New Jersey

Student Learning Assessment—Science

(NJSLA–S)

TECHNICAL REPORT

Grades 5, 8, and 11

2019

March 2020

PTM XXXX.XX

State of New Jersey
Department of Education

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Tables and Figures

Tables and Figures

Tables and Figures

Tables and Figures

Tables and Figures

Tables and Figures

Tables and Figures

Tables and Figures

PART 1: INTRODUCTION

The purpose of this Technical Report is to provide information about the technical characteristics of the 2019 administration of the New Jersey Student Learning Assessment for 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 but 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 2019 test development, administration, scoring, and reporting are provided in Parts 2, 3, 4, and 10, respectively, 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, respectively.

Data for the analyses presented in this Technical Report were collected from the NJSLA–S spring administration in May 2019.

- Analyses in Part 5 of this report, Standard Setting, are based on test results from a priority sample due to the short time between the test administration and the 2019 NJSLA–S Standard Setting meeting. The priority sample was representative of the entire state student population in terms of various demographic information including gender, ethnicity, English learner status, disability status, etc.

- 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) 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 *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), 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 and from students' interactions with teachers on a daily basis, as well as from 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—the results of which form the foundation of this Technical Report—was the assessment's baseline year.

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 towards the students' performance 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 associated with them contribute to the overall scale score; however, there is no specific reported CCC performance indicator for the NJSLA−S.)

1.2.1 Content Domains and Scientific Practices

Although the NJSLA−S is a unidimensional test, six distinct, foundational sub-categories represent the three science content domains (Earth and Space, Life, and Physical) and the three scientific practices (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 performance on 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 as well as the topics that are 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

DCI Topic Description	
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 for 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 three PS DCIs along with the associated detailed topics for each.

Table 1.2.3: Physical Science DCI

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

Scientific practices. The *Framework* (2012) contains eight different 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, Katch-Singer, and Pelletier (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 understanding 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, to engage in argument, and to 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 performance on the NJSLA–S is described using scale scores and performance levels. Each grade level has its own grade-specific scale that represents a composite score of student performance on the three NJSLS–S dimensions (DCIs, SEPs, and CCCs). Student performance is classified into four grade-specific performance levels based on the NJSLA–S Performance 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 performance. A scale score is a conversion of the raw score (that is, the total number of points a student earned on the test as a whole), using a predetermined mathematical algorithm, to permit legitimate and meaningful comparisons over time and across grades. As such, they provide the best generalized information about overall performance. The total scores in science are reported as scale scores with a range of 100 to 300.

- **Performance 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 performance expectations in science. Based on test results, a student's performance is categorized as being at one of four performance levels, each of which is defined by a student's scale score and used to report overall student performance on the NJSLA–S. Grade-appropriate Performance Level Descriptors (PLDs) translate these performance levels into words. They describe the knowledge, skills, and practices that students should know and be able to demonstrate at each of the performance levels, Level 1 through Level 4. Each performance 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, 2019

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 to be below the state minimum level of proficiency. They demonstrate minimal or partial understanding of the DCIs and SEPs. Students at this performance level may need additional instructional support, which could be in the form of individual or programmatic intervention.

Student performance 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 performance classifications are primarily meant to provide teachers, schools, and administrators with feedback as to the specific knowledge, skills, and abilities (KSAs) that their students displayed on the NJSLA–S. Individual students and their parents and teachers receive student level data on these subscores. However, the importance of the individual student-level subscore data is secondary to its aggregated interpretation.

1.3 Organizational Support

The New Jersey Department of Education's Office of Assessments coordinates the development and implementation of 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 contractor for NJSLA–S Grades 5, 8, and 11, is responsible for all aspects of the testing program, including activities such as program management, development of test materials (test items, test booklets, answer documents, and ancillary materials), and psychometric support, including standard setting. MI's other activities include enrollment verification; distribution of all materials; receiving, scanning, editing, and scoring the answer documents; scoring constructed-response items; and creating, generating, and distributing all score reports of test results to students, schools, districts, and the state.

PART 2: TEST DEVELOPMENT

The NJSLS−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) outline 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 2019. The development of procedures and materials for test administration and scoring is covered in Parts 2 and 3. Given that this is the first year of NJSLA–S operational testing, test revisions were not documented.

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 ability levels at the holistic level and at 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 Part 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 either one or two points; each constructed response (CR) item is worth four points. Each constructed response item is scored using an item-specific rubric. The table summarizes the numbers 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 PBA	Grade 5 MSA	Grade 8 PBA	Grade 8 MSA	Grade 11 PBA	Grade 11 MSA
PS	**Investigating** *AQDP, PACI, UMCT*	1–2	3–5	1–2	4–7	1–2	4–8
PS	**Sensemaking** *DUM, AID, CEDS*	1–2	3–5	1–2	4–7	1–2	4–8
PS	**Critiquing** *EAE, OECI*	1–2	3–5	1–2	4–7	1–2	4–8
PS	***Total Items***	*3–5*	*11–13*	*3–5*	*14–18*	*3–5*	*15–21*
LS	**Investigating** *AQDP, PACI, UMCT*	1–2	3–5	1–2	4–7	1–2	4–8
LS	**Sensemaking** *DUM, AID, CEDS*	1–2	3–5	1–2	4–7	1–2	4–8
LS	**Critiquing** *EAE, OECI*	1–2	3–5	1–2	4–7	1–2	4–8
LS	***Total Items***	*3–5*	*11–13*	*3–5*	*14–18*	*3–5*	*15–21*
ESS	**Investigating** *AQDP, PACI, UMCT*	1–2	3–5	1–2	4–7	1–2	4–8
ESS	**Sensemaking** *DUM, AID, CEDS*	1–2	3–5	1–2	4–7	1–2	4–8
ESS	**Critiquing** *EAE, OECI*	1–2	3–5	1–2	4–7	1–2	4–8
ESS	***Total Items***	*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 and one field test. 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 myriad psychometric constraints that are discussed in Part 2.4 of this technical report.

Each MSA and PBA component of a unit is linked to naturally occurring phenomena which provide the impetus for scenarios or simulations. The students are provided the scenario or simulation 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, 4 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	• 2 one-point TEs • 1 two-point TEs • 1 four-point CR
Total # items, MSA: 12 Total points, MSA: 12				Total # items, PBA: 4 Total points, PBA: 8
Total # items, Unit: 16 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 that portion of the assessment which 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 six stimuli, and each stimulus is associated with two to five 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 knowledge, skills, and abilities to a greater degree of cognitive depth; 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. In 2019 NJDOE required that the PBA section contain eight total points, with four of those points coming from the CR item.

2.1.3 Item Types

Three 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; fill in the blank; drag and drop; multiple selection; and ordering. Some TE items are scored dichotomously; others are rubric-dependent and can be worth multiple points.
- CR items are open-ended questions designed to elicit a student response to a range of KSAs that are challenging to measures with traditional MC or TE items. All CR items are rubric-dependent and scored by a human rater.

Table 2.1.2 describes each NJSLA–S item type.

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: Short Answer	Type a brief constrained response to the question.
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: Fill in the Blank	Type a response to fill in a blank within a text-based prompt.
TE: Scatter Plot	Plot one or more points on a graph.
TE: Bar Graph	Drag each bar to the correct length on the graph.
TE: Line Graph	Plot one or more lines on a graph.
TE: Slider	Slide an area within a graphic to change its length.
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 extremely rigorous and involves item writers, content specialists, editors, graphic artists, programmers, scoring experts, and psychometricians. The resulting products are phenomenon-based scenarios (PBS) 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 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, and had to review and approve each item that appears on the NJSLA–S 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, Johnstone, & Thurlow, 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 PBS 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, and geological features, amongst many others.)

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 six 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 student instructions that are clear, that its readability is appropriate, and that its 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, then it's 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 teachers who compose the New Jersey Science Advisory Committee (NJSAC). In 2019 the NJSAC comprised a diverse group of New Jersey science educators representing 19 of the 21 New Jersey counties. The districts each NJSAC member represents, as well as the counties they come from are presented in Appendix B.

The NJSAC are the final authority on universal design principle #2: precisely defined constructs. They 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 predominantly took place in-person at locations approved by NJDOE. Occasionally, it was necessary to conduct the meetings via 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 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 makes sure that all students have the opportunity to show what they know regardless of their background or the test form they took. 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 nine NJBSC members, all nine taught special education status students, seven specialized in teaching students designated as English learners, and five were bilingual. Collectively, they had over 100 years of teaching experience. Just like at the 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 onto 10 to 18 different 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 specialist 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 5,000 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 gets a copy of 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 and Pearson'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 that is 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 being possibly biased against 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 Part 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 Part 2.4.

It should be noted that the 2019 NJSLA–S test construction process was only partially representative of future processes because it had to occur prior to the establishment of the cut scores during standard setting. Future additional statistical constraints will include targeting previous test information functions — described in Part 8.2.1 — to maximize the reliability of performance level classification accuracy.

2.3.1 Test Construction – First Draft

The first step in the NJSLA–S test construction process involves MI's psychometric 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 the sections below.

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 psychometric 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 3 cut score, because it is the place on the scale associated with whether or not 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 3 cut score as possible.

*Item discrimination***.** Item discrimination refers to the ability of the item to discriminate between students who have done well on the test versus those who did not. A poorly discriminating item could indicate ineffective measurement of the NJSLA–S scale and reduces test form reliability. Item discrimination is measured via the item-total correlation, which can range from -1.0 to 1.0; items with item-total correlations that are below 0.2 are only selected for placement on the operational test form if no other viable options are available.

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, discrimination, and lower asymptote. During test construction, the mean item infit, outfit, and 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 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 absolutely 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. Items are selected to minimize the median time spent on the test. If the median time spent on items is greater than the total test time minus 30 minutes, then items that are taking students too long are replaced by items that take less time, unless no other options are available.

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 item difficulty is as close as possible to the Level 3 cut score
Item Discrimination	Items have item-total correlations 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 items < (total test time - 30 minutes)
DIF	• 'B' items are only used if necessary • 'C' items are not used.

2.3.2 Test Construction Content Review

After MI's psychometric 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 review is an iterative process between content specialists and psychometricians. If, during the review, content specialists identify items that are too similar or that cue each other then they alert MI's psychometric staff, and the items are replaced. The content review then resumes until the test matches NJSLA–S' content and statistical constraints.

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 2019 NJSLA–S Test Construction

2019 was the first year of operational testing for the NJSLA–S. Overall, the test construction process achieved forms that matched the balance required by the test blueprint. The science content domains were well-balanced at each grade level. Moreover, all grade levels had three eight-point PBA sections representing each of the three content domains, and all of these met the requirement that no more than 50% of the MSA items be MC. However, there were some constraints that were more difficult to achieve. At all three grade levels, it was challenging to identify enough Critiquing items — that were also acceptable from a content and statistical perspective — to balance out the three scientific practice reporting categories. Furthermore, the content constraint requiring that the PBA section include eight points worth of items meant that some of the items within the PBA sections were subpar.

A final test construction content constraint that was not met was the balance between the three content domains across the three scientific practices reporting categories, as shown in the test blueprint in Section 2.1.1. The items associated with each scientific practice were meant to be balanced across all content domains. Table 2.4.1 shows this lack of balance. At each grade level one content domain was over-represented for each scientific practice. For instance, of the 17 Investigating points available on the grade 5 test, nine were aligned to the Physical Science content domain, whereas only 3 of 17 were aligned to Earth and Space Science.

To iteratively improve the ability of the NJSLA–S to foster valid interpretations and uses of test scores, the test construction issues noted above were addressed by adjusting item development procedures and revising the PBA point total constraints. First, NJSLA–S item development has focused on increasing the proportion of Critiquing and, to a lesser extent, Investigating items. Next, the rules requiring that each PBA have items totaling eight points have been relaxed so that items comprising a PBA can total as few as six points. This rule adjustment was made at the recommendation of content specialists, psychometricians, NJDOE, and the NJSAC to accommodate the differences among the myriad content standards. Finally, balancing the scientific practices across all content domains has been made a primary objective of current NJSLA–S item development. All items currently under development, as described in this paragraph, are scheduled to be field tested in the spring of 2022 and incorporated into the next round of test construction.

Table 2.4.1: Points Available and Intercorrelations by Domain and Practice

Grade	Practice	Earth	Life	Physical
5	Investigating	3	5	9
5	Sensemaking	15	9	5
5	Critiquing	3	8	3
8	Investigating	5	6	12
8	Sensemaking	14	12	8
8	Critiquing	2	7	6
11	Investigating	4	6	12
11	Sensemaking	18	10	9
11	Critiquing	5	9	5

2.4.1 Grade 5 Test Construction

At grade 5 the science content domains were balanced, as illustrated in Table 2.4.2. The least balanced content domain was Physical Science, and it still made up 17 points of the 60 total score points. Each content domain had one PBA section devoted to it. The scientific practices were less balanced, with only 14 out 60 points being allocated to the Critiquing reporting category. Despite being less than ideal, the 14 points were still enough to produce reliable measures of student Critiquing abilities. Other content considerations that were met included: MC items only made up 13 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. Of the 11 major DCI clusters, eight were represented. Table 2.4.2 details the item 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: 2019 NJSLA–S Grade 5 Item and Point Totals by Reporting Category

Domains/Practices	MC Items	TE Items	CR Items	Items	Points
Earth and Space	7	9	1	17	21
Life	3	15	1	19	22
Physical	3	10	1	14	17
Total - Domains	13	34	3	50	60
Investigating	2	10	1	13	17
Sensemaking	7	14	2	23	29
Critiquing	4	10	0	14	14
Total – Practices	13	34	3	50	60

Table 2.4.3: 2019 NJSLA–S Grade 5 DCIs

DCI	Items	Points
ESS1	3	3
ESS2	14	18
ESS3	0	0
LS1	7	10
LS2	6	6
LS3	0	0
LS4	6	6
PS1	3	3
PS2	3	3
PS3	8	11
PS4	0	0

Table 2.4.4: 2019 NJSLA–S Grade 5 SEPs

SEP	Items	Points
AQDP	3	3
PACI	6	9
UMCT	4	5
DUM	3	3
AID	15	18
CEDS	5	8
EAE	12	12
OECI	2	2

Table 2.4.5: 2019 NJSLA–S Grade 5 CCCs

CCC	Items	Points
C & E	9	9
E & M	9	12
Patterns	15	18
S & SM	3	4
S, P, & Q	6	6
SC	3	3
SF	5	8

The statistical constraints for the 2019 Grade 5 NJSLA–S operational test form were met. The item difficulty constraint was relaxed, as this was the first year of operational testing and no cut scores had been established. All items had item-total correlations above the 0.2 threshold, and each of the model fit statistics averaged close to their ideal values. The median test time of 74.8 minutes was well below the 105-minute threshold, and out of 200 DIF classifications there were zero 'C' values and only 5 'B' values. 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: 2019 NJSLA–S Grade 5 Test Construction Statistics

Statistic	Average	Target	Flags
Item Difficulty	0.12	N/A	N/A
IT Correlation	0.41	> 0.35	0
Infit	0.98	1.00	0
Outfit	0.98	1.00	3
PCM Discrim.	1.02	1.00	1
Lower Asymptote	0.02	0.00	3
Median Time	74.80	< 105	N/A

Table 2.4.7: 2019 NJSLA–S Grade 5 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	47	3	0
White/Black	49	1	0
White/Hispanic	50	0	0
White/Asian	49	1	0

2.4.2 Grade 8 Test Construction

The science content domains were almost equal at grade 8. The least balanced content domain was Earth and Space Science, and it still made up 21 points of the 72 total score points. Each content domain had one PBA section devoted to it. The scientific practices were less balanced with only 15 out 72 points allocated to the Critiquing reporting category. Despite being less than ideal, 15 points was still enough to produce reliable measures of student Critiquing abilities. Other content considerations that were met included: MC items only made up 18 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 at least three items. Table 2.4.8 details the item 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: 2019 NJSLA–S Grade 8 Item and Point Totals by Reporting Category

Domains/Practices	MC Items	TE Items	CR Items	Items	Points
Earth and Space	5	11	1	17	21
Life	6	14	1	21	25
Physical	7	14	1	22	26
Total - Domains	18	39	3	60	72
Investigating	13	6	1	20	23
Sensemaking	3	25	1	29	34
Critiquing	2	8	1	11	15
Total – Practices	18	39	3	60	72

Table 2.4.9: 2019 NJSLA–S Grade 8 DCIs

DCI	Items	Points
ESS1	6	10
ESS2	5	5
ESS3	6	6
LS1	5	5
LS2	9	13
LS3	3	3
LS4	4	4
PS1	3	3
PS2	4	4
PS3	10	14
PS4	5	5

Table 2.4.10: 2019 NJSLA–S Grade 8 SEPs

SEP	Items	Points
AQDP	8	8
PACI	5	5
UMCT	7	10
DUM	9	10
AID	11	11
CEDS	9	13
EAE	9	13
OECI	2	2

Table 2.4.11: 2019 NJSLA–S Grade 8 CCCs

CCC	Items	Points
C & E	11	12
E & M	14	15
Patterns	7	7
S & SM	7	11
S, P, & Q	8	11
SC	4	7
SF	9	9

The statistical constraints for the 2019 Grade 8 NJSLA–S operational test form were not ideal. As at grade 5, the item difficulty constraint was relaxed because this was the first year of operational testing and no cut scores had been established. However, six grade 8 items were flagged for having item-total correlations below the 0.2 threshold, including one item with an extremely low value of 0.02, indicating poor or no discrimination. The infit, outfit, and PCM discrimination model fit statistics all drifted from their ideal values of 1.00. On the positive side, the median test time of 82.8 minutes was well below the 105-minute threshold, and out of 240 DIF classifications there were zero “C” values and only 3 “B” values. 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: 2019 NJSLA–S Grade 8 Test Construction Statistics

Statistic	Average	Target	Flags
Item Difficulty	–0.57	N/A	N/A
IT Correlation	0.35	> 0.35	6
Infit	0.96	1.00	1
Outfit	0.94	1.00	2
PCM Discrim.	1.14	1.00	10
Lower Asymptote	0.01	0.00	2
Median Time	82.80	< 105	N/A

Table 2.4.13: 2019 NJSLA–S Grade 8 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	59	1	0
White/Black	58	2	0
White/Hispanic	60	0	0
White/Asian	60	0	0

2.4.3 Grade 11 Test Construction

The grade 11 content domains were the closest of all grades to being equal. Out of 78 total score points the three content domains ranged from 25 to 27 points each. Each content domain had one PBA section. The scientific practices were less balanced with only 20 out 78 points being allocated to the Critiquing reporting category. Despite being less than ideal, the 20 points were still enough to produce reliable measures of student Critiquing abilities. Other content considerations that were met included: MC items only made up 25 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. The seven CCCs were well-balanced except for the SF category which was only represented by 1 item. Table 2.4.14 details the item 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 at grade 11.

Table 2.4.14: 2019 NJSLA–S Grade 11 Item and Point Totals by Reporting Category

Domains/Practices	MC Items	TE Items	CR Items	Items	Points
Earth and Space	8	15	1	24	27
Life	7	13	1	21	25
Physical	10	12	1	23	26
Total – Domains	25	40	3	68	78
Investigating	8	13	0	21	22
Sensemaking	14	19	1	34	37
Critiquing	3	8	2	13	19
Total – Practices	25	40	3	68	78

Table 2.4.15: 2019 NJSLA–S Grade 11 DCIs

DCI	Items	Points
ESS1	5	5
ESS2	7	7
ESS3	12	15
LS1	4	4
LS2	7	11
LS3	3	3
LS4	7	7
PS1	7	7
PS2	8	8
PS3	3	3
PS4	5	8

Table 2.4.16: 2019 NJSLA–S Grade 11 SEPs

SEP	Items	Points
AQDP	6	7
PACI	5	5
UMCT	10	10
DUM	8	8
AID	18	18
CEDS	7	10
EAE	8	14
OECI	6	6

Table 2.4.17: 2019 NJSLA–S Grade 11 CCCs

CCC	Items	Points
C & E	13	16
E & M	6	6
Patterns	11	14
S & SM	11	12
S, P, & Q	12	15
SC	14	14
SF	1	1

The statistical constraints for the 2019 Grade 11 NJSLA–S operational test form were closer to ideal than grade 8. Again, the item difficulty constraint was relaxed because this was the first year of operational testing and no cut scores had been established. No grade 11 items were flagged for having item-total correlations below the 0.2 threshold. The infit, outfit, and PCM discrimination model fit statistics all drifted away from their ideal values of 1.00, but fewer items were flagged than at grade 8. The median test time was only 48.92 minutes, which was 100 minutes below the 150-minute constraint, indicating the high school students' lack of motivation on the field test. Of 272 DIF classifications there were zero "C" values and only 8 "B" values. 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: 2019 NJSLA–S Test Construction Statistics

Statistic	Average	Target	Flags
Item Difficulty	–0.28	N/A	N/A
IT Correlation	0.44	> 0.35	0
Infit	0.97	1.00	1
Outfit	0.94	1.00	1
PCM Discrim.	1.11	1.00	5
Lower Asymptote	0.01	0.00	1
Median Time	48.92	< 150	N/A

Table 2.4.19: 2019 NJSLA–S Test Construction DIF Classifications

Groups	A	B	C
Male/Female	65	3	0
White/Black	67	1	0
White/Hispanic	67	1	0
White/Asian	65	3	0

2.5 2019 NJSLA–S State of the Item Bank

Upon the completion of the 2019 test construction process MI’s psychometricians analyzed the state of 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 2020 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 2019 NJSLA–S 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" (NJDOE, 2019, slide 2). Information about the administration of NJSLA–S is available in the Test Coordinator Manual (TCM). That information is not fully replicated here, but the following elements are of importance to this technical report. The NJSLA–S TCM can be read in full at: NJSLA–S TCM. The DTCs were trained on the following topics:

- 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

Table 3.1.1 shows the NJSLA–S 2019 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 2019 Grades 5, 8 and 11 Science Testing Window

Grade	CBT	PBT	Testing Time
5	5/6/19–6/7/19	5/6/19–5/17/19	45 minutes/4 units
8	5/6/19–6/7/19	5/6/19–5/17/19	45 minutes/4 units
11	5/6/19–6/7/19	5/6/19–5/17/19	60 minutes/4 units

3.2 Test Security and Administration Procedures

This section provides information regarding the NJSLA–S test administration procedures. Descriptions of both the CBT and PBT procedures are detailed below. For a complete description of all test administration activities refer to the NJSLA–S TCM.

3.2.1 Computer-Based Testing

The CBT NJSLA–S test 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 (TA) 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 test 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} to manage each test session; they can monitor the progress of each of their students and lock and unlock units. PearsonAccess^{next} 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 TSCs 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 PearsonAccess^{next}.	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

Tasks	TCM Section(s)
Schedule and supervise make-up testing.	Sections 2.4.2 and 4.1.5
Create make-up test sessions in PearsonAccess^{next}.	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 follow 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 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–S 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 NJSLA–S State Contact 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 Accessibility Features and Accommodations Manual* at the following link: nj.mypearsonsupport.com/resources/manuals/NJSLASpring2019AFA.pdf.

- **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 PearsonAccess^{next} 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): state.nj.us/education/assessment/district
    - Displaying any resource (e.g., posters, models, displays, teaching aids) 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 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. Previously in New Jersey, federal requirements were met by testing grades 4 and 8 students with the NJASK test; grade 11 students were tested via the NJBCT. The NJSLA–S Test Coordinator Manual states:

> All students in grades 5, 8, and 11 must be administered the NJSLA–S, regardless of whether they are enrolled in a science course. Students who are full-time home-

> schooled or full-time at a private or parochial school are **not** eligible to take any statewide assessment, including the NJSLA–S. This excludes special education students who attend an approved private school for the disabled in which tuition is the financial responsibility of the local education agency. (p. viii)

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 Accessibility Features and Accommodations Manual (AF&A Manual)* is available online at nj.mypearsonsupport.com/resources/manuals/NJSLASpring2019AFA.pdf. 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–S *AF&A Manual* (2019) 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 following link: measinc-nj-science.com.

3.4.1.1 Text-to-Speech. The most used NJSLA–S accessibility feature is Test-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 verify 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–S *AF&A Manual* (2019) 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, each student who received a PBT version of the NJSLA–S had an appropriate accommodation. A comprehensive explanation of each NJSLA–S accommodation is presented in the NJSLA–S *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 are 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 Parts 3.4.2.1.1 through 3.4.2.1.3. The Paper-Based Test (PBT) form adaptation process is presented in Part 3.4.2.1.4. Following adaptation, NJDOE verifies each accommodated test form.

3.4.2.1.1 Spanish. All Spanish accommodations were made by Teneo Linguistics Company (TLC). TLC received the NJDOE-approved tests and created the translations within ABBI. Once the items were translated, a NJ teacher committee of 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 2019 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 2 ASL content specialist translators and 1 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 of 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 exactly 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 being 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 don't translate well from English to Spanish, while others are challenging to braille, for example. In 2019 all items were deemed adequately accommodated by external reviewers, content specialists, and NJDOE. However, there were two minor errors in the construction of the accommodation process that required a relatively small number of students to have their scale scores based on one fewer item. One item was incorrectly accommodated both on the grade 5 online Assistive Technology and American Sign Language forms and on the grade 11 online Spanish test form. The two incorrectly accommodated items were removed from the affected forms' scoring criteria, and separate scale score tables were calculated to ensure the students receiving said forms received the proper scale scores. The procedures for calculating the separate scale score tables are detailed in Part 7: Equating and Scaling.

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- and hand-scoring 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. The data from the student responses is then screened via Pearson's Customer Data Quality (CDQ) team. The CDQ team verifies the accuracy of the student responses and metadata within two file types: the Summative File and the item response file (IRF). Verification steps include validating variable acceptable ranges, computing raw overall scores and subscores, validating ID numbers and unique item numbers (UINs), and flagging inconsistent student records for investigation. Once the data have been verified the files are placed on a Secure File Transfer Protocol site from which they are retrieved by MI's IT group who then prepare the files for psychometric analysis and the adjudication process.

4.1.1 Adjudication

Adjudication involves the careful review of all student responses to an item to ensure 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. MI's psychometric department uses the Summative Record File (SRF) and the IRF to analyze 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 thousands of different student responses. The student response data is used to produce one file for each operational item that contains a Response ID, the point-value associated with it (i.e., 0, 1 or 2), the total number and percentage of students selecting each response, the text of the response (retrieved from the item's XML coding), and the item-total correlation associated with each response option that was selected more than 100 times. Item means and item-total correlations are also calculated at the item level, and items are flagged for aberrant behavior. 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 specialist 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 then key changes are submitted to NJDOE, and upon approval subsequently corrected in the SRF and IRF. These steps are essential to ensuring 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 NJ Science assessments. To supplement this core pool, our recruiting team contacts other scorers in our 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, our recruiting team reviews applications—including prospective scorers' resumes, references, proof of degree, and recognition of scorer requirements—before offering employment. All of 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 previous projects and also 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 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, NJ teacher representatives, and MI management personnel, as well as the Scoring Director responsible for each content area.

4.2.3 Field Test Range Finding

Prior to field-test scoring, content committees consisting of NJDOE personnel, NJ teacher representatives, and MI leadership personnel meet in New Jersey 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.

4.2.4 Developing Scoring Guides

After the range finding meeting, 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 training annotations connect to highlighted sections of the student response in PowerPoint presentations used for training, 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 take notes on the training papers in preparation for discussion with the scorers, and the Scoring Directors counsel Team Leaders on application of the rubric and training techniques. 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 which 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 trained on all relevant training materials. Scorers have access to all training materials during live scoring. The same training protocol is followed for both site-based and remote scorers.

After scorers have signed contracts and nondisclosure forms and 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.

After Team Leaders and scorers have thoroughly studied the rubric and anchor responses, all scoring personnel log in to MI's secure Scoring Resource Center (SRC). The SRC 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 of the 2019 NSLA-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	228	9	3
8	161	9	3
11	178	9	3

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 Monitors. 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 our handscoring program is maintained through ongoing monitoring by experienced scoring leadership. Scoring Directors and Team Leaders are skilled in detecting scoring trends and in 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 hand-scoring 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 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 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, we conduct a blank check which resets scores when a 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 when two scorers assign non-adjacent scores, mismatched condition codes, or a combination of a condition code a numeric score provides 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.

Rescoring is conducted automatically for any student who scores within one raw score point of the proficient cut score. MI reviews student responses to constructed response items and verifies the original scores or makes changes if warranted. Scores are never lowered during the automatic rescoring process even if a lower score results. LEAs do not need to request rescoring.

4.2.8 Automatic Rescores

As shown in Part 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 within one raw score point of the proficient cut score. MI reviews student responses to constructed-response items and verifies the original scores or makes changes where warranted. Scores are never lowered during the automatic rescoring process even if a lower score results. LEAs do not need to request rescoring. Table 4.2.2 provides automatic rescoring information for all three grade levels. All open-ended/constructed response item types were scored by a single rater.

Table 4.2.2: 2019 Automatic Rescores

Grade	Eligible for Automatic Rescore	Number of Changes	Percent Changed (of those Eligible)
5	2,665	234	8.78
8	1,736	119	6.85
11	1,365	56	4.10

4.3 Quality Control

During the Quality Control (QC) process MI, Pearson, and NJDOE collaborate to confirm that the processing of student tests and test scores was done correctly, a sample of several hundred students is selected for multiple QC checks. The sample of student tests are then manually reviewed and independently scored by NJDOE staff to compare to the students' score reports and to their records in the Summative Record File (SRF). The sample includes all forms and includes various demographic variables such as gender, ethnicity, EL status, and disability status. Other score reports and the Summary Data File show the distribution of performance levels and the average scores at the school, district, or state levels. These results are verified programmatically using the SRF. The following sections detail the timing and processes used in the three QCs to ensure student test scores are accurate.

4.3.1 QC #1

After testing and scoring were complete, and after standard setting, NJDOE conducted the first QC check. These meetings were held August 12–16, 2019. To begin the QC process, MI provided NJDOE with a Key Information Sheet (KIS) for each student's test. This spreadsheet shows the student information associated with a test (from the SRF), the key for each machine-scored item, and a spot to record the points earned for each item. First, NJDOE verified the student information against the information in PearsonAccess^{next} (PAN). This allows NJDOE to indirectly verify these fields from the SRF against the source of the data. There are over 50 fields involved, and they are organized on the KIS in a way that makes it faster to compare against PAN than scrolling through the SRF.

Next, NJDOE reviews the student's responses to each selected-response item, scores them against the key, and records the score on the KIS. This task may be somewhat complex for technology-enhanced selected-response items which are not as straightforward as multiple-choice items. For open-ended items, MI provides NJDOE with the scores from the handscoring system to record on the spreadsheet. The KIS automatically tallies the student's overall points, as well as the points in each domain and practice. Any discrepancies between these totals and the SRF will require scrutiny of the points earned for each item. The KIS helps narrow down the problem to a particular domain, practice, and unit.

4.3.2 QC #2

QC #2 was conducted from November 18–21, 2019. After the New Jersey Board of Education's approval of the results from standard setting, and the subsequent creation of the scale scores, Pearson imported the scale score tables and produced the score reports. The subsequent QC #2 involved completing the KIS with the student's scale score, overall performance level, and the performance level for each content domain and scientific practice based on the raw scores recorded in QC #1. Afterwards, NJDOE compared this information between the KIS and the Individual Student Report, Individual Student Label and School Student Roster. This stage provided NJDOE with confidence that each piece of student-level information on these reports was accurately derived from the various sources of test data.

4.3.3 QC #3

QC #3 was conducted from November 15–22, 2019. Certain numbers shown on the Individual Student Report and School Student Roster were not calculated directly from individual tests. These numbers include averages of scores at the school, district, or state level; or percentages of students achieving a performance level. In addition, there are several score reports that only show aggregated data. These values were verified in QC #3 by performing the same calculations on the SRF. This part is not necessarily limited to the schools QC sample and the work can be done simultaneously with QC #2, although any problems discovered with the SRF may require the work to be repeated.

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. In accordance with *Standard 5.21,* which 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 sections below provide evidence that the processes and procedures used to accomplish the 2019 NJSLA–S Standard Setting were conducted using best practices. The executive summary from the 2019 NJSLA–S Standard Setting Report is presented in Appendix D.

5.1 Standard Setting Overview

The 2019 NJSLA–S Standard Setting represented the first new performance standards set at each grade band in a minimum of 11 years. The 2019 grade 5 standard setting was the first attempt at setting elementary school performance standards since 2005. The last middle school standard setting took place in 2000; the last high school standard setting (for NJBCT) was in 2008. Given the vast changes to the New Jersey science content standards — represented by the NJSLS–S — it was essential for the valid interpretations and uses of test scores that new performance standards were established.

5.2 Standard-Setting Procedures

The major standard-setting procedures that were utilized by MI prior to and during the 2019 NJSLA–S Standard Setting were reviewed, edited, and eventually approved by both NJDOE and the New Jersey Technical Advisory Committee (NJTAC). Those procedures include the framework and development of the NJSLA–S Performance Level Descriptors (PLDs), the methodology for recommending cut scores, and the way the 2019 NJSLA–S Standard Setting was to be conducted, which included the external review of the process by an NJTAC member. The input of New Jersey educators was instrumental to all major steps in the process. The New Jersey Science Advisory Committee (NJSAC) served as the primary authors of the Performance Level Descriptors (PLDs), and a diverse collection of New Jersey educators participated in standard setting.

5.2.1 Performance Level Descriptors

Performance Level Descriptors (PLDs) are an essential component in the interpretation of educational assessments (Egan, Schneider, & Ferrara, 2012). Cizek and Bunch (2007) define them as “verbal elaborations of the knowledge, skills, or attributes of test takers within a performance level” (p. 46). MI psychometricians and content specialists — with the approval of the NJDOE and NJTAC — worked with educators in New Jersey to construct the NJSLA–S PLDs using Egan et al.’s proposed PLD framework. Four different types of PLDs comprise this framework: Policy PLDs, Range PLDs, Threshold (aka Target) PLDs, and Reporting PLDs. Descriptions of the various PLDs, as well as explanations of how they were constructed, are presented in Parts 5.2.1.1 through 5.2.1.4. All of the various PLDs produced for the NJSLA–S were created and approved by NJDOE and the NJSAC with guidance from MI psychometricians and content specialists.

5.2.1.1 Policy PLDs. The first step of Egan et al.'s (2012) PLD process is to create the Policy PLDs, the role of which is to communicate the overall policy goals of the assessment program. One set of Policy PLDs is used for all grade levels within the assessment program. For the NJSLA–S these Policy PLDs were created by NJDOE during 2018 and amounted to two sentences at each performance level that loosely defined the degree of understanding expected of students. Aside from assisting to conceptualize the performance levels early in the test development process, the Policy PLDs were also presented to the NJSAC as a framework for conceptualizing the Range PLDs and to the standard-setting panelists to help them understand the policy goals of NJDOE. The final NJSLA–S Policy PLDs used in the 2019 NJSLA–S Standard Setting are presented in Appendix E.

5.2.1.2 Range PLDs. The next step in the PLD process after the creation of the Policy PLDs is the construction of the Range PLDs, the purpose of which is to inform test development. Whereas the Policy PLDs are two sentences and assessment program-specific, the Range PLDs are extremely detailed and grade-specific. The end product is meant to define the range of KSAs available to the content specialists, item writers, and NJSAC members charged with either making the test items or providing guidance in their development.

The NJSLA–S Range PLDs were developed over multiple-day meetings during 2018 and 2019. The process involved content specialists, psychometricians, NJDOE staff, and the NJSAC. The initial meetings were led by MI's psychometric department and based on Schneider and Egan's (2014) recommended procedures. Each meeting was held over a two-day period. For each of the 11 DCIs and eight SEPs, the NJSAC worked to describe the KSAs expected of students at each of the four NJSLA–S performance levels. After the initial two-day meetings, MI's content specialists, psychometricians, and NJDOE worked with the NJSAC to finalize the documents prior to the next step in the process: the creation of the Threshold PLDs. Table 5.2.1 summarizes the Range and Threshold PLD development meetings. The March 2018 Grade 8 Range PLD meeting was shortened due to inclement weather.

Table 5.2.1: Summary of PLD Development Meetings

Grade	PLD Type	Dates	NJ educators
5	Range	3/13/18 – 3/14/18	10
8	Range	3/20/2018	9
11	Range	3/27/18 – 3/28/18	13
5	Range + Threshold	4/9/2019 – 4/12/2019	11
8	Range + Threshold	4/23/2019 – 4/26/2019	8
11	Range + Threshold	4/2/2019 – 4/5/2019	11
All Grades	Threshold	5/2/2019 – 5/14/2019	4

5.2.1.3 Threshold PLDs. The role of the Threshold (or Target) PLDs is to provide direction to standard-setting panelists during the standard-setting meeting. The Threshold PLDs represent a subset of the Range PLDs that is specifically targeted to the KSAs associated with students who have just barely entered into a given performance level (except for the lowest level). Standard-setting panelists apply the Threshold PLDs during the standard-setting procedure to determine which students have met NJDOE's performance expectations. All decisions made from the standard-setting meeting must be firmly grounded in the Threshold PLDs; thus, these PLDs are extremely important.

For the NJSLA–S the Threshold PLDs were created during multiple-day meetings after the Range PLD process had completed. Just like the Range PLDs, the Threshold PLD process involved content specialists, psychometricians, NJDOE staff, and the NJSAC. The meetings were led by MI's content specialists with guidance from MI's psychometric department and based on the procedures outlined by Schneider and Egan (2014). The NJSAC members were instructed to identify the subset of KSAs within the Range PLDs representing students who have just barely entered into Levels 2–4. Specific focus was placed on Level 3, because it represents the KSAs needed to have met NJDOE's expectations.

Once the NJSAC completed the Threshold PLDs, a subgroup of NJSAC members who were familiar with multiple grade levels reviewed the documents in unison to ensure they were consistent across grade levels. Those meetings were conducted virtually over two weeks in May 2019. Finally, the Threshold PLDs were submitted for NJDOE approval. The final NJSLA–S Threshold PLDs used in the 2019 NJSLA–S Standard Setting are presented in Appendix E.

5.2.1.4 Reporting PLDs. Reporting PLDs are meant to assist the users of test information with test interpretation. They are created after the completion of the standard-setting meeting and represent the KSAs that test takers displayed. The Reporting PLDs are typically short, grade- and performance-level-specific descriptions of the KSAs students can exhibit at each level. In the context of the NJSLA–S these were written by NJDOE to be only performance level-specific, with the intended audience being parents, teachers, and administrators. The final Reporting PLDs that appear on the NJSLA–S Individual Student Reports (ISR) are presented in Appendix E.

5.2.2 Bookmark Procedure

In accordance with a standard-setting plan approved by both NJDOE and the NJTAC, MI employed the Bookmark procedure for the 2019 NJSLA–S Standard Setting. This widely used standard-setting procedure is thoroughly documented in the NJTAC-approved standard-setting plan and elsewhere (cf., Cizek & Bunch, 2007; Lewis, Mitzel, Green, 1996; Lewis, Mitzel, Mercado, & Shultz, 2012).

In this procedure, panelists review all test items in a specially formatted test booklet (an 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 the Threshold PLDs developed by the NJSAC in conjunction with NJDOE, and over the course of three rounds, panelists place a bookmark at the point in the test booklet at which 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 median of the panelists' bookmarked pages becomes the group bookmark, and the associated raw score becomes the cut score for that level, grade, and round.

For the NJSLA–S Standard Setting, the standard-setting plan called for three rounds of item review in OIBs. Between rounds, panelists were encouraged to discuss the results of the previous round, with facilitation by MI psychometricians. After Round 1, panelists were able to view all bookmark placements for their grade, plus the median bookmark pages and associated cut scores. After Round 2, panelists were able to see all this information plus impact data—that is, the percentages of students who would be classified at each performance level based on Round 2 cut scores. After Round 3, panelists were able to see updated impact data based on Round 3 cut scores but were not allowed to modify their bookmark placements at that time. However, they were able to indicate on the Evaluation Form whether or not they believed Round 3 cut scores to be reasonable. A summary of the results of the NJSLA–S Standard Setting are presented in Part 5.3.

5.2.3 Conducting the Standard Setting Meeting

The 2019 NJSLA–S Standard Setting was conducted at the Princeton Marriott from July 23–25, 2019. The Marriott has also been the site of several other NJSLA–S item review meetings, NJTAC meetings, and other meetings related to the NJSLA–S; consequently, MI and NJDOE staff were quite familiar with the site and considered it highly suited to their purposes. The site contains multiple meeting rooms, AV support, and onsite catering. Furthermore, it is centrally located and has easy access to the New Jersey Turnpike and other major highways, enabling relatively easy access for panelists.

5.2.3.1 Recruiting standard setting panelists. The approved plan called for 12–15 members for each grade level, with ideal composition of a 15-member panel as follows:

- 9 classroom teachers
- 2 administrators
- 2 teachers of students with disabilities (SWD)
- 2 teachers of English-Learners (EL)

Within the classroom teachers' group, MI recommended that at least one be from the next grade higher (i.e., grades 6 and 9 for grades 5 and 8). MI also recommended recruiting multiple SWD and EL teachers at each grade band. Other demographic considerations included balancing panelists by gender, ethnicity, location (i.e., urban vs rural vs suburban), teaching experience, and area of scientific education expertise. This diverse composition enhanced the procedural validity of the cut scores by expanding their generalizability.

NJDOE staff began soliciting potential panelist names and qualifications in April 2019. District superintendents submitted candidates, and NJDOE staff reviewed these submissions in May, contacting those selected in late May and notifying others in July. The gap between notification of those selected and those not selected was to allow time to replace any candidates who turned down the invitation.

The final composition of the three panels for the Standard Setting, July 23–25, was as follows:

- Grade 5, 14 panelists
- Grade 8, 12 panelists
- Grade 11, 13 panelists (one of whom had to leave prior to beginning Round 3)

Table 5.2.2 summarizes the demographic characteristics of these 39 panelists.

Table 5.2.2: Summary of Panelist Demographic Characteristics and Professional Qualifications

Variable	Demographic	Count
Gender	Male	14
Gender	Female	25
Location	Rural	3
Location	Suburban	20
Location	Urban	16
Ethnicity	African American/Black	4
Ethnicity	Arabic	1
Ethnicity	Asian	3
Ethnicity	Hispanic/Latino	5
Ethnicity	Middle-Eastern	1
Ethnicity	Other	3
Ethnicity	White	22
Years Teaching	0–4	1
Years Teaching	5–9	13
Years Teaching	10–14	12
Years Teaching	15+	13
Other Qualifications	Bilingual	19
Other Qualifications	ELL	27
Other Qualifications	General Ed	39
Other Qualifications	Special Ed	33
Other Qualifications	Supervision	7

5.2.3.2 Standard-setting software. The Online Performance Level Setting (OPLS) software was first used in the establishment of cut scores for the Smarter Balanced Assessment Consortium tests in 2014 and has been used to set cut scores on other tests in the intervening years. It can support multiple standard-setting activities and methods, including the bookmark procedure. For this application, the software was modified slightly to accommodate the longer and more complex NJSLA–S item formats.

5.2.3.3 Standard-setting hardware. For this activity, MI loaded all necessary software into Lenovo Chromebooks, shipped them to the meeting site, and set them up four to a table in the breakout rooms. Given the size of the files panelists would be accessing during three rounds of bookmark placements, MI arranged for additional high-speed wireless internet connections for the 45 Chromebooks. A member of MI's IT staff was also present onsite to maintain and troubleshoot the Chromebooks and wireless internet connections.

5.2.3.4 Standard-setting training materials. MI staff prepared PowerPoint presentations for the opening session and for bookmark training. NJDOE prepared a PowerPoint presentation to introduce the test development process to panelists. In addition to these training materials, MI prepared a navigation guide (embedded in the OPLS software) and a facilitator's script. A brief overview of each is provided here.

5.2.3.4.1 Overview and Bookmark PowerPoint. This presentation provides a general orientation to the goals for the three days, the tasks to be completed in a brief introduction to the PLDs and standard-setting activity, ground rules, and panelist assignments. This presentation also provided an opportunity for panelists to ask questions of a more general nature.

5.2.3.4.2 NJSLA–S Test development PowerPoint. This presentation was designed to orient panelists to the process by which the tests were developed, their history, test blueprints, item types, scoring procedures, and a brief introduction to all four types of PLDs.

5.2.3.4.3 OPLS navigation guide. This brief document provided a step-by-step guide to navigating through the OPLS software. It was embedded within the OPLS software as a resource.

5.2.3.3.4 Facilitator script. In advance of the activity, MI psychometricians put together a set of day-by-day, session-by-session talking points, guidelines, and caveats to use. Once the script was approved, all facilitators had a copy handy at all times. They reviewed it at the end of each day (in the daily debriefings) and referred to it frequently during standard setting.

5.2.3.4 Creating the Ordered Item Booklets. Ordered Item Booklets (OIBs) were created for each grade level and subsequently uploaded into OPLS. In order to prepare OIBs, it is first necessary to know the operational statistics of all operational test items. Once the NJSLA–S tests were administered, there was a short amount of time between completion and scoring of the last test and the standard-setting workshop. It was not feasible to include all student responses in the OIB calibration, thus NJDOE and MI agreed upon a representative dataset to calibrate the item difficulty parameters in order to build the OIBs. For this purpose, OIB item difficulty parameters were based on the following counts: 64,419 fifth graders, 88,295 eighth graders, and 76,001 eleventh graders from a representative sample across the state. Demographic characteristics of the students included in the scaling were compared to those of the population. In nearly all instances, the sample subgroups matched the population subgroups within half a percent, the exceptions being grade 5 Hispanic (off by 0.7%), grade 8 economically disadvantaged (off by 0.7%) and grade 11 males and females (off by 0.9%).

Prior to the items being calibrated for placement into the OIB, a standard battery of psychometric tests was conducted in the Preliminary Item Analysis (PIA). The goal was to identify any items that were performing unexpectedly. Item means, item-total correlations, and

DIF statistics were calculated for all items. Content specialists verified the keys of all MC and TE items. One grade 8 item was flagged for having a negative item-total correlation; MI psychometric staff recommended that the aberrant item be removed from the test map and that it not be used to assess student performance. NJDOE approved its removal.

After the PIA an item calibration was conducted in WINSTEPS (Linacre, 2012) using the Item Response Theory (IRT) Partial Credit Model (PCM; Masters, 1982). (The PCM is explained in detail in Part 6 of this Technical Report.) Items were re-tested according to the assumptions of the PCM, and no other misfitting items were found. The resulting item difficulty parameters were then transformed into RP67s by the procedures outlined in Cizek and Bunch (2007). RP67s represent the ability level required for a student to have a 2/3 chance of answering a given item correctly or of obtaining the given score point or higher on polytomous items. The item calibration processes, item difficulty parameters, and RP67s were all independently verified by two MI psychometricians. Once the RP67 were calculated those values were used to construct the OIBs, and they were submitted to NJDOE for approval. The approved items were then entered into OPLS via an item map.

5.3 Summary of Results

The following sections describe the results from all three rounds of the 2019 NJSLA–S Standard Setting.

5.3.1 Round 1

Results for Round 1 are presented in Table 5.3.1 and Figures 5.3.1 through 5.3.3. The dotted vertical lines in the figures represent the median page numbers, which correspond to the raw cuts. These figures are screenshots of the actual figures presented onscreen to panelists.

Table 5.3.1: Summary of Round 1 Bookmark Pages and Raw Score Cuts

Grade	Level	Median Page	Raw Score	Possible Points
Grade 5	Level 2	11	24	60
Grade 5	Level 3	30	35	60
Grade 5	Level 4	49.5	48	60
Grade 8	Level 2	7.5	27	70
Grade 8	Level 3	30	41	70
Grade 8	Level 4	53	52	70
Grade 11	Level 2	9	31	78
Grade 11	Level 3	36	47	78
Grade 11	Level 4	64	61	78

As is frequently the case, and as is depicted in the figures that follow, bookmarks in Round 1 were widely dispersed, and there were two instances of overlap at grade 11. Note in Figure 5.3.3, the highest Level 2 bookmark was set on the same page as the lowest Level 3 bookmark (page 23). One panelist placed a Level 4 bookmark at page 44, six pages lower than the highest Level 3 bookmark.

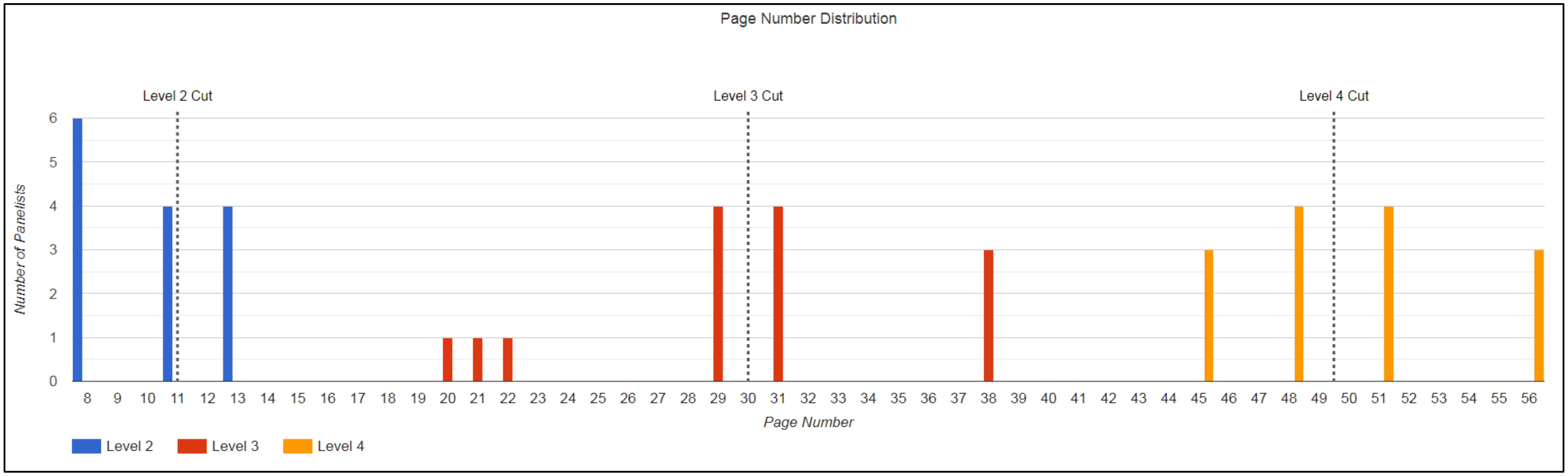


Figure 5.3.1. Distribution of bookmarks after Round 1: Grade 5

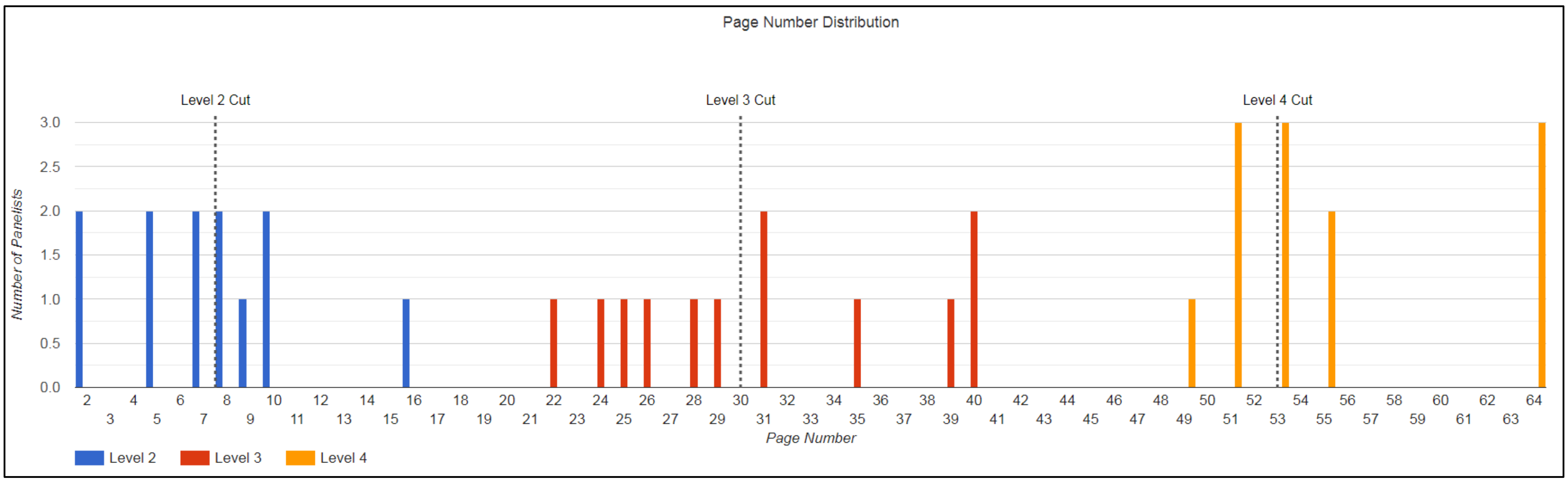


Figure 5.3.2. Distribution of bookmarks after Round 1: Grade 8

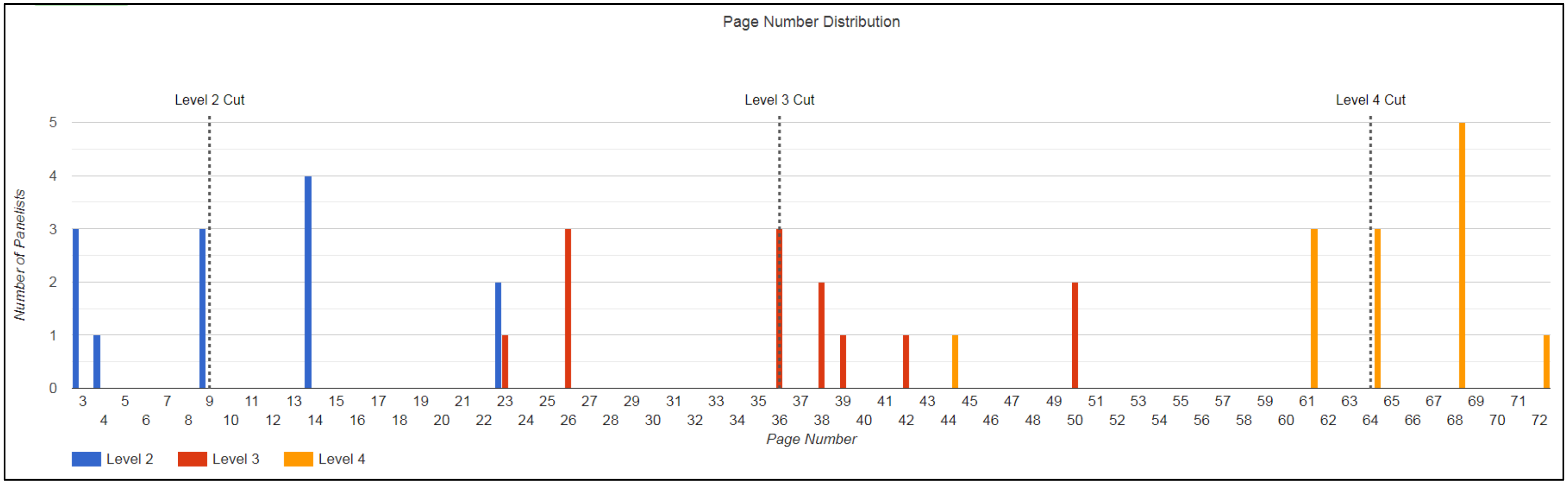


Figure 5.3.3. Distribution of bookmarks after Round 1: Grade 11

5.3.2 Round 2

Results for Round 2 are presented in Table 5.3.2 and Figures 5.3.4 through 5.3.6. These are the data shared with panelists at the beginning of Round 3. As in Round 1, the dotted vertical lines in the figures represent the median bookmark placements**,** and the figures are screenshots of the actual figures presented onscreen to panelists.

Table 5.3.2: Summary of Round 2 Bookmark Pages and Raw Score Cuts

Grade	Level	Median Page	Raw Score	Possible Points
Grade5	Level 2	11.5	24	60
Grade 5	Level 3	34	39	60
Grade 5	Level 4	51	49	60
Grade 8	Level 2	7.5	27	70
Grade 8	Level 3	27	40	70
Grade 8	Level 4	55	52	70
Grade 11	Level 2	14	35	78
Grade 11	Level 3	36	47	78
Grade 11	Level 4	62	60	78

Compared to Round 1, there was a considerable narrowing of the distributions of bookmarks for all levels for all three grades and convergence toward the medians. There were no overlaps in Round 2.

