descriptive Statistics		
			[0, .2)	[.2, .3)	[.3, .4)	[.4, .5)	[.5, 1]	Average	S.D.	Median
MC	NJSLA–S	29	2	10	7	9	1	.34	.10	.36
	Earth and Space	9	1	6	0	2	0	.29	.11	.27
	Life	11	0	3	5	3	0	.36	.08	.37
	Physical	9	1	1	2	4	1	.37	.10	.41
	Critiquing	6	1	3	1	1	0	.28	.11	.26
	Investigating	14	0	3	4	7	0	.38	.07	.40
	Sensemaking	9	1	4	2	1	1	.32	.12	.29
TE	NJSLA–S	38	1	4	13	12	8	.41	.10	.40
	Earth and Space	12	0	2	4	5	1	.38	.08	.39
	Life	10	0	1	4	3	2	.42	.10	.40
	Physical	16	1	1	5	4	5	.43	.12	.41
	Critiquing	14	0	1	5	2	6	.45	.12	.49
	Investigating	11	1	2	4	4	0	.36	.09	.37
	Sensemaking	13	0	1	4	6	2	.41	.08	.40
CR	NJSLA–S	3	0	0	0	0	3	.61	.03	.60
	Earth and Space	1	0	0	0	0	1	.60	N/A	N/A
	Life	1	0	0	0	0	1	.64	N/A	N/A
	Physical	1	0	0	0	0	1	.60	N/A	N/A
	Critiquing	2	0	0	0	0	2	.60	.00	.60
	Investigating	1	0	0	0	0	1	.64	N/A	N/A
	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	6 MC, 10 TE, 1 CR	45
	3	5 MC, 11 TE, 1 CR	45
8	1	6 MC, 15 TE, 1 CR	45
	2	5 MC, 15 TE, 1 CR	45
	3	3 MC, 17 TE, 1 CR	45
11	1	7 MC, 15 TE, 1 CR	60
	2	9 MC, 14 TE, 1 CR	60
	3	13 MC, 9 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	2.62
	2	15	1.24
	3	16	0.45
8	1	21	2.66
	2	20	1.20
	3	20	1.28
11	1	23	1.26
	2	23	0.67
	3	22	1.41

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 detection 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.

Table 6.1.10, Table 6.1.11, and Table 6.1.12 show the DIF classifications for all eight comparison groups for grades 5, 8, and 11, respectively. The results of the operational DIF analysis were positive except for a small number of items classified as “C” for the ML-No/ML-Yes comparison, including one TE item at grade 8 and two CR items at grade 11. For all other comparisons, zero items across all grade levels were classified as “C.” Moreover, each grade level, comparison group, and item type contained minimal classifications of “B” items. All items were classified as “A” for CBT/TTS DIF at grades 5, 8, and 11. The “C” DIF items will be reinvestigated when more test data become available.

Table 6.1.10: Grade 5 DIF Classification by Item Type

Grade	Group	Item Type	A	B	C
5	Male/Female	MC	14	0	0
	Male/Female	TE	34	1	0
	Male/Female	CR	3	0	0
	Male/Female	Total	51	1	0
5	White/Black	MC	14	0	0
	White/Black	TE	35	0	0
	White/Black	CR	3	0	0
	White/Black	Total	52	0	0
5	White/Hispanic	MC	14	0	0
	White/Hispanic	TE	35	0	0
	White/Hispanic	CR	3	0	0
	White/Hispanic	Total	52	0	0
5	White/Asian	MC	14	0	0
	White/Asian	TE	35	0	0
	White/Asian	CR	3	0	0
	White/Asian	Total	52	0	0
5	ML–No/ML–Yes	MC	14	0	0
	ML–No/ML–Yes	TE	33	2	0
	ML–No/ML–Yes	CR	2	1	0
	ML–No/ML–Yes	Total	49	3	0
5	SWD–No/SWD–Yes	MC	14	0	0
	SWD–No/SWD–Yes	TE	35	0	0
	SWD–No/SWD–Yes	CR	3	0	0
	SWD–No/SWD–Yes	Total	52	0	0
5	EconDis–No/EconDis–Yes	MC	14	0	0
	EconDis–No/EconDis–Yes	TE	35	0	0
	EconDis–No/EconDis–Yes	CR	3	0	0
	EconDis–No/EconDis–Yes	Total	52	0	0
5	CBT/TTS	MC	14	0	0
	CBT/TTS	TE	35	0	0
	CBT/TTS	CR	3	0	0
	CBT/TTS	Total	52	0	0

Table 6.1.11: Grade 8 DIF Classification by Item Type

Grade	Group	Item Type	A	B	C
8	Male/Female	MC	14	0	0
	Male/Female	TE	47	0	0
	Male/Female	CR	3	0	0
	Male/Female	Total	64	0	0
8	White/Black	MC	14	0	0
	White/Black	TE	47	0	0
	White/Black	CR	3	0	0
	White/Black	Total	64	0	0
8	White/Hispanic	MC	14	0	0
	White/Hispanic	TE	45	2	0
	White/Hispanic	CR	3	0	0
	White/Hispanic	Total	62	2	0
8	White/Asian	MC	14	0	0
	White/Asian	TE	47	0	0
	White/Asian	CR	3	0	0
	White/Asian	Total	64	0	0
8	ML–No/ML–Yes	MC	11	3	0
	ML–No/ML–Yes	TE	44	2	1
	ML–No/ML–Yes	CR	3	0	0
	ML–No/ML–Yes	Total	58	5	1
8	SWD–No/SWD–Yes	MC	14	0	0
	SWD–No/SWD–Yes	TE	47	0	0
	SWD–No/SWD–Yes	CR	3	0	0
	SWD–No/SWD–Yes	Total	64	0	0
8	EconDis–No/EconDis–Yes	MC	14	0	0
	EconDis–No/EconDis–Yes	TE	47	0	0
	EconDis–No/EconDis–Yes	CR	3	0	0
	EconDis–No/EconDis–Yes	Total	64	0	0
8	CBT/TTS	MC	14	0	0
	CBT/TTS	TE	47	0	0
	CBT/TTS	CR	3	0	0
	CBT/TTS	Total	64	0	0

Table 6.1.12: Grade 11 DIF Classification by Item Type

Grade	Group	Item Type	A	B	C
11	Male/Female	MC	29	0	0
	Male/Female	TE	37	1	0
	Male/Female	CR	2	1	0
	Male/Female	Total	68	2	0
11	White/Black	MC	29	0	0
	White/Black	TE	38	0	0
	White/Black	CR	3	0	0
	White/Black	Total	70	0	0
11	White/Hispanic	MC	29	0	0
	White/Hispanic	TE	38	0	0
	White/Hispanic	CR	3	0	0
	White/Hispanic	Total	70	0	0
11	White/Asian	MC	29	0	0
	White/Asian	TE	38	0	0
	White/Asian	CR	3	0	0
	White/Asian	Total	70	0	0
11	ML–No/ML–Yes	MC	25	4	0
	ML–No/ML–Yes	TE	35	3	0
	ML–No/ML–Yes	CR	1	0	2
	ML–No/ML–Yes	Total	61	7	2
11	SWD–No/SWD–Yes	MC	29	0	0
	SWD–No/SWD–Yes	TE	38	0	0
	SWD–No/SWD–Yes	CR	3	0	0
	SWD–No/SWD–Yes	Total	70	0	0
11	EconDis–No/EconDis–Yes	MC	29	0	0
	EconDis–No/EconDis–Yes	TE	38	0	0
	EconDis–No/EconDis–Yes	CR	3	0	0
	EconDis–No/EconDis–Yes	Total	70	0	0
11	CBT/TTS	MC	29	0	0
	CBT/TTS	TE	38	0	0
	CBT/TTS	CR	3	0	0
	CBT/TTS	Total	70	0	0

6.2 Item Response Theory

The grade-specific NJSLA–S student ability estimates and subsequent scale scores are calibrated via item response theory (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 classical test theory (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; Masters, 1982). 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 does 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- or 0–4-point CR items. Overall, the results of the 2024 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 domains and practices correlated with the total NJSLA–S test score at .90 or above. The lowest correlation between clusters was .77. 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	.91	1.00	-	-
	Life	.94	.78	1.00	-
	Physical	.93	.77	.82	1.00
8	Earth and Space	.94	1.00	-	-
	Life	.93	.80	1.00	-
	Physical	.92	.80	.78	1.00
11	Earth and Space	.92	1.00	-	-
	Life	.94	.81	1.00	-
	Physical	.95	.80	.84	1.00

Table 6.2.2: Correlation Matrix for Practices

Grade	Practice	NJSLA–S	Critiquing	Investigation	Sensemaking
5	Critiquing	.92	1.00	-	-
	Investigating	.93	.79	1.00	-
	Sensemaking	.94	.80	.81	1.00
8	Critiquing	.95	1.00	-	-
	Investigating	.90	.79	1.00	-
	Sensemaking	.93	.82	.78	1.00
11	Critiquing	.95	1.00	-	-
	Investigating	.95	.85	1.00	-
	Sensemaking	.92	.82	.82	1.00

6.2.1.2.2 Principal Component Analysis. Principal Components 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 range from approximately 11–13. Each grade's 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.

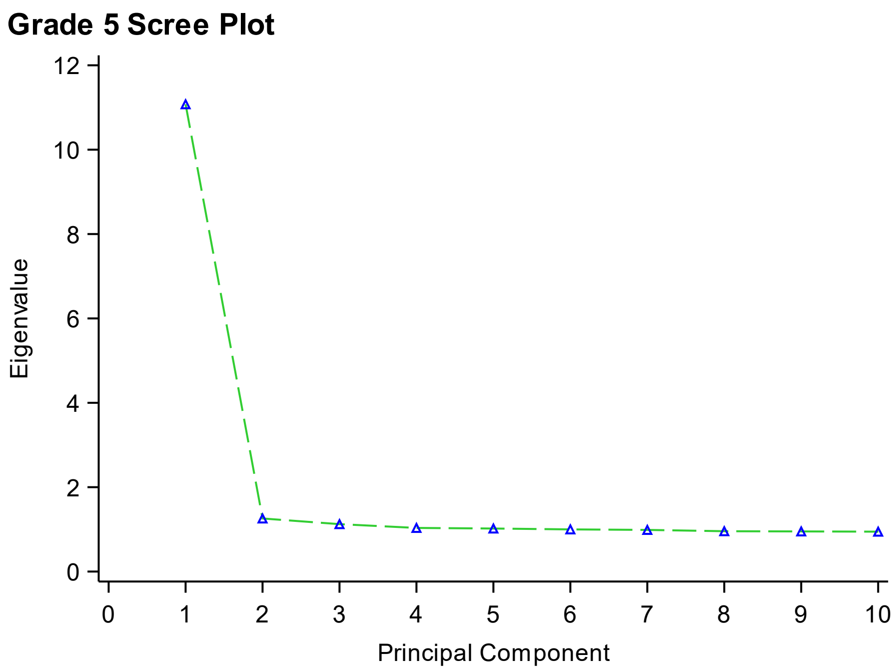


Figure 6.2.1. Grade 5 Scree Plot

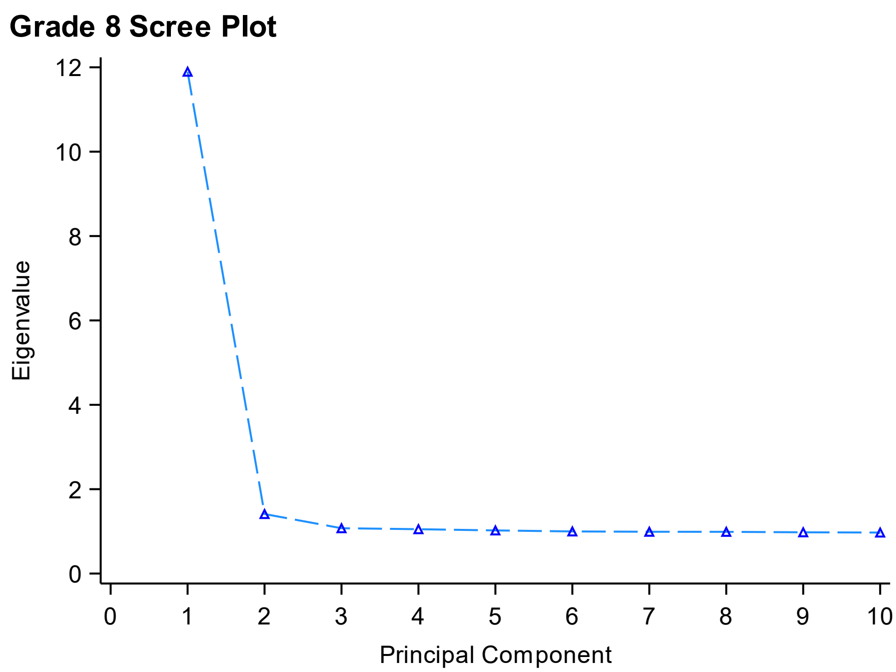


Figure 6.2.2. Grade 8 Scree Plot

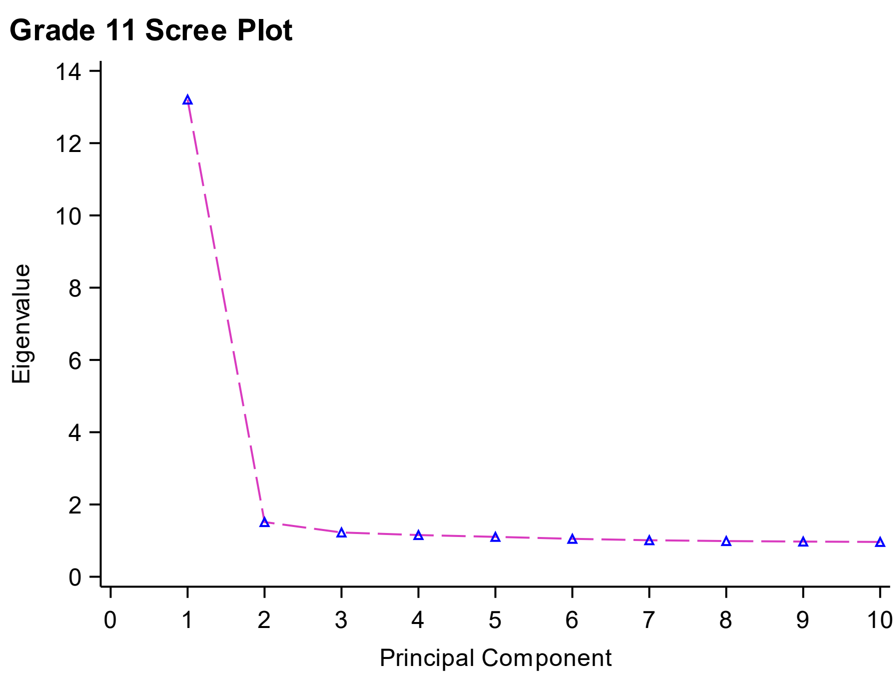


Figure 6.2.3. Grade 11 Scree Plot

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 2024 NJSLA–S test data via Winsteps 3.92.1 (Linacre, 2016); Appendix H of this technical report provides the resulting item fix 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 outfit.Rasch 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 6.2.4 provides a summary of item infit and outfit statistics at each grade level. One grade 5, zero grade 8, and zero grade 11 items were flagged for problematic infit statistics. Slightly more items were flagged based on the outfit statistics, with five, six, and three items being flagged for grades 5, 8, and 11, respectively. However, problematic outfit statistics are less of a threat to the validity of test score interpretations than problematic infit statistics. Thus, while there is clearly room for improving the item outfit, the infit and outfit statistics provide reasonable evidence that the assumptions of the PCM have been met.

Table 6.2.4: Summary of Item Infit and Outfit Statistics

Grade	Fit Statistic	Average	Min	Max	Outside 0.7 to 1.3	% Flagged
5	Infit	1.00	0.77	1.33	1 out of 52	1.9
	Outfit	1.01	0.69	1.61	5 out of 52	9.6
8	Infit	1.00	0.78	1.30	0 out of 64	0.0
	Outfit	1.01	0.67	1.45	6 out of 64	9.4
11	Infit	1.00	0.78	1.27	0 out of 70	0.0
	Outfit	1.02	0.72	1.39	3 out of 70	4.3

6.2.2.2 Rasch discrimination. The 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 6.2.5 provides a summary of Rasch discrimination statistics at each grade level. The Rasch discrimination values were good across each grade. There were three (5.8%), three (4.7%), and two (2.9%) items flagged for having values outside the 0.5 to 1.5 threshold for grades 5, 8, and 11, respectively, with Rasch discrimination values near the thresholds. The Rasch discrimination analysis results provide evidence that the PCM assumptions have been met for NJSLA–S.

Table 6.2.5: Summary of Rasch Discrimination Statistics

Grade	Fit Statistic	Average	Min	Max	Outside 0.5 to 1.5	% Flagged
5	Discrimination	1.01	0.54	1.61	3 out of 52	5.8
8	Discrimination	1.00	0.34	1.56	3 out of 64	4.7
11	Discrimination	0.99	0.30	1.62	2 out of 70	2.9

6.2.2.3 Rasch lower asymptote. The 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 6.2.6 provides a summary of the lower asymptote statistics at each grade level. Each grade level saw only a few items flagged for having a lower asymptote value outside of the .1 threshold including one item in grade 5, two items in grade 8, and four items in grade 11. Three of these items were also flagged for at least one other statistic (i.e., infit, outfit, or discrimination). Unsurprisingly, these items had low item-total correlations. Nevertheless, the Rasch lower asymptote statistics provided evidence that the PCM assumptions have been satisfied, as few items displayed lower asymptote values outside the acceptable threshold.

Table 6.2.6: Summary of Rasch Lower Asymptote Statistics

Grade	Fit Statistic	Average	Min	Max	Greater Than .1	% Flagged
5	Lower Asymptote	0.02	0.00	0.16	1 out of 52	1.9
8	Lower Asymptote	0.02	0.00	0.15	2 out of 64	3.1
11	Lower Asymptote	0.02	0.00	0.15	4 out of 70	5.7

6.2.2.4 Rasch person infit and outfit. PCM 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 have normally missed.
- 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 6.2.7 and 6.2.8 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 6.2.9 and 6.2.10 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 6.2.7, 7.43% of grade 5 students, 2.85% of grade 8 students, and 5.07% 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 their grade level. As shown in Table 6.2.9, the grade 5 PBT forms flagged 14.08% of students for person infit. However, it should be noted that there were only 71 students taking the grade 5 PBT form.

Overall, there were relatively more students flagged for aberrant person outfit statistics than for person infit statistics. As shown in Table 6.2.8, 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, which themselves 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. It is likely that the reason for the large percentages of person-outfit flags is that while these students tended to be lower performing, there were some items that they were able to unexpectedly answer correctly. Moreover, because the students that were flagged were low 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 6.2.7: 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	96,463	168.84	1.02	0.56	3.12	7,164	7.43	181.02
	Male	48,976	169.84	1.02	0.56	3.11	3,580	7.31	184.34
	Female	47,481	167.82	1.02	0.60	3.12	3,583	7.55	177.68
	Am. Indian	188	171.68	1.02	0.71	1.67	13	6.91	184.23
	Asian	10,544	201.93	1.03	0.62	2.73	918	8.71	204.12
	Black	13,482	149.17	1.01	0.59	3.12	897	6.65	169.84
	Hispanic	32,962	151.83	1.02	0.56	3.11	2,226	6.75	166.01
	Pacific Islander	175	174.79	1.03	0.73	1.74	15	8.57	171.60
	White	35,712	181.13	1.02	0.61	2.87	2,839	7.95	188.14
	ML–Yes	10,355	129.91	1.02	0.61	3.11	482	4.65	146.02
	ML–No	86,107	173.53	1.02	0.56	3.12	6,682	7.76	183.54
	EconDis–Yes	37,681	149.18	1.02	0.56	3.12	2,436	6.46	163.90
	EconDis–No	58,782	181.45	1.02	0.60	2.87	4,728	8.04	189.84
	SWD–Yes	20,707	148.43	1.01	0.62	2.87	1,255	6.06	170.17
	SWD–No	75,756	174.42	1.02	0.56	3.12	5,909	7.80	183.32
8	NJSLA–S	99,685	164.27	1.00	0.62	2.45	2,839	2.85	177.52
	Male	51,093	164.73	1.00	0.64	2.44	1,509	2.95	183.78
	Female	48,533	163.77	0.99	0.62	2.45	1,328	2.74	170.43
	Am. Indian	172	160.88	0.99	0.74	1.56	3	1.74	215.33
	Asian	10,785	194.66	1.00	0.68	2.16	429	3.98	200.24
	Black	14,322	147.20	1.00	0.62	2.45	336	2.35	159.27
	Hispanic	33,328	149.95	1.00	0.66	2.44	787	2.36	160.30
	Pacific Islander	184	168.39	1.00	0.76	1.49	7	3.80	183.29
	White	37,959	173.98	0.99	0.63	2.18	1,196	3.15	185.18
	ML–Yes	8,735	134.37	1.01	0.66	2.10	141	1.61	142.19
	ML–No	90,945	167.14	0.99	0.62	2.45	2,698	2.97	179.36
	EconDis–Yes	36,703	148.33	1.00	0.62	2.45	882	2.40	159.45
	EconDis–No	62,964	173.57	0.99	0.63	2.44	1,957	3.11	185.66

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	20,435	147.04	1.00	0.63	2.03	440	2.15	161.20
	SWD–No	79,250	168.71	0.99	0.62	2.45	2,399	3.03	180.51
11	NJSLA–S	98,881	168.89	1.01	0.63	2.66	5,017	5.07	183.26
	Male	50,484	168.63	1.01	0.65	2.66	2,646	5.24	188.56
	Female	48,242	169.07	1.00	0.63	2.40	2,359	4.89	177.32
	Am. Indian	161	163.19	1.01	0.78	1.76	9	5.59	174.22
	Asian	10,214	211.20	1.02	0.66	2.62	708	6.93	214.64
	Black	13,917	146.88	1.01	0.66	2.32	705	5.07	162.81
	Hispanic	32,576	149.77	1.00	0.63	2.66	1,535	4.71	166.25
	Pacific Islander	230	179.57	1.01	0.75	1.67	17	7.39	181.06
	White	39,415	180.88	1.00	0.64	2.47	1,916	4.86	192.50
	ML–Yes	7,090	124.73	1.00	0.67	2.12	185	2.61	135.18
	ML–No	91,785	172.30	1.01	0.63	2.66	4,832	5.26	185.10
	EconDis–Yes	32,659	148.16	1.00	0.63	2.31	1,611	4.93	164.66
	EconDis–No	66,216	179.11	1.01	0.65	2.66	3,406	5.14	192.06
	SWD–Yes	20,089	148.31	1.00	0.67	2.47	810	4.03	166.20
	SWD–No	78,792	174.13	1.01	0.63	2.66	4,207	5.34	186.55

Table 6.2.8: Summary of Person Outfit Statistics by Demographic Group

Grade	Group	N Students	Mean Scale Score	Person Outfit Average	Person Outfit Min	Person Outfit Max	N Flagged	% Flagged	Flagged Mean Scale Score
5	NJSLA–S	96,463	168.84	1.01	0.29	3.37	5,433	5.63	132.46
	Male	48,976	169.84	1.01	0.32	3.37	2,883	5.89	132.98
	Female	47,481	167.82	1.01	0.29	3.37	2,549	5.37	131.89
	Am. Indian	188	171.68	1.04	0.65	1.76	15	7.98	147.33
	Asian	10,544	201.93	1.01	0.38	2.55	406	3.85	209.30
	Black	13,482	149.17	1.03	0.32	3.37	1,109	8.23	116.91
	Hispanic	32,962	151.83	1.02	0.29	3.18	2,513	7.62	116.99
	Pacific Islander	175	174.79	1.05	0.54	2.00	15	8.57	133.60
	White	35,712	181.13	1.00	0.43	3.37	1,209	3.39	150.10
	ML–Yes	10,355	129.91	1.06	0.32	3.18	1,377	13.30	111.86
	ML–No	86,107	173.53	1.01	0.29	3.37	4,055	4.71	139.47
	EconDis–Yes	37,681	149.18	1.03	0.29	3.37	3,036	8.06	116.25
	EconDis–No	58,782	181.45	1.01	0.32	2.80	2,397	4.08	153.00
	SWD–Yes	20,707	148.43	1.03	0.32	3.37	1,916	9.25	116.57
	SWD–No	75,756	174.42	1.01	0.29	3.18	3,517	4.64	141.12
8	NJSLA–S	99,685	164.27	1.01	0.28	2.83	4,428	4.44	126.96
	Male	51,093	164.73	1.01	0.28	2.83	2,515	4.92	126.41
	Female	48,533	163.77	1.00	0.34	2.11	1,910	3.94	127.61
	Am. Indian	172	160.88	1.01	0.55	1.59	6	3.49	106.00
	Asian	10,785	194.66	1.01	0.56	2.83	253	2.35	192.98
	Black	14,322	147.20	1.03	0.38	2.55	1,014	7.08	117.64
	Hispanic	33,328	149.95	1.02	0.28	2.44	2,050	6.15	118.85
	Pacific Islander	184	168.39	1.01	0.73	1.66	5	2.72	124.60
	White	37,959	173.98	1.00	0.52	2.15	991	2.61	135.31
	ML–Yes	8,735	134.37	1.05	0.34	2.15	866	9.91	116.17
	ML–No	90,945	167.14	1.01	0.28	2.83	3,562	3.92	129.58
	EconDis–Yes	36,703	148.33	1.02	0.28	2.55	2,416	6.58	118.41
	EconDis–No	62,964	173.57	1.00	0.34	2.83	2,010	3.19	137.23

Grade	Group	N Students	Mean Scale Score	Person Outfit Average	Person Outfit Min	Person Outfit Max	N Flagged	% Flagged	Flagged Mean Scale Score
	SWD–Yes	20,435	147.04	1.03	0.38	2.41	1,649	8.07	118.11
	SWD–No	79,250	168.71	1.00	0.28	2.83	2,779	3.51	132.21
11	NJSLA–S	98,881	168.89	1.02	0.45	2.71	6,123	6.19	135.19
	Male	50,484	168.63	1.03	0.45	2.71	3,560	7.05	137.08
	Female	48,242	169.07	1.01	0.47	2.60	2,560	5.31	132.50
	Am. Indian	161	163.19	1.03	0.65	1.82	12	7.45	134.08
	Asian	10,214	211.20	1.02	0.45	2.71	428	4.19	220.87
	Black	13,917	146.88	1.03	0.48	2.70	1,250	8.98	118.61
	Hispanic	32,576	149.77	1.03	0.45	2.60	2,575	7.90	118.72
	Pacific Islander	230	179.57	1.01	0.66	1.90	12	5.22	131.92
	White	39,415	180.88	1.01	0.47	2.43	1,729	4.39	149.90
	ML–Yes	7,090	124.73	1.06	0.48	2.35	906	12.78	110.26
	ML–No	91,785	172.30	1.02	0.45	2.71	5,216	5.68	139.52
	EconDis–Yes	32,659	148.16	1.03	0.45	2.55	2,679	8.20	117.95
	EconDis–No	66,216	179.11	1.02	0.45	2.71	3,443	5.20	148.60
	SWD–Yes	20,089	148.31	1.03	0.47	2.55	1,768	8.80	118.36
	SWD–No	78,792	174.13	1.02	0.45	2.71	4,355	5.53	142.02

Table 6.2.9: 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	72,447	175.63	1.02	0.56	3.12	5,696	7.86	184.59
	PBT	71	159.90	1.08	0.72	2.25	10	14.08	153.40
	TTS	20,677	151.54	1.02	0.59	3.11	1,307	6.32	169.84
	SP	1,974	127.53	1.02	0.61	2.43	93	4.71	146.13
	SP TTS	1,089	128.22	1.02	0.62	1.91	44	4.04	139.86
	HR	136	134.17	1.05	0.72	1.80	11	8.09	149.45
8	CBT	80,116	168.58	0.99	0.62	2.45	2,384	2.98	180.75
	PBT	63	142.79	1.00	0.79	1.38	1	1.59	186.00
	TTS	16,308	148.72	1.00	0.65	2.03	403	2.47	162.24
	SP	2,348	136.08	1.01	0.69	1.83	40	1.70	147.95
	SP TTS	755	135.87	1.01	0.71	1.58	10	1.32	143.50
	HR	68	138.46	1.00	0.77	1.32	1	1.47	139.00
11	CBT	87,127	171.94	1.01	0.63	2.62	4,546	5.22	184.48
	PBT	135	165.46	1.02	0.76	1.61	5	3.70	139.40
	TTS	8,679	151.88	1.00	0.67	2.66	398	4.59	176.14
	SP	2,232	126.47	0.99	0.67	1.89	47	2.11	138.94
	SP TTS	424	125.48	0.99	0.70	1.62	8	1.89	150.13
	HR	249	153.69	1.00	0.73	1.74	13	5.22	174.85

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 6.2.10: Summary of Person Outfit Statistics by Form

Grade	Form	N Students	Mean Scale Score	Person Outfit			N Flagged	% Flagged	Flagged Mean Scale Score
				Average	Min	Max			
5	CBT	72,447	175.63	1.01	0.29	3.37	3,190	4.40	142.89
	PBT	71	159.90	1.04	0.57	1.74	7	9.86	144.57
	TTS	20,677	151.54	1.03	0.32	3.37	1,778	8.60	119.14
	SP	1,974	127.53	1.07	0.32	2.62	294	14.89	110.99
	SP TTS	1,089	128.22	1.07	0.52	2.65	143	13.13	111.76
	HR	136	134.17	1.04	0.41	2.05	15	11.03	112.93
8	CBT	80,116	168.58	1.00	0.28	2.83	2,896	3.61	131.91
	PBT	63	142.79	1.05	0.71	1.69	6	9.52	114.83
	TTS	16,308	148.72	1.03	0.38	2.32	1,250	7.66	117.72
	SP	2,348	136.08	1.05	0.48	1.97	207	8.82	117.09
	SP TTS	755	135.87	1.05	0.66	2.02	57	7.55	117.37
	HR	68	138.46	1.08	0.67	1.63	8	11.76	117.38
11	CBT	87,127	171.94	1.02	0.45	2.71	5,045	5.79	138.34
	PBT	135	165.46	1.06	0.77	2.09	10	7.41	109.20
	TTS	8,679	151.88	1.03	0.45	2.55	746	8.60	124.35
	SP	2,232	126.47	1.06	0.48	2.15	247	11.07	111.19
	SP TTS	424	125.48	1.05	0.63	1.81	47	11.08	111.19
	HR	249	153.69	1.00	0.58	1.52	17	6.83	119.88

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

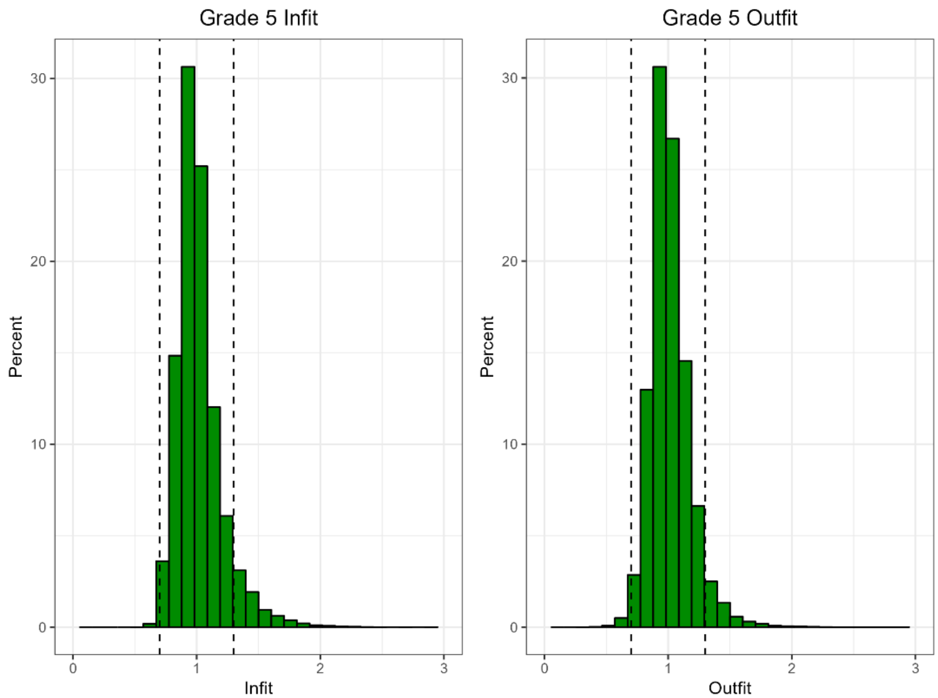


Figure 6.2.4. Grade 5 Person Infit and Outfit Distributions

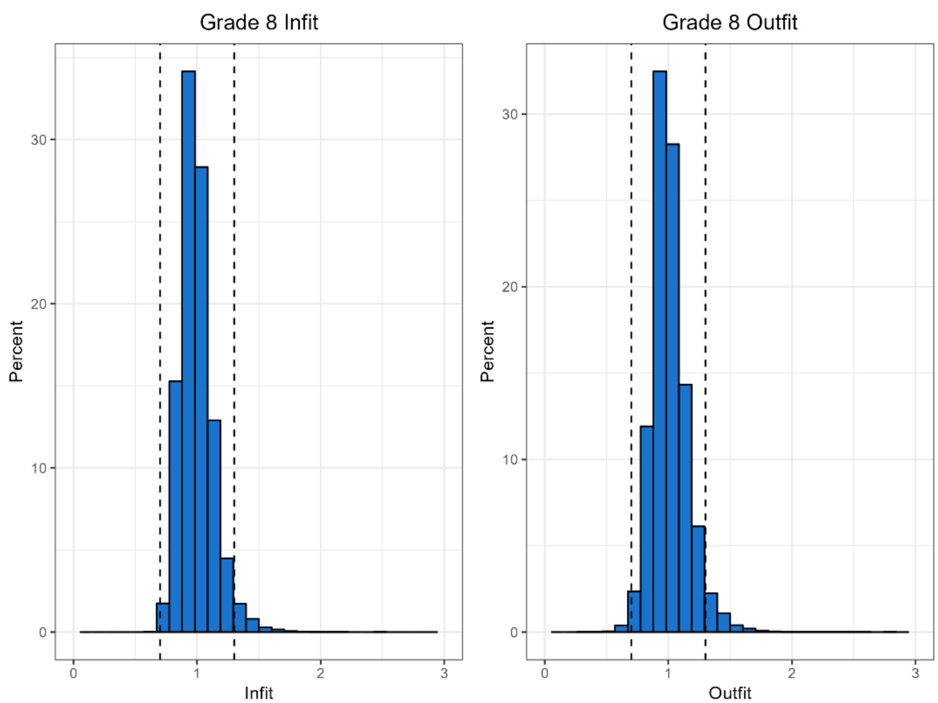


Figure 6.2.5. Grade 8 Person Infit and Outfit Distributions

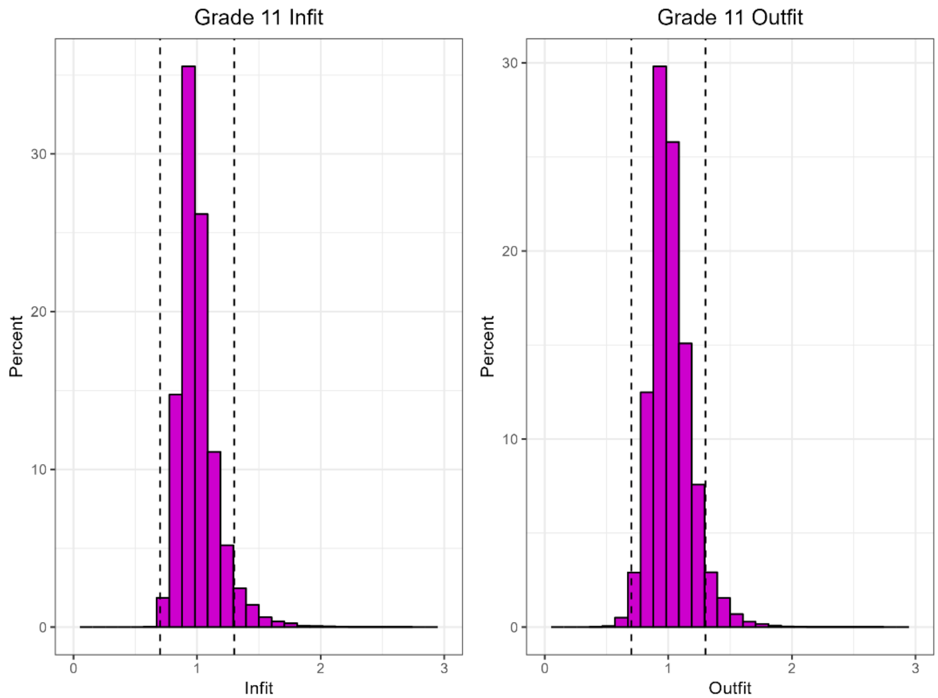


Figure 6.2.6. Grade 11 Person Infit and Outfit Distributions

6.2.2.5 Further investigation of person outfit statistics. Within 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. Notably, underfit is more troublesome than overfit because an excess of expected scores does not necessarily imply invalidity in the measurement process (Engelhard & Wind, 2018; Linacre, 2016). However, an excess of unexpected scores 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 6.2.11 presents the distributions of total fit and total misfit, which further breaks down into overfit and underfit by student cohort for each grade.

