New Jersey
Start Strong Assessment–Science
(NJSSA–S)

TECHNICAL BRIEF
Grades 6, 9, and 12

2021

JANUARY 2023
PTM XXXX.XX

State of New Jersey
Department of Education

Contents

Part 1: Description of the NJSSA–S

The New Jersey Start Strong Assessment–Science (NJSSA–S) assesses students at the beginning of Grades 6, 9, and 12 on their understanding and explanations of scientific phenomena and scenarios. The tests cover a range of material 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. 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), Scientific and Engineering Practices (SEPs—Investigating, Sensemaking, or Critiquing), and Crosscutting Concepts (CCC). Every test item counts toward 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 NJSSA−S.) All items are machine scored and consist of a mixture of multiple-choice (MC) and technology-enhanced (TE) items. Part 1.1 describes in detail the content domains and scientific practices, as well as how they are grouped to form the NJSSA–S reporting categories; next, Part 1.2 briefly describes the Crosscutting Concepts.

1.1 Content Domains and Scientific Practices

Although the NJSSA−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 and engineering 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 NJSSA–S is designed to measure student performance in each of the three science content domains. The test development processes focus on balancing each science content domain equally. Furthermore, within each content domain, each DCI is balanced.

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 1.1.1 shows the three ESS DCIs as well as the topics that are delineated within each.

Table 1.1.1: 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) states that the DCIs "focus on patterns, processes, and relationships of living organisms" (p. 139). Table 1.1.2 presents the four LS DCIs and their underlying topics.

Table 1.1.2: Life Science DCIs

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

3. Physical Science. According to the *Framework* (NRC, 2012), the goal 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 1.1.3 illustrates the three PS DCIs along with the associated detailed topics for each.

Table 1.1.3: Physical Science DCIs

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

Scientific and engineering 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 NJSSA−S, the SEPs are consolidated into three categories of scientific practices: Investigating, Sensemaking, and Critiquing. Table 1.1.4, adapted from the work of McNeill, Katch-Singer, and Pelletier (2015), shows how the eight *Framework* SEPs were consolidated for the purposes of the NJSSA−S.

Table 1.1.4: 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 1.1.5 delineates the *Framework* definition of each of the Investigating Practices.

Table 1.1.5: 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 1.1.6 illustrates the *Framework* definition of each of the Sensemaking Practices.

Table 1.1.6 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 1.1.7 shows the *Framework* definition of each of the Critiquing Practices.

Table 1.1.7 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 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 NJSSA−S and contribute to the overall NJSSA−S scale score. Table 1.2.1 shows the CCCs being measured by the NJSSA−S.

Table 1.2.1: 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.

Part 2: Test Development

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

The Test Design and Development chapter within the *Standards* (2014) outlines a series of five primary phases of the test development process: (1) test specifications; (2) item development and review; (3) assembling and evaluating test forms; (4) development of procedures and materials for test administration and scoring; and (5) test revisions (p. 83). The following sections in Part 2 detail the NJSSA–S test specifications, item development processes, and both the test construction processes and their results in 2021. The development of procedures and materials for test administration and scoring is covered in Parts 3 and 4. No test revisions were 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).

As was described in Part 1 of this document, despite being administered to students entering Grades 6, 9, and 12, the NJSSA−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 the three student support levels (Strong Support, Some Support, and Less Support) 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 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 NJSSA−S test blueprint for all grades. The table summarizes the ideal range of the numbers of items on the operational NJSSA−S for each of the six reporting categories.

Table 2.1.1: NJSSA–S Test Blueprints

Domain	Practice	Grade 6	Grade 9	Grade 12
PS	**Investigating** *AQDP, PACI, UMCT*	2–5	2–5	2–5
PS	**Sensemaking** *DUM, AID, CEDS*	2–5	2–5	2–5
PS	**Critiquing** *EAE, OECI*	2–5	2–5	2–5
PS	***Total Items***	*7–10*	*7–10*	*7–10*
LS	**Investigating** *AQDP, PACI, UMCT*	2–5	2–5	2–5
LS	**Sensemaking** *DUM, AID, CEDS*	2–5	2–5	2–5
LS	**Critiquing** *EAE, OECI*	2–5	2–5	2–5
LS	***Total Items***	*7–10*	*7–10*	*7–10*
ESS	**Investigating** *AQDP, PACI, UMCT*	2–5	2–5	2–5
ESS	**Sensemaking** *DUM, AID, CEDS*	2–5	2–5	2–5
ESS	**Critiquing** *EAE, OECI*	2–5	2–5	2–5
ESS	***Total Items***	*7–10*	*7–10*	*7–10*

2.1.2 Item Types

Two types of items comprise the NJSSA–S: multiple-choice (MC) and technology-enhanced (TE).

- 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. Not all TE interaction types will be used on a single NJSSA–S form.

Table 2.1.2 describes each NJSSA–S item type.

Table 2.1.2: NJSSA–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.