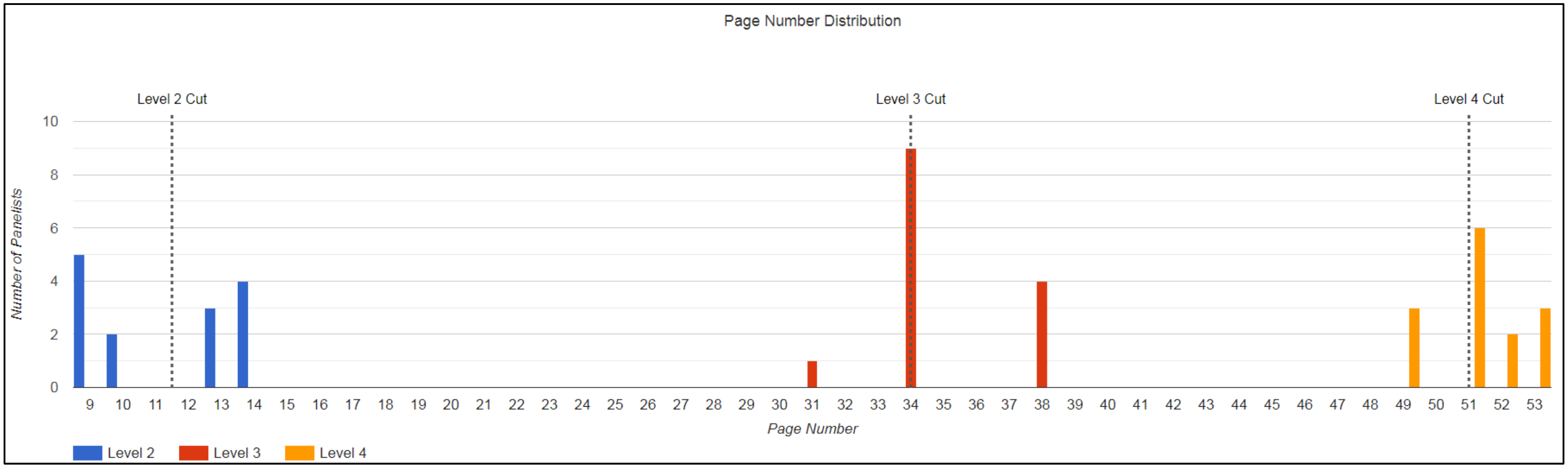


Figure 5.3.4. Distribution of bookmarks after Round 2: Grade 5

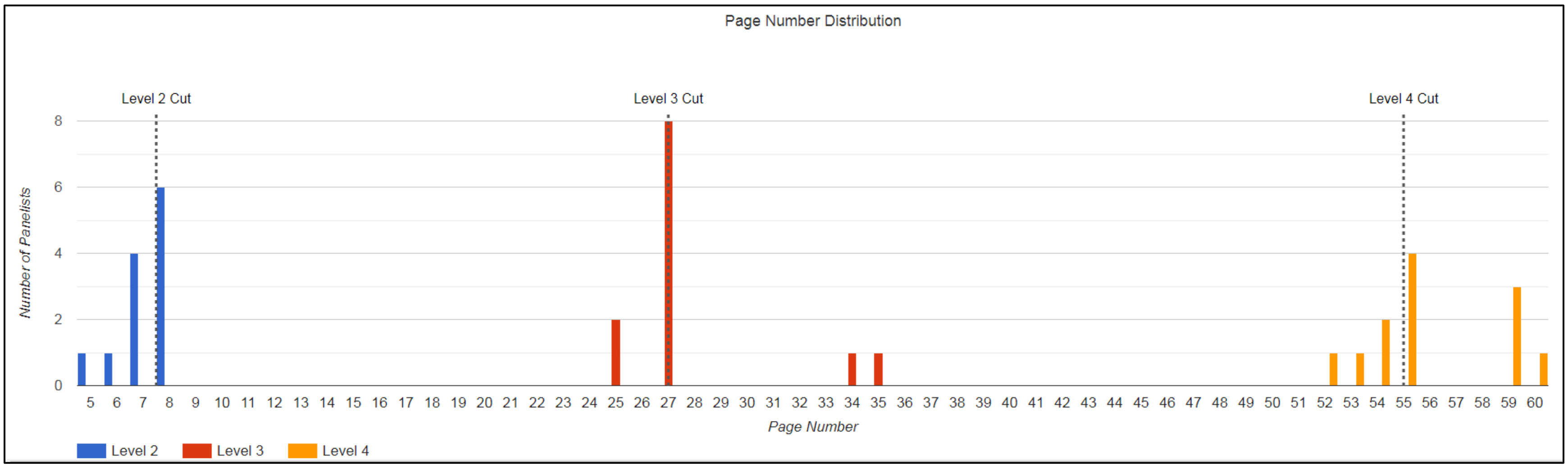


Figure 5.3.5. Distribution of bookmarks after Round 2: Grade 8

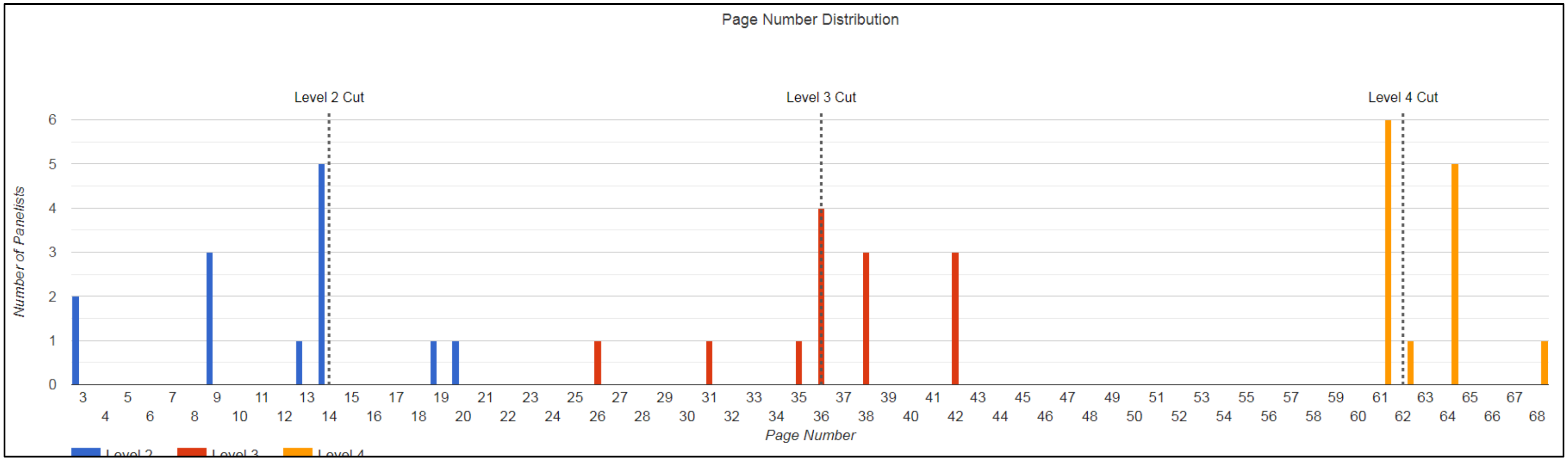


Figure 5.3.6. Distribution of bookmarks after Round 2: Grade 11

Impact data. Impact data for the Round 2 cut scores are shown in Table 5.3.3 and Figures 5.3.7 through 5.3.9. These data (in both tabular and graphical form) were shared with panelists after the discussion of Round 2 bookmark placements.

Table 5.3.3: Impact Data for Round 2 – Percent in each Level

Grade	Level 1	Level 2	Level 3	Level 4	Level 3 or Above
5	35.1	38.6	21.2	5.1	26.3
8	56.5	25.9	13.9	3.7	17.6
11	56.5	20.1	16.4	7.0	23.4

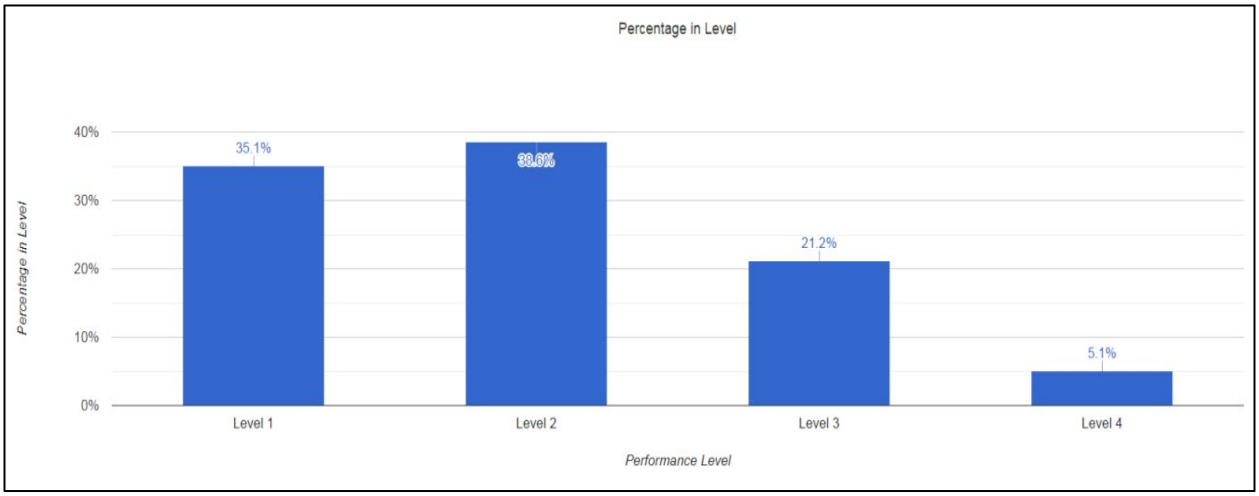


Figure 5.3.7. Percentages of students classified at each level after Round 2: Grade 5

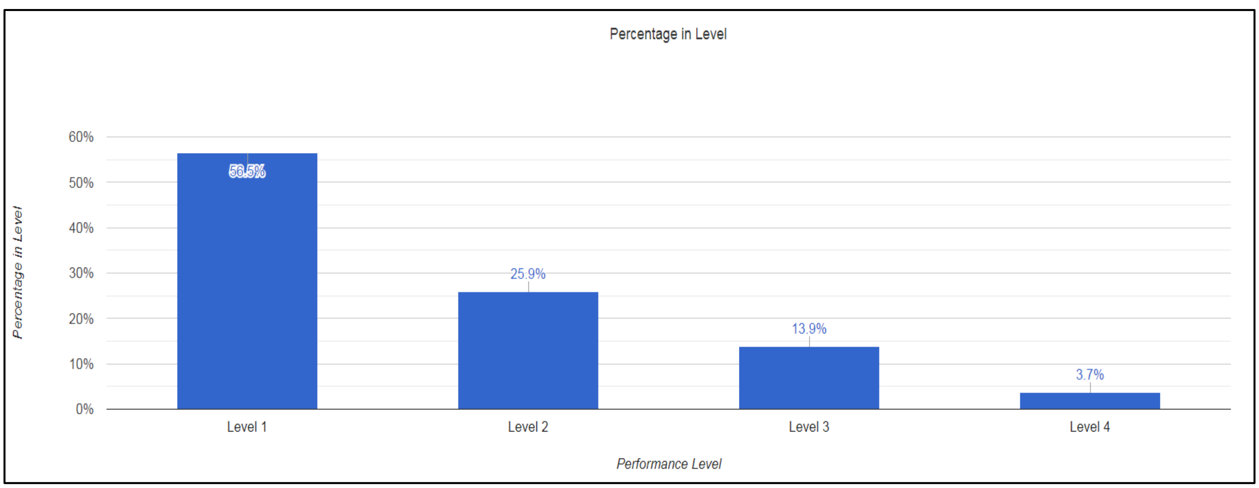


Figure 5.3.8. Percentages of students classified at each level after Round 2: Grade 8

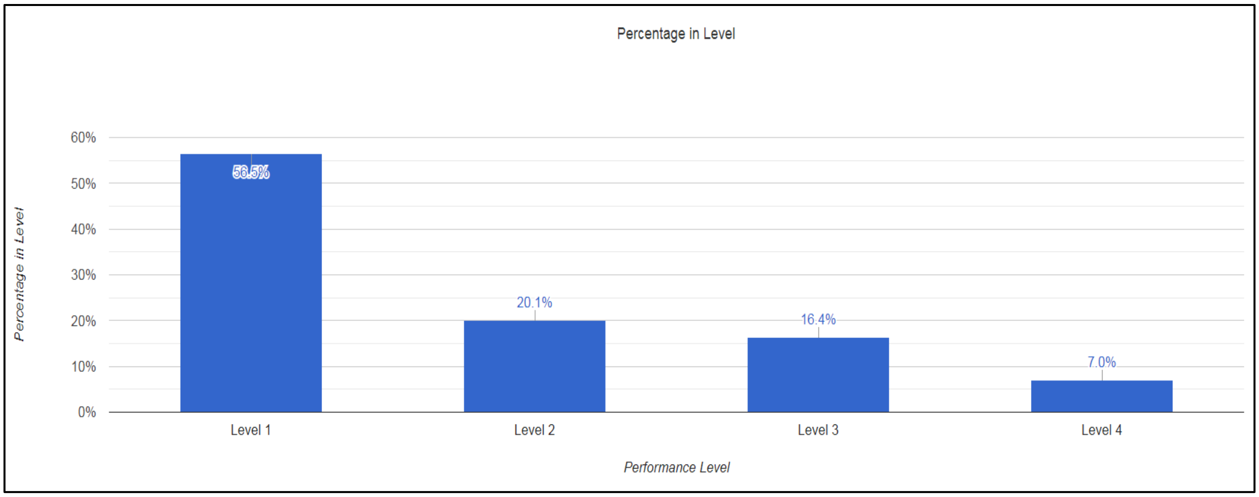


Figure 5.3.9. Percentages of students classified at each level after Round 2: Grade 11

5.3.3 Round 3

Results for Round 3 are presented in Table 5.3.4 and Figures 5.3.10 through 5.3.12. Panelists were able to study these results prior to completing the final evaluation, which contained statements about the reasonableness of the cut scores.

Table 5.3.4: Summary of Round 3 Bookmark Pages and Raw Score Cuts

Grade	Level	Median Page	Raw Score	% At or Above	Possible Points
Grade 5	Level 2	11.5	24	62.5	60
Grade 5	Level 3	34	39	26.3	60
Grade 5	Level 4	51	49	5.1	60
Grade 8	Level 2	7.5	27	61.0	70
Grade 8	Level 3	27	40	17.6	70
Grade 8	Level 4	55	52	3.7	70
Grade 11	Level 2	14	35	50.7	78
Grade 11	Level 3	36	47	26.5	78
Grade 11	Level 4	62	60	7.0	78

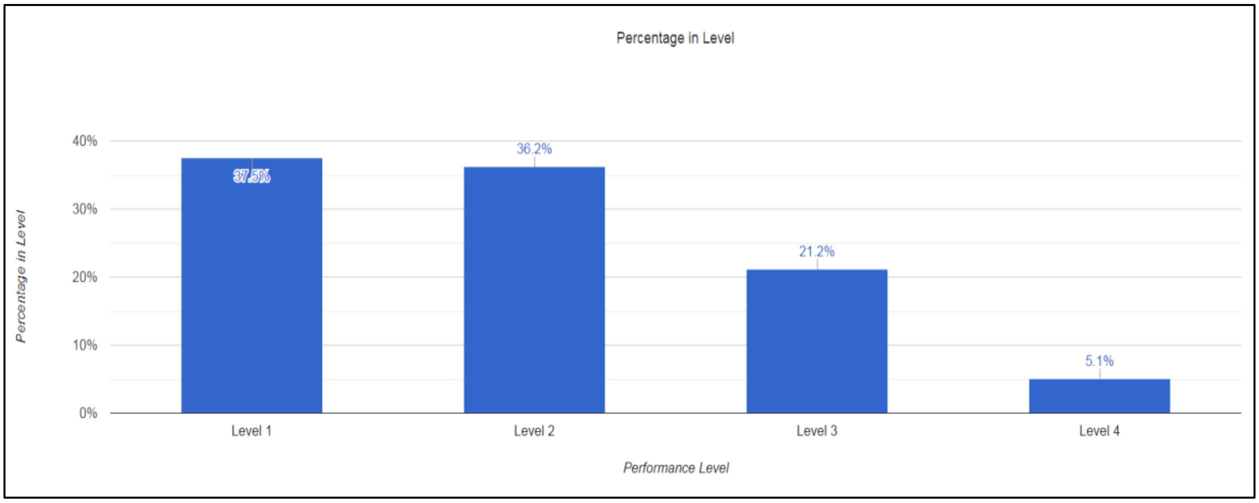


Figure 5.3.10. Percentages of students classified at each level after Round 3: Grade 5

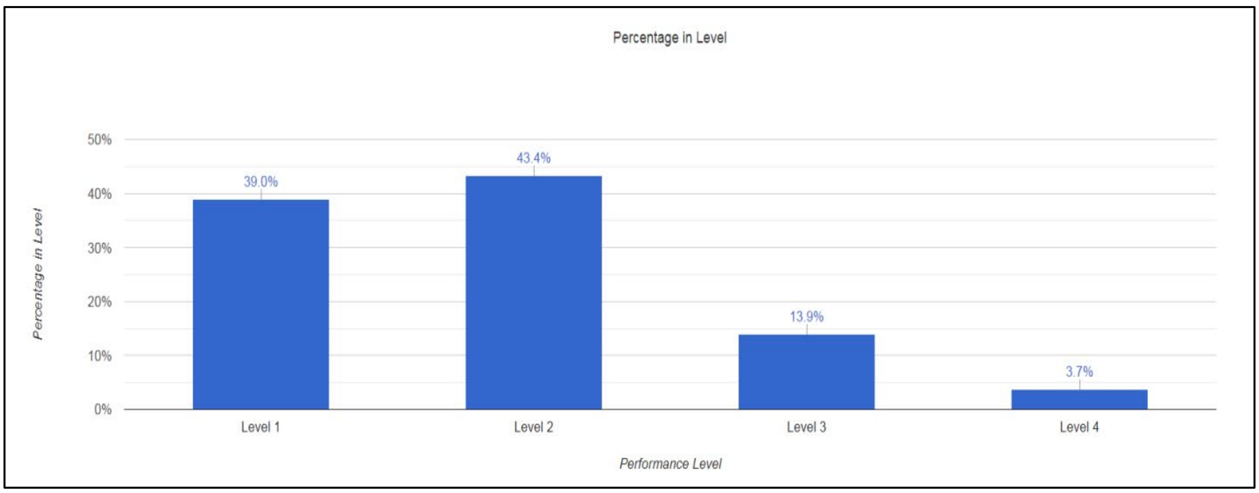


Figure 5.3.11. Percentages of students classified at each level after Round 3: Grade 8

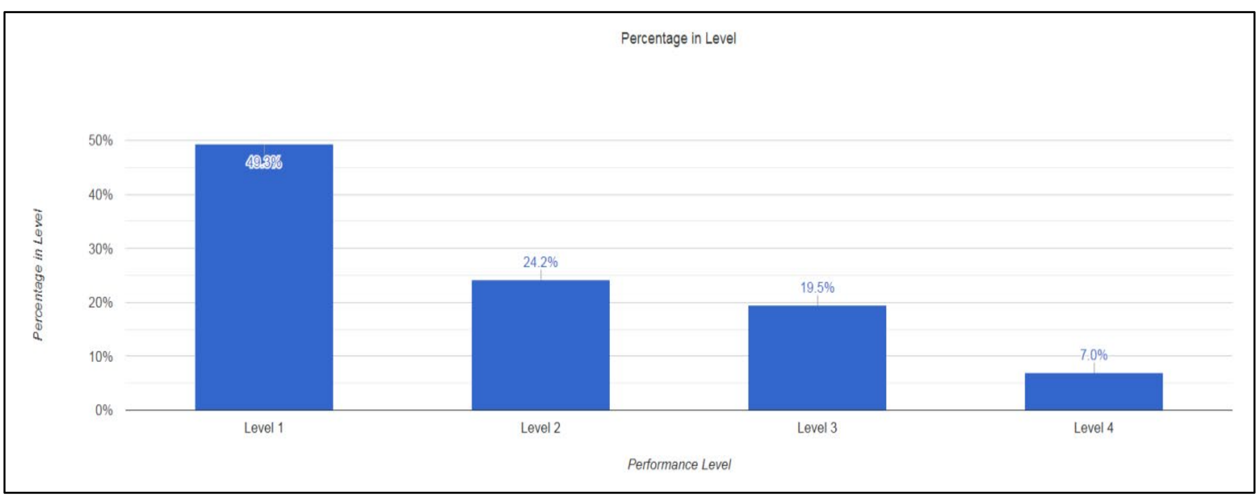


Figure 5.3.12. Percentages of students classified at each level after Round 3: Grade 11

Tables 5.3.5 through 5.3.7 show the percentages of students at each grade level classified at each performance level by gender, ethnicity, EL status, economically disadvantaged (EconDis) status, and students with disabilities (SWD) status. The n-counts of students are reflective of the standard setting sample and are not based on the population of New Jersey students as a whole.

Table 5.3.5: Percentages of Grade 5 Students Classified at Each Performance Level, By Group

Group	Count	Level 1	Level 2	Level 3	Level 4	Level 3 or Higher
All Students	64,419	37.54	36.14	21.19	5.13	26.32
Male	32,839	37.78	34.47	22.01	5.74	27.75
Female	31,580	37.29	37.88	20.34	4.49	24.83
Am. Indian	85	32.94	37.65	25.88	3.53	29.41
Asian	6,887	13.04	31.80	38.73	16.44	55.16
Black	9,532	61.46	29.23	8.39	0.92	9.32
Hispanic	19,447	53.20	34.27	10.98	1.55	12.53
Pacific Islander	109	34.86	39.45	20.18	5.50	25.69
White	26,809	24.42	40.98	28.39	6.21	34.60
EL – No	60,353	34.52	37.54	22.48	5.46	27.94
EL – Yes	4,060	82.41	15.34	2.02	0.22	2.24
EconDis – No	39,528	25.11	38.56	28.66	7.67	36.33
EconDis – Yes	24,882	57.27	32.31	9.33	1.09	10.42
SWD – No	51,523	31.18	38.77	24.10	5.95	30.05
SWD – Yes	12,896	62.95	25.66	9.56	1.83	11.39

Table 5.3.6: Percentages of Grade 8 Students Classified at Each Performance Level, By Group

Group	Count	Level 1	Level 2	Level 3	Level 4	Level 3 or Higher
All Students	88,295	39.05	43.39	13.90	3.67	17.57
Male	45,285	40.87	40.58	14.08	4.47	18.55
Female	43,010	37.13	46.34	13.72	2.82	16.53
Am. Indian	106	47.17	40.57	10.38	1.89	12.26
Asian	8,988	13.44	42.70	30.53	13.33	43.86
Black	12,943	63.66	31.89	3.99	0.46	4.45
Hispanic	25,384	56.37	37.37	5.56	0.70	6.26
Pacific Islander	190	18.95	55.79	20.53	4.74	25.26
White	39,072	25.94	51.23	18.52	4.32	22.83
EL – No	84,406	36.98	44.68	14.50	3.83	18.34
EL – Yes	3,886	83.87	15.23	0.85	0.05	0.90
EconDis – No	56,797	27.23	48.18	19.15	5.44	24.59
EconDis – Yes	31,492	60.35	34.74	4.43	0.47	4.91
SWD – No	70,911	32.73	46.99	16.01	4.27	20.28
SWD – Yes	17,384	64.83	28.68	5.29	1.21	6.50

Table 5.3.7: Percentages of Grade 11 Students Classified at Each Performance Level, By Group

Group	Count	Level 1	Level 2	Level 3	Level 4	Level 3 or Higher
All Students	76,001	49.33	24.17	19.48	7.02	26.49
Male	37,941	50.42	22.29	19.33	7.97	27.30
Female	38,060	48.26	26.05	19.63	6.07	25.70
Am. Indian	94	48.94	24.47	21.28	5.32	26.60
Asian	7,718	22.60	23.02	32.31	22.07	54.38
Black	10,995	72.51	18.28	7.76	1.45	9.20
Hispanic	20,079	65.10	21.69	11.14	2.07	13.20
Pacific Islander	178	30.34	34.27	23.03	12.36	35.39
White	36,000	39.40	27.58	24.87	8.16	33.03
EL – No	72,743	47.40	24.98	20.30	7.33	27.62
EL – Yes	3,245	92.60	6.13	1.20	0.06	1.26
EconDis – No	52,532	41.02	25.94	23.79	9.25	33.04
EconDis – Yes	23,456	67.94	20.23	9.83	2.01	11.84
SWD – No	62,736	44.35	25.96	21.74	7.95	29.69
SWD – Yes	13,265	72.92	15.70	8.78	2.59	11.37

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.

5.3.4 Evaluation survey

In accordance with the NJSLA–S Standard-Setting plan and standard-setting best practices, the standard-setting panelists were required to respond to a short evaluation survey at the end of Round 3. The results of the survey were overwhelmingly positive about the process as well as the outcomes. Of critical importance were the responses to the three reasonableness statements. There were no objections to the Level 2 or Level 3 cut scores at any grade. One grade 5 panelist questioned the Level 4 cut score and indicated that it should be raised by a single point. However, that panelist had checked "Agree" for this statement and seemed to be reflecting on comments made by other panelists during the discussion prior to Round 3. One grade 8 panelist entered a "?" for the reasonableness statement for the Level 2 cut score but did not recommend raising or lowering that cut. Overall, 113 out of 114 responses to the reasonable statements were "Agree" (20%) or "Strongly Agree" (79%). Table 5.3.8 displays the results of the evaluation surveys.

Table 5.3.8: Summary of Evaluation Surveys

Statement	Strongly Disagree	Disagree	"?"	Agree	Strongly Agree	N
Taking and scoring a section of the test was helpful.	0%	0%	0%	8%	92%	38
The discussion of performance level descriptors was useful.	0%	0%	3%	11%	87%	38
The bookmark presentation was clear and helpful.	0%	0%	3%	39%	58%	38
The bookmark practice round was helpful.	0%	3%	3%	18%	76%	38
I was able to navigate the standard setting software successfully.	0%	0%	0%	5%	95%	38
My facilitator was able to answer my questions.	0%	0%	0%	11%	89%	38
The discussion after Round 1 was helpful.	0%	0%	0%	16%	84%	38
The discussion after Round 2 was helpful	0%	0%	3%	8%	89%	38
The process was fair.	0%	0%	3%	16%	82%	38
The process was orderly.	0%	0%	0%	18%	82%	38
My group's final cut score for Level 2 is reasonable.	0%	0%	3%	24%	74%	38
[If you disagree, should it have been higher or lower? Circle one.]						0
My group's final cut score for Level 3 is reasonable.	0%	0%	0%	18%	82%	38
[If you disagree, should it have been higher or lower? Circle one.]						0
My group's final cut score for Level 4 is reasonable.	0%	0%	0%	18%	82%	38
[If you disagree, should it have been higher or lower? Circle one.]				+1		1
The meeting site was a good place to conduct this activity.	0%	3%	0%	11%	87%	38
Food service was good.	0%	0%	0%	16%	84%	38
My personal needs (travel, lodging, accommodations, dietary restrictions) were met.	0%	5%	0%	13%	82%	38

5.4 External Review

The 2019 NJSLA–S Standard Setting was externally reviewed by NJTAC member Stephen Koffler. 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 performance levels, as has been described throughout this Technical Report. The interpretations of the performance level classifications are dependent upon the test performing as intended. As was described in Part 2.3, the NJSLA–S was constructed using a combination of Classical Test Theory (CTT) and Item Response Theory (IRT) statistics, along with the myriad content constraints. The following sections detail how well the 2019 NJSLA–S performed based on those CTT and IRT statistics, along with other criteria. Detailed test maps containing item metadata, various statistics, and Depth-of-Knowledge (DOK; Webb, 1997) and Range PLD alignment are presented in Appendix F.

The data for these and all subsequent analyses were verified by Pearson's Customer Data Quality (CDQ) team. Responses from students who did not attempt enough items or who had their test scores voided were removed from the data set prior to analysis. NJDOE set the threshold for attemptedness as legitimate student responses to 20% of the items. 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 prior to item calibration and scaling. The statistics can be grouped into measures of four simple 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 an item performed in an unintended manner and could negatively impact the reliability or the validity of test score interpretations, it was recommended to NJDOE that the item be removed from the assessment. Descriptions of each type of 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 tending to be too challenging or too easy for a population of test takers, then the reliability and the validity of test score interpretations will suffer. In CTT, dichotomous item difficulty 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 to 1.00; an item with a high p-value is easier to answer correctly, whereas one 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 Part 4.1.1. For polytomous items, such as the 0–2pt TE and 0–4pt CR items, item difficulty is expressed as an item mean. The polytomous 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 a p-value), then flagging the item if its converted p-value falls outside of the .25 to .90 range. It should be noted that the flagging criteria is a recommendation, and many productive items have p-values outside of the .25 to .90 range.

Item discrimination is also extremely important to monitor, because if items are not discriminating between students will high levels of ability in comparison to students with low levels of ability, then both reliability and the validity of test score interpretations can suffer. CTT 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 can range from -1.00 to 1.00. Dichotomous items with values below .2 are flagged for review during the adjudication process. Polytomous items are expected to have higher item-total correlations; as such, the 0–2pt TE items and 0–4pt CR items are flagged with correlations below .25 and .30, respectively. Items with item-total correlations that are negative are considered for removal from the test, because they could be harming both reliability and the validity of test score interpretations. In 2019, one grade 8 item was removed for having a negative item-total correlation.

Two types of tables are presented below. Tables 6.1.1 through 6.1.9 summarize by item type the average item difficulty and discrimination of the 2019 NJSLA–S items. The averages within each of these tables are disaggregated by content domain and scientific practice. Tables 6.1.10 through 6.1.21 summarize frequency distributions for MC and TE item difficulty and discrimination, also disaggregated by content domain and scientific practice.

The average item difficulties and discriminations appear to be productive for measuring students in New Jersey. At grade 5 the average TE and CR item tended to be slightly more challenging, and more discriminating, than the MC items. At grade 8 the pattern for average item difficulties differed from grade 5; the average TE item at grade 8 had an item-total correlation of .37, exactly the same as the average MC item. The grade 8 CR items were highly discriminating, with average item-total correlations of .63. At Grade 11 item difficulties and discriminations displayed the same pattern as at grade 5, with the CR and TE items being, on average, more challenging and more discriminating than the MC items.

The frequency distributions of item total correlations also appear to be productive for discriminating between high and low achieving students. Only two items at both grades 8 and 11 had correlations below .20. Grade 5 had zero items below .20. The p-value distributions, however, were less positive. At grade 11 there were zero items that had p-values above .75, and only three TE items had p-values below .25, indicating almost no items at the easier and harder ends of the scale. Grade 8 items were more heavily concentrated toward the low end, indicating higher difficulty. Nine of 38 TE items had p-values below .25, and another 20 were below .50. Most of the Grade 5 items fell between .25 and .75; however, seven of the 47 MC and TE items had p-values above .90, indicating that some of the items might have been too easy.

Table 6.1.1: Grade 5 Item Difficulty and Discrimination Summary Statistics by Cluster, MC

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	13	.60	.17	.40
Earth and Space	7	.60	.19	.39
Life	3	.52	.17	.39
Physical	3	.68	.12	.45
Critiquing	4	.52	.14	.40
Investigating	2	.68	.16	.32
Sensemaking	7	.62	.19	.43

Table 6.1.2: Grade 5 Item Difficulty and Discrimination Summary Statistics by Cluster, TE

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	34	.52	.18	.43
Earth and Space	9	.56	.20	.45
Life	15	.52	.18	.39
Physical	10	.49	.16	.48
Critiquing	10	.47	.19	.45
Investigating	10	.54	.11	.45
Sensemaking	14	.55	.20	.41

Table 6.1.3: Grade 5 Item Difficulty and Discrimination Summary Statistics by Cluster, CR

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	3	1.53	.85	.60
Earth and Space	1	0.99	-	.56
Life	1	2.51	-	.60
Physical	1	1.10	-	.65
Critiquing	0	-	-	-
Investigating	1	1.10	-	.65
Sensemaking	2	1.75	1.07	.58

Table 6.1.4: Grade 8 Item Difficulty and Discrimination Summary Statistics by Cluster, MC

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	18	.45	.12	.37
Earth and Space	5	.53	.17	.46
Life	6	.44	.09	.35
Physical	7	.41	.09	.32
Critiquing	2	.45	.16	.40
Investigating	13	.42	.09	.35
Sensemaking	3	.59	.15	.45

Table 6.1.5: Grade 8 Item Difficulty and Discrimination Summary Statistics by Cluster, TE

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	38	.40	.18	.37
Earth and Space	11	.41	.15	.35
Life	13	.40	.21	.42
Physical	14	.38	.18	.36
Critiquing	8	.32	.12	.39
Investigating	6	.31	.10	.35
Sensemaking	24	.45	.19	.37

Table 6.1.6: Grade 8 Item Difficulty and Discrimination Summary Statistics by Cluster, CR

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	3	1.23	.30	.63
Earth and Space	1	1.21	-	.61
Life	1	1.54	-	.67
Physical	1	0.95	-	.60
Critiquing	1	1.54	-	.67
Investigating	1	0.95	-	.60
Sensemaking	1	1.21	-	.61

Table 6.1.7: Grade 11 Item Difficulty and Discrimination Summary Statistics by Cluster, MC

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	25	.48	.14	.38
Earth and Space	8	.53	.15	.44
Life	7	.50	.15	.36
Physical	10	.44	.11	.35
Critiquing	3	.40	.11	.41
Investigating	8	.50	.12	.36
Sensemaking	14	.50	.15	.39

Table 6.1.8: Grade 11 Item Difficulty and Discrimination Summary Statistics by Cluster, TE

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	40	.44	.16	.47
Earth and Space	15	.37	.11	.46
Life	13	.50	.16	.46
Physical	12	.47	.16	.49
Critiquing	8	.44	.16	.45
Investigating	13	.41	.19	.50
Sensemaking	19	.47	.13	.46

Table 6.1.9: Grade 11 Item Difficulty and Discrimination Summary Statistics by Cluster, CR

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSLA–S	3	1.02	.23	.64
Earth and Space	1	0.76	-	.57
Life	1	1.17	-	.68
Physical	1	1.14	-	.68
Critiquing	2	1.16	.02	.68
Investigating	0	-	-	-
Sensemaking	1	0.76	-	.57

Table 6.1.10: Grade 5 Difficulty Indices by Cluster, MC

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSLA–S	13	.60	0	4	5	0	4
Earth and Space	7	.60	0	2	2	0	3
Life	3	.43	0	2	1	0	0
Physical	3	.70	0	0	2	0	1
Critiquing	4	.48	0	2	2	0	0
Investigating	2	.68	0	0	1	0	1
Sensemaking	7	.70	0	2	2	0	3

Table 6.1.11: Grade 5 Difficulty Indices by Cluster, TE

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSLA–S	34	.50	2	13	15	1	3
Earth and Space	9	.61	0	4	3	1	1
Life	15	.50	1	5	8	0	1
Physical	10	.48	1	4	4	0	1
Critiquing	10	.47	1	4	5	0	0
Investigating	10	.53	0	4	5	0	1
Sensemaking	14	.55	1	5	5	1	2

Table 6.1.12: Grade 5 Discrimination Indices by Cluster, MC

Domain/Practice	# Items	Median	rpb<.20	.20<=rpb<.30	.30<=rpb<.40	.40<=rpb<.50	rpb>=.50
NJSLA–S	13	.37	0	1	6	4	2
Earth and Space	7	.32	0	1	3	2	1
Life	3	.37	0	0	2	1	0
Physical	3	.48	0	0	1	1	1
Critiquing	4	.40	0	0	2	2	0
Investigating	2	.32	0	0	2	0	0
Sensemaking	7	.40	0	1	2	2	2

Table 6.1.13: Grade 5 Discrimination Indices by Cluster, TE

Domain/Practice	# Items	Median	rpb<.20	.20<=rpb<.30	.30<=rpb<.40	.40<=rpb<.50	rpb>=.50
NJSLA–S	34	.43	0	2	11	11	10
Earth and Space	9	.44	0	0	3	2	4
Life	15	.39	0	2	6	5	2
Physical	10	.47	0	0	2	4	4
Critiquing	10	.46	0	0	3	3	4
Investigating	10	.43	0	1	3	3	3
Sensemaking	14	.41	0	1	5	5	3

Table 6.1.14: Grade 8 Difficulty Indices by Cluster, MC

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSLA–S	18	.45	0	12	6	0	0
Earth and Space	5	.57	0	2	3	0	0
Life	6	.46	0	4	2	0	0
Physical	7	.38	0	6	1	0	0
Critiquing	2	.45	0	1	1	0	0
Investigating	13	.39	0	10	3	0	0
Sensemaking	3	.57	0	1	2	0	0

Table 6.1.15: Grade 8 Difficulty Indices by Cluster, TE

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSLA–S	38	.38	9	20	7	1	1
Earth and Space	11	.46	2	5	4	0	0
Life	13	.37	4	7	1	0	1
Physical	14	.38	3	8	2	1	0
Critiquing	8	.30	3	4	1	0	0
Investigating	6	.36	2	4	0	0	0
Sensemaking	24	.44	4	12	6	1	1

Table 6.1.16: Grade 8 Discrimination Indices by Cluster, MC

Domain/Practice	# Items	Median	rpb<.20	.20<=rpb<.30	.30<=rpb<.40	.40<=rpb<.50	rpb>=.50
NJSLA–S	18	.40	2	3	4	6	3
Earth and Space	5	.47	0	0	1	3	1
Life	6	.37	1	1	1	2	1
Physical	7	.30	1	2	2	1	1
Critiquing	2	.40	0	1	0	0	1
Investigating	13	.39	2	2	3	5	1
Sensemaking	3	.49	0	0	1	1	1

Table 6.1.17: Grade 8 Discrimination Indices by Cluster, TE

Domain/Practice	# Items	Median	rpb<.20	.20<=rpb<.30	.30<=rpb<.40	.40<=rpb<.50	rpb>=.50
NJSLA–S	38	.38	3	6	10	14	5
Earth and Space	11	.30	1	4	1	3	2
Life	13	.44	0	1	3	7	2
Physical	14	.34	2	1	6	4	1
Critiquing	8	.43	0	0	3	5	0
Investigating	6	.35	2	1	0	2	1
Sensemaking	24	.35	1	5	7	7	4

Table 6.1.18: Grade 11 Difficulty Indices by Cluster, MC

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSLA–S	25	.45	0	15	10	0	0
Earth and Space	8	.58	0	3	5	0	0
Life	7	.47	0	4	3	0	0
Physical	10	.43	0	8	2	0	0
Critiquing	3	.45	0	3	0	0	0
Investigating	8	.45	0	5	3	0	0
Sensemaking	14	.50	0	7	7	0	0

Table 6.1.19: Grade 11 Difficulty Indices by Cluster, TE

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSLA–S	40	.47	3	19	18	0	0
Earth and Space	15	.32	1	11	3	0	0
Life	13	.56	1	4	8	0	0
Physical	12	.52	1	4	7	0	0
Critiquing	8	.39	0	5	3	0	0
Investigating	13	.41	3	5	5	0	0
Sensemaking	19	.52	0	9	10	0	0

Table 6.1.20: Grade 11 Discrimination Indices by Cluster, MC

Domain/Practice	# Items	Median	rpb<.20	.20<=rpb<.30	.30<=rpb<.40	.40<=rpb<.50	rpb>=.50
NJSLA–S	25	.40	2	4	5	10	4
Earth and Space	8	.44	0	0	2	3	3
Life	7	.40	1	1	0	5	0
Physical	10	.35	1	3	3	2	1
Critiquing	3	.40	0	0	1	2	0
Investigating	8	.38	1	2	1	3	1
Sensemaking	14	.40	1	2	3	5	3

Table 6.1.21: Grade 11 Discrimination Indices by Cluster, TE

Domain/Practice	# Items	Median	rpb<.20	.20<=rpb<.30	.30<=rpb<.40	.40<=rpb<.50	rpb>=.50
NJSLA–S	40	.50	0	3	7	10	20
Earth and Space	15	.45	0	0	5	4	6
Life	13	.46	0	1	2	4	6
Physical	12	.53	0	2	0	2	8
Critiquing	8	.44	0	1	1	4	2
Investigating	13	.53	0	1	1	3	8
Sensemaking	19	.51	0	1	5	3	10

6.1.2 Speededness

The consequence of time limits on examinees' scores is called speededness. An examination is "speeded" to the degree that those taking the exam score lower than they would have had the test not been timed. A measure of the speededness of a test is the number of items that were not attempted by students. In each separately timed subsection of a test, if a student does not attempt the last item, it can be assumed that the student may have run out of time. 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 to not put effort into answering a constructed response item). Thus, if the percentage of omits is low, that implies that there is little speededness; if a percentage of omits is high, speededness, as well as other factors, may be the cause.

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 number of items and item types composing each operational unit of each test, along with the testing time, are detailed in Table 6.1.22. Table 6.1.23 presents the percentage of students omitting the last TE item in each test section.

Table 6.1.22: Testing Schedule—Items and Time Allocations

Grade	Unit	Items	Time in minutes
5	1	6 MC, 10 TE, 1 CR	45
5	2	4 MC, 12 TE, 1 CR	45
5	3	3 MC, 12 TE, 1 CR	45
8	1	7 MC, 12 TE, 1 CR	45
8	2	6 MC, 13 TE, 1 CR	45
8	3	5 MC, 14 TE, 1 CR	45
11	1	7 MC, 15 TE, 1 CR	60
11	2	10 MC, 12 TE, 1 CR	60
11	3	8 MC, 13 TE, 1 CR	60

Table 6.1.23: Percent of Students Omitting the Last TE Item in Each Unit

Grade	Unit	Location	%
5	1	16	1.0
5	2	17	1.4
5	3	15	0.7
8	1	18	1.9
8	2	19	1.9
8	3	19	1.6
11	1	23	2.4
11	2	22	2.4
11	3	21	5.1

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 are performing differently for sub-groups of students, the test might disadvantage some groups of students over others.

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 (Mantel & Haenszel, 1959) procedure in conjunction with the ETS classification system (Dorans & Holland, 1993). The Mantel-Haenszel (MH) method is a non-parametric approach to DIF. The ETS categorization is applied to flag the significance of DIF effects (Dorans & Holland, 1993). The letters A, B, and C are used to denote the ETS categorizations. A-level indicates a smaller degree of DIF, B-level indicates moderate DIF, and C-level indicates severe DIF and requires a careful review of the item for possible biases. For polytomous items DIF was identified using the Liu-Agresti procedure (Penfield, 2007). The Liu-Agresti cumulative common log-odds ratio allows for the ETS categorization to be applied to polytomous items.

NJSLA–S DIF detection for the field test 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-English learner (EL-No)/English learner (EL-Yes), students with disabilities (SWD-Yes)/ students without disabilities (SWD-No), Not economically disadvantaged (EconDis-No)/economically disadvantaged (EconDis-Yes), and due to the large numbers of students taking the TTS forms CBT/TTS. The traditional CBT test takers were the reference group, whereas the TTS test takers were the focal group.

The results of the operational DIF analysis were very positive with the exception of a very small number of items classified as 'C' for EL-No/EL-Yes. 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. Table 6.1.24 shows the DIF classifications for all eight comparison groups by grade. The small number of 'C' DIF items for EL-No/EL-Yes students are revisited in Part 9.7.3: Future Validity Studies.

Table 6.1.24: DIF Classification by Grade and Item Type

Grade	Group	Item Type	A	B	C
5	Male/Female	MC	13	0	0
5	Male/Female	TE	33	1	0
5	Male/Female	CR	3	0	0
5	Male/Female	Total	49	1	0
5	White/Black	MC	12	1	0
5	White/Black	TE	33	1	0
5	White/Black	CR	3	0	0
5	White/Black	Total	48	2	0
5	White/Hispanic	MC	12	1	0
5	White/Hispanic	TE	34	0	0
5	White/Hispanic	CR	3	0	0
5	White/Hispanic	Total	49	1	0
5	White/Asian	MC	13	0	0
5	White/Asian	TE	34	0	0
5	White/Asian	CR	3	0	0
5	White/Asian	Total	50	0	0
5	EL-No/EL-Yes	MC	12	1	0
5	EL-No/EL-Yes	TE	33	1	0
5	EL-No/EL-Yes	CR	3	0	0
5	EL-No/EL-Yes	Total	48	2	0
5	SWD-No/SWD-Yes	MC	13	0	0
5	SWD-No/SWD-Yes	TE	34	0	0
5	SWD-No/SWD-Yes	CR	3	0	0
5	SWD-No/SWD-Yes	Total	50	0	0
5	EconDis-No/EconDis-Yes	MC	13	0	0
5	EconDis-No/EconDis-Yes	TE	34	0	0
5	EconDis-No/EconDis-Yes	CR	3	0	0
5	EconDis-No/EconDis-Yes	Total	50	0	0
5	CBT/TTS	MC	13	0	0
5	CBT/TTS	TE	34	0	0
5	CBT/TTS	CR	3	0	0
5	CBT/TTS	Total	50	0	0

Grade	Group	Item Type	A	B	C
8	Male/Female	MC	18	0	0
8	Male/Female	TE	37	1	0
8	Male/Female	CR	3	0	0
8	Male/Female	Total	58	1	0
8	White/Black	MC	18	0	0
8	White/Black	TE	38	0	0
8	White/Black	CR	1	2	0
8	White/Black	Total	57	2	0
8	White/Hispanic	MC	18	0	0
8	White/Hispanic	TE	38	0	0
8	White/Hispanic	CR	3	0	0
8	White/Hispanic	Total	59	0	0
8	White/Asian	MC	18	0	0
8	White/Asian	TE	38	0	0
8	White/Asian	CR	3	0	0
8	White/Asian	Total	59	0	0
8	EL-No/EL-Yes	MC	18	0	0
8	EL-No/EL-Yes	TE	34	3	1
8	EL-No/EL-Yes	CR	2	0	1
8	EL-No/EL-Yes	Total	54	3	2
8	SWD-No/SWD-Yes	MC	18	0	0
8	SWD-No/SWD-Yes	TE	37	1	0
8	SWD-No/SWD-Yes	CR	3	0	0
8	SWD-No/SWD-Yes	Total	58	1	0
8	EconDis-No/EconDis-Yes	MC	18	0	0
8	EconDis-No/EconDis-Yes	TE	38	0	0
8	EconDis-No/EconDis-Yes	CR	3	0	0
8	EconDis-No/EconDis-Yes	Total	59	0	0
8	CBT/TTS	MC	18	0	0
8	CBT/TTS	TE	38	0	0
8	CBT/TTS	CR	3	0	0
8	CBT/TTS	Total	59	0	0

Grade	Group	Item Type	A	B	C
11	Male/Female	MC	24	1	0
11	Male/Female	TE	37	3	0
11	Male/Female	CR	2	1	0
11	Male/Female	Total	63	5	0
11	White/Black	MC	25	0	0
11	White/Black	TE	39	1	0
11	White/Black	CR	3	0	0
11	White/Black	Total	67	1	0
11	White/Hispanic	MC	25	0	0
11	White/Hispanic	TE	39	1	0
11	White/Hispanic	CR	3	0	0
11	White/Hispanic	Total	67	1	0
11	White/Asian	MC	24	1	0
11	White/Asian	TE	40	0	0
11	White/Asian	CR	3	0	0
11	White/Asian	Total	67	1	0
11	EL-No/EL-Yes	MC	25	0	0
11	EL-No/EL-Yes	TE	34	5	1
11	EL-No/EL-Yes	CR	3	0	0
11	EL-No/EL-Yes	Total	62	5	1
11	SWD-No/SWD-Yes	MC	25	0	0
11	SWD-No/SWD-Yes	TE	40	0	0
11	SWD-No/SWD-Yes	CR	3	0	0
11	SWD-No/SWD-Yes	Total	68	0	0
11	EconDis-No/EconDis-Yes	MC	25	0	0
11	EconDis-No/EconDis-Yes	TE	40	0	0
11	EconDis-No/EconDis-Yes	CR	3	0	0
11	EconDis-No/EconDis-Yes	Total	68	0	0
11	CBT/TTS	MC	25	0	0
11	CBT/TTS	TE	40	0	0
11	CBT/TTS	CR	3	0	0
11	CBT/TTS	Total	68	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. Part 6.2 of this report explains how IRT is used in the context of the NJSLA–S. The concept of IRT is explained, along with the reasoning as to why it improves upon classical test theory. Then, the specific IRT model used for the NJSLA–S is described in conjunction with the strong assumptions that the model must meet in order to be applicable. The remainder of Part 6.2 evaluates how well the assumptions of IRT are met.

IRT is conceptualized as a family of mathematical models that explain the relationship of student performance on test items to student latent ability level on the construct of interest (Hambleton & Swaminathan, 1985). Latent abilities (e.g., anxiety, intelligence, or mastery of the NJSLS–S) are not directly observable; student responses to items are directly observable. IRT models presume that the directly observable item responses of examinees can be explained by an unobservable latent trait. Within the context of the NJSLA–S the directly observable behaviors are the responses of students to the test items, and the latent trait that we are assuming those items estimate is student understanding of the New Jersey science curriculum: the NJSLS–S.

The logic behind making and meticulously checking these assumptions is that IRT addresses many of the limitations of classical test theory (CTT) and can improve both the construction and uses of tests (Hambleton & van der Linden, 1982); hence IRT can improve the validity of the inferences made from tests. The CTT item statistics that were presented in Part 6.1 are sample-dependent, which means that they are susceptible to substantial changes depending on the students who are answering the items. The sample dependency of CTT makes form-to-form or year-to-year inferences from test scores problematic, because the results take on a different meaning depending on the students who took the items or how hard the items were. The CTT test reliability statistics presented later in Part 8.1 are similarly susceptible to sample dependency and can increase or decrease depending on the sample's heterogeneity. Moreover, CTT reliability is also the same for all examinees, which means that the consistency of students' test performance is assumed to be the same regardless of their ability level.

IRT overcomes these shortcomings, among many others (Hambleton & Swaminathan, 1985). Its item difficulty parameters are independent of the students who took the test; its student ability estimates are independent of the test items. If IRT's assumptions are met, this allows students taking the NJSLA–S years from now, who are taking different items, to be placed onto the same scale as the students who are taking it today, allowing for more meaningful year-to-year and form-to-form comparisons than CTT can offer. Moreover, unlike CTT, the reliability of IRT student ability estimates is different across the student ability spectrum as conceptualized by the test information function (TIF; see Part 8.2 for a more detailed explanation). This allows for test construction to be targeted to specific places on the student ability spectrum where decisions are most important in order to maximize the test's ability to reliably classify examinees.

The increased power of IRT in comparison to CTT comes at a cost. IRT requires that certain strong assumptions be met. When the assumptions of IRT are not met the data and the resulting test scores will be questionable, harming any interpretations of test scores. Thus, it is imperative that assumptions be checked.

The NJSLA–S was constructed to meet the assumptions of a specific IRT model: the Rasch-based (1960) Partial Credit Model (PCM; Masters, 1982). The Rasch family of IRT models is a special case of other IRT models; Rasch models all 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). As was described earlier, the NJSLA–S item types are designed to minimize guessing, and the test contains polytomous items with varying score points (e.g., 0–2pt TE or 0–4pt CR items). If the PCM's assumptions are met, it is likely a good IRT model to use with the NJSLA–S.

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 from the others, and the resulting item parameter estimates are invariant regardless of who answered the items. Each of these five major IRT assumptions will be explained in greater detail in the sections below as they relate to the PCM. The PCM Item Characteristic Curves (ICCs) are also presented to show the relationships between student ability estimates and the probability of achieving a specific score point on the 0–4pt CR items. The final component within this section shows disaggregated descriptive statistics of the scale scores. Overall, the results of the 2019 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 a strong linear relationship 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. The Pearson product-moment correlations among the domains and practices are presented in Tables 6.2.1 through 6.2.3. High correlations would be evidence of a unidimensional test. Generally, more items in a cluster will lead to a higher correlation between that cluster and the total score. Furthermore, because each item is aligned to both a domain and a practice, the domain-to-domain and practice-to-practice intercorrelations will often be lower than the domain-to-practice and practice-to-domain intercorrelations.

At grade 5 all domains and practices correlated with the total test score at 0.91 or above. The lowest correlation among any clusters was 0.79. Relatively high correlations between the domains or practices and the total test score were also present at both grades 8 and 11. The intercorrelations among subscores indicate that the NJSLA–S is a unidimensional test.

Table 6.2.1: Grade 5 Correlation Matrix for Domains and Practices

Content	NJSLA–S	Earth	Life	Physical	Sensemaking	Critiquing	Investigating
NJSLA–S	1	-	-	-	-	-	-
Earth and Space	.92	1	-	-	-	-	-
Life	.93	.79	1	-	-	-	-
Physical	.92	.79	.79	1	-	-	-
Sensemaking	.96	.92	.89	.86	1	-	-
Critiquing	.91	.82	.88	.83	.81	1	-
Investigating	.93	.82	.84	.91	.83	.79	1

Table 6.2.2: Grade 8 Correlation Matrix for Domains and Practices

Content	NJSLA–S	Earth	Life	Physical	Sensemaking	Critiquing	Investigating
NJSLA–S	1	-	-	-	-	-	-
Earth and Space	.92	1	-	-	-	-	-
Life	.93	.79	1	-	-	-	-
Physical	.93	.78	.79	1	-	-	-
Sensemaking	.94	.91	.89	.85	1	-	-
Critiquing	.90	.79	.86	.83	.79	1	-
Investigating	.92	.82	.84	.90	.79	.77	1

Table 6.2.3: Grade 11 Correlation Matrix for Domains and Practices

Content	NJSLA–S	Earth	Life	Physical	Sensemaking	Critiquing	Investigating
NJSLA–S	1	-	-	-	-	-	-
Earth and Space	.95	1	-	-	-	-	-
Life	.95	.85	1	-	-	-	-
Physical	.95	.86	.85	1	-	-	-
Sensemaking	.97	.95	.91	.91	1	-	-
Critiquing	.93	.88	.91	.88	.86	1	-
Investigating	.94	.87	.89	.93	.87	.84	1

6.2.1.2 Principal Component Analysis. Principal Components Analysis (PCA) is a data reduction technique that attempts to account for the variance in measures (Brown, 2006) by converting them into uncorrelated principal components. 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). The resulting principal components can then be plotted and interpreted in a scree plot.

The results of each grade's PCA provide further evidence of the unidimensionality of the NJSLA–S. The scree plots were interpreted by finding the place on the plot where the slope levelled off. Gorsuch (1983) noted that this method of interpretation works well when sample sizes are large, and the factors are well-defined. The principal components to the left of the point on the plot where the slope leveled were deemed practically significant. Each grade's scree plot shows that only one major dimension is practically contributing to the variability in student responses to items. The second most prominent eigenvalue for each grade level is below 2, whereas the most prominent eigenvalues range from approximately 11–16.