Table 6.2.11: Distribution of Person Fit for Grades 5, 8, and 11

Grade	Group	ML (%)	EconDis (%)	SWD (%)	ACC (%)
5	Total	10,355 (100%)	37,681 (100%)	20,707 (100%)	23,947 (100%)
	Fit	8,978 (86.7%)	34,645 (91.9%)	18,791 (90.7%)	21,710 (90.7%)
	Total Misfit	1,377 (13.3%)	3036 (8.1%)	1916 (9.3%)	2,237 (9.3%)
	Overfit	182 (1.8%)	448 (1.2%)	275 (1.3%)	316 (1.3%)
	Underfit	1,195 (11.5%)	2,588 (6.9%)	1,641 (7.9%)	1,921 (8.0%)
8	Total	8,735 (100%)	36,703 (100%)	20,435 (100%)	19,542 (100%)
	Fit	7,869 (90.1%)	34,287 (93.4%)	18,786 (91.9%)	18,014 (92.2%)
	Total Misfit	866 (9.9%)	2,416 (6.6%)	1,649 (8.1%)	1,528 (7.8%)
	Overfit	104 (1.2%)	354 (1.0%)	204 (1.0%)	187 (1.0%)
	Underfit	762 (8.7%)	2,062 (5.6%)	1,445 (7.1%)	1,341 (6.9%)
11	Total	7,090 (100%)	32,659 (100%)	20,089 (100%)	11,719 (100%)
	Fit	6,184 (87.2%)	29,980 (91.8%)	18,321 (91.2%)	10,652 (90.9%)
	Total Misfit	906 (12.8%)	2,679 (8.2%)	1,768 (8.8%)	1,067 (9.1%)
	Overfit	102 (1.4%)	380 (1.2%)	237 (1.2%)	133 (1.1%)
	Underfit	804 (11.3%)	2,299 (7.0%)	1,531 (7.6%)	934 (8.0%)

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

While the partial credit model provides summary statistics to identify overall person misfit (i.e., infit and outfit), other methods are required to identify which items were contributing to the unexpected student performance. To that end, this study 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., Students with Disabilities, Multilingual Learners, Economically Disadvantaged students, 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 6.2.12 exhibits the items flagged resulting from a single round of delta analysis. There were two, four, 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 6.2.12. 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 6.2.12, 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 negative 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 6.2.8) 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 6.2.12: *P-values* of Items Flagged by Delta Method

Flagged Item						ML		EconDis		SWD		ACC	
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	2205B006_06	ESS	CRI	TE	2.114	.05	.26	.10	.31	.11	.31	.10	.30
	2005M015_06	ESS	INV	MC	1.713	.16	.42	.19	.43	.18	.44	.18	.44
8	2208M016_04	LS	INV	TE	-0.993	.36	.05	.50	.06	.46	.06	.46	.06
	2108B006_03	ESS	INV	TE	1.232					.12	.22	.11	.21
	2008B007_09	PS	SEN	TE	-0.954					.49	.08	.47	.07
	2308M017_05	PS	CRI	MC	-1.265							.61	.21
11	2311M023_06	LS	CRI	TE	1.687	.06	.19						
	2311M026_05	PS	INV	TE	1.822	.08	.24	.11	.26				
	2111M004_02	LS	SEN	TE	2.137	.04	.31	.08	.32	.09	.33	.08	.33
	2111B002_11	LS	INV	CR	-0.096	.17	.02	.38	.07	.36	.07	.35	.09
	2211B006_09	ESS	CRI	CR	-0.242			.45	.08	.45	.10	.42	.10

Note. ML = Multilingual learners; EconDis = Students with economically disadvantaged status; SWD = Students with disabilities; ACC = Students taking an accommodated test form; 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 Person correlation of d_i and d_k over all test takers.

Table 6.2.13 summarizes Yen's Q3 statistics for the NJSLA–S test at each grade level. All pairwise combinations of items were checked, and they were flagged if their Q3 value was above .2 or below –.2 (Chen & Thissen, 1997). The results at all grades indicate that the assumption of local independence was met because very few combinations of items displayed Q3 values outside the acceptable threshold.

Table 6.2.13: Summary of Yen's Q3 Statistics

Grade	Average	Min	Max	Outside –.2 to .2	% Flagged
5	–0.02	–0.10	0.23	1 out of 1,326	0.08
8	–0.01	–0.11	0.13	0 out of 2,016	0.00
11	–0.01	–0.10	0.17	0 out of 2,415	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 hand-scored, constructed-response, 0–3- or 0–4-point items are presented in Figures 6.2.7 through 6.2.15 below. 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 6.2.14 shows the percentages of students receiving each score point for all nine CR items.

Table 6.2.14: Constructed-Response Point Distribution Percentages

Grade	Item	%0	%1	%2	%3	%4
5	CR Item 1	49.77	11.50	14.04	14.87	9.83
	CR Item 2	26.69	33.58	17.65	22.07	N/A
	CR Item 3	40.05	19.91	9.30	20.11	10.63
8	CR Item 1	29.40	25.99	24.48	20.13	N/A
	CR Item 2	49.19	31.23	10.82	3.94	4.81
	CR Item 3	28.31	22.92	24.88	16.97	6.92
11	CR Item 1	23.28	12.43	18.57	21.81	23.90
	CR Item 2	31.77	17.20	24.42	26.61	N/A
	CR Item 3	49.25	36.28	7.76	4.19	2.52

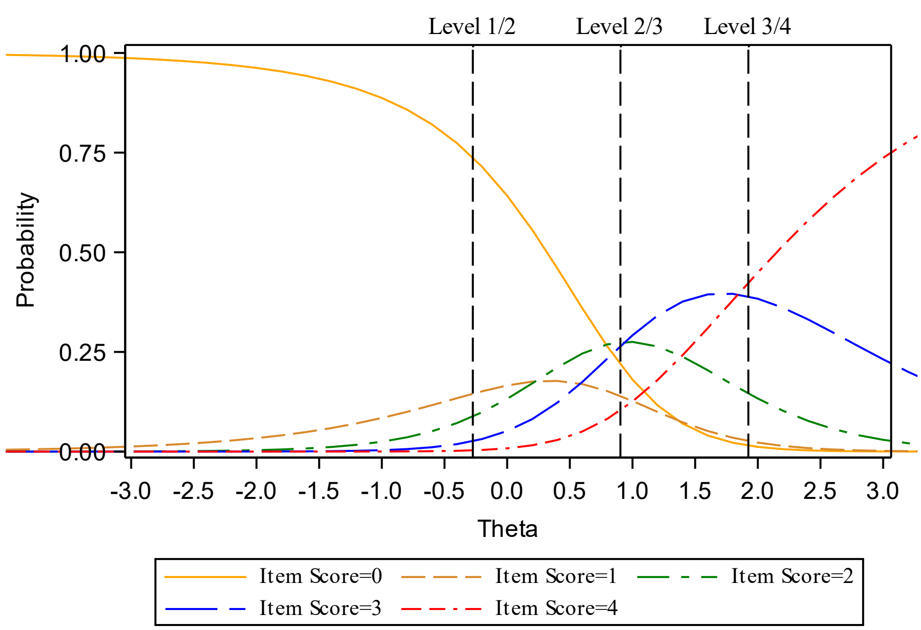


Figure 6.2.7. ICC Plot for Grade 5 Constructed-Response Item 1

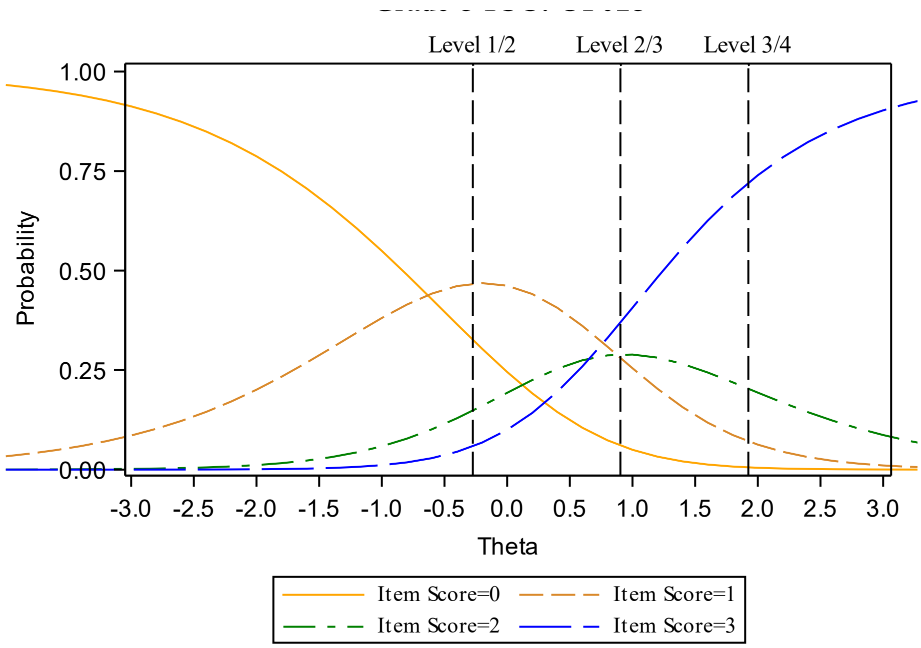


Figure 6.2.8. ICC Plot for Grade 5 Constructed-Response Item 2

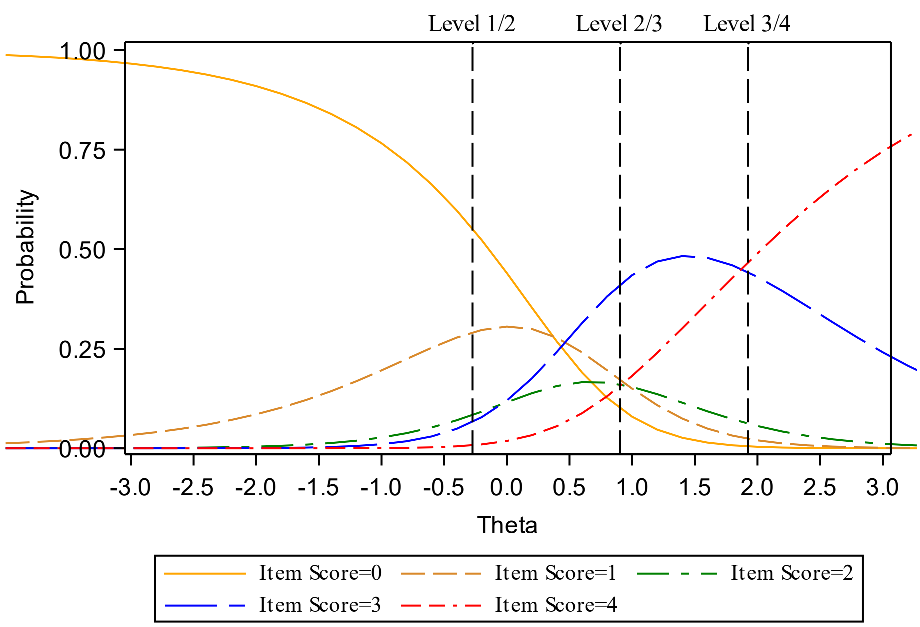


Figure 6.2.9. ICC Plot for Grade 5 Constructed-Response Item 3

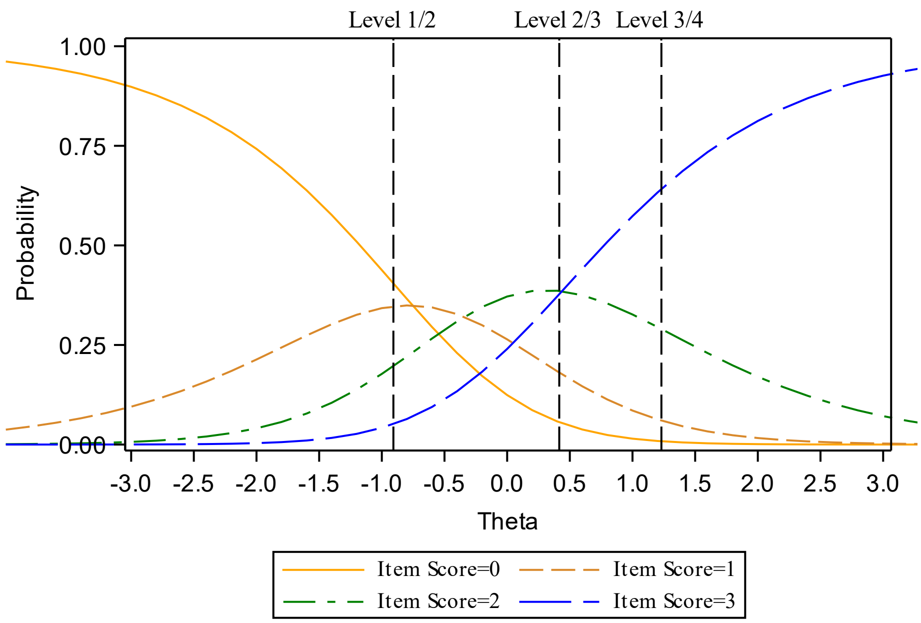


Figure 6.2.10. ICC Plot for Grade 8 Constructed-Response Item 1

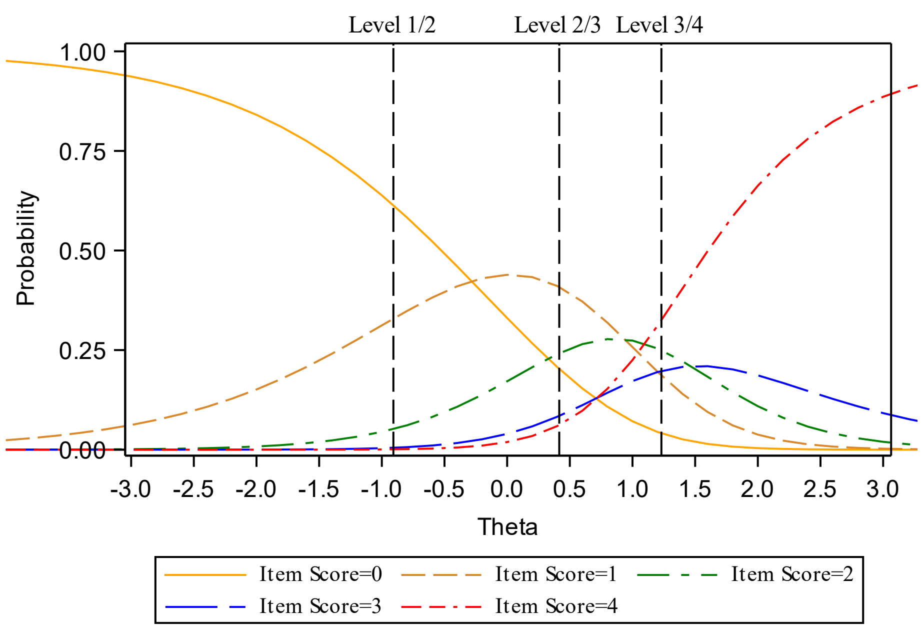


Figure 6.2.11. ICC Plot for Grade 8 Constructed-Response Item 2

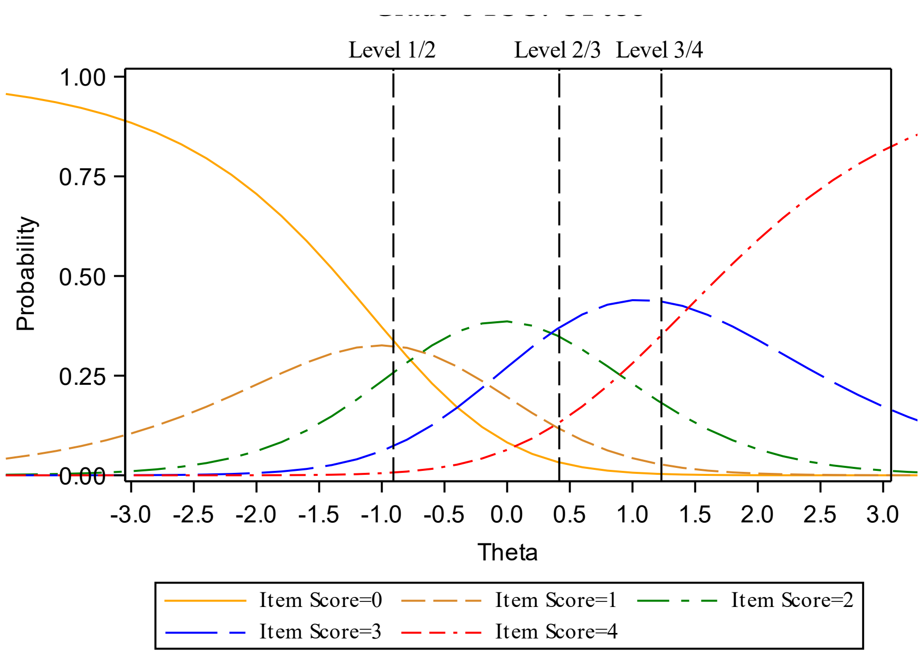


Figure 6.2.12. ICC Plot for Grade 8 Constructed-Response Item 3

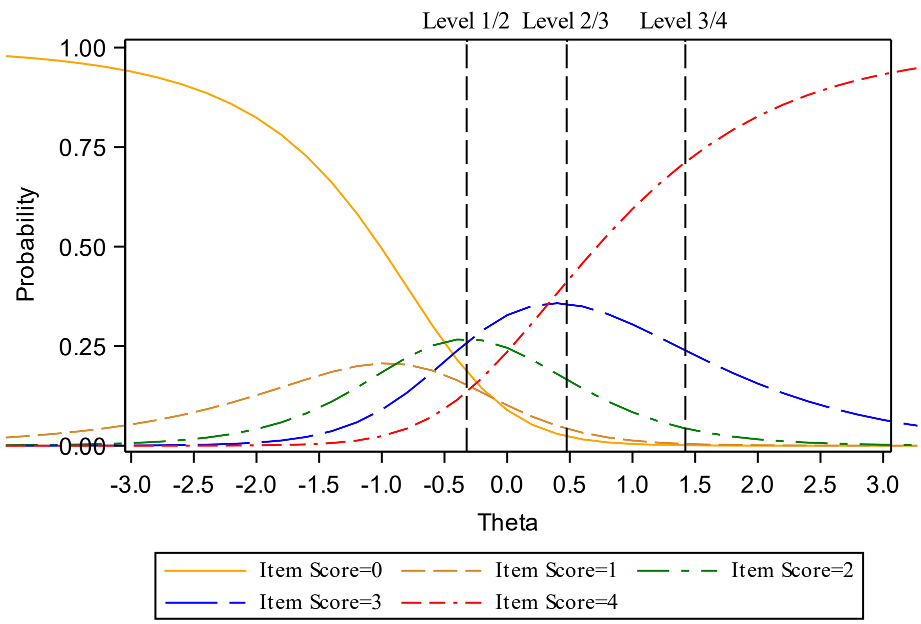


Figure 6.2.13. ICC Plot for Grade 11 Constructed-Response Item 1

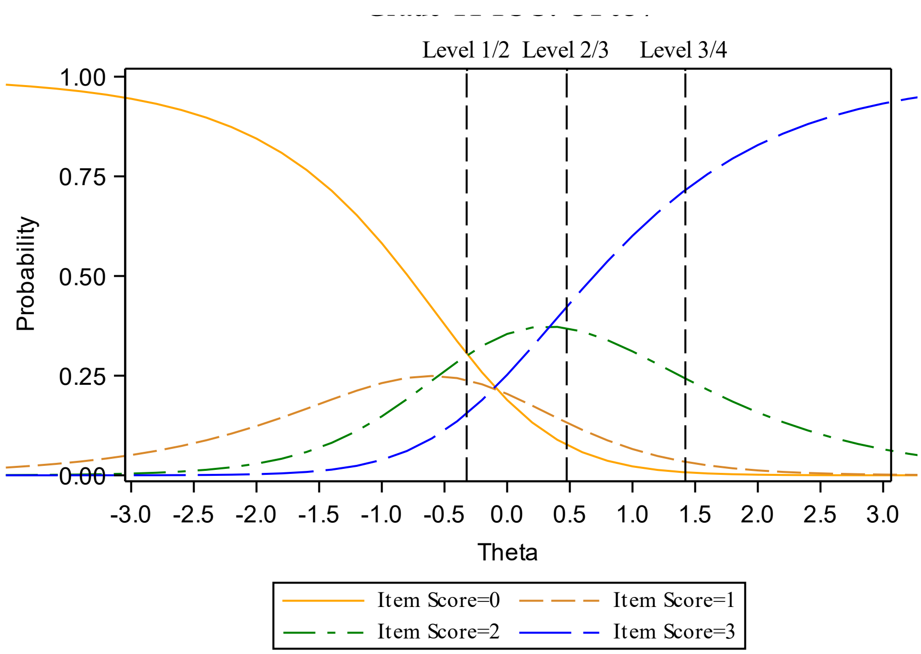


Figure 6.2.14. ICC Plot for Grade 11 Constructed-Response Item 2

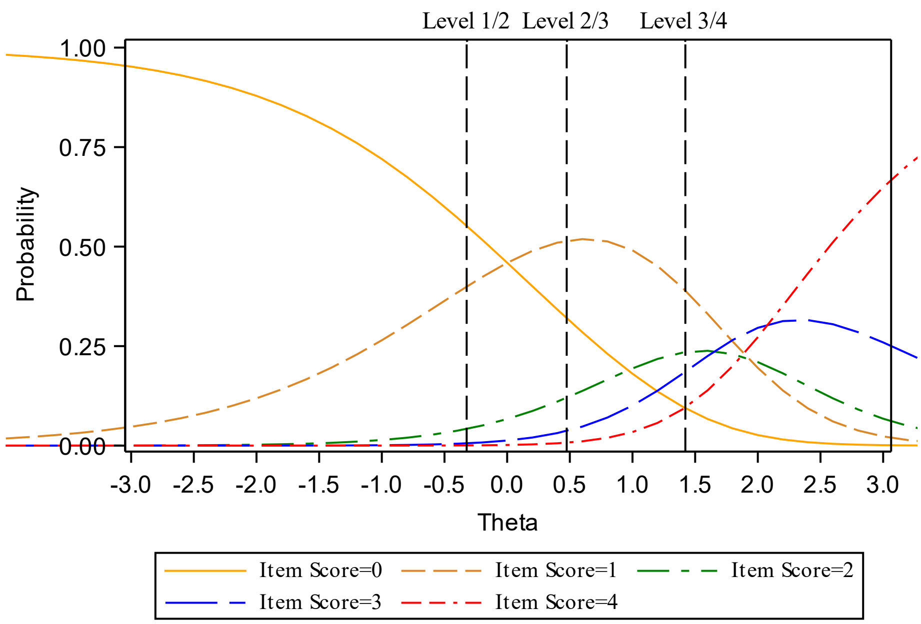


Figure 6.2.15. ICC Plot for Grade 11 Constructed-Response Item 3

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	72,447	175.63	42.29	100	300	28.26	39.56	24.51	7.68
	PBT	71	159.90	39.69	100	300	40.85	40.85	15.49	2.82
	TTS	20,677	151.54	40.35	100	300	53.02	31.38	12.50	3.11
	SP	1,974	127.53	24.49	100	243	81.76	17.02	1.17	0.05
	SP TTS	1,089	128.22	24.59	100	252	80.44	18.46	0.92	0.18
	HR	136	134.17	27.00	100	215	69.85	27.94	2.21	0.00
8	CBT	80,116	168.58	36.05	100	300	30.78	47.71	15.78	5.74
	PBT	63	142.79	29.80	100	223	58.73	38.10	3.17	0.00
	TTS	16,308	148.72	32.32	100	281	54.74	36.30	7.11	1.85
	SP	2,348	136.08	21.43	100	233	73.00	26.24	0.72	0.04
	SP TTS	755	135.87	19.15	100	225	71.66	27.95	0.40	0.00
	HR	68	138.46	28.79	100	228	73.53	22.06	4.41	0.00
11	CBT	87,127	171.94	50.39	100	300	42.45	27.45	20.75	9.36
	PBT	135	165.46	50.81	100	300	48.15	25.19	18.52	8.15
	TTS	8,679	151.88	46.50	100	300	61.24	21.94	11.65	5.17
	SP	2,232	126.47	24.08	100	244	87.19	11.78	1.03	0.00
	SP TTS	424	125.48	24.69	100	234	86.79	12.03	1.18	0.00
	HR	249	153.69	43.11	100	300	59.84	22.49	14.46	3.21

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 percentage of students classified as Below Expectations were relatively consistent across the content domains and the practices, with 51.56% for Physical Science to 54.70% for Life Science and 51.7% for Sensemaking to 55.07% for Investigating. At grade 8, the highest percentages of students classified as Below Expectations were observed for Life Science (67.82%) and Investigation (65.50%). At grade 11, the Below Expectations percentages varied from 50.92% for Earth and Space Science to 55.92% for Physical Science, while the percentage of students classified as Below Expectation were relatively consistent across the practices with 51.42% for Critiquing to 55.95% for Investigating. 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 partial credit model (PCM; Masters, 1982; 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 2024 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 2024 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 it is more than two standard deviation units of the PDs 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 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 2024 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:

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

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 is described in the following section. For the 2024 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 was 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 were taken to ensure the special equating and CBT forms were on the same scale.

1. Anchored item calibration. The inequivalent items were removed prior to the special equating calibrations, and the item parameters and steps of the accommodated test items were 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 assessments reports student proficiency in three content domains/disciplinary core ideas (DCI) including Earth and Space, Life, and Physical. The NJSLA–S also reports proficiency in three scientific 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 statistically significantly above or below the overall scale's Level 3 (proficient) theta cut score (denoted by θ^*). 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, and decision consistency measures, amongst others. The following sections detail evidence supporting the reliability of the NJSLA–S test scores and subscores.

8.1 Classical Test Theory Reliability Estimates

This section describes the Classical Test Theory (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:

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

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:

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

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:

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

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 .76 for the SP and SP TTS forms to .91 for the CBT form. At grade 8, reliability ranged from .67 for the SP TTS forms to .92 for the CBT form. The grade 11 alpha coefficients ranged from .73 for the SP form to .93 for the CBT form. At each grade level, the SP and SP TTS form test takers did poorly on the test. 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	72,447	24.99	12.51	0.91	3.73
	PBT	71	20.27	11.23	0.89	3.68
	TTS	20,677	18.14	11.45	0.91	3.47
	SP	1,974	11.44	6.16	0.76	2.99
	SP TTS	1,089	11.59	6.11	0.76	2.98
	HR	136	13.10	6.83	0.78	3.17
8	CBT	80,116	25.86	13.57	0.92	3.82
	PBT	63	16.79	9.96	0.88	3.49
	TTS	16,308	18.76	11.37	0.90	3.54
	SP	2,348	14.15	6.57	0.75	3.27
	SP TTS	755	13.92	5.64	0.67	3.24
	HR	68	15.37	9.88	0.89	3.27
11	CBT	87,127	29.57	15.13	0.93	4.01
	PBT	135	27.61	15.24	0.93	3.93
	TTS	8,679	23.65	13.77	0.92	3.86
	SP	2,232	16.13	6.66	0.73	3.47
	SP TTS	424	15.92	6.84	0.75	3.45
	HR	249	24.10	12.81	0.91	3.88

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 summarizes the coefficient alpha and SEM of raw scores of the six reporting categories by grade. In general, longer tests yield higher reliability coefficient estimates than shorter tests (Traub & Rowley, 1991). Thus, reporting categories such as Critiquing at grade 5, which had fewer items, tended to have lower reliability measures. For content, the Earth and Space Science subscore tended to have relatively low reliability at grades 5 and 11, while it had the highest reliability at grade 8. For practice, the lowest subscore reliability of .68 was for Critiquing at grade 5, which had the fewest items among the practice for grade 5. However, at grades 8 and 11, the Critiquing subscore had reliability higher than .80. Overall, the subscore reliability for the six reporting categories was deemed strong at each grade level.

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	52	14	35	3	60	0.92	3.68
	Earth and Space	18	5	12	1	21	0.75	2.23
	Life	17	5	11	1	20	0.78	2.16
	Physical	17	4	12	1	19	0.81	1.97
	Critiquing	12	1	9	2	18	0.68	2.29
	Investigating	20	9	11	0	20	0.82	1.91
	Sensemaking	20	4	15	1	22	0.81	2.15
8	NJSLA–S	64	14	47	3	72	0.92	3.77
	Earth and Space	22	4	17	1	25	0.81	2.25
	Life	22	6	15	1	24	0.79	2.18
	Physical	20	4	15	1	23	0.79	2.10
	Critiquing	22	5	14	3	30	0.83	2.58
	Investigating	20	7	13	0	20	0.74	1.9
	Sensemaking	22	2	20	0	22	0.81	1.96
11	NJSLA–S	70	29	38	3	78	0.93	4.00
	Earth and Space	22	9	12	1	25	0.76	2.44
	Life	22	11	10	1	24	0.82	2.21
	Physical	26	9	16	1	29	0.86	2.27
	Critiquing	22	6	14	2	28	0.82	2.47
	Investigating	26	14	11	1	28	0.82	2.42
	Sensemaking	22	9	13	0	22	0.79	2.01

Table 8.1.3 shows the coefficient alpha and SEMs 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 .79, 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} = .76). The pattern for grade 11 was the same as for grade 8. The coefficient alpha values for all groups at grade 11 were above .89 except for the multilingual learners (α_{ML-Yes} = .75).

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	96,463	23.07	12.63	0.92	3.68
	Male	48,976	23.41	13.01	0.92	3.66
	Female	47,481	22.71	12.21	0.91	3.69
	Am. Indian	188	23.97	13.53	0.93	3.69
	Asian	10,544	32.89	12.44	0.91	3.74
	Black	13,482	17.36	10.33	0.89	3.43
	Hispanic	32,962	18.09	10.55	0.89	3.49
	Pacific Islander	175	24.73	12.98	0.91	3.78
	White	35,712	26.61	12.12	0.90	3.77
	ML–Yes	10,355	12.05	6.71	0.79	3.04
	ML–No	86,107	24.39	12.52	0.91	3.72
	EconDis–Yes	37,681	17.32	10.14	0.88	3.45
	EconDis–No	58,782	26.75	12.69	0.91	3.76
	SWD–Yes	20,707	17.29	11.24	0.91	3.39
	SWD–No	75,755	24.65	12.53	0.91	3.73
8	NJSLA–S	99,685	24.32	13.45	0.92	3.77
	Male	51,093	24.57	13.92	0.93	3.76
	Female	48,533	24.04	12.92	0.91	3.77
	Am. Indian	172	23.36	13.81	0.93	3.73
	Asian	10,785	35.91	14.19	0.92	3.96
	Black	14,322	18.12	10.47	0.89	3.52
	Hispanic	33,328	18.97	10.48	0.88	3.57
	Pacific Islander	184	25.88	13.14	0.91	3.84
	White	37,959	27.79	13.11	0.91	3.87
	ML–Yes	8,735	13.69	6.61	0.76	3.23
	ML–No	90,945	25.34	13.49	0.92	3.81
	EconDis–Yes	36,703	18.41	10.17	0.88	3.54
	EconDis–No	62,964	27.76	13.92	0.92	3.87
	SWD–Yes	20,435	18.21	11.27	0.90	3.49
	SWD–No	79,244	25.89	13.51	0.92	3.83
11	NJSLA–S	98,881	28.66	15.09	0.93	4.00
	Male	50,484	28.62	15.88	0.94	3.98
	Female	48,242	28.68	14.20	0.92	4.01
	Am. Indian	161	26.94	13.65	0.91	3.99
	Asian	10,214	41.43	15.64	0.93	4.06
	Black	13,917	22.07	11.68	0.89	3.85
	Hispanic	32,576	22.94	12.15	0.90	3.87
	Pacific Islander	230	31.87	15.68	0.93	4.07
	White	39,415	32.23	14.93	0.93	4.05
	ML–Yes	7,090	15.69	6.86	0.75	3.46

Grade	Group	N Students	Average Raw Score	S.D.	Alpha	SEM
	ML–No	91,785	29.67	15.08	0.93	4.02
	EconDis–Yes	32,659	22.46	11.91	0.90	3.85
	EconDis–No	66,216	31.72	15.54	0.93	4.04
	SWD–Yes	20,089	22.58	13.23	0.92	3.81
	SWD–No	78,786	30.22	15.14	0.93	4.03

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 all close to .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	14	14	6.37	3.00	0.69	1.68
	TE	35	35	12.70	7.49	0.89	2.51
	CR	3	11	4.00	3.15	0.69	1.75
8	MC	14	14	5.76	2.68	0.60	1.70
	TE	47	47	14.85	9.10	0.90	2.84
	CR	3	11	3.71	2.77	0.73	1.45
11	MC	29	29	12.50	5.58	0.81	2.43
	TE	38	38	11.86	7.74	0.90	2.51
	CR	3	11	4.31	2.95	0.72	1.56

8.2 Item Response Theory Reliability

The reliability of the scale scores ascertained from the partial credit model (PCM; Masters, 1982) 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, the 2024 NJSLA–S was 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 (Cronbach, 1951), 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 partial credit model 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:

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

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. At grade 5, the TIF peaked close to the Level 2/3 cut score, but there was a drop in information at the Level 1/2 and Level 3/4 cuts. At grades 8 and 11, the TIF peaked almost directly on the respective Level 2/3 cut scores. However, there was a drop in information at the Level 1/2 cut for grade 8 and at the Level 3/4 cut for grade 11. Overall, the TIFs provide ample evidence that student ability estimates are reliable at the most important decision points. Nonetheless, grade 5 would benefit from more information around the Level 1/2 and Level 3/4 cut scores on future tests. In addition, grade 8 would benefit from more information around the level 1/2 cut score.