2.2 Item Development Processes

All 2021 NJSSA–S items were originally used on the 2019 NJSLA–S operational assessment, and thus they were subjected to the NJSLA–S item development process, which is described in detail in the 2019 NJSLA-S Technical Report. The item development for the 2019 NJSLA–S was conducted by Measurement Incorporated (MI) and Pearson with oversight from NJDOE staff and the New Jersey Science Advisory Committee (NJSAC). The NJSLA–S 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/NJSSA–S reporting categories. The PBSs and their items are all housed in Pearson's Assessment Banking for Building and Interoperability (ABBI) item banking system. ABBI is specifically designed to handle next-generation online, interactive, and accessible content.

The steps in the item development process as well as how they incorporated the principles of universal design (Thompson, Johnstone, & Thurlow, 2002) are detailed in Part 2 of the 2019 NJSLA–S Technical Report details. 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 NJSLA–S/NJSSA–S item multiple times.

2.3 Test Construction Process

The NJSSA–S test construction process ensures that the 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 NJSSA–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.

2.3.1 Test Construction—First Draft

The first step in the NJSSA–S test construction process involves MI's psychometric staff manually selecting items that had been previously used operationally on the NJSLA–S that best matched the NJSSA–S test blueprint and statistical constraints. The process of selecting items is contingent upon the availability of previously operationalized NJSLA–S items at each grade level. If specific content constraints are challenging to fulfill, 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 NJSSA–S operational test form is reliable and fosters valid interpretations, the following statistical requirements are used by MI's psychometric staff during the test construction process. Table 2.3.1 provides a summary of the NJSSA–S test construction requirements.

Item difficulty. Each test form is constructed to a specific difficulty level. The most important decisions made from the NJSSA–S are at the Some Support and Less Support cut scores. To maximize the reliability of those decisions, the test form's target average item difficulty parameter is at the point on the NJSSA–S scale that maximizes test information at both of those decision points.

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 NJSSA–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 NJSSA–S if no other viable options are available.

IRT model fit. The NJSSA–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 necessary.

Time on items. The NJSSA–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 ensure the median time spent on the test is well below the time limit. 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 accounting for their ability levels. NJSSA–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 NJSSA–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 NJSSA–S.

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

Statistical Constraint	Description
Item Difficulty	Average item difficulty maximizes information at both the Some Support and Less Support cut scores.
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 NJSSA–S 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, they alert MI's psychometric staff, and the items are replaced. The content review then resumes until the test matches NJSSA–S content and statistical requirements.

2.3.3 Test Construction NJDOE Review

All NJSSA–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 2020 NJSSA–S Test Construction

2020 was the first year of NJSSA–S test construction. 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 met the requirement that no more than 50% of the 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.

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, at least one scientific practice was over-represented for a given content domain. For instance, of the eight Earth and Space Science points available on the Grade 6 test, six were aligned to the Sensemaking practice, whereas only one point each was aligned to the Investigating or Critiquing practices.

Table 2.4.1: 2021 NJSSA–S Points Available by Domain and Practice

Grade	Practice	Earth	Life	Physical
6	Investigating	1	3	4
6	Sensemaking	6	1	4
6	Critiquing	1	3	2
9	Investigating	4	1	4
9	Sensemaking	2	5	3
9	Critiquing	2	1	3
12	Investigating	3	0	5
12	Sensemaking	5	4	2
12	Critiquing	2	3	1

2.4.1 Grade 6 Test Construction

At Grade 6 the science content domains were balanced, as illustrated in Table 2.4.2. The least balanced content domain was Life Science, and it still made up 7 points of the 25 total score points. The scientific practices were slightly less balanced, with only 6 out of 25 points being allocated to the Critiquing reporting category, and 11 out of 25 points allocated to the Sensemaking reporting category. 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: 2021 NJSSA–S Grade 6 Item and Point Totals by Reporting Category

Domains/Practices	MC Items	TE Items	Items	Points
Earth and Space	4	4	8	8
Life	1	6	7	7
Physical	2	8	10	10
Total—Domains	7	18	25	25
Investigating	2	6	8	8
Sensemaking	4	7	11	11
Critiquing	1	5	6	6
Total—Practices	7	18	25	25

Table 2.4.3: 2021 NJSSA–S Grade 6 DCIs

DCI	Items	Points
ESS1	0	0
ESS2	8	8
ESS3	0	0
LS1	4	4
LS2	0	0
LS3	0	0
LS4	3	3
PS1	3	3
PS2	3	3
PS3	4	4
PS4	0	0

Table 2.4.4: 2021 NJSSA–S Grade 6 SEPs

SEP	Items	Points
AQDP	2	2
PACI	4	4
UMCT	3	3
DUM	1	1
AID	6	6
CEDS	3	3
EAE	6	6
OECI	0	0

Table 2.4.5: 2021 NJSSA–S Grade 6 CCCs

CCC	Items	Points
C & E	5	5
E & M	4	4
Patterns	9	9
S & SM	0	0
S, P, & Q	2	2
SC	1	1
SF	4	4

The statistical constraints for the 2021 Grade 6 NJSSA–S operational test form were generally met. The only exception was that the item difficulty maximized information close to the Some Support cut score, instead of in between the Some and Less Support cuts. Thus, the test was measuring the Some Support cut score at a higher level of reliability than the Less Support cut. Otherwise, all items had item-total correlations above the 0.2 threshold, and very few items were flagged for divergent item fit statistics. The median test time of 30.33 minutes was well below the 45-minute threshold, and out of 100 DIF classifications, there were zero “C” values and only 2 “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: 2021 NJSSA–S Grade 6 Test Construction Statistics