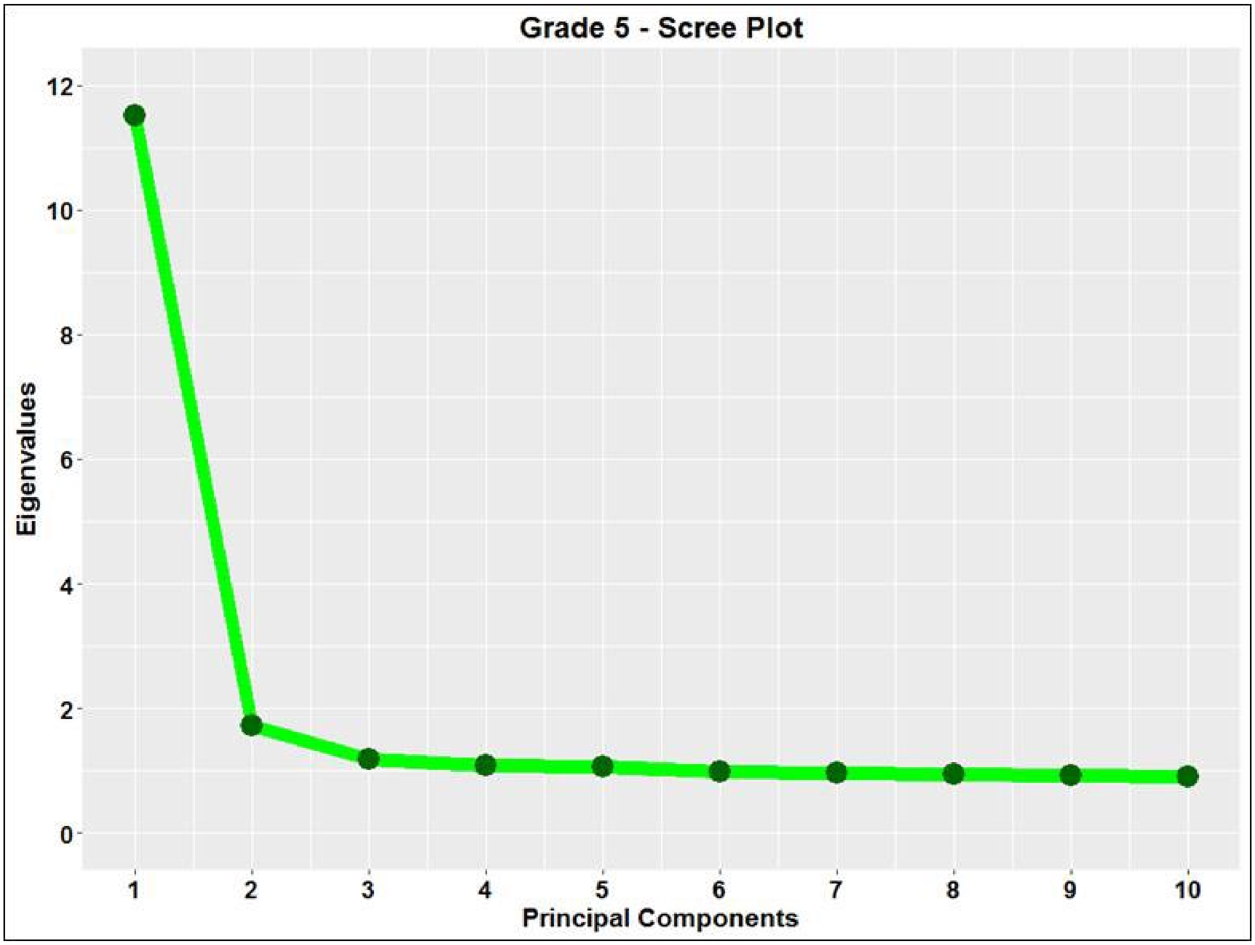


Figure 6.2.1. Grade 5 Scree Plot

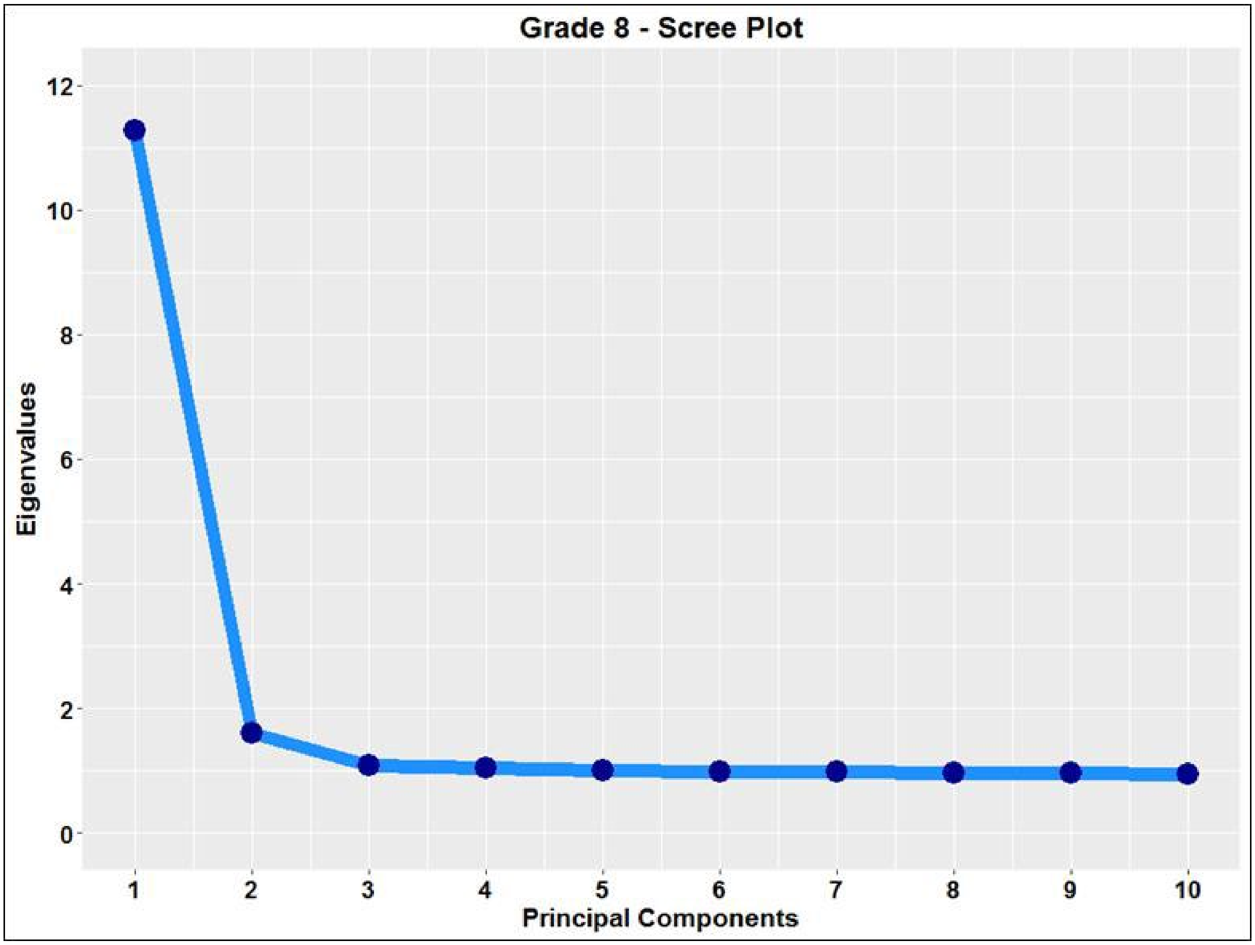


Figure 6.2.2. Grade 8 Scree Plot

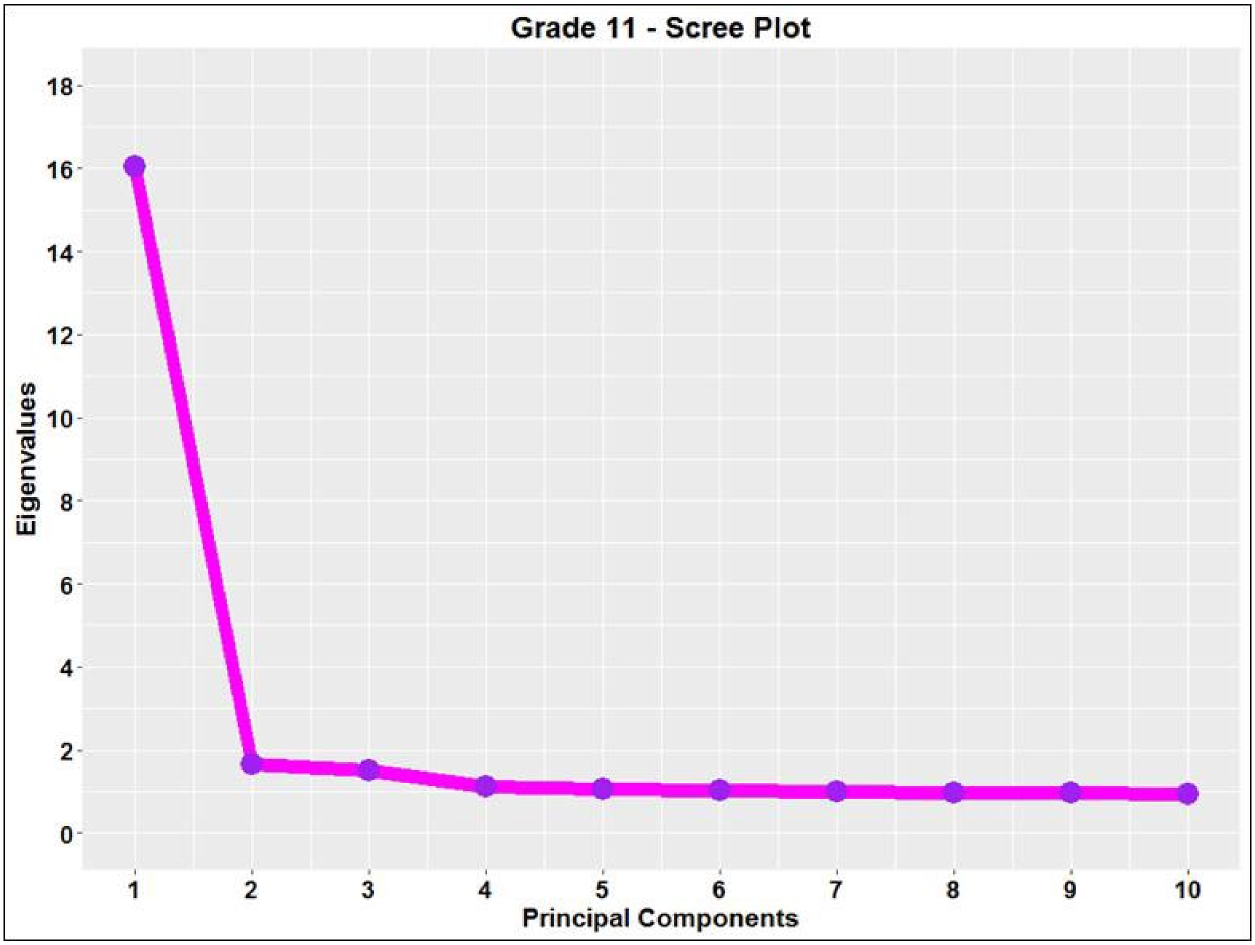


Figure 6.2.3. Grade 11 Scree Plot

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. 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 during the 2019 NJSLA–S IRT calibration processes via Winsteps 3.74 (Linacre, 2012).

Overall, there is ample evidence that both the grade 5 and the grade 8 items fit the assumptions of the PCM. Grade 11 is more complicated. A higher percentage of items were flagged at grade 11 than either of the other two grades for each of the four model fit categories. One possible explanation for the increased model misfit at grade 11 is that student effort on the assessment may have been uneven, or certain high and low performing students performed unexpectedly on items due to knowing specific components within the content, but not others. The grade 11 person fit statistics did not display similar issues, meaning that the student ability estimates were not overly harmed by the increase in percentage of flagged items. The grade 11 model item misfit warrants increased attention as the assessment program evolves.

6.2.2.1 Item infit and outfit. Rasch infit and outfit statistics range from 0 to infinity with 1 representing ideal model fit. 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 are measuring near their ability level (Wright and Masters, 1982). Only three total items across all grades were flagged for problematic infit statistics. Those items were at grade 11.

Outfit statistics are heavily influenced by unexpected student responses to items that are either relatively easy or relatively hard. The NJSLA–S outfit statistics were less positive, with approximately 10–15% of all items being flagged. Grades 5 and 11 displayed the most egregious outfit statistics. The problematic outfit statistics, however, are less of a threat to the validity of test score interpretations than are 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. The following table provides a summary of item infit and outfit statistics at each grade level.

Table 6.2.4: Summary Infit and Outfit Statistics

Grade	Fit Statistic	Mean	Min	Max	Outside 0.7 to 1.3	% Flagged
5	Infit	0.99	0.72	1.24	0 out of 50	0.0%
5	Outfit	1.01	0.53	1.40	6 out of 50	12.0%
8	Infit	1.00	0.81	1.27	0 out of 59	0.0%
8	Outfit	1.00	0.55	1.43	6 out of 59	10.2%
11	Infit	1.00	0.76	1.39	3 out of 68	4.4%
11	Outfit	1.01	0.67	1.64	9 out of 68	13.2%

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. The assumption of equal discrimination can be tested with the Rasch discrimination statistic, as well as the correlations presented earlier in the CTT section. 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 discrimination statistics fall outside of the range of 0.5 to 1.5.

At grade 5 the discrimination statistics looked exceptional. Only one item was flagged for having a value outside the 0.5 to 1.5 threshold. The grade 8 values were good as well. Only five items out of 59 (8.5%) were flagged, and the maximum value was only slightly above the 1.5 threshold. Grade 11 saw 10 out of 68 items flagged (14.7%). Many of these items were also flagged for outfit, providing further evidence that some of the items on the grade 11 test were misfitting the PCM. The following table provides a summary of discrimination statistics at each grade level.

Table 6.2.5: Summary Rasch Discrimination Statistics

Grade	Fit Statistic	Mean	Min	Max	Outside 0.5 to 1.5	% Flagged
5	Discrimination	1.00	0.38	1.48	1 out of 50	2.0%
8	Discrimination	1.01	0.24	1.57	5 out of 59	8.5%
11	Discrimination	1.01	0.11	1.60	10 out of 68	14.7%

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 guessing assumption is met if items remain within a reasonable threshold. The assumption of guessing can be tested with the Rasch lower asymptote statistic. Rasch lower asymptote statistics are ideally 0.0, 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 0.1.

At grade 5 the lower asymptote statistics met the assumption. Only four items out of 50 (8%) were flagged for having values outside the 0.1 threshold. The grade 8 values were even better; only four items out of 59 (6.8%) were flagged. Unsurprisingly, given the results of the infit, outfit, and discrimination statistics, grade 11 had more items flagged. Grade 11 saw eight out of 68 items flagged (11.8%). A majority of these eight items were also flagged for either infit, outfit, or discrimination, again providing further evidence that some of the items on the grade 11 test were misfitting the PCM. The following table provides a summary of the lower asymptote statistics at each grade level.

Table 6.2.6: Summary Rasch Lower Asymptote Statistics

Grade	Fit Statistic	Mean	Min	Max	Greater Than 0.1	% Flagged
5	Lower Asymptote	0.03	0.00	0.16	4 out of 50	8.0%
8	Lower Asymptote	0.03	0.00	0.20	4 out of 59	6.8%
11	Lower Asymptote	0.03	0.00	0.22	8 out of 68	11.8%

6.2.2.4 Rasch person infit and outfit. PCM person fit statistics are useful for evaluating whether student response patterns are reasonable. The 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 measures of 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. Values below 1.0 would indicate that the data are more predictable than anticipated by the PCM; values above 1.0 would indicate that the data are less predictable.

Person fit statistics were evaluated based on the following demographics: gender, ethnicity, English learner (EL) status, economically disadvantaged (EconDis) status, students with disabilities (SWD) status, and by all major forms. Tables 6.2.7 and 6.2.8 show person infit and outfit descriptive statistics by demographic variables. Tables 6.2.9 and 6.2.10 breakdown the person infit and outfit descriptive statistics by CBT, PBT, TTS, Spanish, Spanish TTS, and Human Reader forms. Figures 6.2.4 through 6.2.6 show grade level distributions for all students of both the person infit and outfit statistics.

Overall, there were relatively very few students flagged for aberrant person infit or outfit statistics. Less than 5% of students were flagged for person infit statistics at all combinations of grade and demographic variables. As shown in Table 6.2.9, the grade 5 PBT and Spanish TTS forms flagged 6.19% and 5.43% of students for person infit. However, those percentages only represent 7 PBT and 21 SP TTS test takers.

The person outfit statistics were similarly positive, except those for EL students, students with disabilities, and many of the accommodated forms. The EL students and students with disabilities tended to have higher percentages of students that were flagged for person outfit statistics relative to the other demographic groups at their grade level. Within those groups the students that were flagged also tended to be lower performing. They also were more likely to have taken accommodated forms, which themselves had much higher percentages of students flagged for person outfit than did the CBT forms. As stated earlier, aberrant person infit statistics are more of a threat to the validity of the inferences than are aberrant person outfit 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 so low performing, it is unlikely that the misfit was having any meaningful impact on the reliability of the student proficiency classification. That being said, a deeper investigation into the person outfit statistics for EL students, students with disabilities, and the accommodated forms is warranted and described in greater detail in Part 9.7.3: Future Studies.

Table 6.2.7: Person Infit Statistics by Demographic Group

Grade	Group	N	Mean Scale Score	Person Infit Mean	Person Infit Min	Person Infit Max	Flagged N	Flagged Percent	Flagged Mean Scale Score
5	NJSLAS	101,220	169.79	1.02	0.46	3.04	3,730	3.69	180.71
5	Male	51,656	170.14	1.03	0.46	3.04	1,967	3.81	180.97
5	Female	49,564	169.42	1.01	0.51	2.92	1,763	3.56	180.42
5	Am. Indian	133	168.25	1.01	0.66	1.71	3	2.26	176.33
5	Asian	10,859	203.08	1.02	0.50	2.56	537	4.95	209.34
5	Black	15,345	143.78	1.03	0.51	2.79	447	2.91	153.17
5	Hispanic	29,836	152.10	1.03	0.48	2.80	1,015	3.40	161.08
5	Pacific Islander	189	177.38	1.02	0.57	2.00	6	3.17	176.67
5	White	42,442	182.59	1.01	0.46	3.04	1,633	3.85	191.61
5	EL – Yes	5,830	125.02	1.07	0.58	2.54	187	3.21	131.94
5	EL – No	95,382	172.53	1.01	0.46	3.04	3,543	3.71	183.28
5	EconDis – Yes	38,634	147.50	1.03	0.48	2.71	1,254	3.25	157.76
5	EconDis – No	62,586	183.54	1.01	0.46	3.04	2,476	3.96	192.33
5	SWD – Yes	20,499	143.14	1.05	0.48	2.94	664	3.24	158.14
5	SWD – No	80,721	176.56	1.01	0.46	3.04	3,066	3.80	185.59
8	NJSLAS	99,852	165.46	1.02	0.54	3.14	1,674	1.68	175.14
8	Male	51,124	164.81	1.02	0.54	3.14	905	1.77	179.93
8	Female	48,728	166.14	1.01	0.57	2.27	769	1.58	169.51
8	Am. Indian	117	159.35	1.01	0.72	2.11	2	1.71	122.00
8	Asian	10,346	194.07	1.01	0.57	2.48	219	2.12	203.93
8	Black	14,452	144.41	1.04	0.61	2.23	195	1.35	156.26
8	Hispanic	28,176	149.72	1.02	0.59	3.14	417	1.48	156.84
8	Pacific Islander	207	179.94	0.99	0.70	1.77	6	2.90	185.33
8	White	44,716	175.15	1.00	0.54	2.48	791	1.77	181.55
8	EL – Yes	4,381	129.12	1.08	0.65	2.48	80	1.83	136.00
8	EL – No	95,468	167.13	1.01	0.54	3.14	1,594	1.67	177.11
8	EconDis – Yes	34,908	146.73	1.03	0.61	2.27	489	1.40	152.53
8	EconDis – No	64,944	175.53	1.01	0.54	3.14	1,185	1.82	184.48

Grade	Group	N	Mean Scale Score	Person Infit Mean	Person Infit Min	Person Infit Max	Flagged N	Flagged Percent	Flagged Mean Scale Score
8	SWD – Yes	19,664	144.42	1.05	0.60	2.27	294	1.50	154.63
8	SWD – No	80,188	170.62	1.01	0.54	3.14	1,380	1.72	179.51
11	NJSLAS	90,024	163.40	1.02	0.56	3.79	2,164	2.40	180.32
11	Male	45,733	162.37	1.02	0.58	3.79	1,024	2.24	190.52
11	Female	44,291	164.48	1.03	0.56	3.20	1,140	2.57	171.16
11	Am. Indian	111	162.65	1.03	0.70	1.57	2	1.80	240.50
11	Asian	9,097	203.56	1.02	0.58	3.21	301	3.31	217.17
11	Black	12,935	136.07	1.03	0.58	3.79	251	1.94	146.59
11	Hispanic	23,417	143.79	1.04	0.59	3.20	636	2.72	149.70
11	Pacific Islander	223	184.22	1.04	0.59	2.72	10	4.48	149.50
11	White	43,112	173.54	1.02	0.56	2.92	936	2.17	197.31
11	EL – Yes	3,878	115.24	1.05	0.69	3.20	67	1.73	129.18
11	EL – No	86,132	165.58	1.02	0.56	3.79	2,097	2.43	181.95
11	EconDis – Yes	27,411	140.93	1.04	0.58	3.79	684	2.50	148.21
11	EconDis – No	62,613	173.24	1.02	0.56	3.21	1,480	2.36	195.16
11	SWD – Yes	16,414	135.74	1.04	0.59	2.48	295	1.80	160.91
11	SWD – No	73,610	169.57	1.02	0.56	3.79	1,869	2.54	183.38

Table 6.2.8: Person Outfit Statistics by Demographic Group

Grade	Group	N	Mean Scale Score	Person Outfit Mean	Person Outfit Min	Person Outfit Max	Flagged N	Flagged Percent	Flagged Mean Scale Score
5	NJSLAS	101,220	169.79	1.01	0.21	5.69	5,500	5.43	152.52
5	Male	51,656	170.14	1.02	0.22	5.69	3,263	6.32	150.61
5	Female	49,564	169.42	1.00	0.21	4	2,237	4.51	155.30
5	Am. Indian	133	168.25	1.02	0.36	1.92	8	6.02	155.88
5	Asian	10,859	203.08	0.98	0.21	4.06	651	6.00	231.98
5	Black	15,345	143.78	1.06	0.34	4.44	1,146	7.47	110.93
5	Hispanic	29,836	152.10	1.04	0.3	4.1	1,887	6.32	118.57
5	Pacific Islander	189	177.38	0.97	0.43	1.7	7	3.70	135.57
5	White	42,442	182.59	0.98	0.22	5.69	1,680	3.96	185.91
5	EL – Yes	5,830	125.02	1.16	0.36	5.69	836	14.34	103.69
5	EL – No	95,382	172.53	1.00	0.21	4.44	4,663	4.89	161.28
5	EconDis – Yes	38,634	147.50	1.05	0.23	4.44	2,614	6.77	113.39
5	EconDis – No	62,586	183.54	0.99	0.21	5.69	2,886	4.61	187.96
5	SWD – Yes	20,499	143.14	1.09	0.22	4.37	2,043	9.97	113.05
5	SWD – No	80,721	176.56	0.99	0.21	5.69	3,457	4.28	175.85
8	NJSLAS	99,852	165.46	1.00	0.3	5.26	2,711	2.72	118.34
8	Male	51,124	164.81	1.01	0.3	5.26	1,662	3.25	117.22
8	Female	48,728	166.14	1.00	0.38	3.36	1,049	2.15	120.13
8	Am. Indian	117	159.35	0.99	0.68	2.11	5	4.27	119.60
8	Asian	10,346	194.07	0.97	0.35	5.26	116	1.12	158.05
8	Black	14,452	144.41	1.06	0.47	5.06	715	4.95	114.02
8	Hispanic	28,176	149.72	1.04	0.3	4.77	1,165	4.13	113.90
8	Pacific Islander	207	179.94	0.95	0.67	1.36	0	0.00	
8	White	44,716	175.15	0.98	0.43	3.36	670	1.50	123.34
8	EL – Yes	4,381	129.12	1.13	0.3	3.18	435	9.93	111.12
8	EL – No	95,468	167.13	1.00	0.35	5.26	2,276	2.38	119.72
8	EconDis – Yes	34,908	146.73	1.05	0.35	5.06	1,589	4.55	113.91
8	EconDis – No	64,944	175.53	0.98	0.3	5.26	1,122	1.73	124.62
8	SWD – Yes	19,664	144.42	1.07	0.49	3.66	1,231	6.26	112.62
8	SWD – No	80,188	170.62	0.99	0.3	5.26	1,480	1.85	123.11
11	NJSLAS	90,024	163.40	1.00	0.22	7.7	2,296	2.55	124.70

Grade	Group	N	Mean Scale Score	Person Outfit Mean	Person Outfit Min	Person Outfit Max	Flagged N	Flagged Percent	Flagged Mean Scale Score
11	Male	45,733	162.37	1.00	0.22	7.7	1,285	2.81	128.59
11	Female	44,291	164.48	1.01	0.28	5.31	1,011	2.28	119.75
11	Am. Indian	111	162.65	1.00	0.52	1.99	2	1.80	100.00
11	Asian	9,097	203.56	0.95	0.22	2.58	173	1.90	218.65
11	Black	12,935	136.07	1.05	0.37	7.7	510	3.94	106.54
11	Hispanic	23,417	143.79	1.04	0.43	5.31	809	3.45	108.81
11	Pacific Islander	223	184.22	1.00	0.51	2.92	7	3.14	127.43
11	White	43,112	173.54	0.99	0.26	3.75	766	1.78	132.56
11	EL – Yes	3,878	115.24	1.08	0.59	5.31	198	5.11	102.37
11	EL – No	86,132	165.58	1.00	0.22	7.7	2,098	2.44	126.81
11	EconDis – Yes	27,411	140.93	1.04	0.28	7.7	997	3.64	109.88
11	EconDis – No	62,613	173.24	0.99	0.22	5.31	1,299	2.07	136.07
11	SWD – Yes	16,414	135.74	1.05	0.41	4.7	666	4.06	106.42
11	SWD – No	73,610	169.57	0.99	0.22	7.7	1,630	2.21	132.17

Table 6.2.9: Person Infit Statistics by Form

Grade	Group	N	Mean Scale Score	Person Infit Mean	Person Infit Min	Person Infit Max	Flagged N	Flagged Percent	Flagged Mean Scale Score
5	CBT	81,996	175.89	1.01	0.46	3.04	2,961	3.61	184.37
5	PBT	113	133.24	1.12	0.60	2.29	7	6.19	144.43
5	TTS	17,551	145.59	1.04	0.48	2.80	580	3.30	160.10
5	SP	825	120.70	1.08	0.64	2.02	26	3.15	124.65
5	SP TTS	387	121.51	1.10	0.70	2.25	21	5.43	127.19
5	Human Reader	303	135.03	1.06	0.68	1.99	8	2.64	171.50
8	CBT	85,985	169.05	1.01	0.56	3.14	1,460	1.70	178.21
8	PBT	92	132.76	1.08	0.72	1.79	3	3.26	155.00
8	TTS	12,100	145.74	1.04	0.54	2.48	181	1.50	157.48
8	SP	1,126	124.18	1.11	0.66	1.79	20	1.78	132.80
8	SP TTS	432	128.09	1.10	0.74	1.92	16	3.70	132.31
8	Human Reader	84	127.29	1.06	0.77	1.47	0	0.00	N/A
11	CBT	84,298	165.27	1.02	0.57	3.79	1,882	2.23	187.21
11	PBT	154	135.82	1.10	0.59	1.80	6	3.90	162.17
11	TTS	4,408	141.60	1.03	0.59	2.57	66	1.50	166.83
11	SP	761	115.46	1.06	0.69	1.68	10	1.31	121.40
11	SP TTS	262	111.52	1.06	0.79	3.20	4	1.53	114.25
11	Human Reader	113	113.97	1.06	0.83	1.78	3	2.65	146.67

Table 6.2.10: Person Outfit Statistics by Form

Grade	Group	N	Mean Scale Score	Person Outfit Mean	Person Outfit Min	Person Outfit Max	Flagged N	Flagged Percent	Flagged Mean Scale Score
5	CBT	81,996	175.89	0.99	0.21	5.75	3,551	4.33	171.15
5	PBT	113	133.24	1.24	0.51	2.83	22	19.47	102.68
5	TTS	17,551	145.59	1.08	0.27	4.11	1647	9.38	116.26
5	SP	825	120.70	1.16	0.47	2.97	117	14.18	102.82
5	SP TTS	387	121.51	1.16	0.57	2.74	56	14.47	102.41
5	Human Reader	303	135.03	1.12	0.45	2.97	32	10.56	106.53
8	CBT	85,985	169.05	0.99	0.35	5.25	1,849	2.15	121.14
8	PBT	92	132.76	1.16	0.72	2.20	12	13.04	114.08
8	TTS	12,100	145.74	1.06	0.48	3.20	650	5.37	113.07
8	SP	1,126	124.18	1.16	0.30	3.19	121	10.75	107.38
8	SP TTS	432	128.09	1.12	0.61	2.70	37	8.56	111.27
8	Human Reader	84	127.29	1.13	0.57	2.33	8	9.52	107.38
11	CBT	84,298	165.27	1.00	0.22	7.45	1,952	2.32	126.42
11	PBT	154	135.82	1.16	0.59	2.04	13	8.44	104.38
11	TTS	4,408	141.60	1.04	0.44	3.04	161	3.65	106.95
11	SP	761	115.46	1.11	0.67	2.52	43	5.65	101.70
11	SP TTS	262	111.52	1.17	0.64	5.14	30	11.45	101.70
11	Human Reader	113	113.97	1.09	0.76	1.89	5	4.42	103.40

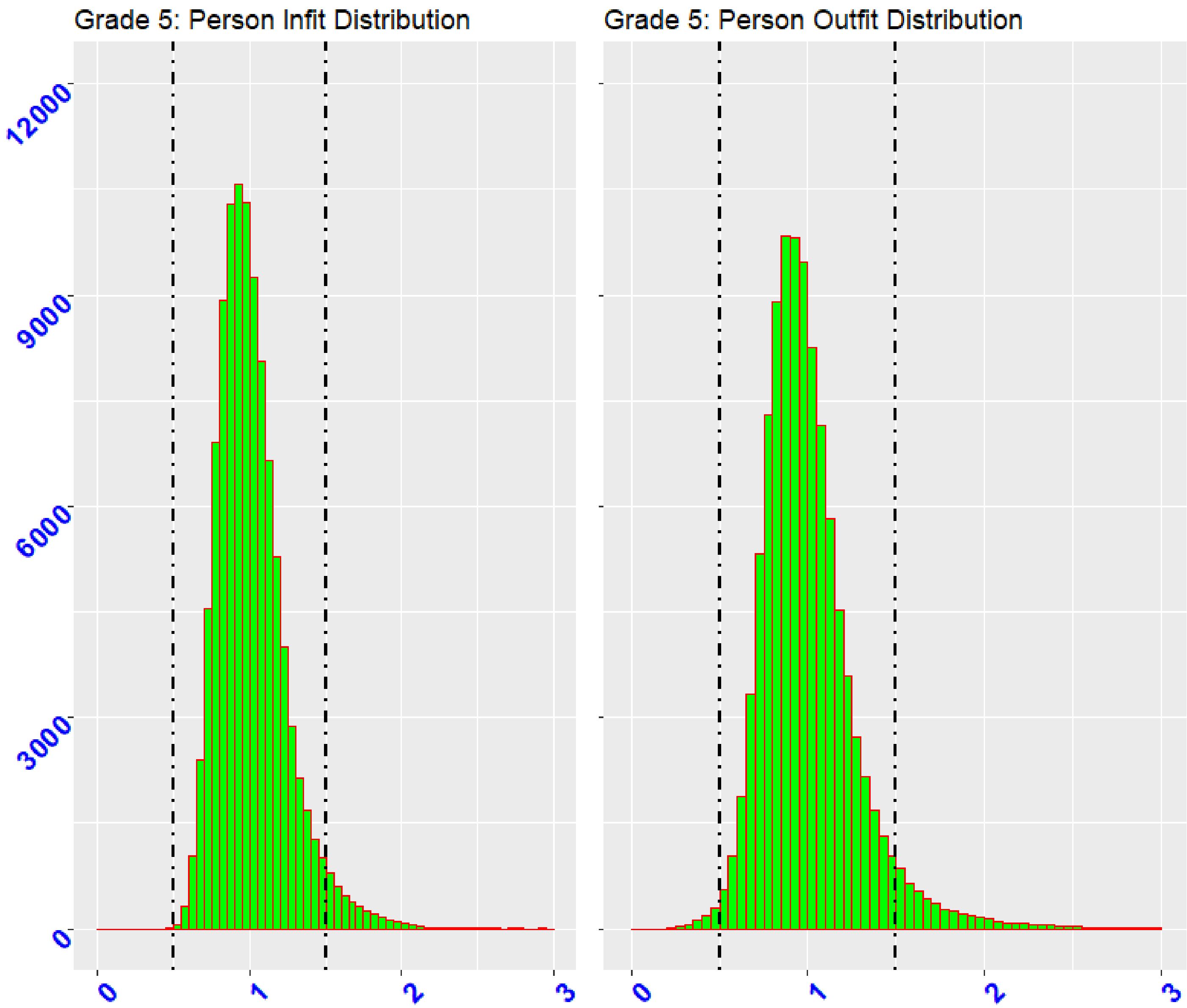


Figure 6.2.4. Grade 5 Person Infit and Outfit Distributions

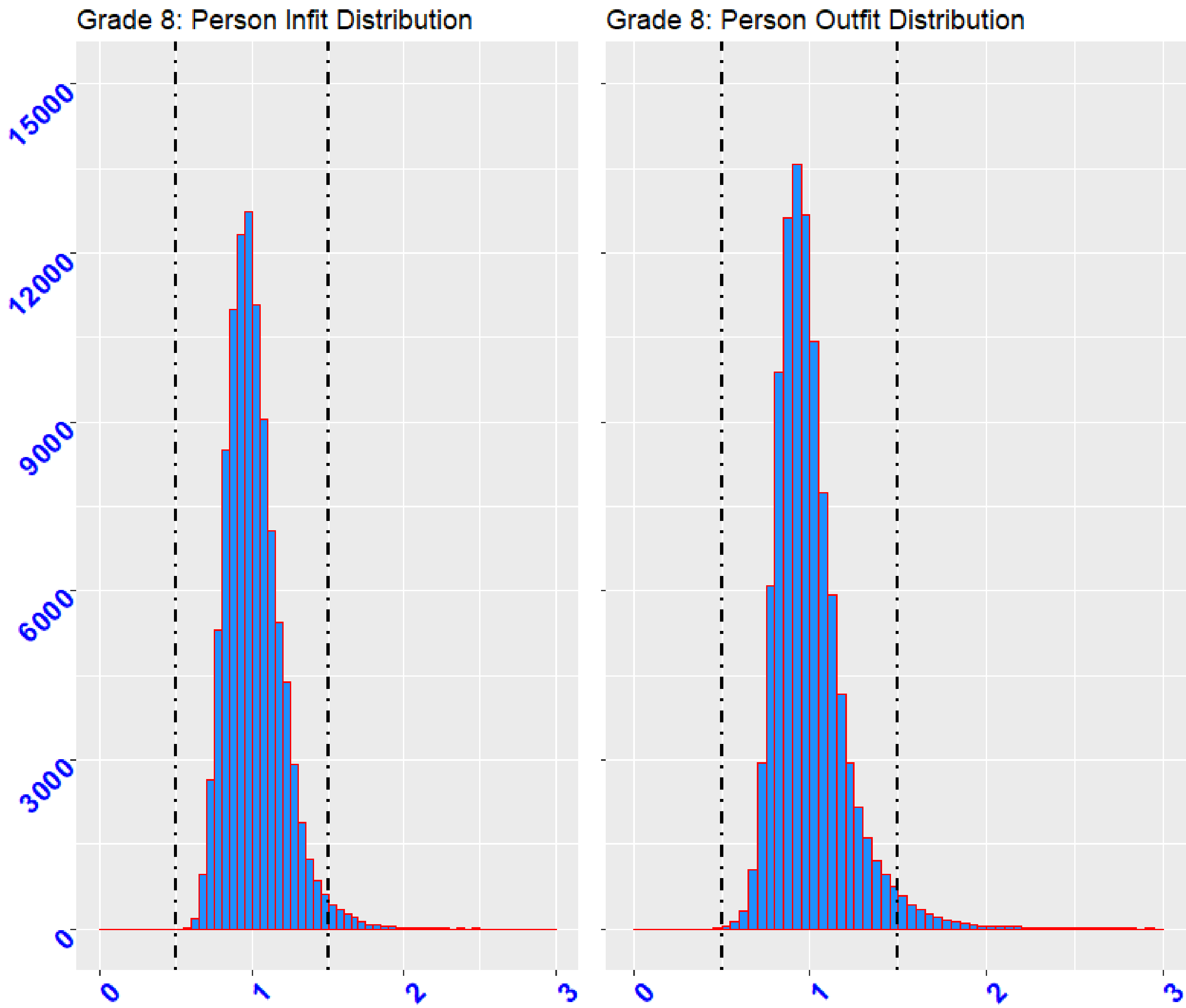


Figure 6.2.5. Grade 8 Person Infit and Outfit Distributions

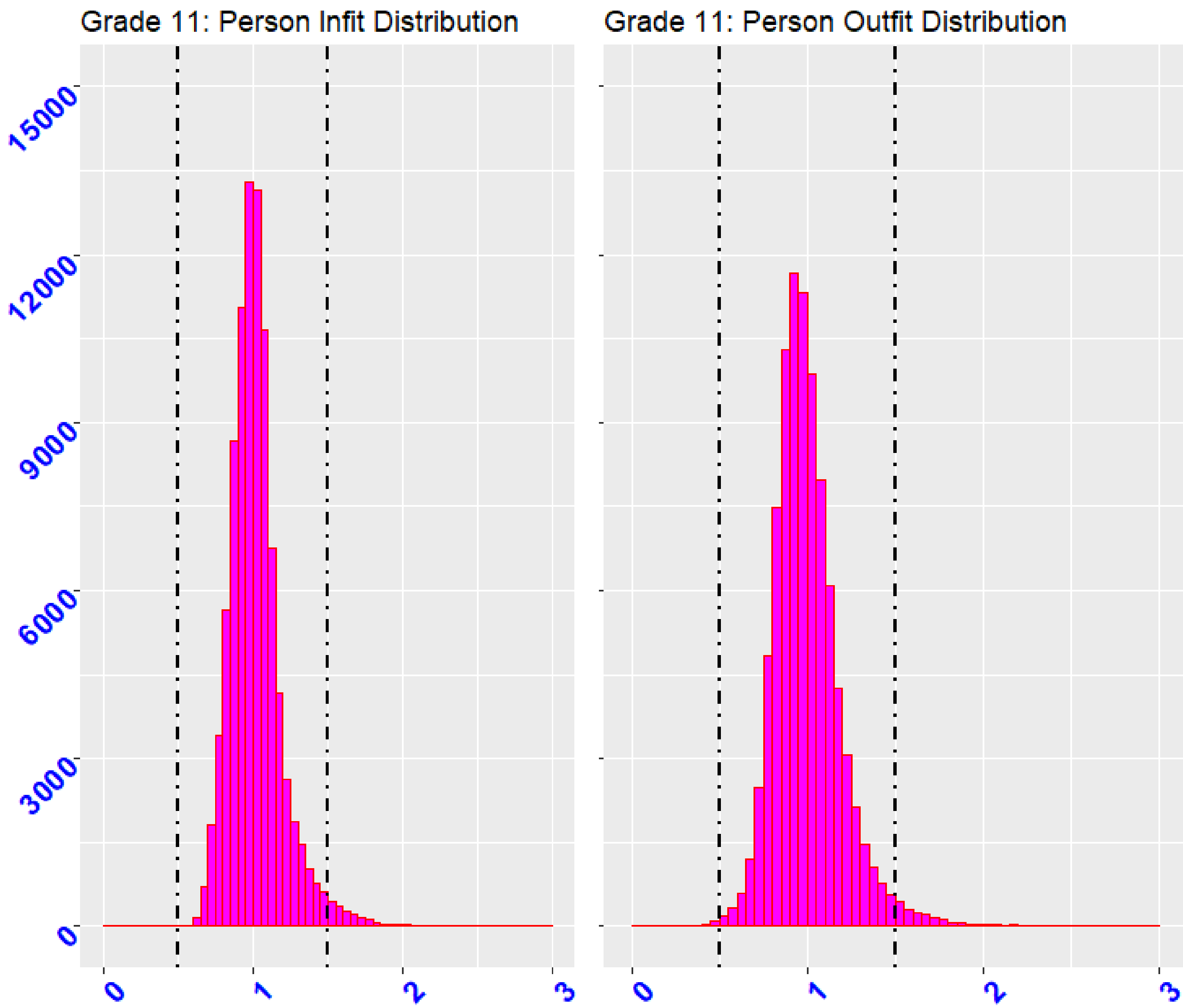


Figure 6.2.6. Grade 11 Person Infit and Outfit Distributions

6.2.3 Local Independence

The PCM assumes that student responses to items are independent from 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 that could pose a threat to the validity of inferences made from test scores, 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 (1984) Q3, which is a residual correlation. All 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 6.2.11 summarizes Yen's Q3 statistics at each grade level.

Table 6.2.11: Summary Yen's Q3 Statistics

Grade	Mean	Min	Max	Outside -0.2 to 0.2	% Flagged
5	–0.02	–0.11	0.30	4 out of 1,225	0.3%
8	–0.02	–0.10	0.19	0 out of 1,711	0.0%
11	–0.01	–0.13	0.25	3 out of 2,278	0.1%

6.2.4 Item Characteristic Curves – CR Items

Item characteristic curves (ICCs) show the relationship between latent student ability (theta) and the probability of achieving a specific score point on a given item. The ICCs for each of the 0–4 point, hand-scored, constructed-response items are presented in Figures 6.2.7 through 6.2.15 below. The vertical dashed lines represent from left to right the Levels 2–4 cut scores. Table 6.2.12 shows the percentages of students receiving each score point for all nine CR items.

Table 6.2.12: Constructed-Response Point Distribution Percentages

Grade	Item	%0	%1	%2	%3	%4
5	CR Item 1	44.8	21.7	18.6	9.0	5.9
5	CR Item 2	12.0	12.5	19.2	25.0	31.3
5	CR Item 3	36.0	39.1	17.8	3.9	3.2
8	CR Item 1	35.3	16.3	18.2	19.2	11.0
8	CR Item 2	57.3	12.7	14.8	8.5	6.7
8	CR Item 3	27.8	44.0	15.0	5.7	7.6
11	CR Item 1	45.8	15.3	22.6	12.0	4.4
11	CR Item 2	62.6	15.5	9.8	7.1	5.0
11	CR Item 3	53.1	11.6	12.2	11.8	11.3