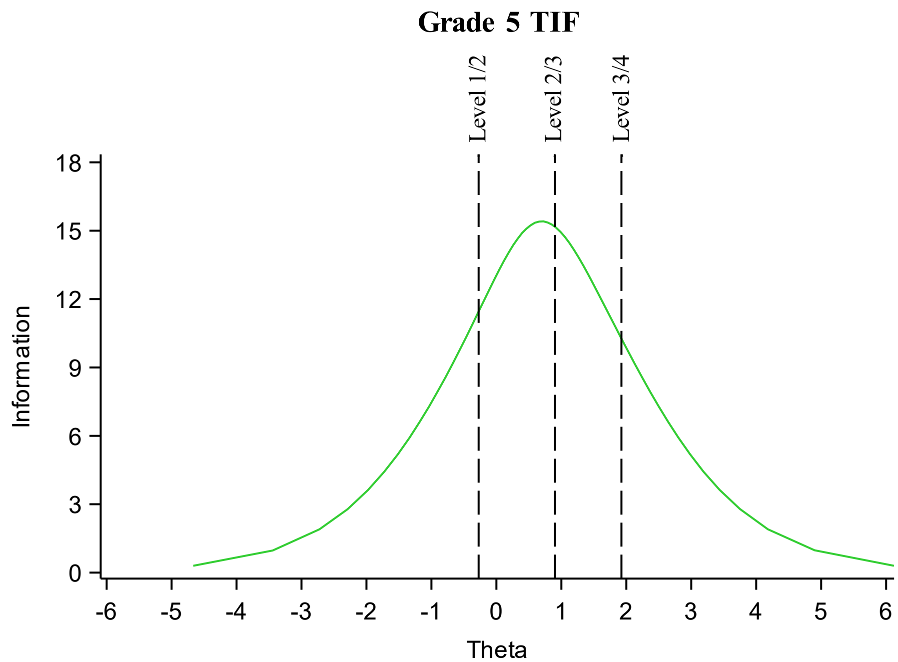


Figure 8.2.1. Grade 5 Test Information Function

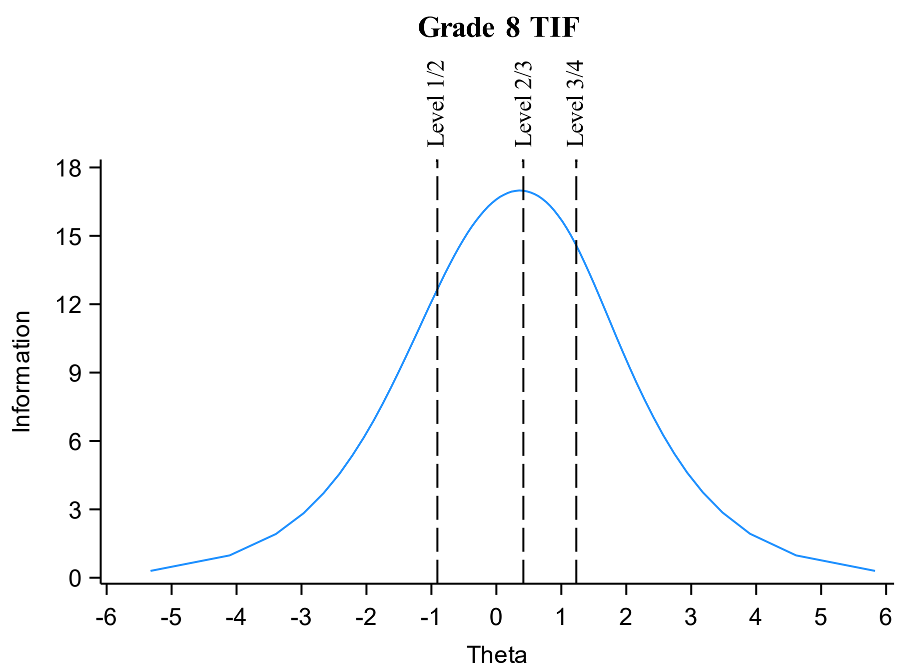


Figure 8.2.2. Grade 8 Test Information Function

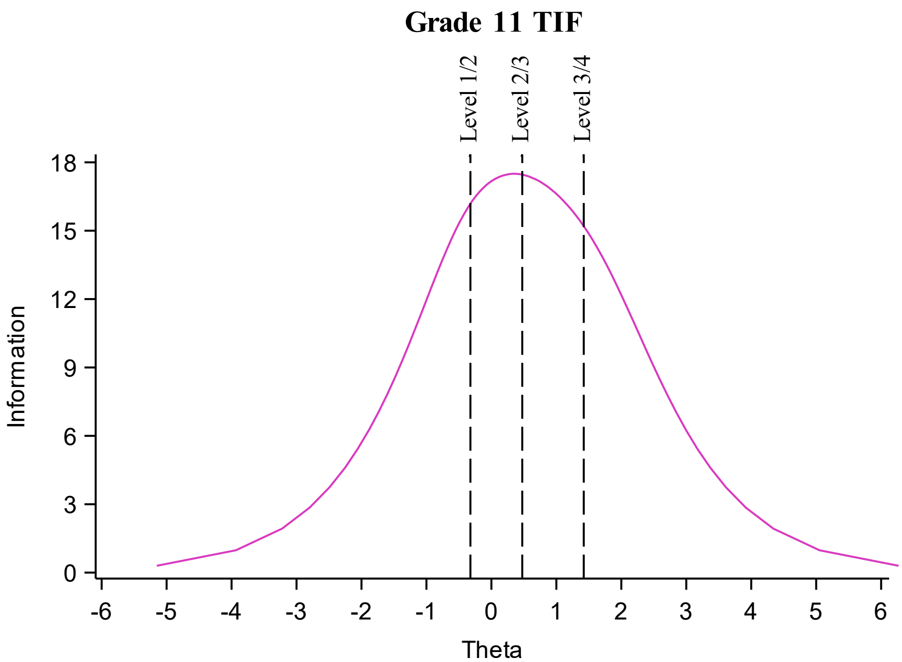


Figure 8.2.3. Grade 11 Test Information Function

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:

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

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.

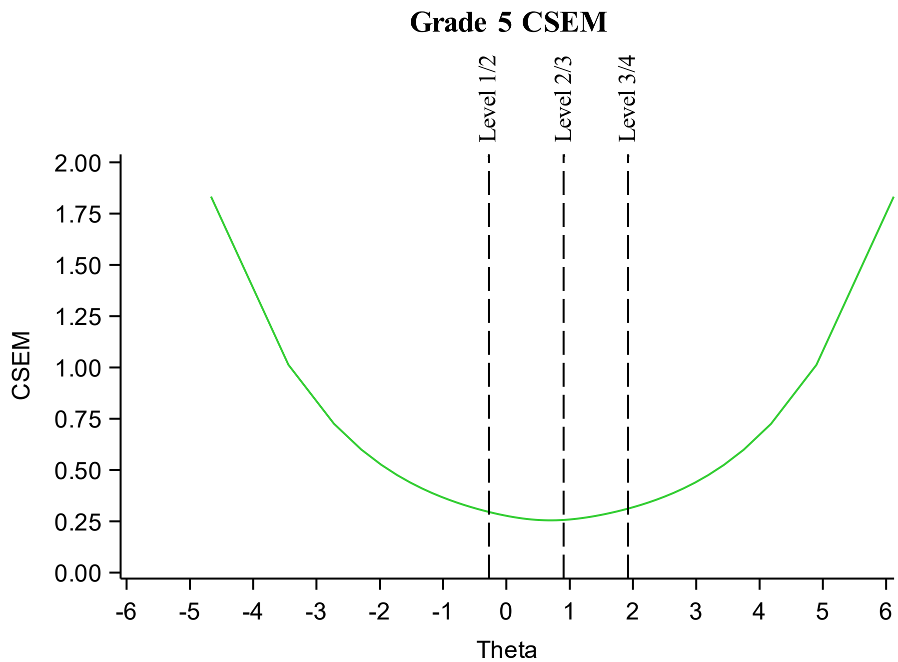


Figure 8.2.4. Grade 5 Conditional Standard Error of Measurement

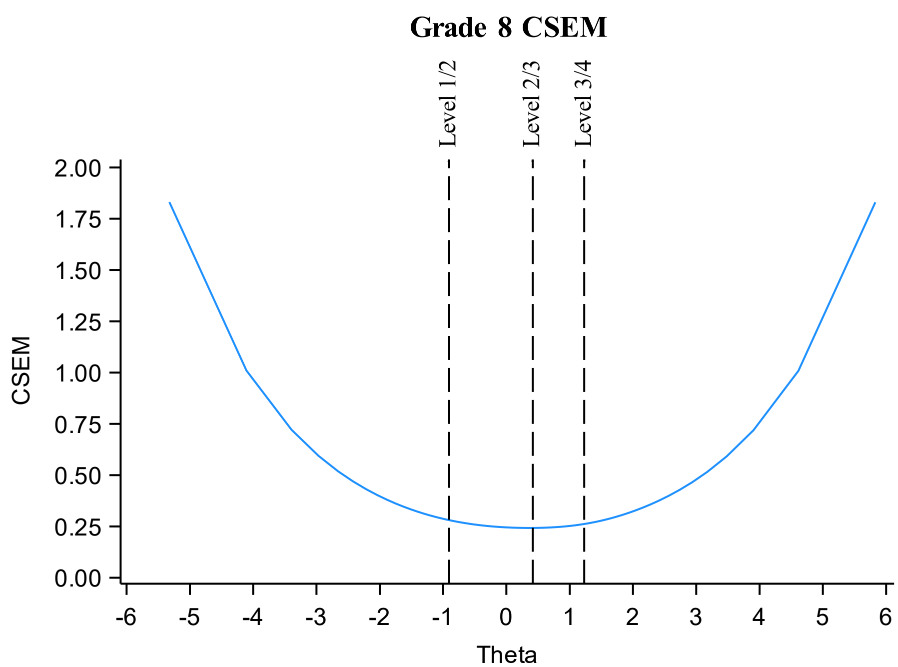


Figure 8.2.5. Grade 8 Conditional Standard Error of Measurement

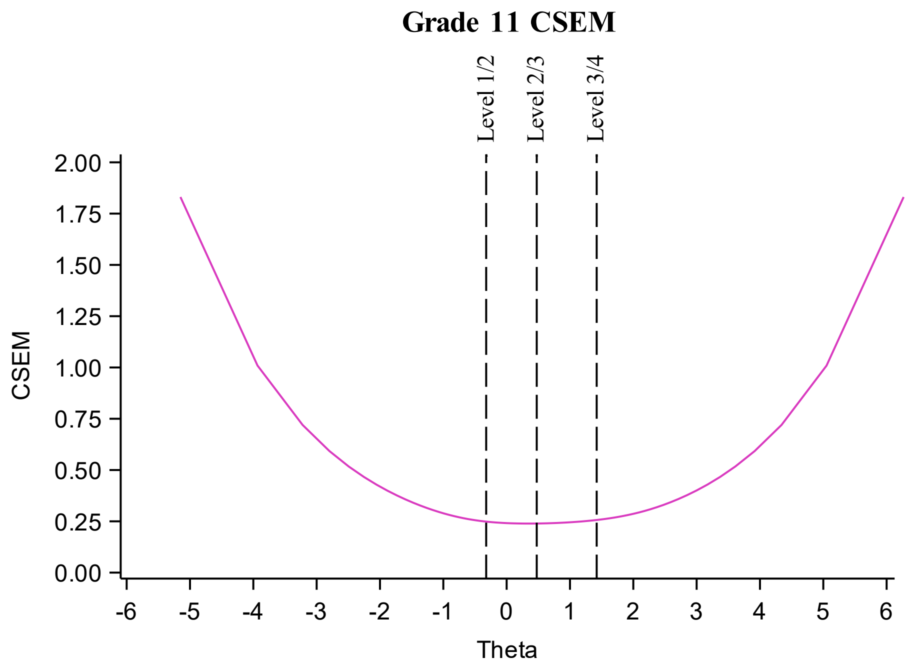


Figure 8.2.6. Grade 11 Conditional Standard Error of Measurement

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 show the 2024 NJSLA–S item maps for grade levels 5, 8, and 11, respectively. Each item map figure is delineated into two panels, the top containing the item difficulty estimate distribution and the lower containing the ability (theta) distribution. As shown in the figures, at each grade level, NJSLA–S items were appropriately targeted to the student ability distribution. At grade 5, the item difficulty distributions peaked between the Level 1/2 and the Level 2/3 cut scores; the theta distributions peaked around the Level 1/2 cut score. At grade 5, there were few students above the Level 3/4 cut score and very few items along that part of the scale. The grade 8 item difficulty distribution was lacking items at the lower (easier) part of the scale in comparison to the student ability distribution. At grade 11, the item difficulty distribution peaked at the Level 2/3 cut and saw several items at the upper (harder) part of the scale, while the student ability distribution peaked near the Level 1/2 cut.

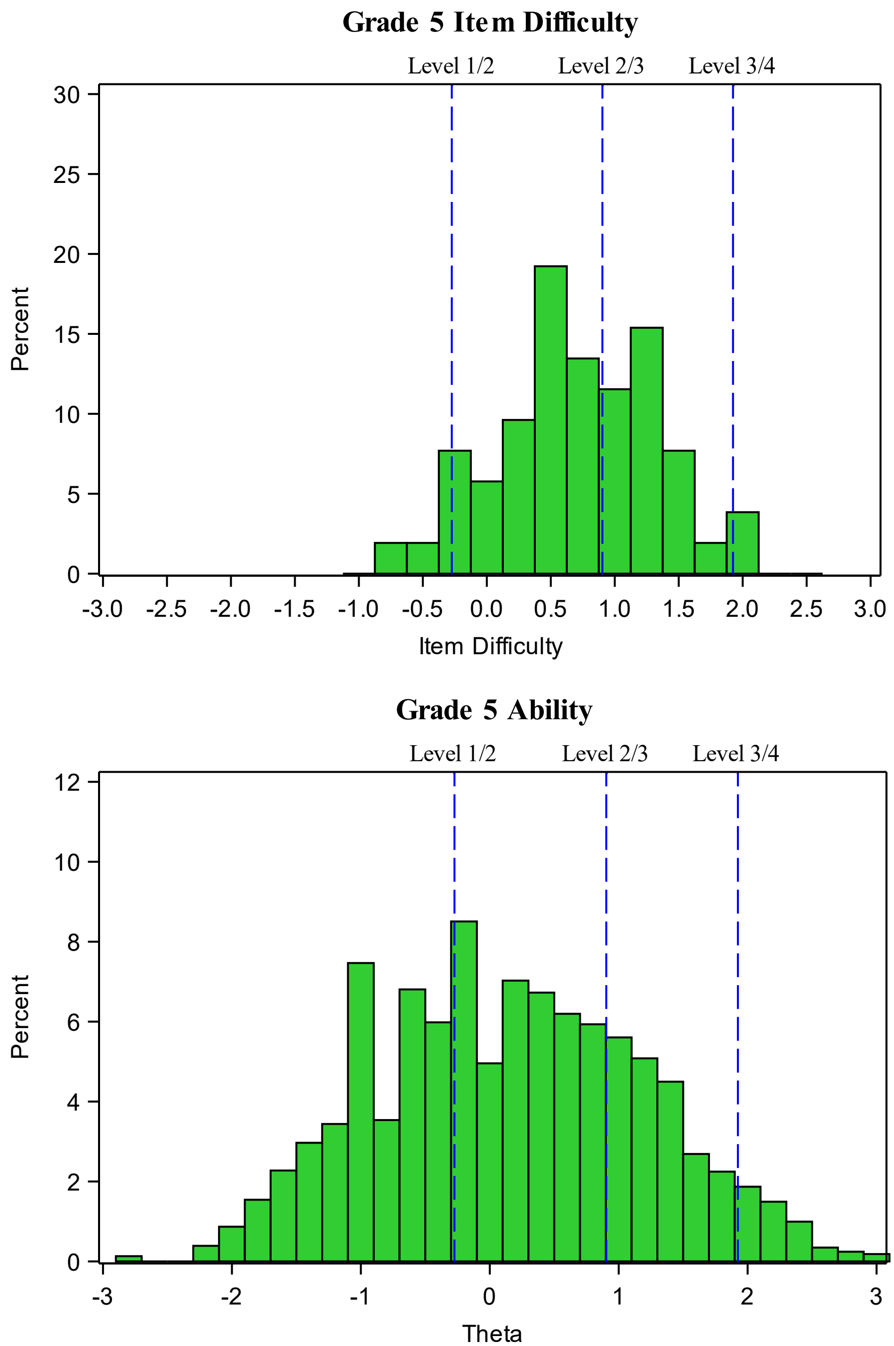
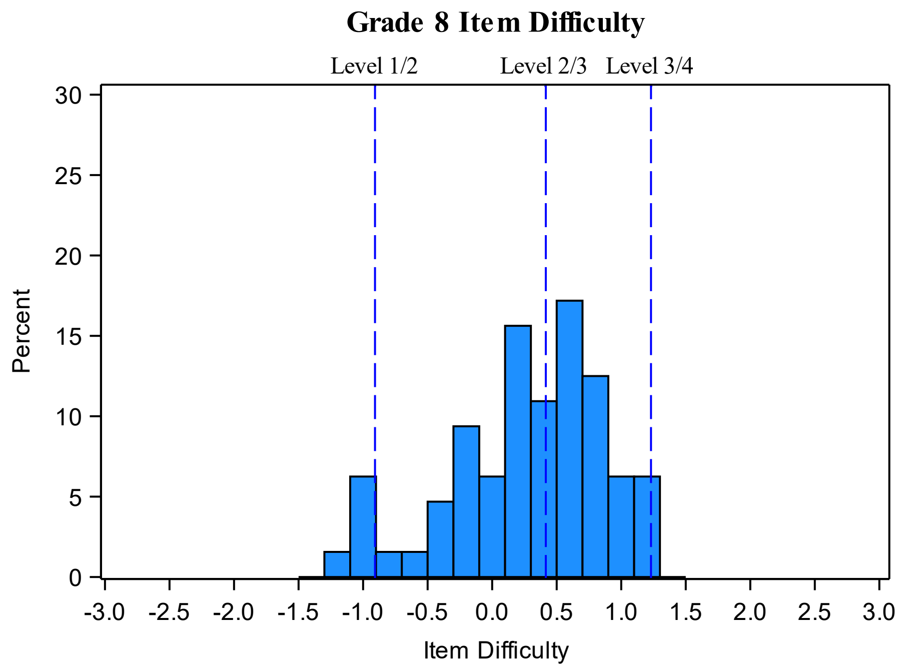


Figure 8.2.7. Grade 5 Item Difficulty and Student Ability Distributions



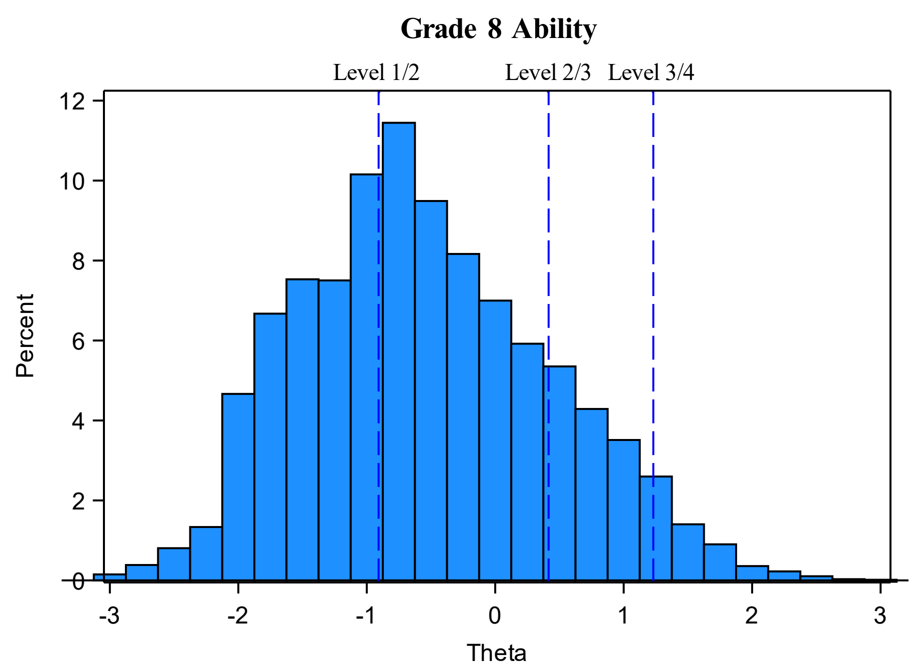
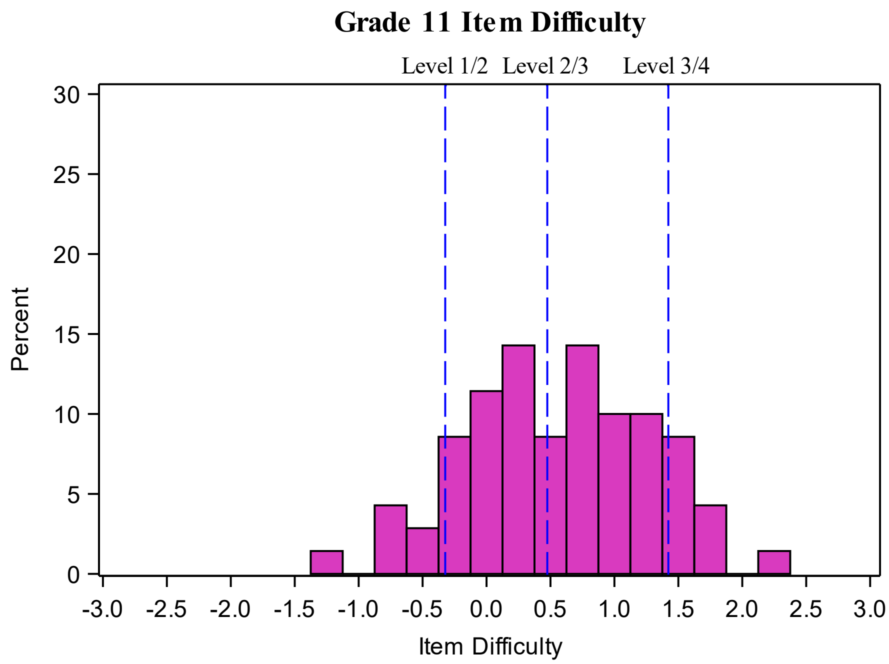


Figure 8.2.8. Grade 8 Item Difficulty and Student Ability Distributions



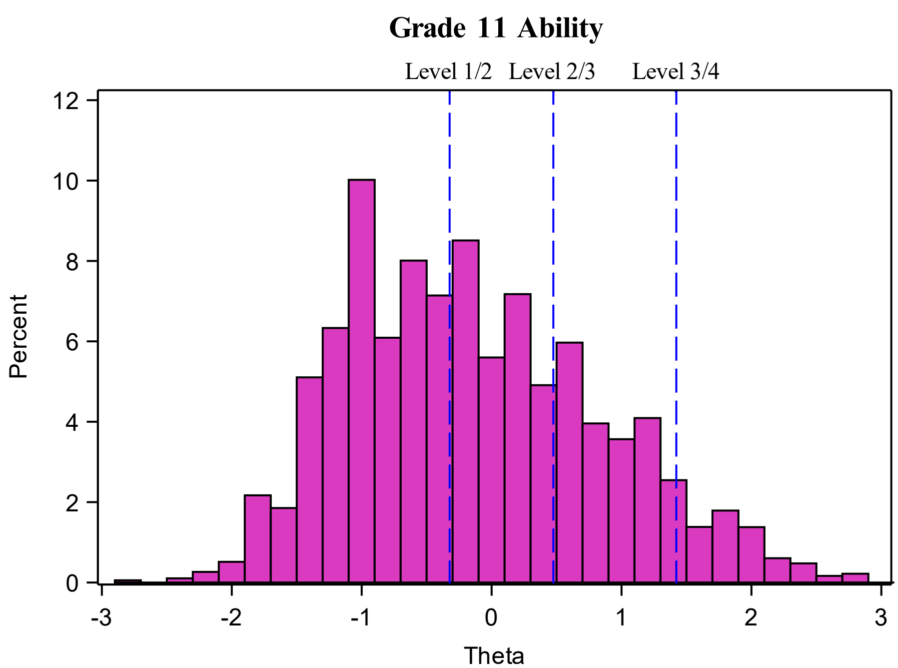


Figure 8.2.9. Grade 11 Item Difficulty and Student Ability Distributions

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 2024 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 2024 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 grade 8, the Level 1/2 cut score's CSEM was slightly higher than that at the Level 3/4, meaning that there was slightly less error in the scale score at 150 than at 231. At grades 5 and 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. In 2024, no overlap between the upper or lower bound of a cut score and another cut score was found for any 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	12.6	124.8	175.2
	Level 2/3	200	10.9	178.3	221.7
	Level 3/4	243	12.9	217.1	268.9
8	Level 1/2	150	10.7	128.5	171.5
	Level 2/3	200	9.2	181.6	218.4
	Level 3/4	231	9.9	211.3	250.7
11	Level 1/2	158	13.2	131.6	184.4
	Level 2/3	200	12.7	174.7	225.3
	Level 3/4	250	13.5	223.1	276.9

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:

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

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 .77. 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 .92, indicating 90% to 92% probability of being correctly classified as Level 3 or above. The overall NJSLA–S proficiency classification should be interpreted as 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	.64	.77	.92
11	158	200	250	.63	.75	.91

8.4 Reliability of Subscore Proficiency Classifications

As discussed in Section 7.3, in each content domain and practice, subscore proficiency are 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 .66, 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	Theta	CSEM
5	Earth and Space	0.48	0.70	Near/Met	8	0.414	0.440
				Above	14	1.545	0.464
	Life	0.53	0.74	Near/Met	8	0.297	0.444
				Above	15	1.788	0.533
	Physical	0.54	0.73	Near/Met	8	0.264	0.473
				Above	14	1.652	0.530
	Critiquing	0.41	0.66	Near/Met	6	0.377	0.445
				Above	12	1.554	0.492
	Investigating	0.56	0.75	Near/Met	8	0.322	0.475
				Above	14	1.686	0.508
	Sensemaking	0.54	0.74	Near/Met	10	0.285	0.436
				Above	17	1.723	0.513
8	Earth and Space	0.55	0.77	Near/Met	10	–0.131	0.417
				Above	17	1.091	0.440
	Life	0.57	0.80	Near/Met	10	–0.053	0.422
				Above	17	1.241	0.462
	Physical	0.54	0.76	Near/Met	9	–0.200	0.454
				Above	16	1.124	0.441
	Critiquing	0.56	0.78	Near/Met	13	–0.128	0.378
				Above	21	1.017	0.393
	Investigating	0.51	0.77	Near/Met	8	–0.081	0.476
				Above	14	1.27	0.503
	Sensemaking	0.58	0.79	Near/Met	8	–0.161	0.455
				Above	15	1.228	0.467
11	Earth and Space	0.48	0.70	Near/Met	10	–0.094	0.410
				Above	17	1.159	0.452
	Life	0.56	0.74	Near/Met	9	–0.015	0.434
				Above	16	1.321	0.460
	Physical	0.60	0.77	Near/Met	10	–0.036	0.416
				Above	17	1.088	0.397
	Critiquing	0.55	0.73	Near/Met	9	–0.106	0.408
				Above	17	1.193	0.405
	Investigating	0.57	0.75	Near/Met	12	0.019	0.396
				Above	19	1.164	0.428
	Sensemaking	0.53	0.73	Near/Met	8	–0.072	0.467
				Above	14	1.182	0.469

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 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 re-scored. 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 75.2% to 83.5% for grade 5, from 65.8% to 77.8% for grade 8, and from 66.8% to 83.7% for grade 11. While there was only one grade 8 CR item that showed weighted Kappa above .90, all the CR items in grade 5 and two of the CR items in grade 11 had weighted Kappas above .90. Overall, rater agreement on the NJSLA–S CR items was excellent.

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	83.5	16.5	0.93
	CR 2	75.2	24.8	0.94
	CR 3	76.7	23.3	0.94
8	CR 1	77.8	22.2	0.90
	CR 2	71.2	28.8	0.87
	CR 3	65.8	34.2	0.88
11	CR 1	69.5	30.5	0.88
	CR 2	83.7	16.3	0.91
	CR 3	66.8	33.2	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 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 less 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 under-represented. 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, the 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 2019 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 his/her thinking as they worked to answer each item. The study included both a concurrent account of problem-solving, as well as a retrospective cognitive interview. Because the study was conducted in the fall, an off-grade approach was used to ensure 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. However, at the subscore level, results from a confirmatory factor analysis provided some evidence that the internal structure was not performing as intended.

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 .77 at grade 5 between Physical Science and Earth and Space Science. Among the scientific practices, the intercorrelation ranged from .79 to .81.

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 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 .92 to .93 for the CBT form. At the subscore level the reliability coefficients were relatively impressive, with the lowest estimates of .75 for both the grade 5 Physical Science and the grade 8 Investigating subscores.

9.3.3 Confirmatory Factor Analysis Using the 2023 NJSLA–S Tests

To provide further evidence supporting the internal subscore structure of the NJSLA–S, confirmatory factor analyses (CFA) were conducted using the 2023 operational NJSLA–S test results. CFA is a powerful statistical technique used to verify proposed measurement models based on the underlying covariance structure of the data. With this technique, a measurement model is specified, the data are fit to the specified model, and then fit indices (and other criteria) can be examined to determine how well the model fits the data (Brown, 2006; Kline, 2011).

Figure 9.3.1 presents the *a priori* test structure specified for the content domain CFA across all grades. Figure 9.3.2 presents the *a priori* test structure that was specified for the scientific practice CFA across all grades. Adjusted unweighted least squares with means and variances adjusted (ULSMV) was used to estimate these models. ULSMV is a robust estimation method that is appropriate to use when data are categorical or ordinal and there are potential concerns about non-normality in the latent variables (Beauducel & Herzberg, 2006; Muthén, 1993; Muthén et al., 1997).

To assess model fit, the model chi-square test was used as a test of exact model fit. However, it is well documented that the model chi-square test is sensitive to sample size (Cheung & Rensvold, 2002), so approximate fit indices were used to supplement the model chi-square. Three approximate fit indices were examined for each model: the comparative fit index (CFI; Bentler 1990), the root mean square error of approximation (RMSEA; Steiger, 1990), and the standardized root mean square residual (SRMR; Jöreskog & Sörbom, 1988). Following the recommendations of Brown and Cudeck (1993) and Hu and Bentler (1999), values of CFI greater than .90 and .95 were considered evidence of acceptable model fit and good model fit, respectively; values of RMSEA less than .08 and .05 were considered evidence of acceptable model fit and good model fit, respectively; and values of SRMR less than .08 were considered evidence of good model fit.

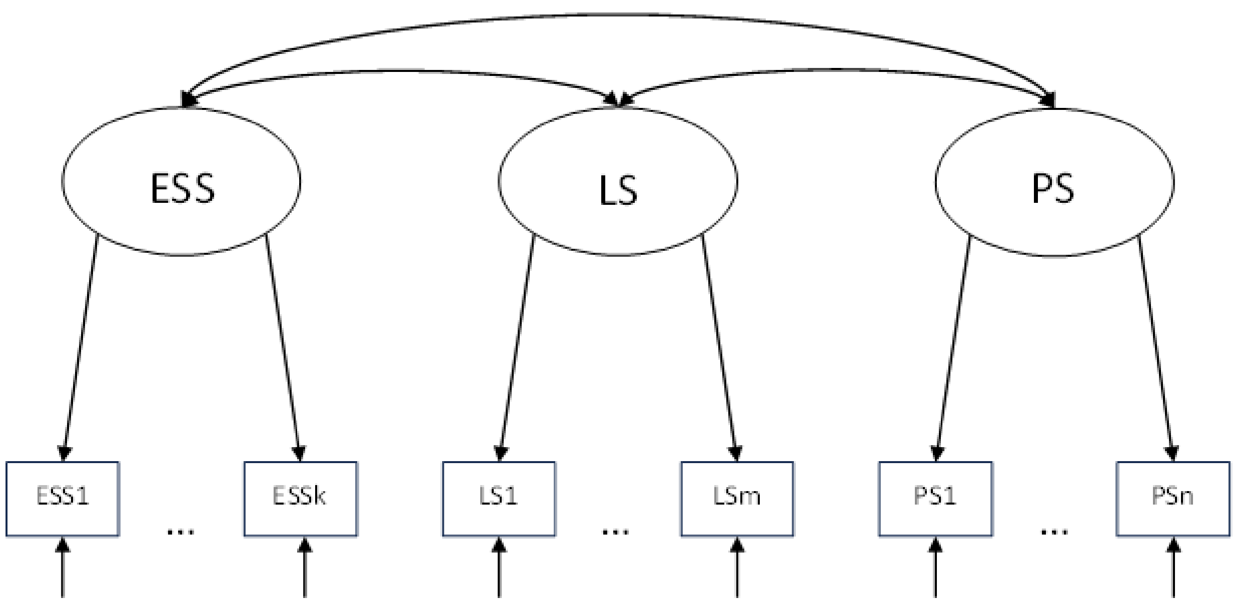


Figure 9.3.1. Domain Subscore Structure

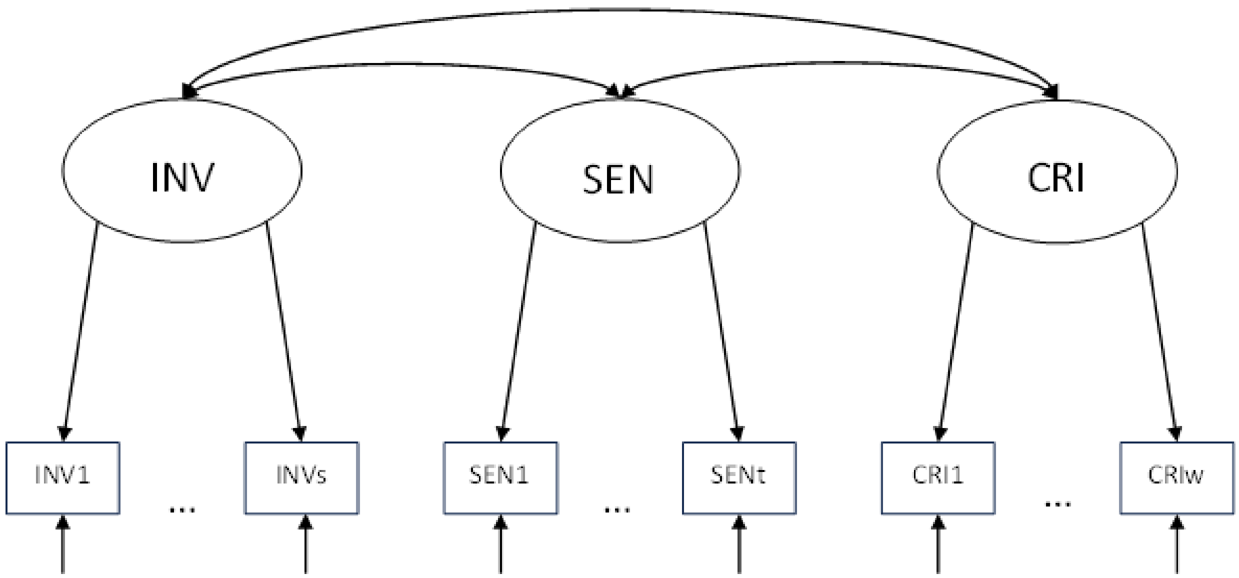


Figure 9.3.2. Practice Subscore Structure

There are a few considerations regarding the use of the model-fit statistics in this study. First, since a robust estimation method was used, the scaled chi-square statistic is reported (Satorra & Bentler, 1994). The scaled chi-square statistic is simply the chi-square statistic modified by a scaling parameter to adjust for violations of normality. However, since the CFI and the RMSEA are functions of the chi-square statistic, they are calculated using this scaled chi-square statistic. Second, studies have shown that the CFI and RMSEA are impacted by the estimation method. Specifically, ULSMV model estimation has been shown to result in higher CFI values and lower RMSEA values when compared to maximum likelihood estimation methods (Garrido et al., 2016; Nye & Drasgow, 2011; Shi & Maydeu-Olivares, 2020; Xia & Yang, 2019). This can result in a lower likelihood of identifying a poorly fitting model. However, the SRMR has been shown to be consistent across estimation methods (Shi & Maydeu-Olivares, 2020), so it was given considerable weight when judging model fit.

9.3.3.1. Domain results. The domain model converged without issues for all three grades. The model-fit indices for each grade are presented in Table 9.3.1. For all three grades, the chi-square test indicated that the differences between the observed and model-predicted covariances were statistically significant. However, the approximate model-fit indices indicated that the domain model was a good representation of the data. As shown in Table 9.3.2, the correlations between the latent subscores were very high due to the underlying unidimensionality of the NJSLA–S, but the acceptable model fit indices suggest that they were also empirically distinct from one another. The parameter estimates obtained from fitting the domain model are provided in Appendix O for all three grades.