Statistics	Average	Target	Flags
Item Difficulty	–0.29	0.31	N/A
Item Total Correlation	0.42	> 0.35	0
Infit	0.97	1.00	0
Outfit	0.99	1.00	2
Item Discrimination	1.05	1.00	0
Lower Asymptote	0.02	0.00	1
Median Time	30.33	< 45	N/A

Table 2.4.7: 2021 NJSSA–S Grade 6 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	23	2	0
White/Black	25	0	0
White/Hispanic	25	0	0
White/Asian	25	0	0

2.4.2 Grade 9 Test Construction

The Grade 9 NJSSA–S content constraints were generally met. The science content domains were balanced. Each content domain was represented by between 7 to 10 points worth of items. The scientific practices were slightly less balanced with only 6 of 25 points allocated to the Critiquing reporting category. All eleven DCIs, seven of eight SEPs, and all seven CCCs were represented. 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: 2021 NJSSA–S Grade 9 Item and Point Totals by Reporting Category

Domains/Practices	MC Items	TE Items	Items	Points
Earth and Space	4	4	8	8
Life	1	6	7	7
Physical	3	6	9	10
Total—Domains	8	16	24	25
Investigating	6	3	9	9
Sensemaking	1	9	10	10
Critiquing	1	4	5	6
Total—Practices	8	16	24	25

Table 2.4.9: 2021 NJSSA–S Grade 9 DCIs

DCI	Items	Points
ESS1	2	2
ESS2	3	3
ESS3	3	3
LS1	2	2
LS2	3	3
LS3	1	1
LS4	1	1
PS1	1	1
PS2	3	3
PS3	3	4
PS4	2	2

Table 2.4.10: 2021 NJSSA–S Grade 9 SEPs

SEP	Items	Points
AQDP	5	5
PACI	2	2
UMCT	2	2
DUM	4	4
AID	3	3
CEDS	3	3
EAE	5	6
OECI	0	0

Table 2.4.11: 2021 NJSSA–S Grade 9 CCCs

CCC	Items	Points
C & E	5	5
E & M	5	6
Patterns	4	4
S & SM	2	2
S, P, & Q	2	2
SC	1	1
SF	5	5

The statistical constraints for the 2021 Grade 9 NJSSA–S operational test form were met. The average item difficulty parameter was only 0.05 logits from the target value. Only one Grade 9 item was flagged for having item-total correlations below the 0.2 threshold. The infit, outfit, and PCM item discrimination model-fit statistics were all close to their ideal values of 1.00. The median test time of 28.67 minutes was well below the 45-minute threshold, and out of 96 DIF classifications, there were zero “C” values and only one “B” value. The “B” DIF item was 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: 2021 NJSSA–S Grade 9 Test Construction Statistics

Statistics	Average	Target	Flags
Item Difficulty	–0.20	–0.25	N/A
Item Total Correlation	0.39	> 0.35	1
Infit	0.97	1.00	0
Outfit	0.97	1.00	2
Item Discrimination	1.06	1.00	1
Lower Asymptote	0.02	0.00	0
Median Time	28.67	< 45	N/A

Table 2.4.13: 2021 NJSSA–S Grade 9 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	23	1	0
White/Black	24	0	0
White/Hispanic	24	0	0
White/Asian	24	0	0

2.4.3 Grade 12 Test Construction

The Grade 12 NJSSA–S content constraints were generally met. The content domains were close to being equal. Each content domain had between 7 and 10 points worth of items. Similar as they were in the other two grade levels, the scientific practices were less balanced, with only 6 out of 25 points being allocated to the Critiquing reporting category; the Sensemaking practice was overrepresented with 11 out of 25 points. 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 12.

Table 2.4.14: 2021 NJSSA–S Grade 12 Item and Point Totals by Reporting Category

Domains/Practices	MC Items	TE Items	Items	Points
Earth and Space	5	5	10	10
Life	1	6	7	7
Physical	4	4	8	8
Total – Domains	10	15	25	25
Investigating	4	4	8	8
Sensemaking	5	6	11	11
Critiquing	1	5	6	6
Total – Practices	10	15	25	25

Table 2.4.15: 2021 NJSSA–S Grade 12 DCIs

DCI	Items	Points
ESS1	3	3
ESS2	3	3
ESS3	4	4
LS1	0	0
LS2	3	3
LS3	0	0
LS4	4	4
PS1	3	3
PS2	2	2
PS3	3	3
PS4	0	0

Table 2.4.16: 2021 NJSSA–S Grade 12 SEPs

SEP	Items	Points
AQDP	2	2
PACI	2	2
UMCT	4	4
DUM	3	3
AID	7	7
CEDS	1	1
EAE	2	2
OECI	4	4

Table 2.4.17: 2021 NJSSA–S Grade 12 CCCs

CCC	Items	Points
C & E	4	4
E & M	1	1
Patterns	5	5
S & SM	8	8
S, P, & Q	4	4
SC	3	3
SF	0	0

The statistical constraints for the 2021 Grade 12 NJSSA–S operational test form were more challenging to meet than for the other two grades. The average item difficulty parameter was 0.36 logits below the target, meaning the test was more reliable towards the Some Support cut score than at the Less Support cut score. One Grade 12 item was flagged for having an item-total correlation below the 0.2 threshold. The infit, outfit, PCM item discrimination, and lower asymptote model fit statistics all did not meet their goals, and more items were flagged than at Grades 6 and 9 combined. The increase in model-fit flags was due to the relatively large percentage of items that were flagged for the 2019 NJSLA–S operational test, as was documented in the 2019 NJSLA–S Technical Report (NJDOE, 2019). The median test time was

only 20.28 minutes, which was over 20 minutes below the 45-minute constraint. Of 100 DIF classifications, there were zero "C" values and seven "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 12.