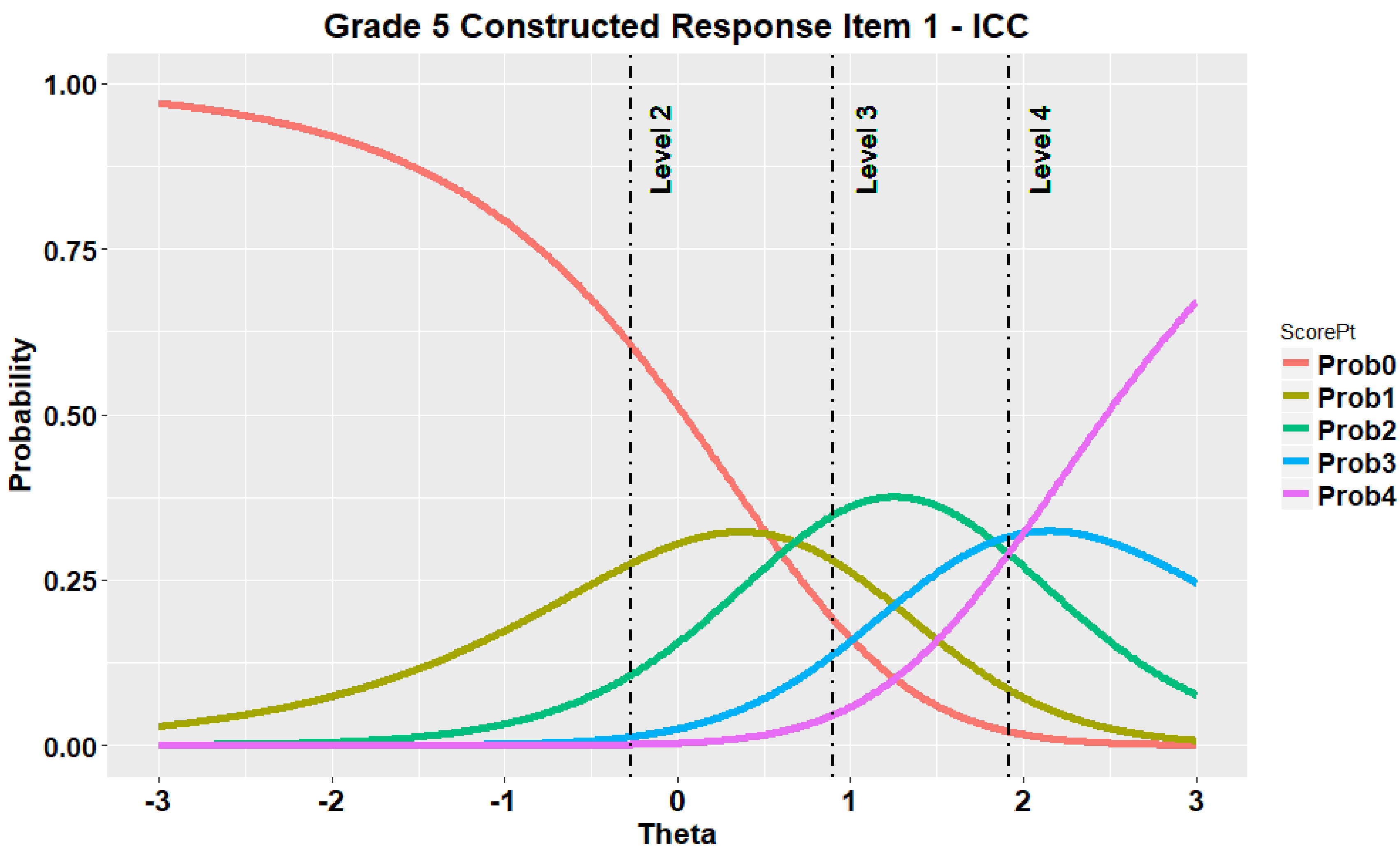


Figure 6.2.7. Grade 5 Constructed Response Item 1 ICC

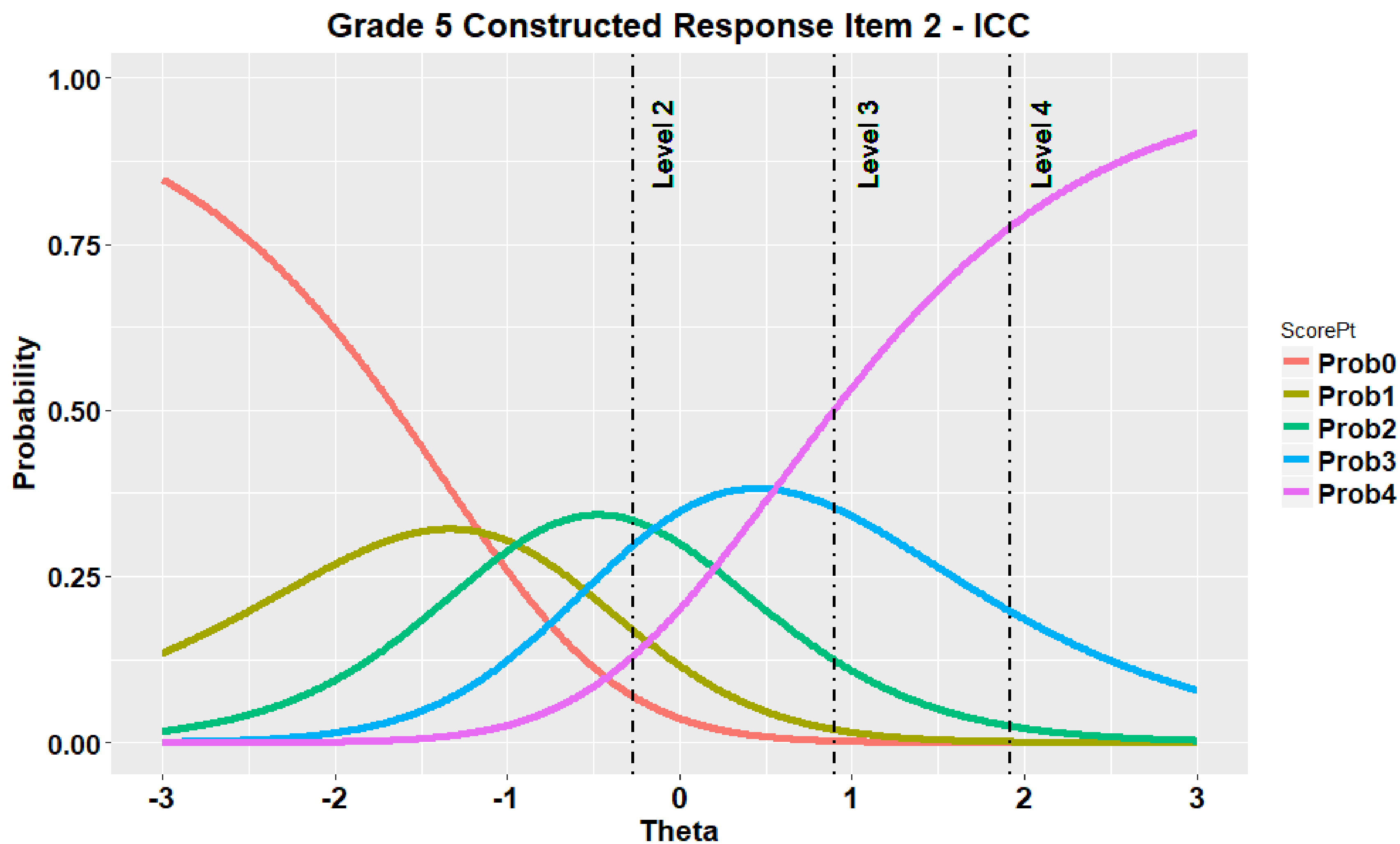


Figure 6.2.8. Grade 5 Constructed Response Item 2 ICC

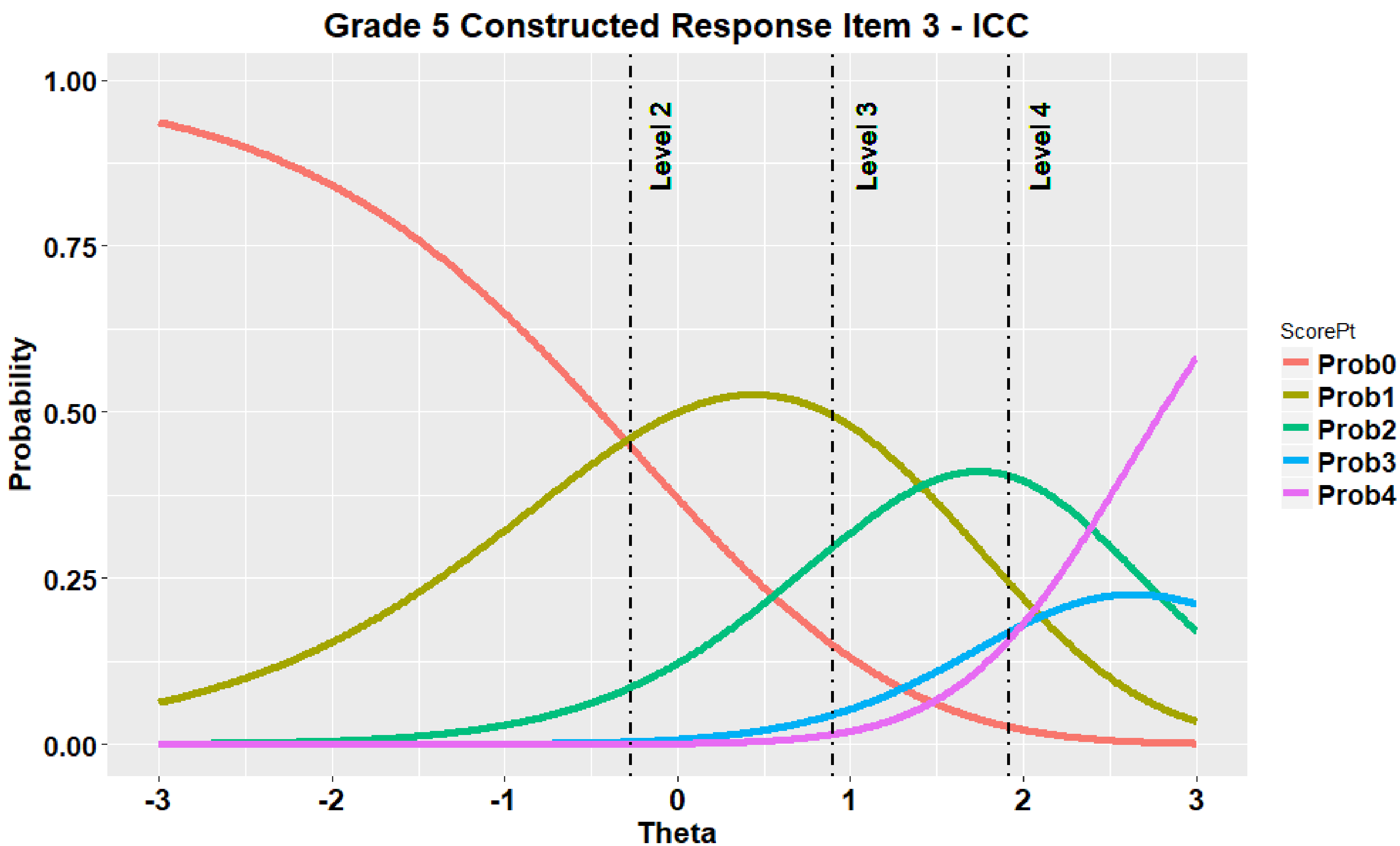


Figure 6.2.9. Grade 5 Constructed Response Item 3 ICC

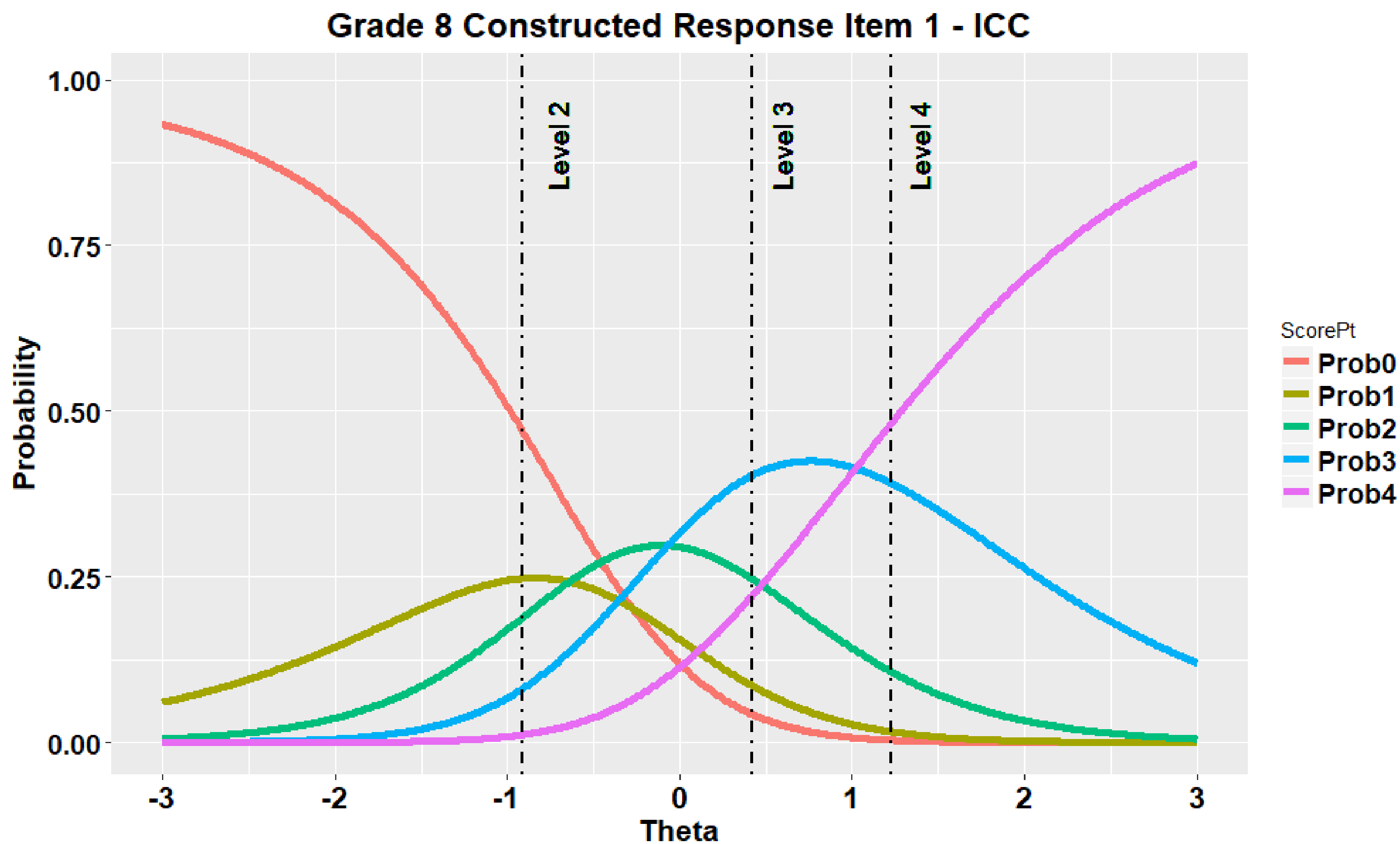


Figure 6.2.10. Grade 8 Constructed Response Item 1 ICC

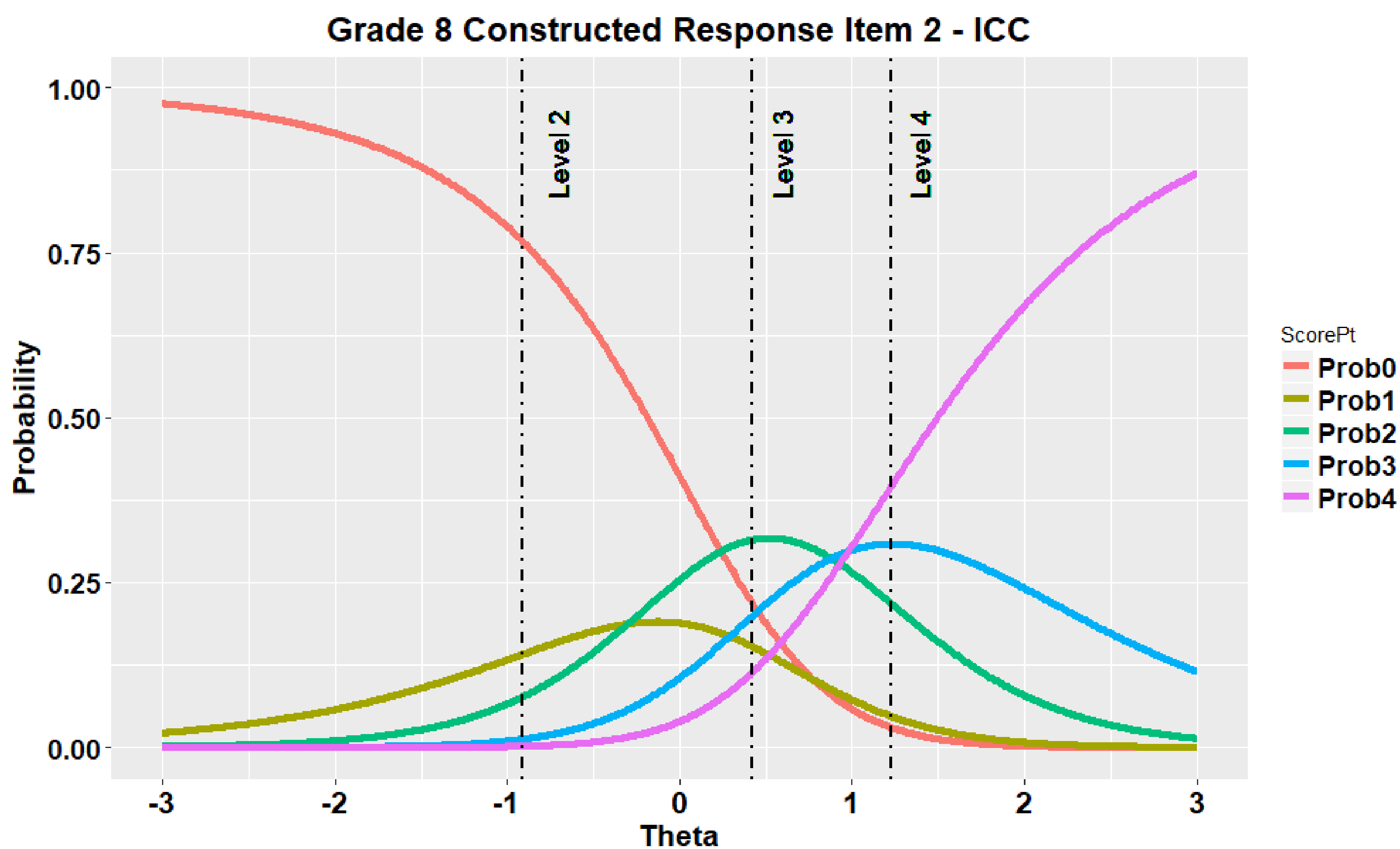


Figure 6.2.11. Grade 8 Constructed Response Item 2 ICC

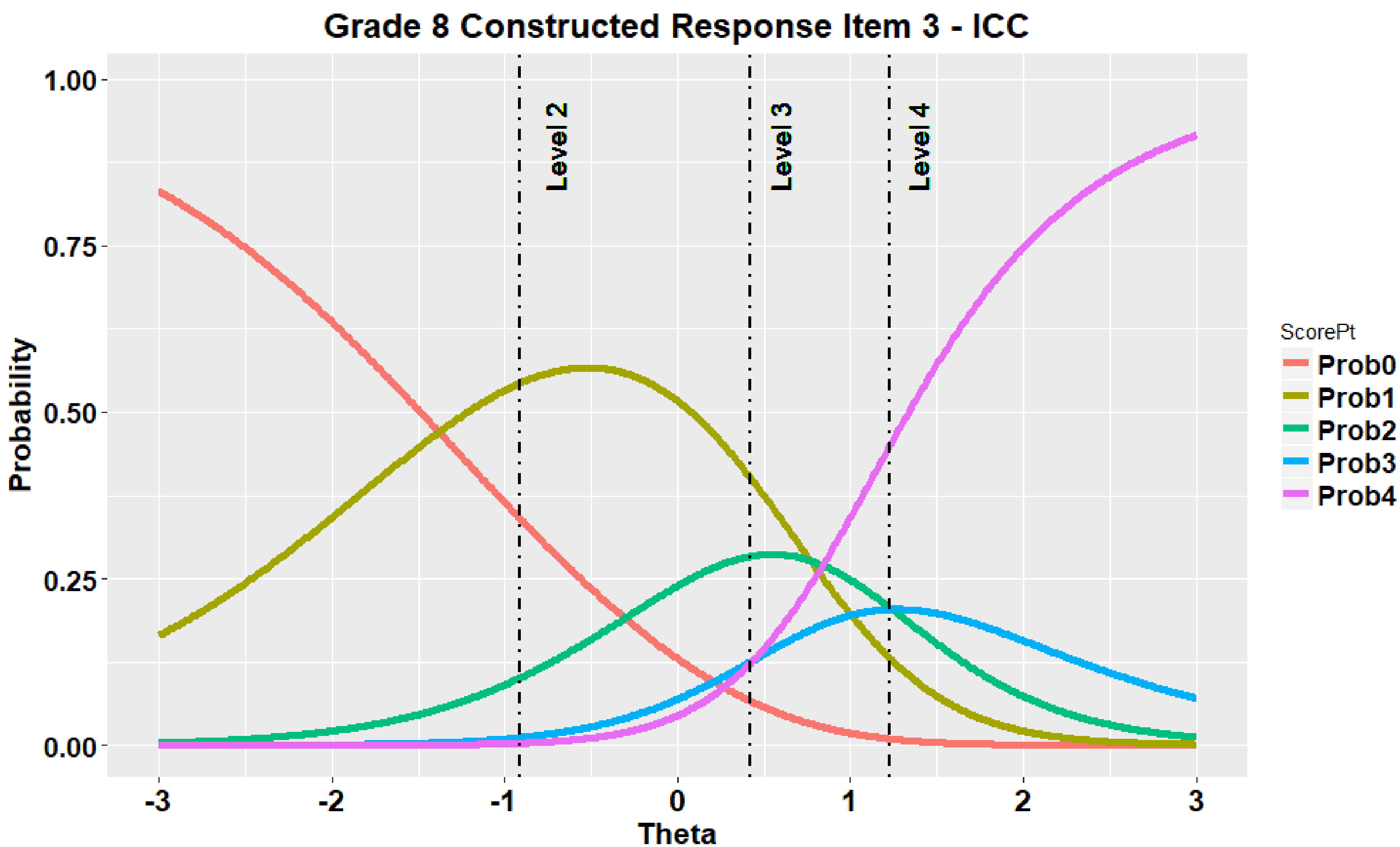


Figure 6.2.12. Grade 8 Constructed Response Item 3 ICC

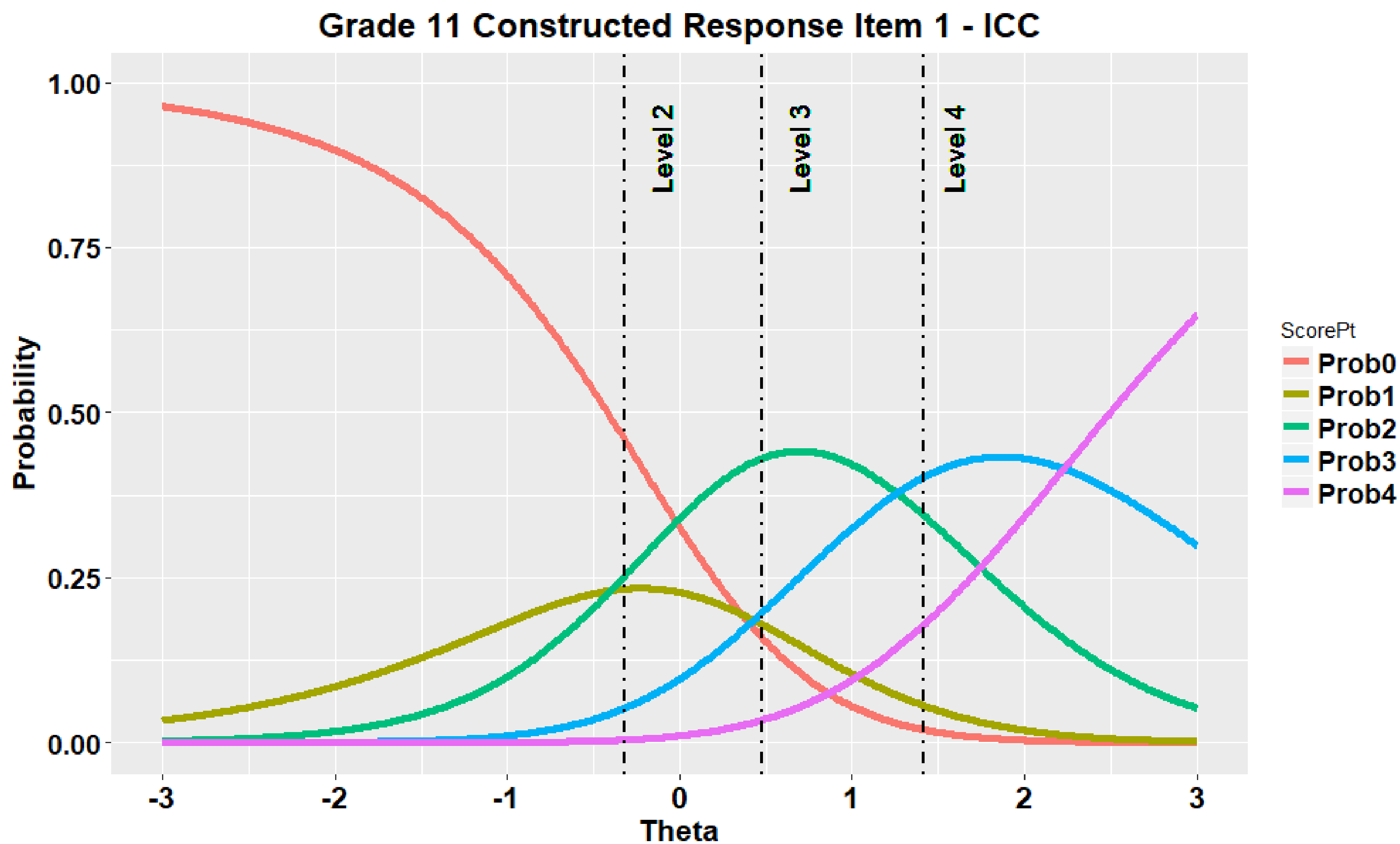


Figure 6.2.13. Grade 11 Constructed Response Item 1 ICC

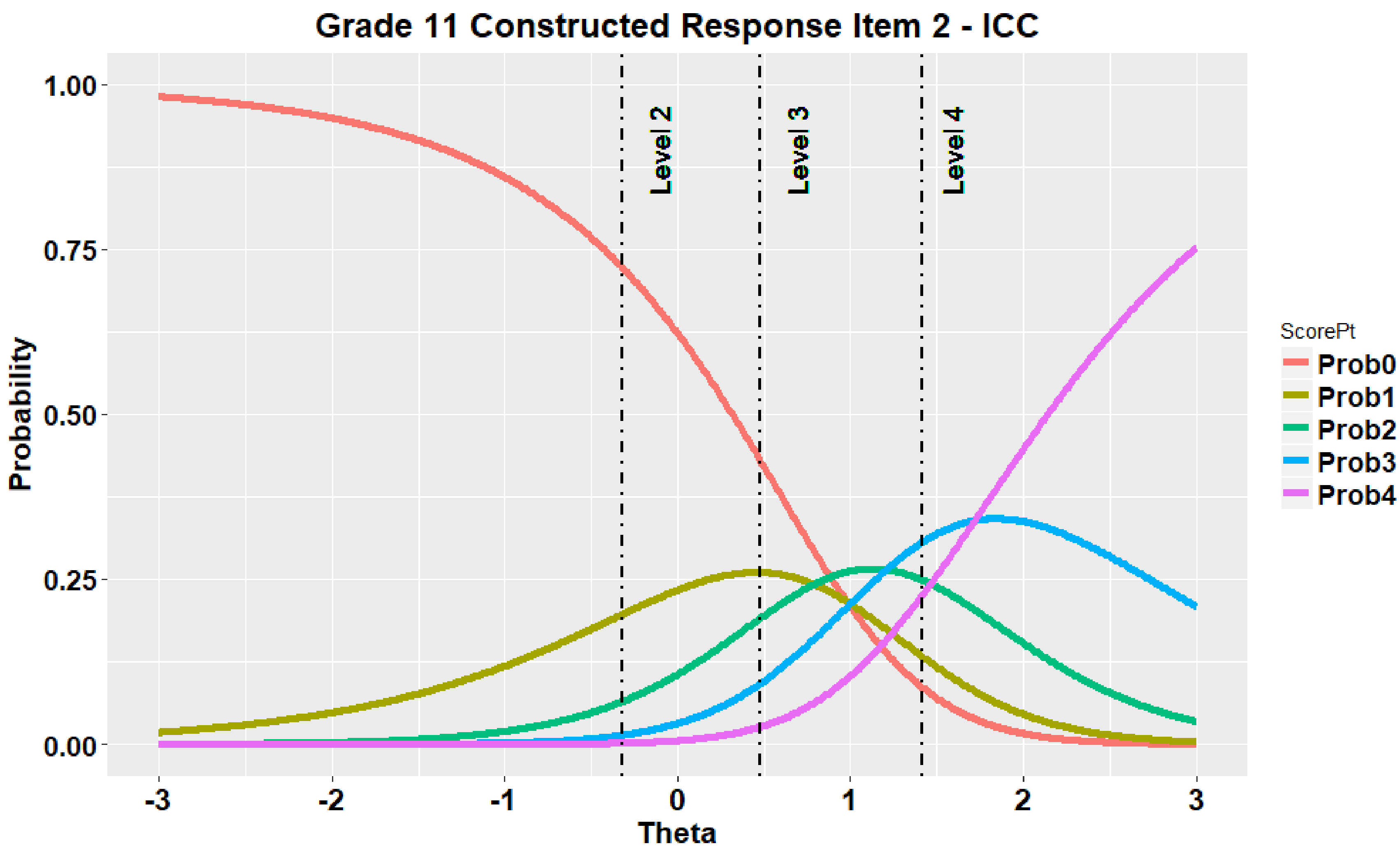


Figure 6.2.14. Grade 11 Constructed Response Item 2 ICC

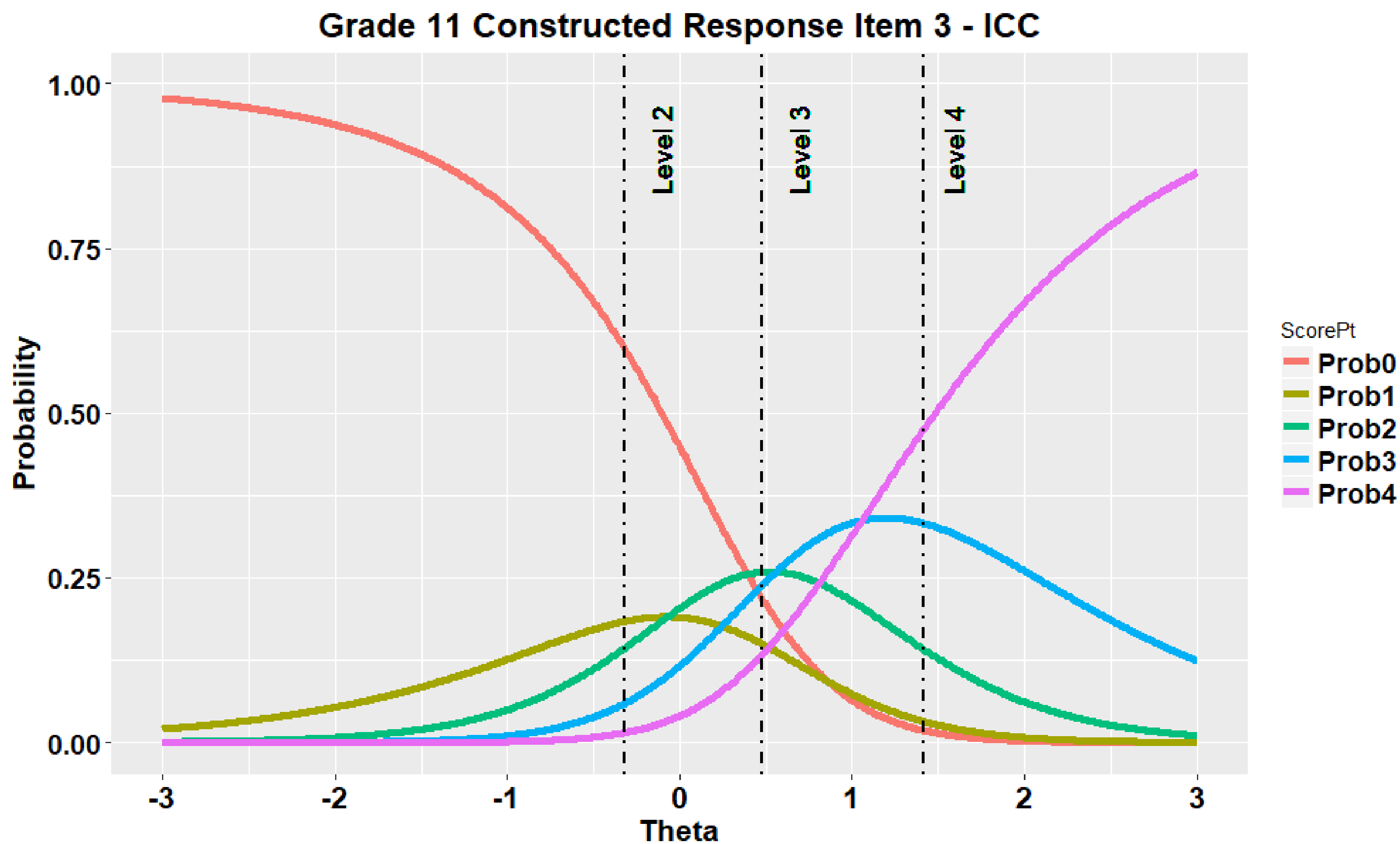


Figure 6.2.15. Grade 11 Constructed Response Item 3 ICC

6.2.5 Descriptive Statistics - Scale Score

Descriptive statistics for scale scores and performance level distributions by form and demographic groups are presented in the following sections.

6.2.5.1 Scale score distributions by grade. Descriptive statistics for scale scores and percentage distributions of students' performance levels are summarized by grade in Table 6.2.13. For all test forms, scale scores have a range of 100 to 300. The Level 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 Levels 2 and 4 cut score ranges are more complex and can be found in Part 1 of this technical report.

Table 6.2.13: Descriptive Statistics of Students' Performance Levels by Grade

Grade	Form	N+	Mean	SD	Min	Max	%L1	%L2	%L3	%L4
5	CBT	81,996	175.89	44.61	100	300	28.86	38.17	25.50	7.48
5	PBT	113	133.24	38.65	100	249	71.68	18.58	7.96	1.77
5	TTS	17,551	145.59	43.20	100	300	58.28	27.51	11.33	2.89
5	SP	825	120.70	26.02	100	238	83.39	15.39	1.21	0.00
5	SP TTS	387	121.51	27.87	100	249	83.72	14.73	1.29	0.26
5	HR	303	135.03	33.45	100	243	69.31	25.08	5.28	0.33
8	CBT	85,985	169.05	36.09	100	300	31.27	46.83	16.83	5.06
8	PBT	92	132.76	28.08	100	215	75.00	21.74	3.26	0.00
8	TTS	12,100	145.74	33.07	100	283	59.95	32.08	6.57	1.40
8	SP	1,126	124.18	20.94	100	217	87.66	12.08	0.27	0.00
8	SP TTS	432	128.09	21.84	100	215	83.80	15.05	1.16	0.00
8	HR	84	127.29	21.44	100	188	84.52	15.48	0.00	0.00
11	CBT	84,298	165.27	54.14	100	300	47.50	24.13	20.23	8.15
11	PBT	154	135.82	45.75	100	300	72.73	16.23	7.79	3.25
11	TTS	4,408	141.60	46.83	100	300	67.38	18.38	11.21	3.04
11	SP	761	115.46	23.49	100	257	92.38	6.57	0.92	0.13
11	SP TTS	262	111.52	20.91	100	237	96.18	3.44	0.38	0.00
11	HR	113	113.97	22.08	100	185	92.92	7.08	0.00	0.00

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

6.2.5.2 Scale score distributions by demographic group. Descriptive statistics of scale scores and percentage distributions of students' Performance by Demographic Groups can be found on the New Jersey Statewide Assessment Reports webpage. Scale score cumulative frequency distributions are attached as Appendix G.

6.2.5.3 Subscore proficiency classifications. There are no scale scores for the various subscores. As is explained in Part 7, student performance on the subscore categories was classified into three levels: Below, Near/Met, and Above Expectations. Overall, at each grade level the percentages of students who were Below Expectations were consistent across the subscores.

Approximately 50% of all grade 5 students were classified as Below Expectations for all six subscores. At grade 8 there was more variance in the percentages with 54% of students being classified as Below Expectations for Earth and Space Science, and the percentage increasing to 65% for Physical Science. The Below Expectations percentages at grade 11 varied from 52% for Critiquing to 60% for Sensemaking. The percentages of students who were placed in the three subscore proficiency classifications are presented in Appendix K. The data are disaggregated by form type, gender, ethnicity, and other demographic variables.

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). Given that 2019 was the year performance standards were set, year-to-year equating procedures were not necessary. However, due to the minor errors in accommodated form development previously described in Part 3.4.2, two separate score tables were created for a small number of students who received the affected accommodated forms. This section details the scaling procedures that were used after the 2019 NJSLA–S standard setting as well as the equating procedures used for placing the two separate scores tables onto the scales.

7.1 Summary of Scaling Procedures

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 state-wide assessments; in the past, with only three performance levels, scale scores of 200 and 250 represented proficient and advanced proficient performance, respectively (NJDOE, 2017). 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 performance 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 2 and Level 3 cut scores be anchored during the linear transformation and at grade 11, the Level 3 and Level 4 cut scores be anchored. The linear transformation is described in greater detail below.

At all grades, a scale score of 200 still represents the proficient cut point (i.e., Level 3). 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 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 at grade 5 the scale score cut was 243, and at grade 8 it was 231. The scale score ranges are reflected below:

Table 7.1.1: Proficiency Levels by Grade and Cut Score

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

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

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

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. 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

Grade	Cut Score	Theta	Scale Score	Slope	Intercept
5	Level 2	–0.27392	150	42.46393	161.6317
5	Level 3	0.90355	200	42.46393	161.6317
5	Level 4	1.92436	243	42.46393	161.6317
8	Level 2	–0.90778	150	37.78004	184.2960
8	Level 3	0.41567	200	37.78004	184.2960
8	Level 4	1.23068	231	37.78004	184.2960
11	Level 2	– 0.32307	158	52.81895	174.9036
11	Level 3	0.47514	200	52.81895	174.9036
11	Level 4	1.42177	250	52.81895	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 the following rounding rules apply:

- Scale scores below 100 are rounded up to 100.
- Scale scores above 300 are rounded down to 300
- For each performance level:
    a. If the highest raw score that maps to an unrounded scale score less than or equal to the scale score cut (e.g., less than or equal to 200 for Level 3) is greater than the scale score cut minus 0.501 (e.g., greater than 199.499 for Level 3), that raw score is the cut score.
    b. If the highest raw score that maps to an unrounded scale score less than or equal to the scale score cut is also less than or equal to the scale score cut minus 0.501 (e.g., less than or equal to 199.499 for Level 3), that raw score is the cut score, and the associated scale score will be assigned a value of exactly the scale score cut (e.g., 200 for Level 3).

c. In the unlikely event that two raw scores both map to unrounded scale scores less than or equal to the scale score cut and greater than the scale score cut minus 0.501, the lower of the two will be the cut score.
d. When the implementation of these rounding rules results in two raw scores mapping to a *rounded* scale score of exactly the scale score cut, the scale score associated with the higher of the two raw scores will be adjusted upwards by one (1) scale score point.

7.2 Accommodative Form Equivalence

NJDOE (2017) has traditionally used the same score tables for their accommodative forms as for their traditional operational test forms, a decision which 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 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 was explained Part 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. In 2019 this assumption was violated for two types of accommodative forms. The special equatings for those forms are described in the following section.

7.2.1 Special Equatings

Two special equatings were needed for the NJSLA–S in 2019. Errors during the test construction process led to the removal of one item from the grade 5 ASL and Screen-Reader forms and one item from the grade 11 online Spanish forms. In order to place the students who received those forms onto a scale equivalent to that underlying the other CBT forms, special equatings that required re-calculations of score tables were conducted using the item parameters from the standard setting process. The following steps were taken to ensure the special equatings 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 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 Part 7.1.1 were applied to the transformed scale scores, resulting in a separate raw-to-scale score table for each special equating that could be interpreted exactly the same as the other operational forms.

The resulting score tables for the affected accommodated forms were based on one fewer item. However, the average item difficulty parameters were hardly affected, as is illustrated in Table 7.2.1. The raw cut scores at each level, with the exception of the grade 11 Level 2 cut, all decreased by one score point.

Table 7.2.1: Special Equatings

NJSLA–S	Grade	Total Items	Raw Score Range	Average Item Difficulty	L2 Raw Cut Score	L3 Raw Cut Score	L4 Raw Cut Score
CBT	5	49	0–60	0.000	25	39	49
Special Equating	5	48	0–59	0.005	24	38	48
CBT	11	68	0–78	0.000	31	45	60
Online Spanish	11	67	0–77	–0.010	31	44	59

7.3 Subscore Performance Levels

NJDOE policy decisions require that the NJSLA–S report student performance in three content domains (Earth and Space, Life, and Physical) and three scientific practices (Investigating, Sensemaking, and Critiquing). The subscores for these six reporting categories are themselves described in Part 1 of this Technical Report; this section details the processes used to create the NJSLA–S subscore performance level classifications.

NJSLA–S test-takers were labeled as being either “Below,” “Near/Met,” or “Above” expectations in each of the three content domains and the three scientific practices. The process for classifying student performance at the subscore level involved creating separate score tables for each combination of grade and reporting category. During each subscore calibration the item difficulty parameters of each item associated with a given domain or practice were anchored (i.e., held constant), and those values were used to create subscore tables. The subscore performance 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) cut score. Statistical significance was determined based on the conditional standard error of measurement (CSEM) within each subscore table (Smarter Balanced Assessment Consortium, 2017). Each subscore table, then, consists of the raw subscore, its associated theta, the CSEM, a lower and upper bound, and a classification level. The “Lower” and “Upper” columns represent the subscore theta +/– (1.5*CSEM). The subscore score tables for each combination of grade and reporting category, as well as the special equating versions, are presented in Appendix J.

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 2019 NJSLA–S. Part 8.1.1 describes the concept of reliability in the CTT framework, and Part 8.1.2 displays the results.

8.1.1 Reliability and 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. 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:

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

Estimating the size of the measurement error associated with the true score is the key to estimating reliability. 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).

The definitions or assumptions in CTT lead to several important properties. For example, it can be demonstrated that

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2, \qquad \textit{Equation 8.2}$$

or observed score variance (σ^2_x) equals the sum of true score variance (σ^2_t) and error variance (σ^2_e). The relationships among the variance terms (i.e., σ^2_x, σ^2_t, σ^2_e) are critical to a more thorough understanding of important CTT concepts, including reliability and the standard error of measurement. For example, CTT reliability (ρ) is defined as the correlation between observed scores on parallel forms (x_1, x_2), which is equal to true score variance (σ^2_t) divided by observed score variance (σ^2_x):

$$\rho_{x_1 x_2} = \sigma_t^2 \big/ \sigma_x^2 . \qquad \textit{Equation 8.3}$$

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

$$\sigma_e = \sigma_X \sqrt{1 - \rho_{X_1 X_2}}. \qquad \textit{Equation 8.4}$$

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 error.

Separate analyses were performed for each grade level. Scores from all item types were used in the computations. Coefficient alpha was estimated using the following formula:

$$\alpha_{\text{Cronbach}} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_{Y_i}^2}{\sigma_X^2}\right], \qquad \textit{Equation 8.5}$$

where n is the number of items, $\sigma^2_{Y_i}$ (read "sigma squared sub Y sub i") is the variance of item i, and σ^2_X ("sigma squared sub X") is the variance of observed total score, X. SEMs were calculated using the following formula:

$$SEM = S_X \sqrt{1 - \alpha_{\text{Cronbach}}}, \qquad \textit{Equation 8.6}$$

where S_X ("S sub X") is the standard deviation of observed total scores.

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. Reliability measures decrease when the students taking a given test form are more homogeneous in their test performance. Thus, it would be expected that the Spanish and Human Reader forms would have lower reliability coefficients than the CBT forms.

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. The results at grade 5 stand out as particularly exceptional given that the grade 5 test is significantly shorter than either the grade 8 or 11 tests. The grade 5 reliability coefficients ranged from .84 to .92. The most likely reason for the better results at grade 5, despite it being a shorter test, is that the grade 5 items were closer to the ability levels of the grade 5 students, thereby increasing the variance among test scores. At grade 8, where the distribution of test scores was heavily skewed towards the low end of the ability spectrum, reliability ranged from .74 to .91. The relatively low reliability measures for the Spanish, Spanish TTS, and Human Reader forms is due to those populations doing very poorly on the test, which limits the amounts of variance in test scores. The grade 11 alpha coefficients ranged from .80 to .94.

Table 8.1.1: Coefficient Alpha and SEM, by Form

Grade	Form*	N-Count	Mean	SD	Alpha	SEM
5	CBT	81,996	31.81	11.92	0.92	3.46
5	PBT	113	19.81	11.20	0.91	3.34
5	TTS	17,551	23.29	12.19	0.92	3.40
5	SP	825	16.21	8.03	0.84	3.21
5	SP TTS	387	16.33	8.61	0.86	3.24
5	HR	303	20.70	9.57	0.88	3.36
8	CBT	85,985	28.16	13.10	0.91	3.86
8	PBT	92	15.06	8.63	0.86	3.27
8	TTS	12,100	20.07	11.28	0.90	3.56
8	SP	1126	13.06	6.11	0.74	3.12
8	SP TTS	432	14.13	6.57	0.76	3.21
8	HR	84	13.99	6.31	0.75	3.12
11	CBT	84,298	33.38	17.03	0.94	4.02
11	PBT	154	23.93	14.67	0.93	3.84
11	TTS	4,408	25.81	15.15	0.94	3.82
11	SP	761	17.21	8.17	0.82	3.45
11	SP TTS	262	15.74	7.70	0.81	3.34
11	HR	113	17.00	7.72	0.80	3.46

* 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 SEMs of the six reporting categories, by grade. It should be noted that reliability coefficients are commonly low when based upon small numbers of items (Traub & Rowley, 2008). Thus, reporting categories such as Critiquing and Investigating, which had fewer items, tended to have lower reliability measures. The lowest subscore reliability of .68 was for Critiquing at grade 8. The reliability measures in Table 8.1.2 are based on all test takers.

Table 8.1.2: Coefficient Alpha and SEM by Reporting Category

Grade	Reporting Category	Total Items	MC Items	TE1 Items	TE2 Items	CR Items	Max Points	Alpha	SEM
5	Total	50	13	33	1	3	60	.92	3.46
5	Earth and Space	17	7	8	1	1	21	.80	1.93
5	Life	19	3	15	0	1	22	.79	2.21
5	Physical	14	3	10	0	1	17	.82	1.83
5	Sensemaking	23	7	14	0	2	29	.83	2.43
5	Critiquing	14	4	10	0	0	14	.78	1.57
5	Investigating	13	2	9	1	1	17	.76	1.91
8	Total	59	18	36	2	3	70	.92	3.83
8	Earth and Space	17	5	10	1	1	21	.77	2.04
8	Life	20	6	13	0	1	23	.80	2.24
8	Physical	22	7	13	1	1	26	.77	2.34
8	Sensemaking	28	3	23	1	1	32	.84	2.45
8	Critiquing	11	2	7	1	1	15	.68	1.88
8	Investigating	20	13	6	0	1	23	.76	2.27
11	Total	68	25	39	1	3	78	.94	4.00
11	Earth and Space	24	8	15	0	1	27	.86	2.27
11	Life	21	7	12	1	1	25	.83	2.37
11	Physical	23	10	12	0	1	26	.85	2.29
11	Sensemaking	34	14	19	0	1	37	.89	2.70
11	Critiquing	13	3	8	0	2	19	.78	2.20
11	Investigating	21	8	12	1	0	22	.85	1.97

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 .87, for English learner (EL) students, which is still very strong. At grade 8 the coefficient alphas were all very close to .9, excepting the EL students (α_{EL-Yes} = .80). The same pattern was evident at grade 11, where all the coefficient alphas hovered close to .93, except for ELA students (α_{EL-Yes} = .84)

Table 8.1.3: Coefficient Alpha and SEM by Demographic Group

Grade	Group	N	Mean	SD	Alpha	SEM
5	NJSLAS	101,220	30.09	12.46	.92	3.46
5	Male	51,656	30.11	12.91	.93	3.46
5	Female	49,564	30.07	11.97	.92	3.47
5	Am. Indian	133	29.84	12.12	.92	3.51
5	Asian	10,859	38.80	11.04	.91	3.38
5	Black	15,345	22.93	11.14	.91	3.40
5	Hispanic	29,836	25.34	11.32	.91	3.44
5	Pacific Islander	189	32.19	11.97	.92	3.44
5	White	42,442	33.65	11.29	.91	3.44

Grade	Group	N	Mean	SD	Alpha	SEM
5	EL - Yes	5,830	17.50	8.92	.87	3.27
5	EL - No	95,382	30.86	12.23	.92	3.46
5	EconDis - Yes	38,634	24.03	11.14	.91	3.42
5	EconDis - No	62,586	33.83	11.74	.91	3.45
5	SWD - Yes	20,499	22.56	12.14	.92	3.37
5	SWD - No	80,721	32.00	11.80	.91	3.46
8	NJSLAS	99,852	26.92	13.20	.92	3.83
8	Male	51,124	26.77	13.77	.92	3.80
8	Female	48,728	27.07	12.57	.91	3.84
8	Am. Indian	117	24.72	12.16	.91	3.73
8	Asian	10,346	37.43	13.23	.91	3.95
8	Black	14,452	19.47	10.16	.88	3.53
8	Hispanic	28,176	21.24	10.66	.89	3.61
8	Pacific Islander	207	31.96	12.12	.89	3.95
8	White	44,716	30.32	12.59	.90	3.90
8	EL–Yes	4,381	14.55	7.19	.80	3.23
8	EL–No	95,468	27.49	13.13	.91	3.84
8	EconDis–Yes	34,908	20.21	10.19	.88	3.56
8	EconDis–No	64,944	30.52	13.22	.91	3.91
8	SWD–Yes	19,664	19.62	11.12	.90	3.52
8	SWD–No	80,188	28.71	13.05	.91	3.88
11	NJSLAS	90,024	32.78	17.03	.94	4.00
11	Male	45,733	32.35	17.69	.95	3.95
11	Female	44,291	33.22	16.31	.94	4.05
11	Am. Indian	111	32.66	16.50	.94	4.00
11	Asian	9,097	45.16	17.05	.94	4.06
11	Black	12,935	24.00	13.73	.92	3.77
11	Hispanic	23,417	26.63	14.49	.93	3.89
11	Pacific Islander	223	39.29	16.13	.93	4.12
11	White	43,112	36.06	16.62	.94	4.06
11	EL–Yes	3,878	17.18	8.73	.84	3.44
11	EL–No	86,132	33.48	16.98	.94	4.02
11	EconDis–Yes	27,411	25.65	14.34	.93	3.85
11	EconDis–No	62,613	35.90	17.18	.94	4.05
11	SWD–Yes	16,414	23.83	14.72	.94	3.74
11	SWD–No	73,610	34.77	16.87	.94	4.05

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 increases, 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 .9. The alphas associated with each grade levels' CR items were all close to .7, 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	Items	Points	Mean	SD	Alpha	SEM
5	MC	13	13	7.78	2.92	.73	1.52
5	TE	34	35	17.71	7.69	.90	2.49
5	CR	3	12	4.60	2.81	.67	1.60
8	MC	18	18	8.16	3.88	.76	1.89
8	TE	38	40	15.06	7.32	.87	2.64
8	CR	3	12	3.70	3.08	.72	1.64
11	MC	25	25	12.08	5.22	.82	2.21
11	TE	40	41	17.74	9.68	.92	2.65
11	CR	3	12	3.07	3.18	.75	1.59

8.2 Item Response Theory Reliability

The reliability of the scale scores ascertained from the Partial Credit Model (PCM) was assessed in multiple ways. Test information functions (TIFs), item maps, and person fit statistics were evaluated at each grade level. Overall, the 2019 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 (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 specifically at the most important decision points (i.e., the Levels 2–4 cut scores).

The TIF consists of the summation of all the item information functions (IIF; Lord & Novick, 1968; Hambleton, 1989) on a given test. An IIF is the probability of a correct response multiplied by the probability of an incorrect response. Item information functions (I_{ij}) for every item (j) at every level of student ability (i) can be calculated for each item using the following equation:

$$I_{ij}\left(\theta_i, \delta_j\right) = P_{ij} * \left(1 - P_{ij}\right) \qquad \textit{Equation 8.7}$$

The total test information function is simply the sum of all the item information functions. Thus, each item contributes to the TIF, and proper selection of items during the test construction process will lead to TIFs that maximize information at important decision points.

Figures 8.2.1 to 8.2.3 illustrate the TIFs for grades 5, 8, and 11 at person ability estimates ranging from –4 to + 4. More information at a specific ability level implies less measurement error. Ideally, the Level 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 2 and 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 3 cut score, and whether there was a precipitous drop in information at the Level 2 and 4 cut scores. Within each figure there are three lines representing the cut scores.

At grade 5 the TIF peaked in between the Level 2 and 3 cut scores. There was a large drop in information at the Level 4 cut. The 2019 NJSLA–S tests were all constructed without the benefit of knowing where the cut scores would eventually reside. Grade 5 item and test development has since been focused on shifting the TIF slightly to the right in the coming years in order to maximize the consistency of performance level classifications. At grade 8 the TIF peaked almost directly on the Level 3 cut score. Information at the Levels 2 and 4 cut scores was relatively even. Overall, the grade 8 TIF is very close to being ideal. Similarly, the grade 11 TIF peaked almost directly at the Level 3 cut score. However, unlike in grade 8 the information was heavily skewed towards the Level 2 cut score. Overall, the TIFs provide ample evidence that student ability estimates are reliable at the most important decision points. However, both grades 5 and 11 need more information around the Level 4 cut score on future tests.

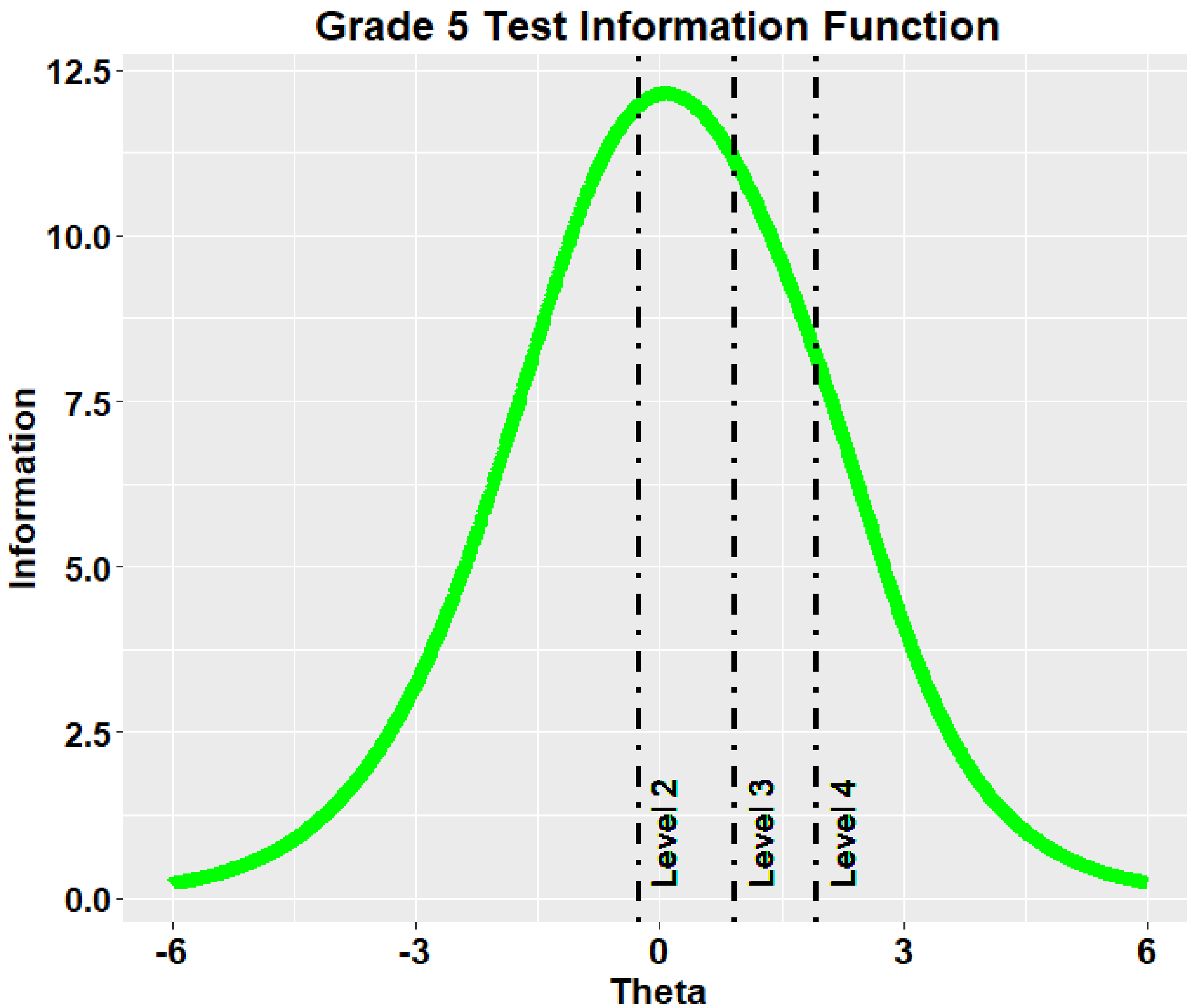


Figure 8.2.1. Grade 5 Test Information Function

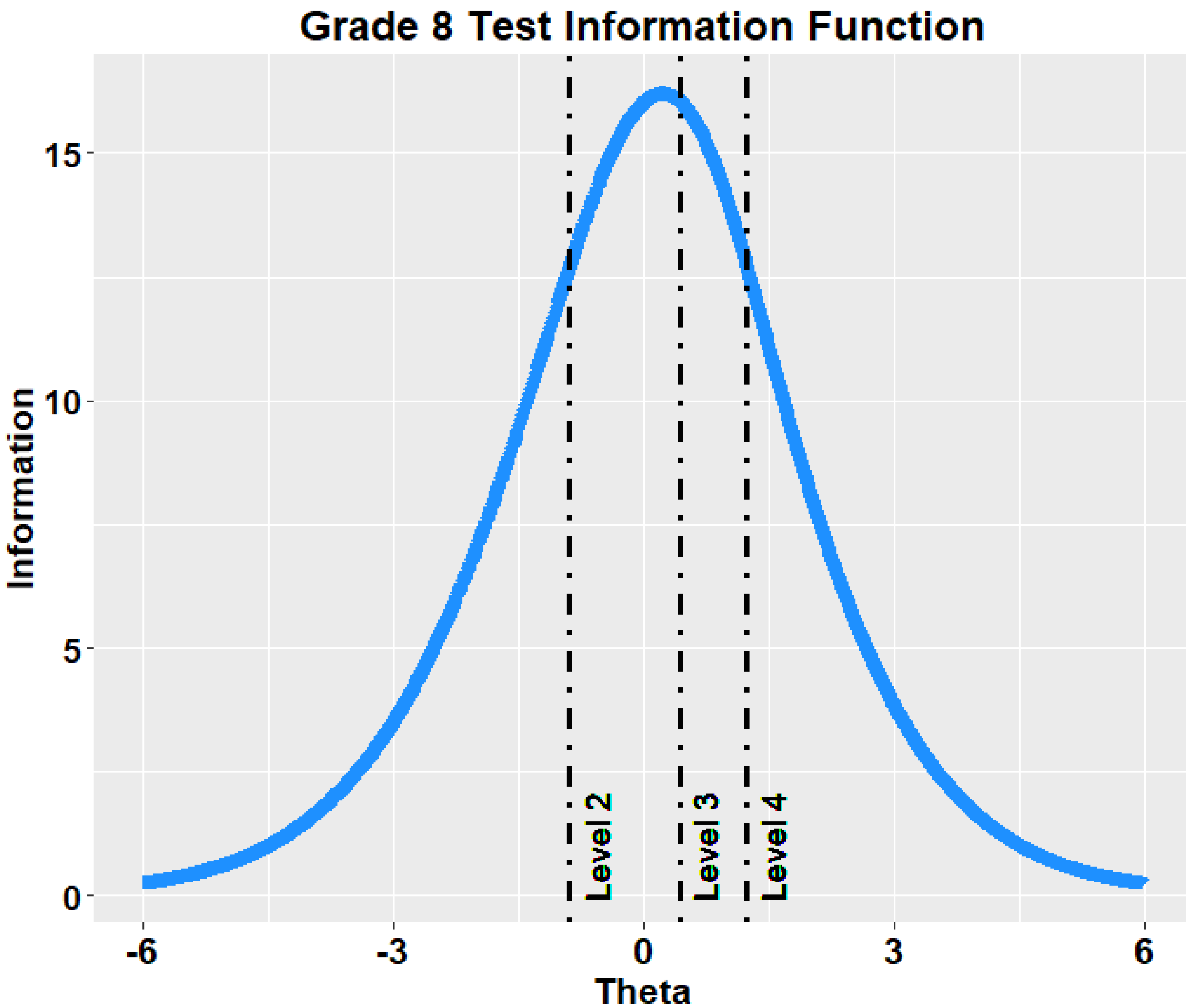


Figure 8.2.2. Grade 8 Test Information Function

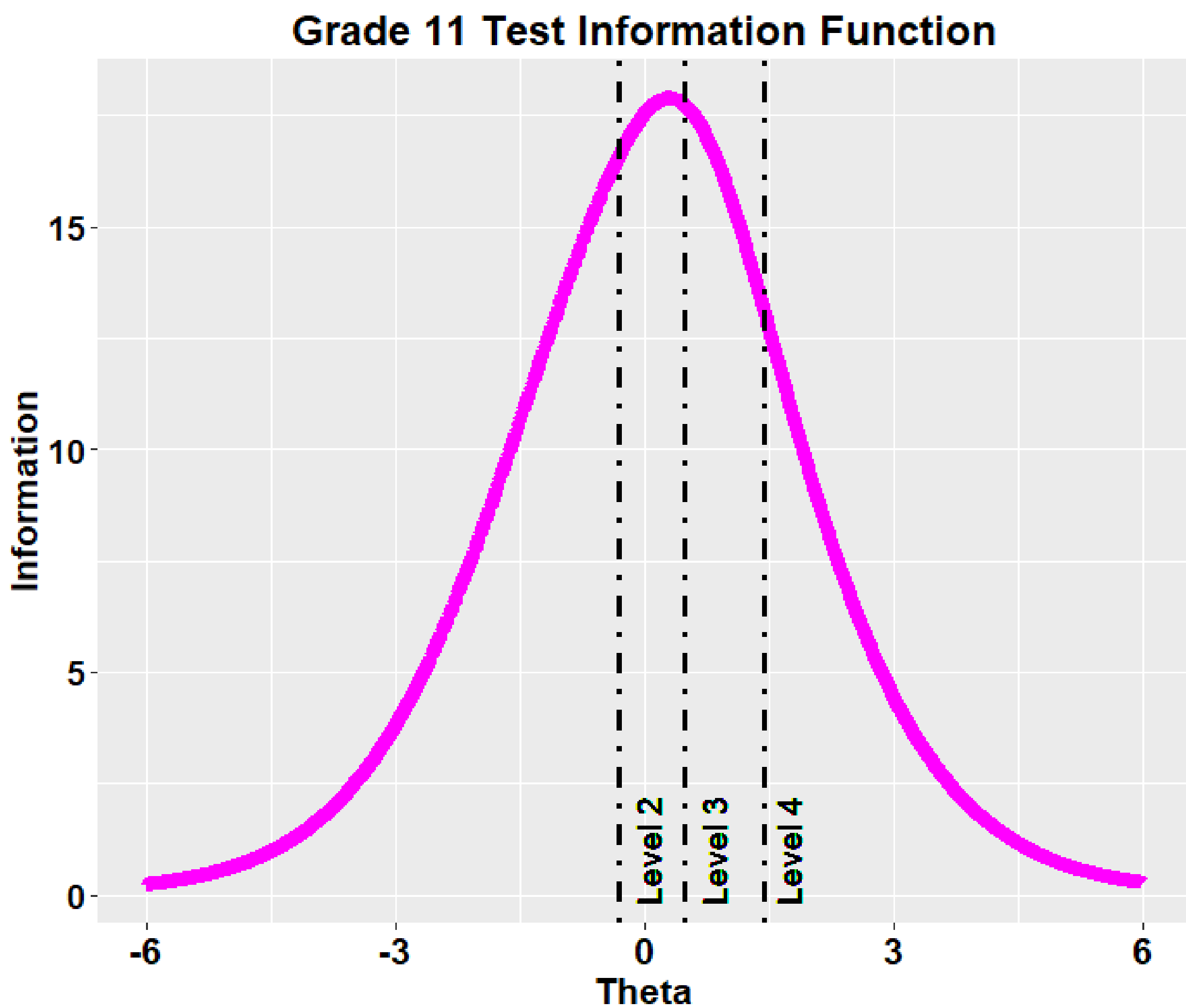


Figure 8.2.3. Grade 11 Test Information Function

8.2.3 Item Maps

Item maps indicate how well the item difficulties and person ability levels match. Items that are targeted to the ability levels of the students taking the test will result in more reliable measures of student ability. Figures 8.2.4 through 8.2.6 show the 2019 NJSLA–S item maps. The grade 5 item difficulty distribution peaks at the Level 2 cut score, while the grade 8 and 11 distributions show item difficulty peaks at or near the Level 3 cut score. At grade 5 the theta distribution was normally distributed with student ability peaking in between the Level 2 and 3 cut scores. The theta distributions at grade 8 and 11 were skewed towards the lower end of the ability spectrum, peaking around the Level 2 cut score. The grade 5 item distribution matched the ability levels of the students better than either grades 8 or 11. The grade 8 item distribution, as the grade 8 TIF showed, matched the decision points on the scale very well. However, there were many students below the Level 2 cut score, and very few items along that part of the scale. The grade 11 item distribution was lacking items at both the upper and lower parts of the scales in comparison to the ability levels of the students.