Table 9.3.1. Model-Fit Indices for the Domain Model

Grade	Chi-square	CFI	RMSEA	SRMR
5	χ_S^2 (1221) = 43140.93*	0.985	0.019	0.024
8	χ_S^2 (2012) = 40174.26*	0.985	0.014	0.020
11	χ_S^2 (2346) = 74085.05*	0.975	0.018	0.026

Note. N_5 = 96,392; N_8 = 101,478; N_{11} = 94,023; * = p < .001

Table 9.3.2. Correlations between the Latent Subscores Implied by the Domain Model

	Grade 5			Grade 8			Grade 11	
	ESS	LS		ESS	LS		ESS	LS
ESS	-	-	ESS	-	-	ESS	-	-
LS	.98	-	LS	.98	-	LS	.99	-
PS	.97	.99	PS	.99	.97	PS	.99	.98

9.3.3.2. Practice results. The practice model converged without issues for all three grades but returned nonpositive definite latent covariance matrices in each grade. A nonpositive definite latent covariance matrix suggests that some aspect of the model is not accurately reflecting the covariance structure observed in the data (i.e., the model is misspecified). For the NJSLA–S, this misspecification was determined to be caused by collinearity between the practice subscore categories. That is, the practice subscore categories were not empirically distinct enough to be treated as separate subscore categories. Since the practice model returned a nonpositive definite latent covariance matrix for all three grades, the parameter estimates should not be interpreted. Therefore, the parameter estimates obtained from the practice models are omitted from Appendix O.

9.3.3.3. Follow-up analysis. The item subscore correlations (rpb_{sub}) were examined to further examine the subscore categories. The item subscore correlations are analogous to item total correlations except the raw subscore rather than total test raw score is used for calculation. For each item, the item subscore correlation was calculated for all three domains (i.e., rpb_{ESS}, rpb_{LS}, and rpb_{PS}) and all three practices (i.e., rpb_{CRI}, rpb_{INV}, and rpb_{SEN}).

Table 9.3.3 and Table 9.3.4 show the number of items with the highest rpb_{sub} on their assigned domain and practice, respectively. For all three grades, at least 50% of the items showed the highest rpb_{sub} on their assigned domain. Conversely, only at least 50% of the Grade 5 Sensemaking items, the Grade 8 Critiquing and Sensemaking items, and the Grade 11 Sensemaking items showed the highest rpb_{sub} on their assigned practice.

Table 9.3.3. Number of Items with Highest rpb_{sub} on Assigned Domain

Grade	Domain	Item Count	Number (%) of Items with Highest rpb_{sub} on Assigned Domain
5	ESS	18	18 (100.00%)
	LS	17	17 (100.00%)
	PS	16	16 (100.00%)
8	ESS	20	20 (100.00%)
	LS	24	14 (58.33%)
	PS	21	21 (100.00%)
11	ESS	20	20 (100.00%)
	LS	23	15 (62.55%)
	PS	26	26 (100.00%)

Note. ESS = Earth and Space Science; LS = Life Science; PS = Physical Science

Table 9.3.4. Number of Items with Highest rpb_{sub} on Assigned Practice

Grade	Practice	Item Count	Number (%) of Items with Highest rpb_{sub} on Assigned Practice
5	CRI	18	4 (22.22%)
	INV	16	1 (6.25%)
	SEN	17	10 (58.82%)
8	CRI	22	11 (50.00%)
	INV	20	1 (5.00%)
	SEN	23	13 (56.52%)
11	CRI	22	6 (27.27%)
	INV	23	6 (26.09%)
	SEN	24	14 (58.33%)

Note. CRI = Critiquing; INV = Investigating; SEN = Sensemaking

The results of the CFAs and the item subscore correlations support the structure of the NJSLA–S in terms of the content domains. Specifically, the results suggest that while the domain subscore categories are highly related, they are empirically distinct enough to provide some unique information in the form of subscores. However, the results of the CFAs and the item subscore correlations did not support the structure of the scientific practices. The presence of nonpositive definite latent covariance matrices and much commingling between the rpb_{sub} values suggest that the practice categories are not empirically distinct from one another. Explaining this finding requires further information from the item and test development processes of the NJSLA–S, which is presented below.

Each content domain: Earth and Space Science (ESS), Life Science (LS), and Physical Science (PS) is divided into a subset of fundamental, core ideas that are necessary for understanding a given science discipline. The core ideas are woven throughout the K–12 standards, providing themes that build in complexity and allow for a more interdisciplinary approach to learning science. As a result, each DCI standard is a discrete scientific concept that can be directly assessed, and the cluster design of the NJSLA–S allows a DCI to be assessed multiple times. The discreteness of the definitions and the directness with which the content domains can be measured has resulted in distinct, well-defined subscore categories.

However, while all eight of the practices are defined in the standards, the SEP standards are not as discreetly defined as the DCI standards. This is likely because the DCIs are based on content knowledge and not application/skill-based standards. Overlap among the practices and perhaps more importantly between reporting categories exists within the standards, which is most likely why collinearity was observed between the SEP subscores. Additionally, the skills characterized in the SEPs are inherently connected. There is no way for a student to critique (Critiquing) a dataset without analyzing the data (Sensemaking) and probably hypothesizing about it (Investigating) first. This inherent connectedness is likely another reason why the SEP subscores displayed collinearity.

9.4 Evidence Based on Relationships to Other Variables

Evidence based on relationships to other variables takes the form of relationships between test scores and other variables that are external to the test (AERA, APA, NCME, 2014). This evidence can come from investigating the relationships among tests that measure similar constructs, tests that measure different constructs, or other outcomes that a test purports to predict. In 2024, NJDOE conducted an internal validity study that investigated the relationships among the NJSLA–S and other New Jersey large-scale, statewide subject scale scores (i.e., NJSLA–ELA and NJSLA–Math). The results indicated that the scientific KSAs the NJSLA–S is intended to measure comprise a construct distinct from other disciplines measured by the New Jersey statewide assessment program.

The results at grade 5 are displayed in Table 9.4.1. ELA consists of two major claims: Reading Complex Text and Writing. The scale score for those two major claims were added to the correlation matrix. Grade 5 Science showed correlations of .80 and .83 with ELA (ELA 5) and math (math 5), respectively. The correlation between science and ELA Writing was .65, and the correlation between science and ELA Reading was .82.

Table 9.4.1: Grade 5 Intercorrelations by Content Area

Content Areas	N	Correlation
SCI 5 - ELA 5	94,505	.80
SCI 5 - ELA 5 Reading	94,505	.82
SCI 5 - ELA 5 Writing	94,505	.65
SCI 5 - MAT 5	96,375	.83
MAT 5 - ELA 5	94,644	.77
MAT 5 - ELA 5 Reading	94,644	.82
MAT 5 - E5 LA Writing	94,644	.65

The results at grade 8 are displayed in Table 9.4.2. The only difference in calculating the grade 8 intercorrelation matrix in comparison to grade 5 pertained to the math scale scores. Depending on which course a student was enrolled in, there were four different math assessments that grade 8 students could have taken: Math 8 (MAT 8), Algebra I (ALG 1), Algebra II (ALG 2), or Geometry (GEO). Therefore, instead of one math scale score the grade 8 intercorrelation matrix is based on four distinct math scale scores. It is impossible for students to have scale scores on two different math tests; thus, those correlations were not included in Table 9.4.2.

The correlation between Grade 8 science (SCI 8) and ELA 8 was .77. The correlations between science and various math test scores ranged from .64 to .78. This is most likely due to the higher- and lower-achieving students taking different assessments, which could decrease the scale score variance for each math test. Thus, the magnitude of the correlations between science and the various math tests appears reasonable when considering that math achievement is more homogeneous within each subgroup than if all students at all ability levels were taking the same assessment.

Table 9.4.2: Grade 8 Intercorrelations by Content Area

Content Areas	N	Correlation
SCI 8 - ELA 8	97,836	.77
SCI 8 - ELA 8 Reading	97,836	.79
SCI 8 - ELA 8 Writing	97,836	.65
SCI 8 - MAT 8	65,599	.76
SCI 8 - ALG I	28,554	.78
SCI 8 - GEO 1	4,760	.75
SCI 8 - ALG 2	600	.64
MAT 8 - ELA 8	64,312	.67
MAT 8 - ELA 8 Reading	64,312	.62
MAT 8 - ELA 8 Writing	64,312	.55
ALG 1 - ELA 8	28,805	.65
ALG 1 - ELA 8 Reading	28,805	.67
ALG1 - ELA 8 Writing	28,805	.53
GEO 1 - ELA 8	4,812	.56
GEO 1 - ELA 8 Reading	4,812	.50
GEO 1 - ELA 8 Writing	4,812	.43
ALG 2 - ELA 8	610	.36
ALG 2 - ELA 8 Reading	610	.45
ALG 2 - ELA 8 Writing	610	.20

The results for 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 math was .78. This was consistent with the relationships between science, math, and ELA seen in other grades and is in line with expectations. The relationships between science and ELA Reading GPA and ELA Writing GPA were .72 and .59, respectively, showing similar relationships to those seen in grade 5 and grade 8.

Table 9.4.3: Grade 11 Intercorrelations by Content Area

Content Areas	N	Correlations
SCI 11 - ELA GPA	94,475	.70
SCI 11 - ELA Reading GPA	94,475	.72
SCI 11 - ELA Writing GPA	94,475	.59
SCI 11 - MAT GPA	97,796	.78
MAT GPA - ELA GPA	102,043	.74
MAT GPA - ELA Reading GPA	102,043	.73
MAT GPA - ELA Writing GPA	102,043	.62

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 were used to evaluate the NJSLA–S at both the test and item level, including:

- summaries of item performance across grade level, content domain, scientific practice, and item type to verify that the items are 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 is for portions of the population
- evaluating the IRT assumptions of the PCM to ensure it is 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 reviewed 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 a 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 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 is reporting 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 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 over-interpretation 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.

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

FIRSTNAME004 C. LASTNAME004
Spring 2024 Grade: 11
SID: **1200001006** DOB: **04/25/2006**
Local Student Identification: **674316**
NJ SAMPLE DISTRICT NAME
NJ SAMPLE SCHOOL NAME

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.

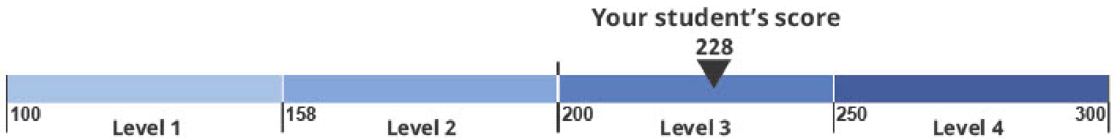
Rz8Nppysqw7T

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

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

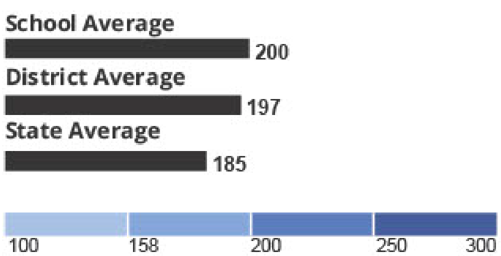
Proficient

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

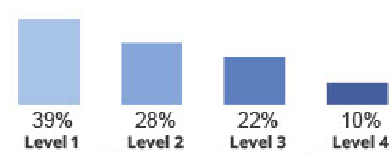


FIRSTNAME004's score on the NJSLA-S indicates that your student is at Level 3.
Students who are at Level 3 demonstrated appropriate grade-level understanding of the New Jersey Student Learning Standards-Science (NJSLS-S) by comprehending information from a variety of sources (e.g., text, charts, graphs, 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 10042023-Z9999999-R-000001-111-S-000001-111 - 0000000

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.

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

! 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.

✓ 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.

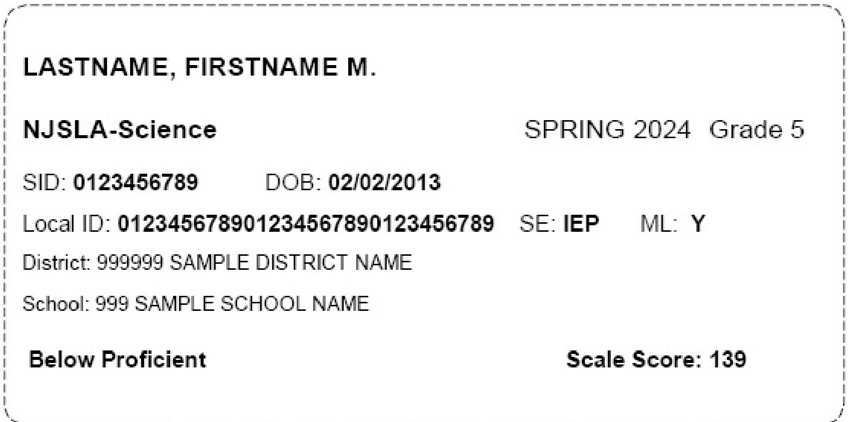
Learn More about the New Jersey Learning Standards
Explore your school website, or ask your principal, for information on your school's annual assessment schedule; the curriculum chosen by your district to give students more hands-on learning experiences that meet state standards; and to learn more about 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

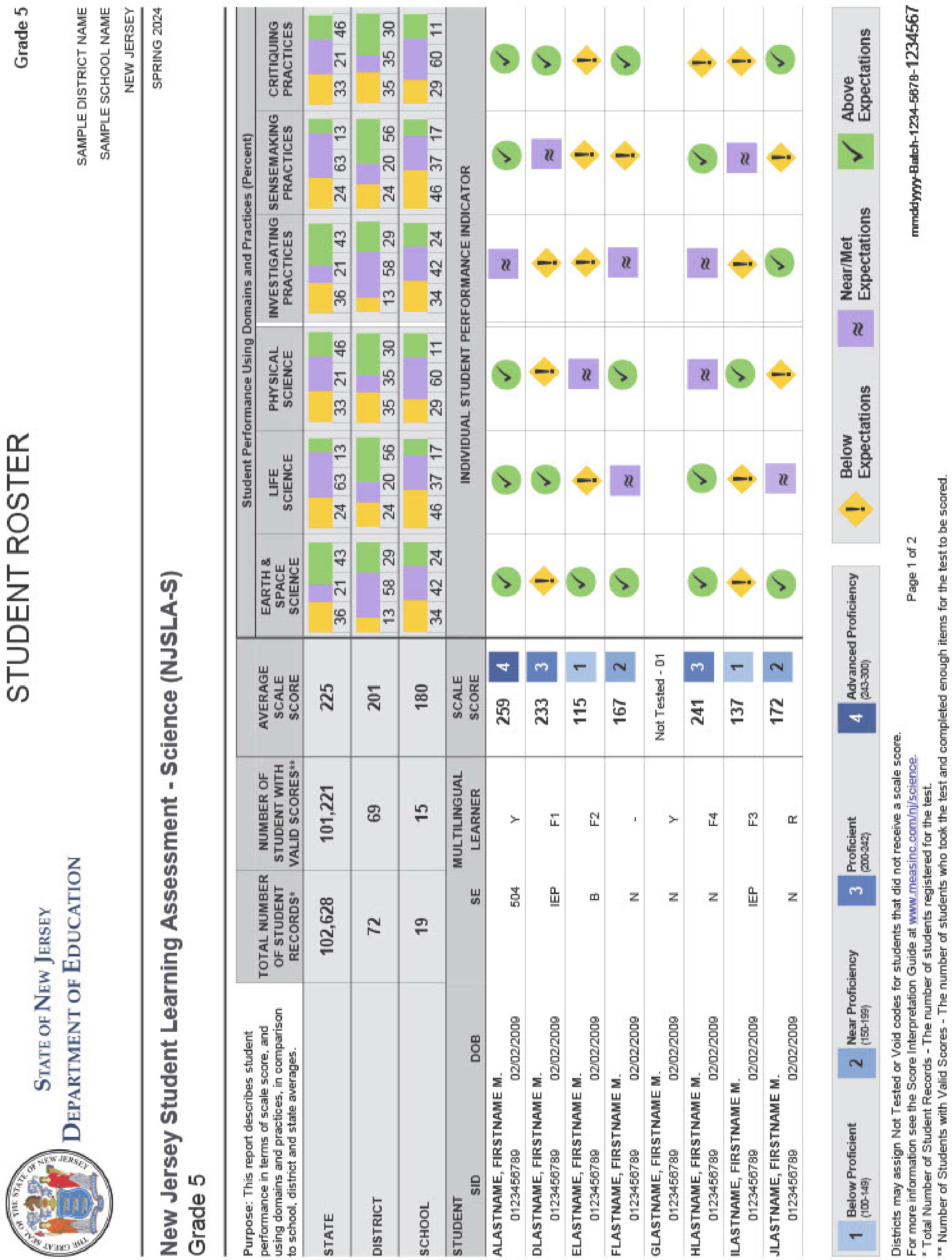
School: 999 SAMPLE SCHOOL NAME

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.



STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

STUDENT ROSTER

Grade 5

SAMPLE DISTRICT NAME
SAMPLE SCHOOL NAME
NEW JERSEY
SPRING 2024

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

Purpose: This report describes student performance in terms of scale score, and using domains and practices, in comparison to school, district and state averages.	TOTAL NUMBER OF STUDENT RECORDS*	NUMBER OF STUDENT WITH VALID SCORES**	AVERAGE SCALE SCORE	EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
STATE	102,628	101,221	225	36 21 43	24 63 13	33 21 46	36 21 43	24 63 13	33 21 46
DISTRICT	72	69	201	13 58 29	24 20 56	35 35 30	13 58 29	24 20 56	35 35 30
SCHOOL	19	15	180	34 42 24	46 37 17	29 60 11	34 42 24	46 37 17	29 60 11

Student Performance Using Domains and Practices (Percent)

STUDENT SID / DOB	SE	MULTILINGUAL LEARNER	SCALE SCORE	EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
ALASTNAME, FIRSTNAME M. 0123456789 02/02/2009	504	Y	259 4	✓	✓	✓	≈	✓	✓
DLASTNAME, FIRSTNAME M. 0123456789 02/02/2009	IEP	F1	233 3	!	✓	!	!	≈	✓
ELASTNAME, FIRSTNAME M. 0123456789 02/02/2009	B	F2	115 1	✓	!	≈	!	!	!
FLASTNAME, FIRSTNAME M. 0123456789 02/02/2009	N	-	167 2	✓	≈	✓	≈	!	✓
GLASTNAME, FIRSTNAME M. 0123456789 02/02/2009	N	Y	Not Tested - 01						
HLASTNAME, FIRSTNAME M. 0123456789 02/02/2009	N	F4	241 3	✓	✓	≈	≈	✓	!
ILASTNAME, FIRSTNAME M. 0123456789 02/02/2009	IEP	F3	137 1	!	!	✓	!	≈	!
JLASTNAME, FIRSTNAME M. 0123456789 02/02/2009	N	R	172 2	✓	≈	!	✓	!	✓

INDIVIDUAL STUDENT PERFORMANCE INDICATOR

1 Below Proficient (100-149) 2 Near Proficiency (150-199) 3 Proficient (200-242) 4 Advanced Proficiency (243-300)

! Below Expectations ≈ Near/Met Expectations ✓ Above Expectations

Districts may assign Not Tested or Void codes for students that did not receive a scale score.
For more information see the Score Interpretation Guide at www.measinc.com/nj/science.
* Total Number of Student Records - The number of students registered for the test.
** Number of Students with Valid Scores - The number of students who took the test and completed enough items for the test to be scored.

Page 1 of 2

mmddyyyy-Batch-1234-5678-1234567

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.

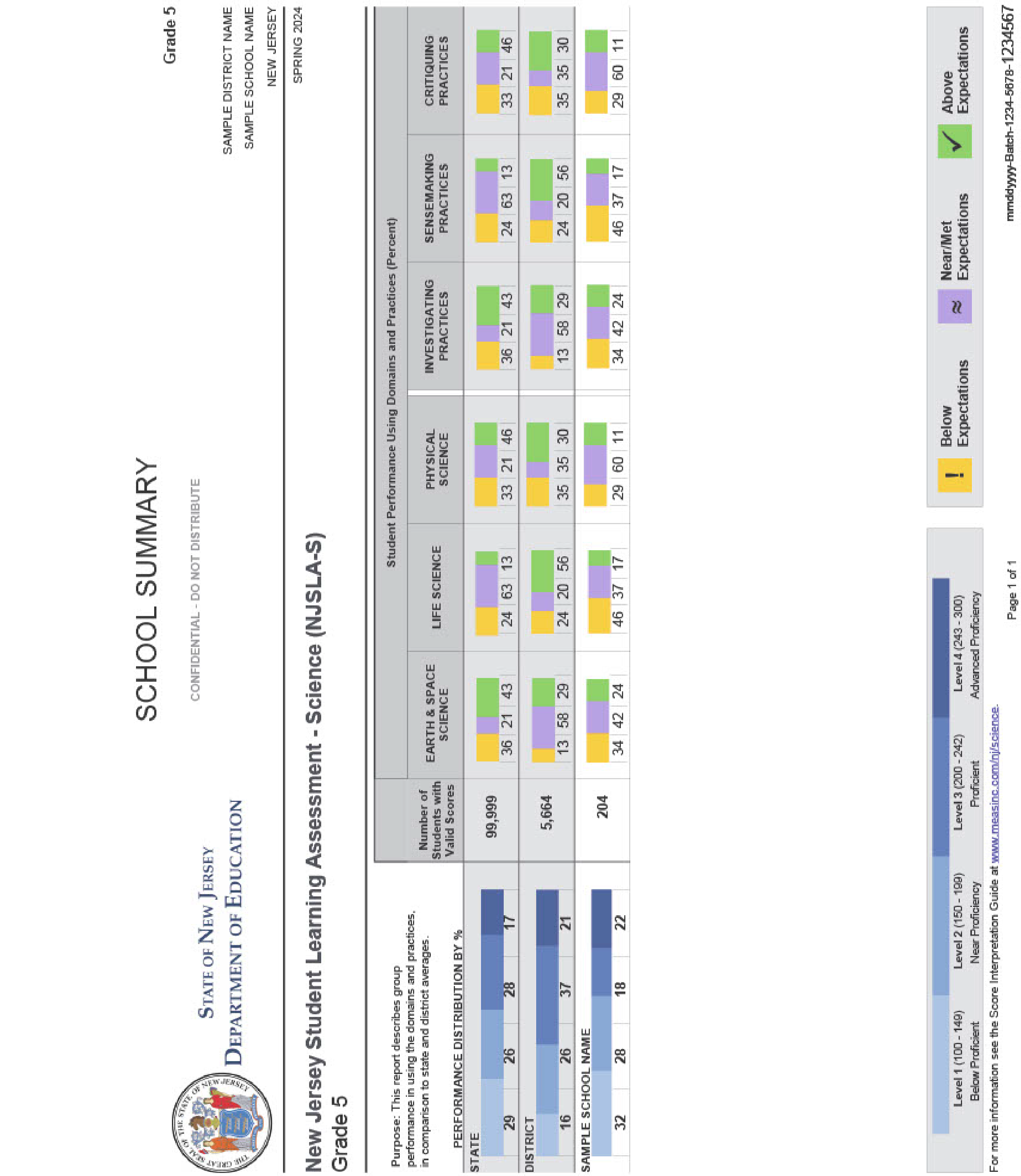
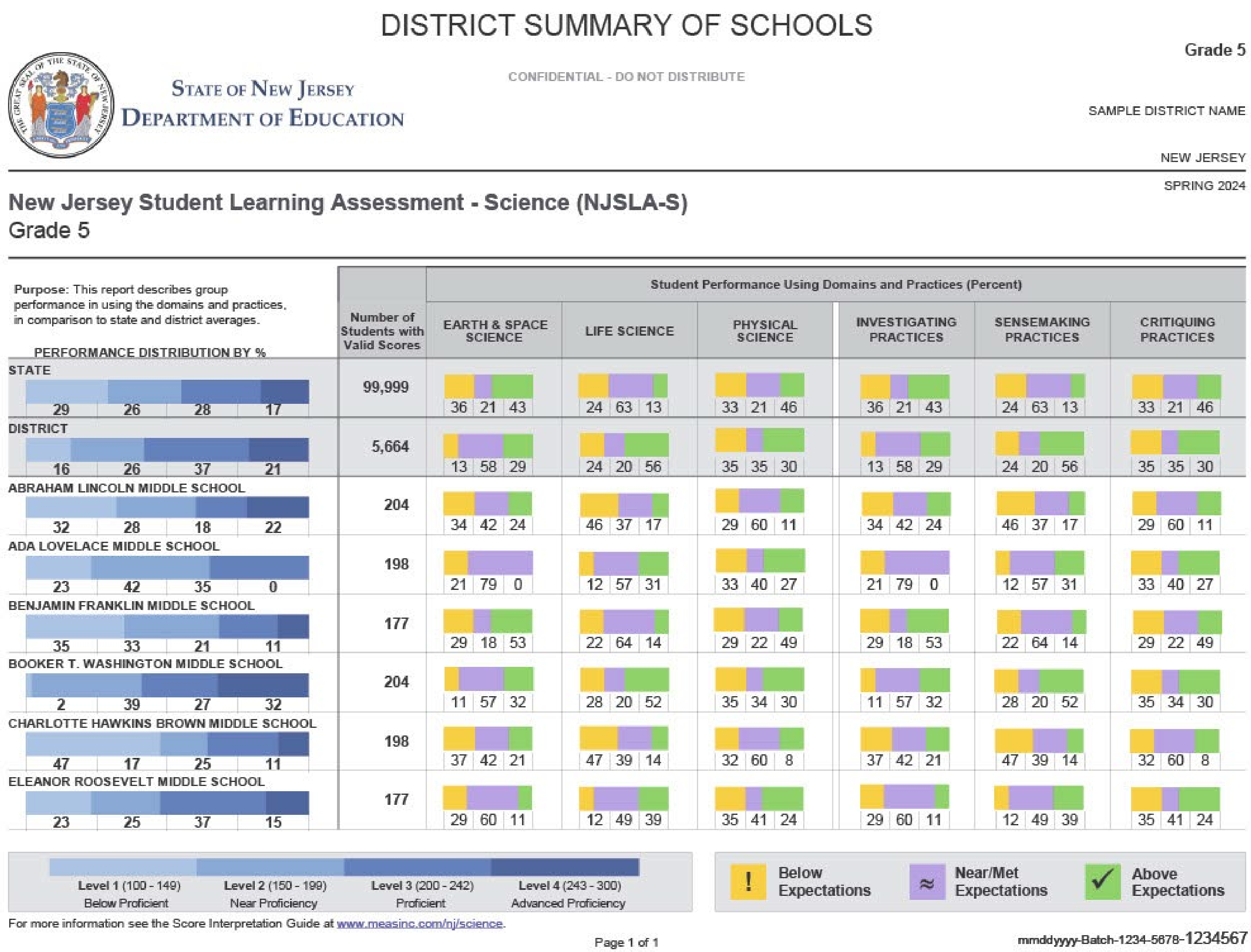


Figure 10.4.1. Sample School Performance-Level Summary Report–Domains and Practices



DISTRICT SUMMARY OF SCHOOLS

CONFIDENTIAL - DO NOT DISTRIBUTE

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

Grade 5

SAMPLE DISTRICT NAME

NEW JERSEY

SPRING 2024

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

Purpose: This report describes group performance in using the domains and practices, in comparison to state and district averages.

PERFORMANCE DISTRIBUTION BY %	Number of Students with Valid Scores	EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
		Student Performance Using Domains and Practices (Percent)					
STATE: 29 26 28 17	99,999	36 21 43	24 63 13	33 21 46	36 21 43	24 63 13	33 21 46
DISTRICT: 16 26 37 21	5,664	13 58 29	24 20 56	35 35 30	13 58 29	24 20 56	35 35 30
ABRAHAM LINCOLN MIDDLE SCHOOL: 32 28 18 22	204	34 42 24	46 37 17	29 60 11	34 42 24	46 37 17	29 60 11
ADA LOVELACE MIDDLE SCHOOL: 23 42 35 0	198	21 79 0	12 57 31	33 40 27	21 79 0	12 57 31	33 40 27
BENJAMIN FRANKLIN MIDDLE SCHOOL: 35 33 21 11	177	29 18 53	22 64 14	29 22 49	29 18 53	22 64 14	29 22 49
BOOKER T. WASHINGTON MIDDLE SCHOOL: 2 39 27 32	204	11 57 32	28 20 52	35 34 30	11 57 32	28 20 52	35 34 30
CHARLOTTE HAWKINS BROWN MIDDLE SCHOOL: 47 17 25 11	198	37 42 21	47 39 14	32 60 8	37 42 21	47 39 14	32 60 8
ELEANOR ROOSEVELT MIDDLE SCHOOL: 23 25 37 15	177	29 60 11	12 49 39	35 41 24	29 60 11	12 49 39	35 41 24

Level 1 (100 - 149) Below Proficient | Level 2 (150 - 199) Near Proficiency | Level 3 (200 - 242) Proficient | Level 4 (243 - 300) Advanced Proficiency

! Below Expectations | ≈ Near/Met Expectations | ✓ Above Expectations

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

Page 1 of 1

mmddyyyy-Batch-1234-5678-1234567

Figure 10.4.2. 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

Grade 5

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

CONFIDENTIAL - DO NOT DISTRIBUTE

SAMPLE DISTRICT NAME
SAMPLE SCHOOL NAME
NEW JERSEY
SPRING 2024

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 3	
					Level 1 Below Proficient		Level 2 Near Proficiency		Level 3 Proficient		Level 4 Advanced Proficiency			
					#	%	#	%	#	%	#	%	#	%
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	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%
School	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%
Gender														
Female	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%
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

mmddyyyy-Batch-1234-5678-1234567

Figure 10.5.1. Sample School Performance-Level Summary Report

DISTRICT PERFORMANCE LEVEL SUMMARY

Grade 5

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

CONFIDENTIAL - DO NOT DISTRIBUTE

SAMPLE DISTRICT NAME

NEW JERSEY

SPRING 2024

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	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%
Gender														
Female	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%
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

mmddyyyy-Batch-1234-5678-1234567

Figure 10.5.2. Sample District 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.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Beauducel, A., & Herzberg, P.Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186–203.

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.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: The Guilford Press.

Brown, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

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.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

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.

Garrido, L.E., Abad, F.J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via monte carlo simulation. *Psychological Methods, 21*(1), 93–111.

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.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jöreskog, K.G., & Sörbom, D. (1988). *LISERL 7. A guide to the program and applications* (2nd ed.). International Education Services.

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.

Kline, R.B. (2011). *Principles and practice of structural equation modeling* (3rd edition). The Guilford Press.

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.

Muthén, B.O. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Sage.

Muthén, B.O., du Toit, S.H.C, & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes [Unpublished manuscript]. https://www.statmodel.com/wlscv.shtml

National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts 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 (2024). *Spring 2024 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/2024_NJSLA_NJGPA_DTC_Training_Part1_V1_1.pdf

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

Nye, C.D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*(3), 548–570.

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.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage Publications, Inc.

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement, 80*(3), 421–445.

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/

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.

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.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation models with ordered categorical data: The story they tell depends on the estimation methods. *Behavioral Research Methods, 51*, 409–428.

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
CFA	Confirmatory Factor Analysis
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 Inc.
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

Abbreviation	Definition
NRC	National Research Council
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	Wallington Public School District	Bergen
2	Glen Rock Public Schools	Bergen
3	Saddle River School District	Bergen
4	Northern Burlington County Regional Schools District	Burlington
5	Lindenwold Public Schools	Camden
6	Camden City School District	Camden
7	Avalon School District	Cape May
8	Orange School District	Essex
9	Swedesboro-Woolwich School District	Gloucester
10	Jersey City Public Schools	Hudson
11	Readington Township School District	Hunterdon
12	Alexandria Township School District	Hunterdon
13	West-Windsor-Plainsboro Regional School District	Mercer
14	Metuchen Public School District	Middlesex
15	Freehold Township Public Schools	Monmouth
16	Rumson School District	Monmouth
17	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
5	Newark School District	Essex
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
11	Matawan Aberdeen Regional School District	Monmouth
12	Matawan Aberdeen Regional School District	Monmouth
13	Mount Olive Township Schools	Morris
14	Elmwood Park School District	Ocean
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
4	Delran Township School District	Burlington
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	Sayreville Public Schools	Middlesex
13	Mount Olive Township School District	Morris
14	Point Pleasant Borough School District	Ocean
15	Passaic City Schools	Passaic
16	Patterson Public Schools	Passaic
17	Pennsville School District	Salem
18	Somerset County Vocational & Technical School District	Somerset
19	Lenape Valley Regional School District	Sussex
20	Elizabeth Public Schools	Union
21	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	Pemberton Township School District	Burlington
4	Cherry Hill Public Schools	Camden
5	Millburn Township Public Schools	Essex
6	Jersey City Global Charter Schools	Hudson
7	Metuchen School District	Middlesex
8	Middlesex Borough School District	Middlesex
9	Freehold Township School District	Monmouth
10	Long Branch Public School District	Monmouth
11	Morris County Vocational School District	Morris
12	Clifton Public Schools	Passaic
13	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 Measurement Incorporated 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.