Table 2.4.18: 2021 NJSSA–S Grade 12 Test Construction Statistics

Statistics	Average	Target	Flags
Item Difficulty	–0.28	0.08	N/A
Item Total Correlation	0.43	> 0.35	1
Infit	1.02	1.00	1
Outfit	1.01	1.00	3
Item Discrimination	0.95	1.00	4
Lower Asymptote	0.05	0.00	6
Median Time	20.28	< 45	N/A

Table 2.4.19: 2021 NJSSA–S Grade 12 Test Construction DIF Classifications

Groups	A	B	C
Male/Female	24	1	0
White/Black	23	2	0
White/Hispanic	23	2	0
White/Asian	23	2	0

2.5 Test Administration

The Start Strong Assessments were available for administration using Pearson's TestNav online delivery system. Support for educators, students, and caregivers was available at https://nj.mypearsonsupport.com/startStrong/. The Getting Started Online section included tutorials and practice tests to familiarize students with the online testing experience.

2.6 Test Registration

Student registration consisted of a simplified version of the normal NJSLA summative assessment registration process, with a streamlined Student Registration/Personal Needs Profile (SR/PNP) using PearsonAccess^{next}. Teachers created sessions at the classroom level, generated testing tickets, and provided login information for students to take the assessment at home or in the classroom. Students accessed the assessments online with the teacher-provided usernames and passwords.

2.7 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. To ensure that the diverse population of students taking the NJSSA–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.

2.7.1 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.

2.7.1.1 Text-to-Speech

The most used NJSSA–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.

2.7.2 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 NJSSA–S had an appropriate accommodation. No physical test materials were automatically shipped for Start Strong. Test coordinators placed orders in PAN for braille and large-print test kits.

A comprehensive explanation of each NJSSA–S accommodation is presented in the NJSLA–S *AF&A Manual*. The NJSSA–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, which include:

- Braille
- Large Print
- Spanish
- Spanish Large Print

2.7.2.1 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.

2.7.2.1.1 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, an 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.

2.7.2.1.2 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 2021 NJSSA–S, all items were able to be brailled.

2.7.2.1.3 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.

2.7.2.1.4 Paper-Based Test. The conversion of the NJSSA–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.

2.8 Administration

Administration procedures were standardized as much as possible. Two administration guides were made available—Start Strong PBT User Guide and Start Strong CBT User Guide, as well as Start Strong Administration Policies. The user guides provide detailed instructions for starting and ending the administration, as well as allowable and unallowable supports that may be provided to students while taking the assessment. Like the NJSLA summative assessment, assistance that supported student responses was discouraged because any deviation from normal administration conditions threatens inferences made from the results.

Additionally, an app-based testing lockdown of the desktop may have occurred to provide a focused testing experience if recommended by the teacher. However, it was not a required functionality for the Start Strong assessments. Therefore, it is important to note when interpreting the results that the Start Strong administrations are considered nonsecure.

2.9 Scores and Score Reports

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, which then prepares the files for psychometric analysis.

2.9.1 Scores

Student performance is reported using an overall raw score (i.e., number of points earned). While the raw score can be used to compare students who took the same assessment (e.g., Grade 6 Science), it cannot be used to compare students from a Science assessment to students who took the mathematics assessment, nor can it be used to compare students in 6th grade to 9th grade. Because the Start Strong Assessment is a classroom assessment for gauging where students are in their learning of previous content standards, converting the raw score to a percent correct for the purpose of assigning a grade is not appropriate.

2.9.2 Support Level

Students are categorized into one of three support levels based on their individual total raw scores. Each support level is defined by a range of overall raw scores. There are three support levels for the Start Strong Assessment:

- Level 3–Less Support May Be Needed
- Level 2–Some Support May Be Needed
- Level 1–Strong Support May Be Needed

Students performing at a Level 3 may not require additional academic/instructional support in the tested content area while students in Level 1 will likely benefit from additional academic/instructional support in the tested content area.

The Start Strong performance levels are meant to indicate the amount of support a student might require. While these performance levels leverage the NJSLA summative cut scores, they are not intended to assign proficiency or mastery, because the purpose and blueprint of the Start Strong assessments are different from those of the NJSLA.

Part 3: 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 NJSSA–S was designed to support inferences based on the classification of students into three support levels, as has been described throughout this technical brief. The interpretations of the support level classifications are dependent upon the test performing as intended. As was described in Part 2.3, the NJSSA–S was constructed using a combination of Classical Test Theory (CTT) and Item Response Theory (IRT) statistics, along with the numerous content constraints. The following sections detail how well the 2021 NJSSA–S performed based on those CTT and IRT statistics, along with other criteria. Detailed test maps containing item metadata and various statistics are presented in Appendix A.