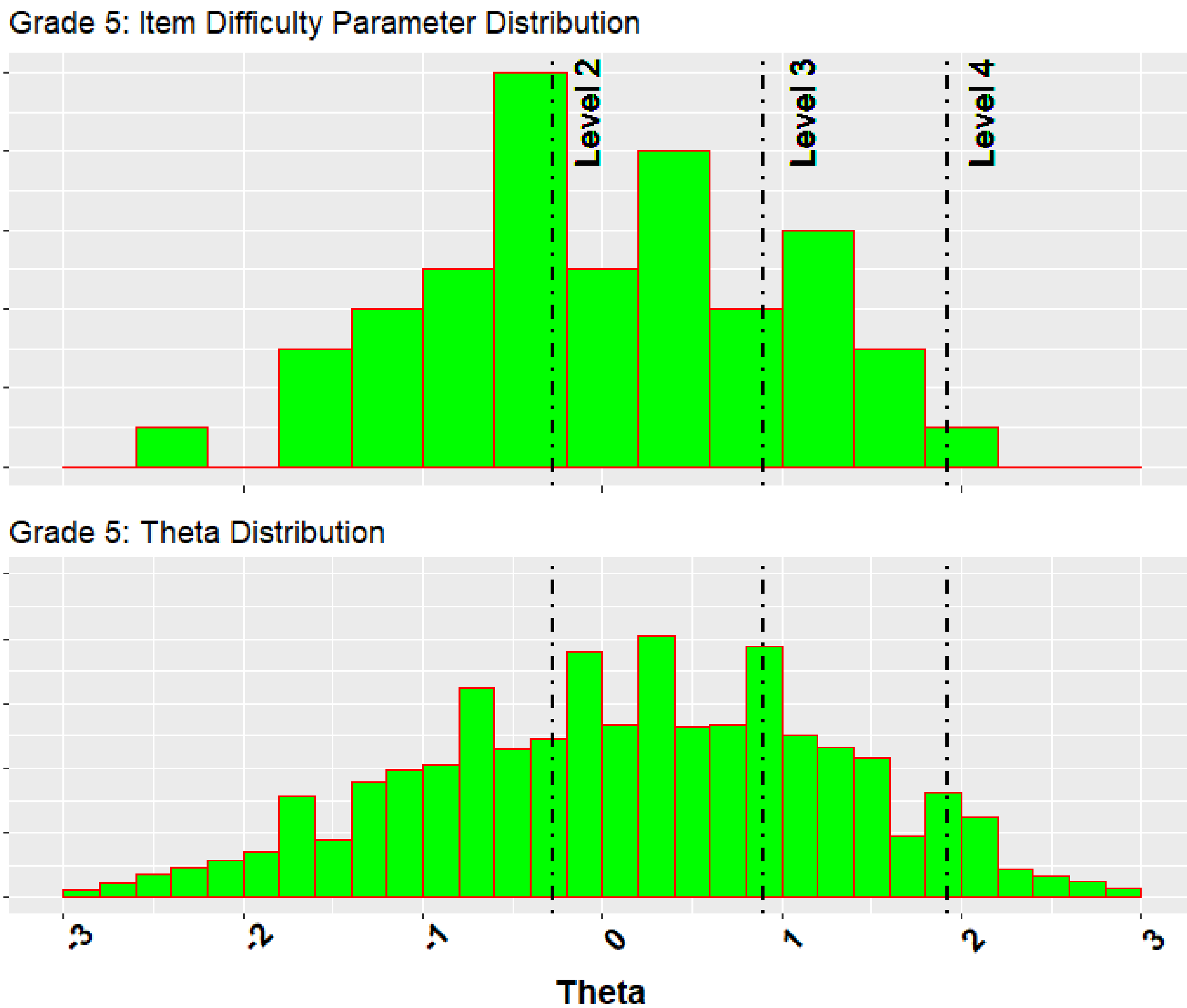


Figure 8.2.4. Grade 5 Item Difficulty and Student Ability Distributions

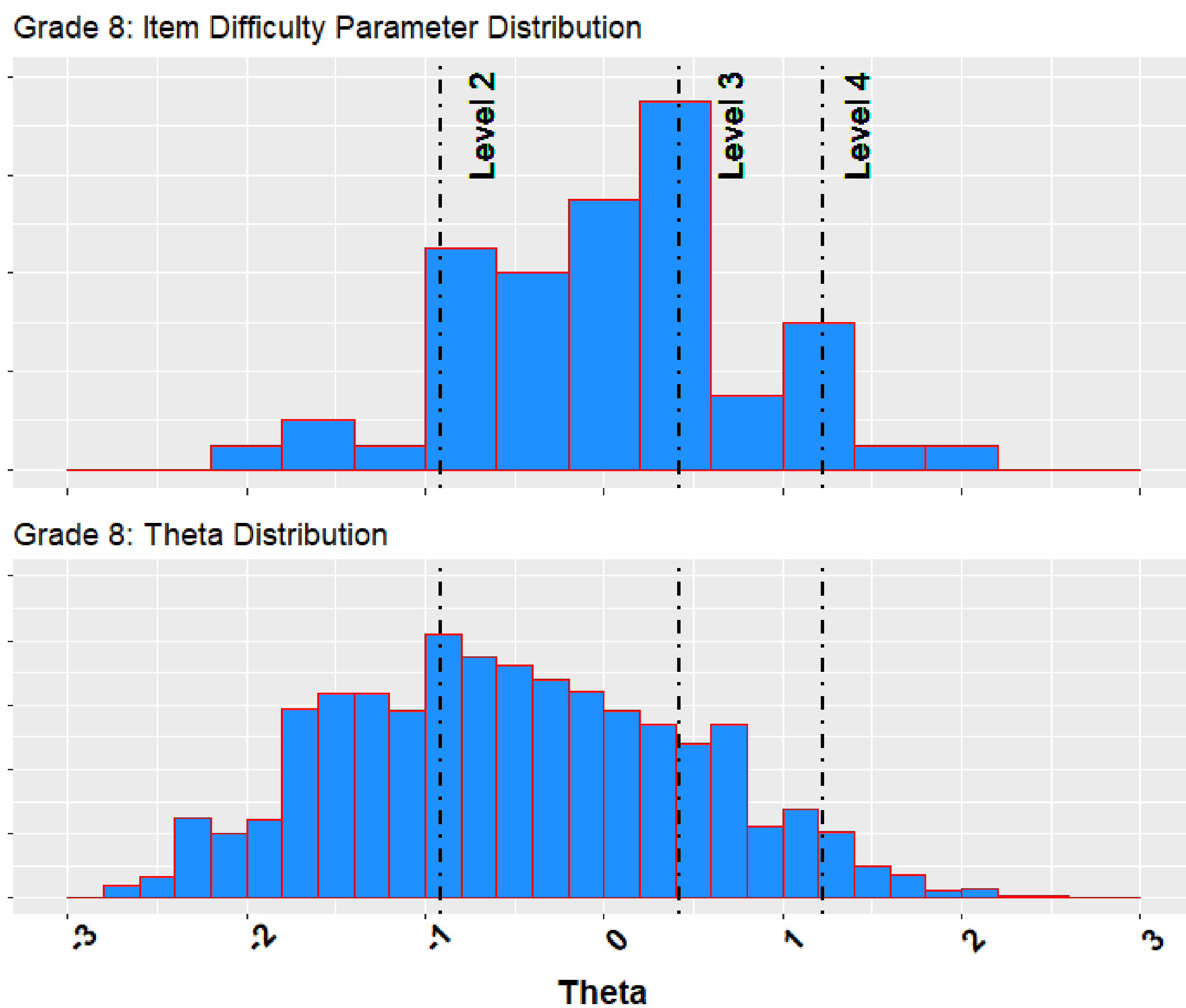


Figure 8.2.5. Grade 8 Item Difficulty and Student Ability Distributions

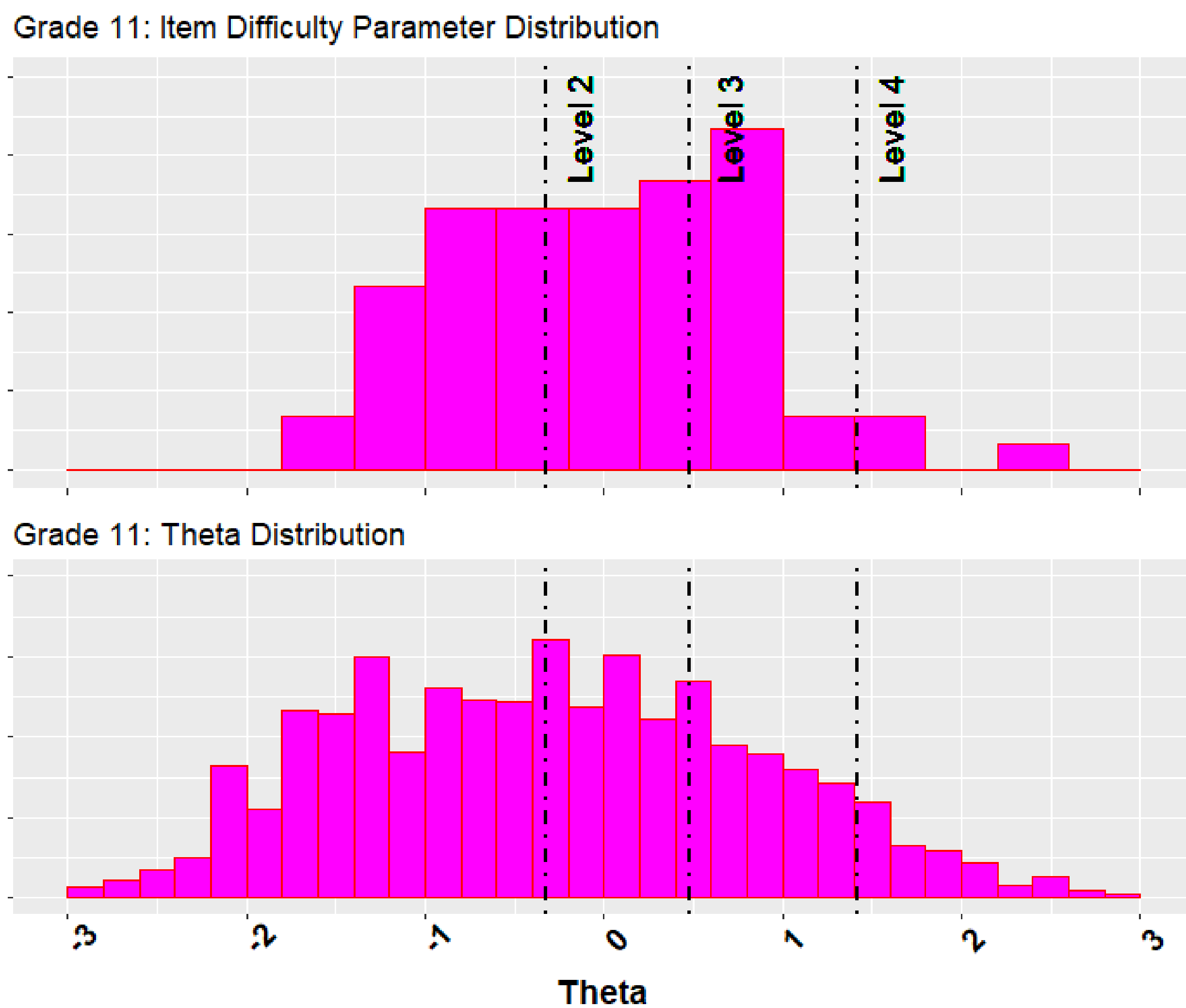


Figure 8.2.6. Grade 11 Item Difficulty and Student Ability Distributions

8.3 Reliability of Performance Classifications

The reliability of the performance 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 performance level classification consistency indices. The results of both methods indicate that the 2019 NJSLA–S performance level classifications were reliable.

8.3.1 Conditional Estimate of Error at Each Cut-Score

WINSTEPS calculates the standard error (SE) at each score point using the information function. The equation for the standard error at each value of theta (ability) is given by:

$$SE\left(\hat{\theta}\right) = \frac{1}{\sqrt{I(\theta)}} \qquad \textit{Equation 8.8}$$

where *I(ϑ)* ("I theta") is the information function for a test at θ (theta).

The 2019 NJSLA–S cut scores and the corresponding conditional standard error of measurement (CSEM) are summarized in Table 8.3.1, and the CSEM tables for all raw and scale scores are presented in Appendix I. The values in Table 8.3.1 have been placed on the same scale as the scale score. Given that the CSEMs are the inverse of the TIF, their interpretations are similar. At grade 5 the Level 3 cut score's CSEM was slightly higher than at Level 2, meaning that there was slightly less error in the scale score at 150 than at 200. At grades 8 and 11 the cut score with the least amount of error was the Level 3 cut score. Error bands 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. Any overlap between the upper or lower bounds and one of the other cut scores could indicate reliability problems among the performance level classifications. For all nine cut scores there was no overlap between either their upper or lower boundaries and another cut score.

Table 8.3.1: Cut Scores with Conditional Standard Error of Measurement

Grade	Level	Cut score	CSEM	Lower Bound	Upper Bound
5	Level 2	150	12.3	125.4	174.6
5	Level 3	200	12.7	174.6	225.4
5	Level 4	243	14.9	213.2	272.8
8	Level 2	150	10.6	128.8	171.2
8	Level 3	200	9.4	181.2	218.8
8	Level 4	231	10.6	209.8	252.2
11	Level 2	158	13.0	132.0	184.0
11	Level 3	200	12.5	175.0	225.0
11	Level 4	250	14.6	220.8	279.2

8.3.2 Classification Consistency Indices

The reliability index for proficiency classifications (kappa) is an estimate of how reliably the test classifies students into the performance categories (i.e. Levels 1–4). Kappa was computed with the BB-CLASS program (Brennan, 2004) based on the beta-binomial model. Coefficient kappa is given by:

$$\kappa = \frac{\varphi - \varphi_c}{1 - \varphi_c}, \qquad \textit{Equation 8.9}$$

where j is the probability of a consistent classification and j_c is the probability of a consistent classification by chance. A classification consistency index can be regarded as the percentage of examinees that would hypothetically be assigned to the same achievement level if the same test was administered a second time or an equivalent test was administered under the same conditions.

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 ranged from .73 to .78. Thus, if the NJSLA–S had been administered a second time, approximately 75% of the students would have been classified at the exact same performance level. The most important decision is at the Level 3 cut score because it demarcates the point along the scale where students are deemed proficient or not. The decision consistency at the Level 3 cut score or above was exceptional at .89 to .92, indicating a 89% to 92% probability of being correctly classified as Level 3 or above. The overall NJSLA–S classification should be interpreted as being consistent.

Table 8.3.2: Performance Level Classification Reliability

Grade	Alpha	SEM	Level 2 Cut	Level 3 Cut	Level 4 Cut	Kappa	[illegible]	Level 3 or above [illegible]
5	.92	3.46	25	39	49	.61	.73	.89
8	.92	3.83	20	40	52	.65	.77	.92
11	.94	4.00	31	45	60	.67	.78	.92

8.4 Reliability of Subscore Performance Classifications

The methodology used to create the subscore performance level classifications was dependent on the CSEMs in each subscore's raw-to-theta subscore tables. Subscores associated with large CSEMs would indicate unreliable subscore performance level classifications and would ensure that most students be classified as "Near/Met Expectations" regardless of their actual ability on the KSAs as measured by the subscores. Table 8.4.1 shows that the CSEMs for each combination of grade level and subscore were relatively small, indicating reliable subscore classifications. The complete raw-to-theta subscore tables, including the ones used for the special equatings, are presented in Appendix J.

Table 8.4.1: Subscore Conditional Standard Errors of Measurement

Grade	Subscore	Level	Raw Score	Theta	CSEM
5	Earth and Space	Near/Met	11	0.37858	0.51245
5	Earth and Space	Above	16	1.77873	0.55484
5	Life	Near/Met	13	0.33391	0.46331
5	Life	Above	19	2.02244	0.65577
5	Physical	Near/Met	8	0.18774	0.53340
5	Physical	Above	14	2.01676	0.64033
5	Investigating	Near/Met	8	0.24907	0.51905
5	Investigating	Above	14	2.02960	0.63913
5	Sensemaking	Near/Met	17	0.37065	0.43294
5	Sensemaking	Above	23	1.66845	0.50546
5	Critiquing	Near/Met	7	0.25610	0.58395
5	Critiquing	Above	12	2.32234	0.79776

Grade	Subscore	Level	Raw Score	Theta	CSEM
8	Earth and Space	Near/Met	9	–0.27876	0.48384
8	Earth and Space	Above	16	1.29852	0.51289
8	Life	Near/Met	11	–0.13403	0.42725
8	Life	Above	18	1.34807	0.52899
8	Physical	Near/Met	11	–0.05404	0.40590
8	Physical	Above	18	1.09606	0.43385
8	Investigating	Near/Met	10	–0.06677	0.41694
8	Investigating	Above	17	1.18512	0.46892
8	Sensemaking	Near/Met	16	–0.07384	0.39357
8	Sensemaking	Above	24	1.18548	0.42339
8	Critiquing	Near/Met	6	–0.11147	0.51864
8	Critiquing	Above	11	1.39714	0.61545
11	Earth and Space	Near/Met	11	–0.08407	0.42276
11	Earth and Space	Above	19	1.24892	0.42018
11	Life	Near/Met	12	–0.07289	0.41203
11	Life	Above	20	1.41625	0.51009
11	Physical	Near/Met	12	–0.10654	0.41165
11	Physical	Above	20	1.41384	0.49651
11	Investigating	Near/Met	10	–0.09436	0.46383
11	Investigating	Above	16	1.28357	0.52016
11	Sensemaking	Near/Met	19	0.06408	0.35483
11	Sensemaking	Above	27	1.09553	0.37543
11	Critiquing	Near/Met	7	–0.12386	0.45773
11	Critiquing	Above	14	1.32870	0.52074

8.5 Rater Reliability

For constructed-response (CR) items, raters used item-specific scoring rubrics with a score range of 0 to 4. 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 a second read was non-adjacent, then the scores for the response were erased and the paper was re-scored. Thus, all scores in the 10% read-behinds were either perfect or adjacent agreement. Table 8.5.1 shows the percentages of constructed-response items scored with exact and adjacent agreement. At the test level, the exact agreement rates ranged from 71.7% to 81.3%. At the item level there was only one item per grade level that had an agreement rate below 75%. Overall, rater agreement on the NJSLA–S 0–4 point CR items was excellent.

Table 8.5.1: Scoring Consistency of Constructed-Response Items

Grade	Item	% Raters in Exact Agreement	% Raters in Adjacent Agreement
NJSLA–S	Total	76.4	23.6
5	Total	71.7	28.3
5	519001_02	77.3	22.7
5	519003_06	75.5	24.5
5	519013_08	62.1	37.9
8	Total	77.9	22.1
8	818002_04	60.8	39.2
8	818003_04	87.2	12.8
8	818015_05	86.3	13.7
11	Total	81.3	18.7
11	HS19003_07	71.9	28.1
11	HS19004_09	82.6	17.4
11	HS19011_07	90.4	9.6

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 on-going, 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 2019 NJSLA–S fosters valid interpretations and uses of test scores as they pertain to the overall performance 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 through use of 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. Part 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 three points of being perfectly balanced. The scientific practices (i.e., Investigating, Sensemaking, and Critiquing) were less balanced. At each grade level the Sensemaking scientific practice was over-represented and Critiquing and Investigating were under-represented. At a more granular level all DCIs, SEPs, and CCCs were represented on each grade level's test, with the exception of grade 5, which was missing three DCIs. 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 over-represented by certain components within the NJSLS–S, which could affect interpretations of test scores at both the overall and subscore level.

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 myriad 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), and at present time the NJSLA–S evidence based on response processes is limited to judgments from the NJSAC and content specialists. Part 9.7.3: Future Validity Studies discusses the prospects for providing evidence based on the response processes of students.

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 Part 5.2.1: Performance Level Descriptors, 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 extremely detailed descriptions of the knowledge, skills, and abilities (KSAs) that a student needs to display in order 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
5	SEP	12	22	16	0
8	DCI	6	27	22	4
8	SEP	10	33	10	6
11	DCI	19	29	13	7
11	SEP	18	34	15	1

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, unexpected patterns of correlations provide 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 among the total test score and its subscores. If the subscores are highly correlated, then that provides evidence that the test is unidimensional. Part 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. The lowest correlation among all subscores at all grade levels was .77 at grade 8 between the Investigating and Critiquing scientific practice categories.

One pattern that was identified within the intercorrelations that could show slight dependencies across the content domains and scientific practices is that certain domains always correlated higher with certain practices. At all three grade levels, Sensemaking displayed a higher correlation with Earth and Space Science than with either Life or Physical Science. At all three grade levels Investigating correlated higher with Physical Science, and Critiquing correlated higher with Life Science. This possible dependency is evidence that the internal structure of the NJSLA–S subscores is not performing as expected.

Table 9.3.1 presents a likely explanation for the unexpected correlational pattern. For Investigating at all three grade levels there were more Physical Science points available than Earth and Space and Life combined. For Sensemaking at each grade level there were more Earth and Space Science points available, and for Critiquing there were more Life Science points available. When the domains and practices are too intertwined, that will lead to those subscores being overly dependent on each other. For example, at grade 5 nine out of 17 Investigating points are dual-aligned to Physical Science. Thus, over 50% of a given student's Investigating points are dependent upon their KSAs in Physical Science. Test score interpretations of subscores that are so highly dependent upon each other could be misleading.

Table 9.3.1: Points Available and Intercorrelations by Domain and Practice

Grade	Practice	Earth	Life	Physical	Earth	Life	Physical
5	Investigating	3	5	9	.82	.84	.91
5	Sensemaking	15	9	5	.92	.89	.86
5	Critiquing	3	8	3	.82	.88	.83
8	Investigating	5	6	12	.82	.84	.90
8	Sensemaking	14	10	8	.91	.89	.85
8	Critiquing	2	7	6	.79	.86	.83
11	Investigating	4	6	12	.87	.89	.93
11	Sensemaking	18	10	9	.95	.91	.91
11	Critiquing	5	9	5	.88	.91	.88

9.3.2 Other Internal Structure Evidence

Evidence of the internal structure of the NJSLA–S was also presented via a principal component analysis (PCA). Its results are presented in Part 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 test scores.

Part 8 of this Technical Report provides ample evidence to support NJSLA–S reliability. Reliability is a measure of internal consistency that provides a sign as to whether the internal structure of the NJSLA–S is unidimensional. The grade level reliability coefficients presented in Part 8.1 were strong, ranging from .92 to .94. At the subscore level the reliability coefficients were relatively impressive, with only grade 8 Critiquing falling below .70.

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. 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 study only included grades 5 and 8, because at grade 11 the NJSLA–S is the only large-scale statewide assessment delivered to all students. The results indicate 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. Moreover, they are very consistent with the results from previous assessments such as NJASK (NJDOE, 2014).

The results at grade 5 are displayed below in Table 9.4.1. The intercorrelation matrix was calculated by correlating students' valid scale scores in ELA, math, and science. ELA consists of two major claims: Reading Complex Text and Writing. The scale scores for those two major claims were added to the matrix. The relationships among science, ELA, and math are consistent with expectations and very similar to results from previous large-scale, statewide assessments such as NJASK which consistently showed correlations among science and ELA, ELA reading, and math of approximately 0.80. The correlation of 0.65 between science and ELA writing was also very consistent with previous results (NJDOE, 2014).

Table 9.4.1: Grade 5 Intercorrelations by Content Area

Content Area	N count	Science	ELA	ELA-R	ELA-W	Math
Science	99,832	1.00	-	-	-	-
ELA	99,832	0.80	1.00	-	-	-
ELA Reading	99,832	0.81	0.95	1.00	-	-
ELA Writing	99,832	0.65	0.87	0.71	1.00	-
Math	99,832	0.83	0.76	0.74	0.64	1.00

The results at grade 8 are displayed below 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, Algebra I, Algebra II, or Geometry. Thus, 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 cells in the correlation matrix are represented by N/A. Overall, the results for science and ELA were very similar to grade 5 and previous New Jersey assessments such as NJASK (NJDOE, 2014). The correlations among science and the myriad math scale scores are lower than both those at grade 5 and the results from NJASK. However, that is most likely due to the higher- and lower-achieving students taking different assessments, which could have the effect of decreasing the scale score variance for each math test. Thus, the magnitude of the correlations among science and the various math tests all appear reasonable when considering that math achievement is more homogeneous within each sub-group than if all students at all ability levels were taking the same assessment.

Table 9.4.2: Grade 8 Intercorrelations by Content Area

Content Area	N count	Science	ELA	ELA-R	ELA-W	Math 8	Alg. I	Alg. II	Geo.
Science	98,184	1.00	-	-	-	-	-	-	-
ELA	98,184	0.78	1.00	-	-	-	-	-	-
ELA-Reading	98,184	0.80	0.94	1.00	-	-	-	-	-
ELA-Writing	98,184	0.65	0.91	0.71	1.00	-	-	-	-
Math 8	62,044	0.72	0.67	0.74	0.64	1.00	-	-	-
Algebra I	32,061	0.78	0.67	0.67	0.55	N/A	1.00	-	-
Algebra II	410	0.73	0.48	0.52	0.31	N/A	N/A	1.00	-
Geometry	3,669	0.67	0.44	0.45	0.33	N/A	N/A	N/A	1.00

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 following 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 exists at the moment. Future NJSLA–S validity studies, including evidence based on consequences are detailed below in Part 9.7.3.

9.6 Other Validity Evidence

Each section within this technical report contributes evidence relevant to validity. The following is a summary of evidence within each section:

Part 1: Introduction — This section 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 section describes the processes used to design and develop the NJSLA–S including:

- the steps taken to link test development to the NJSLA–S' intended inferences and uses
- 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 section 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 section 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 section describes 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 a NJTAC member
- documentation of all results in the NJSLA–S Standard-Setting Report

Part 6: Item and Test Statistics – This section 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 section describes the methods used to ensure all students at a given grade level received scale scores that were comparable including:

- documenting the scaling and special equating procedures
- results of the special equatings

Part 8: Reliability — This section describes the myriad 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
- 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 judgement 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 judgement with such a diverse assortment of evidence is extremely challenging given that the validity process is ongoing and exists through the duration of the testing program. Overall, there is ample evidence that the NJSLA–S will foster 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 re-evaluating their effectiveness is needed. After all, 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 NJSLS–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. Special equatings were performed on test forms that had abnormalities in their test administration. 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 response processes contained limited evidence. Without having a degree of evidence that student responses to test items are indeed measuring what the test is intending to measure, the validity argument is incomplete. After all, even if content experts and the NJSAC say an item is measuring a specific skill, that claim should be verified with evidence from the students who actually have to answer the item. 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. More pressingly, while there is ample validity evidence presented in both Part 2 of this document and in this validity section, the validity evidence would be more conclusive with an alignment study from an outside evaluator. Finally, as noted in Part 5.2.1.4, 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 non-existent. After all, the NJSLA–S is a standards-based assessment, and thus 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). However, there is clearly the need for studying the issues noted above to 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 valid 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. Finally, the subscores displayed adequate reliability coefficients and CSEMs.

The intercorrelations presented in Part 6.2.1.1 and revisited in Part 9.3.1 of this Technical Report show evidence that the proposed interpretations of the subscores should be undertaken with caution; the internal subscore structure displayed dependencies between the content domains and scientific practices that were unintended. At each grade level and for each scientific practice, approximately 50% of score points were dually aligned to the same content domain. To use Critiquing as an example, it would be expected that all Critiquing points be balanced among Earth and Space, Life, and Physical Sciences and that the intercorrelations between Critiquing and each of those three content domains would be relatively similar (because theoretically, Critiquing skills are applicable across all content domains). A possible solution to alleviating these issues involves conscientiously developing a balance of scientific practices across all content domains during the item development process so that the test construction can be similarly balanced.

A similar issue with the subscore intercorrelations is apparent in that the correlations among all the content domains and scientific practices were very high, even when they did not share many dually aligned items. This provides evidence that the dual alignment of items to both the content domains and scientific practices may make the subscores too interdependent and could lead to misinterpretations of subscores. Factor analytic methods could be employed to test the dual-alignment of items and to provide evidence for or against the current dual-alignment structure.

Another issue affecting the validity of subscore interpretations includes the lack of evidence based on response processes. This is especially important with the dually aligned items because it is not known whether the content domain or the scientific practice is driving the difficulty of the item. For example, if an item is dually aligned to the Earth and Space Science content domain and the Sensemaking scientific practice, but the item is predominantly measuring KSAs associated with Sensemaking while the Earth and Space Science KSAs are secondary, then reporting that item with the Earth and Space Science subscore could be misleading.

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 on 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 in both 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 Part 2.4, on-going, 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. There are numerous future validity studies that are planned; they are detailed in the sections below.

9.7.3.1 Alignment Study. NJDOE is scheduled to conduct an independent alignment study with external vendor, edCount. The study seeks to verify the alignment of the NJSLA–S to NJSLS–S by answering the three following alignment questions (edCount, 2020; p.3):

1. How well does the blueprint represent the multi-dimensional standards?
2. How well do the set of items on a test form match the blueprint?
3. How well do the set of items on a test form reflect the full range of performance described in the Performance Level Descriptors?

The evidence the alignment study intends to collect in order to answer those questions will provide extremely important information pertaining to validity evidence based on test content and will complement the evidence presented in Part 9.1 of this report. Moreover, as part of answering alignment question #2, edCount’s independent panelists will review the cognitive complexity of the phenomenon-based stimuli (PBS) and their items providing limited validity evidence based on response processes.

9.7.3.2 Cognitive Labs. The current validity evidence based on response processes grounded in the Range PLDs and the response processes evidence to be collected in the alignment study are by nature limited forms of evidence because they are dependent on the expert judgment of adults, not students. NJDOE plans to conduct a cognitive lab study with students to provide stronger validity evidence based on response processes. Given present uncertainties associated with the impact of the COVID-19 situation upon school operations, the timing of that study is unknown.

9.7.3.3 Confirmatory Factor Analysis. The validity evidence based on internal structure is comprehensive and decisive as it pertains to the unidimensionality of the NJSLA–S. However, to confirm the existence of the theoretical internal structure of the subscores requires a confirmatory factor analysis (CFA), as opposed to the principal components analysis used for monitoring the unidimensionality. CFA is a powerful tool for providing validity evidence that the internal structure of a construct fits the theoretical model (Brown, 2006). The theory behind the NJSLA–S asserts that the DCIs and SEPs can be grouped into, respectively, three content domains and three categories of scientific practices. A CFA can provide insight into whether the DCI and SEP groupings are justified. As noted in Part 9.7.2., another use for a CFA includes evaluating the interdependency among the three content domains and scientific practices. Finally, a CFA could be used to evaluate measurement invariance across test forms and sub-groups, with particular attention being paid to the PBT, Spanish, and Spanish TTS forms, as well as sub-groups that had larger percentages of students flagged for person infit and outfit, such as English learner (EL) students or students with disabilities.

9.7.3.4 Accommodated Test Form Equivalence. Aside from the CFA suggested above, more evidence pertaining to the equivalence of the myriad accommodated test forms could be acquired by a more detailed review of the IRT results. It is known that some forms had a disproportionately large number of students flagged for person infit and outfit statistics (see Part 6.2.2: Partial Credit Model Fit Statistics). What is not currently known is which items were causing the misfit for those groups. By increasing the depth of the measurement invariance analysis for the item difficulty parameters, the items causing the misfit could be identified and assessed for patterns to determine whether certain characteristics of the items were more likely to lead to certain forms having more students flagged for person infit and outfit. To the extent that information is gleaned from the deeper analysis, then in the spirit of iteratively improving the NJSLA–S, that information could be incorporated into the item development processes to ensure the validity of the NJSLA–S test score inferences.

9.7.3.5 Consequences of the NJSLA–S. Two of the goals of the NJSLA–S are to influence adoption of the NJSLS–S curriculum and inform instruction, which will in turn improve the educational opportunities for New Jersey students. As described in Part 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. NJDOE is committed to evaluating the effects of the NJSLA–S; given present uncertainties associated with the impact of the COVID-19 situation upon school operations, the timing of that study is unknown.

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 minimize possible misinterpretations of test scores, and they were accompanied by the NJSLA–S Score Interpretation Guide (SIG). This section describes how the scale scores, performance level classifications, and subscores were presented to New Jersey stakeholders, including students, parents, teachers, administrators, and districts. There were five different reports produced for dissemination to the public. A sample score report for each of the five reports is explained and presented below. More comprehensive descriptions of each component within the myriad reports can be found in the NJSLA–S Score Interpretation Guide: measinc-nj-science.com.

10.1 Individual Student Report

The Individual Student Report (ISR) is a two-sided document intended for use by students, parents, teachers, and administrators. It contains the student scale score; the Reporting PLD associated with the student's performance; data for comparison across the state, district and school; subscore performance; and descriptions of subscore expectations. Figures 10.1.1 and 10.1.2 show an example of the information housed within the ISRs.

State of New Jersey
Department of Education

FIRSTNAME M. LASTNAME
Spring 2019 Grade: 5
SID: **0123456789** DOB: **01/01/9999**
Local Student Identification: **0123456789**
SAMPLE DISTRICT NAME
SAMPLE SCHOOL NAME

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

This report shows how FIRSTNAME performed on the [elementary/middle/high] school science assessment. **This assessment is just one measure of how well your child is performing academically.**

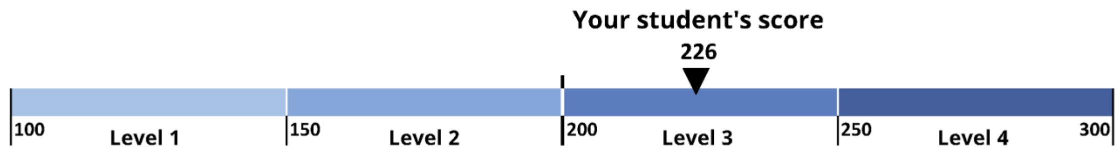
To learn more about the test and to view sample questions and practice tests, visit the Score Interpretation Guide (SIG) at www.measinc.com/nj/science.

How Can You Use This Report?
Ask your child's teachers:
- What do you see as my child's academic strengths and areas for improvement?
- How will you use these test results to help my child make progress this school year?

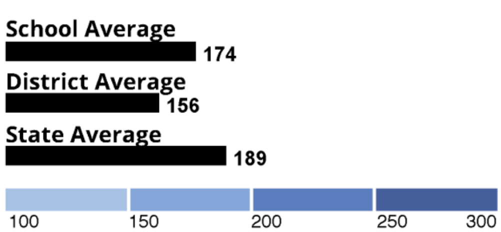
See side 2 of this report for specific information on your student's performance using domains and practices.

How did **FIRSTNAME** perform on the NJSLA-S?
Your student's score: **226**
Performance: **Level 3**

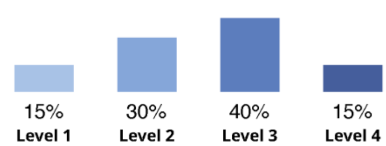
Level 1 (100 – 149)
Level 2 (150 – 199)
Level 3 (200 – 249)
Level 4 (250 – 300)



FIRSTNAME'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.



How Students Statewide Performed



Percentage of students at each performance level

Page 1 of 2 mmddccyy-Z9999999-999999-999-9999999

Figure 10.1.1. Sample Individual Student Report – Page 1

FIRSTNAME M. LASTNAME

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.

✓ 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.

✓ Life Science

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

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

! 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.

✓ Physical Science

Your student's performance is **Above 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.

≈ 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 his/her academic performance. The results also give your school and school district important information to make improvements to the education program and to teaching.

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/aps/cccs/science/.

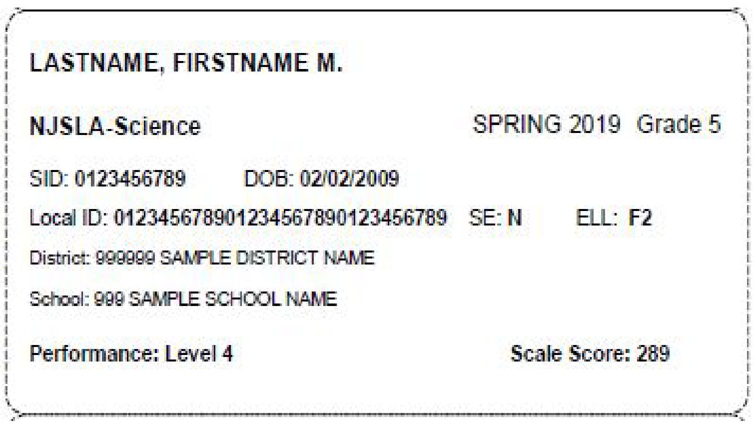
Districts may assign Not Tested or Void codes for students that did not receive a scale score. For more information see the Score Interpretation Guide at www.measinc.com/nj/science.

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 assist teachers, schools, and administrators with the identification of student assessment records. 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 2019 Grade 5

SID: 0123456789 DOB: 02/02/2009

Local ID: 012345678901234567890123456789 SE: N ELL: F2

District: 999999 SAMPLE DISTRICT NAME

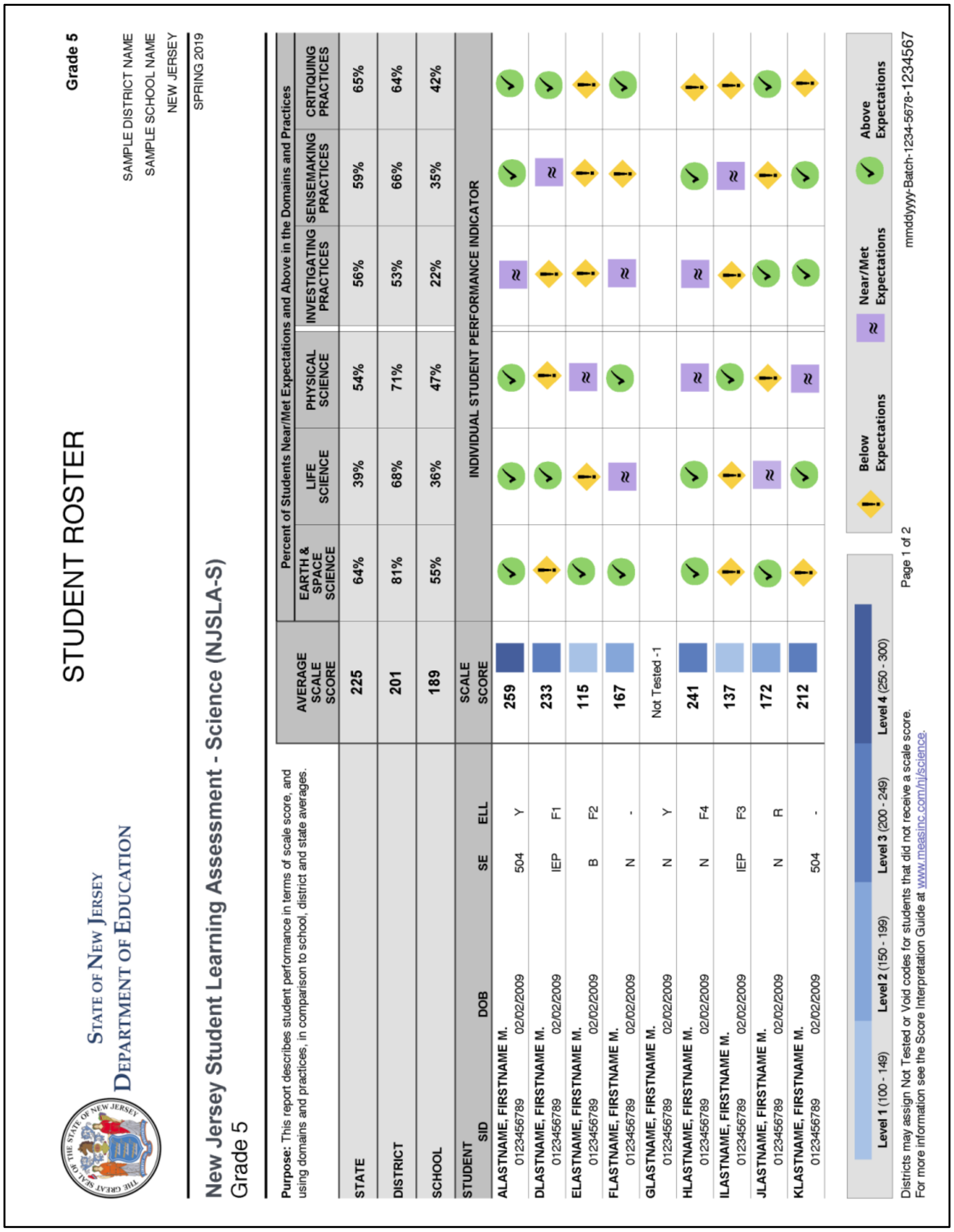
School: 999 SAMPLE SCHOOL NAME

Performance: Level 4 Scale Score: 289

Figure 10.2.1. Sample Student Label

10.3 Student Roster

The NJSLA–S Student Roster reports display student test results within a school. They are used by teachers, schools, and administrators. The Student Roster contains state, district, and school level average scale scores and percentages of subscores that are Near/Met Expectations or above. It also contains student information, including names, Special Education status, EL status, scale score, performance level classifications for the overall scale and the subscores. Figure10.3.1 shows a sample Student Roster.



STUDENT ROSTER

State of New Jersey
Department of Education

Grade 5

SAMPLE DISTRICT NAME
SAMPLE SCHOOL NAME
NEW JERSEY
SPRING 2019

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.

	AVERAGE SCALE SCORE	EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
		Percent of Students Near/Met Expectations and Above in the Domains and Practices					
STATE	225	64%	39%	54%	56%	59%	65%
DISTRICT	201	81%	68%	71%	53%	66%	64%
SCHOOL	189	55%	36%	47%	22%	35%	42%

STUDENT / SID	DOB	SE	ELL	SCALE SCORE	EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
					INDIVIDUAL STUDENT PERFORMANCE INDICATOR					
ALASTNAME, FIRSTNAME M. 0123456789	02/02/2009	504	Y	259	✓	✓	✓	≈	✓	✓
DLASTNAME, FIRSTNAME M. 0123456789	02/02/2009	IEP	F1	233	!	✓	!	!	≈	✓
ELASTNAME, FIRSTNAME M. 0123456789	02/02/2009	B	F2	115	✓	!	≈	!	!	!
FLASTNAME, FIRSTNAME M. 0123456789	02/02/2009	N	-	167	✓	≈	✓	≈	!	✓
GLASTNAME, FIRSTNAME M. 0123456789	02/02/2009	N	Y	Not Tested - 1						
HLASTNAME, FIRSTNAME M. 0123456789	02/02/2009	N	F4	241	✓	✓	≈	≈	✓	!
ILASTNAME, FIRSTNAME M. 0123456789	02/02/2009	IEP	F3	137	!	!	✓	!	≈	!
JLASTNAME, FIRSTNAME M. 0123456789	02/02/2009	N	R	172	✓	≈	!	✓	!	✓
KLASTNAME, FIRSTNAME M. 0123456789	02/02/2009	504	-	212	!	✓	≈	✓	✓	!

Level 1 (100 - 149) Level 2 (150 - 199) Level 3 (200 - 249) Level 4 (250 - 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.

Page 1 of 2

mmddyyyy-Batch-1234-5678-1234567

Figure 10.3.1. Sample Student Roster

10.4 School Performance Level Summary

The NJSLA–S School Performance Level Summary reports display aggregate student performance at the state, district, and school level. Other aggregations include gender, ethnicity, disability status, and EL status. Aggregate student performance is displayed via average scale score, the percentages of students in each performance level classification, and breakdowns of the subscore performance levels. Figures 10.4.1 and 10.4.2 display examples of the School 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 2019

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

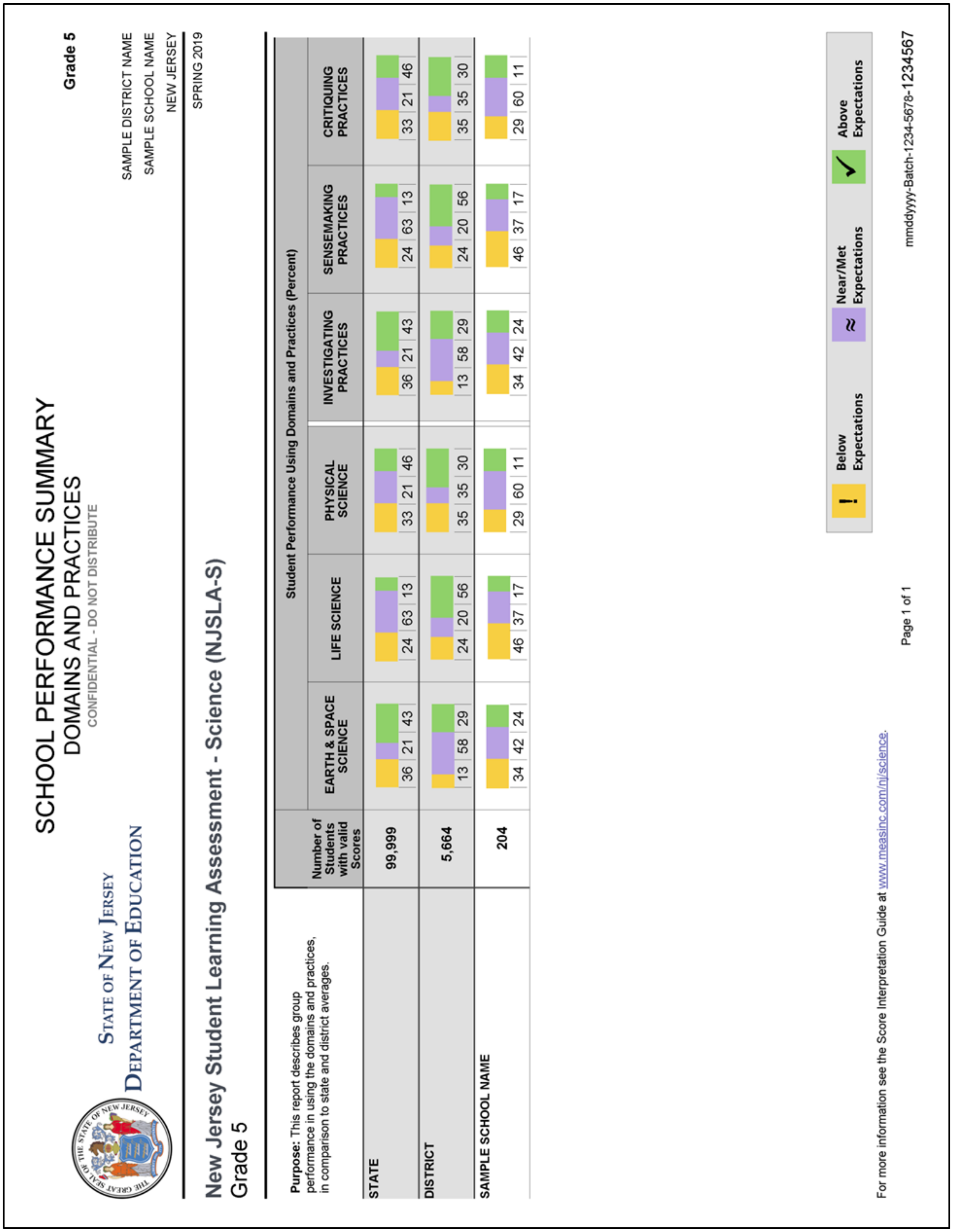
Purpose: This report describes group achievement in terms of average scale scores and performance levels.	Total Number of Student Records	No Scores Reported	Number of Students with Valid Scores	Average Scale Score	Performance Levels								≥ Level 3	
					Level 1		Level 2		Level 3		Level 4			
					#	%	#	%	#	%	#	%	#	%
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%
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%
English Language Learner														
Current EL	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 EL	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 10.4.1. Sample School Performance Level Summary Report



SCHOOL PERFORMANCE SUMMARY
DOMAINS AND PRACTICES
CONFIDENTIAL - DO NOT DISTRIBUTE

State of New Jersey
Department of Education

Grade 5

SAMPLE DISTRICT NAME
SAMPLE SCHOOL NAME
NEW JERSEY
SPRING 2019

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.

	Number of Students with valid Scores	Student Performance Using Domains and Practices (Percent)					
		EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
STATE	99,999	36 21 43	24 63 13	33 21 46	36 21 43	24 63 13	33 21 46
DISTRICT	5,664	13 58 29	24 20 56	35 35 30	13 58 29	24 20 56	35 35 30
SAMPLE SCHOOL NAME	204	34 42 24	46 37 17	29 60 11	34 42 24	46 37 17	29 60 11

! 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 School Performance Level Summary Report – Domains and Practices

10.5 District Performance Level Summary

The NJSLA–S District Performance Level Summary reports display aggregate student performance at the state, district, and school level. Other district-level aggregations include gender, ethnicity, disability status, and EL status. District-level aggregate student performance is displayed via average scale score, the percentages of students in each performance level classification, and breakdowns of the subscore performance levels. Figures 10.5.1 and 10.5.2 display examples of the District Performance Level Summary reports.

DISTRICT PERFORMANCE LEVEL SUMMARY

Grade 5

STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

CONFIDENTIAL - DO NOT DISTRIBUTE

SAMPLE DISTRICT NAME

NEW JERSEY
SPRING 2019

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

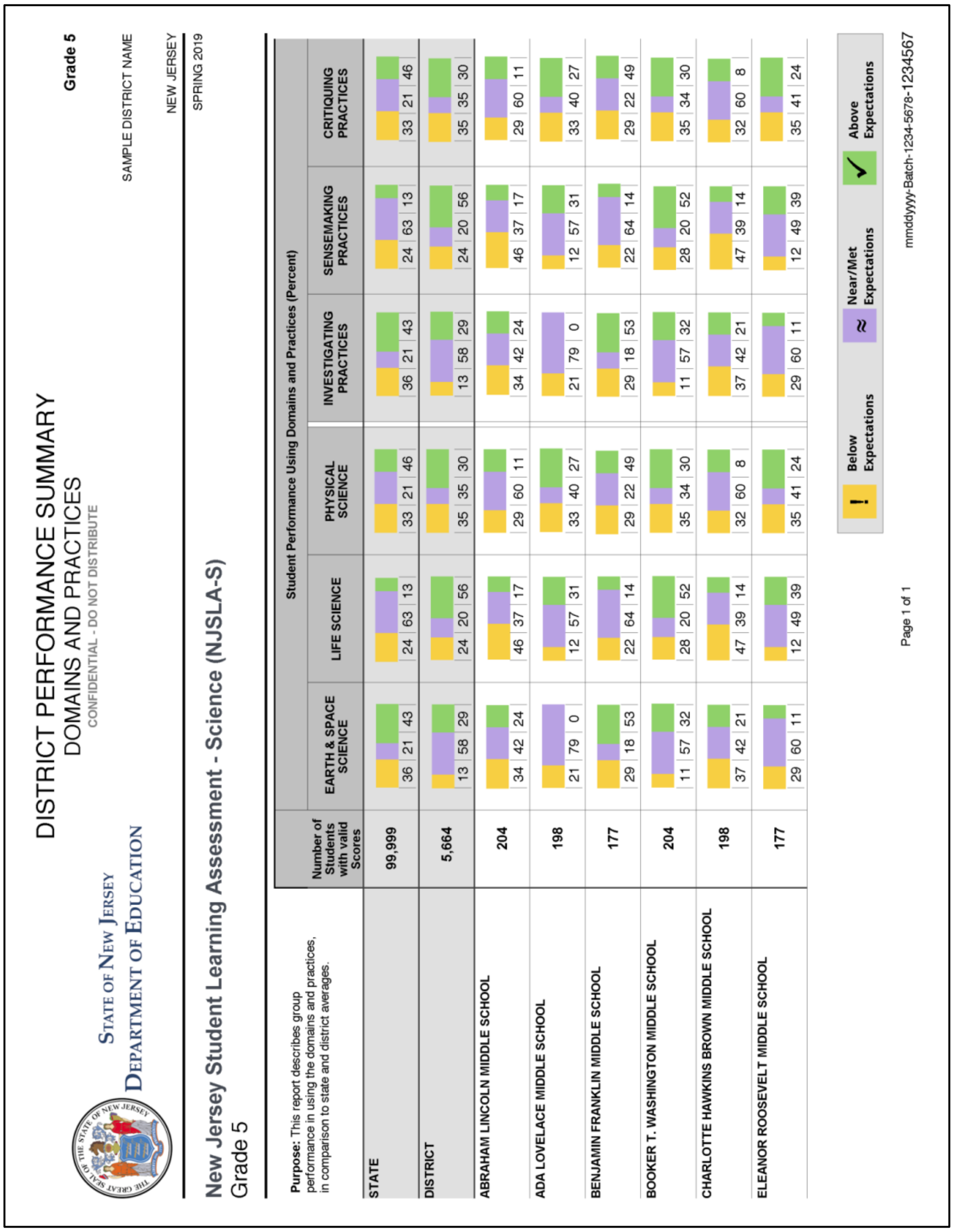
Purpose: This report describes group achievement in terms of average scale scores and performance levels.	Total Number of Student Records	No Scores Reported	Number of Students with Valid Scores	Average Scale Score	Performance Levels								≥ Level 3	
					Level 1		Level 2		Level 3		Level 4			
					#	%	#	%	#	%	#	%	#	%
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%
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%
English Language Learner														
Current EL	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 EL	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 District Performance Level Summary Report



STATE OF NEW JERSEY
DEPARTMENT OF EDUCATION

DISTRICT PERFORMANCE SUMMARY
DOMAINS AND PRACTICES
CONFIDENTIAL - DO NOT DISTRIBUTE

Grade 5
SAMPLE DISTRICT NAME
NEW JERSEY
SPRING 2019

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.

Student Performance Using Domains and Practices (Percent)

	Number of Students with valid Scores	EARTH & SPACE SCIENCE	LIFE SCIENCE	PHYSICAL SCIENCE	INVESTIGATING PRACTICES	SENSEMAKING PRACTICES	CRITIQUING PRACTICES
STATE	99,999	36 21 43	24 63 13	33 21 46	36 21 43	24 63 13	33 21 46
DISTRICT	5,664	13 58 29	24 20 56	35 35 30	13 58 29	24 20 56	35 35 30
ABRAHAM LINCOLN MIDDLE SCHOOL	204	34 42 24	46 37 17	29 60 11	34 42 24	46 37 17	29 60 11
ADA LOVELACE MIDDLE SCHOOL	198	21 79 0	12 57 31	33 40 27	21 79 0	12 57 31	33 40 27
BENJAMIN FRANKLIN MIDDLE SCHOOL	177	29 18 53	22 64 14	29 22 49	29 18 53	22 64 14	29 22 49
BOOKER T. WASHINGTON MIDDLE SCHOOL	204	11 57 32	28 20 52	35 34 30	11 57 32	28 20 52	35 34 30
CHARLOTTE HAWKINS BROWN MIDDLE SCHOOL	198	37 42 21	47 39 14	32 60 8	37 42 21	47 39 14	32 60 8
ELEANOR ROOSEVELT MIDDLE SCHOOL	177	29 60 11	12 49 39	35 41 24	29 60 11	12 49 39	35 41 24

! Below Expectations ≈ Near/Met Expectations ✓ Above Expectations

Page 1 of 1

mmddyyyy-Batch-1234-5678-1234567

Figure 10.5.2. Sample District Performance Level Summary Report – Domains and Practices

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.

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.

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

Cizek, G.J. and Bunch, M.B. (2007). *Standard Setting*. Thousand Oaks, CA: 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.

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.

Cronbach, L.J., Schönemann, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational and Psychological Measurement, 25*, 291-312.

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

edCount. (2020). A proposal to conduct and independent evaluation of the alignment quality of the New Jersey Student Learning Assessment in Science. Alexandria, VA: edCount.

Educational Testing Service. (2015). *ETS guidelines for fair tests and communications.* Princeton, NJ: Author.

Egan, K. L., (2001). *Validity and defensibility of cutscores established by the Bookmark standard setting method.* Paper presented at the 2001 Council of Chief State School Officers Conference on Large-Scale Assessment, Houston.

Egan, K. L., Schneider, M. C., and Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Eds.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 79-106). New York, NY: Routledge.

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

Gorsuch, R. L. (1983). Factor Analysis. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Hambleton, R. K., Pitoniak, M. J., Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 79-106). New York, NY: Routledge.

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

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

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

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). Mahwah, N.J.: Erlbaum

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

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

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

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

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

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

Lewis, D.M., Mitzel, H.C., & Green, D. R. (1996, June). Standard setting: A Bookmark approach. symposium presented at the Council of the Chief State Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D.M., Mitzel, H.C., Mercado, R.L., and Shultz, E.M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Eds.), Setting Performance Standards: Foundations, Methods, and Innovations (pp. 225-254). New York, NY: Routledge.

Lin, J. (2004). The Bookmark Standard Setting Procedure: Strengths and Weaknesses. The Centre for Research in Applied Measurement and Evaluation. The University of Alberta. Alberta, Canada.

Linacre, J. M. (2012). A *User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Programs*. Chicago, IL.

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*. Reading MA: 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). New York, NY: American Council on Education.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts and core ideas.* Washington, DC: National Academies Press.

New Jersey Department of Education (2014). NJASK Technical Report. Trenton, NJ.

New Jersey Department of Education (2017). NJASK Technical Report. Trenton, NJ.

New Jersey Department of Education (2019). 2019 statewide assessment district test coordinator and district technology coordinator training [PowerPoint slides]. Trenton, NJ.

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). New York, NY: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

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

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

Schneider M. C. and Egan K. L. (2014). A Handbook for Creating Range and Target Performance Level Descriptors. Retrieved from: https://www.nciea.org/sites/default/files/publications/Handbook_091914.pdf

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.* Boston, MA: Massachusetts Department of Elementary and Secondary Education.

Sireci, S.G., Hauger, J., Lewis, C., Wells, C, Zenisky, A.L. and Delton, J. (2009). Evaluation of the Standard Setting on the 2005 Grade 12 National Assessment of Educational Progress Mathematics Test. In Buckendahl, C.W., et. al., *Evaluation of the National Assessment of Educational Progress*. Washington, D.C.: U.S. Department of Education.

Smarter Balanced Assessment Consortium (2017). Smarter Balanced Assessment Consortium: 2016-17 Technical Report. Retrieved from: https://portal.smarterbalanced.org/library/en/2016-17-summative-assessment-technical-report.pdf

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal Design Applied to Large Scale Assessments. Synthesis Report.

Traub, R. E. and Rowley, G. L. (2008). Understanding reliability. Instructional topics in educational measurement. Madison, WI: National Council on Measurement and Education 176-177.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

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*, Chicago: MESA Press.

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

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

APPENDIX A: Glossary of Abbreviations

Table A.1: Glossary of NJSLA–S Abbreviations

Abbreviations	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-Scoreable Assessment
NBP	National Braille Press
NCME	National Council on Measurement in Education
NJASK	New Jersey Assessment of Skills and Knowledge
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

Abbreviations	Definition
NJSLS–S	New Jersey Student Learning Standards – Science
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
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	School County
1	Cranford Public School District	Union
2	Spring Lake School District	Monmouth
3	Brick Township Public School District	Ocean
4	Newark Public Schools	Essex
5	Stillwater Township School District	Sussex
6	Brick Public Schools	Ocean
7	Metuchen Public School District	Middlesex
8	Swedesboro-Woolwich School District	Gloucester
9	Rumson Borough School District	Monmouth
10	Readington Township School District	Hunterdon
11	Collingswood Public Schools	Camden
12	Burlington Township School District	Burlington
13	Avalon School District	Cape May County
14	Lawrence Township Public School District	Mercer

Table B.2 Grade 8 NJSAC District and County Representation

Number	District	School County
1	Newark Public Schools	Essex
2	Cinnaminson Township School District	Burlington
3	Piscataway Township School District	Middlesex
4	Moorestown Township	Burlington
5	Cherry Hill School District	Camden
6	West Windsor-Plainsboro Regional School District	Mercer
7	Tabernacle Township School District	Burlington
8	Brick Public Schools	Ocean
9	Berkeley Heights School District	Union
10	East Brunswick Township School District	Middlesex
11	Lakewood School District	Ocean
12	City of Orange Township School District	Essex
13	Paterson Public School District	Passaic
14	Knowlton Township School District	Warren

Table B.3 Grade 11 NJSAC District and County Representation

Number	District	School County
1	Cranford Public School District	Union
2	Pascack Valley Regional High School District	Bergen
3	Piscataway Township School District	Middlesex
4	Elizabeth	Union
5	Paramus Public	Bergen
6	Union City School District	Hudson
7	Parsippany-Troy Hills Township	Morris
8	Lenape Regional High School District	Burlington
9	Jersey City Public Schools	Hudson
10	Cherry Hill School District	Camden
11	Toms River Regional School District	Ocean
12	Passaic City School District	Passaic
13	Greater Egg Harbor Regional High School District	Atlantic
14	North Hunterdon-Voorhees Regional High School District	Hunterdon
15	Pennsville Public Schools	Salem

Table B.4 NJBSC District and County Representation

Number	District	School County
1	Roxbury School District	Morris
2	East Brunswick Township School District	Middlesex
3	Northern Burlington County School District	Burlington
4	Roselle Park School District	Union
5	Red Bank Borough School District	Monmouth
6	Egg Harbor City School District	Atlantic
7	Jersey City Global Charter	Hudson
8	Mantua Township School District	Gloucester
9	Unity Charter	Morris

APPENDIX C: Statistical Review Reference Sheet

P-VALUE **(1-point items)**	Proportion correct. P-value is a measure of item difficulty for dichotomous items. P-values can range from 0 to 1.
****FLAGGED IF***	***P < .25 (too hard)*** ***P > .90 (too easy)***

ITEM MEAN **(2-/4-point items)**	Average item score. Item mean is a measure of item difficulty for multi-point items.