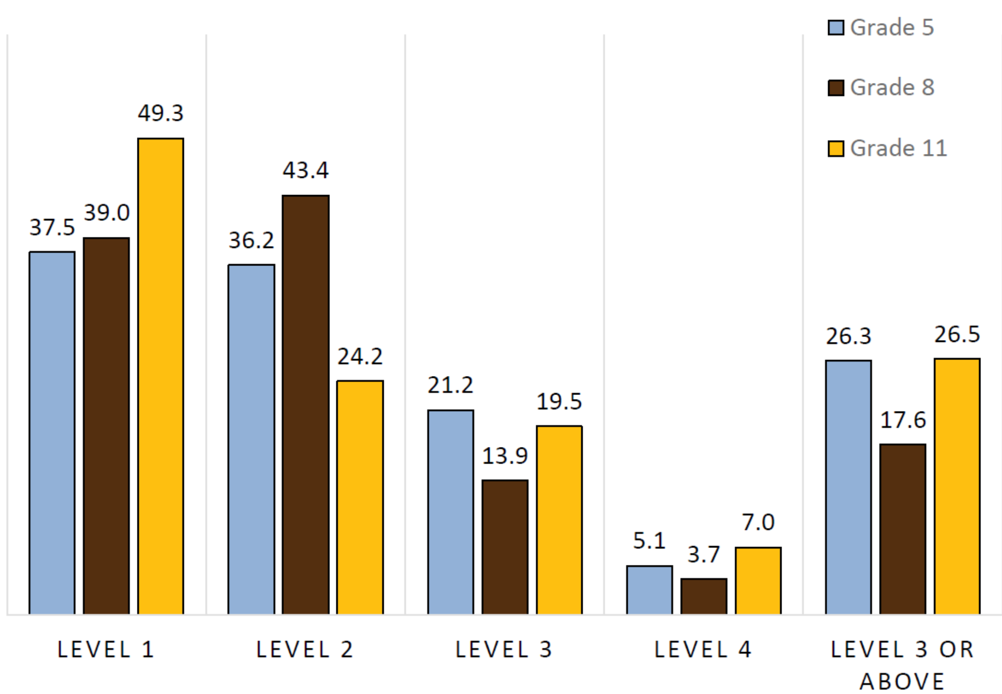


Figure ES.1. Percentages of Students Classified at Each Level after Round 3

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**
1905M069_01	1	TE	UMCT	PS1.B	S, P & Q	Physical	Investigating	0.400	0.45	0.55	172
1905M069_02	1	TE	AID	PS1.B	PAT	Physical	Sensemaking	1.131	0.31	0.54	65
1905M069_05	1	TE	UMCT	PS1.B	PAT	Physical	Investigating	1.569	0.24	0.54	86
1905M076_01	1	MC	EAE	PS3.A	E & M	Physical	Critiquing	–0.352	0.61	0.24	94
1905M076_03	1	TE	CEDS	PS2.A	PAT	Physical	Sensemaking	0.472	0.47	0.52	91
1905M076_05	1	TE	CEDS	PS3.B	PAT	Physical	Sensemaking	1.204	0.32	0.52	57
2005B008_01	1	TE	DUM	PS3.D	C & E	Physical	Sensemaking	–0.225	0.57	0.57	74
2005B008_02	1	TE	OECI	PS3.B	S & SM	Physical	Critiquing	1.582	0.23	0.54	82
2005B008_03	1	TE	AQDP	PS3.B	C & E	Physical	Investigating	1.320	0.28	0.45	59
2005B008_04	1	TE	PACI	PS3.B	E & M	Physical	Investigating	–0.256	0.58	0.40	39
2005B008_05	3	CR	CEDS	PS3.B	S & SM	Physical	Sensemaking	0.300	0.45	0.47	384
2005M015_02	1	MC	AID	ESS3.B	C & E	Earth and Space	Sensemaking	–0.415	0.61	0.39	87
2005M015_04	1	MC	CEDS	ESS3.B	C & E	Earth and Space	Sensemaking	1.009	0.33	0.36	67
2005M015_05	1	TE	CEDS	ESS3.B	SF	Earth and Space	Sensemaking	0.934	0.34	0.36	71
2005M015_06	1	MC	AQDP	ESS3.B	SC	Earth and Space	Investigating	1.713	0.22	0.10	52
2005M015_07	1	TE	UMCT	ESS3.B	PAT	Earth and Space	Investigating	0.696	0.39	0.58	85
2005M015_10	1	TE	DUM	ESS3.B	S & SM	Earth and Space	Sensemaking	0.875	0.35	0.38	55
2105M003_01	1	TE	AQDP	LS3.A	PAT	Life	Investigating	1.183	0.30	0.49	98
2105M003_07	1	MC	CEDS	LS3.A	PAT	Life	Sensemaking	–0.783	0.68	0.40	35
2105M003_09	1	TE	EAE	LS3.A	PAT	Life	Critiquing	1.315	0.28	0.54	54
2205B000_02	1	TE	AQDP	LS2.C	C & E	Life	Investigating	2.044	0.17	0.40	99
2205B000_03	1	MC	PACI	LS2.C	PAT	Life	Investigating	0.409	0.44	0.40	51
2205B000_04	1	TE	PACI	LS4.C	C & E	Life	Investigating	0.409	0.44	0.62	42
2205B000_07	4	CR	EAE	LS2.C	C & E	Life	Critiquing	0.793	0.35	0.74	376
2205B000_10	1	MC	AID	LS2.C	PAT	Life	Sensemaking	1.114	0.31	0.27	67

Grade 5											
UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	Rasch B	*p-value*	*rpb*	Median Time
2205B006_01	1	MC	AQDP	ESS1.C	PAT	Earth and Space	Investigating	0.025	0.52	0.30	93
2205B006_03	1	TE	EAE	ESS1.C	PAT	Earth and Space	Critiquing	1.140	0.31	0.31	92
2205B006_06	1	TE	EAE	ESS1.C	PAT	Earth and Space	Critiquing	2.114	0.16	0.29	105
2205B009_01	1	TE	DUM	ESS2.A	S & SM	Earth and Space	Sensemaking	–0.298	0.58	0.50	53
2205B009_02	1	TE	AQDP	ESS2.A	S & SM	Earth and Space	Investigating	1.339	0.28	0.46	46
2205B009_03	1	TE	PACI	ESS2.A	S & SM	Earth and Space	Investigating	1.430	0.26	0.24	44
2205B009_04	1	TE	DUM	ESS2.A	S & SM	Earth and Space	Sensemaking	0.372	0.45	0.43	38
2205B009_05	4	CR	EAE	ESS2.A	S & SM	Earth and Space	Critiquing	1.091	0.31	0.64	395
2205M000_01	1	TE	CEDS	ESS2.E	SF	Earth and Space	Sensemaking	1.123	0.31	0.22	70
2205M000_02	1	TE	CEDS	ESS2.E	C & E	Earth and Space	Sensemaking	0.752	0.38	0.31	96
2205M000_03	1	MC	AQDP	ESS2.E	C & E	Earth and Space	Investigating	0.457	0.44	0.43	72
2205M000_04	1	TE	EAE	ESS2.E	SC	Earth and Space	Critiquing	0.804	0.37	0.30	108
2205M004_03	1	TE	CEDS	LS1.D	SF	Life	Sensemaking	0.265	0.44	0.59	157
2205M004_05	1	TE	EAE	LS2.C	C & E	Life	Critiquing	1.162	0.28	0.40	117
2205M004_07	1	TE	DUM	LS1.D	SF	Life	Sensemaking	0.273	0.45	0.30	72
2205M006_01	1	MC	AQDP	LS2.B	S & SM	Life	Investigating	0.030	0.51	0.42	106
2205M006_02	1	TE	EAE	LS2.B	S & SM	Life	Critiquing	0.591	0.36	0.23	101
2205M006_05	1	TE	DUM	LS2.B	S & SM	Life	Sensemaking	0.103	0.49	0.43	70
2205M022_01	1	MC	AQDP	PS4.B	S & SM	Physical	Investigating	0.433	0.45	0.44	74
2205M022_03	1	MC	PACI	PS4.B	SF	Physical	Investigating	0.362	0.47	0.48	54
2205M022_05	1	TE	PACI	PS4.B	S, P & Q	Physical	Investigating	0.447	0.43	0.26	41
2305B006_01	1	TE	DUM	PS4.C	S & SM	Physical	Sensemaking	0.508	0.43	0.45	95
2305B006_03	1	MC	AQDP	PS4.C	C & E	Physical	Investigating	0.812	0.37	0.41	61
2305B006_04	1	TE	EAE	PS4.A	C & E	Physical	Critiquing	0.875	0.35	0.22	55
2305M008_01	1	MC	AQDP	LS2.C	PAT	Life	Investigating	0.593	0.41	0.25	95
2305M008_03	1	TE	EAE	LS4.B	C & E	Life	Critiquing	1.459	0.25	0.16	65
2305M008_04	1	TE	CEDS	LS4.A	PAT	Life	Sensemaking	0.726	0.38	0.38	78

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**
2008B007_03	1	TE	DUM	PS3.A	S & SM	Physical	Sensemaking	0.797	0.24	0.30	34
2008B007_07	4	CR	EAE	PS3.A	C & E	Physical	Critiquing	0.713	0.21	0.68	310
2008B007_09	1	TE	AID	PS3.A	C & E	Physical	Sensemaking	–0.954	0.58	0.39	45
2008B007_11	1	TE	AQDP	PS3.B	PAT	Physical	Investigating	0.513	0.29	0.39	69
2008M000_01	1	TE	DUM	LS1.A	SF	Life	Sensemaking	0.640	0.26	0.21	46
2008M000_02	1	TE	OECI	LS1.A	SF	Life	Critiquing	1.107	0.17	0.47	85
2008M000_04	1	MC	AQDP	LS1.C	E & M	Life	Investigating	0.291	0.30	0.26	37
2108B006_03	1	TE	AQDP	ESS2.C	S & SM	Earth and Space	Investigating	1.232	0.16	0.30	91
2108B006_06	1	TE	CEDS	ESS1.B	PAT	Earth and Space	Sensemaking	0.543	0.29	0.38	79
2108B006_09	1	TE	OECI	ESS2.C	PAT	Earth and Space	Critiquing	0.175	0.38	0.53	33
2108B007_03	1	MC	AQDP	PS3.A	S, P & Q	Physical	Investigating	0.980	0.21	0.14	54
2108B007_07	1	TE	CEDS	PS1.B	PAT	Physical	Sensemaking	0.988	0.20	0.24	35
2108B007_08	1	TE	OECI	PS1.B	PAT	Physical	Critiquing	–1.045	0.58	0.41	95
2108M000_02	1	TE	AQDP	ESS3.C	C & E	Earth and Space	Investigating	0.213	0.35	0.43	89
2108M000_03	1	TE	AID	ESS3.D	C & E	Earth and Space	Sensemaking	0.698	0.26	0.49	47
2108M000_06	1	TE	EAE	ESS3.D	SC	Earth and Space	Critiquing	0.893	0.23	0.24	66
2208B003_01	1	TE	AID	LS2.C	C & E	Life	Sensemaking	–0.204	0.41	0.48	63
2208B003_05	1	TE	CEDS	LS4.D	E & M	Life	Sensemaking	–0.107	0.42	0.49	48
2208B003_07	1	TE	PACI	LS2.C	PAT	Life	Investigating	1.196	0.20	0.48	65
2208B003_09	1	TE	EAE	LS2.C	C & E	Life	Critiquing	0.590	0.29	0.23	44
2208B003_11	3	CR	OECI	LS2.C	SC	Life	Critiquing	–0.220	0.45	0.59	327
2208M016_03	1	MC	EAE	LS4.B	C & E	Life	Critiquing	0.207	0.35	0.38	120
2208M016_04	1	TE	PACI	LS3.A	C & E	Life	Investigating	–0.993	0.59	0.53	53
2208M016_05	1	TE	UMCT	LS3.B	PAT	Life	Investigating	0.402	0.31	0.07	50
2208M016_08	1	MC	OECI	LS3.A	C & E	Life	Critiquing	–0.190	0.42	0.21	53
2208M021_03	1	TE	OECI	ESS1.C	PAT	Earth and Space	Critiquing	0.581	0.28	0.38	106
2208M021_05	1	TE	AID	ESS1.C	C & E	Earth and Space	Sensemaking	0.104	0.38	0.38	55

Grade 8											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p-value***	***rpb***	**Median Time**
2208M021_09	1	MC	AID	ESS1.C	PAT	Earth and Space	Sensemaking	0.443	0.32	0.34	70
2208M021_10	1	TE	EAE	ESS1.C	PAT	Earth and Space	Critiquing	0.065	0.39	0.57	48
2208M051_13	1	TE	AQDP	PS4.A	C & E	Physical	Investigating	0.091	0.38	0.39	83
2208M051_16	1	TE	EAE	PS3.B	PAT	Physical	Critiquing	–0.379	0.45	0.45	53
2208M051_17	1	TE	OECI	PS4.A	C & E	Physical	Critiquing	0.344	0.32	0.44	67
2308B003_01	1	TE	DUM	ESS2.A	S & SM	Earth and Space	Sensemaking	0.044	0.38	0.55	60
2308B003_03	1	TE	AQDP	ESS2.B	S & SM	Earth and Space	Investigating	0.811	0.24	0.41	54
2308B003_05	4	CR	OECI	ESS2.A	SC	Earth and Space	Critiquing	0.064	0.38	0.62	266
2308B003_07	1	TE	PACI	ESS2.B	S, P & Q	Earth and Space	Investigating	0.937	0.22	0.38	32
2308B003_11	1	TE	EAE	ESS3.B	S & SM	Earth and Space	Critiquing	–0.704	0.53	0.46	39
2308B006_01	1	TE	CEDS	LS4.C	C & E	Life	Sensemaking	1.133	0.20	0.33	73
2308B006_07	1	TE	CEDS	LS2.B	C & E	Life	Sensemaking	1.027	0.21	0.18	41
2308B006_08	1	TE	AID	LS2.A	PAT	Life	Sensemaking	0.837	0.24	0.35	37
2308M000_04	1	MC	AID	LS3.A	PAT	Life	Sensemaking	–0.232	0.43	0.34	78
2308M000_06	1	TE	EAE	LS3.A	C & E	Life	Critiquing	0.819	0.24	0.45	42
2308M000_08	1	MC	AQDP	LS3.B	PAT	Life	Investigating	0.495	0.29	0.54	47
2308M017_01	1	TE	CEDS	PS2.A	C & E	Physical	Sensemaking	0.406	0.31	0.39	106
2308M017_02	1	MC	UMCT	PS3.A	S, P & Q	Physical	Investigating	–0.150	0.42	0.43	71
2308M017_05	1	MC	OECI	PS2.A	SF	Physical	Critiquing	–1.265	0.64	0.23	51
2308M020_05	1	TE	EAE	PS4.B	E & M	Physical	Critiquing	0.212	0.35	0.41	91
2308M020_09	1	TE	UMCT	PS3.B	PAT	Physical	Investigating	0.410	0.31	0.36	54
2308M020_11	1	MC	AQDP	PS4.A	C & E	Physical	Investigating	–1.025	0.60	0.33	39
2308M023_02	1	MC	EAE	ESS3.C	PAT	Earth and Space	Critiquing	–0.555	0.50	0.33	88
2308M023_03	1	TE	UMCT	ESS3.C	S, P & Q	Earth and Space	Investigating	0.833	0.24	0.52	99
2308M023_07	1	TE	CEDS	ESS3.C	C & E	Earth and Space	Sensemaking	0.568	0.28	0.15	51
2308M027_01	1	TE	AQDP	PS4.B	C & E	Physical	Investigating	0.189	0.35	0.31	32
2308M027_02	1	TE	PACI	PS4.B	C & E	Physical	Investigating	0.592	0.28	0.26	57
2308M027_03	1	TE	EAE	PS4.B	E & M	Physical	Critiquing	0.778	0.25	0.43	47

Grade 8											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p-value***	***rpb***	**Median Time**
2308M027_05	1	TE	DUM	PS4.B	PAT	Physical	Sensemaking	0.340	0.32	0.61	53
2308M028_04	1	TE	DUM	ESS2.C	SC	Earth and Space	Sensemaking	0.628	0.27	0.38	70
2308M028_05	1	MC	AQDP	ESS3.B	C & E	Earth and Space	Investigating	0.253	0.34	0.16	65
2308M028_06	1	TE	OECI	ESS2.C	C & E	Earth and Space	Critiquing	0.133	0.36	0.26	103
2308M028_09	1	MC	EAE	ESS3.D	S & SM	Earth and Space	Critiquing	–0.388	0.46	0.31	87
2308M049_01	1	TE	DUM	LS4.B	C & E	Life	Sensemaking	0.605	0.28	0.52	74
2308M049_03	1	TE	AID	PS4.B	SF	Life	Sensemaking	0.295	0.33	0.50	88
2308M049_05	1	TE	AID	LS2.A	E & M	Life	Sensemaking	0.634	0.27	0.39	98
2308M049_07	1	MC	AQDP	LS4.B	SF	Life	Investigating	–0.391	0.47	0.17	54

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**
2011M003_01	1	TE	AID	LS2.C	SC	Life	Sensemaking	0.955	0.29	0.47	71
2011M003_03	1	MC	OECI	LS2.C	SC	Life	Critiquing	1.147	0.26	0.32	63
2011M003_04	1	MC	AQDP	LS2.C	SC	Life	Investigating	0.788	0.33	0.40	28
2011M003_05	1	TE	EAE	LS2.C	C & E	Life	Critiquing	0.497	0.39	0.53	67
2011M003_06	1	MC	PACI	LS2.C	C & E	Life	Investigating	0.034	0.47	0.42	87
2011M028_01	1	MC	AQDP	LS4.D	C & E	Life	Investigating	–0.682	0.61	0.27	56
2011M028_03	1	TE	CEDS	ESS3.C	S & SM	Earth and Space	Sensemaking	0.237	0.42	0.40	51
2011M028_04	1	MC	EAE	LS4.D	S & SM	Life	Critiquing	1.348	0.22	0.24	60
2011M028_05	1	TE	PACI	ESS3.C	C & E	Earth and Space	Investigating	0.025	0.46	0.38	31
2011M031_03	1	TE	CEDS	ESS1.C	SC	Earth and Space	Sensemaking	1.447	0.21	0.32	103
2011M031_04	1	TE	EAE	ESS2.B	SC	Earth and Space	Critiquing	0.798	0.31	0.26	66
2011M031_05	1	MC	DUM	ESS1.C	S & SM	Earth and Space	Sensemaking	0.756	0.32	0.24	66
2011M031_06	1	MC	AQDP	ESS2.B	S & SM	Earth and Space	Investigating	0.050	0.46	0.27	42
2011M054_04	1	TE	PACI	LS3.B	SF	Life	Investigating	1.449	0.21	0.43	208
2011M054_06	1	MC	OECI	LS3.B	S, P & Q	Life	Critiquing	0.139	0.44	0.49	84
2011M054_08	1	TE	EAE	LS3.B	PAT	Life	Critiquing	0.327	0.40	0.34	72
2011M071_01	1	MC	AID	PS3.D	S, P & Q	Physical	Sensemaking	–0.598	0.52	0.50	32
2011M071_03	1	TE	EAE	PS3.D	E & M	Physical	Critiquing	0.569	0.30	0.50	43
2011M071_04	1	MC	DUM	PS3.D	S & SM	Physical	Sensemaking	–0.262	0.47	0.28	53
2011M071_05	1	MC	UMCT	PS3.D	E & M	Physical	Investigating	–0.311	0.47	0.41	46
2111B002_01	1	TE	EAE	LS4.C	C & E	Life	Critiquing	0.732	0.32	0.61	57
2111B002_03	1	TE	AID	LS2.A	SC	Physical	Sensemaking	0.278	0.41	0.40	69
2111B002_07	1	TE	AQDP	LS2.A	SC	Life	Investigating	0.760	0.32	0.37	29
2111B002_10	1	TE	CEDS	LS2.A	C & E	Life	Sensemaking	1.177	0.25	0.49	68
2111B002_11	3	CR	PACI	LS2.C	C & E	Life	Investigating	–0.096	0.49	0.64	269
2111M004_02	1	TE	AID	LS4.C	S, P & Q	Life	Sensemaking	2.137	0.14	0.28	59
2111M004_03	1	MC	UMCT	LS2.A	C & E	Life	Investigating	1.189	0.32	0.45	84

Grade 11											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p-value***	***rpb***	**Median Time**
2111M004_05	1	MC	CEDS	LS1.A	PAT	Life	Sensemaking	–0.260	0.55	0.29	105
2111M004_06	1	TE	EAE	LS4.C	C & E	Life	Critiquing	1.328	0.27	0.37	80
2111M020_01	1	MC	UMCT	PS2.A	S, P & Q	Physical	Investigating	–0.593	0.59	0.41	63
2111M020_03	1	TE	CEDS	PS2.A	SF	Physical	Sensemaking	1.345	0.22	0.36	64
2111M020_05	1	TE	EAE	PS2.A	PAT	Physical	Critiquing	–0.057	0.48	0.54	45
2211B001_01	1	TE	AQDP	PS3.B	C & E	Physical	Investigating	0.922	0.29	0.35	40
2211B001_03	1	TE	PACI	PS3.B	C & E	Physical	Investigating	1.071	0.27	0.33	36
2211B001_05	1	TE	OECI	PS3.A	S & SM	Physical	Critiquing	0.039	0.46	0.60	27
2211B001_07	1	MC	UMCT	PS3.B	S & SM	Physical	Investigating	0.734	0.32	0.41	39
2211B001_10	4	CR	EAE	PS3.B	E & M	Physical	Critiquing	1.417	0.19	0.60	219
2211B006_02	1	TE	AQDP	ESS3.B	C & E	Earth and Space	Investigating	0.825	0.33	0.44	72
2211B006_05	1	TE	PACI	ESS3.B	SF	Earth and Space	Investigating	0.971	0.29	0.48	47
2211B006_06	1	TE	PACI	ESS3.B	C & E	Earth and Space	Investigating	–1.220	0.70	0.44	38
2211B006_09	4	CR	EAE	ESS3.B	C & E	Earth and Space	Critiquing	–0.242	0.53	0.60	254
2211B006_12	1	TE	EAE	PS1.A	SF	Physical	Critiquing	0.526	0.35	0.55	43
2211M008_01	1	TE	AID	ESS1.B	S & SM	Earth and Space	Sensemaking	1.421	0.19	0.40	82
2211M008_03	1	MC	EAE	ESS1.B	S & SM	Earth and Space	Critiquing	0.242	0.42	0.27	46
2211M008_07	1	TE	PACI	ESS1.B	S & SM	Earth and Space	Investigating	0.896	0.28	0.24	53
2211M016_01	1	TE	DUM	PS3.A	E & M	Physical	Sensemaking	0.548	0.36	0.38	123
2211M016_02	1	MC	UMCT	PS3.B	S & SM	Physical	Investigating	–0.239	0.52	0.33	90
2211M016_06	1	TE	OECI	ESS3.C	C & E	Earth and Space	Critiquing	0.711	0.33	0.30	55
2211M016_08	1	TE	CEDS	PS3.D	E & M	Physical	Sensemaking	1.537	0.20	0.43	87
2211M022_01	1	MC	AQDP	ESS2.E	C & E	Earth and Space	Investigating	0.542	0.36	0.49	42
2211M022_02	1	TE	DUM	ESS2.E	S & SM	Earth and Space	Sensemaking	0.895	0.30	0.50	36
2211M022_08	1	MC	EAE	ESS2.E	C & E	Earth and Space	Critiquing	0.748	0.32	0.22	36
2311M011_03	1	TE	EAE	PS2.B	C & E	Physical	Critiquing	0.990	0.28	0.48	64
2311M011_05	1	MC	DUM	ESS1.B	S & SM	Earth and Space	Sensemaking	–0.699	0.61	0.26	39
2311M011_07	1	MC	AQDP	ESS1.B	C & E	Earth and Space	Investigating	–0.355	0.54	0.42	22

Grade 11											
UIN	**Points**	**Item Type**	**SEP**	**DCI**	**CCC**	**Domain**	**Practice**	**Rasch B**	***p-value***	***rpb***	**Median Time**
2311M016_01	1	TE	CEDS	PS4.A	S, P & Q	Physical	Sensemaking	0.216	0.42	0.35	147
2311M016_02	1	TE	EAE	PS4.A	S, P & Q	Physical	Critiquing	1.232	0.24	0.55	81
2311M016_04	1	MC	AID	LS4.C	C & E	Life	Sensemaking	0.132	0.44	0.37	63
2311M016_07	1	TE	UMCT	PS4.A	PAT	Physical	Investigating	1.444	0.21	0.19	73
2311M021_02	1	TE	EAE	ESS2.A	SC	Earth and Space	Critiquing	1.657	0.18	0.35	89
2311M021_04	1	MC	AID	ESS2.A	C & E	Earth and Space	Sensemaking	0.344	0.40	0.13	47
2311M021_06	1	MC	PACI	ESS2.A	SC	Earth and Space	Investigating	0.223	0.42	0.30	26
2311M023_01	1	MC	UMCT	LS1.C	S, P & Q	Life	Investigating	–0.671	0.60	0.36	55
2311M023_03	1	MC	DUM	LS1.A	PAT	Life	Sensemaking	0.077	0.45	0.38	66
2311M023_06	1	TE	EAE	LS1.C	C & E	Life	Critiquing	1.687	0.18	0.35	57
2311M026_01	1	MC	DUM	PS1.C	E & M	Physical	Sensemaking	0.358	0.40	0.47	98
2311M026_02	1	MC	OECI	PS1.C	E & M	Physical	Critiquing	0.822	0.31	0.16	56
2311M026_03	1	TE	CEDS	PS1.C	SC	Physical	Sensemaking	0.082	0.45	0.55	58
2311M026_05	1	TE	PACI	PS1.C	S & SM	Physical	Investigating	1.822	0.16	0.27	57
2311M026_06	1	MC	PACI	PS1.C	S, P & Q	Physical	Investigating	0.552	0.36	0.40	53

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	4	0.00	0.01	0.00	0.00	0.01	0.00	0.00
1	100	36	0.04	0.03	0.04	0.02	0.11	0.05	0.01
2	100	162	0.17	0.15	0.19	0.03	0.39	0.24	0.08
3	100	537	0.56	0.45	0.66	0.12	1.10	0.86	0.25
4	100	1,372	1.42	1.26	1.58	0.33	2.61	2.24	0.64
5	100	2,856	2.96	2.61	3.30	0.70	5.44	4.60	1.31
6	100	5,044	5.23	4.75	5.69	1.17	9.46	8.16	2.31
7	105	7,899	8.19	7.48	8.88	1.98	14.54	12.70	3.76
8	111	11,207	11.62	10.74	12.47	2.88	20.29	17.83	5.60
9	118	14,750	15.29	14.35	16.21	4.02	25.95	23.35	7.66
10	123	18,389	19.06	18.19	19.91	5.23	31.84	28.70	10.00
11	128	21,791	22.59	21.84	23.32	6.27	36.64	33.92	12.28
12	133	25,131	26.05	25.48	26.61	7.69	41.75	38.62	14.64
13	138	28,338	29.38	28.94	29.80	8.94	46.14	43.10	17.17
14	142	31,217	32.36	32.17	32.55	10.33	49.88	47.04	19.45
15	146	34,093	35.34	35.37	35.32	11.69	53.61	50.78	21.98
16	150	36,919	38.27	38.47	38.09	12.97	57.05	54.30	24.74
17	153	39,699	41.15	41.51	40.82	14.58	60.28	57.66	27.46
18	157	42,276	43.83	44.40	43.28	16.19	63.15	60.78	30.01
19	160	44,711	46.35	46.99	45.74	17.65	65.79	63.50	32.63
20	163	47,044	48.77	49.58	47.98	19.27	68.28	66.02	35.16
21	167	49,274	51.08	52.01	50.18	20.79	70.61	68.37	37.65
22	170	51,602	53.49	54.62	52.41	22.56	73.06	70.71	40.33
23	173	53,803	55.78	57.00	54.60	24.50	75.37	72.86	42.84
24	176	55,991	58.04	59.38	56.75	26.26	77.42	74.99	45.46
25	178	58,169	60.30	61.83	58.82	28.32	79.13	76.98	48.24
26	181	60,273	62.48	64.08	60.94	30.43	80.91	78.87	50.83
27	184	62,289	64.57	66.29	62.91	32.44	82.60	80.65	53.39
28	187	64,313	66.67	68.49	64.92	34.88	84.03	82.32	56.06
29	190	66,233	68.66	70.51	66.88	36.99	85.36	83.80	58.71
30	192	68,164	70.66	72.47	68.92	39.34	86.74	85.34	61.27
31	195	69,862	72.42	74.26	70.65	41.64	87.98	86.59	63.51
32	200	71,941	74.58	76.44	72.78	44.48	89.30	88.05	66.38
33	201	73,849	76.56	78.46	74.72	47.20	90.46	89.45	68.98
34	203	75,601	78.37	80.26	76.55	49.94	91.52	90.53	71.45
35	206	77,333	80.17	82.02	78.38	53.03	92.62	91.53	73.81
36	209	78,996	81.89	83.67	80.18	56.14	93.47	92.53	76.11
37	212	80,648	83.61	85.20	82.06	59.03	94.28	93.46	78.50
38	215	82,221	85.24	86.75	83.77	62.01	95.25	94.26	80.72
39	218	83,723	86.79	88.20	85.43	64.77	95.91	95.09	82.86
40	222	85,164	88.29	89.59	87.03	67.65	96.57	95.80	84.92
41	225	86,547	89.72	90.91	88.58	70.67	97.13	96.51	86.82

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
42	228	87,868	91.09	92.17	90.05	74.00	97.58	97.17	88.53
43	232	89,133	92.40	93.36	91.47	77.19	98.04	97.69	90.30
44	236	90,254	93.56	94.44	92.71	80.09	98.36	98.15	91.85
45	243	91,294	94.64	95.43	93.88	83.06	98.76	98.57	93.17
46	244	92,203	95.58	96.32	94.87	85.52	99.01	98.90	94.44
47	248	93,092	96.51	97.14	95.89	88.16	99.34	99.20	95.63
48	252	93,836	97.28	97.81	96.76	90.42	99.46	99.45	96.65
49	257	94,529	98.00	98.36	97.64	92.92	99.65	99.59	97.55
50	262	95,060	98.55	98.82	98.28	94.66	99.76	99.70	98.26
51	268	95,486	98.99	99.19	98.79	96.26	99.84	99.80	98.78
52	274	95,819	99.33	99.45	99.22	97.37	99.91	99.87	99.23
53	281	96,053	99.57	99.63	99.52	98.37	99.96	99.93	99.49
54	288	96,230	99.76	99.78	99.74	99.03	99.98	99.97	99.71
55	297	96,352	99.88	99.89	99.88	99.51	99.99	99.98	99.88
56	300	96,416	99.95	99.95	99.95	99.79	100.00	99.99	99.95
57	300	96,447	99.98	99.98	99.98	99.92	100.00	100.00	99.99
58	300	96,460	100.00	100.00	100.00	99.98	100.00	100.00	100.00
59	300	96,463	100.00	100.00	100.00	100.00	100.00	100.00	100.00
60	300	96,463	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	2	0.00	0.00	0.00	0.00	0.00	0.01	0.00
1	100	27	0.03	0.02	0.03	0.01	0.03	0.06	0.00
2	100	84	0.08	0.07	0.10	0.02	0.15	0.15	0.03
3	100	233	0.23	0.18	0.29	0.05	0.47	0.34	0.10
4	100	616	0.62	0.48	0.75	0.14	1.31	0.88	0.27
5	100	1,418	1.42	1.17	1.66	0.28	2.99	2.06	0.62
6	100	2,747	2.76	2.40	3.10	0.48	5.44	4.10	1.25
7	107	4,753	4.77	4.29	5.23	0.95	9.08	7.10	2.28
8	113	7,392	7.42	6.79	8.02	1.50	13.71	11.08	3.61
9	118	10,519	10.55	9.68	11.39	2.11	18.98	15.69	5.43
10	123	14,036	14.08	13.02	15.10	2.98	24.81	20.93	7.42
11	127	17,757	17.81	16.62	18.96	4.02	30.62	26.46	9.66
12	131	21,536	21.60	20.39	22.77	5.28	36.51	31.86	12.06
13	135	25,302	25.38	24.15	26.56	6.45	42.02	37.07	14.70
14	139	29,008	29.10	27.99	30.17	7.78	46.93	42.19	17.47
15	142	32,632	32.74	31.80	33.64	9.29	51.51	46.97	20.44
16	145	35,946	36.06	35.40	36.70	10.72	55.67	51.28	23.16
17	150	39,121	39.24	38.81	39.68	12.18	59.45	55.30	25.90
18	151	42,149	42.28	42.09	42.49	13.63	62.70	59.00	28.72
19	154	45,134	45.28	45.20	45.37	15.28	65.91	62.42	31.70
20	157	47,921	48.07	48.26	47.92	16.86	68.62	65.41	34.70
21	160	50,519	50.68	50.99	50.41	18.56	71.16	68.25	37.40
22	162	53,094	53.26	53.74	52.83	20.46	73.70	70.80	40.24
23	165	55,489	55.66	56.26	55.12	22.17	75.72	73.25	42.97
24	168	57,759	57.94	58.62	57.32	23.97	77.89	75.46	45.53
25	170	59,969	60.16	60.99	59.40	25.90	79.67	77.45	48.16
26	172	62,154	62.35	63.32	61.44	28.18	81.34	79.29	50.80
27	175	64,203	64.41	65.51	63.37	30.19	83.03	81.09	53.25
28	177	66,223	66.43	67.64	65.30	32.42	84.30	82.72	55.87
29	180	68,100	68.32	69.68	67.04	34.61	85.48	84.21	58.26
30	182	69,928	70.15	71.53	68.85	36.83	86.66	85.55	60.62
31	184	71,712	71.94	73.36	70.60	39.17	87.94	86.84	62.86
32	186	73,434	73.67	75.09	72.33	41.40	88.99	88.02	65.13
33	189	75,069	75.31	76.74	73.96	43.49	89.99	89.18	67.30
34	191	76,688	76.93	78.37	75.59	45.82	90.99	90.21	69.48
35	193	78,269	78.52	79.99	77.13	48.19	91.83	91.22	71.63
36	195	79,806	80.06	81.48	78.73	50.61	92.65	92.15	73.68
37	198	80,964	81.22	82.65	79.88	52.48	93.37	92.83	75.23
38	200	82,552	82.81	84.28	81.43	55.01	94.17	93.71	77.49
39	202	83,897	84.16	85.59	82.82	57.30	94.78	94.48	79.30
40	204	85,103	85.37	86.79	84.04	59.61	95.31	95.13	80.96
41	207	86,294	86.57	87.97	85.25	61.96	95.87	95.65	82.60
42	209	87,481	87.76	89.08	86.52	64.49	96.31	96.17	84.26

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
43	211	88,528	88.81	90.04	87.65	66.60	96.76	96.65	85.68
44	213	89,543	89.83	90.96	88.77	68.82	97.14	97.07	87.09
45	216	90,564	90.85	91.89	89.88	71.15	97.59	97.48	88.47
46	218	91,519	91.81	92.76	90.92	73.44	97.90	97.85	89.78
47	220	92,402	92.69	93.64	91.81	75.76	98.21	98.12	91.00
48	223	93,260	93.55	94.46	92.71	77.92	98.46	98.43	92.14
49	225	94,061	94.36	95.19	93.59	80.09	98.78	98.63	93.24
50	228	94,787	95.09	95.87	94.35	82.15	98.97	98.85	94.22
51	231	95,466	95.77	96.50	95.08	84.11	99.16	99.05	95.08
52	233	96,077	96.38	97.05	95.75	85.92	99.24	99.25	95.88
53	236	96,648	96.95	97.53	96.41	87.95	99.36	99.37	96.59
54	238	97,183	97.49	97.96	97.05	89.76	99.50	99.51	97.24
55	241	97,635	97.94	98.36	97.55	91.43	99.62	99.61	97.76
56	244	98,045	98.35	98.72	98.01	93.04	99.71	99.69	98.24
57	247	98,390	98.70	99.01	98.41	94.35	99.76	99.76	98.67
58	251	98,684	99.00	99.25	98.75	95.57	99.84	99.82	98.98
59	254	98,943	99.26	99.45	99.08	96.63	99.87	99.87	99.26
60	258	99,137	99.45	99.59	99.32	97.36	99.92	99.91	99.49
61	262	99,299	99.61	99.72	99.51	98.05	99.94	99.94	99.66
62	266	99,417	99.73	99.79	99.67	98.55	99.97	99.95	99.79
63	271	99,523	99.84	99.89	99.79	99.11	99.98	99.96	99.88
64	276	99,596	99.91	99.94	99.88	99.51	99.99	99.98	99.94
65	281	99,626	99.94	99.96	99.92	99.64	99.99	99.99	99.97
66	288	99,656	99.97	99.98	99.96	99.83	99.99	99.99	99.99
67	295	99,671	99.99	100.00	99.98	99.92	99.99	100.00	99.99
68	300	99,673	99.99	100.00	99.98	99.93	99.99	100.00	100.00
69	300	99,683	100.00	100.00	100.00	99.99	100.00	100.00	100.00
70	300	99,685	100.00	100.00	100.00	100.00	100.00	100.00	100.00
71	300	99,685	100.00	100.00	100.00	100.00	100.00	100.00	100.00
72	300	99,685	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	7	0.01	0.01	0.01	0.00	0.01	0.01	0.01
2	100	29	0.03	0.02	0.04	0.00	0.04	0.05	0.02
3	100	84	0.08	0.06	0.11	0.01	0.16	0.14	0.04
4	100	187	0.19	0.14	0.24	0.03	0.39	0.28	0.09
5	100	448	0.45	0.36	0.54	0.06	0.87	0.68	0.23
6	100	956	0.97	0.82	1.11	0.14	1.86	1.42	0.51
7	100	1,812	1.83	1.54	2.11	0.32	3.36	2.75	0.95
8	100	3,097	3.13	2.61	3.63	0.61	5.51	4.63	1.76
9	100	4,926	4.98	4.17	5.77	1.00	8.54	7.36	2.88
10	100	7,187	7.27	6.15	8.35	1.51	12.29	10.88	4.16
11	103	9,963	10.08	8.70	11.40	2.13	16.53	15.12	5.88
12	108	12,953	13.10	11.40	14.75	3.01	21.00	19.56	7.83
13	113	16,212	16.40	14.40	18.32	3.91	26.18	24.30	9.95
14	118	19,505	19.73	17.48	21.90	4.68	31.03	28.96	12.33
15	123	22,902	23.16	20.73	25.52	5.81	35.69	33.69	14.87
16	127	26,098	26.39	23.78	28.93	7.16	40.35	37.92	17.29
17	131	29,176	29.51	26.85	32.08	8.57	44.77	41.86	19.74
18	135	32,108	32.47	29.89	34.98	9.76	48.62	45.68	22.21
19	139	34,860	35.25	32.63	37.82	11.04	52.17	49.10	24.60
20	143	37,530	37.95	35.41	40.44	12.21	55.54	52.42	26.95
21	146	40,012	40.46	38.07	42.81	13.51	58.42	55.50	29.17
22	150	42,550	43.03	40.83	45.19	14.97	61.08	58.49	31.65
23	153	44,857	45.36	43.26	47.44	16.32	63.68	61.22	33.79
24	158	47,059	47.59	45.64	49.51	17.59	65.94	63.84	35.98
25	160	49,287	49.84	48.03	51.64	18.93	68.43	66.18	38.35
26	163	51,367	51.95	50.31	53.58	20.45	70.62	68.30	40.53
27	166	53,415	54.02	52.60	55.44	22.10	72.68	70.36	42.71
28	169	55,457	56.08	54.89	57.28	23.51	74.77	72.39	44.98
29	172	57,332	57.98	57.01	58.97	25.20	76.51	74.22	47.04
30	175	59,203	59.87	59.04	60.74	26.85	78.23	76.01	49.14
31	178	60,982	61.67	61.12	62.27	28.44	79.71	77.68	51.20
32	182	62,853	63.56	63.24	63.93	30.09	81.38	79.41	53.39
33	185	64,679	65.41	65.37	65.52	32.15	82.97	80.94	55.51
34	188	66,392	67.14	67.28	67.07	33.93	84.29	82.48	57.55
35	191	68,062	68.83	69.16	68.57	35.60	85.69	83.85	59.60
36	194	69,706	70.49	71.09	69.98	37.27	87.10	85.16	61.64
37	197	71,081	71.89	72.65	71.21	38.97	88.12	86.22	63.35
38	200	72,907	73.73	74.69	72.87	41.20	89.34	87.54	65.71
39	203	74,460	75.30	76.35	74.36	43.22	90.33	88.62	67.75
40	206	75,986	76.85	78.02	75.78	45.12	91.33	89.62	69.81
41	209	77,380	78.26	79.58	77.04	47.09	92.02	90.55	71.71
42	212	78,797	79.69	81.14	78.35	49.31	92.82	91.46	73.57