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 any items or who had their test scores voided were removed from the data set prior to analysis. NJDOE set the threshold for attemptedness as any student who made a legitimate student response to at least one item. Student responses were voided for cheating, security breaches, or other reasons.

3.1 Classical Test Theory Statistics

For each administration, a set of statistics based on CTT was generated prior to item calibration and scaling. The statistics can be grouped into measures of four 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. Descriptions of each type of statistic appear in the following sections.

3.1.1 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 tend to be too challenging or too easy for a population of test-takers, then the reliability and 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. For 0–2-pt TE items, item difficulty is expressed as an item mean. The polytomous flagging criteria involve 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 are intended as a

recommendation, and many productive polytomous items have p-values outside of the .25 to .90 range.

Item discrimination is also important to monitor, because if items are not discriminating between students with 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 (ITOTC), 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 0–2-pt TE items are flagged with correlations below .25.

Two types of tables are presented below. Tables 3.1.1 through 3.1.6 summarize by item type the average item difficulty and discrimination of the 2021 NJSSA–S items. The averages within each of these tables are disaggregated by content domain and scientific practice. Tables 3.1.7 through 3.1.18 summarize frequency distributions for MC and TE item difficulty and discrimination; they are also disaggregated by content domain and scientific practice.

The average MC and TE item difficulties and discriminations indicate that the items are productive for measuring students in New Jersey. At Grade 6, the average TE item tended to be more challenging, and more discriminating, than the MC items. At Grade 9, the TE items were more challenging than the MC items; however, the MC items were slightly more discriminating. At Grade 12, item difficulties and discriminations displayed the same pattern as at Grade 6, with the TE items being, on average, more challenging and more discriminating than the MC items.

The frequency distributions of item-total correlations also indicate that the items are productive for discriminating between high- and low-achieving students. Only one item at Grade 9 and three at Grade 12 had correlations below .20. Grade 6 had zero items below .20. The p-value distributions, however, were less positive. At Grade 12 there was only one item that had a p-value above .75, and zero below .25, indicating almost no items at the easier and harder ends of the scale. At Grade 9, five of the 16 TE items had p-values below .25, meaning that almost 33% of the Grade 9 TE items were extremely challenging for New Jersey students. Most of the Grade 6 items fell between .25 and .75; only two of the TE items had p-values below .25.

Table 3.1.1: Grade 6 Item Difficulty and Discrimination Summary Statistics by Domain/Practice, MC

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSSA–S	7	.66	.10	.43
Earth and Space	4	.63	.11	.40
Life	1	.72	N/A	.49
Physical	2	.69	.08	.45
Investigating	2	.62	.18	.38
Sensemaking	4	.67	.07	.44
Critiquing	1	.72	N/A	.49

Table 3.1.2: Grade 6 Item Difficulty and Discrimination Summary Statistics by Domain/Practice, TE

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSSA–S	18	.48	.17	.45
Earth and Space	4	.43	.17	.39
Life	6	.56	.11	.42
Physical	8	.45	.19	.49
Investigating	7	.48	.07	.48
Sensemaking	6	.55	.22	.40
Critiquing	5	.41	.19	.44

Table 3.1.3: Grade 9 Item Difficulty and Discrimination Summary Statistics by Domain/Practice, MC

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSSA–S	8	.47	.15	.37
Earth and Space	4	.53	.19	.40
Life	1	.38	N/A	.26
Physical	3	.43	.11	.37
Investigating	6	.41	.10	.34
Sensemaking	1	.75	N/A	.45
Critiquing	1	.55	N/A	.49

Table 3.1.4: Grade 9 Item Difficulty and Discrimination Summary Statistics by Domain/Practice, TE

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSSA–S	16	.38	.20	.34
Earth and Space	4	.40	.18	.30
Life	6	.48	.26	.41
Physical	6	.28	.10	.29
Investigating	3	.19	.06	.28
Sensemaking	9	.49	.20	.36
Critiquing	4	.30	.13	.33

Table 3.1.5: Grade 12 Item Difficulty and Discrimination Summary Statistics by Domain/Practice, MC

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSSA–S	10	.54	.12	.31
Earth and Space	5	.60	.10	.38
Life	1	.65	N/A	.37
Physical	4	.43	.03	.20
Investigating	4	.56	.13	.26
Sensemaking	5	.54	.12	.35
Critiquing	1	.41	N/A	.30

Table 3.1.6: Grade 12 Item Difficulty and Discrimination Summary Statistics by Domain/Practice, TE

Domain/Practice	# Items	Item Difficulty Mean	Item Difficulty S.D.	Item Discrimination Mean
NJSSA–S	15	.49	.12	.40
Earth and Space	5	.43	.12	.43
Life	6	.58	.11	.38
Physical	4	.43	.07	.39
Investigating	4	.48	.07	.46
Sensemaking	6	.51	.11	.37
Critiquing	5	.49	.19	.39