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***

RASCH VALUE **(all items)**	A measure of item difficulty based on item response theory (IRT) with possible values ranging from negative infinity to positive infinity. Higher values indicate greater difficulty (reverse of p-value).

ITEM-TOTAL CORRELATION	Correlation between this item and the total test score. This correlation measures 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. The RPB can range from –1 to +1.
****FLAGGED IF:***	***RPB < .20 (1-point items)*** ***RPB < .25 (2-point items*** ***RPB < .30 (4-point items)***

DIF CATEGORY	Differential item functioning (DIF) categorization is a means of detecting potential item bias. DIF looks at the extent to which an item performs differently with different groups — in this case Male/Female, White/Black, White/Hispanic, and White/Asian — controlling for the groups' ability (again, 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

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 report (Table 1.2). 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, we analyzed data available as of 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.

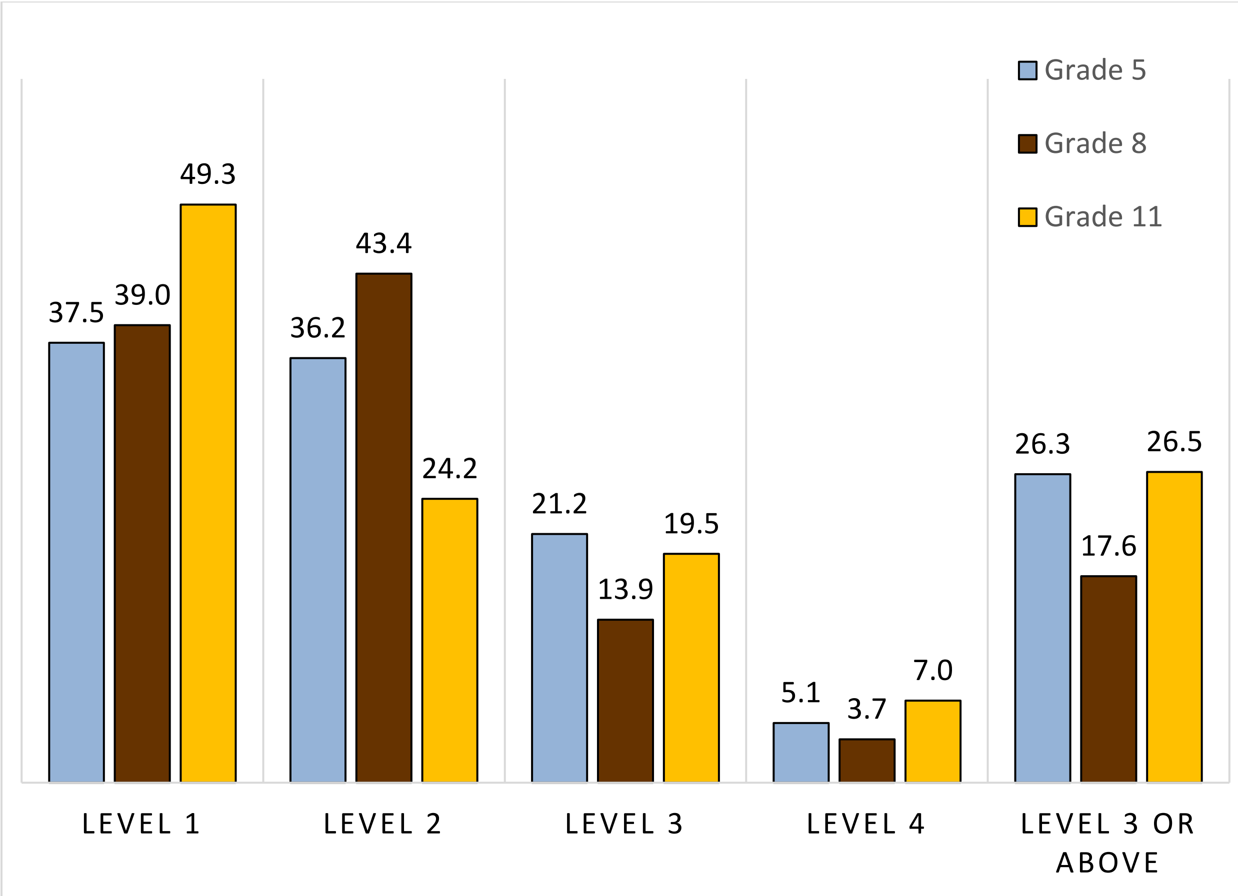


Figure ES-1. Percentages of students classified at each level after Round 3

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. In particular, we wished to know 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 impact 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. We therefore recommend that the cut score recommendations presented here be given strong consideration for approval.

APPENDIX E: NJSLA–S Performance Level Descriptors

E.1 Policy PLDs

DRAFT

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 doesn't 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 depends 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 the 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 the 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 non-renewable 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 the 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, determining 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 is 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 non-renewable 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 non-living 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 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 the 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 the 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 the 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 the 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 feedbacks between the Earth's systems and biosphere cause 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 mange 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-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-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-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 choose 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.

E.3.4 Reporting PLDs – Level 4

Students who are at Level 4 demonstrate advanced understanding of the New Jersey Student Learning Standards-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 non-essential 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

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Level
518043_01	1	TE	OECI	ESS2	PAT	Earth and Space	Critiquing	2	DCI = B1; SEP = B3
518043_03	1	MC	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = B1; SEP = B3
518043_05	1	MC	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = B1; SEP = B3
518008_01	1	MC	PACI	PS2	C and E	Physical	Investigating	2	DCI = B2; SEP = D2
518008_02	1	TE	CEDS	PS2	C and E	Physical	Sensemaking	2	DCI = B2; SEP = B1
518008_06	1	TE	EAE	PS2	C and E	Physical	Critiquing	2	DCI = B2; SEP = E2
518010_01	1	MC	AID	PS1	SC	Physical	Sensemaking	2	DCI = B2; SEP = B3
518010_03	1	TE	AID	PS1	PAT	Physical	Sensemaking	2	DCI = B2; SEP = B3
518010_05	1	TE	PACI	PS1	PAT	Physical	Investigating	2	DCI = B2; SEP = B3
518054_01	1	TE	DUM	ESS1	PAT	Earth and Space	Sensemaking	2	DCI = B2; SEP = D2
518054_03	1	MC	DUM	ESS1	PAT	Earth and Space	Sensemaking	2	DCI = B1; SEP = D2
518054_06	1	MC	EAE	ESS1	C and E	Earth and Space	Critiquing	2	DCI = B2; SEP = E2
519003_01a	1	TE	AQDP	LS1	SF	Life	Investigating	2	DCI = B1; SEP = A1
519003_02a	1	TE	PACI	LS1	PAT	Life	Investigating	2	DCI = B1; SEP = A1
519003_04a	1	TE	UMCT	LS1	S, P, and Q	Life	Investigating	2	DCI = A1; SEP = D1
519003_05a	1	TE	CEDS	LS1	C and E	Life	Sensemaking	2	DCI = B1; SEP = B2
519003_06a	4	CR	CEDS	LS1	SF	Life	Sensemaking	2	DCI = B3; SEP = B3
518011_01	1	TE	AID	ESS2	S, P, and Q	Earth and Space	Sensemaking	2	DCI = A2; SEP = A1
518011_04	1	MC	AID	ESS2	S, P, and Q	Earth and Space	Sensemaking	2	DCI = A1; SEP = A1
518011_06	1	MC	UMCT	ESS2	S, P, and Q	Earth and Space	Investigating	2	DCI = A1; SEP = B2
518011_09	1	TE	EAE	ESS2	C and E	Earth and Space	Critiquing	2	DCI = A1; SEP = D2
518059_02	1	TE	AID	LS4	PAT	Life	Sensemaking	2	DCI = B1; SEP = A1
518059_04	1	TE	AID	LS4	S, P, and Q	Life	Sensemaking	2	DCI = B1; SEP = A2
518059_05	1	TE	EAE	LS4	SC	Life	Critiquing	2	DCI = B3; SEP = C2
518035_01	1	TE	EAE	LS1	E&M	Life	Critiquing	2	DCI = B2; SEP = D3

Table F.1: Grade 5 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Level
518035_04	1	MC	EAE	LS1	E&M	Life	Critiquing	2	DCI = B2; SEP = B2
518060_01	1	MC	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = C2; SEP = A1
518060_02	1	TE	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = C1; SEP = A1
518060_03	1	TE	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = C3; SEP = A1
519001_01a	1	TE	AQDP	PS3	E&M	Physical	Investigating	2	DCI = A2; SEP = C2
519001_02a	4	CR	PACI	PS3	E&M	Physical	Investigating	2	DCI = A4; SEP = D3
519001_07b	1	TE	AID	PS3	E&M	Physical	Sensemaking	2	DCI = A3; SEP = E2
519001_08b	1	TE	EAE	PS3	E&M	Physical	Critiquing	2	DCI = B2; SEP = D2
519001_10b	1	TE	PACI	PS3	E&M	Physical	Investigating	2	DCI = A3; SEP = D3
519000_01a	1	TE	AQDP	LS2	C and E	Life	Investigating	2	DCI = A2; SEP = C2
519000_05a	1	TE	UMCT	LS2	S, P, and Q	Life	Investigating	2	DCI = B1; SEP = C3
519000_07a	1	TE	EAE	LS2	C and E	Life	Critiquing	2	DCI = B2; SEP = D2
518004_01	1	TE	CEDS	PS3	E&M	Physical	Sensemaking	2	DCI = B1; SEP = A2
518004_03	1	MC	EAE	PS3	E&M	Physical	Critiquing	2	DCI = B1; SEP = C2
518004_05	1	TE	PACI	PS3	SC	Physical	Investigating	2	DCI = B1; SEP = D3
518000_01	1	TE	DUM	LS2	S & SM	Life	Sensemaking	2	DCI = B2; SEP = C3
518000_03	1	TE	OECI	LS2	S & SM	Life	Critiquing	2	DCI = C3; SEP = C2
518000_06	1	MC	CEDS	LS2	C and E	Life	Sensemaking	2	DCI = F2; SEP = B3
518012_07	1	TE	EAE	LS4	SF	Life	Critiquing	2	DCI = C2; SEP = A2
518012_04	1	MC	EAE	LS4	SF	Life	Critiquing	2	DCI = A2; SEP = C2
518012_02	1	TE	EAE	LS4	SF	Life	Critiquing	2	DCI = A2; SEP = C2
519013_02b	1	TE	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = B3; SEP = B1
519013_04b	1	TE	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = A2; SEP = B1
519013_06b	2	TE	UMCT	ESS2	S & SM	Earth and Space	Investigating	2	DCI = B1; SEP = B3
519013_08b	4	CR	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = B3; SEP = C3

Table F.2: Grade 5 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
518043_01	1	TE	–0.53023	0.63	0.54	174
518043_03	1	MC	–1.31270	0.76	0.40	79
518043_05	1	MC	–1.34630	0.77	0.49	54
518008_01	1	MC	–1.53790	0.79	0.32	67
518008_02	1	TE	–0.95565	0.70	0.45	32
518008_06	1	TE	1.70235	0.23	0.36	86
518010_01	1	MC	–0.96460	0.70	0.55	94
518010_03	1	TE	–0.31411	0.59	0.59	103
518010_05	1	TE	0.41557	0.45	0.59	73
518054_01	1	TE	1.02603	0.34	0.53	225
518054_03	1	MC	1.32364	0.28	0.28	29
518054_06	1	MC	0.62693	0.41	0.32	69
519003_01a	1	TE	–0.60579	0.64	0.39	71
519003_02a	1	TE	0.18245	0.50	0.45	45
519003_04a	1	TE	–0.23129	0.57	0.29	68
519003_05a	1	TE	–1.66820	0.81	0.34	40
519003_06a	4	CR	–0.43043	2.51	0.60	549
518011_01	1	TE	–2.50290	0.90	0.44	66
518011_04	1	MC	–1.48750	0.78	0.61	60
518011_06	1	MC	–0.22986	0.57	0.32	43
518011_09	1	TE	–0.50623	0.63	0.39	82
518059_02	1	TE	–0.52643	0.63	0.36	89
518059_04	1	TE	1.88462	0.20	0.34	90
518059_05	1	TE	1.45832	0.26	0.41	78
518035_01	1	TE	1.23432	0.30	0.32	127
518035_04	1	MC	0.61445	0.41	0.30	49
518060_01	1	MC	–0.39235	0.60	0.31	90
518060_02	1	TE	1.25079	0.30	0.31	78
518060_03	1	TE	0.85600	0.37	0.31	97
519001_01a	1	TE	–0.09625	0.55	0.48	78
519001_02a	4	CR	1.26325	1.10	0.65	402

Table F.2: Grade 5 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
519001_07b	1	TE	–1.27660	0.75	0.40	40
519001_08b	1	TE	1.21804	0.30	0.53	64
519001_10b	1	TE	0.56024	0.42	0.60	53
519000_01a	1	TE	0.52987	0.43	0.40	112
519000_05a	1	TE	–0.69901	0.66	0.39	56
519000_07a	1	TE	–1.15990	0.74	0.50	37
518004_01	1	TE	0.17780	0.50	0.41	77
518004_03	1	MC	–0.11391	0.55	0.48	44
518004_05	1	TE	0.52075	0.43	0.37	32
518000_01	1	TE	0.33558	0.46	0.23	77
518000_03	1	TE	0.44543	0.44	0.44	37
518000_06	1	MC	0.48982	0.43	0.37	72
518012_07	1	TE	0.12630	0.50	0.47	117
518012_04	1	MC	–0.97657	0.71	0.49	46
518012_02	1	TE	–0.51762	0.63	0.57	46
519013_02b	1	TE	0.22098	0.49	0.44	111
519013_04b	1	TE	–-0.42386	0.61	0.54	60
519013_06b	2	TE	0.90368	0.75	0.54	150
519013_08b	4	CR	1.43906	0.99	0.56	641

Table F.3: Grade 8 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Level
818004_01b	1	TE	AID	PS3	E&M	Physical	Sensemaking	3	DCI = C1; SEP = A2
818004_02a	1	TE	EAE	PS3	E&M	Physical	Critiquing	3	DCI = D2; SEP = C1
818004_03b	1	TE	AID	PS3	E&M	Physical	Sensemaking	3	DCI = D2; SEP = A2
818004_04a	1	TE	CEDS	PS3	S & SM	Physical	Sensemaking	3	DCI = C3; SEP = C3
818083	1	MC	AQDP	PS2	S, P, and Q	Physical	Investigating	2	DCI = B2; SEP = B1
818251	1	MC	AQDP	PS1	S, P, and Q	Physical	Investigating	2	DCI = B2; SEP = A2
818077	1	TE	EAE	ESS2	PAT	Earth and Space	Critiquing	3	DCI = A3; SEP = C1

Table F.3: Grade 8 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Levels
818079_01	1	TE	AID	ESS2	E&M	Earth and Space	Sensemaking	2	DCI = B2; SEP = B2
818181	1	TE	AID	ESS3	C and E	Earth and Space	Sensemaking	2	DCI = B2; SEP = A2
818307_01	1	TE	CEDS	ESS3	SF	Earth and Space	Sensemaking	2	DCI = A3; SEP = A2
818336_01	1	MC	AID	ESS3	C and E	Earth and Space	Sensemaking	2	DCI = A1; SEP = G3
818283	1	MC	EAE	PS2	C and E	Physical	Critiquing	2	DCI = B2; SEP = C1
818288	1	MC	PACI	PS3	E&M	Physical	Investigating	2	DCI = E3; SEP = D2
818314	1	MC	OECI	PS4	SF	Physical	Critiquing	2	DCI = A3; SEP = A1
818333	1	TE	PACI	PS3	E&M	Physical	Investigating	2	DCI = A2; SEP = B4
818033_02	1	MC	AQDP	PS4	E&M	Physical	Investigating	3	DCI = B2; SEP = A2
818002_01b	1	TE	DUM	LS2	C and E	Life	Sensemaking	3	DCI = B4; SEP = E2
818002_02a	1	TE	CEDS	LS2	E&M	Life	Sensemaking	3	DCI = B3; SEP = B3
818002_04a	4	CR	EAE	LS2	SC	Life	Critiquing	2	DCI = B4; SEP = C4
818055_02	1	TE	DUM	LS2	E&M	Life	Sensemaking	2	DCI = A2; SEP = E2
818055_01	1	TE	CEDS	LS2	C and E	Life	Sensemaking	2	DCI = C3; SEP = B2
818055_03	1	TE	DUM	LS2	SC	Life	Sensemaking	3	DCI = B3; SEP = E2
818067_01	1	MC	PACI	LS1	SF	Life	Investigating	2	DCI = C3; SEP = A3
818067_02	1	TE	EAE	LS1	SF	Life	Critiquing	2	DCI = C2; SEP = C3
818309	1	TE	DUM	ESS3	S, P, and Q	Earth and Space	Sensemaking	2	DCI = B3; SEP = E2
818095_01	1	MC	CEDS	ESS3	C and E	Earth and Space	Sensemaking	3	DCI = A2; SEP = B2
818300_01	1	TE	UMCT	ESS1	S, P, and Q	Earth and Space	Investigating	2	DCI = B2; SEP = C2
818306_01	1	MC	AQDP	ESS3	SF	Earth and Space	Investigating	2	DCI = A2; SEP = A1
818064	1	MC	AQDP	LS1	SF	Life	Investigating	2	DCI = A2; SEP = E2
818351	1	TE	UMCT	LS2	E&M	Life	Investigating	2	DCI = A4; SEP = D2
818114_01	1	TE	AID	LS2	E&M	Life	Sensemaking	3	DCI = D1; SEP = D3
818302	1	MC	AQDP	ESS1	S, P, and Q	Earth and Space	Investigating	2	DCI = A2; SEP = C2
818334	1	TE	AID	ESS2	SC	Earth and Space	Sensemaking	2	DCI = B3; SEP = D4
818267	1	MC	AQDP	ESS2	S & SM	Earth and Space	Investigating	3	DCI = A3; SEP = B2

Table F.3: Grade 8 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Levels
818271	1	TE	EAE	ESS2	PAT	Earth and Space	Critiquing	2	DCI = A3; SEP = C3
818003_02a	1	TE	AID	PS3	PAT	Physical	Sensemaking	2	DCI = A3; SEP = A2
818003_01a	1	TE	DUM	PS3	E&M	Physical	Sensemaking	2	DCI = B3; SEP = D2
818003_03a	2	TE	EAE	PS3	E&M	Physical	Critiquing	3	DCI = A3; SEP = B3
818003_04a	4	CR	UMCT	PS3	S, P, and Q	Physical	Investigating	4	DCI = B4; SEP = C4
818109	1	TE	AID	LS4	C and E	Life	Sensemaking	2	DCI = B2; SEP = D2
818296_02	1	TE	DUM	LS3	PAT	Life	Sensemaking	2	DCI = A2; SEP = E2
818065	1	TE	EAE	LS1	SF	Life	Critiquing	2	DCI = C3; SEP = C3
818062	1	MC	PACI	LS1	E&M	Life	Investigating	2	DCI = A1; SEP = A3
818244	1	TE	AID	PS4	PAT	Physical	Sensemaking	2	DCI = A3; SEP = A2
818250	1	TE	PACI	PS4	SF	Physical	Investigating	2	DCI = B2; SEP = B4
818285	1	TE	UMCT	PS2	S & SM	Physical	Investigating	3	DCI = B2; SEP = D2
818089_01	1	MC	AQDP	PS2	C and E	Physical	Investigating	2	DCI = A1; SEP = A1
818041_02	1	TE	CEDS	LS4	PAT	Life	Sensemaking	3	DCI = B3; SEP = B2
818041_03a	1	MC	CEDS	LS4	PAT	Life	Sensemaking	2	DCI = A2; SEP = B2
818041_04b	1	TE	EAE	LS4	C and E	Life	Critiquing	2	DCI = A2; SEP = C2
818197_03	1	MC	UMCT	LS3	S, P, and Q	Life	Investigating	2	DCI = A2; SEP = D2
818197_02	1	MC	UMCT	LS3	C and E	Life	Investigating	2	DCI = C2; SEP = D2
818028	1	TE	AID	PS1	SF	Physical	Sensemaking	2	DCI = C2; SEP = D1
818082	1	TE	DUM	PS1	S & SM	Physical	Sensemaking	2	DCI = I2; SEP = E1
818165	1	TE	OECI	PS4	SC	Physical	Critiquing	2	DCI = A3; SEP = B1
818015_01a	1	TE	DUM	ESS1	S & SM	Earth and Space	Sensemaking	3	DCI = A2; SEP = E2
818015_02a	1	TE	UMCT	ESS1	S, P, and Q	Earth and Space	Investigating	2	DCI = B1; SEP = D2
818015_03a	2	TE	DUM	ESS1	C and E	Earth and Space	Sensemaking	3	DCI = A3; SEP = E2
818015_05b	4	CR	CEDS	ESS1	S & SM	Earth and Space	Sensemaking	3	DCI = C3; SEP = B4

Table F.4: Grade 8 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
818004_01b	1	TE	–1.95470	0.77	0.40	98
818004_02a	1	TE	0.58734	0.29	0.30	95
818004_03b	1	TE	–0.85213	0.57	0.19	74
818004_04a	1	TE	2.05769	0.10	0.33	80
818083	1	MC	–0.24646	0.44	0.13	81
818251	1	MC	0.21866	0.35	0.35	79
818077	1	TE	–0.53246	0.50	0.30	100
818079_01	1	TE	–0.73936	0.55	0.29	66
818181	1	TE	–0.95953	0.59	0.52	77
818307_01	1	TE	–0.94620	0.59	0.42	85
818336_01	1	MC	–0.83361	0.57	0.56	104
818283	1	MC	–0.81809	0.56	0.55	55
818288	1	MC	0.05498	0.38	0.21	57
818314	1	MC	0.25410	0.34	0.25	90
818333	1	TE	–0.05846	0.41	0.47	68
818033_02	1	MC	0.40108	0.32	0.47	60
818002_01b	1	TE	0.54821	0.29	0.34	128
818002_02a	1	TE	1.16574	0.20	0.30	44
818002_04a	4	CR	0.01609	1.54	0.67	361
818055_02	1	TE	–3.11420	0.90	0.35	60
818055_01	1	TE	–0.18681	0.43	0.47	95
818055_03	1	TE	–0.35735	0.47	0.50	59
818067_01	1	MC	0.59223	0.28	0.12	95
818067_02	1	TE	1.48503	0.16	0.42	79
818309	1	TE	–0.30320	0.46	0.50	70
818095_01	1	MC	–1.77250	0.74	0.49	49
818300_01	1	TE	1.16099	0.20	0.25	105
818306_01	1	MC	–1.03570	0.61	0.47	59
818064	1	MC	–0.61878	0.52	0.49	56
818351	1	TE	0.10521	0.37	0.45	72
818114_01	1	TE	0.08253	0.38	0.49	62

Table F.4: Grade 8 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
818302	1	MC	0.12636	0.37	0.39	77
818334	1	TE	1.13709	0.20	0.28	80
818267	1	MC	0.24607	0.35	0.40	53
818271	1	TE	0.44726	0.31	0.43	64
818003_02a	1	TE	–0.17777	0.43	0.35	82
818003_01a	1	TE	–0.06862	0.41	0.27	44
818003_03a	2	TE	0.98992	0.47	0.34	94
818003_04a	4	CR	0.58747	0.95	0.60	286
818109	1	TE	–1.62630	0.72	0.44	67
818296_02	1	TE	–0.23481	0.44	0.50	98
818065	1	TE	0.89336	0.24	0.45	53
818062	1	MC	–0.00256	0.39	0.25	80
818244	1	TE	–0.84105	0.57	0.31	50
818250	1	TE	1.38845	0.17	0.19	74
818285	1	TE	0.31647	0.34	0.57	95
818089_01	1	MC	–0.39005	0.47	0.30	54
818041_02	1	TE	0.84910	0.24	0.28	105
818041_03a	1	MC	–0.27809	0.45	0.30	44
818041_04b	1	TE	0.22858	0.35	0.44	74
818197_03	1	MC	–0.37892	0.47	0.50	64
818197_02	1	MC	–0.65825	0.53	0.43	41
818028	1	TE	0.53728	0.29	0.32	58
818082	1	TE	0.47632	0.31	0.48	60
818165	1	TE	1.04296	0.21	0.46	75
818015_01a	1	TE	0.52711	0.30	0.45	85
818015_02a	1	TE	0.09657	0.37	0.17	37
818015_03a	2	TE	1.09743	0.48	0.21	71
818015_05b	4	CR	0.26823	1.21	0.61	356

Table F.5: Grade 11 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Level
HS18038_02	1	TE	AID	LS4	S, P, and Q	Life	Sensemaking	2	DCI = A1; SEP = A3
HS18038_10	1	TE	AID	LS4	S & SM	Life	Sensemaking	2	DCI = A1; SEP = E2
HS18038_12	1	TE	EAE	LS4	S, P, and Q	Life	Critiquing	2	DCI = A2; SEP = B2
HS18038_16	1	TE	OECI	LS4	S & SM	Life	Critiquing	2	DCI = A2; SEP = B2
HS18047_02	1	TE	AID	LS1	SC	Life	Sensemaking	2	DCI = D1; SEP = A1
HS18047_03	1	MC	OECI	LS1	PAT	Life	Critiquing	2	DCI = D1; SEP = A1
HS18047_05	1	MC	AQDP	LS1	S, P, and Q	Life	Investigating	2	DCI = D1; SEP = G1
HS18047_07	1	TE	UMCT	LS1	SC	Life	Investigating	2	DCI = D1; SEP = F2
HS18089_01	1	TE	AQDP	ESS3	SC	Earth and Space	Investigating	2	DCI = B2; SEP = D2
HS18089_04	1	MC	AID	ESS3	SC	Earth and Space	Sensemaking	2	DCI = B1; SEP = A2
HS18089_06	1	TE	EAE	ESS3	SC	Earth and Space	Critiquing	2	DCI = B2; SEP = B2
HS18051_02	1	MC	DUM	ESS2	E&M	Earth and Space	Sensemaking	2	DCI = A3; SEP = C2
HS18051_04	1	TE	AID	ESS2	E&M	Earth and Space	Sensemaking	2	DCI = A3; SEP = A2
HS18051_05	1	TE	CEDS	ESS2	E&M	Earth and Space	Sensemaking	2	DCI = A4; SEP = B2
HS18051_08	1	MC	CEDS	ESS2	E&M	Earth and Space	Sensemaking	2	DCI = A3; SEP = A3
HS18004_01	1	MC	OECI	PS1	PAT	Physical	Critiquing	2	DCI = B2; SEP = A1
HS18004_04	1	TE	DUM	PS1	PAT	Physical	Sensemaking	2	DCI = B2; SEP = C2
HS18004_05	1	MC	DUM	PS1	S & SM	Physical	Sensemaking	2	DCI = B2; SEP = F1
HS19003_01A	1	TE	DUM	PS4	C and E	Physical	Sensemaking	2	DCI = A2; SEP = C2
HS19003_03A	1	TE	UMCT	PS4	S & SM	Physical	Investigating	2	DCI = A1; SEP = D2
HS19003_05A	1	TE	AQDP	PS4	S & SM	Physical	Investigating	2	DCI = A2; SEP = F3
HS19003_07A	4	CR	EAE	PS4	C and E	Physical	Critiquing	2	DCI = A4; SEP = F3
HS19003_09A	1	TE	AID	PS4	PAT	Physical	Sensemaking	2	DCI = A4; SEP = E3
HS18006_01	1	MC	AID	LS3	S, P, and Q	Life	Sensemaking	2	DCI = C2; SEP = B2
HS18006_05	1	MC	EAE	LS3	SC	Life	Critiquing	2	DCI = C3; SEP = E3
HS18006_07	1	TE	CEDS	LS3	C and E	Life	Sensemaking	2	DCI = C3; SEP = D3
HS18069_01	1	TE	AID	LS2	S,P, and Q	Life	Sensemaking	2	DCI = A1; SEP = A1

Table F.5: Grade 11 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Levels
HS18069_04	1	MC	AID	LS2	SC	Life	Sensemaking	2	DCI = C1; SEP = A1
HS18069_07	1	TE	OECI	LS2	SC	Life	Critiquing	2	DCI = C2; SEP = A1
HS18039_02	1	MC	AID	PS2	PAT	Physical	Sensemaking	2	DCI = A2; SEP = A1
HS18039_03	1	TE	UMCT	PS2	S, P, and Q	Physical	Investigating	2	DCI = A2; SEP = F2
HS18039_05	1	TE	UMCT	PS2	PAT	Physical	Investigating	2	DCI = A2; SEP = F2
HS18082_03	1	MC	PACI	PS2	C and E	Physical	Investigating	2	DCI = C3; SEP = D3
HS18082_06	1	TE	PACI	PS2	C and E	Physical	Investigating	2	DCI = C4; SEP = E3
HS18082_07	1	MC	UMCT	PS2	C and E	Physical	Investigating	2	DCI = C4; SEP = F3
HS18013_01	1	MC	UMCT	ESS1	S & SM	Earth and Space	Investigating	2	DCI = A3; SEP = D3
HS18013_03	1	MC	UMCT	ESS1	S & SM	Earth and Space	Investigating	2	DCI = A4; SEP = F3
HS18013_05	1	MC	DUM	ESS1	S, P, and Q	Earth and Space	Sensemaking	2	DCI = B1; SEP = C2
HS18040_01	1	MC	AQDP	PS3	E&M	Physical	Investigating	2	DCI = C3; SEP = A2
HS18040_03	1	TE	PACI	PS3	C and E	Physical	Investigating	2	DCI = C4; SEP = E1
HS18040_04	1	TE	PACI	PS3	C and E	Physical	Investigating	2	DCI = C3; SEP = E1
HS19004_01a	1	TE	AQDP	ESS3	S & SM	Earth and Space	Investigating	2	DCI = B2; SEP = D4
HS19004_03a	1	TE	CEDS	ESS3	S & SM	Earth and Space	Sensemaking	2	DCI = A2; SEP = E1
HS19004_06b	1	TE	OECI	ESS3	C and E	Earth and Space	Critiquing	2	DCI = A2; SEP = A1
HS19004_07a	1	TE	EAE	ESS3	S & SM	Earth and Space	Critiquing	2	DCI = B3; SEP = E3
HS19004_09a	4	CR	CEDS	ESS3	S, P, and Q	Earth and Space	Sensemaking	2	DCI = B3; SEP = C2
HS18005_02	1	MC	AID	PS1	C and E	Physical	Sensemaking	2	DCI = A2; SEP = A1
HS18005_03	1	MC	AID	PS1	C and E	Physical	Sensemaking	2	DCI = A2; SEP = A2
HS18005_05	1	TE	CEDS	PS1	S, P, and Q	Physical	Sensemaking	2	DCI = A3; SEP = C3
HS18001_01	1	TE	UMCT	PS2	C and E	Physical	Investigating	2	DCI = A2; SEP = F2
HS18001_07	1	MC	UMCT	PS2	SC	Physical	Investigating	2	DCI = A3; SEP = F2
HS18059_03	1	MC	DUM	PS1	S, P, and Q	Physical	Sensemaking	2	DCI = A2; SEP = C2
HS18059_05	1	TE	DUM	PS3	E&M	Earth and Space	Sensemaking	2	DCI = A1; SEP = C2
HS18059_07	1	TE	OECI	PS3	SC	Earth and Space	Critiquing	2	DCI = A2; SEP = A1

Table F.5: Grade 11 Test Map – Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	DOK	Range PLD Levels
HS18071_01	1	MC	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = D2; SEP = A2
HS18071_04	1	MC	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = D2; SEP = A1
HS18071_05	1	TE	AID	ESS2	PAT	Earth and Space	Sensemaking	2	DCI = D2; SEP = A2
HS18018_02	1	TE	AID	ESS3	SC	Earth and Space	Sensemaking	2	DCI = A1; SEP = A1
HS18018_03	1	TE	AID	ESS3	SC	Earth and Space	Sensemaking	2	DCI = A1; SEP = A1
HS18018_05	1	TE	EAE	ESS3	SC	Earth and Space	Critiquing	2	DCI = A1; SEP = E2
HS18018_08	1	TE	CEDS	ESS3	SC	Earth and Space	Sensemaking	2	DCI = A1; SEP = B2
HS18011_01	1	MC	UMCT	LS4	S,P, and Q	Life	Investigating	2	DCI = B1; SEP = F3
HS18011_05	1	MC	DUM	LS4	PAT	Life	Sensemaking	3	DCI = B2; SEP = C2
HS18011_06	1	TE	CEDS	LS4	C and E	Life	Sensemaking	2	DCI = B2; SEP = D2
HS19011_02b	2	TE	AQDP	LS2	S & SM	Life	Investigating	2	DCI = A2; SEP = E2
HS19011_04b	1	TE	PACI	LS2	SF	Life	Investigating	3	DCI = A1; SEP = B2
HS19011_06b	1	TE	UMCT	LS2	S, P, and Q	Life	Investigating	2	DCI = A1; SEP = F2
HS19011_07a	4	CR	EAE	LS2	PAT	Life	Critiquing	3	DCI = A2; SEP = E2

Table F.6: Grade 11 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
HS18038_02	1	TE	–0.68282	0.58	0.34	74
HS18038_10	1	TE	–0.21273	0.49	0.33	67
HS18038_12	1	TE	–1.42390	0.71	0.46	50
HS18038_16	1	TE	–0.75146	0.59	0.29	76
HS18047_02	1	TE	–0.47471	0.53	0.56	106
HS18047_03	1	MC	–0.11909	0.47	0.40	54
HS18047_05	1	MC	–0.55767	0.55	0.42	39
HS18047_07	1	TE	–1.18490	0.66	0.54	42
HS18089_01	1	TE	1.58741	0.18	0.36	112
HS18089_04	1	MC	–1.21120	0.67	0.53	61
HS18089_06	1	TE	–0.50771	0.53	0.62	68

Table F.6: Grade 11 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
HS18051_02	1	MC	0.83006	0.29	0.31	81
HS18051_04	1	TE	–0.02665	0.45	0.34	60
HS18051_05	1	TE	0.76907	0.30	0.37	41
HS18051_08	1	MC	0.48224	0.35	0.35	26
HS18004_01	1	MC	–0.03307	0.45	0.38	41
HS18004_04	1	TE	0.19559	0.41	0.24	39
HS18004_05	1	MC	–0.01170	0.45	0.25	38
HS19003_01A	1	TE	–0.41693	0.52	0.53	50
HS19003_03A	1	TE	–0.44804	0.52	0.65	43
HS19003_05A	1	TE	2.26983	0.11	0.23	40
HS19003_07A	4	CR	0.84865	1.14	0.68	290
HS19003_09A	1	TE	–0.69006	0.57	0.52	56
HS18006_01	1	MC	–1.51770	0.72	0.48	55
HS18006_05	1	MC	0.97803	0.27	0.46	74
HS18006_07	1	TE	–0.62552	0.56	0.41	43
HS18069_01	1	TE	–0.65162	0.56	0.54	59
HS18069_04	1	MC	–0.93538	0.62	0.40	36
HS18069_07	1	TE	0.26127	0.39	0.59	79
HS18039_02	1	MC	–1.10260	0.65	0.44	42
HS18039_03	1	TE	–0.56647	0.54	0.63	68
HS18039_05	1	TE	–1.00680	0.63	0.56	40
HS18082_03	1	MC	0.30253	0.39	0.34	62
HS18082_06	1	TE	0.80561	0.30	0.44	38
HS18082_07	1	MC	0.22545	0.40	0.48	69
HS18013_01	1	MC	–1.09850	0.65	0.51	65
HS18013_03	1	MC	–1.33090	0.69	0.45	37
HS18013_05	1	MC	–0.51066	0.54	0.40	34
HS18040_01	1	MC	0.19436	0.41	0.26	32
HS18040_03	1	TE	0.51802	0.34	0.40	26
HS18040_04	1	TE	0.12950	0.41	0.59	35
HS19004_01a	1	TE	–0.21588	0.48	0.53	80

Table F.6: Grade 11 Test Map – Item Statistics

UIN	Points	Item Type	Rasch	Mean	RPB	Median Time
HS19004_03a	1	TE	0.53253	0.34	0.51	36
HS19004_06b	1	TE	0.65995	0.32	0.45	43
HS19004_07a	1	TE	0.85302	0.28	0.42	55
HS19004_09a	4	CR	1.15350	0.76	0.57	274
HS18005_02	1	MC	0.94272	0.27	0.29	76
HS18005_03	1	MC	–0.87176	0.60	0.56	43
HS18005_05	1	TE	–1.32110	0.68	0.52	21
HS18001_01	1	TE	–0.89782	0.60	0.62	56
HS18001_07	1	MC	–0.08102	0.46	0.17	53
HS18059_03	1	MC	0.64001	0.32	0.36	43
HS18059_05	1	TE	0.77813	0.30	0.60	49
HS18059_07	1	TE	0.31003	0.38	0.38	33
HS18071_01	1	MC	0.05550	0.43	0.42	47
HS18071_04	1	MC	–0.95385	0.62	0.51	38
HS18071_05	1	TE	–0.64984	0.56	0.54	21
HS18018_02	1	TE	–0.42561	0.52	0.56	29
HS18018_03	1	TE	1.04112	0.26	0.34	33
HS18018_05	1	TE	0.82941	0.29	0.41	27
HS18018_08	1	TE	0.78624	0.29	0.49	45
HS18011_01	1	MC	0.08545	0.43	0.26	51
HS18011_05	1	MC	0.22390	0.41	0.13	27
HS18011_06	1	TE	0.90811	0.28	0.45	34
HS19011_02b	2	TE	0.55566	0.69	0.53	89
HS19011_04b	1	TE	1.65016	0.17	0.43	43
HS19011_06b	1	TE	0.52029	0.34	0.53	32
HS19011_07a	4	CR	0.59221	1.17	0.68	217

APPENDIX G: Scale Score Cumulative Frequency Distributions

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

Raw Score	Scale Score	All Cum. #	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	12	0.01	0.00	0.02	0.00	0.01	0.02	0.01
2	100	47	0.05	0.02	0.08	0.02	0.11	0.05	0.03
3	100	164	0.16	0.10	0.22	0.06	0.47	0.18	0.06
4	100	375	0.37	0.24	0.49	0.09	1.00	0.47	0.15
5	100	817	0.81	0.56	1.05	0.17	1.97	1.13	0.35
6	100	1,519	1.50	1.13	1.86	0.31	3.49	2.16	0.66
7	100	2,460	2.43	1.90	2.94	0.56	5.49	3.56	1.06
8	100	3,633	3.59	2.83	4.32	0.83	8.06	5.25	1.58
9	100	5,031	4.97	3.96	5.94	1.19	10.88	7.36	2.22
10	100	6,544	6.47	5.33	7.56	1.70	13.88	9.70	2.86
11	100	8,198	8.10	6.74	9.40	2.11	17.06	12.22	3.65
12	100	9,970	9.85	8.30	11.34	2.53	20.68	14.79	4.51
13	102	11,709	11.57	9.96	13.11	3.03	23.78	17.42	5.43
14	107	13,556	13.39	11.74	14.99	3.55	27.15	20.09	6.45
15	111	15,473	15.29	13.70	16.81	4.14	30.74	22.73	7.56
16	116	17,488	17.28	15.63	18.86	4.87	33.97	25.69	8.76
17	120	19,544	19.31	17.68	20.87	5.50	37.29	28.56	10.12
18	124	21,620	21.36	19.79	22.87	6.28	40.53	31.51	11.48
19	128	23,714	23.43	21.98	24.83	6.90	43.58	34.33	13.07
20	132	25,868	25.56	24.27	26.79	7.66	46.68	37.36	14.59
21	135	28,102	27.77	26.69	28.80	8.71	49.49	40.52	16.24
22	139	30,440	30.08	29.14	30.97	9.73	52.41	43.52	18.18
23	143	32,719	32.33	31.57	33.05	10.66	55.28	46.35	20.18
24	146	35,217	34.80	34.29	35.29	11.79	58.21	49.62	22.24
25	150	37,646	37.20	36.90	37.48	13.20	61.06	52.54	24.35
26	154	40,143	39.66	39.58	39.75	14.64	63.71	55.59	26.61
27	157	42,619	42.11	42.23	42.00	16.22	66.38	58.45	28.90
28	161	45,159	44.62	44.90	44.35	17.81	68.98	61.28	31.38
29	164	47,770	47.20	47.70	46.72	19.60	71.52	64.08	34.02
30	168	50,486	49.88	50.64	49.16	21.84	74.05	66.87	36.82
31	171	53,204	52.57	53.45	51.73	24.01	76.37	69.63	39.76
32	175	55,940	55.27	56.28	54.30	26.10	78.88	72.29	42.72
33	178	58,582	57.88	58.97	56.84	28.30	80.80	74.71	45.80
34	182	61,234	60.50	61.70	59.35	30.60	82.61	77.08	48.95
35	185	63,874	63.11	64.36	61.91	33.25	84.53	79.30	52.06
36	189	66,509	65.72	67.06	64.42	35.90	86.16	81.44	55.35
37	193	69,192	68.37	69.80	66.99	38.97	87.94	83.40	58.65
38	196	71,626	70.77	72.27	69.33	41.80	89.23	85.20	61.73
39	200	74,374	73.49	75.00	72.03	45.16	90.64	87.18	65.24
40	204	76,946	76.03	77.48	74.63	48.56	92.04	88.87	68.56

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

Raw Score	Scale Score	All Cum. #	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
41	208	79,424	78.48	79.97	77.05	51.78	93.37	90.47	71.81
42	212	81,960	80.98	82.37	79.64	55.68	94.57	91.97	75.16
43	216	84,257	83.25	84.63	81.93	59.23	95.44	93.22	78.32
44	220	86,631	85.60	86.85	84.39	63.24	96.36	94.51	81.46
45	224	88,845	87.78	88.98	86.64	67.36	97.11	95.60	84.41
46	229	90,970	89.88	90.96	88.85	71.69	97.82	96.57	87.17
47	233	92,855	91.75	92.64	90.89	75.87	98.32	97.32	89.68
48	238	94,561	93.43	94.16	92.73	80.09	98.79	97.97	91.89
49	243	96,097	94.95	95.54	94.38	83.98	99.11	98.51	93.87
50	249	97,440	96.28	96.75	95.83	87.43	99.34	98.95	95.65
51	254	98,571	97.39	97.77	97.03	90.74	99.60	99.29	97.02
52	260	99,443	98.26	98.54	97.99	93.38	99.80	99.59	98.05
53	267	100,134	98.94	99.10	98.78	95.73	99.92	99.79	98.84
54	275	100,626	99.42	99.51	99.35	97.61	99.97	99.88	99.40
55	283	100,923	99.72	99.76	99.68	98.73	99.98	99.94	99.74
56	293	101,076	99.87	99.88	99.86	99.40	99.99	99.96	99.89
57	300	101,167	99.96	99.97	99.95	99.78	100.00	99.99	99.97
58	300	101,198	99.99	99.99	99.99	99.94	100.00	100.00	100.00
59	300	101,207	100.00	100.00	100.00	100.00	100.00	100.00	100.00
60	300	101,208	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. #	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
0	100	6	0.01	0.01	0.01	0.00	0.02	0.01	0.00
1	100	20	0.02	0.02	0.02	0.00	0.07	0.03	0.00
2	100	50	0.05	0.04	0.06	0.00	0.17	0.07	0.01
3	100	134	0.13	0.09	0.18	0.03	0.42	0.18	0.04
4	100	324	0.32	0.21	0.43	0.03	0.89	0.50	0.11
5	100	718	0.72	0.52	0.91	0.11	1.76	1.16	0.26
6	100	1,361	1.36	0.94	1.76	0.24	3.11	2.15	0.58
7	100	2,409	2.41	1.72	3.08	0.52	5.17	3.78	1.14
8	101	3,874	3.88	2.96	4.75	0.86	8.18	6.20	1.80
9	107	5,915	5.92	4.60	7.18	1.34	12.36	9.33	2.86
10	112	8,338	8.35	6.71	9.91	1.87	17.09	13.15	4.15
11	117	11,164	11.18	9.17	13.10	2.64	22.61	17.56	5.60
12	122	14,242	14.26	12.00	16.42	3.42	28.24	22.20	7.46
13	126	17,415	17.44	15.03	19.74	4.46	33.84	27.04	9.36
14	130	20,636	20.67	18.13	23.09	5.42	38.83	32.00	11.47
15	134	23,933	23.97	21.40	26.42	6.48	44.04	36.83	13.74
16	137	27,018	27.06	24.64	29.36	7.70	48.73	41.12	16.00
17	141	30,015	30.06	27.78	32.23	8.90	53.12	45.14	18.32
18	144	32,863	32.91	30.71	35.01	10.31	56.76	49.14	20.58
19	147	35,665	35.72	33.60	37.73	11.71	59.89	52.81	23.07
20	150	38,365	38.42	36.46	40.29	13.15	63.05	56.14	25.53
21	153	41,072	41.13	39.32	42.86	14.51	66.30	59.39	28.07
22	156	43,653	43.72	42.06	45.30	15.91	69.08	62.44	30.59
23	159	46,112	46.18	44.72	47.58	17.53	71.51	65.09	33.16
24	161	48,576	48.65	47.41	49.83	19.16	74.03	67.64	35.78
25	164	50,979	51.05	49.93	52.12	20.67	76.10	70.12	38.44
26	167	53,432	53.51	52.55	54.42	22.75	78.00	72.50	41.20
27	169	55,800	55.88	55.04	56.68	24.52	79.86	74.75	43.94
28	172	58,141	58.23	57.62	58.81	26.55	81.55	76.94	46.69
29	174	60,409	60.50	60.02	60.96	28.48	83.28	78.85	49.42
30	176	62,605	62.70	62.45	62.93	30.68	84.77	80.64	52.10
31	179	64,739	64.83	64.80	64.87	32.58	86.39	82.31	54.77
32	181	66,810	66.91	67.08	66.75	34.88	87.68	83.94	57.32
33	184	68,936	69.04	69.39	68.70	37.12	88.92	85.53	60.03
34	186	70,924	71.03	71.53	70.55	39.45	90.14	86.95	62.55
35	188	72,843	72.95	73.56	72.37	41.87	91.21	88.31	64.96
36	191	74,739	74.85	75.62	74.12	44.56	92.15	89.52	67.42
37	193	76,626	76.74	77.64	75.88	47.14	93.14	90.73	69.89
38	195	78,366	78.48	79.47	77.54	49.98	93.94	91.73	72.14
39	198	80,051	80.17	81.17	79.22	52.63	94.76	92.66	74.34
40	200	81,800	81.92	82.96	80.93	55.36	95.41	93.60	76.71
41	202	83,457	83.58	84.64	82.57	58.18	96.03	94.43	78.93
42	205	84,909	85.03	86.18	83.94	60.43	96.59	95.11	80.94

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

Raw Score	Scale Score	All Cum. #	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
43	207	86,342	86.47	87.64	85.35	62.94	96.99	95.79	82.92
44	210	87,728	87.86	89.09	86.68	65.73	97.42	96.39	84.77
45	212	89,066	89.20	90.46	87.99	68.51	97.88	96.99	86.51
46	215	90,315	90.45	91.70	89.25	71.38	98.17	97.44	88.19
47	217	91,474	91.61	92.85	90.42	74.30	98.50	97.79	89.71
48	220	92,555	92.69	93.83	91.61	76.83	98.83	98.21	91.09
49	222	93,602	93.74	94.81	92.73	79.67	99.00	98.57	92.42
50	225	94,508	94.65	95.67	93.68	82.02	99.22	98.88	93.58
51	228	95,332	95.47	96.40	94.59	84.21	99.38	99.12	94.65
52	231	96,067	96.21	97.04	95.41	86.42	99.49	99.30	95.58
53	234	96,771	96.91	97.62	96.24	88.62	99.58	99.45	96.47
54	237	97,404	97.55	98.14	96.99	90.68	99.64	99.61	97.24
55	240	97,964	98.11	98.65	97.59	92.54	99.77	99.71	97.90
56	244	98,400	98.55	99.00	98.11	94.00	99.85	99.80	98.43
57	247	98,784	98.93	99.28	98.60	95.21	99.90	99.89	98.91
58	251	99,100	99.25	99.52	98.99	96.44	99.93	99.93	99.26
59	256	99,351	99.50	99.69	99.32	97.61	99.95	99.96	99.52
60	260	99,502	99.65	99.79	99.51	98.26	99.96	99.97	99.68
61	265	99,624	99.77	99.86	99.68	98.85	99.96	99.98	99.81
62	270	99,722	99.87	99.92	99.82	99.29	99.99	99.99	99.89
63	276	99,785	99.93	99.97	99.90	99.61	99.99	100.00	99.95
64	283	99,820	99.97	99.98	99.95	99.83	99.99	100.00	99.98
65	291	99,834	99.98	99.99	99.97	99.89	99.99	100.00	99.99
66	300	99,843	99.99	100.00	99.99	99.94	99.99	100.00	100.00
67	300	99,850	100.00	100.00	100.00	99.98	100.00	100.00	100.00
68	300	99,851	100.00	100.00	100.00	99.99	100.00	100.00	100.00
69	300	99,852	100.00	100.00	100.00	100.00	100.00	100.00	100.00
70	300	99,852	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. #	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
0	100	7	0.01	0.00	0.01	0.00	0.02	0.00	0.01
1	100	38	0.04	0.04	0.05	0.00	0.12	0.05	0.02
2	100	118	0.13	0.13	0.13	0.02	0.35	0.17	0.08
3	100	242	0.27	0.24	0.30	0.04	0.77	0.33	0.15
4	100	430	0.48	0.43	0.53	0.11	1.26	0.57	0.29
5	100	694	0.78	0.64	0.91	0.15	1.91	0.93	0.49
6	100	1,140	1.28	0.98	1.57	0.29	2.78	1.63	0.85
7	100	1,848	2.08	1.62	2.52	0.53	4.30	2.71	1.41
8	100	2,841	3.19	2.51	3.85	0.88	6.44	4.24	2.16
9	100	4,329	4.86	3.88	5.82	1.46	9.70	6.55	3.25
10	100	6,114	6.87	5.49	8.21	2.04	13.38	9.29	4.68
11	100	8,199	9.21	7.46	10.91	2.72	17.46	12.52	6.41
12	100	10,462	11.75	9.71	13.74	3.61	21.94	16.12	8.19
13	100	12,745	14.32	11.89	16.67	4.39	26.31	19.55	10.14
14	100	15,044	16.90	14.15	19.57	5.27	30.59	23.32	11.96
15	100	17,239	19.37	16.45	22.20	6.28	34.61	26.83	13.73
16	102	19,352	21.74	18.67	24.72	7.01	38.59	30.10	15.52
17	106	21,268	23.90	20.81	26.88	7.99	41.57	33.11	17.23
18	111	23,136	26.00	22.90	28.99	8.88	44.54	35.88	18.99
19	115	24,987	28.07	25.04	31.02	9.92	47.37	38.71	20.69
20	119	26,703	30.00	26.94	32.97	10.82	49.97	41.18	22.37
21	123	28,512	32.04	29.04	34.94	11.93	52.68	43.83	24.10
22	127	30,187	33.92	31.02	36.73	13.03	55.26	46.16	25.69
23	131	31,864	35.80	32.96	38.55	14.15	57.58	48.49	27.38
24	134	33,482	37.62	34.75	40.40	15.33	59.92	50.58	29.05
25	138	35,097	39.43	36.71	42.07	16.32	62.01	52.70	30.80
26	141	36,757	41.30	38.72	43.80	17.43	63.97	55.03	32.56
27	145	38,381	43.12	40.72	45.45	18.56	65.97	57.21	34.30
28	148	40,041	44.99	42.84	47.07	19.84	68.03	59.33	36.08
29	151	41,635	46.78	44.78	48.71	21.09	69.75	61.37	37.87
30	155	43,250	48.59	46.78	50.35	22.34	71.52	63.46	39.69
31	158	44,859	50.40	48.80	51.96	23.60	73.23	65.46	41.55
32	161	46,512	52.26	50.78	53.69	25.02	75.01	67.31	43.54
33	164	48,061	54.00	52.73	55.23	26.35	76.63	69.05	45.40
34	167	49,615	55.75	54.63	56.83	27.59	78.26	70.88	47.25
35	170	51,094	57.41	56.40	58.39	28.88	79.67	72.41	49.13
36	173	52,640	59.15	58.32	59.94	30.56	81.08	74.18	50.94
37	176	54,149	60.84	60.24	61.42	32.22	82.36	75.75	52.85
38	179	55,630	62.50	62.04	62.96	33.54	83.62	77.35	54.74
39	182	57,157	64.22	63.82	64.61	35.22	84.86	78.85	56.70
40	185	58,680	65.93	65.73	66.12	37.09	86.09	80.19	58.71

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

Raw Score	Scale Score	All Cum. #	All Cum. %	Female Cum. %	Male Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	White Cum. %
41	188	60,183	67.62	67.55	67.68	38.76	87.30	81.76	60.62
42	191	61,680	69.30	69.37	69.23	40.67	88.42	83.20	62.53
43	194	63,106	70.90	71.15	70.67	42.38	89.40	84.50	64.46
44	197	64,434	72.40	72.77	72.04	44.23	90.26	85.69	66.22
45	200	65,833	73.97	74.47	73.48	46.17	91.02	86.86	68.17
46	203	67,201	75.51	76.08	74.95	48.02	91.86	87.93	70.11
47	206	68,505	76.97	77.65	76.31	50.21	92.57	88.95	71.88
48	209	69,764	78.39	79.21	77.59	52.16	93.30	89.87	73.66
49	212	71,102	79.89	80.77	79.03	54.12	93.98	90.86	75.57
50	215	72,315	81.25	82.18	80.35	56.16	94.54	91.69	77.31
51	218	73,542	82.63	83.59	81.71	58.21	95.13	92.49	79.06
52	222	74,709	83.94	84.96	82.96	60.39	95.60	93.29	80.68
53	225	75,911	85.29	86.38	84.24	62.52	96.07	94.13	82.41
54	228	77,005	86.52	87.61	85.46	64.94	96.57	94.82	83.87
55	232	78,118	87.77	88.84	86.74	67.25	97.01	95.45	85.47
56	235	79,124	88.90	89.90	87.94	69.46	97.31	95.98	86.93
57	239	80,125	90.03	91.08	89.01	71.58	97.72	96.52	88.37
58	242	81,093	91.11	92.11	90.15	73.85	98.12	97.03	89.70
59	246	81,994	92.13	93.08	91.20	76.03	98.44	97.48	90.94
60	250	82,848	93.09	93.94	92.26	78.30	98.65	97.90	92.12
61	254	83,652	93.99	94.81	93.19	80.15	98.88	98.29	93.29
62	258	84,369	94.80	95.60	94.02	82.30	99.10	98.60	94.22
63	263	85,047	95.56	96.28	94.86	84.23	99.27	98.85	95.17
64	268	85,663	96.25	96.95	95.57	86.16	99.37	99.04	96.03
65	273	86,278	96.94	97.53	96.37	88.27	99.51	99.23	96.83
66	278	86,844	97.58	98.08	97.09	90.47	99.63	99.45	97.51
67	284	87,291	98.08	98.52	97.66	92.29	99.69	99.59	98.05
68	290	87,723	98.56	98.90	98.24	94.05	99.78	99.71	98.57
69	296	88,051	98.93	99.21	98.66	95.54	99.85	99.78	98.94
70	300	88,364	99.28	99.52	99.06	96.90	99.91	99.83	99.31
71	300	88,585	99.53	99.69	99.38	97.82	99.95	99.92	99.57
72	300	88,753	99.72	99.81	99.63	98.69	99.95	99.96	99.74
73	300	88,871	99.85	99.91	99.80	99.30	99.98	99.98	99.87
74	300	88,955	99.95	99.96	99.94	99.66	100.00	100.00	99.97
75	300	88,985	99.98	99.98	99.98	99.88	100.00	100.00	99.99
76	300	88,997	100.00	99.99	100.00	99.97	100.00	100.00	100.00
77	300	88,999	100.00	100.00	100.00	99.98	100.00	100.00	100.00
78	300	89,001	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