Raw Score	Scale Score	All Cum. N	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
43	215	80,160	81.07	82.62	79.63	51.52	93.55	92.33	75.38
44	218	81,468	82.39	84.05	80.85	53.62	94.25	93.07	77.18
45	221	82,703	83.64	85.39	82.01	55.95	94.91	93.81	78.76
46	224	83,906	84.86	86.72	83.12	58.50	95.52	94.47	80.29
47	228	85,071	86.03	87.91	84.29	60.78	96.01	95.04	81.89
48	231	86,222	87.20	89.08	85.43	63.27	96.52	95.61	83.41
49	234	87,319	88.31	90.24	86.50	65.61	96.98	96.14	84.88
50	237	88,337	89.34	91.24	87.55	67.81	97.49	96.62	86.21
51	241	89,341	90.35	92.24	88.58	69.94	97.84	97.07	87.60
52	244	90,260	91.28	93.11	89.56	72.09	98.17	97.41	88.90
53	250	91,113	92.14	93.91	90.48	74.16	98.43	97.75	90.06
54	251	91,971	93.01	94.69	91.43	76.54	98.66	98.05	91.22
55	254	92,773	93.82	95.44	92.29	78.59	98.87	98.37	92.33
56	258	93,463	94.52	96.03	93.09	80.69	99.04	98.59	93.25
57	262	94,138	95.20	96.59	93.89	82.66	99.15	98.81	94.18
58	265	94,790	95.86	97.14	94.66	84.46	99.33	99.02	95.07
59	269	95,382	96.46	97.59	95.39	86.28	99.42	99.22	95.85
60	273	95,905	96.99	97.99	96.05	88.09	99.55	99.39	96.45
61	277	96,412	97.50	98.38	96.67	89.76	99.68	99.56	97.08
62	281	96,846	97.94	98.67	97.26	91.37	99.76	99.68	97.59
63	286	97,265	98.37	98.94	97.82	92.97	99.77	99.76	98.14
64	291	97,585	98.69	99.18	98.23	94.36	99.83	99.81	98.51
65	296	97,862	98.97	99.38	98.58	95.53	99.89	99.87	98.83
66	300	98,105	99.22	99.55	98.90	96.35	99.94	99.91	99.14
67	300	98,332	99.44	99.68	99.22	97.38	99.96	99.94	99.40
68	300	98,495	99.61	99.79	99.44	98.13	99.99	99.96	99.58
69	300	98,624	99.74	99.88	99.61	98.73	99.99	99.98	99.73
70	300	98,711	99.83	99.94	99.72	99.18	99.99	99.98	99.82
71	300	98,787	99.90	99.97	99.84	99.54	99.99	99.98	99.92
72	300	98,837	99.96	99.98	99.93	99.77	100.00	99.99	99.96
73	300	98,860	99.98	99.99	99.97	99.89	100.00	100.00	99.98
74	300	98,869	99.99	99.99	99.98	99.93	100.00	100.00	99.99
75	300	98,877	100.00	100.00	99.99	99.97	100.00	100.00	100.00
76	300	98,880	100.00	100.00	100.00	99.99	100.00	100.00	100.00
77	300	98,881	100.00	100.00	100.00	100.00	100.00	100.00	100.00
78	300	98,881	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***
1905M069_01	0.400	0.85	0.80	0.55	1.42	0.00	0.45
1905M069_02	1.131	0.85	0.78	0.54	1.28	0.00	0.31
1905M069_05	1.569	0.83	0.71	0.54	1.26	0.00	0.24
1905M076_01	–0.352	1.14	1.17	0.24	0.63	0.16	0.61
1905M076_03	0.472	0.89	0.86	0.52	1.29	0.00	0.47
1905M076_05	1.204	0.91	0.87	0.52	1.16	0.00	0.32
2005B008_01	–0.225	0.80	0.75	0.57	1.55	0.00	0.57
2005B008_02	1.582	0.84	0.69	0.54	1.25	0.00	0.23
2005B008_03	1.320	0.92	0.94	0.45	1.11	0.00	0.28
2005B008_04	–0.256	0.97	1.00	0.40	1.07	0.00	0.58
2005B008_05	0.300	1.33	1.35	0.47	0.57	0.07	0.45
2005M015_02	–0.415	0.99	0.98	0.39	1.03	0.00	0.61
2005M015_04	1.009	1.04	1.08	0.36	0.91	0.02	0.33
2005M015_05	0.934	1.04	1.05	0.36	0.91	0.02	0.34
2005M015_06	1.713	1.24	1.61	0.10	0.63	0.07	0.22
2005M015_07	0.696	0.82	0.77	0.58	1.44	0.00	0.39
2005M015_10	0.875	1.02	1.00	0.38	0.97	0.00	0.35
2105M003_01	1.183	0.90	0.86	0.49	1.19	0.00	0.30
2105M003_07	–0.783	0.95	0.96	0.40	1.09	0.00	0.68
2105M003_09	1.315	0.84	0.80	0.54	1.26	0.00	0.28
2205B000_02	2.044	0.94	0.85	0.40	1.07	0.00	0.17
2205B000_03	0.409	1.02	1.05	0.40	0.94	0.05	0.44
2205B000_04	0.409	0.77	0.72	0.62	1.61	0.00	0.44
2205B000_07	0.793	0.82	0.76	0.74	1.16	0.00	0.35
2205B000_10	1.114	1.12	1.23	0.27	0.74	0.05	0.31
2205B006_01	0.025	1.11	1.14	0.30	0.68	0.09	0.52
2205B006_03	1.140	1.09	1.12	0.31	0.84	0.03	0.31
2205B006_06	2.114	1.02	1.16	0.29	0.96	0.01	0.16
2205B009_01	–0.298	0.88	0.82	0.51	1.34	0.00	0.58
2205B009_02	1.339	0.92	0.92	0.46	1.12	0.00	0.28
2205B009_03	1.430	1.13	1.27	0.24	0.77	0.05	0.26
2205B009_04	0.372	0.97	0.96	0.43	1.08	0.00	0.45
2205B009_05	1.091	1.24	1.15	0.64	0.85	0.00	0.31
2205M000_01	1.123	1.17	1.25	0.22	0.66	0.06	0.31
2205M000_02	0.752	1.10	1.12	0.31	0.78	0.04	0.38
2205M000_03	0.457	0.98	0.97	0.43	1.06	0.00	0.44
2205M000_04	0.804	1.10	1.14	0.30	0.76	0.06	0.37

NJSLA–S Grade 5							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p-value***
2205M004_03	0.265	0.80	0.76	0.59	1.56	0.00	0.44
2205M004_05	1.162	0.95	0.96	0.40	1.09	0.00	0.28
2205M004_07	0.273	1.10	1.16	0.30	0.69	0.02	0.45
2205M006_01	0.030	0.98	0.96	0.43	1.07	0.00	0.51
2205M006_02	0.591	1.15	1.18	0.23	0.62	0.04	0.36
2205M006_05	0.103	0.96	0.96	0.43	1.10	0.00	0.49
2205M022_01	0.433	0.97	0.96	0.44	1.08	0.00	0.45
2205M022_03	0.362	0.93	0.91	0.48	1.20	0.00	0.47
2205M022_05	0.447	1.15	1.19	0.26	0.60	0.08	0.43
2305B006_01	0.508	0.95	0.92	0.45	1.14	0.00	0.43
2305B006_03	0.812	0.99	1.01	0.41	1.02	0.00	0.37
2305B006_04	0.875	1.19	1.34	0.22	0.54	0.10	0.35
2305M008_01	0.593	1.16	1.23	0.25	0.58	0.10	0.41
2305M008_03	1.459	1.21	1.41	0.16	0.65	0.07	0.25
2305M008_04	0.726	1.03	1.10	0.38	0.90	0.04	0.38

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*
2008B007_03	0.797	1.06	1.14	0.30	0.90	0.02	0.24
2008B007_07	0.713	0.87	0.79	0.68	1.17	0.00	0.21
2008B007_09	–0.954	0.95	0.96	0.39	1.13	0.00	0.58
2008B007_11	0.513	0.99	0.95	0.39	1.03	0.00	0.29
2008M000_01	0.640	1.12	1.24	0.21	0.78	0.05	0.26
2008M000_02	1.107	0.79	0.67	0.47	1.26	0.00	0.17
2008M000_04	0.291	1.07	1.09	0.26	0.85	0.01	0.30
2108B006_03	1.232	0.93	0.98	0.31	1.06	0.00	0.16
2108B006_06	0.543	0.99	0.97	0.38	1.02	0.00	0.29
2108B006_09	0.175	0.88	0.84	0.53	1.28	0.00	0.38
2108B007_03	0.980	1.18	1.34	0.14	0.75	0.05	0.21
2108B007_07	0.988	1.04	1.19	0.24	0.91	0.02	0.20
2108B007_08	–1.045	0.95	0.97	0.41	1.14	0.00	0.58
2108M000_02	0.213	0.95	0.93	0.43	1.11	0.00	0.35
2108M000_03	0.698	0.89	0.79	0.49	1.20	0.00	0.26
2108M000_06	0.893	1.11	1.14	0.24	0.85	0.03	0.23
2208B003_01	–0.204	0.90	0.88	0.48	1.28	0.00	0.41
2208B003_05	–0.107	0.91	0.88	0.49	1.26	0.00	0.42
2208B003_07	1.196	0.93	0.84	0.48	1.09	0.00	0.20
2208B003_09	0.590	1.15	1.23	0.23	0.74	0.05	0.29
2208B003_11	–0.220	1.03	1.02	0.59	0.95	0.00	0.45
2208M016_03	0.207	1.00	1.02	0.38	0.99	0.01	0.35
2208M016_04	–0.993	0.82	0.77	0.53	1.56	0.00	0.59
2208M016_05	0.402	1.30	1.45	0.07	0.40	0.13	0.31
2208M016_08	–0.190	1.17	1.20	0.21	0.53	0.09	0.42
2208M021_03	0.581	0.99	0.99	0.38	1.01	0.00	0.28
2208M021_05	0.104	1.01	1.03	0.38	0.97	0.01	0.38
2208M021_09	0.443	1.07	1.12	0.34	0.86	0.04	0.32
2208M021_10	0.065	0.83	0.77	0.57	1.42	0.00	0.39
2208M051_13	0.091	1.00	1.01	0.39	0.99	0.02	0.38
2208M051_16	–0.379	0.93	0.91	0.45	1.23	0.00	0.45
2208M051_17	0.344	0.93	0.90	0.45	1.15	0.00	0.32
2308B003_01	0.044	0.83	0.81	0.55	1.40	0.00	0.38
2308B003_03	0.811	0.96	0.92	0.41	1.07	0.00	0.24
2308B003_05	0.064	1.00	0.99	0.62	1.01	0.00	0.38
2308B003_07	0.937	0.97	1.04	0.38	1.02	0.00	0.22
2308B003_11	–0.704	0.90	0.88	0.46	1.33	0.00	0.53
2308B006_01	1.133	1.00	1.07	0.33	0.98	0.00	0.20
2308B006_07	1.027	1.15	1.32	0.18	0.79	0.04	0.21

NJSLA–S Grade 8							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p-value***
2308B006_08	0.837	1.01	1.04	0.35	0.98	0.00	0.24
2308M000_04	–0.232	1.06	1.08	0.34	0.83	0.07	0.43
2308M000_06	0.819	0.91	0.92	0.45	1.11	0.00	0.24
2308M000_08	0.495	0.84	0.82	0.54	1.27	0.00	0.29
2308M017_01	0.406	0.99	0.97	0.39	1.03	0.00	0.31
2308M017_02	–0.150	0.96	0.97	0.43	1.10	0.00	0.42
2308M017_05	–1.265	1.09	1.20	0.23	0.71	0.04	0.64
2308M020_05	0.212	0.97	0.96	0.41	1.06	0.00	0.35
2308M020_09	0.410	1.02	1.01	0.36	0.97	0.00	0.31
2308M020_11	–1.025	1.01	1.05	0.33	0.95	0.00	0.60
2308M023_02	–0.555	1.06	1.06	0.33	0.82	0.07	0.50
2308M023_03	0.833	0.85	0.77	0.52	1.22	0.00	0.24
2308M023_07	0.568	1.21	1.34	0.15	0.61	0.08	0.28
2308M027_01	0.189	1.07	1.06	0.31	0.86	0.02	0.35
2308M027_02	0.592	1.10	1.25	0.26	0.80	0.05	0.28
2308M027_03	0.778	0.93	0.90	0.44	1.10	0.00	0.25
2308M027_05	0.340	0.78	0.70	0.62	1.46	0.00	0.32
2308M028_04	0.628	0.99	0.99	0.38	1.01	0.00	0.27
2308M028_05	0.253	1.21	1.32	0.16	0.53	0.10	0.34
2308M028_06	0.133	1.11	1.13	0.27	0.74	0.05	0.36
2308M028_09	–0.388	1.08	1.11	0.31	0.74	0.09	0.46
2308M049_01	0.605	0.86	0.80	0.52	1.24	0.00	0.28
2308M049_03	0.295	0.89	0.88	0.50	1.22	0.00	0.33
2308M049_05	0.634	0.98	1.03	0.39	1.02	0.00	0.27
2308M049_07	–0.391	1.21	1.25	0.17	0.34	0.15	0.47

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***
2011M003_01	0.955	0.92	0.90	0.47	1.13	0.00	0.29
2011M003_03	1.147	1.07	1.14	0.32	0.88	0.03	0.26
2011M003_04	0.788	1.02	1.05	0.40	0.96	0.01	0.33
2011M003_05	0.497	0.89	0.86	0.53	1.25	0.00	0.39
2011M003_06	0.034	0.98	0.98	0.42	1.05	0.01	0.47
2011M028_01	–0.682	1.10	1.13	0.27	0.71	0.15	0.61
2011M028_03	0.237	1.00	1.00	0.40	1.00	0.03	0.42
2011M028_04	1.348	1.10	1.30	0.24	0.82	0.04	0.22
2011M028_05	0.025	1.00	1.01	0.38	0.99	0.00	0.46
2011M031_03	1.447	1.04	1.10	0.32	0.94	0.01	0.21
2011M031_04	0.798	1.12	1.21	0.26	0.75	0.06	0.31
2011M031_05	0.756	1.15	1.21	0.24	0.70	0.06	0.32
2011M031_06	0.050	1.11	1.17	0.27	0.65	0.04	0.46
2011M054_04	1.449	0.94	0.85	0.43	1.09	0.00	0.21
2011M054_06	0.139	0.91	0.88	0.49	1.27	0.00	0.44
2011M054_08	0.327	1.05	1.06	0.34	0.87	0.01	0.40
2011M071_01	–0.598	0.93	0.90	0.50	1.23	0.00	0.52
2011M071_03	0.569	0.83	0.81	0.50	1.36	0.00	0.30
2011M071_04	–0.262	1.12	1.15	0.28	0.62	0.03	0.47
2011M071_05	–0.311	0.99	1.03	0.41	1.01	0.00	0.47
2111B002_01	0.732	0.78	0.72	0.61	1.43	0.00	0.32
2111B002_03	0.278	0.99	0.98	0.40	1.02	0.00	0.41
2111B002_07	0.760	1.03	1.01	0.37	0.95	0.00	0.32
2111B002_10	1.177	0.90	0.78	0.49	1.17	0.00	0.25
2111B002_11	–0.096	0.92	0.92	0.64	1.09	0.00	0.49
2111M004_02	2.137	1.08	1.24	0.28	0.92	0.01	0.14
2111M004_03	1.189	1.13	1.13	0.45	0.83	0.04	0.32
2111M004_05	–0.260	1.09	1.10	0.29	0.72	0.11	0.55
2111M004_06	1.328	1.14	1.23	0.37	0.81	0.05	0.27
2111M020_01	–0.593	0.96	0.93	0.41	1.13	0.00	0.59
2111M020_03	1.345	0.99	1.12	0.36	0.98	0.01	0.22
2111M020_05	–0.057	0.85	0.82	0.54	1.46	0.00	0.48
2211B001_01	0.922	1.02	1.12	0.35	0.93	0.02	0.29
2211B001_03	1.071	1.05	1.10	0.33	0.91	0.02	0.27
2211B001_05	0.039	0.79	0.75	0.60	1.62	0.00	0.46
2211B001_07	0.734	0.98	0.98	0.41	1.04	0.00	0.32
2211B001_10	1.417	1.00	0.92	0.60	1.04	0.00	0.19
2211B006_02	0.825	0.99	1.00	0.44	1.02	0.01	0.33
2211B006_05	0.971	0.93	0.95	0.48	1.11	0.00	0.29

NJSLA–S Grade 11							
Item	**Rasch B**	**Infit**	**Outfit**	***rpb***	**Rasch Discrim.**	**Lower Asym.**	***p-value***
2211B006_06	–1.220	0.89	0.84	0.44	1.23	0.00	0.70
2211B006_09	–0.242	1.09	1.18	0.60	0.82	0.00	0.53
2211B006_12	0.526	0.83	0.79	0.56	1.39	0.00	0.35
2211M008_01	1.421	0.89	0.80	0.40	1.15	0.00	0.19
2211M008_03	0.242	1.12	1.15	0.28	0.67	0.06	0.42
2211M008_07	0.896	1.11	1.14	0.24	0.81	0.03	0.28
2211M016_01	0.548	1.01	1.00	0.38	0.97	0.00	0.36
2211M016_02	–0.239	1.05	1.05	0.33	0.84	0.04	0.52
2211M016_06	0.711	1.09	1.13	0.31	0.82	0.04	0.33
2211M016_08	1.537	0.93	0.90	0.43	1.08	0.00	0.20
2211M022_01	0.542	0.90	0.89	0.49	1.21	0.00	0.36
2211M022_02	0.895	0.89	0.82	0.50	1.20	0.00	0.30
2211M022_08	0.748	1.18	1.22	0.22	0.66	0.06	0.32
2311M011_03	0.990	0.91	0.83	0.48	1.16	0.00	0.28
2311M011_05	–0.699	1.10	1.17	0.26	0.69	0.13	0.61
2311M011_07	–0.355	0.97	0.95	0.42	1.10	0.02	0.54
2311M016_01	0.216	1.04	1.06	0.36	0.88	0.03	0.42
2311M016_02	1.232	0.83	0.72	0.55	1.25	0.00	0.24
2311M016_04	0.132	1.02	1.03	0.37	0.94	0.00	0.44
2311M016_07	1.444	1.15	1.37	0.19	0.78	0.04	0.21
2311M021_02	1.657	1.01	0.94	0.35	1.01	0.00	0.18
2311M021_04	0.344	1.27	1.36	0.13	0.30	0.15	0.40
2311M021_06	0.223	1.10	1.12	0.30	0.73	0.07	0.42
2311M023_01	–0.671	1.01	0.98	0.36	0.99	0.05	0.60
2311M023_03	0.077	1.01	1.02	0.38	0.97	0.01	0.45
2311M023_06	1.687	0.98	1.11	0.35	0.99	0.00	0.18
2311M026_01	0.358	0.93	0.91	0.47	1.18	0.00	0.40
2311M026_02	0.822	1.23	1.39	0.16	0.55	0.10	0.31
2311M026_03	0.082	0.84	0.80	0.55	1.47	0.00	0.45
2311M026_05	1.822	1.06	1.19	0.27	0.92	0.01	0.16
2311M026_06	0.552	1.00	1.01	0.40	1.00	0.00	0.36

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	–4.668	1.833	-36.603	100	19	100	119
1	–3.446	1.013	15.314	100	19	100	119
2	–2.726	0.726	45.879	100	19	100	119
3	–2.294	0.600	64.224	100	19	100	119
4	–1.980	0.526	77.566	100	19	100	119
5	–1.730	0.476	88.165	100	19	100	119
6	–1.521	0.439	97.027	100	19	100	119
7	–1.341	0.411	104.692	105	17	100	122
8	–1.181	0.389	111.473	111	17	100	128
9	–1.037	0.370	117.584	118	16	102	134
10	–0.906	0.355	123.159	123	15	108	138
11	–0.785	0.342	128.306	128	15	113	143
12	–0.672	0.330	133.092	133	14	119	147
13	–0.566	0.320	137.576	138	14	124	152
14	–0.467	0.311	141.805	142	13	129	155
15	–0.372	0.303	145.810	146	13	133	159
16	–0.283	0.296	149.623	150	13	137	163
17	–0.197	0.290	153.271	153	12	141	165
18	–0.114	0.284	156.765	157	12	145	169
19	–0.035	0.279	160.133	160	12	148	172
20	0.041	0.275	163.386	163	12	151	175
21	0.116	0.271	166.536	167	11	156	178
22	0.188	0.267	169.602	170	11	159	181
23	0.258	0.264	172.596	173	11	162	184
24	0.327	0.261	175.522	176	11	165	187
25	0.395	0.259	178.401	178	11	167	189
26	0.462	0.258	181.233	181	11	170	192
27	0.528	0.256	184.036	184	11	173	195
28	0.593	0.255	186.813	187	11	176	198
29	0.658	0.255	189.577	190	11	179	201
30	0.723	0.255	192.333	192	11	181	203
31	0.788	0.255	195.098	195	11	184	206
32	0.854	0.256	197.870	200	11	189	211
33	0.919	0.257	200.665	201	11	190	212
34	0.986	0.259	203.488	203	11	192	214
35	1.053	0.261	206.351	206	11	195	217
36	1.122	0.263	209.259	209	11	198	220

NJSLA–S Grade 5							
Raw Score	**Theta**	**Standard Error**	**Unrounded Scale Score**	**Scale Score (SS)**	**SS CSEM**	**SS Lower Bound**	**SS Upper Bound**
37	1.192	0.266	212.228	212	11	201	223
38	1.263	0.269	215.259	215	11	204	226
39	1.336	0.272	218.372	218	12	206	230
40	1.412	0.276	221.570	222	12	210	234
41	1.489	0.281	224.865	225	12	213	237
42	1.570	0.286	228.279	228	12	216	240
43	1.653	0.291	231.816	232	12	220	244
44	1.740	0.298	235.502	236	13	223	249
45	1.830	0.305	239.349	243	13	230	256
46	1.925	0.312	243.383	244	13	231	257
47	2.026	0.321	247.638	248	14	234	262
48	2.132	0.331	252.139	252	14	238	266
49	2.245	0.342	256.938	257	15	242	272
50	2.366	0.355	262.084	262	15	247	277
51	2.497	0.370	267.656	268	16	252	284
52	2.641	0.388	273.754	274	16	258	290
53	2.800	0.411	280.518	281	17	264	298
54	2.980	0.439	288.153	288	19	269	300
55	3.188	0.475	296.981	297	20	277	300
56	3.436	0.525	307.542	300	22	278	300
57	3.749	0.599	320.833	300	22	278	300
58	4.180	0.725	339.127	300	22	278	300
59	4.898	1.012	369.629	300	22	278	300
60	6.120	1.833	421.503	300	22	278	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	–5.327	1.831	–16.947	100	16	100	116
1	–4.109	1.010	29.065	100	16	100	116
2	–3.396	0.721	56.003	100	16	100	116
3	–2.971	0.594	72.067	100	16	100	116
4	–2.663	0.519	83.680	100	16	100	116
5	–2.420	0.469	92.857	100	16	100	116
6	–2.218	0.432	100.492	100	16	100	116
7	–2.044	0.403	107.062	107	15	100	122
8	–1.891	0.381	112.858	113	14	100	127
9	–1.753	0.362	118.064	118	14	104	132
10	–1.628	0.347	122.805	123	13	110	136
11	–1.512	0.334	127.173	127	13	114	140
12	–1.405	0.322	131.230	131	12	119	143
13	–1.304	0.313	135.035	135	12	123	147
14	–1.209	0.304	138.628	139	11	128	150
15	–1.119	0.297	142.032	142	11	131	153
16	–1.033	0.290	145.281	145	11	134	156
17	–0.950	0.284	148.394	150	11	139	161
18	–0.871	0.279	151.386	151	11	140	162
19	–0.795	0.274	154.276	154	10	144	164
20	–0.721	0.270	157.072	157	10	147	167
21	–0.649	0.266	159.788	160	10	150	170
22	–0.579	0.263	162.437	162	10	152	172
23	–0.510	0.260	165.021	165	10	155	175
24	–0.443	0.257	167.548	168	10	158	178
25	–0.378	0.255	170.030	170	10	160	180
26	–0.313	0.253	172.467	172	10	162	182
27	–0.250	0.251	174.866	175	9	166	184
28	–0.187	0.249	177.235	177	9	168	186
29	–0.125	0.248	179.574	180	9	171	189
30	–0.064	0.247	181.886	182	9	173	191
31	–0.003	0.246	184.175	184	9	175	193
32	0.057	0.245	186.450	186	9	177	195
33	0.117	0.244	188.709	189	9	180	198
34	0.176	0.244	190.957	191	9	182	200
35	0.236	0.243	193.193	193	9	184	202
36	0.295	0.243	195.426	195	9	186	204
37	0.354	0.243	197.655	198	9	189	207
38	0.413	0.243	199.884	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	0.472	0.243	202.113	202	9	193	211
40	0.531	0.243	204.350	204	9	195	213
41	0.590	0.244	206.590	207	9	198	216
42	0.650	0.245	208.846	209	9	200	218
43	0.710	0.245	211.112	211	9	202	220
44	0.770	0.247	213.398	213	9	204	222
45	0.831	0.248	215.706	216	9	207	225
46	0.893	0.249	218.037	218	9	209	227
47	0.956	0.251	220.402	220	9	211	229
48	1.019	0.253	222.801	223	10	213	233
49	1.084	0.255	225.246	225	10	215	235
50	1.150	0.258	227.736	228	10	218	238
51	1.217	0.261	230.282	231	10	221	241
52	1.286	0.265	232.896	233	10	223	243
53	1.358	0.269	235.586	236	10	226	246
54	1.431	0.273	238.359	238	10	228	248
55	1.507	0.279	241.234	241	11	230	252
56	1.586	0.284	244.227	244	11	233	255
57	1.669	0.291	247.355	247	11	236	258
58	1.756	0.299	250.638	251	11	240	262
59	1.848	0.307	254.106	254	12	242	266
60	1.945	0.317	257.790	258	12	246	270
61	2.050	0.329	261.730	262	12	250	274
62	2.162	0.342	265.976	266	13	253	279
63	2.284	0.358	270.601	271	14	257	285
64	2.419	0.377	275.690	276	14	262	290
65	2.570	0.400	281.372	281	15	266	296
66	2.741	0.428	287.832	288	16	272	300
67	2.940	0.466	295.354	295	18	277	300
68	3.180	0.517	304.418	300	20	280	300
69	3.484	0.592	315.922	300	20	280	300
70	3.906	0.719	331.872	300	20	280	300
71	4.616	1.008	358.692	300	20	280	300
72	5.832	1.830	404.622	300	20	280	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	–5.146	1.831	–96.913	100	18	100	118
1	–3.929	1.009	–32.638	100	18	100	118
2	–3.218	0.720	4.932	100	18	100	118
3	–2.795	0.592	27.290	100	18	100	118
4	–2.489	0.517	43.411	100	18	100	118
5	–2.249	0.466	56.103	100	18	100	118
6	–2.050	0.428	66.625	100	18	100	118
7	–1.879	0.399	75.641	100	18	100	118
8	–1.729	0.376	83.559	100	18	100	118
9	–1.595	0.357	90.636	100	18	100	118
10	–1.474	0.341	97.048	100	18	100	118
11	–1.363	0.327	102.927	103	17	100	120
12	–1.260	0.315	108.373	108	17	100	125
13	–1.163	0.305	113.449	113	16	100	129
14	–1.073	0.296	118.218	118	16	102	134
15	–0.988	0.288	122.729	123	15	108	138
16	–0.907	0.282	127.018	127	15	112	142
17	–0.829	0.276	131.111	131	15	116	146
18	–0.755	0.270	135.046	135	14	121	149
19	–0.683	0.266	138.844	139	14	125	153
20	–0.613	0.262	142.515	143	14	129	157
21	–0.546	0.258	146.086	146	14	132	160
22	–0.480	0.255	149.561	150	13	137	163
23	–0.415	0.252	152.963	153	13	140	166
24	–0.352	0.250	156.296	158	13	145	171
25	–0.290	0.248	159.576	160	13	147	173
26	–0.229	0.246	162.803	163	13	150	176
27	–0.169	0.245	165.988	166	13	153	179
28	–0.109	0.244	169.136	169	13	156	182
29	–0.050	0.243	172.263	172	13	159	185
30	0.009	0.242	175.358	175	13	162	188
31	0.067	0.241	178.437	178	13	165	191
32	0.125	0.240	181.495	182	13	169	195
33	0.183	0.240	184.548	185	13	172	198
34	0.240	0.240	187.585	188	13	175	201
35	0.298	0.240	190.623	191	13	178	204
36	0.355	0.240	193.654	194	13	181	207
37	0.412	0.240	196.681	197	13	184	210
38	0.470	0.240	199.718	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	0.527	0.240	202.755	203	13	190	216
40	0.585	0.240	205.797	206	13	193	219
41	0.643	0.241	208.856	209	13	196	222
42	0.701	0.241	211.919	212	13	199	225
43	0.759	0.242	215.004	215	13	202	228
44	0.818	0.243	218.104	218	13	205	231
45	0.877	0.243	221.221	221	13	208	234
46	0.936	0.244	224.363	224	13	211	237
47	0.996	0.245	227.527	228	13	215	241
48	1.057	0.247	230.723	231	13	218	244
49	1.118	0.248	233.955	234	13	221	247
50	1.180	0.249	237.219	237	13	224	250
51	1.242	0.251	240.521	241	13	228	254
52	1.306	0.253	243.869	244	13	231	257
53	1.370	0.255	247.271	250	13	237	263
54	1.436	0.257	250.725	251	14	237	265
55	1.502	0.259	254.243	254	14	240	268
56	1.570	0.262	257.835	258	14	244	272
57	1.640	0.265	261.506	262	14	248	276
58	1.711	0.269	265.261	265	14	251	279
59	1.784	0.272	269.127	269	14	255	283
60	1.859	0.277	273.105	273	15	258	288
61	1.937	0.282	277.219	277	15	262	292
62	2.018	0.287	281.492	281	15	266	296
63	2.102	0.294	285.940	286	16	270	300
64	2.191	0.301	290.604	291	16	275	300
65	2.284	0.309	295.510	296	16	280	300
66	2.382	0.319	300.713	300	17	283	300
67	2.487	0.330	306.270	300	17	283	300
68	2.600	0.343	312.249	300	17	283	300
69	2.723	0.359	318.745	300	17	283	300
70	2.859	0.377	325.892	300	17	283	300
71	3.010	0.400	333.862	300	17	283	300
72	3.181	0.429	342.926	300	17	283	300
73	3.381	0.466	353.479	300	17	283	300
74	3.622	0.517	366.193	300	17	283	300
75	3.927	0.593	382.318	300	17	283	300
76	4.350	0.720	404.682	300	17	283	300
77	5.062	1.009	442.247	300	17	283	300
78	6.278	1.831	506.512	300	17	283	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	–3.426	1.847	–6.197	–0.655	Below
1	–2.167	1.039	–3.724	–0.609	Below
2	–1.395	0.760	–2.535	–0.256	Below
3	–0.914	0.639	–1.872	0.045	Below
4	–0.552	0.568	–1.404	0.299	Below
5	–0.259	0.519	–1.037	0.520	Below
6	–0.008	0.484	–0.734	0.717	Below
7	0.213	0.458	–0.474	0.900	Below
8	0.414	0.440	–0.246	1.073	Near/Met
9	0.602	0.429	–0.041	1.245	Near/Met
10	0.783	0.424	0.147	1.419	Near/Met
11	0.963	0.426	0.325	1.601	Near/Met
12	1.147	0.433	0.498	1.796	Near/Met
13	1.339	0.445	0.671	2.007	Near/Met
14	1.545	0.464	0.850	2.241	Above
15	1.771	0.488	1.039	2.504	Above
16	2.026	0.521	1.244	2.807	Above
17	2.320	0.567	1.470	3.171	Above
18	2.679	0.636	1.725	3.633	Above
19	3.155	0.755	2.023	4.288	Above
20	3.918	1.034	2.367	5.469	Above
21	5.170	1.844	2.404	7.937	Above

Table J.2: Grade 5 Life Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.558	1.849	–6.330	–0.785	Below
1	–2.296	1.040	–3.856	–0.735	Below
2	–1.522	0.761	–2.663	–0.380	Below
3	–1.038	0.640	–1.998	–0.078	Below
4	–0.676	0.568	–1.528	0.177	Below
5	–0.381	0.520	–1.160	0.398	Below
6	–0.130	0.485	–0.857	0.597	Below
7	0.093	0.460	–0.597	0.783	Below
8	0.297	0.444	–0.370	0.963	Near/Met
9	0.490	0.436	–0.165	1.144	Near/Met
10	0.680	0.436	0.026	1.333	Near/Met
11	0.872	0.442	0.209	1.534	Near/Met
12	1.072	0.455	0.390	1.754	Near/Met
13	1.287	0.474	0.577	1.997	Near/Met
14	1.523	0.499	0.774	2.271	Near/Met
15	1.788	0.533	0.989	2.587	Above
16	2.096	0.579	1.228	2.964	Above
17	2.469	0.647	1.498	3.439	Above
18	2.959	0.765	1.812	4.107	Above
19	3.737	1.041	2.175	5.299	Above
20	5.000	1.848	2.227	7.772	Above

Table J.3: Grade 5 Physical Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.664	1.846	–6.433	–0.894	Below
1	–2.407	1.038	–3.963	–0.851	Below
2	–1.637	0.760	–2.777	–0.496	Below
3	–1.152	0.643	–2.116	–0.188	Below
4	–0.784	0.576	–1.648	0.081	Below
5	–0.477	0.533	–1.277	0.323	Below
6	–0.209	0.505	–0.966	0.548	Below
7	0.035	0.485	–0.692	0.763	Below
8	0.264	0.473	–0.445	0.973	Near/Met
9	0.484	0.466	–0.216	1.183	Near/Met
10	0.701	0.466	0.002	1.399	Near/Met
11	0.919	0.471	0.213	1.626	Near/Met
12	1.146	0.482	0.423	1.869	Near/Met
13	1.387	0.501	0.635	2.139	Near/Met
14	1.652	0.530	0.857	2.446	Above
15	1.954	0.572	1.095	2.813	Above
16	2.318	0.640	1.359	3.277	Above
17	2.798	0.758	1.662	3.935	Above
18	3.565	1.036	2.011	5.118	Above
19	4.819	1.845	2.051	7.587	Above