Table 3.1.7: Grade 6 Difficulty Indices by Domain/Practice, MC

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSSA–S	7	.72	0	1	6	0	0
Earth and Space	4	.65	0	1	3	0	0
Life	1	.72	0	0	1	0	0
Physical	2	.69	0	0	2	0	0
Investigating	2	.62	0	1	1	0	0
Sensemaking	4	.67	0	0	4	0	0
Critiquing	1	.72	0	0	1	0	0

Table 3.1.8: Grade 6 Difficulty Indices by Domain/Practice, TE

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSSA–S	18	.51	2	7	8	1	0
Earth and Space	4	.44	0	2	2	0	0
Life	6	.55	0	2	3	1	0
Physical	8	.47	2	3	3	0	0
Investigating	7	.50	0	4	3	0	0
Sensemaking	6	.63	0	2	3	1	0
Critiquing	5	.47	2	1	2	0	0

Table 3.1.9: Grade 6 Discrimination Indices by Domain/Practice, MC

Domain/Practice	# Items	Median	ITOTC **<.20**	**.20<=** ITOTC **<.30**	**.30<=** ITOTC **<.40**	**.40<=** ITOTC **<.50**	ITOTC **>=.50**
NJSSA–S	7	.41	0	1	2	2	2
Earth and Space	4	.39	0	1	1	1	1
Life	1	.49	0	0	0	1	0
Physical	2	.45	0	0	1	0	1
Investigating	2	.38	0	0	1	1	0
Sensemaking	4	.45	0	1	1	0	2
Critiquing	1	.49	0	0	0	1	0

Table 3.1.10: Grade 6 Discrimination Indices by Domain/Practice, TE

Domain/Practice	# Items	Median	ITOTC <.20	.20<= ITOTC <.30	.30<= ITOTC <.40	.40<= ITOTC <.50	ITOTC >=.50
NJSSA–S	18	.45	0	2	3	7	6
Earth and Space	4	.36	0	2	0	1	1
Life	6	.42	0	0	2	3	1
Physical	8	.50	0	0	1	3	4
Investigating	7	.53	0	0	1	2	4
Sensemaking	6	.41	0	2	1	2	1
Critiquing	5	.46	0	0	1	3	1

Table 3.1.11: Grade 9 Difficulty Indices by Domain/Practice, MC

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSSA–S	8	.40	0	5	3	0	0
Earth and Space	4	.57	0	2	2	0	0
Life	1	.38	0	1	0	0	0
Physical	3	.41	0	2	1	0	0
Investigating	6	.38	0	5	1	0	0
Sensemaking	1	.75	0	0	1	0	0
Critiquing	1	.55	0	0	1	0	0

Table 3.1.12: Grade 9 Difficulty Indices by Domain/Practice, TE

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSSA–S	16	.36	5	8	2	1	0
Earth and Space	4	.40	1	2	1	0	0
Life	6	.40	1	3	1	1	0
Physical	6	.26	3	3	0	0	0
Investigating	3	.19	3	0	0	0	0
Sensemaking	9	.42	0	6	2	1	0
Critiquing	4	.28	2	2	0	0	0

Table 3.1.13: Grade 9 Discrimination Indices by Domain/Practice, MC

Domain/Practice	# Items	Median	ITOTC **<.20**	**.20<=** ITOTC **<.30**	**.30<=** ITOTC **<.40**	**.40<=** ITOTC **<.50**	ITOTC **>=.50**
NJSSA–S	8	.39	0	2	2	4	0
Earth and Space	4	.41	0	0	2	2	0
Life	1	.26	0	1	0	0	0
Physical	3	.41	0	1	0	2	0
Investigating	6	.36	0	2	2	2	0
Sensemaking	1	.45	0	0	0	1	0
Critiquing	1	.49	0	0	0	1	0

Table 3.1.14: Grade 9 Discrimination Indices by Domain/Practice, TE

Domain/Practice	# Items	Median	ITOTC **<.20**	**.20<=** ITOTC **<.30**	**.30<=** ITOTC **<.40**	**.40<=** ITOTC **<.50**	ITOTC **>=.50**
NJSSA–S	16	.35	1	3	8	3	1
Earth and Space	4	.30	0	2	2	0	0
Life	6	.42	0	0	3	3	0
Physical	6	.30	1	1	3	0	1
Investigating	3	.22	1	1	0	0	1
Sensemaking	9	.36	0	1	5	3	0
Critiquing	4	.34	0	1	3	0	0

Table 3.1.15: Grade 12 Difficulty Indices by Domain/Practice, MC

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSSA–S	10	.52	0	5	5	0	0
Earth and Space	5	.63	0	1	4	0	0
Life	1	.65	0	0	1	0	0
Physical	4	.42	0	4	0	0	0
Investigating	4	.55	0	2	2	0	0
Sensemaking	5	.55	0	2	3	0	0
Critiquing	1	.41	0	1	0	0	0

Table 3.1.16: Grade 12 Difficulty Indices by Domain/Practice, TE

Domain/Practice	# Items	Median	p<.25	.25<=p<.50	.50<=p<.75	.75<=p<.90	p>=.90
NJSSA–S	15	.52	0	7	7	1	0
Earth and Space	5	.38	0	3	2	0	0
Life	6	.59	0	1	4	1	0
Physical	4	.42	0	3	1	0	0
Investigating	4	.48	0	2	2	0	0
Sensemaking	6	.55	0	2	4	0	0
Critiquing	5	.43	0	3	1	1	0