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
1	0.33558	1.24	1.37	0.23	0.37	0.16	0.46
2	0.44543	0.99	0.97	0.44	1.03	0.00	0.44
3	0.48982	1.07	1.15	0.37	0.81	0.06	0.43
4	0.17780	1.03	1.05	0.41	0.92	0.05	0.50
5	−0.11391	0.96	0.93	0.48	1.11	0.00	0.55
6	0.52075	1.07	1.14	0.37	0.81	0.06	0.43
7	−1.53790	1.06	1.14	0.32	0.91	0.03	0.79
8	−0.95565	0.97	0.90	0.45	1.06	0.03	0.70
9	1.70235	1.00	1.11	0.36	0.98	0.01	0.23
10	−0.96460	0.84	0.74	0.55	1.28	0.00	0.70
11	−0.31411	0.81	0.76	0.59	1.43	0.00	0.59
12	0.41557	0.81	0.76	0.59	1.46	0.00	0.45
13	−2.50290	0.81	0.55	0.44	1.17	0.00	0.90
14	−1.48750	0.72	0.53	0.61	1.37	0.00	0.78
15	−0.22986	1.14	1.23	0.32	0.65	0.04	0.57
16	−0.50623	1.04	1.08	0.39	0.90	0.01	0.63
17	−0.51762	0.83	0.79	0.57	1.35	0.00	0.63
18	−0.97657	0.91	0.83	0.49	1.17	0.00	0.71
19	0.12630	0.96	0.95	0.47	1.09	0.00	0.50
20	1.23432	1.11	1.18	0.32	0.80	0.03	0.30
21	0.61445	1.15	1.31	0.30	0.60	0.10	0.41
22	−0.53023	0.87	0.82	0.54	1.27	0.00	0.63
23	−1.31270	0.98	0.96	0.40	1.02	0.01	0.76
24	−1.34630	0.87	0.80	0.49	1.18	0.00	0.77
25	1.02603	0.85	0.78	0.53	1.30	0.00	0.34
26	1.32364	1.09	1.40	0.28	0.77	0.05	0.28
27	0.62693	1.12	1.23	0.32	0.69	0.08	0.41
28	−0.52643	1.10	1.10	0.36	0.80	0.16	0.63
29	1.88462	1.01	1.04	0.34	0.98	0.00	0.20
30	1.45832	0.95	1.12	0.41	1.04	0.00	0.26
31	−0.39235	1.15	1.26	0.31	0.64	0.13	0.60
32	1.25079	1.07	1.31	0.31	0.81	0.04	0.30
33	0.85600	1.13	1.26	0.31	0.69	0.07	0.37
34	0.52987	1.05	1.06	0.40	0.89	0.01	0.43
35	−0.69901	1.05	1.07	0.39	0.90	0.12	0.66
36	−1.15990	0.89	0.79	0.50	1.18	0.00	0.74
37	−0.09625	0.96	0.94	0.48	1.11	0.00	0.55
38	1.26325	0.96	0.93	0.65	1.05	0.00	1.10

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

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
39	–1.27660	1.00	0.97	0.40	1.01	0.00	0.75
40	1.21804	0.84	0.81	0.53	1.29	0.00	0.30
41	0.56024	0.80	0.75	0.60	1.48	0.00	0.42
42	–0.60579	1.05	1.04	0.39	0.91	0.03	0.64
43	0.18245	0.99	0.98	0.45	1.02	0.00	0.50
44	–0.23129	1.17	1.31	0.29	0.57	0.03	0.57
45	–1.66820	1.00	1.12	0.34	0.97	0.00	0.81
46	–0.43043	1.24	1.28	0.60	0.72	0.00	2.51
47	0.22098	1.00	0.99	0.44	1.01	0.00	0.49
48	–0.42386	0.88	0.82	0.54	1.27	0.00	0.61
49	0.90368	0.97	0.97	0.54	1.05	0.00	0.75
50	1.43906	1.07	1.02	0.56	0.97	0.00	0.99

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

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
1	0.54821	1.04	0.99	0.34	0.95	0.00	0.29
2	1.16574	1.04	1.02	0.30	0.96	0.01	0.20
3	0.01609	0.98	0.93	0.67	1.04	0.00	1.54
4	–0.06862	1.12	1.15	0.27	0.68	0.06	0.41
5	–0.17777	1.04	1.04	0.35	0.89	0.03	0.43
6	0.98992	1.14	1.22	0.34	0.81	0.07	0.47
7	0.58747	1.11	1.06	0.60	0.94	0.00	0.95
8	–1.95470	0.90	0.82	0.40	1.15	0.00	0.77
9	0.58734	1.07	1.11	0.30	0.87	0.03	0.29
10	–0.85213	1.20	1.24	0.19	0.42	0.20	0.57
11	2.05769	0.93	0.92	0.33	1.05	0.00	0.10
12	0.52711	0.92	0.88	0.45	1.15	0.00	0.30
13	0.09657	1.21	1.33	0.17	0.44	0.13	0.37
14	1.09743	1.27	1.36	0.21	0.63	0.13	0.48
15	0.26823	0.99	0.98	0.61	1.02	0.00	1.21
16	0.53728	1.05	1.10	0.32	0.90	0.02	0.29
17	0.40108	0.91	0.90	0.47	1.17	0.00	0.32
18	0.84910	1.05	1.23	0.28	0.88	0.03	0.24
19	–0.27809	1.10	1.13	0.30	0.71	0.11	0.45
20	0.22858	0.94	0.98	0.44	1.10	0.00	0.35
21	–0.18681	0.92	0.92	0.47	1.21	0.00	0.43
22	–3.11420	0.85	0.55	0.35	1.14	0.00	0.90
23	–0.35735	0.88	0.85	0.50	1.36	0.00	0.47
24	–0.00256	1.13	1.20	0.25	0.64	0.07	0.39
25	–0.61878	0.89	0.87	0.49	1.33	0.00	0.52

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

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
26	0.89336	0.92	0.86	0.45	1.13	0.00	0.24
27	0.59223	1.23	1.39	0.12	0.56	0.09	0.28
28	1.48503	0.91	0.82	0.42	1.10	0.00	0.16
29	–0.53246	1.09	1.12	0.30	0.72	0.07	0.50
30	–0.73936	1.08	1.11	0.29	0.76	0.01	0.55
31	0.47632	0.89	0.91	0.48	1.19	0.00	0.31
32	–0.24646	1.26	1.35	0.13	0.23	0.17	0.44
33	–0.39005	1.09	1.13	0.30	0.70	0.10	0.47
34	–1.77250	0.84	0.70	0.49	1.30	0.00	0.74
35	–1.62630	0.87	0.82	0.44	1.24	0.00	0.72
36	0.08253	0.90	0.89	0.49	1.24	0.00	0.38
37	1.04296	0.89	0.85	0.46	1.15	0.00	0.21
38	–0.95953	0.84	0.80	0.52	1.46	0.00	0.59
39	–0.65825	0.96	0.95	0.43	1.12	0.02	0.53
40	–0.37892	0.89	0.88	0.50	1.32	0.00	0.47
41	–0.84105	1.07	1.09	0.31	0.79	0.07	0.57
42	1.38845	1.09	1.43	0.19	0.85	0.03	0.17
43	0.21866	1.02	1.10	0.35	0.91	0.04	0.35
44	0.24607	0.98	1.01	0.40	1.03	0.00	0.35
45	0.44726	0.95	0.91	0.43	1.10	0.00	0.31
46	–0.81809	0.82	0.78	0.55	1.53	0.00	0.56
47	0.31647	0.82	0.76	0.57	1.39	0.00	0.34
48	0.05498	1.18	1.23	0.21	0.55	0.08	0.38
49	–0.23481	0.89	0.86	0.50	1.33	0.00	0.44
50	1.16099	1.06	1.25	0.25	0.90	0.02	0.20
51	0.12636	0.99	1.00	0.39	1.01	0.00	0.37
52	–1.03570	0.92	0.86	0.47	1.25	0.00	0.61
53	–0.94620	0.95	0.93	0.42	1.15	0.00	0.59
54	–0.30320	0.89	0.88	0.50	1.31	0.00	0.46
55	0.25410	1.13	1.15	0.25	0.71	0.04	0.34
56	–0.05846	0.92	0.89	0.47	1.22	0.00	0.41
57	1.13709	1.04	1.08	0.28	0.94	0.01	0.20
58	–0.83361	0.81	0.76	0.56	1.57	0.00	0.57
59	0.10521	0.94	0.91	0.45	1.16	0.00	0.37

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

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
1	–0.89782	0.77	0.70	0.62	1.52	0.00	0.60
2	–0.08102	1.34	1.48	0.17	0.16	0.23	0.46
3	–0.03307	1.09	1.11	0.38	0.79	0.05	0.45
4	0.19559	1.25	1.36	0.24	0.42	0.13	0.41
5	–0.01170	1.24	1.33	0.25	0.41	0.15	0.45
6	0.94272	1.14	1.35	0.29	0.74	0.05	0.27
7	–0.87176	0.85	0.80	0.56	1.35	0.00	0.60
8	–1.32110	0.87	0.77	0.52	1.27	0.00	0.68
9	–1.51770	0.87	0.84	0.48	1.21	0.00	0.72
10	0.97803	0.94	0.93	0.46	1.08	0.00	0.27
11	–0.62552	1.04	1.04	0.41	0.92	0.02	0.56
12	0.08545	1.23	1.35	0.26	0.45	0.15	0.43
13	0.22390	1.39	1.64	0.13	0.09	0.22	0.41
14	0.90811	0.96	1.00	0.45	1.05	0.00	0.28
15	–1.09850	0.91	0.83	0.51	1.21	0.00	0.65
16	–1.33090	0.93	0.94	0.45	1.12	0.00	0.69
17	–0.51066	1.06	1.05	0.40	0.86	0.08	0.54
18	–0.42561	0.85	0.84	0.56	1.35	0.00	0.52
19	1.04112	1.10	1.10	0.34	0.86	0.02	0.26
20	0.82941	1.02	1.03	0.41	0.96	0.00	0.29
21	0.78624	0.94	0.84	0.49	1.13	0.00	0.29
22	–0.68282	1.12	1.13	0.34	0.72	0.10	0.58
23	–0.21273	1.15	1.19	0.33	0.64	0.12	0.49
24	–1.42390	0.91	0.85	0.46	1.17	0.00	0.71
25	–0.75146	1.18	1.24	0.29	0.57	0.17	0.59
26	–1.10260	0.97	0.95	0.44	1.06	0.00	0.65
27	–0.56647	0.77	0.72	0.63	1.56	0.00	0.54
28	–1.00680	0.83	0.74	0.56	1.38	0.00	0.63
29	0.19436	1.23	1.33	0.26	0.47	0.12	0.41
30	0.51802	1.05	1.03	0.40	0.91	0.01	0.34
31	0.12950	0.84	0.78	0.59	1.37	0.00	0.41
32	–0.47471	0.87	0.81	0.56	1.34	0.00	0.53
33	–0.11909	1.06	1.10	0.40	0.83	0.07	0.47
34	–0.55767	1.03	1.02	0.42	0.94	0.04	0.55
35	–1.18490	0.84	0.76	0.54	1.33	0.00	0.66
36	0.83006	1.12	1.30	0.31	0.75	0.05	0.29
37	–0.02665	1.13	1.17	0.34	0.69	0.07	0.45
38	0.76907	1.07	1.11	0.37	0.88	0.03	0.30
39	0.48224	1.11	1.16	0.35	0.77	0.05	0.35
40	0.64001	1.08	1.22	0.36	0.81	0.04	0.32

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

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
41	0.77813	0.80	0.67	0.60	1.36	0.00	0.30
42	0.31003	1.07	1.10	0.38	0.84	0.03	0.38
43	–0.65162	0.87	0.84	0.54	1.31	0.00	0.56
44	–0.93538	1.03	1.03	0.40	0.93	0.06	0.62
45	0.26127	0.84	0.76	0.59	1.37	0.00	0.39
46	0.05550	1.04	1.08	0.42	0.90	0.05	0.43
47	–0.95385	0.91	0.86	0.51	1.21	0.00	0.62
48	–0.64984	0.88	0.83	0.54	1.30	0.00	0.56
49	0.30253	1.14	1.24	0.34	0.68	0.08	0.39
50	0.80561	0.98	1.17	0.44	0.98	0.02	0.30
51	0.22545	0.96	0.98	0.48	1.07	0.01	0.40
52	1.58741	1.03	1.01	0.36	0.97	0.00	0.18
53	–1.21120	0.86	0.77	0.53	1.29	0.00	0.67
54	–0.50771	0.78	0.72	0.62	1.53	0.00	0.53
55	–0.41693	0.90	0.86	0.53	1.25	0.00	0.52
56	–0.44804	0.76	0.69	0.65	1.60	0.00	0.52
57	2.26983	1.09	1.33	0.23	0.91	0.01	0.11
58	0.84865	1.06	1.01	0.68	0.98	0.00	1.14
59	–0.69006	0.91	0.87	0.52	1.23	0.00	0.57
60	–0.21588	0.90	0.86	0.53	1.24	0.00	0.48
61	0.53253	0.92	0.85	0.51	1.16	0.00	0.34
62	0.65995	0.98	0.93	0.45	1.05	0.00	0.32
63	0.85302	1.02	0.94	0.42	0.99	0.00	0.28
64	1.15350	1.30	1.21	0.57	0.87	0.00	0.76
65	0.55566	1.04	1.07	0.53	0.92	0.05	0.69
66	1.65016	0.92	0.86	0.43	1.09	0.00	0.17
67	0.52029	0.89	0.82	0.53	1.22	0.00	0.34
68	0.59221	1.12	1.15	0.68	0.94	0.01	1.17

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

Table I.1: Grade 5 – Operational

Raw Score	Theta	CSEM	Slope	Intercept	Unrounded	Scale Score	Scale CSEM	Lower SS	Upper SS
0	−5.64711	1.83752	42.46393	161.6317	-78.1668	100	14.68	100	115
1	−4.41336	1.02076	42.46393	161.6317	-25.7769	100	14.68	100	115
2	−3.67899	0.73522	42.46393	161.6317	5.407326	100	14.68	100	115
3	−3.23357	0.61052	42.46393	161.6317	24.32161	100	14.68	100	115
4	−2.90697	0.5371	42.46393	161.6317	38.19033	100	14.68	100	115
5	−2.64574	0.48757	42.46393	161.6317	49.28318	100	14.68	100	115
6	−2.42600	0.45145	42.46393	161.6317	58.61421	100	14.68	100	115
7	−2.23494	0.42373	42.46393	161.6317	66.72736	100	14.68	100	115
8	−2.06489	0.4017	42.46393	161.6317	73.94836	100	14.68	100	115
9	−1.91086	0.38375	42.46393	161.6317	80.48907	100	14.68	100	115
10	−1.76941	0.36883	42.46393	161.6317	86.4956	100	14.68	100	115
11	−1.63808	0.35628	42.46393	161.6317	92.07239	100	14.68	100	115
12	−1.51501	0.34559	42.46393	161.6317	97.29842	100	14.68	100	115
13	−1.39879	0.33644	42.46393	161.6317	102.2336	102	14.29	100	116
14	−1.28830	0.32855	42.46393	161.6317	106.9254	107	13.95	100	121
15	−1.18263	0.32174	42.46393	161.6317	111.4126	111	13.66	100	125
16	−1.08105	0.31584	42.46393	161.6317	115.7261	116	13.41	103	129
17	−0.98294	0.31073	42.46393	161.6317	119.8922	120	13.19	107	133
18	−0.88779	0.30632	42.46393	161.6317	123.9326	124	13.01	111	137
19	−0.79515	0.30251	42.46393	161.6317	127.8665	128	12.85	115	141
20	−0.70464	0.29925	42.46393	161.6317	131.7099	132	12.71	119	145
21	−0.61594	0.29647	42.46393	161.6317	135.4765	135	12.59	122	148
22	−0.52876	0.29412	42.46393	161.6317	139.1785	139	12.49	127	151
23	−0.44285	0.29217	42.46393	161.6317	142.8265	143	12.41	131	155
24	−0.35797	0.29057	42.46393	161.6317	146.4309	146	12.34	134	158
25	−0.27392	0.28929	42.46393	161.6317	150	150	12.28	138	162
26	−0.19053	0.28832	42.46393	161.6317	153.541	154	12.24	142	166
27	−0.10761	0.28763	42.46393	161.6317	157.0622	157	12.21	145	169
28	−0.02502	0.2872	42.46393	161.6317	160.5693	161	12.2	149	173
29	0.05741	0.28703	42.46393	161.6317	164.0696	164	12.19	152	176
30	0.1398	0.2871	42.46393	161.6317	167.5682	168	12.19	156	180
31	0.22231	0.28742	42.46393	161.6317	171.0719	171	12.2	159	183
32	0.30507	0.28797	42.46393	161.6317	174.5862	175	12.23	163	187
33	0.38821	0.28876	42.46393	161.6317	178.1166	178	12.26	166	190
34	0.47189	0.2898	42.46393	161.6317	181.67	182	12.31	170	194
35	0.55623	0.29109	42.46393	161.6317	185.2514	185	12.36	173	197
36	0.6414	0.29264	42.46393	161.6317	188.8681	189	12.43	177	201

Table I.1: Grade 5 – Operational

Raw Score	Theta	CSEM	Slope	Intercept	Unrounded	Scale Score	Scale CSEM	Lower SS	Upper SS
37	0.72757	0.29447	42.46393	161.6317	192.5272	193	12.5	181	206
38	0.81489	0.29658	42.46393	161.6317	196.2351	196	12.59	183	209
39	0.90355	0.29901	42.46393	161.6317	200	200	12.7	187	213
40	0.99377	0.30177	42.46393	161.6317	203.8311	204	12.81	191	217
41	1.08576	0.30488	42.46393	161.6317	207.7373	208	12.95	195	221
42	1.17977	0.30839	42.46393	161.6317	211.7294	212	13.1	199	225
43	1.27607	0.31233	42.46393	161.6317	215.8186	216	13.26	203	229
44	1.37498	0.31676	42.46393	161.6317	220.0188	220	13.45	207	233
45	1.47687	0.32173	42.46393	161.6317	224.3454	224	13.66	210	238
46	1.58216	0.32735	42.46393	161.6317	228.8164	229	13.9	215	243
47	1.69138	0.33374	42.46393	161.6317	233.4543	233	14.17	219	247
48	1.80517	0.34107	42.46393	161.6317	238.2863	238	14.48	224	252
49	1.92436	0.3496	42.46393	161.6317	243.3476	243	14.85	228	258
50	2.05004	0.3597	42.46393	161.6317	248.6845	249	15.27	234	264
51	2.1837	0.37187	42.46393	161.6317	254.3602	254	15.79	238	270
52	2.32744	0.38691	42.46393	161.6317	260.4639	260	16.43	244	276
53	2.48432	0.40597	42.46393	161.6317	267.1257	267	17.24	250	284
54	2.65895	0.43086	42.46393	161.6317	274.5412	275	18.3	257	293
55	2.85869	0.46459	42.46393	161.6317	283.0229	283	19.73	263	300
56	3.0961	0.51252	42.46393	161.6317	293.1043	293	21.76	271	300
57	3.39492	0.58566	42.46393	161.6317	305.7933	300	24.87	275	300
58	3.8087	0.71208	42.46393	161.6317	323.3641	300	24.87	275	300
59	4.50794	1.00264	42.46393	161.6317	353.0565	300	24.87	275	300
60	5.71542	1.82707	42.46393	161.6317	404.3309	300	24.87	275	300

Table I.2: Grade 5 – Special Equating

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
0	–5.63205	1.83781	42.46393	161.6317	-77.5273	100	14.73	100	115
1	–4.39756	1.02127	42.46393	161.6317	-25.106	100	14.73	100	115
2	–3.66216	0.7359	42.46393	161.6317	6.121994	100	14.73	100	115
3	–3.21576	0.61132	42.46393	161.6317	25.07789	100	14.73	100	115
4	–2.88819	0.53799	42.46393	161.6317	38.9878	100	14.73	100	115
5	–2.62601	0.48854	42.46393	161.6317	50.121	100	14.73	100	115
6	–2.40533	0.45248	42.46393	161.6317	59.49194	100	14.73	100	115
7	–2.21335	0.42482	42.46393	161.6317	67.64416	100	14.73	100	115
8	–2.04237	0.40285	42.46393	161.6317	74.90464	100	14.73	100	115
9	–1.88741	0.38495	42.46393	161.6317	81.48485	100	14.73	100	115
10	–1.74503	0.3701	42.46393	161.6317	87.53087	100	14.73	100	115
11	–1.61276	0.35761	42.46393	161.6317	93.14757	100	14.73	100	115
12	–1.48873	0.34699	42.46393	161.6317	98.41437	100	14.73	100	115
13	–1.37153	0.33791	42.46393	161.6317	103.3911	103	14.35	100	117
14	–1.26004	0.33009	42.46393	161.6317	108.1254	108	14.02	100	122
15	–1.15334	0.32336	42.46393	161.6317	112.6564	113	13.73	100	127
16	–1.05069	0.31754	42.46393	161.6317	117.0153	117	13.48	104	130
17	–0.95149	0.31252	42.46393	161.6317	121.2277	121	13.27	108	134
18	–0.8552	0.3082	42.46393	161.6317	125.3165	125	13.09	112	138
19	–0.76138	0.30448	42.46393	161.6317	129.3005	129	12.93	116	142
20	–0.66966	0.30131	42.46393	161.6317	133.1953	133	12.79	120	146
21	–0.5797	0.29863	42.46393	161.6317	137.0154	137	12.68	124	150
22	–0.49121	0.29637	42.46393	161.6317	140.773	141	12.59	128	154
23	–0.40394	0.29451	42.46393	161.6317	144.4788	144	12.51	131	157
24	–0.31767	0.29301	42.46393	161.6317	148.1422	150	12.44	138	162
25	–0.23217	0.29183	42.46393	161.6317	151.7728	152	12.39	140	164
26	–0.14728	0.29095	42.46393	161.6317	155.3776	155	12.35	143	167
27	–0.06281	0.29036	42.46393	161.6317	158.9645	159	12.33	147	171
28	0.02139	0.29003	42.46393	161.6317	162.54	163	12.32	151	175
29	0.10548	0.28996	42.46393	161.6317	166.1108	166	12.31	154	178
30	0.18959	0.29014	42.46393	161.6317	169.6824	170	12.32	158	182
31	0.27389	0.29057	42.46393	161.6317	173.2621	173	12.34	161	185
32	0.35851	0.29125	42.46393	161.6317	176.8554	177	12.37	165	189
33	0.44359	0.29217	42.46393	161.6317	180.4683	180	12.41	168	192
34	0.52929	0.29335	42.46393	161.6317	184.1074	184	12.46	172	196
35	0.61575	0.2948	42.46393	161.6317	187.7789	188	12.52	175	201
36	0.70316	0.29653	42.46393	161.6317	191.4906	191	12.59	178	204
37	0.79167	0.29855	42.46393	161.6317	195.2491	195	12.68	182	208
38	0.88149	0.30088	42.46393	161.6317	199.0632	200	12.78	187	213
39	0.97281	0.30355	42.46393	161.6317	202.941	203	12.89	190	216
40	1.06586	0.30658	42.46393	161.6317	206.8923	207	13.02	194	220

Table I.2: Grade 5 – Special Equating

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
41	1.16088	0.31001	42.46393	161.6317	210.9272	211	13.16	198	224
42	1.25817	0.31387	42.46393	161.6317	215.0585	215	13.33	202	228
43	1.35802	0.31821	42.46393	161.6317	219.2986	219	13.51	205	233
44	1.46082	0.32311	42.46393	161.6317	223.6639	224	13.72	210	238
45	1.56698	0.32865	42.46393	161.6317	228.1718	228	13.96	214	242
46	1.67703	0.33496	42.46393	161.6317	232.845	233	14.22	219	247
47	1.79162	0.34221	42.46393	161.6317	237.7109	238	14.53	223	253
48	1.91156	0.35065	42.46393	161.6317	242.8041	243	14.89	228	258
49	2.03795	0.36065	42.46393	161.6317	248.1711	248	15.31	233	263
50	2.17227	0.37272	42.46393	161.6317	253.8748	254	15.83	238	270
51	2.31661	0.38764	42.46393	161.6317	260.0041	260	16.46	244	276
52	2.47403	0.40658	42.46393	161.6317	266.6887	267	17.26	250	284
53	2.64912	0.43136	42.46393	161.6317	274.1237	274	18.32	256	292
54	2.84925	0.46496	42.46393	161.6317	282.6221	283	19.74	263	300
55	3.08695	0.51277	42.46393	161.6317	292.7157	293	21.77	271	300
56	3.38599	0.5858	42.46393	161.6317	305.4141	300	24.88	275	300
57	3.7999	0.71215	42.46393	161.6317	322.9904	300	24.88	275	300
58	4.4992	1.00266	42.46393	161.6317	352.6854	300	24.88	275	300
59	5.70665	1.82705	42.46393	161.6317	403.9585	300	24.88	275	300

Table I.3: Grade 8 – Operational

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
0	–5.83055	1.84298	37.78004	184.296	-35.9824	100	16.02	100	116
1	–4.58421	1.02856	37.78004	184.296	11.10436	100	16.02	100	116
2	–3.8363	0.74286	37.78004	184.296	39.36043	100	16.02	100	116
3	–3.38142	0.61687	37.78004	184.296	56.54582	100	16.02	100	116
4	–3.0484	0.54188	37.78004	184.296	69.12733	100	16.02	100	116
5	–2.78307	0.49078	37.78004	184.296	79.1515	100	16.02	100	116
6	–2.56101	0.45317	37.78004	184.296	87.54094	100	16.02	100	116
7	–2.36906	0.42404	37.78004	184.296	94.79282	100	16.02	100	116
8	–2.19931	0.40067	37.78004	184.296	101.206	101	15.14	100	116
9	–2.04659	0.38142	37.78004	184.296	106.9757	107	14.41	100	121
10	–1.90736	0.36524	37.78004	184.296	112.2359	112	13.8	100	126
11	–1.77907	0.3514	37.78004	184.296	117.0827	117	13.28	104	130
12	–1.65984	0.33943	37.78004	184.296	121.5872	122	12.82	109	135
13	–1.54823	0.32894	37.78004	184.296	125.8038	126	12.43	114	138
14	–1.4431	0.31969	37.78004	184.296	129.7756	130	12.08	118	142
15	–1.34356	0.31146	37.78004	184.296	133.5362	134	11.77	122	146
16	–1.24886	0.3041	37.78004	184.296	137.114	137	11.49	126	148
17	–1.15841	0.2975	37.78004	184.296	140.5312	141	11.24	130	152
18	–1.0717	0.29155	37.78004	184.296	143.8071	144	11.01	133	155
19	–0.98828	0.28618	37.78004	184.296	146.9587	147	10.81	136	158
20	–0.90778	0.28133	37.78004	184.296	150	150	10.63	139	161
21	–0.82988	0.27695	37.78004	184.296	152.9431	153	10.46	143	163
22	–0.75429	0.27299	37.78004	184.296	155.7989	156	10.31	146	166
23	–0.68075	0.26942	37.78004	184.296	158.5772	159	10.18	149	169
24	–0.60905	0.2662	37.78004	184.296	161.2861	161	10.06	151	171
25	–0.53896	0.26332	37.78004	184.296	163.9341	164	9.95	154	174
26	–0.47031	0.26074	37.78004	184.296	166.5277	167	9.85	157	177
27	–0.40293	0.25845	37.78004	184.296	169.0733	169	9.76	159	179
28	–0.33667	0.25643	37.78004	184.296	171.5766	172	9.69	162	182
29	–0.27138	0.25465	37.78004	184.296	174.0433	174	9.62	164	184
30	–0.20693	0.25312	37.78004	184.296	176.4782	176	9.56	166	186
31	–0.1432	0.2518	37.78004	184.296	178.8859	179	9.51	169	189
32	-0.08008	0.25071	37.78004	184.296	181.2706	181	9.47	172	190
33	-0.01746	0.24982	37.78004	184.296	183.6364	184	9.44	175	193
34	0.04477	0.24914	37.78004	184.296	185.9874	186	9.41	177	195
35	0.10672	0.24866	37.78004	184.296	188.3279	188	9.39	179	197
36	0.16847	0.24839	37.78004	184.296	190.6608	191	9.38	182	200
37	0.23014	0.24831	37.78004	184.296	192.9907	193	9.38	184	202
38	0.29183	0.24845	37.78004	184.296	195.3213	195	9.39	186	204
39	0.35363	0.2488	37.78004	184.296	197.6562	198	9.4	189	207
40	0.41567	0.24937	37.78004	184.296	200	200	9.42	191	209

Table I.3: Grade 8 – Operational

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
41	0.47804	0.25018	37.78004	184.296	202.3564	202	9.45	193	211
42	0.54089	0.25124	37.78004	184.296	204.7308	205	9.49	196	214
43	0.60433	0.25255	37.78004	184.296	207.1276	207	9.54	197	217
44	0.6685	0.25415	37.78004	184.296	209.552	210	9.6	200	220
45	0.73357	0.25605	37.78004	184.296	212.0103	212	9.67	202	222
46	0.79969	0.25828	37.78004	184.296	214.5083	215	9.76	205	225
47	0.86705	0.26087	37.78004	184.296	217.0532	217	9.86	207	227
48	0.93586	0.26384	37.78004	184.296	219.6528	220	9.97	210	230
49	1.00635	0.26723	37.78004	184.296	222.3159	222	10.1	212	232
50	1.07878	0.2711	37.78004	184.296	225.0524	225	10.24	215	235
51	1.15344	0.27548	37.78004	184.296	227.873	228	10.41	218	238
52	1.23068	0.28043	37.78004	184.296	230.7911	231	10.59	220	242
53	1.31086	0.28602	37.78004	184.296	233.8203	234	10.81	223	245
54	1.39445	0.29232	37.78004	184.296	236.9784	237	11.04	226	248
55	1.48196	0.29944	37.78004	184.296	240.2845	240	11.31	229	251
56	1.574	0.30748	37.78004	184.296	243.7618	244	11.62	232	256
57	1.67131	0.31658	37.78004	184.296	247.4382	247	11.96	235	259
58	1.77478	0.32695	37.78004	184.296	251.3473	251	12.35	239	263
59	1.88551	0.33882	37.78004	184.296	255.5306	256	12.8	243	269
60	2.0049	0.35253	37.78004	184.296	260.0412	260	13.32	247	273
61	2.13476	0.36856	37.78004	184.296	264.9473	265	13.92	251	279
62	2.27752	0.3876	37.78004	184.296	270.3408	270	14.64	255	285
63	2.43655	0.41066	37.78004	184.296	276.349	276	15.51	260	292
64	2.61678	0.43936	37.78004	184.296	283.1581	283	16.6	266	300
65	2.82578	0.47645	37.78004	184.296	291.0541	291	18	273	300
66	3.07626	0.52694	37.78004	184.296	300.5172	300	19.91	280	300
67	3.39195	0.60141	37.78004	184.296	312.444	300	19.91	280	300
68	3.82614	0.72738	37.78004	184.296	328.8477	300	19.91	280	300
69	4.54875	1.01481	37.78004	184.296	356.148	300	19.91	280	300
70	5.7738	1.83403	37.78004	184.296	402.4304	300	19.91	280	300

Table I.4: Grade 11 – Operational

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
0	–5.74013	1.832	52.81895	174.9036	-128.284	100	16.02	100	116
1	–4.5201	1.01143	52.81895	174.9036	-63.8433	100	16.02	100	116
2	–3.80387	0.7233	52.81895	174.9036	-26.0128	100	16.02	100	116
3	–3.37516	0.59721	52.81895	174.9036	-3.36881	100	16.02	100	116
4	–3.06407	0.52296	52.81895	174.9036	13.06264	100	16.02	100	116
5	–2.81736	0.47292	52.81895	174.9036	26.0936	100	16.02	100	116
6	–2.6113	0.43645	52.81895	174.9036	36.97748	100	16.02	100	116
7	–2.43324	0.40846	52.81895	174.9036	46.38242	100	16.02	100	116
8	–2.27564	0.3862	52.81895	174.9036	54.70668	100	16.02	100	116
9	–2.13363	0.368	52.81895	174.9036	62.2075	100	16.02	100	116
10	–2.00387	0.35282	52.81895	174.9036	69.06129	100	16.02	100	116
11	–1.884	0.33993	52.81895	174.9036	75.3927	100	16.02	100	116
12	–1.77226	0.32885	52.81895	174.9036	81.29469	100	16.02	100	116
13	–1.66733	0.3192	52.81895	174.9036	86.83698	100	16.02	100	116
14	–1.56818	0.31074	52.81895	174.9036	92.07398	100	16.02	100	116
15	–1.47397	0.30324	52.81895	174.9036	97.05005	100	16.02	100	116
16	–1.38407	0.29656	52.81895	174.9036	101.7985	102	15.66	100	118
17	–1.29792	0.29056	52.81895	174.9036	106.3488	106	15.35	100	121
18	–1.21508	0.28516	52.81895	174.9036	110.7244	111	15.06	100	126
19	–1.13517	0.28026	52.81895	174.9036	114.9451	115	14.8	100	130
20	–1.05789	0.27581	52.81895	174.9036	119.027	119	14.57	104	134
21	–0.98295	0.27174	52.81895	174.9036	122.9852	123	14.35	109	137
22	–0.91013	0.26802	52.81895	174.9036	126.8315	127	14.16	113	141
23	–0.83922	0.2646	52.81895	174.9036	130.5769	131	13.98	117	145
24	–0.77005	0.26146	52.81895	174.9036	134.2304	134	13.81	120	148
25	–0.70245	0.25856	52.81895	174.9036	137.8009	138	13.66	124	152
26	–0.6363	0.25589	52.81895	174.9036	141.2949	141	13.52	127	155
27	–0.57146	0.25343	52.81895	174.9036	144.7197	145	13.39	132	158
28	–0.50781	0.25116	52.81895	174.9036	148.0816	148	13.27	135	161
29	–0.44526	0.24907	52.81895	174.9036	151.3854	151	13.16	138	164
30	–0.38371	0.24715	52.81895	174.9036	154.6364	155	13.05	142	168
31	–0.32307	0.24539	52.81895	174.9036	157.8394	158	12.96	145	171
32	–0.26325	0.2438	52.81895	174.9036	160.999	161	12.88	148	174
33	–0.20417	0.24236	52.81895	174.9036	164.1196	164	12.8	151	177
34	–0.14575	0.24107	52.81895	174.9036	167.2052	167	12.73	154	180
35	–0.08791	0.23993	52.81895	174.9036	170.2603	170	12.67	157	183
36	–0.03059	0.23895	52.81895	174.9036	173.2879	173	12.62	160	186
37	0.0263	0.23812	52.81895	174.9036	176.2927	176	12.58	163	189
38	0.08284	0.23744	52.81895	174.9036	179.2791	179	12.54	166	192
39	0.13909	0.23693	52.81895	174.9036	182.2502	182	12.51	169	195
40	0.19514	0.23658	52.81895	174.9036	185.2107	185	12.5	173	198

Table I.4: Grade 11 – Operational

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
41	0.25106	0.2364	52.81895	174.9036	188.1643	188	12.49	176	200
42	0.30694	0.2364	52.81895	174.9036	191.1158	191	12.49	179	203
43	0.36285	0.23657	52.81895	174.9036	194.069	194	12.5	182	207
44	0.41889	0.23692	52.81895	174.9036	197.0289	197	12.51	184	210
45	0.47514	0.23746	52.81895	174.9036	200	200	12.54	187	213
46	0.5317	0.23819	52.81895	174.9036	202.9874	203	12.58	190	216
47	0.58865	0.23913	52.81895	174.9036	205.9955	206	12.63	193	219
48	0.6461	0.24027	52.81895	174.9036	209.0299	209	12.69	196	222
49	0.70415	0.24163	52.81895	174.9036	212.0961	212	12.76	199	225
50	0.7629	0.24321	52.81895	174.9036	215.1992	215	12.85	202	228
51	0.82249	0.24504	52.81895	174.9036	218.3467	218	12.94	205	231
52	0.88303	0.24711	52.81895	174.9036	221.5443	222	13.05	209	235
53	0.94467	0.24946	52.81895	174.9036	224.8001	225	13.18	212	238
54	1.00755	0.2521	52.81895	174.9036	228.1213	228	13.32	215	241
55	1.07183	0.25505	52.81895	174.9036	231.5165	232	13.47	219	245
56	1.13772	0.25835	52.81895	174.9036	234.9968	235	13.65	221	249
57	1.2054	0.26203	52.81895	174.9036	238.5716	239	13.84	225	253
58	1.27512	0.26613	52.81895	174.9036	242.2541	242	14.06	228	256
59	1.34714	0.27069	52.81895	174.9036	246.0581	246	14.3	232	260
60	1.42177	0.27577	52.81895	174.9036	250	250	14.57	235	265
61	1.49936	0.28143	52.81895	174.9036	254.0982	254	14.86	239	269
62	1.58032	0.28776	52.81895	174.9036	258.3744	258	15.2	243	273
63	1.66515	0.29485	52.81895	174.9036	262.8551	263	15.57	247	279
64	1.75441	0.30282	52.81895	174.9036	267.5697	268	15.99	252	284
65	1.8488	0.31181	52.81895	174.9036	272.5553	273	16.47	257	289
66	1.94918	0.32203	52.81895	174.9036	277.8572	278	17.01	261	295
67	2.05661	0.33374	52.81895	174.9036	283.5316	284	17.63	266	300
68	2.17245	0.34727	52.81895	174.9036	289.6501	290	18.34	272	300
69	2.29848	0.36313	52.81895	174.9036	296.3069	296	19.18	277	300
70	2.43711	0.382	52.81895	174.9036	303.6292	300	20.18	280	300
71	2.59166	0.40494	52.81895	174.9036	311.7924	300	20.18	280	300
72	2.76704	0.43359	52.81895	174.9036	321.0557	300	20.18	280	300
73	2.97081	0.47071	52.81895	174.9036	331.8187	300	20.18	280	300
74	3.21565	0.52137	52.81895	174.9036	344.7509	300	20.18	280	300
75	3.5253	0.59619	52.81895	174.9036	361.1062	300	20.18	280	300
76	3.95303	0.72279	52.81895	174.9036	383.6985	300	20.18	280	300
77	4.66878	1.01133	52.81895	174.9036	421.5037	300	20.18	280	300
78	5.88877	1.83202	52.81895	174.9036	485.9422	300	20.18	280	300

Table I.5: Grade 11 – Computer-Based Spanish

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
0	–5.73455	1.83206	52.81895	174.9036	–127.989	100	16.05	100	116
1	–4.51436	1.01155	52.81895	174.9036	–63.5402	100	16.05	100	116
2	–3.7979	0.72346	52.81895	174.9036	–25.6975	100	16.05	100	116
3	–3.36895	0.59742	52.81895	174.9036	–3.0408	100	16.05	100	116
4	–3.05761	0.52321	52.81895	174.9036	13.40385	100	16.05	100	116
5	–2.81063	0.4732	52.81895	174.9036	26.44907	100	16.05	100	116
6	–2.60431	0.43676	52.81895	174.9036	37.34668	100	16.05	100	116
7	–2.42598	0.40881	52.81895	174.9036	46.76588	100	16.05	100	116
8	–2.26809	0.38657	52.81895	174.9036	55.10547	100	16.05	100	116
9	–2.12578	0.36841	52.81895	174.9036	62.62213	100	16.05	100	116
10	–1.99572	0.35325	52.81895	174.9036	69.49177	100	16.05	100	116
11	–1.87554	0.34039	52.81895	174.9036	75.83955	100	16.05	100	116
12	–1.76348	0.32934	52.81895	174.9036	81.75844	100	16.05	100	116
13	–1.65823	0.31973	52.81895	174.9036	87.31763	100	16.05	100	116
14	–1.55873	0.31129	52.81895	174.9036	92.57312	100	16.05	100	116
15	–1.46418	0.30382	52.81895	174.9036	97.56715	100	16.05	100	116
16	–1.37392	0.29716	52.81895	174.9036	102.3346	102	15.7	100	118
17	–1.28741	0.2912	52.81895	174.9036	106.904	107	15.38	100	122
18	–1.20419	0.28582	52.81895	174.9036	111.2995	111	15.1	100	126
19	–1.12391	0.28095	52.81895	174.9036	115.5399	116	14.84	101	131
20	–1.04623	0.27652	52.81895	174.9036	119.6428	120	14.61	105	135
21	–0.97089	0.27248	52.81895	174.9036	123.6222	124	14.39	110	138
22	–0.89766	0.26879	52.81895	174.9036	127.4901	127	14.2	113	141
23	–0.82634	0.26539	52.81895	174.9036	131.2572	131	14.02	117	145
24	–0.75674	0.26228	52.81895	174.9036	134.9334	135	13.85	121	149
25	–0.68871	0.25941	52.81895	174.9036	138.5267	139	13.7	125	153
26	–0.62211	0.25676	52.81895	174.9036	142.0444	142	13.56	128	156
27	–0.55681	0.25433	52.81895	174.9036	145.4935	145	13.43	132	158
28	–0.49271	0.25209	52.81895	174.9036	148.8792	149	13.32	136	162
29	–0.42968	0.25003	52.81895	174.9036	152.2084	152	13.21	139	165
30	–0.36765	0.24814	52.81895	174.9036	155.4847	155	13.11	142	168
31	–0.30651	0.24642	52.81895	174.9036	158.7141	159	13.02	146	172
32	–0.24617	0.24486	52.81895	174.9036	161.9012	162	12.93	149	175
33	–0.18657	0.24346	52.81895	174.9036	165.0492	165	12.86	152	178
34	–0.12761	0.24221	52.81895	174.9036	168.1634	168	12.79	155	181
35	–0.06921	0.24112	52.81895	174.9036	171.248	171	12.74	158	184
36	–0.01131	0.24018	52.81895	174.9036	174.3062	174	12.69	161	187
37	0.04619	0.23941	52.81895	174.9036	177.3433	177	12.65	164	190
38	0.10336	0.2388	52.81895	174.9036	180.363	180	12.61	167	193
39	0.16027	0.23836	52.81895	174.9036	183.3689	183	12.59	170	196
40	0.21701	0.23808	52.81895	174.9036	186.3658	186	12.58	173	199

Table I.5: Grade 11 – Computer-Based Spanish

Raw Score	Theta	Standard Error	Slope	Intercept	Unrounded	Scale Score	Scale SE	Lower SS	Upper SS
41	0.27367	0.23799	52.81895	174.9036	189.3586	189	12.57	176	202
42	0.33032	0.23807	52.81895	174.9036	192.3508	192	12.57	179	205
43	0.38705	0.23834	52.81895	174.9036	195.3472	195	12.59	182	208
44	0.44396	0.2388	52.81895	174.9036	198.3531	200	12.61	187	213
45	0.50113	0.23946	52.81895	174.9036	201.3728	201	12.65	188	214
46	0.55868	0.24032	52.81895	174.9036	204.4125	204	12.69	191	217
47	0.61668	0.2414	52.81895	174.9036	207.476	207	12.75	194	220
48	0.67526	0.24269	52.81895	174.9036	210.5701	211	12.82	198	224
49	0.73452	0.24422	52.81895	174.9036	213.7002	214	12.9	201	227
50	0.79459	0.24599	52.81895	174.9036	216.873	217	12.99	204	230
51	0.85559	0.24801	52.81895	174.9036	220.095	220	13.1	207	233
52	0.91766	0.25031	52.81895	174.9036	223.3734	223	13.22	210	236
53	0.98095	0.2529	52.81895	174.9036	226.7163	227	13.36	214	240
54	1.04563	0.25581	52.81895	174.9036	230.1327	230	13.51	216	244
55	1.11189	0.25907	52.81895	174.9036	233.6325	234	13.68	220	248
56	1.17994	0.26271	52.81895	174.9036	237.2268	237	13.88	223	251
57	1.25	0.26677	52.81895	174.9036	240.9273	241	14.09	227	255
58	1.32236	0.27129	52.81895	174.9036	244.7493	245	14.33	231	259
59	1.39731	0.27634	52.81895	174.9036	248.708	250	14.6	235	265
60	1.47522	0.28198	52.81895	174.9036	252.8232	253	14.89	238	268
61	1.55649	0.28829	52.81895	174.9036	257.1158	257	15.23	242	272
62	1.64161	0.29536	52.81895	174.9036	261.6117	262	15.6	246	278
63	1.73117	0.30331	52.81895	174.9036	266.3422	266	16.02	250	282
64	1.82586	0.3123	52.81895	174.9036	271.3436	271	16.5	255	288
65	1.92655	0.32251	52.81895	174.9036	276.6619	277	17.03	260	294
66	2.03429	0.33421	52.81895	174.9036	282.3527	282	17.65	264	300
67	2.15045	0.34774	52.81895	174.9036	288.4881	288	18.37	270	300
68	2.27681	0.36359	52.81895	174.9036	295.1623	295	19.2	276	300
69	2.41578	0.38246	52.81895	174.9036	302.5026	300	20.2	280	300
70	2.57069	0.4054	52.81895	174.9036	310.6847	300	20.2	280	300
71	2.74645	0.43404	52.81895	174.9036	319.9682	300	20.2	280	300
72	2.95061	0.47114	52.81895	174.9036	330.7517	300	20.2	280	300
73	3.19587	0.52178	52.81895	174.9036	343.7061	300	20.2	280	300
74	3.50595	0.59656	52.81895	174.9036	360.0842	300	20.2	280	300
75	3.93415	0.72312	52.81895	174.9036	382.7013	300	20.2	280	300
76	4.65036	1.01157	52.81895	174.9036	420.5307	300	20.2	280	300
77	5.87074	1.83218	52.81895	174.9036	484.9899	300	20.2	280	300

APPENDIX J: Raw Score-to-Theta Subscore Tables

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

Raw Score	Theta	CSEM	Lower	Upper	Level
0	−4.77171	1.86696	−7.57215	−1.97127	Below
1	−3.46317	1.07081	−5.06939	−1.85696	Below
2	−2.62678	0.80015	−3.82701	−1.42656	Below
3	−2.08324	0.68532	−3.11122	−1.05526	Below
4	−1.66019	0.62029	−2.59063	−0.72976	Below
5	−1.30222	0.57896	−2.17066	−0.43378	Below
6	−0.98375	0.55139	−1.81084	−0.15667	Below
7	−0.69050	0.53288	−1.48982	0.10882	Below
8	−0.41335	0.52095	−1.19478	0.36807	Below
9	−0.14590	0.51412	−0.91708	0.62528	Below
10	0.11675	0.51149	−0.65049	0.88398	Below
11	0.37858	0.51245	−0.39010	1.14725	Near/Met
12	0.64300	0.51647	−0.13171	1.41770	Near/Met
13	0.91295	0.52303	0.12840	1.69749	Near/Met
14	1.19083	0.53151	0.39356	1.98809	Near/Met
15	1.47858	0.54164	0.66612	2.29104	Near/Met
16	1.77873	0.55484	0.94647	2.61099	Above
17	2.09759	0.57669	1.23255	2.96262	Above
18	2.45286	0.62057	1.52200	3.38371	Above
19	2.89269	0.71845	1.81501	3.97036	Above
20	3.58221	0.98788	2.10039	5.06403	Above
21	4.75938	1.81156	2.04204	7.47672	Above

Table J.2: Grade 5 Life Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.52455	1.84494	–7.29196	–1.75714	Below
1	–3.27196	1.03369	–4.8225	–1.72143	Below
2	–2.51133	0.75277	–3.64049	–1.38218	Below
3	–2.03955	0.63209	–2.98769	–1.09142	Below
4	–1.68526	0.56326	–2.53015	–0.84037	Below
5	–1.39371	0.51942	–2.17284	–0.61458	Below
6	–1.13972	0.49032	–1.8752	–0.40424	Below
7	–0.9093	0.47102	–1.61583	–0.20277	Below
8	–0.69364	0.45877	–1.3818	–0.00549	Below
9	–0.48666	0.45191	–1.16453	0.19120	Below
10	–0.28388	0.44939	–0.95797	0.39020	Below
11	–0.08164	0.4506	–0.75754	0.59426	Below
12	0.12324	0.45524	–0.55962	0.8061	Below
13	0.33391	0.46331	–0.36106	1.02887	Near/Met
14	0.55378	0.47516	–0.15896	1.26652	Near/Met
15	0.78703	0.49156	0.04969	1.52437	Near/Met
16	1.03924	0.51391	0.26837	1.81010	Near/Met
17	1.31857	0.54468	0.50155	2.13559	Near/Met
18	1.63823	0.5886	0.75533	2.52113	Near/Met
19	2.02244	0.65577	1.03878	3.00609	Above
20	2.5249	0.77288	1.36558	3.68422	Above
21	3.31585	1.04799	1.74386	4.88783	Above
22	4.58889	1.85267	1.80988	7.36789	Above

Table J.3: Grade 5 Physical Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.19997	1.86021	–6.99029	–1.40966	Below
1	–2.90584	1.06326	–4.50073	–1.31095	Below
2	–2.07982	0.79697	–3.27528	–0.88437	Below
3	–1.53703	0.68756	–2.56837	–0.50569	Below
4	–1.10799	0.6271	–2.04864	–0.16734	Below
5	–0.73981	0.5889	–1.62316	0.14354	Below
6	–0.40894	0.56301	–1.25346	0.43557	Below
7	–0.10255	0.54511	–0.92022	0.71511	Below
8	0.18774	0.5334	–0.61236	0.98784	Near/Met
9	0.46858	0.52746	–0.32261	1.25977	Near/Met
10	0.74636	0.52775	–0.04527	1.53798	Near/Met
11	1.02823	0.53545	0.22505	1.83140	Near/Met
12	1.3233	0.55279	0.49411	2.15248	Near/Met
13	1.64495	0.5843	0.7685	2.5214	Near/Met
14	2.01676	0.64033	1.05626	2.97725	Above
15	2.49109	0.74888	1.36777	3.61441	Above
16	3.23657	1.02161	1.70415	4.76898	Above
17	4.46688	1.83439	1.71529	7.21846	Above

Table J.4: Grade 5 Sensemaking Score Table

–	–	–	–	–	–
0	–5.15316	1.8501	–7.92831	–2.37801	Below
1	–3.8876	1.04234	–5.45111	–2.32409	Below
2	–3.1099	0.76339	–4.25499	–1.96482	Below
3	–2.62318	0.64282	–3.58741	–1.65895	Below
4	–2.25673	0.57252	–3.11551	–1.39795	Below
5	–1.95652	0.52581	–2.74524	–1.16781	Below
6	–1.69804	0.4926	–2.43694	–0.95914	Below
7	–1.46778	0.46821	–2.1701	–0.76547	Below
8	–1.25734	0.45013	–1.93254	–0.58215	Below
9	–1.06094	0.43689	–1.71628	–0.40561	Below
10	–0.87437	0.42755	–1.5157	–0.23305	Below
11	–0.69437	0.42146	–1.32656	–0.06218	Below
12	–0.51832	0.41813	–1.14552	0.10887	Below
13	–0.34403	0.4172	–0.96983	0.28177	Below
14	–0.16963	0.41838	–0.7972	0.45794	Below
15	0.00659	0.42148	–0.62563	0.63881	Below
16	0.18617	0.42636	–0.45337	0.82571	Below
17	0.37065	0.43294	–0.27876	1.02006	Near/Met
18	0.56156	0.44117	–0.1002	1.22331	Near/Met
19	0.76045	0.45102	0.08392	1.43698	Near/Met
20	0.96894	0.46241	0.27532	1.66255	Near/Met
21	1.18864	0.47523	0.47579	1.90148	Near/Met
22	1.42118	0.48942	0.68705	2.15531	Near/Met
23	1.66845	0.50546	0.91026	2.42664	Above
24	1.93368	0.52547	1.14547	2.72188	Above
25	2.22439	0.5551	1.39174	3.05704	Above
26	2.55911	0.60703	1.64856	3.46965	Above
27	2.98628	0.7124	1.91768	4.05488	Above
28	3.67137	0.98786	2.18958	5.15316	Above
29	4.85071	1.81324	2.13085	7.57057	Above

Table J.5: Grade 5 Critiquing Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–3.9153	1.85859	–6.70319	–1.12742	Below
1	–2.62491	1.06111	–4.21658	–1.03325	Below
2	–1.80198	0.79609	–2.99612	–0.60785	Below
3	–1.2584	0.68992	–2.29328	–0.22352	Below
4	–0.82306	0.63477	–1.77522	0.12909	Below
5	–0.4411	0.6042	–1.3474	0.4652	Below
6	–0.08662	0.58857	–0.96948	0.79623	Below
7	0.2561	0.58395	–0.61983	1.13202	Near/Met
8	0.59921	0.58923	–0.28464	1.48305	Near/Met
9	0.95482	0.60542	0.04669	1.86295	Near/Met
10	1.33856	0.63639	0.38397	2.29314	Near/Met
11	1.77618	0.6917	0.73863	2.81373	Near/Met
12	2.32234	0.79776	1.1257	3.51898	Above
13	3.14794	1.06236	1.5544	4.74148	Above
14	4.44014	1.85927	1.65123	7.22904	Above

Table J.6: Grade 5 Investigating Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.01408	1.85907	–6.80269	–1.22548	Below
1	–2.72459	1.05906	–4.31318	–1.136	Below
2	–1.91024	0.78805	–3.09232	–0.72817	Below
3	–1.38347	0.67464	–2.39543	–0.37151	Below
4	–0.97306	0.61144	–1.89022	–0.0559	Below
5	–0.62456	0.57187	–1.48237	0.23324	Below
6	–0.31316	0.5459	–1.13201	0.50569	Below
7	–0.02498	0.52899	–0.81847	0.76850	Below
8	0.24907	0.51905	–0.52951	1.02764	Near/Met
9	0.51603	0.51533	–0.25697	1.28902	Near/Met
10	0.78239	0.51799	0.00540	1.55937	Near/Met
11	1.05522	0.52802	0.26319	1.84725	Near/Met
12	1.34347	0.54757	0.52211	2.16482	Near/Met
13	1.66041	0.58114	0.7887	2.53212	Near/Met
14	2.0296	0.63913	1.07090	2.98829	Above
15	2.50363	0.74959	1.37924	3.62801	Above
16	3.25204	1.02414	1.71583	4.78825	Above
17	4.4876	1.8372	1.7318	7.2434	Above

Table J.7: Grade 5 Earth and Space Science Score Table – Special Equating

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.77171	1.86696	–7.57215	–1.97127	Below
1	–3.46317	1.07081	–5.06939	–1.85696	Below
2	–2.62678	0.80015	–3.82701	–1.42656	Below
3	–2.08324	0.68532	–3.11122	–1.05526	Below
4	–1.66019	0.62029	–2.59063	–0.72976	Below
5	–1.30222	0.57896	–2.17066	–0.43378	Below
6	–0.98375	0.55139	–1.81084	–0.15667	Below
7	–0.6905	0.53288	–1.48982	0.10882	Below
8	–0.41335	0.52095	–1.19478	0.36807	Below
9	–0.1459	0.51412	–0.91708	0.62528	Below
10	0.11675	0.51149	–0.65049	0.88398	Below
11	0.37858	0.51245	–0.3901	1.14725	Near/Met
12	0.643	0.51647	–0.13171	1.41770	Near/Met
13	0.91295	0.52303	0.12840	1.69749	Near/Met
14	1.19083	0.53151	0.39356	1.98809	Near/Met
15	1.47858	0.54164	0.66612	2.29104	Near/Met
16	1.77873	0.55484	0.94647	2.61099	Above
17	2.09759	0.57669	1.23255	2.96262	Above
18	2.45286	0.62057	1.52200	3.38371	Above
19	2.89269	0.71845	1.81501	3.97036	Above
20	3.58221	0.98788	2.10039	5.06403	Above
21	4.75938	1.81156	2.04204	7.47672	Above

Table J.8: Grade 5 Life Science Score Table – Special Equating

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.47758	1.84616	–7.24682	–1.70834	Below
1	–3.22194	1.03572	–4.77552	–1.66836	Below
2	–2.45727	0.75538	–3.59034	–1.3242	Below
3	–1.98159	0.63519	–2.93438	–1.02881	Below
4	–1.62326	0.56698	–2.47373	–0.77279	Below
5	–1.32723	0.52401	–2.11325	–0.54122	Below
6	–1.06808	0.49599	–1.81207	–0.3241	Below
7	–0.83156	0.47799	–1.54855	–0.11458	Below
8	–0.60868	0.46724	–1.30954	0.09218	Below
9	–0.39317	0.46204	–1.08623	0.29989	Below
10	–0.18032	0.46138	–0.87239	0.51175	Below
11	0.03378	0.46466	–0.66321	0.73077	Below
12	0.25269	0.47173	–0.45491	0.96028	Near/Met
13	0.48013	0.48277	–0.24403	1.20428	Near/Met
14	0.72043	0.49847	–0.02728	1.46813	Near/Met
15	0.97929	0.52016	0.19905	1.75953	Near/Met
16	1.26492	0.5503	0.43947	2.09037	Near/Met
17	1.59059	0.59356	0.70025	2.48093	Near/Met
18	1.98053	0.66001	0.99051	2.97054	Above
19	2.48844	0.77631	1.32397	3.65290	Above
20	3.28456	1.05038	1.70899	4.86013	Above
21	4.56103	1.85396	1.78009	7.34197	Above