Table J.4: Grade 5 Critiquing Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–2.837	1.850	–5.612	–0.062	Below
1	–1.578	1.034	–3.128	–0.027	Below
2	–0.830	0.736	–1.934	0.275	Below
3	–0.391	0.600	–1.291	0.510	Below
4	–0.079	0.522	–0.862	0.703	Below
5	0.167	0.474	–0.544	0.877	Below
6	0.377	0.445	–0.291	1.045	Near/Met
7	0.568	0.430	–0.078	1.213	Near/Met
8	0.750	0.426	0.112	1.389	Near/Met
9	0.933	0.430	0.288	1.578	Near/Met
10	1.122	0.443	0.459	1.786	Near/Met
11	1.327	0.463	0.632	2.022	Near/Met
12	1.554	0.492	0.816	2.292	Above
13	1.814	0.530	1.020	2.609	Above
14	2.120	0.579	1.252	2.989	Above
15	2.495	0.649	1.521	3.469	Above
16	2.989	0.768	1.838	4.141	Above
17	3.772	1.043	2.207	5.337	Above
18	5.037	1.850	2.263	7.812	Above

Table J.5: Grade 5 Investigating Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.578	1.844	–6.344	–0.813	Below
1	–2.328	1.033	–3.878	–0.778	Below
2	–1.566	0.755	–2.699	–0.432	Below
3	–1.088	0.638	–2.045	–0.130	Below
4	–0.724	0.573	–1.583	0.135	Below
5	–0.420	0.532	–1.218	0.377	Below
6	–0.153	0.504	–0.910	0.604	Below
7	0.092	0.486	–0.638	0.821	Below
8	0.322	0.475	–0.390	1.034	Near/Met
9	0.544	0.469	–0.158	1.247	Near/Met
10	0.763	0.467	0.063	1.463	Near/Met
11	0.982	0.470	0.277	1.686	Near/Met
12	1.205	0.477	0.490	1.920	Near/Met
13	1.438	0.489	0.704	2.171	Near/Met
14	1.686	0.508	0.924	2.447	Above
15	1.957	0.536	1.153	2.761	Above
16	2.265	0.577	1.399	3.131	Above
17	2.635	0.643	1.670	3.599	Above
18	3.119	0.760	1.979	4.259	Above
19	3.888	1.037	2.333	5.443	Above
20	5.144	1.846	2.375	7.913	Above

Table J.6: Grade 5 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.955	1.844	–6.721	–1.189	Below
1	–2.704	1.034	–4.254	–1.154	Below
2	–1.942	0.755	–3.074	–0.810	Below
3	–1.466	0.636	–2.420	–0.513	Below
4	–1.107	0.568	–1.959	–0.256	Below
5	–0.811	0.523	–1.596	–0.026	Below
6	–0.554	0.493	–1.292	0.185	Below
7	–0.322	0.470	–1.028	0.383	Below
8	–0.109	0.454	–0.791	0.572	Below
9	0.092	0.443	–0.573	0.757	Below
10	0.285	0.436	–0.369	0.939	Near/Met
11	0.474	0.433	–0.176	1.123	Near/Met
12	0.661	0.434	0.011	1.312	Near/Met
13	0.851	0.438	0.193	1.508	Near/Met
14	1.047	0.447	0.376	1.718	Near/Met
15	1.253	0.462	0.560	1.946	Near/Met
16	1.476	0.483	0.751	2.200	Near/Met
17	1.723	0.513	0.953	2.493	Above
18	2.008	0.558	1.172	2.845	Above
19	2.356	0.626	1.416	3.295	Above
20	2.819	0.747	1.700	3.939	Above
21	3.569	1.028	2.028	5.110	Above
22	4.812	1.841	2.050	7.573	Above

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

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.151	1.836	–6.905	–1.397	Below
1	–2.920	1.019	–4.449	–1.391	Below
2	–2.187	0.736	–3.291	–1.084	Below
3	–1.739	0.614	–2.661	–0.818	Below
4	–1.406	0.545	–2.224	–0.588	Below
5	–1.134	0.501	–1.885	–0.382	Below
6	–0.898	0.471	–1.605	–0.192	Below
7	–0.687	0.450	–1.361	–0.012	Below
8	–0.492	0.435	–1.144	0.160	Below
9	–0.308	0.424	–0.944	0.328	Below
10	–0.131	0.417	–0.757	0.494	Near/Met
11	0.041	0.413	–0.578	0.660	Near/Met
12	0.210	0.411	–0.406	0.827	Near/Met
13	0.379	0.412	–0.238	0.996	Near/Met
14	0.550	0.415	–0.072	1.171	Near/Met
15	0.724	0.420	0.094	1.353	Near/Met
16	0.903	0.428	0.261	1.546	Near/Met
17	1.091	0.440	0.431	1.752	Above
18	1.292	0.456	0.607	1.976	Above
19	1.510	0.479	0.792	2.228	Above
20	1.753	0.510	0.989	2.518	Above
21	2.035	0.554	1.203	2.866	Above
22	2.379	0.623	1.444	3.313	Above
23	2.838	0.744	1.723	3.953	Above
24	3.583	1.025	2.045	5.121	Above
25	4.822	1.839	2.063	7.581	Above

Table J.8: Grade 8 Life Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.132	1.840	–6.891	–1.373	Below
1	–2.893	1.025	–4.430	–1.355	Below
2	–2.149	0.742	–3.262	–1.036	Below
3	–1.693	0.621	–2.623	–0.762	Below
4	–1.352	0.551	–2.179	–0.526	Below
5	–1.075	0.506	–1.833	–0.316	Below
6	–0.835	0.475	–1.548	–0.122	Below
7	–0.620	0.454	–1.300	0.061	Below
8	–0.421	0.439	–1.079	0.237	Below
9	–0.233	0.429	–0.876	0.410	Below
10	–0.053	0.422	–0.686	0.581	Near/Met
11	0.124	0.419	–0.505	0.753	Near/Met
12	0.300	0.419	–0.329	0.928	Near/Met
13	0.476	0.421	–0.156	1.108	Near/Met
14	0.656	0.427	0.016	1.295	Near/Met
15	0.841	0.435	0.189	1.493	Near/Met
16	1.035	0.446	0.365	1.704	Near/Met
17	1.241	0.462	0.547	1.934	Above
18	1.464	0.484	0.738	2.191	Above
19	1.713	0.515	0.941	2.486	Above
20	2.000	0.559	1.161	2.839	Above
21	2.350	0.628	1.408	3.291	Above
22	2.815	0.747	1.694	3.935	Above
23	3.565	1.028	2.023	5.107	Above
24	4.808	1.841	2.047	7.569	Above

Table J.9: Grade 8 Physical Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.352	1.847	–7.123	–1.581	Below
1	–3.092	1.039	–4.651	–1.533	Below
2	–2.318	0.763	–3.462	–1.174	Below
3	–1.830	0.645	–2.798	–0.862	Below
4	–1.459	0.579	–2.327	–0.591	Below
5	–1.150	0.535	–1.953	–0.347	Below
6	–0.880	0.505	–1.638	–0.123	Below
7	–0.637	0.483	–1.361	0.088	Below
8	–0.412	0.466	–1.111	0.288	Below
9	–0.200	0.454	–0.881	0.480	Near/Met
10	0.001	0.444	–0.665	0.667	Near/Met
11	0.195	0.437	–0.460	0.850	Near/Met
12	0.383	0.431	–0.264	1.030	Near/Met
13	0.567	0.428	–0.075	1.209	Near/Met
14	0.750	0.428	0.109	1.392	Near/Met
15	0.935	0.432	0.287	1.582	Near/Met
16	1.124	0.441	0.463	1.785	Above
17	1.325	0.458	0.639	2.012	Above
18	1.547	0.486	0.818	2.276	Above
19	1.803	0.530	1.008	2.599	Above
20	2.121	0.602	1.218	3.025	Above
21	2.556	0.728	1.464	3.649	Above
22	3.282	1.017	1.756	4.808	Above
23	4.511	1.836	1.757	7.266	Above

Table J.10: Grade 8 Critiquing Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.525	1.838	–7.282	–1.769	Below
1	–3.291	1.021	–4.823	–1.759	Below
2	–2.555	0.737	–3.660	–1.450	Below
3	–2.107	0.613	–3.027	–1.188	Below
4	–1.777	0.541	–2.589	–0.965	Below
5	–1.510	0.494	–2.251	–0.769	Below
6	–1.283	0.461	–1.974	–0.591	Below
7	–1.082	0.437	–1.737	–0.427	Below
8	–0.900	0.418	–1.527	–0.272	Below
9	–0.731	0.405	–1.337	–0.124	Below
10	–0.571	0.394	–1.163	0.020	Below
11	–0.419	0.387	–.999	0.161	Below
12	–0.272	0.381	–0.843	0.300	Below
13	–0.128	0.378	–0.694	0.439	Near/Met
14	0.014	0.375	–0.549	0.577	Near/Met
15	0.155	0.374	–0.407	0.716	Near/Met
16	0.294	0.374	–0.267	0.856	Near/Met
17	0.435	0.375	–0.128	0.997	Near/Met
18	0.576	0.377	0.011	1.141	Near/Met
19	0.719	0.380	0.149	1.290	Near/Met
20	0.866	0.385	0.288	1.443	Near/Met
21	1.017	0.393	0.427	1.606	Above
22	1.175	0.404	0.569	1.781	Above
23	1.344	0.420	0.714	1.974	Above
24	1.530	0.443	0.866	2.195	Above
25	1.740	0.476	1.026	2.455	Above
26	1.989	0.525	1.203	2.776	Above
27	2.302	0.599	1.404	3.200	Above
28	2.733	0.726	1.645	3.822	Above
29	3.454	1.015	1.932	4.976	Above
30	4.680	1.835	1.928	7.431	Above

Table J.11: Grade 8 Investigating Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.041	1.848	–6.813	–1.270	Below
1	–2.780	1.040	–4.340	–1.221	Below
2	–2.005	0.763	–3.150	–0.861	Below
3	–1.517	0.646	–2.486	–0.548	Below
4	–1.144	0.580	–2.014	–0.275	Below
5	–0.834	0.537	–1.640	–0.028	Below
6	–0.562	0.509	–1.324	0.201	Below
7	–0.313	0.489	–1.047	0.420	Below
8	–0.081	0.476	–0.795	0.634	Near/Met
9	0.142	0.469	–0.561	0.845	Near/Met
10	0.360	0.466	–0.339	1.058	Near/Met
11	0.577	0.467	–0.124	1.278	Near/Met
12	0.798	0.473	0.088	1.508	Near/Met
13	1.027	0.485	0.300	1.754	Near/Met
14	1.270	0.503	0.516	2.024	Above
15	1.536	0.530	0.741	2.331	Above
16	1.837	0.571	0.980	2.694	Above
17	2.199	0.637	1.244	3.155	Above
18	2.676	0.754	1.544	3.807	Above
19	3.436	1.032	1.888	4.985	Above
20	4.686	1.843	1.921	7.450	Above

Table J.12: Grade 8 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.991	1.843	–6.756	–1.226	Below
1	–2.741	1.032	–4.289	–1.193	Below
2	–1.983	0.752	–3.111	–0.854	Below
3	–1.511	0.633	–2.460	–0.561	Below
4	–1.155	0.565	–2.002	–0.307	Below
5	–0.861	0.521	–1.643	–0.079	Below
6	–0.605	0.491	–1.342	0.132	Below
7	–0.375	0.470	–1.080	0.330	Below
8	–0.161	0.455	–0.844	0.522	Near/Met
9	0.042	0.446	–0.627	0.710	Near/Met
10	0.237	0.440	–0.423	0.897	Near/Met
11	0.429	0.438	–0.227	1.086	Near/Met
12	0.621	0.439	–0.037	1.280	Near/Met
13	0.816	0.444	0.150	1.482	Near/Met
14	1.017	0.453	0.337	1.696	Near/Met
15	1.228	0.467	0.527	1.928	Above
16	1.455	0.487	0.724	2.186	Above
17	1.706	0.517	0.931	2.481	Above
18	1.995	0.560	1.155	2.834	Above
19	2.344	0.628	1.403	3.285	Above
20	2.809	0.747	1.689	3.929	Above
21	3.558	1.027	2.017	5.099	Above
22	4.800	1.840	2.040	7.561	Above

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

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.050	1.842	–6.813	–1.287	Below
1	–2.806	1.026	–4.346	–1.267	Below
2	–2.065	0.738	–3.172	–0.958	Below
3	–1.618	0.610	–2.532	–0.703	Below
4	–1.293	0.535	–2.095	–0.491	Below
5	–1.034	0.486	–1.763	–0.305	Below
6	–0.814	0.454	–1.495	–0.134	Below
7	–0.619	0.433	–1.268	0.031	Below
8	–0.437	0.420	–1.067	0.193	Below
9	–0.264	0.413	–0.884	0.356	Below
10	–0.094	0.410	–0.710	0.521	Near/Met
11	0.074	0.411	–0.542	0.690	Near/Met
12	0.243	0.413	–0.376	0.863	Near/Met
13	0.415	0.417	–0.210	1.041	Near/Met
14	0.591	0.422	–0.042	1.225	Near/Met
15	0.773	0.430	0.128	1.417	Near/Met
16	0.961	0.439	0.303	1.620	Near/Met
17	1.159	0.452	0.482	1.837	Above
18	1.370	0.468	0.668	2.072	Above
19	1.599	0.490	0.864	2.334	Above
20	1.854	0.521	1.073	2.634	Above
21	2.147	0.564	1.300	2.993	Above
22	2.502	0.632	1.553	3.450	Above
23	2.973	0.751	1.846	4.099	Above
24	3.729	1.031	2.183	5.275	Above
25	4.977	1.843	2.213	7.740	Above

Table J.14: Grade 11 Life Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.902	1.840	–6.661	–1.143	Below
1	–2.663	1.024	–4.200	–1.127	Below
2	–1.922	0.740	–3.032	–0.811	Below
3	–1.468	0.617	–2.395	–0.542	Below
4	–1.132	0.547	–1.953	–0.311	Below
5	–0.858	0.503	–1.612	–0.103	Below
6	–0.620	0.474	–1.331	0.090	Below
7	–0.406	0.454	–1.087	0.275	Below
8	–0.206	0.441	–0.868	0.456	Below
9	–0.015	0.434	–0.665	0.636	Near/Met
10	0.171	0.430	–0.474	0.816	Near/Met
11	0.356	0.429	–0.288	0.999	Near/Met
12	0.540	0.431	–0.106	1.186	Near/Met
13	0.727	0.434	0.076	1.378	Near/Met
14	0.918	0.440	0.258	1.578	Near/Met
15	1.115	0.449	0.442	1.788	Near/Met
16	1.321	0.460	0.631	2.012	Above
17	1.540	0.476	0.826	2.254	Above
18	1.777	0.498	1.030	2.523	Above
19	2.039	0.528	1.247	2.830	Above
20	2.339	0.571	1.483	3.196	Above
21	2.702	0.638	1.745	3.660	Above
22	3.181	0.757	2.046	4.316	Above
23	3.946	1.035	2.393	5.498	Above
24	5.199	1.845	2.431	7.966	Above

Table J.15: Grade 11 Physical Science Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.144	1.840	–6.904	–1.383	Below
1	–2.902	1.027	–4.442	–1.362	Below
2	–2.154	0.745	–3.272	–1.036	Below
3	–1.692	0.625	–2.630	–0.755	Below
4	–1.346	0.556	–2.180	–0.512	Below
5	–1.063	0.511	–1.829	–0.297	Below
6	–0.819	0.479	–1.538	–0.100	Below
7	–0.600	0.456	–1.285	0.084	Below
8	–0.400	0.439	–1.059	0.258	Below
9	–0.214	0.426	–0.853	0.426	Below
10	–0.036	0.416	–0.661	0.588	Near/Met
11	0.134	0.409	–0.480	0.747	Near/Met
12	0.298	0.403	–0.307	0.903	Near/Met
13	0.459	0.400	–0.140	1.058	Near/Met
14	0.618	0.397	0.022	1.213	Near/Met
15	0.775	0.396	0.181	1.369	Near/Met
16	0.931	0.396	0.338	1.525	Near/Met
17	1.088	0.397	0.493	1.684	Above
18	1.247	0.399	0.648	1.845	Above
19	1.407	0.403	0.803	2.012	Above
20	1.572	0.409	0.958	2.186	Above
21	1.743	0.419	1.115	2.371	Above
22	1.924	0.433	1.275	2.573	Above
23	2.120	0.454	1.439	2.800	Above
24	2.339	0.485	1.612	3.065	Above
25	2.595	0.531	1.799	3.391	Above
26	2.913	0.603	2.009	3.818	Above
27	3.349	0.729	2.256	4.442	Above
28	4.075	1.017	2.550	5.600	Above
29	5.304	1.836	2.550	8.057	Above

Table J.16: Grade 11 Critiquing Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.726	1.831	–6.472	–0.980	Below
1	–2.511	1.007	–4.021	–1.001	Below
2	–1.806	0.715	–2.878	–0.734	Below
3	–1.390	0.587	–2.271	–0.509	Below
4	–1.089	0.515	–1.862	–0.316	Below
5	–0.847	0.471	–1.554	–0.141	Below
6	–0.639	0.443	–1.304	0.025	Below
7	–0.451	0.425	–1.089	0.187	Below
8	–0.275	0.415	–0.897	0.347	Below
9	–0.106	0.408	–0.719	0.506	Near/Met
10	0.059	0.405	–0.548	0.666	Near/Met
11	0.222	0.403	–0.382	0.826	Near/Met
12	0.384	0.402	–0.219	0.986	Near/Met
13	0.545	0.401	–0.057	1.147	Near/Met
14	0.706	0.402	0.104	1.308	Near/Met
15	0.867	0.402	0.264	1.470	Near/Met
16	1.029	0.403	0.425	1.634	Near/Met
17	1.193	0.405	0.585	1.800	Above
18	1.358	0.409	0.745	1.971	Above
19	1.527	0.414	0.906	2.148	Above
20	1.702	0.423	1.068	2.336	Above
21	1.886	0.436	1.232	2.539	Above
22	2.084	0.456	1.400	2.768	Above
23	2.305	0.486	1.576	3.033	Above
24	2.562	0.532	1.765	3.359	Above
25	2.881	0.603	1.976	3.786	Above
26	3.317	0.729	2.224	4.410	Above
27	4.043	1.017	2.518	5.568	Above
28	5.272	1.836	2.518	8.025	Above

Table J.17: Grade 11 Investigating Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–4.337	1.841	–7.098	–1.576	Below
1	–3.095	1.027	–4.635	–1.555	Below
2	–2.348	0.744	–3.464	–1.233	Below
3	–1.890	0.621	–2.822	–0.959	Below
4	–1.551	0.549	–2.375	–0.727	Below
5	–1.276	0.502	–2.029	–0.523	Below
6	–1.041	0.469	–1.744	–0.339	Below
7	–0.834	0.444	–1.500	–0.167	Below
8	–0.644	0.427	–1.284	–0.004	Below
9	–0.468	0.414	–1.089	0.153	Below
10	–0.301	0.405	–0.908	0.307	Below
11	–0.139	0.399	–0.738	0.460	Below
12	0.019	0.396	–0.575	0.613	Near/Met
13	0.175	0.395	–0.417	0.768	Near/Met
14	0.332	0.396	–0.262	0.926	Near/Met
15	0.490	0.399	–0.109	1.088	Near/Met
16	0.650	0.403	0.046	1.255	Near/Met
17	0.815	0.409	0.201	1.428	Near/Met
18	0.985	0.417	0.360	1.611	Near/Met
19	1.164	0.428	0.522	1.805	Above
20	1.352	0.441	0.690	2.014	Above
21	1.554	0.459	0.866	2.243	Above
22	1.775	0.482	1.052	2.499	Above
23	2.023	0.514	1.252	2.794	Above
24	2.309	0.559	1.470	3.148	Above
25	2.659	0.628	1.717	3.601	Above
26	3.124	0.748	2.002	4.246	Above
27	3.877	1.029	2.333	5.420	Above
28	5.121	1.842	2.359	7.883	Above

Table J.18: Grade 11 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower Bound	Upper Bound	Level
0	–3.954	1.844	–6.720	–1.189	Below
1	–2.703	1.034	–4.254	–1.153	Below
2	–1.941	0.756	–3.074	–0.807	Below
3	–1.463	0.638	–2.420	–0.506	Below
4	–1.100	0.571	–1.957	–0.243	Below
5	–0.799	0.529	–1.593	–0.006	Below
6	–0.535	0.500	–1.286	0.215	Below
7	–0.295	0.480	–1.016	0.425	Below
8	–0.072	0.467	–0.772	0.629	Near/Met
9	0.142	0.458	–0.545	0.829	Near/Met
10	0.349	0.453	–0.331	1.029	Near/Met
11	0.554	0.452	–0.124	1.232	Near/Met
12	0.759	0.454	0.078	1.440	Near/Met
13	0.967	0.460	0.278	1.657	Near/Met
14	1.182	0.469	0.479	1.886	Above
15	1.409	0.483	0.684	2.134	Above
16	1.652	0.504	0.896	2.408	Above
17	1.920	0.533	1.120	2.719	Above
18	2.226	0.576	1.362	3.089	Above
19	2.594	0.642	1.630	3.557	Above
20	3.077	0.760	1.938	4.217	Above
21	3.847	1.037	2.291	5.403	Above
22	5.104	1.846	2.335	7.873	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	96,463	54.42	33.80	11.77	54.70	35.30	10.00	51.56	34.33	14.11
Male	48,976	52.95	34.35	12.70	54.34	34.72	10.94	50.41	33.81	15.77
Female	47,481	55.95	33.24	10.81	55.08	35.89	9.03	52.76	34.85	12.39
Am. Indian	188	55.85	28.19	15.96	58.51	29.26	12.23	51.06	31.91	17.02
Asian	10,544	25.70	44.13	30.17	24.48	48.06	27.47	23.54	41.73	34.73
Black	13,482	71.09	24.71	4.20	73.75	22.91	3.34	70.57	24.71	4.72
Hispanic	32,962	70.32	25.15	4.53	70.71	25.50	3.79	67.77	26.59	5.64
Pacific Islander	175	53.71	29.71	16.57	47.43	41.14	11.43	40.00	41.14	18.86
White	35,712	42.86	41.77	15.36	42.58	44.79	12.64	38.63	42.61	18.75
ML–Yes	10,355	90.99	8.46	0.55	90.67	8.89	0.43	87.84	11.33	0.83
ML–No	86,107	50.03	36.85	13.12	50.37	38.47	11.15	47.20	37.09	15.70
EconDis–Yes	37,681	72.82	23.50	3.68	73.50	23.33	3.17	70.37	24.91	4.72
EconDis–No	58,782	42.63	40.41	16.97	42.65	42.97	14.38	39.51	40.36	20.13
SWD–Yes	20,707	72.56	21.88	5.56	74.27	21.16	4.56	70.19	22.82	6.99
SWD–No	75,755	49.46	37.06	13.47	49.35	39.16	11.49	46.47	37.47	16.05
CBT	72,447	48.25	37.82	13.93	48.34	39.83	11.83	45.31	38.08	16.61
PBT	71	63.38	30.99	5.63	63.38	32.39	4.23	57.75	33.80	8.45
TTS	20,677	69.89	24.06	6.05	71.24	23.63	5.13	67.52	24.93	7.54
SP	1,974	93.16	6.64	0.20	91.34	8.36	0.30	89.26	10.44	0.30
SP TTS	1,089	95.22	4.50	0.28	90.45	9.00	0.55	89.07	10.74	0.18
Human Reader	136	80.88	18.38	0.74	91.91	8.09	0.00	88.97	11.03	0.00

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

Group	N	Critiquing % Below	Critiquing % Near/Met	Critiquing % Above	Investigating % Below	Investigating % Near/Met	Investigating % Above	Sensemaking % Below	Sensemaking % Near/Met	Sensemaking % Above
All Students	96,463	53.79	34.55	11.66	55.07	31.85	13.08	51.70	37.93	10.37
Male	48,976	53.01	34.44	12.55	54.20	31.74	14.06	50.38	37.71	11.91
Female	47,481	54.59	34.67	10.74	55.96	31.97	12.07	53.07	38.15	8.78
Am. Indian	188	54.79	29.26	15.96	53.72	27.66	18.62	52.13	34.57	13.30
Asian	10,544	24.27	46.08	29.65	25.41	40.71	33.89	22.52	50.22	27.26
Black	13,482	72.62	23.60	3.78	72.62	22.23	5.15	70.85	26.06	3.09
Hispanic	32,962	69.65	26.07	4.28	71.40	23.47	5.13	67.78	28.14	4.08
Pacific Islander	175	53.14	30.29	16.57	48.57	33.14	18.29	49.14	37.14	13.71
White	35,712	41.66	42.80	15.54	43.08	40.22	16.69	39.25	47.30	13.45
ML–Yes	10,355	90.07	9.49	0.43	90.94	8.54	0.52	89.08	10.36	0.56
ML–No	86,107	49.43	37.56	13.01	50.75	34.66	14.59	47.21	41.24	11.55
EconDis–Yes	37,681	72.45	24.04	3.51	73.69	21.92	4.39	70.97	25.70	3.33
EconDis–No	58,782	41.83	41.29	16.88	43.13	38.22	18.65	39.36	45.76	14.88
SWD–Yes	20,707	72.83	21.72	5.45	73.73	20.10	6.16	70.71	24.21	5.08
SWD–No	75,755	48.58	38.06	13.36	49.97	35.06	14.97	46.51	41.68	11.82
CBT	72,447	47.53	38.67	13.81	48.84	35.73	15.42	45.27	42.48	12.24
PBT	71	56.34	38.03	5.63	59.15	36.62	4.23	66.20	28.17	5.63
TTS	20,677	69.84	24.21	5.95	70.95	22.12	6.93	68.07	26.53	5.41
SP	1,974	90.93	8.76	0.30	93.06	6.79	0.15	90.83	8.92	0.25
SP TTS	1,089	92.75	6.98	0.28	92.38	7.25	0.37	90.63	9.00	0.37
Human Reader	136	82.35	17.65	0.00	90.44	8.82	0.74	87.50	12.50	0.00

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	99,685	62.99	28.58	8.43	67.82	25.23	6.95	60.22	32.13	7.66
Male	51,093	61.41	29.25	9.33	67.36	24.92	7.72	59.76	31.69	8.56
Female	48,533	64.67	27.86	7.47	68.33	25.54	6.12	60.73	32.58	6.69
Am. Indian	172	66.86	26.74	6.40	68.60	23.26	8.14	61.63	31.98	6.40
Asian	10,785	30.46	44.02	25.52	35.89	41.85	22.25	28.17	48.28	23.55
Black	14,322	81.20	16.56	2.24	83.99	14.09	1.92	78.24	19.45	2.31
Hispanic	33,328	79.04	18.32	2.64	82.18	15.72	2.10	76.90	20.98	2.12
Pacific Islander	184	53.26	36.96	9.78	62.50	32.07	5.43	50.54	40.76	8.70
White	37,959	52.06	37.25	10.69	58.96	32.61	8.43	48.60	41.70	9.70
ML–Yes	8,735	94.40	5.28	0.32	94.73	5.08	0.18	93.27	6.43	0.30
ML–No	90,945	59.97	30.82	9.21	65.24	27.16	7.60	57.04	34.60	8.36
EconDis–Yes	36,703	80.64	17.02	2.34	83.81	14.42	1.77	78.28	19.81	1.91
EconDis–No	62,964	52.69	35.33	11.98	58.50	31.53	9.97	49.68	39.32	11.01
SWD–Yes	20,435	80.59	16.09	3.32	82.85	14.15	3.00	78.90	18.30	2.80
SWD–No	79,244	58.45	31.81	9.75	63.95	28.08	7.97	55.40	35.69	8.91
CBT	80,116	58.49	31.83	9.68	63.83	28.16	8.00	55.43	35.64	8.93
PBT	63	80.95	17.46	1.59	87.30	11.11	1.59	82.54	17.46	0.00
TTS	16,308	78.88	17.22	3.89	82.32	14.56	3.12	76.90	20.17	2.92
SP	2,348	93.91	5.75	0.34	93.14	6.73	0.13	94.29	5.66	0.04
SP TTS	755	96.42	3.44	0.13	95.50	4.37	0.13	95.89	3.97	0.13
Human Reader	68	86.76	11.76	1.47	88.24	8.82	2.94	89.71	10.29	0.00

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	99,685	62.96	28.24	8.80	65.50	28.82	5.68	62.87	29.45	7.68
Male	51,093	62.74	27.91	9.35	64.87	28.57	6.55	61.07	30.01	8.91
Female	48,533	63.22	28.58	8.20	66.17	29.08	4.76	64.77	28.86	6.37
Am. Indian	172	65.12	26.16	8.72	66.86	30.23	2.91	64.53	28.49	6.98
Asian	10,785	29.41	43.96	26.63	34.01	46.76	19.23	31.71	45.18	23.11
Black	14,322	81.80	15.53	2.67	80.65	17.80	1.55	80.95	16.99	2.05
Hispanic	33,328	79.14	18.17	2.68	80.04	18.45	1.51	78.84	18.86	2.30
Pacific Islander	184	56.52	32.61	10.87	62.50	32.61	4.89	59.24	34.24	6.52
White	37,959	52.02	36.94	11.04	56.67	36.49	6.84	51.58	38.59	9.83
ML–Yes	8,735	94.49	5.24	0.26	92.88	6.95	0.17	93.96	5.87	0.17
ML–No	90,945	59.93	30.45	9.62	62.87	30.92	6.21	59.88	31.72	8.40
EconDis–Yes	36,703	81.19	16.49	2.31	81.06	17.55	1.39	80.42	17.67	1.92
EconDis–No	62,964	52.33	35.09	12.58	56.42	35.40	8.18	52.63	36.33	11.04
SWD–Yes	20,435	81.43	15.20	3.37	81.42	16.36	2.22	79.54	17.08	3.38
SWD–No	79,244	58.20	31.60	10.20	61.39	32.03	6.58	58.56	32.65	8.79
CBT	80,116	58.29	31.51	10.19	61.45	31.97	6.57	58.34	32.77	8.89
PBT	63	80.95	17.46	1.59	84.13	15.87	0.00	80.95	17.46	1.59
TTS	16,308	79.72	16.65	3.63	80.03	17.54	2.42	78.92	17.85	3.23
SP	2,348	93.65	6.05	0.30	92.04	7.88	0.09	93.95	6.01	0.04
SP TTS	755	96.29	3.58	0.13	93.77	6.23	0.00	95.36	4.50	0.13
Human Reader	68	89.71	5.88	4.41	85.29	14.71	0.00	88.24	8.82	2.94

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	98,881	50.92	37.21	11.87	53.26	33.41	13.33	55.92	28.34	15.74
Male	50,484	51.95	34.62	13.44	55.23	30.47	14.30	56.65	25.82	17.54
Female	48,242	49.90	39.92	10.17	51.26	36.46	12.27	55.23	30.93	13.83
Am. Indian	161	55.90	37.27	6.83	56.52	33.54	9.94	63.35	26.09	10.56
Asian	10,214	22.80	46.62	30.58	23.10	41.58	35.32	23.92	32.70	43.38
Black	13,917	68.89	27.36	3.74	72.17	23.78	4.05	72.50	22.37	5.13
Hispanic	32,576	65.21	30.21	4.58	68.92	26.17	4.91	71.60	22.32	6.08
Pacific Islander	230	42.61	37.39	20.00	43.91	40.00	16.09	46.52	31.74	21.74
White	39,415	40.62	43.72	15.66	41.98	40.43	17.59	45.83	34.17	20.00
ML–Yes	7,090	88.29	11.45	0.25	91.83	7.98	0.18	92.47	7.08	0.45
ML–No	91,785	48.03	39.21	12.77	50.28	35.38	14.35	53.10	29.98	16.92
EconDis–Yes	32,659	66.97	28.79	4.24	70.62	24.91	4.47	72.82	21.64	5.54
EconDis–No	66,216	42.99	41.37	15.63	44.69	37.61	17.70	47.59	31.64	20.77
SWD–Yes	20,089	66.69	26.65	6.67	71.11	22.25	6.65	74.10	18.42	7.48
SWD–No	78,786	46.89	39.91	13.20	48.71	36.26	15.03	51.29	30.87	17.84
CBT	87,127	48.35	38.90	12.75	50.55	35.13	14.32	53.33	29.79	16.88
PBT	135	56.30	31.11	12.59	56.30	30.37	13.33	60.74	23.70	15.56
TTS	8,679	64.92	28.28	6.81	68.45	23.83	7.72	70.24	20.46	9.30
SP	2,232	87.46	12.46	0.09	90.28	9.68	0.04	91.89	7.97	0.13
SP TTS	424	87.74	12.26	0.00	89.86	9.67	0.47	94.10	5.42	0.47
Human Reader	249	63.86	30.92	5.22	69.88	24.50	5.62	70.28	20.88	8.84

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	98,881	51.42	34.94	13.65	55.95	30.07	13.98	52.65	33.72	13.64
Male	50,484	53.33	31.59	15.08	57.04	27.59	15.37	53.24	31.23	15.53
Female	48,242	49.48	38.42	12.10	54.87	32.64	12.49	52.09	36.29	11.62
Am. Indian	161	57.14	34.78	8.07	63.35	27.33	9.32	57.76	34.16	8.07
Asian	10,214	21.59	40.86	37.56	23.84	37.57	38.59	23.14	40.70	36.16
Black	13,917	69.33	26.71	3.97	74.28	21.41	4.31	69.61	25.68	4.71
Hispanic	32,576	66.96	27.92	5.12	71.82	22.97	5.21	67.37	27.20	5.42
Pacific Islander	230	38.26	43.48	18.26	46.96	34.35	18.70	40.87	40.87	18.26
White	39,415	40.52	41.78	17.70	45.19	36.82	17.99	42.65	39.84	17.51
ML–Yes	7,090	91.40	8.38	0.23	93.72	6.09	0.18	86.70	12.83	0.47
ML–No	91,785	48.32	36.99	14.69	53.03	31.92	15.05	50.01	35.33	14.65
EconDis–Yes	32,659	68.54	26.89	4.58	73.31	21.95	4.75	68.47	26.46	5.06
EconDis–No	66,216	42.97	38.91	18.12	47.38	34.08	18.54	44.84	37.30	17.87
SWD–Yes	20,089	69.37	23.85	6.78	73.39	19.83	6.78	69.84	23.21	6.95
SWD–No	78,786	46.84	37.77	15.40	51.50	32.68	15.82	48.26	36.40	15.34
CBT	87,127	48.71	36.66	14.63	53.25	31.73	15.02	50.18	35.23	14.59
PBT	135	49.63	36.30	14.07	61.48	23.70	14.81	57.78	28.89	13.33
TTS	8,679	66.06	25.81	8.13	70.94	21.00	8.05	66.82	24.78	8.40
SP	2,232	90.41	9.45	0.13	93.28	6.68	0.04	84.81	15.10	0.09
SP TTS	424	91.98	7.78	0.24	93.87	5.90	0.24	87.50	11.56	0.94
Human Reader	249	64.26	27.71	8.03	69.48	23.69	6.83	65.86	28.11	6.02

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. The 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, multi-faceted 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 his or her 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 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*