Table 3.1.17: Grade 12 Discrimination Indices by Domain/Practice, MC

Domain/Practice	# Items	Median	ITOTC **<.20**	**.20<=** ITOTC **<.30**	**.30<=** ITOTC **<.40**	**.40<=** ITOTC **<.50**	ITOTC **>=.50**
NJSSA–S	10	.36	2	1	6	1	0
Earth and Space	5	.37	0	0	4	1	0
Life	1	.37	0	0	1	0	0
Physical	4	.22	2	1	1	0	0
Investigating	4	.30	1	1	2	0	0
Sensemaking	5	.37	1	0	3	1	0
Critiquing	1	.30	0	0	1	0	0

Table 3.1.18: Grade 12 Discrimination Indices by Domain/Practice, TE

Domain/Practice	**# Items**	**Median**	ITOTC **<.20**	**.20<=** ITOTC **<.30**	**.30<=** ITOTC **<.40**	**.40<=** ITOTC **<.50**	ITOTC **>=.50**
NJSSA–S	15	.42	1	2	4	6	2
Earth and Space	5	.46	0	0	2	3	0
Life	6	.37	0	2	1	3	0
Physical	4	.42	1	0	1	0	2
Investigating	4	.49	0	0	1	1	2
Sensemaking	6	.39	1	1	1	3	0
Critiquing	5	.39	0	1	2	2	0

3.1.2 Speededness

The consequence of time limits on examinees' scores is called speededness. A test is "speeded" to the degree that those taking the test 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 an 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.

NJSSA–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. NJSSA–S assessments were administered during a testing window with 60 minutes of testing time at each grade level. Students were assumed to have enough time to complete the test.

That assumption was tested by calculating the percentage of students omitting the last item on the test. As shown in Table 3.1.19, Grade 6 had the highest percentage of students omitting the last item at only 1.22%. This is clear evidence of the NJSSA–S being a non-speeded power test at all grade levels.

Table 3.1.19: Percentage of Students Omitting the Last Item

Grade	Location	%
6	25	1.22
9	24	1.16
12	25	0.55

3.1.3 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; Zieky, 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 negligible 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 0–2pt TE 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.

DIF detection for the NJSSA–S operational test focused on seven comparisons of students.

- Male/Female
- White/Black
- White/Hispanic
- White/Asian
- 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)

The results of the DIF analyses were positive with the exception of a small number of items classified as "C." There were no C-DIF classifications at Grade 6, and three each at Grades 9 and 12. At Grade 9, the EL-No/EL-Yes comparison was the most problematic with two C-DIF and three B-DIF items. Moreover, one C-DIF item was identified for the SWD-Yes/SWD-No comparison. At Grade 12, C-DIF was identified for one White/Asian and two EL-No/EL-Yes comparisons. No C-DIF classifications were identified for the Male/Female, White/Black, White/Hispanic, and not economically disadvantaged/economically disadvantaged comparisons. Table 3.1.20 shows the DIF classifications for all seven comparison groups by grade.

Table 3.1.20: DIF Classification by Grade and Item Type

Grade	Group	Item Type	A	B	C
6	Male/Female	MC	7	0	0
6	Male/Female	TE	17	1	0
6	Male/Female	Total	24	1	0
6	White/Black	MC	6	1	0
6	White/Black	TE	18	0	0
6	White/Black	Total	24	1	0
6	White/Hispanic	MC	6	1	0
6	White/Hispanic	TE	18	0	0
6	White/Hispanic	Total	24	1	0
6	White/Asian	MC	7	0	0
6	White/Asian	TE	18	0	0
6	White/Asian	Total	25	0	0
6	EL-No/EL-Yes	MC	6	1	0
6	EL-No/EL-Yes	TE	16	2	0
6	EL-No/EL-Yes	Total	22	3	0
6	SWD-No/SWD-Yes	MC	7	0	0
6	SWD-No/SWD-Yes	TE	18	0	0
6	SWD-No/SWD-Yes	Total	25	0	0
6	EconDis-No/EconDis-Yes	MC	7	0	0
6	EconDis-No/EconDis-Yes	TE	18	0	0
6	EconDis-No/EconDis-Yes	Total	25	0	0

Grade	Group	Item Type	A	B	C
9	Male/Female	MC	8	0	0
9	Male/Female	TE	16	0	0
9	Male/Female	Total	24	0	0
9	White/Black	MC	7	1	0
9	White/Black	TE	15	1	0
9	White/Black	Total	22	2	0
9	White/Hispanic	MC	8	0	0
9	White/Hispanic	TE	16	0	0
9	White/Hispanic	Total	24	0	0
9	White/Asian	MC	8	0	0
9	White/Asian	TE	16	0	0
9	White/Asian	Total	24	0	0
9	EL-No/EL-Yes	MC	6	2	0
9	EL-No/EL-Yes	TE	13	1	2
9	EL-No/EL-Yes	Total	19	3	2
9	SWD-No/SWD-Yes	MC	8	0	0
9	SWD-No/SWD-Yes	TE	15	0	1
9	SWD-No/SWD-Yes	Total	23	0	1
9	EconDis-No/EconDis-Yes	MC	8	0	0
9	EconDis-No/EconDis-Yes	TE	16	0	0
9	EconDis-No/EconDis-Yes	Total	24	0	0