Table J.9: Grade 5 Physical Science Score Table – Special Equating

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.19997	1.86021	–6.99029	–1.40966	Below
1	–2.90584	1.06326	–4.50073	–1.31095	Below
2	–2.07982	0.79697	–3.27528	–0.88437	Below
3	–1.53703	0.68756	–2.56837	–0.50569	Below
4	–1.10799	0.6271	–2.04864	–0.16734	Below
5	–0.73981	0.5889	–1.62316	0.14354	Below
6	–0.40894	0.56301	–1.25346	0.43557	Below
7	–0.10255	0.54511	–0.92022	0.71511	Below
8	0.18774	0.5334	–0.61236	0.98784	Near/Met
9	0.46858	0.52746	–0.32261	1.25977	Near/Met
10	0.74636	0.52775	–0.04527	1.53798	Near/Met
11	1.02823	0.53545	0.22505	1.83140	Near/Met
12	1.3233	0.55279	0.49411	2.15248	Near/Met
13	1.64495	0.5843	0.7685	2.5214	Near/Met
14	2.01676	0.64033	1.05626	2.97725	Above
15	2.49109	0.74888	1.36777	3.61441	Above
16	3.23657	1.02161	1.70415	4.76898	Above
17	4.46688	1.83439	1.71529	7.21846	Above

Table J.10: Grade 5 Sensemaking Score Table – Special Equating

Raw Score	Theta	CSEM	Lower	Upper	Level
0	−5.15316	1.8501	−7.92831	−2.37801	Below
1	−3.8876	1.04234	−5.45111	−2.32409	Below
2	−3.1099	0.76339	−4.25499	−1.96482	Below
3	−2.62318	0.64282	−3.58741	−1.65895	Below
4	−2.25673	0.57252	−3.11551	−1.39795	Below
5	−1.95652	0.52581	−2.74524	−1.16781	Below
6	−1.69804	0.4926	−2.43694	−0.95914	Below
7	−1.46778	0.46821	−2.1701	−0.76547	Below
8	−1.25734	0.45013	−1.93254	−0.58215	Below
9	−1.06094	0.43689	−1.71628	−0.40561	Below
10	−0.87437	0.42755	−1.5157	−0.23305	Below
11	−0.69437	0.42146	−1.32656	−0.06218	Below
12	−0.51832	0.41813	−1.14552	0.10887	Below
13	−0.34403	0.4172	−0.96983	0.28177	Below
14	−0.16963	0.41838	−0.7972	0.45794	Below
15	0.00659	0.42148	−0.62563	0.63881	Below
16	0.18617	0.42636	−0.45337	0.82571	Below
17	0.37065	0.43294	−0.27876	1.02006	Near/Met
18	0.56156	0.44117	−0.1002	1.22331	Near/Met
19	0.76045	0.45102	0.08392	1.43698	Near/Met
20	0.96894	0.46241	0.27532	1.66255	Near/Met
21	1.18864	0.47523	0.47579	1.90148	Near/Met
22	1.42118	0.48942	0.68705	2.15531	Near/Met
23	1.66845	0.50546	0.91026	2.42664	Above
24	1.93368	0.52547	1.14547	2.72188	Above
25	2.22439	0.5551	1.39174	3.05704	Above
26	2.55911	0.60703	1.64856	3.46965	Above
27	2.98628	0.7124	1.91768	4.05488	Above
28	3.67137	0.98786	2.18958	5.15316	Above
29	4.85071	1.81324	2.13085	7.57057	Above

Table J.11: Grade 5 Critiquing Score Table – Special Equating

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–3.9153	1.85859	–6.70319	–1.12742	Below
1	–2.62491	1.06111	–4.21658	–1.03325	Below
2	–1.80198	0.79609	–2.99612	–0.60785	Below
3	–1.2584	0.68992	–2.29328	–0.22352	Below
4	–0.82306	0.63477	–1.77522	0.12909	Below
5	–0.4411	0.6042	–1.3474	0.4652	Below
6	–0.08662	0.58857	–0.96948	0.79623	Below
7	0.2561	0.58395	–0.61983	1.13202	Near/Met
8	0.59921	0.58923	–0.28464	1.48305	Near/Met
9	0.95482	0.60542	0.04669	1.86295	Near/Met
10	1.33856	0.63639	0.38397	2.29314	Near/Met
11	1.77618	0.6917	0.73863	2.81373	Near/Met
12	2.32234	0.79776	1.1257	3.51898	Above
13	3.14794	1.06236	1.5544	4.74148	Above
14	4.44014	1.85927	1.65123	7.22904	Above

Table J.12: Grade 5 Investigating Score Table – Special Equating

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–3.93397	1.86281	–6.72819	–1.13976	Below
1	–2.63487	1.06543	–4.23302	–1.03673	Below
2	–1.80729	0.79621	–3.00161	–0.61298	Below
3	–1.26786	0.68378	–2.29353	–0.24219	Below
4	–0.84525	0.62117	–1.77701	0.08650	Below
5	–0.48491	0.58203	–1.35796	0.38813	Below
6	–0.16181	0.55654	–0.99662	0.673	Below
7	0.13832	0.54046	–0.67237	0.94901	Near/Met
8	0.42525	0.53207	–0.37286	1.22335	Near/Met
9	0.70715	0.53108	–0.08947	1.50377	Near/Met
10	0.99221	0.5382	0.18491	1.79951	Near/Met
11	1.29016	0.5554	0.45706	2.12326	Near/Met
12	1.61482	0.58702	0.73429	2.49535	Near/Met
13	1.99008	0.64328	1.02516	2.955	Above
14	2.46871	0.75216	1.34047	3.59695	Above
15	3.22027	1.02535	1.68224	4.75829	Above
16	4.45728	1.83767	1.70077	7.21378	Above

Table J.13: Grade 8 Earth and Space Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.60421	1.8507	–7.38026	–1.82816	Below
1	–3.33546	1.04565	–4.90394	–1.76699	Below
2	–2.54761	0.77209	–3.70575	–1.38948	Below
3	–2.04445	0.65778	–3.03112	–1.05778	Below
4	–1.65564	0.59412	–2.54682	–0.76446	Below
5	–1.32756	0.55404	–2.15862	–0.4965	Below
6	–1.03619	0.52709	–1.82683	–0.24556	Below
7	–0.76876	0.50818	–1.53103	–0.00649	Below
8	–0.51779	0.49439	–1.25938	0.22379	Below
9	-0.27876	0.48384	–1.00452	0.447	Near/Met
10	-0.04885	0.47541	–0.76197	0.66426	Near/Met
11	0.17386	0.46875	–0.52927	0.87698	Near/Met
12	0.39134	0.46444	–0.30532	1.088	Near/Met
13	0.60647	0.46402	–0.08956	1.3025	Near/Met
14	0.8239	0.46983	0.11915	1.52864	Near/Met
15	1.05089	0.48482	0.32366	1.77812	Near/Met
16	1.29852	0.51289	0.52918	2.06785	Above
17	1.58435	0.55981	0.74463	2.42406	Above
18	1.93851	0.63576	0.98487	2.89215	Above
19	2.42178	0.76548	1.27356	3.57	Above
20	3.2109	1.05194	1.63299	4.78881	Above
21	4.49486	1.85881	1.70664	7.28307	Above

Table J.14: Grade 8 Life Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.94304	1.89401	–7.78406	–2.10203	Below
1	–3.57406	1.10285	–5.22834	–1.91979	Below
2	–2.69172	0.8151	–3.91437	–1.46907	Below
3	–2.14109	0.67962	–3.16052	–1.12166	Below
4	–1.73711	0.59647	–2.63182	–0.84241	Below
5	–1.41592	0.53985	–2.2257	–0.60615	Below
6	–1.14687	0.49951	–1.89614	–0.39761	Below
7	–0.91234	0.47053	–1.61814	–0.20655	Below
8	–0.70095	0.45024	–1.37631	–0.02559	Below
9	–0.50461	0.43699	–1.1601	0.15087	Below
10	–0.31725	0.42961	–0.96167	0.32716	Below
11	–0.13403	0.42725	–0.77491	0.50684	Near/Met
12	0.04908	0.42924	–0.59478	0.69294	Near/Met
13	0.23557	0.43508	–0.41705	0.88819	Near/Met
14	0.42874	0.44452	–0.23804	1.09552	Near/Met
15	0.63193	0.45763	–0.05452	1.31837	Near/Met
16	0.84904	0.475	0.13654	1.56154	Near/Met
17	1.0852	0.49793	0.33830	1.83209	Near/Met
18	1.34807	0.52899	0.55458	2.14155	Above
19	1.65033	0.57306	0.79074	2.50992	Above
20	2.01571	0.64058	1.05484	2.97658	Above
21	2.49751	0.75873	1.35941	3.63560	Above
22	3.26571	1.03656	1.71087	4.82055	Above
23	4.52152	1.84581	1.75280	7.29023	Above

Table J.15: Grade 8 Physical Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.50239	1.8522	–7.28069	–1.72409	Below
1	–3.23195	1.04533	–4.79995	–1.66396	Below
2	–2.44888	0.7664	–3.59848	–1.29928	Below
3	–1.95816	0.64549	–2.9264	–0.98993	Below
4	–1.58875	0.57467	–2.45076	–0.72675	Below
5	–1.28661	0.52711	–2.07728	–0.49595	Below
6	–1.02745	0.4925	–1.7662	–0.2887	Below
7	–0.79823	0.46602	–1.49726	–0.0992	Below
8	–0.591	0.44515	–1.25873	0.07672	Below
9	–0.40042	0.42855	–1.04325	0.24240	Below
10	–0.22252	0.41556	–0.84586	0.40082	Below
11	–0.05404	0.4059	–0.66289	0.55481	Near/Met
12	0.10791	0.39949	–0.49133	0.70714	Near/Met
13	0.26603	0.39635	–0.3285	0.86055	Near/Met
14	0.42299	0.39657	–0.17187	1.01784	Near/Met
15	0.58148	0.40024	–0.01888	1.18184	Near/Met
16	0.74435	0.4075	0.1331	1.3556	Near/Met
17	0.91469	0.41858	0.28682	1.54256	Near/Met
18	1.09606	0.43385	0.44528	1.74683	Above
19	1.29276	0.45399	0.61177	1.97374	Above
20	1.51049	0.48028	0.79007	2.23091	Above
21	1.75743	0.5151	0.98478	2.53008	Above
22	2.04676	0.56313	1.20206	2.89145	Above
23	2.40262	0.63454	1.45081	3.35443	Above
24	2.87865	0.75621	1.74433	4.01296	Above
25	3.64504	1.03667	2.09003	5.20004	Above
26	4.90204	1.84665	2.13206	7.67201	Above

Table J.16: Grade 8 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–5.45902	1.85898	–8.24749	–2.67055	Below
1	–4.17265	1.05506	–5.75524	–2.59006	Below
2	–3.37199	0.77615	–4.53622	–2.20777	Below
3	–2.86819	0.65427	–3.8496	–1.88679	Below
4	–2.48856	0.5827	–3.36261	–1.61451	Below
5	–2.1777	0.53493	–2.9801	–1.37531	Below
6	–1.91036	0.50074	–2.66147	–1.15925	Below
7	–1.67277	0.47519	–2.38556	–0.95999	Below
8	–1.45654	0.45558	–2.13991	–0.77317	Below
9	–1.25617	0.44026	–1.91656	–0.59578	Below
10	–1.06781	0.42818	–1.71008	–0.42554	Below
11	–0.88871	0.4186	–1.51661	–0.26081	Below
12	–0.71677	0.41099	–1.33326	–0.10029	Below
13	–0.55042	0.40496	–1.15786	0.05702	Below
14	–0.38842	0.4002	–0.98872	0.21188	Below
15	–0.22982	0.39645	–0.8245	0.36485	Below
16	–0.07384	0.39357	–0.6642	0.51651	Near/Met
17	0.08018	0.39146	–0.50701	0.66737	Near/Met
18	0.23286	0.39017	–0.3524	0.81811	Near/Met
19	0.38492	0.38993	–0.19998	0.96981	Near/Met
20	0.53731	0.39109	–0.04933	1.12394	Near/Met
21	0.69134	0.39423	0.09999	1.28268	Near/Met
22	0.84886	0.40006	0.24877	1.44895	Near/Met
23	1.01242	0.40944	0.39826	1.62658	Near/Met
24	1.18548	0.42339	0.55039	1.82056	Above
25	1.37277	0.44317	0.70801	2.03752	Above
26	1.58084	0.47041	0.87522	2.28645	Above
27	1.81912	0.50755	1.05779	2.58044	Above
28	2.10213	0.55894	1.26372	2.94054	Above
29	2.45524	0.63399	1.50425	3.40622	Above
30	2.93276	0.75866	1.79477	4.07075	Above
31	3.70494	1.0404	2.14434	5.26554	Above
32	4.96818	1.84936	2.19414	7.74222	Above

Table J.17: Grade 8 Critiquing Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–3.47818	1.84542	–6.24631	–0.71005	Below
1	–2.22819	1.02948	–3.77241	–0.68397	Below
2	–1.48077	0.74206	–2.59386	–0.36768	Below
3	–1.0245	0.62129	–1.95644	–0.09257	Below
4	–0.67939	0.55982	–1.51912	0.16034	Below
5	–0.38479	0.52954	–1.1791	0.40952	Below
6	–0.11147	0.51864	–0.88943	0.66649	Near/Met
7	0.15769	0.52075	–0.62344	0.93881	Near/Met
8	0.43407	0.53197	–0.36389	1.23202	Near/Met
9	0.72635	0.55042	–0.09928	1.55198	Near/Met
10	1.04321	0.57683	0.17796	1.90845	Near/Met
11	1.39714	0.61545	0.47396	2.32031	Above
12	1.81149	0.67666	0.7965	2.82648	Above
13	2.33936	0.78759	1.15797	3.52074	Above
14	3.15077	1.05653	1.56597	4.73556	Above
15	4.43518	1.85658	1.65031	7.22005	Above

Table J.18: Grade 8 Investigating Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.23295	1.84313	–6.99765	–1.46826	Below
1	–2.98407	1.03176	–4.53171	–1.43643	Below
2	–2.22589	0.75202	–3.35392	–1.09786	Below
3	–1.75442	0.63229	–2.70286	–0.80599	Below
4	–1.39988	0.56323	–2.24473	–0.55504	Below
5	–1.10919	0.51755	–1.88552	–0.33287	Below
6	–0.85874	0.48483	–1.58599	–0.1315	Below
7	–0.6359	0.46027	–1.32631	0.05450	Below
8	–0.43297	0.44145	–1.09515	0.22920	Below
9	–0.24463	0.42718	–0.8854	0.39614	Below
10	–0.06677	0.41694	–0.69218	0.55864	Near/Met
11	0.10418	0.41062	–0.51175	0.72011	Near/Met
12	0.27155	0.40827	–0.34086	0.88395	Near/Met
13	0.43868	0.4101	–0.17647	1.05383	Near/Met
14	0.60916	0.41645	–0.01551	1.23383	Near/Met
15	0.78698	0.42779	0.14529	1.42866	Near/Met
16	0.97693	0.44485	0.30965	1.64420	Near/Met
17	1.18512	0.46892	0.48174	1.8885	Above
18	1.42014	0.50231	0.66667	2.17360	Above
19	1.69543	0.54961	0.87101	2.51984	Above
20	2.03527	0.62099	1.10378	2.96675	Above
21	2.4933	0.74356	1.37796	3.60864	Above
22	3.23975	1.02654	1.69994	4.77956	Above
23	4.48165	1.84064	1.72069	7.24261	Above

Table J.19: Grade 11 Earth and Space Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.54337	1.84526	–7.31126	–1.77548	Below
1	–3.28884	1.03576	–4.84248	–1.7352	Below
2	–2.52209	0.75791	–3.65896	–1.38523	Below
3	–2.04137	0.63984	–3.00113	–1.08161	Below
4	–1.67681	0.57244	–2.53547	–0.81815	Below
5	–1.37517	0.52849	–2.16791	–0.58244	Below
6	–1.11276	0.49754	–1.85907	–0.36645	Below
7	–0.87696	0.47463	–1.58891	–0.16502	Below
8	–0.66026	0.45705	–1.34584	0.02531	Below
9	–0.45789	0.44317	–1.12265	0.20686	Below
10	–0.26658	0.43196	–0.91452	0.38136	Below
11	–0.08407	0.42276	–0.71821	0.55007	Near/Met
12	0.09138	0.4152	–0.53142	0.71418	Near/Met
13	0.26115	0.40911	–0.35252	0.87481	Near/Met
14	0.42655	0.40456	–0.18029	1.03339	Near/Met
15	0.58899	0.40183	–0.01376	1.19173	Near/Met
16	0.75009	0.40135	0.14806	1.35211	Near/Met
17	0.91192	0.40374	0.30631	1.51753	Near/Met
18	1.07708	0.40972	0.4625	1.69166	Near/Met
19	1.24892	0.42018	0.61865	1.87919	Above
20	1.43188	0.43629	0.77744	2.08631	Above
21	1.63199	0.45967	0.94248	2.32149	Above
22	1.85801	0.49291	1.11864	2.59737	Above
23	2.12373	0.5407	1.31268	2.93478	Above
24	2.45384	0.61313	1.53414	3.37353	Above
25	2.90225	0.73713	1.79655	4.00794	Above
26	3.63921	1.0219	2.10636	5.17206	Above
27	4.87439	1.83799	2.11740	7.63137	Above

Table J.20: Grade 11 Life Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.70883	1.84533	–7.47683	–1.94084	Below
1	–3.45409	1.03594	–5.008	–1.90018	Below
2	–2.68683	0.75832	–3.82431	–1.54935	Below
3	–2.20537	0.64053	–3.16617	–1.24458	Below
4	–1.83981	0.57342	–2.69994	–0.97968	Below
5	–1.53696	0.5297	–2.33151	–0.74241	Below
6	–1.27324	0.49882	–2.02147	–0.52501	Below
7	–1.03628	0.47568	–1.7498	–0.32276	Below
8	–0.81886	0.45748	–1.50508	–0.13264	Below
9	–0.61651	0.44259	–1.2804	0.04737	Below
10	–0.42624	0.43015	–1.07147	0.21898	Below
11	–0.24574	0.4199	–0.87559	0.38411	Below
12	–0.07289	0.41203	–0.69094	0.54515	Near/Met
13	0.0946	0.40701	–0.51592	0.70511	Near/Met
14	0.25938	0.40548	–0.34884	0.8676	Near/Met
15	0.42455	0.4081	–0.1876	1.0367	Near/Met
16	0.59379	0.41553	–0.02951	1.21708	Near/Met
17	0.77148	0.42846	0.12879	1.41417	Near/Met
18	0.96291	0.44766	0.29142	1.6344	Near/Met
19	1.17481	0.47422	0.46348	1.88614	Near/Met
20	1.41625	0.51009	0.65111	2.18138	Above
21	1.70091	0.55938	0.86184	2.53998	Above
22	2.05296	0.63178	1.10529	3.00063	Above
23	2.52561	0.754	1.39461	3.65661	Above
24	3.2885	1.03485	1.73622	4.84077	Above
25	4.54266	1.84548	1.77444	7.31088	Above

Table J.21: Grade 11 Physical Science Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.62453	1.84272	–7.38861	–1.86045	Below
1	–3.37666	1.03109	–4.9233	–1.83003	Below
2	–2.61977	0.75123	–3.74662	–1.49293	Below
3	–2.14937	0.63155	–3.0967	–1.20205	Below
4	–1.79557	0.56276	–2.63971	–0.95143	Below
5	–1.50513	0.5176	–2.28153	–0.72873	Below
6	–1.25425	0.48568	–1.98277	–0.52573	Below
7	–1.03015	0.46215	–1.72338	–0.33693	Below
8	–0.82502	0.44445	–1.4917	–0.15835	Below
9	–0.63361	0.43119	–1.2804	0.01317	Below
10	–0.45204	0.42157	–1.0844	0.18031	Below
11	–0.27723	0.41513	–0.89993	0.34546	Below
12	–0.10654	0.41165	–0.72402	0.51093	Near/Met
13	0.06247	0.41102	–0.55406	0.679	Near/Met
14	0.23211	0.4132	–0.38769	0.85191	Near/Met
15	0.40474	0.41826	–0.22265	1.03213	Near/Met
16	0.58284	0.42629	–0.0566	1.22227	Near/Met
17	0.76916	0.43754	0.11285	1.42547	Near/Met
18	0.96689	0.45241	0.28827	1.64550	Near/Met
19	1.18	0.47163	0.47255	1.88744	Near/Met
20	1.41384	0.49651	0.66907	2.15860	Above
21	1.67626	0.52949	0.88202	2.47049	Above
22	1.98008	0.57532	1.1171	2.84306	Above
23	2.34912	0.64428	1.3827	3.31554	Above
24	2.83684	0.76344	1.69168	3.982	Above
25	3.61366	1.04148	2.05144	5.17588	Above
26	4.87759	1.84935	2.10356	7.65161	Above

Table J.22: Grade 11 Sensemaking Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–5.14989	1.83789	–7.90673	–2.39306	Below
1	–3.91453	1.02234	–5.44804	–2.38102	Below
2	–3.17571	0.73905	–4.28429	–2.06714	Below
3	–2.72352	0.61693	–3.64892	–1.79813	Below
4	–2.38801	0.54626	–3.2074	–1.56862	Below
5	–2.11588	0.49959	–2.86527	–1.3665	Below
6	–1.88335	0.46638	–2.58292	–1.18378	Below
7	–1.67771	0.44161	–2.34013	–1.0153	Below
8	–1.49132	0.42255	–2.12515	–0.8575	Below
9	–1.31926	0.40758	–1.93063	–0.70789	Below
10	–1.15814	0.39565	–1.75162	–0.56467	Below
11	–1.00549	0.38607	–1.5846	–0.42639	Below
12	–0.85951	0.37834	–1.42702	–0.292	Below
13	–0.7188	0.37212	–1.27698	–0.16062	Below
14	–0.58224	0.36713	–1.13294	–0.03155	Below
15	–0.44897	0.36316	–0.99371	0.09577	Below
16	–0.31825	0.36006	–0.85834	0.22184	Below
17	–0.1895	0.3577	–0.72605	0.34705	Below
18	–0.0622	0.35598	–0.59617	0.47177	Below
19	0.06408	0.35483	–0.46817	0.59632	Near/Met
20	0.18973	0.35422	–0.3416	0.72106	Near/Met
21	0.31516	0.35418	–0.21611	0.84643	Near/Met
22	0.44077	0.35477	–0.09139	0.97292	Near/Met
23	0.56707	0.35616	0.03283	1.10131	Near/Met
24	0.6947	0.35855	0.15687	1.23252	Near/Met
25	0.82451	0.36227	0.28110	1.36791	Near/Met
26	0.95761	0.36772	0.40603	1.50919	Near/Met
27	1.09553	0.37543	0.53238	1.65867	Above
28	1.24028	0.38603	0.66123	1.81932	Above
29	1.39463	0.40037	0.79407	1.99518	Above
30	1.56235	0.41958	0.93298	2.19172	Above
31	1.74887	0.44532	1.08089	2.41685	Above
32	1.96231	0.48029	1.24187	2.68274	Above
33	2.21588	0.5294	1.42178	3.00998	Above
34	2.5338	0.603	1.6293	3.4383	Above
35	2.96966	0.72839	1.87707	4.06224	Above
36	3.69354	1.01533	2.17054	5.21653	Above
37	4.91921	1.83423	2.16786	7.67055	Above

Table J.23: Grade 11 Critiquing Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–3.84252	1.85875	–6.63065	–1.0544	Below
1	–2.55649	1.05485	–4.13877	–0.97422	Below
2	–1.75741	0.77405	–2.91849	–0.59634	Below
3	–1.25926	0.64797	–2.23122	–0.28731	Below
4	–0.89034	0.57137	–1.7474	–0.03329	Below
5	–0.5944	0.51934	–1.37341	0.18461	Below
6	–0.34429	0.48283	–1.06854	0.37995	Below
7	–0.12386	0.45773	–0.81046	0.56273	Near/Met
8	0.07788	0.44199	–0.58511	0.74086	Near/Met
9	0.26939	0.43456	–0.38245	0.92123	Near/Met
10	0.45787	0.43505	–0.19471	1.11044	Near/Met
11	0.65023	0.44343	–0.01492	1.31537	Near/Met
12	0.85367	0.46002	0.16364	1.5437	Near/Met
13	1.07645	0.48538	0.34838	1.80452	Near/Met
14	1.3287	0.52074	0.54759	2.10981	Above
15	1.62444	0.5692	0.77064	2.47824	Above
16	1.98713	0.63955	1.02780	2.94645	Above
17	2.46811	0.75817	1.33085	3.60536	Above
18	3.2342	1.03462	1.68227	4.78613	Above
19	4.48566	1.84342	1.72053	7.25079	Above

Table J.24: Grade 11 Investigating Score Table

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.47899	1.84599	–7.24798	–1.71001	Below
1	–3.22242	1.03729	–4.77836	–1.66649	Below
2	–2.45206	0.76059	–3.59295	–1.31118	Below
3	–1.96671	0.64392	–2.93259	–1.00083	Below
4	–1.5962	0.5783	–2.46365	–0.72875	Below
5	–1.2869	0.53661	–2.09182	–0.48199	Below
6	–1.01468	0.50854	–1.77749	–0.25187	Below
7	–0.76635	0.48926	–1.50024	–0.03246	Below
8	–0.53373	0.47623	–1.24808	0.18061	Below
9	–0.31116	0.46802	–1.01319	0.39087	Below
10	–0.09436	0.46383	–0.79011	0.60138	Near/Met
11	0.12024	0.46325	–0.57464	0.81511	Near/Met
12	0.33592	0.46616	–0.36332	1.03516	Near/Met
13	0.55598	0.47268	–0.15304	1.265	Near/Met
14	0.78409	0.48323	0.05924	1.50893	Near/Met
15	1.02467	0.4986	0.27677	1.77257	Near/Met
16	1.28357	0.52016	0.50333	2.06381	Above
17	1.56923	0.55037	0.74367	2.39478	Above
18	1.89512	0.59389	1.00428	2.78595	Above
19	2.2857	0.66072	1.29462	3.27678	Above
20	2.79492	0.77742	1.62879	3.96105	Above
21	3.59326	1.0517	2.01571	5.17081	Above
22	4.87197	1.85494	2.08956	7.65438	Above

Table J.25: Grade 11 Earth and Space Science Score Table – Computer-Based Spanish

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.52457	1.8459	–7.29342	–1.75572	Below
1	–3.26833	1.03698	–4.8238	–1.71286	Below
2	–2.49888	0.75977	–3.63854	–1.35923	Below
3	–2.01521	0.64227	–2.97862	–1.05181	Below
4	–1.64739	0.5754	–2.51049	–0.78429	Below
5	–1.34221	0.53197	–2.14017	–0.54426	Below
6	–1.07596	0.50152	–1.82824	–0.32368	Below
7	–0.83604	0.47909	–1.55468	–0.11741	Below
8	–0.61497	0.46193	–1.30787	0.07792	Below
9	–0.408	0.44842	–1.08063	0.26463	Below
10	–0.21194	0.43751	–0.86821	0.44432	Below
11	–0.02453	0.42857	–0.66739	0.61832	Near/Met
12	0.15591	0.42124	–0.47595	0.78777	Near/Met
13	0.33082	0.41547	–0.29239	0.95402	Near/Met
14	0.50163	0.41144	–0.11553	1.11879	Near/Met
15	0.66999	0.4096	0.05559	1.28439	Near/Met
16	0.83795	0.41058	0.22208	1.45382	Near/Met
17	1.00813	0.41517	0.38537	1.63088	Near/Met
18	1.18398	0.42435	0.54745	1.82050	Above
19	1.37002	0.43935	0.71099	2.02904	Above
20	1.57246	0.46186	0.87967	2.26525	Above
21	1.80025	0.49449	1.05851	2.54198	Above
22	2.06741	0.54195	1.25448	2.88033	Above
23	2.39886	0.61425	1.47748	3.32023	Above
24	2.84873	0.73819	1.74144	3.95601	Above
25	3.5874	1.0228	2.0532	5.1216	Above
26	4.82392	1.83853	2.06612	7.58171	Above

Table J.26: Grade 11 Life Science Score Table – Computer-Based Spanish

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.70883	1.84533	–7.47683	–1.94084	Below
1	–3.45409	1.03594	–5.008	–1.90018	Below
2	–2.68683	0.75832	–3.82431	–1.54935	Below
3	–2.20537	0.64053	–3.16617	–1.24458	Below
4	–1.83981	0.57342	–2.69994	–0.97968	Below
5	–1.53696	0.5297	–2.33151	–0.74241	Below
6	–1.27324	0.49882	–2.02147	–0.52501	Below
7	–1.03628	0.47568	–1.7498	–0.32276	Below
8	–0.81886	0.45748	–1.50508	–0.13264	Below
9	–0.61651	0.44259	–1.2804	0.04737	Below
10	–0.42624	0.43015	–1.07147	0.21898	Below
11	–0.24574	0.4199	–0.87559	0.38411	Below
12	–0.07289	0.41203	–0.69094	0.54515	Near/Met
13	0.0946	0.40701	–0.51592	0.70511	Near/Met
14	0.25938	0.40548	–0.34884	0.8676	Near/Met
15	0.42455	0.4081	–0.1876	1.0367	Near/Met
16	0.59379	0.41553	–0.02951	1.21708	Near/Met
17	0.77148	0.42846	0.12879	1.41417	Near/Met
18	0.96291	0.44766	0.29142	1.6344	Near/Met
19	1.17481	0.47422	0.46348	1.88614	Near/Met
20	1.41625	0.51009	0.65111	2.18138	Above
21	1.70091	0.55938	0.86184	2.53998	Above
22	2.05296	0.63178	1.10529	3.00063	Above
23	2.52561	0.754	1.39461	3.65661	Above
24	3.2885	1.03485	1.73622	4.84077	Above
25	4.54266	1.84548	1.77444	7.31088	Above

Table J.27: Grade 11 Physical Science Score Table – Computer-Based Spanish

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.62453	1.84272	–7.38861	–1.86045	Below
1	–3.37666	1.03109	–4.9233	–1.83003	Below
2	–2.61977	0.75123	–3.74662	–1.49293	Below
3	–2.14937	0.63155	–3.0967	–1.20205	Below
4	–1.79557	0.56276	–2.63971	–0.95143	Below
5	–1.50513	0.5176	–2.28153	–0.72873	Below
6	–1.25425	0.48568	–1.98277	–0.52573	Below
7	–1.03015	0.46215	–1.72338	–0.33693	Below
8	–0.82502	0.44445	–1.4917	–0.15835	Below
9	–0.63361	0.43119	–1.2804	0.01317	Below
10	–0.45204	0.42157	–1.0844	0.18031	Below
11	–0.27723	0.41513	–0.89993	0.34546	Below
12	–0.10654	0.41165	–0.72402	0.51093	Near/Met
13	0.06247	0.41102	–0.55406	0.679	Near/Met
14	0.23211	0.4132	–0.38769	0.85191	Near/Met
15	0.40474	0.41826	–0.22265	1.03213	Near/Met
16	0.58284	0.42629	–0.0566	1.22227	Near/Met
17	0.76916	0.43754	0.11285	1.42547	Near/Met
18	0.96689	0.45241	0.28827	1.64550	Near/Met
19	1.18	0.47163	0.47255	1.88744	Near/Met
20	1.41384	0.49651	0.66907	2.15860	Above
21	1.67626	0.52949	0.88202	2.47049	Above
22	1.98008	0.57532	1.1171	2.84306	Above
23	2.34912	0.64428	1.3827	3.31554	Above
24	2.83684	0.76344	1.69168	3.982	Above
25	3.61366	1.04148	2.05144	5.17588	Above
26	4.87759	1.84935	2.10356	7.65161	Above

Table J.28: Grade 11 Sensemaking Score Table – Computer-Based Spanish

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–5.14989	1.83789	–7.90673	–2.39306	Below
1	–3.91453	1.02234	–5.44804	–2.38102	Below
2	–3.17571	0.73905	–4.28429	–2.06714	Below
3	–2.72352	0.61693	–3.64892	–1.79813	Below
4	–2.38801	0.54626	–3.2074	–1.56862	Below
5	–2.11588	0.49959	–2.86527	–1.3665	Below
6	–1.88335	0.46638	–2.58292	–1.18378	Below
7	–1.67771	0.44161	–2.34013	–1.0153	Below
8	–1.49132	0.42255	–2.12515	–0.8575	Below
9	–1.31926	0.40758	–1.93063	–0.70789	Below
10	–1.15814	0.39565	–1.75162	–0.56467	Below
11	–1.00549	0.38607	–1.5846	–0.42639	Below
12	–0.85951	0.37834	–1.42702	–0.292	Below
13	–0.7188	0.37212	–1.27698	–0.16062	Below
14	–0.58224	0.36713	–1.13294	–0.03155	Below
15	–0.44897	0.36316	–0.99371	0.09577	Below
16	–0.31825	0.36006	–0.85834	0.22184	Below
17	–0.1895	0.3577	–0.72605	0.34705	Below
18	–0.0622	0.35598	–0.59617	0.47177	Below
19	0.06408	0.35483	–0.46817	0.59632	Near/Met
20	0.18973	0.35422	–0.3416	0.72106	Near/Met
21	0.31516	0.35418	–0.21611	0.84643	Near/Met
22	0.44077	0.35477	–0.09139	0.97292	Near/Met
23	0.56707	0.35616	0.03283	1.10131	Near/Met
24	0.6947	0.35855	0.15687	1.23252	Near/Met
25	0.82451	0.36227	0.28110	1.36791	Near/Met
26	0.95761	0.36772	0.40603	1.50919	Near/Met
27	1.09553	0.37543	0.53238	1.65867	Above
28	1.24028	0.38603	0.66123	1.81932	Above
29	1.39463	0.40037	0.79407	1.99518	Above
30	1.56235	0.41958	0.93298	2.19172	Above
31	1.74887	0.44532	1.08089	2.41685	Above
32	1.96231	0.48029	1.24187	2.68274	Above
33	2.21588	0.5294	1.42178	3.00998	Above
34	2.5338	0.603	1.6293	3.4383	Above
35	2.96966	0.72839	1.87707	4.06224	Above
36	3.69354	1.01533	2.17054	5.21653	Above
37	4.91921	1.83423	2.16786	7.67055	Above

Table J.29: Grade 11 Critiquing Score Table – Computer-Based Spanish

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–3.80388	1.86087	–6.59519	–1.01258	Below
1	–2.51254	1.05826	–4.09993	–0.92515	Below
2	–1.70699	0.77771	–2.87356	–0.54043	Below
3	–1.20397	0.65113	–2.18067	–0.22728	Below
4	–0.83153	0.57402	–1.69256	0.0295	Below
5	–0.5328	0.5219	–1.31565	0.25005	Below
6	–0.2799	0.48593	–1.0088	0.44899	Below
7	–0.05602	0.46211	–0.74919	0.63714	Near/Met
8	0.15056	0.44847	–0.52215	0.82326	Near/Met
9	0.34907	0.44413	–0.31713	1.01526	Near/Met
10	0.54775	0.44887	–0.12556	1.22105	Near/Met
11	0.7549	0.46297	0.06044	1.44935	Near/Met
12	0.97977	0.48708	0.24915	1.71039	Near/Met
13	1.23363	0.52241	0.45001	2.01724	Near/Met
14	1.5316	0.57169	0.67406	2.38913	Above
15	1.89788	0.64293	0.93348	2.86227	Above
16	2.3838	0.76172	1.24122	3.52638	Above
17	3.15541	1.03723	1.59956	4.71125	Above
18	4.4104	1.84466	1.64341	7.17739	Above

Table J.30: Grade 11 Investigating Score Table – Computer-Based Spanish

Raw Score	Theta	CSEM	Lower	Upper	Level
0	–4.47899	1.84599	–7.24798	–1.71001	Below
1	–3.22242	1.03729	–4.77836	–1.66649	Below
2	–2.45206	0.76059	–3.59295	–1.31118	Below
3	–1.96671	0.64392	–2.93259	–1.00083	Below
4	–1.5962	0.5783	–2.46365	–0.72875	Below
5	–1.2869	0.53661	–2.09182	–0.48199	Below
6	–1.01468	0.50854	–1.77749	–0.25187	Below
7	–0.76635	0.48926	–1.50024	–0.03246	Below
8	–0.53373	0.47623	–1.24808	0.18061	Below
9	–0.31116	0.46802	–1.01319	0.39087	Below
10	–0.09436	0.46383	–0.79011	0.60138	Near/Met
11	0.12024	0.46325	–0.57464	0.81511	Near/Met
12	0.33592	0.46616	–0.36332	1.03516	Near/Met
13	0.55598	0.47268	–0.15304	1.265	Near/Met
14	0.78409	0.48323	0.05924	1.50893	Near/Met
15	1.02467	0.4986	0.27677	1.77257	Near/Met
16	1.28357	0.52016	0.50333	2.06381	Above
17	1.56923	0.55037	0.74367	2.39478	Above
18	1.89512	0.59389	1.00428	2.78595	Above
19	2.2857	0.66072	1.29462	3.27678	Above
20	2.79492	0.77742	1.62879	3.96105	Above
21	3.59326	1.0517	2.01571	5.17081	Above
22	4.87197	1.85494	2.08956	7.65438	Above

APPENDIX K: Subscore Proficiency Classifications

Table K.1: Grade 5 Earth and Space Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	101,220	50.42	38.28	11.30
Male	51,656	50.82	37.48	11.71
Female	49,564	50.01	39.11	10.87
Am. Indian	133	51.13	39.85	9.02
Asian	10,859	22.57	47.81	29.62
Black	15,345	73.86	23.47	2.67
Hispanic	29,836	66.23	29.33	4.44
Pacific Islander	189	43.92	40.74	15.34
White	42,442	38.37	47.23	14.40
EL–Yes	5,830	88.64	10.48	0.87
EL–No	95,382	48.09	39.98	11.94
EconDis–Yes	38,634	70.25	26.22	3.53
EconDis–No	62,586	38.18	45.72	16.09
SWD–Yes	20,499	73.18	22.56	4.26
SWD–No	80,721	44.64	42.27	13.09
CBT	81,996	45.17	42.04	12.80
PBT	113	81.42	18.58	0.00
TTS	17,551	71.27	23.38	5.35
SP	825	92.12	7.64	0.24
SP TTS	387	90.70	9.04	0.26
Human Reader	303	82.18	16.50	1.32

Table K.2: Grade 5 Life Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	101,220	52.34	39.70	7.96
Male	51,656	52.07	39.39	8.54
Female	49,564	52.62	40.02	7.36
Am. Indian	133	57.14	35.34	7.52
Asian	10,859	25.72	54.11	20.17
Black	15,345	73.87	23.74	2.39
Hispanic	29,836	67.48	29.17	3.35
Pacific Islander	189	42.86	48.15	8.99
White	42,442	41.15	48.89	9.96
EL–Yes	5,830	90.12	9.33	0.55
EL–No	95,382	50.03	41.56	8.42
EconDis–Yes	38,634	71.58	25.90	2.52
EconDis–No	62,586	40.46	48.21	11.32
SWD–Yes	20,499	73.85	22.92	3.23
SWD–No	80,721	46.88	43.96	9.17
CBT	81,996	47.23	43.70	9.06
PBT	113	78.76	18.58	2.65
TTS	17,551	72.63	23.83	3.54
SP	825	93.21	6.42	0.36
SP TTS	387	92.25	7.49	0.26
Human Reader	303	79.87	19.47	0.66

Table K.3: Grade 5 Physical Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	101,220	47.99	39.03	12.98
Male	51,656	46.32	38.99	14.69
Female	49,564	49.73	39.07	11.19
Am. Indian	133	51.13	35.34	13.53
Asian	10,859	22.16	46.74	31.11
Black	15,345	70.76	25.50	3.74
Hispanic	29,836	64.21	30.70	5.09
Pacific Islander	189	43.39	41.80	14.81
White	42,442	35.42	47.60	16.98
EL–Yes	5,830	87.39	11.82	0.79
EL–No	95,382	45.58	40.70	13.72
EconDis–Yes	38,634	68.01	27.92	4.07
EconDis–No	62,586	35.63	45.89	18.48
SWD–Yes	20,499	70.36	24.12	5.51
SWD–No	80,721	42.31	42.82	14.87
CBT	81,996	42.73	42.51	14.76
PBT	113	76.99	16.81	6.19
TTS	17,551	68.77	25.45	5.78
SP	825	93.09	6.42	0.48
SP TTS	387	87.34	12.66	0.00
Human Reader	303	78.22	20.13	1.65

Table K.4: Grade 5 Investigating Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	101,220	47.75	43.80	8.44
Male	51,656	46.98	43.65	9.37
Female	49,564	48.56	43.96	7.48
Am. Indian	133	50.38	40.60	9.02
Asian	10,859	21.48	57.32	21.21
Black	15,345	70.62	27.16	2.23
Hispanic	29,836	64.54	32.32	3.14
Pacific Islander	189	41.80	50.26	7.94
White	42,442	34.81	54.19	11.00
EL–Yes	5,830	87.96	11.58	0.46
EL–No	95,382	45.30	45.77	8.93
EconDis–Yes	38,634	68.12	29.40	2.48
EconDis–No	62,586	35.19	52.69	12.12
SWD–Yes	20,499	70.99	25.55	3.46
SWD–No	80,721	41.85	48.44	9.71
CBT	81,996	42.31	48.07	9.62
PBT	113	74.34	22.12	3.54
TTS	17,551	69.44	26.84	3.71
SP	825	90.67	9.09	0.24
SP TTS	387	88.63	11.37	0.00
Human Reader	303	79.54	19.47	0.99

Table K.5: Grade 5 Sensemaking Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	101,220	53.25	34.37	12.38
Male	51,656	53.09	33.92	12.99
Female	49,564	53.41	34.85	11.75
Am. Indian	133	52.63	37.59	9.77
Asian	10,859	24.53	42.93	32.54
Black	15,345	76.57	20.40	3.03
Hispanic	29,836	69.32	25.89	4.79
Pacific Islander	189	44.97	37.04	17.99
White	42,442	41.32	42.97	15.71
EL–Yes	5,830	90.45	8.71	0.84
EL–No	95,382	50.97	35.94	13.09
EconDis–Yes	38,634	73.48	22.77	3.75
EconDis–No	62,586	40.76	41.53	17.71
SWD–Yes	20,499	75.65	19.69	4.66
SWD–No	80,721	47.56	38.10	14.34
CBT	81,996	48.06	37.90	14.04
PBT	113	85.84	13.27	0.88
TTS	17,551	73.86	20.39	5.75
SP	825	93.09	6.91	0.00
SP TTS	387	91.99	7.24	0.78
Human Reader	303	86.47	11.88	1.65

Table K.6: Grade 5 Critiquing Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	101,220	47.93	42.98	9.09
Male	51,656	47.35	42.34	10.31
Female	49,564	48.54	43.63	7.83
Am. Indian	133	45.11	45.86	9.02
Asian	10,859	23.12	53.92	22.96
Black	15,345	69.73	27.85	2.42
Hispanic	29,836	62.52	33.70	3.77
Pacific Islander	189	44.44	44.97	10.58
White	42,442	36.56	51.93	11.52
EL–Yes	5,830	86.84	12.45	0.70
EL–No	95,382	45.55	44.84	9.60
EconDis–Yes	38,634	66.60	30.61	2.79
EconDis–No	62,586	36.41	50.61	12.98
SWD–Yes	20,499	69.79	26.19	4.02
SWD–No	80,721	42.38	47.24	10.38
CBT	81,996	42.78	46.93	10.29
PBT	113	76.99	21.24	1.77
TTS	17,551	68.20	27.51	4.29
SP	825	91.76	8.00	0.24
SP TTS	387	89.92	9.82	0.26
Human Reader	303	78.88	19.47	1.65

Table K.7: Grade 8 Earth and Space Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	99,852	53.98	39.67	6.35
Male	51,124	53.83	38.44	7.73
Female	48,728	54.13	40.97	4.90
Am. Indian	117	56.41	40.17	3.42
Asian	10,346	25.34	56.23	18.42
Black	14,452	76.97	21.92	1.11
Hispanic	28,176	70.72	27.55	1.74
Pacific Islander	207	38.65	55.07	6.28
White	44,716	43.00	48.99	8.01
EL–Yes	4,381	90.57	9.08	0.34
EL–No	95,468	52.30	41.08	6.62
EconDis–Yes	34,908	74.44	24.37	1.19
EconDis–No	64,944	42.98	47.90	9.12
SWD–Yes	19,664	76.06	21.42	2.52
SWD–No	80,188	48.56	44.15	7.29
CBT	85,985	50.35	42.65	6.99
PBT	92	86.96	11.96	1.09
TTS	12,100	73.97	23.37	2.66
SP	1,126	94.49	5.51	0.00
SP TTS	432	92.59	6.94	0.46
Human Reader	84	94.05	5.95	0.00

Table K.8: Grade 8 Life Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	99,852	59.96	33.05	7.00
Male	51,124	61.49	31.42	7.09
Female	48,728	58.35	34.76	6.90
Am. Indian	117	62.39	32.48	5.13
Asian	10,346	28.99	49.40	21.61
Black	14,452	81.35	17.13	1.52
Hispanic	28,176	77.21	20.77	2.02
Pacific Islander	207	45.41	43.96	10.63
White	44,716	49.77	41.90	8.33
EL–Yes	4,381	94.70	5.09	0.21
EL–No	95,468	58.36	34.33	7.31
EconDis–Yes	34,908	80.03	18.47	1.50
EconDis–No	64,944	49.17	40.88	9.95
SWD–Yes	19,664	81.38	16.55	2.07
SWD–No	80,188	54.71	37.09	8.20
CBT	85,985	56.49	35.74	7.77
PBT	92	86.96	13.04	0.00
TTS	12,100	79.31	18.18	2.50
SP	1,126	96.80	3.11	0.09
SP TTS	432	96.53	3.47	0.00
Human Reader	84	94.05	5.95	0.00

Table K.9: Grade 8 Physical Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	99,852	65.08	27.82	7.10
Male	51,124	64.66	27.11	8.22
Female	48,728	65.52	28.55	5.93
Am. Indian	117	70.94	26.50	2.56
Asian	10,346	35.15	43.51	21.33
Black	14,452	85.70	13.04	1.26
Hispanic	28,176	82.22	16.02	1.76
Pacific Islander	207	51.69	38.16	10.14
White	44,716	54.98	36.12	8.91
EL–Yes	4,381	95.85	3.83	0.32
EL–No	95,468	63.67	28.92	7.42
EconDis–Yes	34,908	84.50	14.16	1.33
EconDis–No	64,944	54.64	35.15	10.21
SWD–Yes	19,664	84.25	13.33	2.42
SWD–No	80,188	60.38	31.37	8.25
CBT	85,985	61.81	30.24	7.94
PBT	92	93.48	6.52	0.00
TTS	12,100	83.62	14.21	2.17
SP	1,126	97.42	2.58	0.00
SP TTS	432	96.99	3.01	0.00
Human Reader	84	96.43	3.57	0.00

Table K.10: Grade 8 Investigating Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	99,852	64.07	29.61	6.32
Male	51,124	64.12	28.69	7.18
Female	48,728	64.01	30.57	5.42
Am. Indian	117	71.79	25.64	2.56
Asian	10,346	33.37	47.22	19.42
Black	14,452	84.84	13.97	1.19
Hispanic	28,176	80.71	17.70	1.59
Pacific Islander	207	51.21	39.13	9.66
White	44,716	54.41	37.79	7.80
EL–Yes	4,381	94.80	4.82	0.39
EL–No	95,468	62.66	30.75	6.59
EconDis–Yes	34,908	83.40	15.44	1.15
EconDis–No	64,944	53.68	37.22	9.10
SWD–Yes	19,664	84.21	13.68	2.11
SWD–No	80,188	59.13	33.51	7.36
CBT	85,985	60.87	32.11	7.02
PBT	92	90.22	9.78	0.00
TTS	12,100	82.07	15.65	2.28
SP	1,126	96.54	3.37	0.09
SP TTS	432	97.69	2.08	0.23
Human Reader	84	98.81	1.19	0.00

Table K.11: Grade 8 Sensemaking Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	99,852	62.40	30.53	7.07
Male	51,124	62.00	29.82	8.18
Female	48,728	62.82	31.28	5.90
Am. Indian	117	67.52	26.50	5.98
Asian	10,346	32.03	47.01	20.95
Black	14,452	83.68	15.10	1.22
Hispanic	28,176	79.48	18.68	1.85
Pacific Islander	207	46.86	45.89	7.25
White	44,716	52.24	38.91	8.85
EL–Yes	4,381	95.05	4.72	0.23
EL–No	95,468	60.90	31.72	7.38
EconDis–Yes	34,908	82.14	16.60	1.27
EconDis–No	64,944	51.79	38.02	10.19
SWD–Yes	19,664	82.42	14.99	2.59
SWD–No	80,188	57.49	34.34	8.17
CBT	85,985	59.08	33.06	7.86
PBT	92	90.22	9.78	0.00
TTS	12,100	80.99	16.51	2.50
SP	1,126	97.34	2.66	0.00
SP TTS	432	96.30	3.47	0.23
Human Reader	84	94.05	5.95	0.00

Table K.12: Grade 8 Critiquing Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	99,852	58.84	34.43	6.73
Male	51,124	60.30	32.58	7.13
Female	48,728	57.32	36.38	6.31
Am. Indian	117	64.10	32.48	3.42
Asian	10,346	28.74	51.13	20.13
Black	14,452	80.64	17.87	1.49
Hispanic	28,176	76.07	21.87	2.06
Pacific Islander	207	44.44	43.48	12.08
White	44,716	48.27	43.65	8.08
EL–Yes	4,381	93.81	5.96	0.23
EL–No	95,468	57.24	35.74	7.02
EconDis–Yes	34,908	78.99	19.41	1.60
EconDis–No	64,944	48.02	42.50	9.48
SWD–Yes	19,664	79.63	18.04	2.33
SWD–No	80,188	53.75	38.45	7.80
CBT	85,985	55.27	37.25	7.47
PBT	92	85.87	13.04	1.09
TTS	12,100	78.93	18.69	2.38
SP	1,126	95.74	4.09	0.18
SP TTS	432	95.14	4.86	0.00
Human Reader	84	95.24	4.76	0.00

Table K.13: Grade 11 Earth and Space Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	90,024	52.85	34.98	12.17
Male	45,733	54.98	32.28	12.75
Female	44,291	50.66	37.78	11.57
Am. Indian	111	55.86	32.43	11.71
Asian	9,097	26.61	43.31	30.08
Black	12,935	73.78	22.42	3.80
Hispanic	23,417	67.31	27.61	5.07
Pacific Islander	223	34.98	48.43	16.59
White	43,112	44.40	40.95	14.65
EL–Yes	3,878	91.08	8.43	0.49
EL–No	86,132	51.13	36.18	12.69
EconDis–Yes	27,411	69.78	25.61	4.61
EconDis–No	62,613	45.44	39.08	15.48
SWD–Yes	16,414	74.39	20.76	4.86
SWD–No	73,610	48.05	38.15	13.80
CBT	84,298	51.44	35.89	12.67
PBT	154	74.68	20.78	4.55
TTS	4,408	68.56	25.48	5.97
SP	761	92.12	7.49	0.39
SP TTS	262	92.75	6.87	0.38
Human Reader	113	94.69	5.31	0.00

Table K.14: Grade 11 Life Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	90,024	56.23	33.05	10.72
Male	45,733	57.40	31.28	11.32
Female	44,291	55.02	34.88	10.10
Am. Indian	111	57.66	32.43	9.91
Asian	9,097	28.62	42.44	28.93
Black	12,935	77.63	19.64	2.74
Hispanic	23,417	71.42	24.46	4.12
Pacific Islander	223	38.12	46.19	15.70
White	43,112	47.58	39.66	12.76
EL–Yes	3,878	93.99	5.91	0.10
EL–No	86,132	54.53	34.27	11.20
EconDis–Yes	27,411	74.14	22.17	3.68
EconDis–No	62,613	48.39	37.81	13.80
SWD–Yes	16,414	77.48	18.59	3.93
SWD–No	73,610	51.49	36.27	12.24
CBT	84,298	54.76	34.07	11.18
PBT	154	76.62	17.53	5.84
TTS	4,408	73.28	21.73	4.99
SP	761	95.53	4.47	0.00
SP TTS	262	97.33	2.29	0.38
Human Reader	113	93.81	6.19	0.00

Table K.15: Grade 11 Physical Science Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	90,024	54.02	35.66	10.32
Male	45,733	54.53	33.74	11.74
Female	44,291	53.50	37.65	8.85
Am. Indian	111	51.35	40.54	8.11
Asian	9,097	25.77	44.47	29.77
Black	12,935	75.34	22.13	2.53
Hispanic	23,417	69.67	26.78	3.55
Pacific Islander	223	35.43	49.78	14.80
White	43,112	45.30	42.59	12.12
EL–Yes	3,878	92.78	6.88	0.34
EL–No	86,132	52.27	36.96	10.77
EconDis–Yes	27,411	71.87	24.77	3.36
EconDis–No	62,613	46.21	40.43	13.36
SWD–Yes	16,414	76.04	20.09	3.87
SWD–No	73,610	49.11	39.14	11.75
CBT	84,298	52.50	36.73	10.77
PBT	154	77.27	16.88	5.84
TTS	4,408	71.62	23.93	4.45
SP	761	93.56	6.31	0.13
SP TTS	262	96.56	3.44	0.00
Human Reader	113	93.81	6.19	0.00

Table K.16: Grade 11 Investigating Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	90,024	54.24	31.66	14.10
Male	45,733	56.20	29.08	14.71
Female	44,291	52.21	34.32	13.47
Am. Indian	111	54.95	28.83	16.22
Asian	9,097	25.98	37.59	36.43
Black	12,935	74.94	21.01	4.05
Hispanic	23,417	69.88	24.53	5.59
Pacific Islander	223	35.87	44.39	19.73
White	43,112	45.67	37.46	16.88
EL–Yes	3,878	93.63	5.96	0.41
EL–No	86,132	52.46	32.82	14.72
EconDis–Yes	27,411	72.03	22.78	5.19
EconDis–No	62,613	46.45	35.55	18.00
SWD–Yes	16,414	77.12	17.78	5.09
SWD–No	73,610	49.13	34.75	16.11
CBT	84,298	52.71	32.57	14.72
PBT	154	75.97	20.13	3.90
TTS	4,408	71.69	22.07	6.24
SP	761	95.93	3.94	0.13
SP TTS	262	96.56	3.05	0.38
Human Reader	113	96.46	3.54	0.00

Table K.17: Grade 11 Sensemaking Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	90,024	59.60	26.61	13.79
Male	45,733	59.61	25.40	14.98
Female	44,291	59.60	27.85	12.56
Am. Indian	111	61.26	26.13	12.61
Asian	9,097	31.74	33.37	34.89
Black	12,935	80.26	15.95	3.80
Hispanic	23,417	74.57	19.75	5.69
Pacific Islander	223	47.53	32.29	20.18
White	43,112	51.31	32.08	16.61
EL–Yes	3,878	95.05	4.64	0.31
EL–No	86,132	58.01	27.59	14.40
EconDis–Yes	27,411	76.58	18.28	5.14
EconDis–No	62,613	52.17	30.25	17.58
SWD–Yes	16,414	79.55	14.94	5.51
SWD–No	73,610	55.16	29.21	15.64
CBT	84,298	58.23	27.40	14.37
PBT	154	79.87	13.64	6.49
TTS	4,408	75.45	18.01	6.53
SP	761	96.58	3.29	0.13
SP TTS	262	98.47	1.53	0.00
Human Reader	113	94.69	5.31	0.00

Table K.18: Grade 11 Critiquing Disaggregated Subscore Proficiency Classifications

Group	Ncount	Below	Near/Met	Above
All Students	90,024	52.00	36.35	11.65
Male	45,733	54.06	33.73	12.21
Female	44,291	49.88	39.06	11.07
Am. Indian	111	48.65	39.64	11.71
Asian	9,097	26.02	45.10	28.88
Black	12,935	73.95	22.96	3.09
Hispanic	23,417	66.40	29.12	4.48
Pacific Islander	223	34.53	50.67	14.80
White	43,112	43.27	42.38	14.35
EL–Yes	3,878	90.43	9.36	0.21
EL–No	86,132	50.27	37.57	12.16
EconDis–Yes	27,411	69.39	26.53	4.08
EconDis–No	62,613	44.39	40.65	14.96
SWD–Yes	16,414	72.90	22.31	4.79
SWD–No	73,610	47.35	39.48	13.18
CBT	84,298	50.55	37.30	12.15
PBT	154	70.78	22.73	6.49
TTS	4,408	69.10	25.61	5.29
SP	761	88.96	10.91	0.13
SP TTS	262	93.13	6.49	0.38
Human Reader	113	92.92	7.08	0.00