UIN	Domain	Practice	ItemType	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2205M006_01	Life Science	Investigating	MC	0.298	0.110	0.427	0.153	0.435	0.159	0.416	0.151
2205M006_02	Life Science	Critiquing	TE	0.278	0.145	0.320	0.162	0.334	0.141	0.324	0.154
2205M006_05	Life Science	Sensemaking	TE	0.277	0.096	0.410	0.118	0.398	0.110	0.393	0.117
2205M000_02	Earth and Space Science	Sensemaking	TE	0.203	0.121	0.301	0.142	0.322	0.151	0.301	0.156
2205M000_01	Earth and Space Science	Sensemaking	TE	0.210	0.185	0.267	0.206	0.260	0.180	0.258	0.204
2205M000_03	Earth and Space Science	Investigating	MC	0.255	0.128	0.354	0.131	0.348	0.146	0.351	0.148
2205M000_04	Earth and Space Science	Critiquing	TE	0.245	0.174	0.309	0.195	0.315	0.199	0.304	0.200
2205M004_03	Life Science	Sensemaking	TE	0.192	0.063	0.324	0.079	0.307	0.075	0.317	0.088
2205M004_05	Life Science	Critiquing	TE	0.145	0.132	0.199	0.145	0.217	0.168	0.209	0.167
2205M004_07	Life Science	Sensemaking	TE	0.340	0.213	0.400	0.189	0.385	0.189	0.397	0.205
2205B006_01	Earth and Space Science	Investigating	MC	0.378	0.197	0.454	0.226	0.475	0.249	0.459	0.232
2205B006_03	Earth and Space Science	Critiquing	TE	0.179	0.152	0.242	0.158	0.251	0.160	0.238	0.166
2205B006_06	Earth and Space Science	Critiquing	TE	0.054	0.264	0.098	0.305	0.107	0.313	0.102	0.301
2005B008_01	Physical Science	Sensemaking	TE	0.292	0.074	0.452	0.099	0.458	0.107	0.446	0.102
2005B008_02	Physical Science	Critiquing	TE	0.036	0.013	0.124	0.039	0.157	0.041	0.139	0.040
2005B008_03	Physical Science	Investigating	TE	0.122	0.171	0.187	0.169	0.212	0.189	0.206	0.196
2005B008_04	Physical Science	Investigating	TE	0.327	0.076	0.489	0.119	0.480	0.125	0.470	0.109
2005B008_05	Physical Science	Sensemaking	CR	0.346	0.140	0.375	0.132	0.381	0.128	0.389	0.135
1905M076_01	Physical Science	Critiquing	MC	0.505	0.278	0.583	0.321	0.567	0.277	0.573	0.308
1905M076_03	Physical Science	Sensemaking	TE	0.213	0.073	0.351	0.102	0.355	0.099	0.346	0.103
1905M076_05	Physical Science	Sensemaking	TE	0.129	0.098	0.209	0.116	0.221	0.111	0.224	0.109
2005M015_02	Earth and Space Science	Sensemaking	MC	0.404	0.196	0.531	0.214	0.546	0.229	0.541	0.223
2005M015_05	Earth and Space Science	Sensemaking	TE	0.196	0.121	0.275	0.138	0.286	0.151	0.285	0.151
2005M015_04	Earth and Space Science	Sensemaking	MC	0.216	0.206	0.253	0.182	0.271	0.208	0.274	0.212
2005M015_06	Earth and Space Science	Investigating	MC	0.158	0.418	0.186	0.432	0.183	0.437	0.180	0.440

UIN	Domain	Practice	ItemType	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2005M015_07	Earth and Space Science	Investigating	TE	0.114	0.054	0.264	0.083	0.261	0.083	0.255	0.084
2005M015_10	Earth and Space Science	Sensemaking	TE	0.162	0.103	0.270	0.118	0.292	0.119	0.273	0.121
1905M069_01	Physical Science	Investigating	TE	0.211	0.056	0.334	0.082	0.301	0.079	0.322	0.088
1905M069_02	Physical Science	Sensemaking	TE	0.106	0.047	0.215	0.074	0.197	0.091	0.208	0.091
1905M069_05	Physical Science	Investigating	TE	0.064	0.073	0.135	0.075	0.147	0.082	0.146	0.085
2205B000_02	Life Science	Investigating	TE	0.055	0.132	0.104	0.143	0.105	0.124	0.109	0.134
2205B000_03	Life Science	Investigating	MC	0.307	0.218	0.360	0.213	0.382	0.227	0.377	0.231
2205B000_04	Life Science	Investigating	TE	0.143	0.033	0.311	0.061	0.312	0.067	0.307	0.068
2205B000_07	Life Science	Critiquing	CR	0.118	0.018	0.234	0.039	0.226	0.042	0.237	0.045
2205B000_10	Life Science	Sensemaking	MC	0.218	0.230	0.262	0.247	0.252	0.243	0.253	0.233
2205M022_01	Physical Science	Investigating	MC	0.277	0.172	0.363	0.181	0.355	0.168	0.370	0.191
2205M022_03	Physical Science	Investigating	MC	0.265	0.139	0.375	0.143	0.358	0.147	0.377	0.160
2205M022_05	Physical Science	Investigating	TE	0.349	0.203	0.404	0.210	0.373	0.202	0.393	0.211
2105M003_01	Life Science	Investigating	TE	0.102	0.089	0.195	0.118	0.198	0.122	0.202	0.126
2105M003_07	Life Science	Sensemaking	MC	0.471	0.173	0.624	0.197	0.594	0.200	0.591	0.194
2105M003_09	Life Science	Critiquing	TE	0.095	0.102	0.162	0.117	0.183	0.115	0.174	0.117
2305B006_01	Physical Science	Sensemaking	TE	0.158	0.064	0.329	0.116	0.338	0.112	0.307	0.111
2305B006_03	Physical Science	Investigating	MC	0.212	0.134	0.274	0.142	0.298	0.160	0.292	0.161
2305B006_04	Physical Science	Critiquing	TE	0.282	0.293	0.318	0.303	0.340	0.319	0.337	0.319
2305M008_01	Life Science	Investigating	MC	0.301	0.204	0.368	0.231	0.376	0.253	0.375	0.255
2305M008_03	Life Science	Critiquing	TE	0.177	0.289	0.223	0.298	0.214	0.282	0.210	0.295
2305M008_04	Life Science	Sensemaking	TE	0.265	0.225	0.310	0.237	0.315	0.246	0.323	0.242
2205B009_01	Earth and Space Science	Sensemaking	TE	0.320	0.091	0.472	0.147	0.480	0.143	0.463	0.136
2205B009_02	Earth and Space Science	Investigating	TE	0.122	0.156	0.187	0.171	0.191	0.162	0.191	0.169
2205B009_03	Earth and Space Science	Investigating	TE	0.184	0.259	0.206	0.254	0.215	0.249	0.213	0.259
2205B009_04	Earth and Space Science	Sensemaking	TE	0.249	0.105	0.383	0.144	0.344	0.141	0.353	0.137
2205B009_05	Earth and Space Science	Critiquing	CR	0.086	0.022	0.203	0.066	0.195	0.060	0.203	0.061

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	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2308M017_01	Physical Science	Sensemaking	TE	0.105	0.081	0.208	0.093	0.237	0.096	0.204	0.092
2308M017_02	Physical Science	Investigating	MC	0.207	0.185	0.317	0.187	0.334	0.203	0.303	0.197
2308M017_05	Physical Science	Critiquing	MC	0.586	0.222	0.615	0.225	0.605	0.207	0.607	0.210
2308M028_06	Earth and Space Science	Critiquing	TE	0.259	0.114	0.304	0.129	0.305	0.140	0.302	0.125
2308M028_04	Earth and Space Science	Sensemaking	TE	0.141	0.144	0.192	0.133	0.209	0.124	0.201	0.133
2308M028_09	Earth and Space Science	Critiquing	MC	0.371	0.231	0.407	0.226	0.413	0.234	0.408	0.221
2308M028_05	Earth and Space Science	Investigating	MC	0.298	0.256	0.304	0.265	0.318	0.273	0.315	0.265
2308M020_05	Physical Science	Critiquing	TE	0.184	0.087	0.272	0.114	0.249	0.108	0.254	0.109
2308M020_09	Physical Science	Investigating	TE	0.168	0.106	0.226	0.126	0.226	0.107	0.221	0.117
2308M020_11	Physical Science	Investigating	MC	0.420	0.108	0.535	0.136	0.520	0.126	0.512	0.123
2308M049_03	Life Science	Sensemaking	TE	0.172	0.118	0.235	0.112	0.240	0.124	0.235	0.119
2308M049_05	Life Science	Sensemaking	TE	0.132	0.194	0.190	0.191	0.203	0.189	0.194	0.182
2308M049_01	Life Science	Sensemaking	TE	0.087	0.064	0.181	0.064	0.196	0.073	0.176	0.060
2308M049_07	Life Science	Investigating	MC	0.395	0.239	0.440	0.245	0.456	0.244	0.445	0.247
2108B006_03	Earth and Space Science	Investigating	TE	0.066	0.176	0.107	0.197	0.117	0.217	0.107	0.205
2108B006_06	Earth and Space Science	Sensemaking	TE	0.142	0.104	0.197	0.090	0.223	0.096	0.209	0.102
2108B006_09	Earth and Space Science	Critiquing	TE	0.091	0.039	0.238	0.052	0.281	0.065	0.248	0.059
2208B003_01	Life Science	Sensemaking	TE	0.220	0.072	0.317	0.104	0.328	0.104	0.310	0.086
2208B003_05	Life Science	Sensemaking	TE	0.233	0.080	0.323	0.089	0.321	0.085	0.315	0.078
2208B003_07	Life Science	Investigating	TE	0.068	0.094	0.121	0.102	0.129	0.104	0.121	0.093
2208B003_09	Life Science	Critiquing	TE	0.220	0.230	0.239	0.201	0.241	0.202	0.239	0.221
2208B003_11	Life Science	Critiquing	CR	0.289	0.065	0.350	0.065	0.360	0.068	0.350	0.063
2208M016_03	Life Science	Critiquing	MC	0.205	0.144	0.261	0.133	0.289	0.171	0.276	0.164
2208M016_05	Life Science	Investigating	TE	0.297	0.261	0.303	0.290	0.308	0.295	0.309	0.280
2208M016_04	Life Science	Investigating	TE	0.359	0.046	0.499	0.064	0.461	0.058	0.464	0.055
2208M016_08	Life Science	Critiquing	MC	0.425	0.213	0.404	0.193	0.367	0.168	0.410	0.206

UIN	Domain	Practice	Item Type	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2208M021_03	Earth and Space Science	Critiquing	TE	0.164	0.115	0.217	0.124	0.197	0.116	0.206	0.122
2208M021_05	Earth and Space Science	Sensemaking	TE	0.252	0.155	0.307	0.137	0.314	0.145	0.309	0.147
2208M021_09	Earth and Space Science	Sensemaking	MC	0.221	0.164	0.256	0.167	0.261	0.177	0.260	0.169
2208M021_10	Earth and Space Science	Critiquing	TE	0.132	0.026	0.276	0.040	0.244	0.042	0.249	0.036
2108B007_08	Physical Science	Critiquing	TE	0.353	0.083	0.504	0.100	0.471	0.100	0.480	0.107
2108B007_03	Physical Science	Investigating	MC	0.180	0.226	0.189	0.260	0.167	0.246	0.185	0.242
2108B007_07	Physical Science	Sensemaking	TE	0.111	0.243	0.154	0.266	0.155	0.263	0.153	0.257
2108M000_02	Earth and Space Science	Investigating	TE	0.184	0.080	0.267	0.100	0.265	0.111	0.260	0.107
2108M000_06	Earth and Space Science	Critiquing	TE	0.135	0.164	0.188	0.177	0.183	0.165	0.178	0.166
2108M000_03	Earth and Space Science	Sensemaking	TE	0.077	0.042	0.167	0.047	0.185	0.060	0.168	0.060
2008M000_02	Life Science	Critiquing	TE	0.030	0.050	0.096	0.064	0.104	0.078	0.095	0.066
2008M000_04	Life Science	Investigating	MC	0.262	0.178	0.260	0.172	0.240	0.166	0.258	0.175
2008M000_01	Life Science	Sensemaking	TE	0.211	0.219	0.231	0.227	0.228	0.240	0.234	0.233
2008B007_11	Physical Science	Investigating	TE	0.181	0.089	0.239	0.089	0.207	0.084	0.218	0.078
2008B007_03	Physical Science	Sensemaking	TE	0.129	0.154	0.194	0.177	0.181	0.158	0.177	0.164
2008B007_09	Physical Science	Sensemaking	TE	0.364	0.064	0.506	0.103	0.486	0.078	0.469	0.066
2008B007_07	Physical Science	Critiquing	CR	0.073	0.007	0.127	0.014	0.127	0.023	0.125	0.016
2208M051_13	Physical Science	Investigating	TE	0.220	0.133	0.310	0.145	0.296	0.152	0.310	0.147
2208M051_16	Physical Science	Critiquing	TE	0.207	0.083	0.332	0.081	0.369	0.100	0.333	0.092
2208M051_17	Physical Science	Critiquing	TE	0.152	0.080	0.228	0.090	0.227	0.084	0.221	0.085
2308M023_02	Earth and Space Science	Critiquing	MC	0.384	0.210	0.435	0.209	0.443	0.234	0.433	0.221
2308M023_03	Earth and Space Science	Investigating	TE	0.079	0.066	0.140	0.072	0.155	0.073	0.154	0.075
2308M023_07	Earth and Space Science	Sensemaking	TE	0.223	0.239	0.252	0.224	0.253	0.193	0.241	0.211
2308M000_04	Life Science	Sensemaking	MC	0.328	0.213	0.371	0.233	0.372	0.249	0.376	0.243
2308M000_06	Life Science	Critiquing	TE	0.100	0.109	0.160	0.136	0.165	0.149	0.165	0.152
2308M000_08	Life Science	Investigating	MC	0.162	0.133	0.200	0.119	0.197	0.116	0.218	0.124
2308M027_02	Physical Science	Investigating	TE	0.165	0.231	0.211	0.253	0.239	0.286	0.224	0.277
2308M027_03	Physical Science	Critiquing	TE	0.110	0.104	0.170	0.107	0.167	0.089	0.173	0.087

UIN	Domain	Practice	Item Type	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2308M027_05	Physical Science	Sensemaking	TE	0.090	0.022	0.190	0.035	0.188	0.046	0.188	0.033
2308M027_01	Physical Science	Investigating	TE	0.235	0.104	0.305	0.116	0.275	0.113	0.284	0.110
2308B006_01	Life Science	Sensemaking	TE	0.083	0.206	0.131	0.209	0.135	0.209	0.134	0.207
2308B006_08	Life Science	Sensemaking	TE	0.138	0.178	0.180	0.176	0.174	0.154	0.178	0.154
2308B006_07	Life Science	Sensemaking	TE	0.152	0.235	0.176	0.258	0.179	0.253	0.169	0.256
2308B003_01	Earth and Space Science	Sensemaking	TE	0.171	0.068	0.268	0.092	0.275	0.095	0.266	0.089
2308B003_03	Earth and Space Science	Investigating	TE	0.104	0.096	0.175	0.102	0.166	0.122	0.157	0.113
2308B003_11	Earth and Space Science	Critiquing	TE	0.270	0.046	0.443	0.079	0.415	0.077	0.407	0.062
2308B003_07	Earth and Space Science	Investigating	TE	0.101	0.220	0.146	0.206	0.168	0.198	0.155	0.193
2308B003_05	Earth and Space Science	Critiquing	CR	0.191	0.019	0.296	0.027	0.268	0.028	0.281	0.028

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	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2211M016_01	Physical Science	Sensemaking	TE	0.161	0.077	0.276	0.161	0.289	0.170	0.260	0.167
2211M016_08	Physical Science	Sensemaking	TE	0.057	0.101	0.116	0.128	0.126	0.122	0.134	0.160
2211M016_02	Physical Science	Investigating	MC	0.340	0.184	0.444	0.234	0.440	0.236	0.430	0.230
2211M016_06	Earth and Space Science	Critiquing	TE	0.181	0.159	0.259	0.195	0.294	0.216	0.266	0.201
2011M054_04	Life Science	Investigating	TE	0.058	0.032	0.138	0.078	0.145	0.093	0.143	0.092
2011M054_06	Life Science	Critiquing	MC	0.229	0.107	0.360	0.161	0.344	0.165	0.332	0.170
2011M054_08	Life Science	Critiquing	TE	0.280	0.128	0.324	0.156	0.335	0.166	0.334	0.194
2011M031_03	Earth and Space Science	Sensemaking	TE	0.103	0.164	0.153	0.195	0.167	0.209	0.163	0.206
2011M031_05	Earth and Space Science	Sensemaking	MC	0.284	0.243	0.285	0.225	0.282	0.208	0.298	0.252
2011M031_04	Earth and Space Science	Critiquing	TE	0.206	0.215	0.268	0.263	0.276	0.263	0.257	0.233
2011M031_06	Earth and Space Science	Investigating	MC	0.384	0.225	0.418	0.236	0.389	0.221	0.417	0.256
2311M016_01	Physical Science	Sensemaking	TE	0.300	0.201	0.355	0.235	0.344	0.222	0.353	0.257
2311M016_02	Physical Science	Critiquing	TE	0.039	0.024	0.145	0.070	0.149	0.084	0.146	0.091
2311M016_04	Life Science	Sensemaking	MC	0.284	0.143	0.375	0.185	0.366	0.175	0.352	0.172
2311M016_07	Physical Science	Investigating	TE	0.129	0.281	0.163	0.284	0.185	0.291	0.166	0.275
2311M023_01	Life Science	Investigating	MC	0.439	0.280	0.522	0.285	0.526	0.281	0.516	0.322
2311M023_06	Life Science	Critiquing	TE	0.055	0.193	0.109	0.239	0.131	0.253	0.114	0.249
2311M023_03	Life Science	Sensemaking	MC	0.253	0.184	0.370	0.219	0.396	0.241	0.366	0.258
2211B006_02	Earth and Space Science	Investigating	TE	0.155	0.123	0.238	0.164	0.269	0.166	0.238	0.175
2211B006_05	Earth and Space Science	Investigating	TE	0.114	0.147	0.201	0.170	0.224	0.195	0.205	0.213
2211B006_06	Earth and Space Science	Investigating	TE	0.512	0.170	0.635	0.222	0.632	0.226	0.612	0.207
2211B006_09	Earth and Space Science	Critiquing	CR	0.274	0.065	0.446	0.081	0.448	0.097	0.416	0.101
2211B006_12	Physical Science	Critiquing	TE	0.109	0.046	0.240	0.100	0.253	0.106	0.229	0.107
2311M026_01	Physical Science	Sensemaking	MC	0.220	0.163	0.314	0.187	0.302	0.195	0.295	0.190
2311M026_03	Physical Science	Sensemaking	TE	0.203	0.086	0.336	0.134	0.333	0.142	0.328	0.156
2311M026_02	Physical Science	Critiquing	MC	0.246	0.236	0.285	0.291	0.279	0.301	0.276	0.297
2311M026_05	Physical Science	Investigating	TE	0.080	0.239	0.109	0.259	0.115	0.238	0.110	0.251

UIN	Domain	Practice	Item Type	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2311M026_06	Physical Science	Investigating	MC	0.187	0.192	0.279	0.199	0.283	0.219	0.274	0.202
2111M004_05	Life Science	Sensemaking	MC	0.479	0.294	0.512	0.297	0.502	0.287	0.510	0.307
2111M004_03	Life Science	Investigating	MC	0.159	0.165	0.244	0.207	0.232	0.213	0.230	0.214
2111M004_02	Life Science	Sensemaking	TE	0.041	0.307	0.080	0.321	0.085	0.329	0.081	0.332
2111M004_06	Life Science	Critiquing	TE	0.130	0.195	0.211	0.240	0.205	0.206	0.205	0.241
2011M028_01	Life Science	Investigating	MC	0.481	0.318	0.561	0.358	0.561	0.351	0.536	0.333
2011M028_03	Earth and Space Science	Sensemaking	TE	0.284	0.184	0.358	0.229	0.360	0.223	0.367	0.253
2011M028_04	Life Science	Critiquing	MC	0.204	0.290	0.187	0.274	0.190	0.280	0.212	0.313
2011M028_05	Earth and Space Science	Investigating	TE	0.192	0.090	0.369	0.156	0.401	0.156	0.353	0.158
2111M020_01	Physical Science	Investigating	MC	0.414	0.245	0.517	0.287	0.492	0.263	0.510	0.302
2111M020_03	Physical Science	Sensemaking	TE	0.127	0.254	0.158	0.254	0.162	0.253	0.170	0.244
2111M020_05	Physical Science	Critiquing	TE	0.172	0.068	0.358	0.136	0.364	0.142	0.347	0.152
2211M008_01	Earth and Space Science	Sensemaking	TE	0.072	0.073	0.122	0.100	0.126	0.108	0.122	0.121
2211M008_03	Earth and Space Science	Critiquing	MC	0.375	0.255	0.392	0.248	0.372	0.238	0.390	0.255
2211M008_07	Earth and Space Science	Investigating	TE	0.194	0.168	0.233	0.185	0.237	0.173	0.233	0.172
2111B002_01	Life Science	Critiquing	TE	0.086	0.049	0.201	0.089	0.222	0.095	0.207	0.110
2111B002_03	Physical Science	Sensemaking	TE	0.229	0.109	0.348	0.143	0.330	0.132	0.327	0.145
2111B002_07	Life Science	Investigating	TE	0.172	0.103	0.253	0.124	0.253	0.138	0.252	0.149
2111B002_10	Life Science	Sensemaking	TE	0.055	0.040	0.163	0.077	0.176	0.071	0.154	0.079
2111B002_11	Life Science	Investigating	CR	0.166	0.017	0.384	0.073	0.364	0.071	0.352	0.086
2011M003_06	Life Science	Investigating	MC	0.228	0.179	0.394	0.239	0.398	0.236	0.368	0.236
2011M003_01	Life Science	Sensemaking	TE	0.111	0.122	0.195	0.131	0.214	0.128	0.199	0.158
2011M003_03	Life Science	Critiquing	MC	0.157	0.198	0.206	0.203	0.220	0.236	0.211	0.222
2011M003_04	Life Science	Investigating	MC	0.172	0.180	0.247	0.209	0.265	0.222	0.254	0.239
2011M003_05	Life Science	Critiquing	TE	0.175	0.112	0.291	0.144	0.300	0.143	0.283	0.150
2311M021_02	Earth and Space Science	Critiquing	TE	0.055	0.080	0.127	0.121	0.137	0.107	0.105	0.124
2311M021_04	Earth and Space Science	Sensemaking	MC	0.334	0.305	0.378	0.316	0.387	0.300	0.367	0.293
2311M021_06	Earth and Space Science	Investigating	MC	0.274	0.233	0.366	0.284	0.399	0.277	0.336	0.236

UIN	Domain	Practice	Item Type	ML		Econ Dis		SWD		ACC	
				P_FT	P_UF	P_FT	P_UF	P_FT	P_UF	P_FT	P_UF
2011M071_04	Physical Science	Sensemaking	MC	0.445	0.277	0.444	0.262	0.420	0.235	0.449	0.283
2011M071_01	Physical Science	Sensemaking	MC	0.343	0.187	0.443	0.204	0.431	0.195	0.426	0.238
2011M071_03	Physical Science	Critiquing	TE	0.100	0.095	0.208	0.138	0.216	0.153	0.191	0.148
2011M071_05	Physical Science	Investigating	MC	0.308	0.160	0.410	0.187	0.403	0.185	0.393	0.221
2311M011_03	Physical Science	Critiquing	TE	0.104	0.060	0.211	0.104	0.203	0.097	0.200	0.124
2311M011_05	Earth and Space Science	Sensemaking	MC	0.523	0.367	0.581	0.363	0.579	0.334	0.564	0.360
2311M011_07	Earth and Space Science	Investigating	MC	0.439	0.275	0.479	0.260	0.462	0.239	0.498	0.285
2211M022_01	Earth and Space Science	Investigating	MC	0.179	0.121	0.278	0.186	0.286	0.172	0.274	0.170
2211M022_02	Earth and Space Science	Sensemaking	TE	0.117	0.082	0.216	0.105	0.212	0.094	0.212	0.117
2211M022_08	Earth and Space Science	Critiquing	MC	0.294	0.228	0.304	0.220	0.294	0.223	0.313	0.236
2211B001_01	Physical Science	Investigating	TE	0.178	0.214	0.216	0.201	0.250	0.226	0.235	0.233
2211B001_03	Physical Science	Investigating	TE	0.131	0.116	0.203	0.173	0.219	0.193	0.201	0.184
2211B001_07	Physical Science	Investigating	MC	0.200	0.174	0.253	0.190	0.248	0.204	0.255	0.206
2211B001_05	Physical Science	Critiquing	TE	0.174	0.071	0.353	0.126	0.343	0.118	0.340	0.139
2211B001_10	Physical Science	Critiquing	CR	0.044	0.010	0.128	0.045	0.125	0.050	0.127	0.064

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

APPENDIX O: Parameter Estimates from the Confirmatory Factor Analyses Using the 2023 NJSLA–S Tests

Table O.1: Grade 5 Domain Model Parameter Estimates

Earth and Space Science				
UIN	Estimate (standardized)	R-Squared	Standard Error	z
1905M008_01	0.724	0.525	0.003	280.960
1905M008_05	0.738	0.545	0.003	290.327
1905M008_06	0.684	0.468	0.003	240.680
2105M015_06	0.502	0.252	0.004	136.144
2105M015_05	0.116	0.013	0.004	26.8990
2105M015_04	0.472	0.223	0.004	121.544
1905M005_01	0.619	0.383	0.003	198.879
1905M005_03	0.726	0.527	0.003	266.165
1905M005_04	0.333	0.111	0.004	84.6760
2205B003_05	0.659	0.434	0.003	225.515
2205M011_04	0.485	0.235	0.004	134.361
2205M011_02	0.713	0.508	0.003	252.868
2205M011_01	0.364	0.133	0.005	80.3304
2205B009_01	0.667	0.445	0.003	235.220
2205B009_02	0.578	0.335	0.004	155.195
2205B009_03	0.313	0.098	0.004	70.6840
2205B009_04	0.544	0.296	0.003	160.471
2205B009_05	0.733	0.538	0.002	336.396
Life Science				
UIN	Estimate (standardized)	R-Squared	Standard Error	z
1905B007_01	0.810	0.655	0.002	349.585
1905B007_03	0.781	0.610	0.003	301.870
1905B007_05	0.759	0.576	0.003	302.071
1905B007_08	0.717	0.514	0.002	360.551
1905B007_10	0.837	0.700	0.002	408.575
2205M006_01	0.518	0.268	0.003	153.024
2205M006_02	0.340	0.116	0.004	86.502
2205M006_05	0.584	0.341	0.003	181.298
2205M004_03	0.778	0.605	0.002	331.180
2205M004_05	0.535	0.286	0.004	143.259
2205M004_07	0.418	0.175	0.004	110.983
1905B009_01	0.799	0.639	0.002	357.005
1905B009_02	0.872	0.760	0.002	475.842
1905B009_05	0.367	0.135	0.004	83.307

1905M044_02	0.711	0.506	0.003	268.190
1905M044_03	0.602	0.363	0.003	178.558
1905M044_04	0.525	0.276	0.004	144.707
		Physical Science		
UIN	**Estimate (standardized)**	**R-Squared**	**Standard Error**	**z**
1905M040_01	0.616	0.380	0.003	190.444
1905M040_03	0.577	0.333	0.003	179.148
1905M040_05	0.523	0.274	0.004	147.132
2205M012_01	0.661	0.436	0.003	224.461
2205M012_03	0.598	0.358	0.003	181.922
2205M012_04	0.076	0.006	0.005	16.033
2205B003_01	0.179	0.032	0.004	42.191
2205B003_02	0.548	0.300	0.004	155.338
2205B003_03	0.350	0.122	0.004	84.022
2205B003_04	0.691	0.478	0.003	272.448
2205M022_01	0.569	0.324	0.003	171.595
2205M022_03	0.610	0.373	0.003	194.492
2205M022_05	0.338	0.114	0.004	85.539
1905M076_01	0.298	0.089	0.004	75.711
1905M076_03	0.705	0.497	0.003	252.510
1905M076_05	0.691	0.478	0.003	219.227

Note. All parameter estimates were significant at $p < .001$; R-squared is the squared standardized estimate and is interpreted as the proportion of variance in the item explained by the latent subscore group (Kline, 2011).

Table O.2: Grade 8 Domain Model Parameter Estimates

Earth and Space Science				
UIN	**Estimate (standardized)**	**R-Squared**	**Standard Error**	**z**
1908M033_02	0.435	0.189	0.004	116.435
1908M033_03	0.382	0.146	0.004	100.991
1908M033_04	0.477	0.228	0.003	138.596
1908M026_01	0.627	0.393	0.004	161.802
1908M026_04	0.443	0.196	0.004	120.141
1908M026_06	0.393	0.154	0.004	103.104
2208M028_06	0.329	0.108	0.004	83.593
2108M015_06	0.650	0.423	0.003	223.841
2108M015_09	0.652	0.426	0.003	222.697
2108M015_10	0.287	0.082	0.004	68.763
2108M015_02	0.550	0.303	0.003	171.560
2208M021_03	0.491	0.242	0.004	127.687
2208M021_05	0.491	0.241	0.004	134.589
2208M021_09	0.442	0.196	0.004	112.478
2208M021_10	0.743	0.552	0.003	285.193
2108B006_01	0.764	0.584	0.003	296.508
2108B006_03	0.491	0.241	0.004	109.766
2108B006_06	0.536	0.287	0.004	133.312
2108B006_09	0.701	0.491	0.003	243.530
2108B006_11	0.706	0.498	0.003	240.758
Life Science				
UIN	**Estimate (standardized)**	**R-Squared**	**Standard Error**	**z**
1908M030_01	0.499	0.249	0.004	138.545
1908M030_02	0.530	0.281	0.003	151.711
1908M030_05	0.520	0.271	0.004	143.633
2208M028_02	0.163	0.027	0.005	34.705
2208M028_07	0.368	0.135	0.005	76.997
2208M028_09	0.398	0.158	0.004	103.650
2008M000_02	0.658	0.433	0.004	183.757
2008M000_04	0.340	0.115	0.004	82.986
2008M000_01	0.270	0.073	0.005	59.943
2208B003_01	0.635	0.403	0.003	211.358
2208B003_05	0.615	0.378	0.003	195.354
2208B003_07	0.678	0.460	0.004	186.370
2208B003_09	0.297	0.088	0.004	68.478
2208B003_11	0.673	0.452	0.002	304.369
1908M003_02	0.372	0.139	0.005	76.257

1908M003_07	0.445	0.198	0.005	91.829
1908M003_08	0.561	0.314	0.004	159.331
2008M015_04	0.463	0.214	0.004	109.834
2008M015_05	0.287	0.082	0.004	65.560
2008M015_06	0.638	0.407	0.003	208.105
2008M015_09	0.473	0.223	0.004	126.428
2108M027_07	0.364	0.132	0.004	93.636
2108M027_01	0.632	0.399	0.003	183.158
2108M027_03	0.649	0.422	0.003	218.901
		Physical Science		
UIN	**Estimate (standardized)**	**R-Squared**	**Standard Error**	**z**
2108B003_02	0.524	0.275	0.003	162.369
2108B003_08	0.348	0.121	0.004	89.924
2108B003_07	0.452	0.204	0.004	110.297
1908B000_03	0.437	0.191	0.005	95.233
1908B000_04	0.543	0.295	0.004	146.188
1908B000_08	0.559	0.312	0.004	143.319
1908B000_11	0.771	0.595	0.002	431.725
1908B000_12	0.515	0.266	0.004	134.023
1908M005_02	0.532	0.283	0.004	133.031
1908M005_03	0.676	0.457	0.004	155.138
1908M005_05	0.507	0.257	0.004	143.554
2008M001_05	0.535	0.286	0.003	168.818
2008M001_01	0.766	0.587	0.003	295.965
2008M001_08	0.459	0.211	0.004	119.642
2208M051_13	0.490	0.240	0.004	135.586
2208M051_16	0.573	0.329	0.003	180.145
2208M051_17	0.563	0.317	0.003	161.488
2108B007_01	0.640	0.410	0.003	204.429
2108B007_08	0.537	0.288	0.003	157.762
2108B007_03	0.169	0.029	0.005	35.636
2108B007_07	0.372	0.139	0.005	80.845

Note. All parameter estimates were significant at $p < .001$; R-squared is the squared standardized estimate and is interpreted as the proportion of variance in the item explained by the latent subscore group (Kline, 2011).

Table O.3: Grade 11 Domain Model Parameter Estimates

UIN	Estimate (standardized)	R-Squared	Standard Error	z
Earth and Space Science				
2111M000_06	0.440	0.193	0.004	118.810
2111M000_02	0.642	0.412	0.003	214.218
2111M000_09	0.577	0.333	0.003	179.757
2111M000_08	0.596	0.355	0.003	185.437
1911M002_01	0.458	0.210	0.004	103.654
1911M002_04	0.335	0.112	0.004	74.815
1911M002_05	0.356	0.126	0.004	90.458
1911M119_06	0.671	0.450	0.003	223.175
1911M119_02	0.361	0.130	0.004	86.318
1911M119_05	0.586	0.343	0.003	175.468
1911M079_02	0.296	0.088	0.004	71.518
1911M079_03	0.613	0.376	0.003	192.913
1911M079_04	0.640	0.410	0.004	168.825
2211M008_01	0.524	0.274	0.004	129.244
2211M008_03	0.364	0.132	0.004	91.943
2211M008_07	0.328	0.108	0.004	77.210
2211B006_02	0.577	0.333	0.004	159.979
2211B006_05	0.622	0.387	0.004	172.225
2211B006_06	0.646	0.418	0.003	215.614
2211B006_09	0.678	0.460	0.002	309.365
Life Science				
UIN	**Estimate (standardized)**	**R-Squared**	**Standard Error**	**z**
1911B009_01A	0.687	0.472	0.003	253.523
1911B009_03A	0.806	0.650	0.002	329.176
1911B009_05A	0.633	0.400	0.004	173.082
1911B009_07A	0.638	0.407	0.002	275.169
1911B009_09A	0.513	0.263	0.005	112.466
1911M023_02	0.666	0.443	0.003	217.491
1911M023_05	0.694	0.482	0.003	248.230
1911M023_06	0.352	0.124	0.004	87.764
1911M023_07	0.284	0.081	0.004	68.173
2211M003_01	0.465	0.216	0.004	124.552
2211M003_03	0.468	0.219	0.004	119.172
2211M003_04	0.572	0.327	0.005	115.560
2211M003_02	0.291	0.085	0.004	68.385
2211M003_05	0.644	0.415	0.004	169.881
2011M003_06	0.523	0.274	0.003	151.310

2011M003_01	0.657	0.432	0.003	193.732
2011M003_03	0.403	0.162	0.004	92.935
2011M003_04	0.527	0.278	0.004	138.844
2011M003_05	0.699	0.488	0.003	237.381
2111M004_05	0.368	0.135	0.004	95.651
2111M004_02	0.417	0.174	0.005	79.571
2111M004_03	0.618	0.382	0.004	169.605
2111M004_06	0.434	0.189	0.004	98.416
		Physical Science		
UIN	**Estimate (standardized)**	**R-Squared**	**Standard Error**	**z**
1911M028_01	0.349	0.122	0.004	91.325
1911M028_03	0.601	0.361	0.004	158.105
1911M028_04	0.279	0.078	0.004	68.087
1911M028_06	0.332	0.110	0.004	80.447
HS18060_01	0.756	0.572	0.003	245.410
HS18060_03	0.587	0.344	0.003	181.617
HS18060_04	0.659	0.434	0.003	234.292
HS18060_06	0.529	0.280	0.005	114.533
1911M124_02	0.617	0.381	0.003	179.744
1911M124_05	0.284	0.080	0.004	72.038
1911M124_10	0.621	0.385	0.003	201.757
1911M124_01	0.284	0.081	0.004	64.201
2211B000_01	0.542	0.294	0.004	151.947
2211B000_07	0.515	0.265	0.004	131.015
2211B000_03	0.298	0.089	0.004	71.049
2211B000_12	0.727	0.528	0.002	302.092
2011M071_04	0.378	0.143	0.004	99.830
2011M071_01	0.666	0.444	0.003	234.901
2011M071_02	0.425	0.180	0.004	97.043
2011M071_03	0.635	0.403	0.003	193.968
2011M071_05	0.561	0.314	0.003	164.288
2211B006_12	0.732	0.536	0.003	257.220
2011M010_01	0.259	0.067	0.004	57.557
2011M010_02	0.447	0.200	0.004	115.284
2011M010_03	0.357	0.128	0.004	88.614
2011M010_05	0.402	0.162	0.005	86.636

Note. All parameter estimates were significant at $p < .001$; R-squared is the squared standardized estimate and is interpreted as the proportion of variance in the item explained by the latent subscore group (Kline, 2011).