Grade	Group	Item Type	A	B	C
12	Male/Female	MC	9	1	0
12	Male/Female	TE	14	1	0
12	Male/Female	Total	23	2	0
12	White/Black	MC	10	0	0
12	White/Black	TE	14	1	0
12	White/Black	Total	24	1	0
12	White/Hispanic	MC	10	0	0
12	White/Hispanic	TE	14	1	0
12	White/Hispanic	Total	24	1	0
12	White/Asian	MC	10	0	0
12	White/Asian	TE	14	0	1
12	White/Asian	Total	24	0	1
12	EL-No/EL-Yes	MC	9	1	0
12	EL-No/EL-Yes	TE	9	4	2
12	EL-No/EL-Yes	Total	18	5	2
12	SWD-No/SWD-Yes	MC	10	0	0
12	SWD-No/SWD-Yes	TE	15	0	0
12	SWD-No/SWD-Yes	Total	25	0	0
12	EconDis-No/EconDis-Yes	MC	10	0	0
12	EconDis-No/EconDis-Yes	TE	15	0	0
12	EconDis-No/EconDis-Yes	Total	25	0	0

3.2 Item Response Theory

The grade-specific NJSSA–S student ability estimates are calibrated via Item Response Theory (IRT) statistical processes. Part 3.2 of this report explains how IRT is used in the context of the NJSSA–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 NJSSA–S is described in conjunction with the assumptions that the model must meet in order to be applicable. The remainder of Part 3.2 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 NJSSA–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 3.1 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 5.1 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 (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 NJSSA–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 5.2 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 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 NJSSA–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 NJSSA–S item types are designed to minimize guessing, and the test contains polytomous items (e.g., 0–2pt TE items). If the PCM's assumptions are met, it is likely a good IRT model to use with the NJSSA–S.

The main assumptions of the PCM as they apply to the NJSSA–S are that the test is unidimensional, the items discriminate relatively equally, guessing on items is minimal, the response to each individual item is independent of the others, and the resulting item parameter estimates are invariant regardless of who answered the items. Each of these five major IRT assumptions will be explained in greater detail in the sections below as they relate to the PCM. The final component within this section shows disaggregated descriptive statistics of the raw scores. Overall, the results of the 2021 NJSSA–S indicate that the assumptions of the PCM were adequately met.

3.2.1 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 NJSSA–S is a unidimensional test and that the PCM assumption of unidimensionality has been met.

3.2.1.1 Intercorrelations. The Pearson product-moment correlations among the domains and practices are presented in Tables 3.2.1 through 3.2.3. 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 6, all domains and practices correlated with the total test score at 0.85 or above. Relatively high correlations between the domains or practices and the total test score were also present at both Grades 9 and 12. The lowest correlation among any subscore and the total test score was with Critiquing at Grade 9. The intercorrelations among subscores provide strong evidence that the NJSSA–S is a unidimensional test.

Table 3.2.1: Grade 6 Correlation Matrix for Domains and Practices

Content	NJSSA–S	Earth	Life	Physical	Investigating	Sensemaking	Critiquing
NJSSA–S	1	N/A	N/A	N/A	N/A	N/A	N/A
Earth and Space	.86	1	N/A	N/A	N/A	N/A	N/A
Life	.86	.63	1	N/A	N/A	N/A	N/A
Physical	.92	.68	.68	1	N/A	N/A	N/A
Investigating	.92	.72	.81	.89	1	N/A	N/A
Sensemaking	.91	.90	.71	.81	.74	1	N/A
Critiquing	.85	.68	.81	.76	.69	.67	1

Table 3.2.2: Grade 9 Correlation Matrix for Domains and Practices

Content	NJSSA–S	Earth	Life	Physical	Investigating	Sensemaking	Critiquing
NJSSA–S	1	N/A	N/A	N/A	N/A	N/A	N/A
Earth and Space	.86	1	N/A	N/A	N/A	N/A	N/A
Life	.84	.61	1	N/A	N/A	N/A	N/A
Physical	.86	.58	.58	1	N/A	N/A	N/A
Investigating	.85	.77	.64	.75	1	N/A	N/A
Sensemaking	.89	.71	.85	.73	.59	1	N/A
Critiquing	.78	.69	.60	.70	.53	.57	1

Table 3.2.3: Grade 12 Correlation Matrix for Domains and Practices

Content	NJSSA–S	Earth	Life	Physical	Investigating	Sensemaking	Critiquing
NJSSA–S	1	N/A	N/A	N/A	N/A	N/A	N/A
Earth and Space	.90	1	N/A	N/A	N/A	N/A	N/A
Life	.83	.63	1	N/A	N/A	N/A	N/A
Physical	.81	.58	.52	1	N/A	N/A	N/A
Investigating	.86	.76	.59	.82	1	N/A	N/A
Sensemaking	.91	.83	.79	.67	.64	1	N/A
Critiquing	.81	.71	.76	.59	.58	.61	1

3.2.1.2 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 NJSSA–S. The scree plots were interpreted by finding the place on the plot where the slope leveled 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 close to 1, whereas the most prominent eigenvalues range from approximately 5–6.

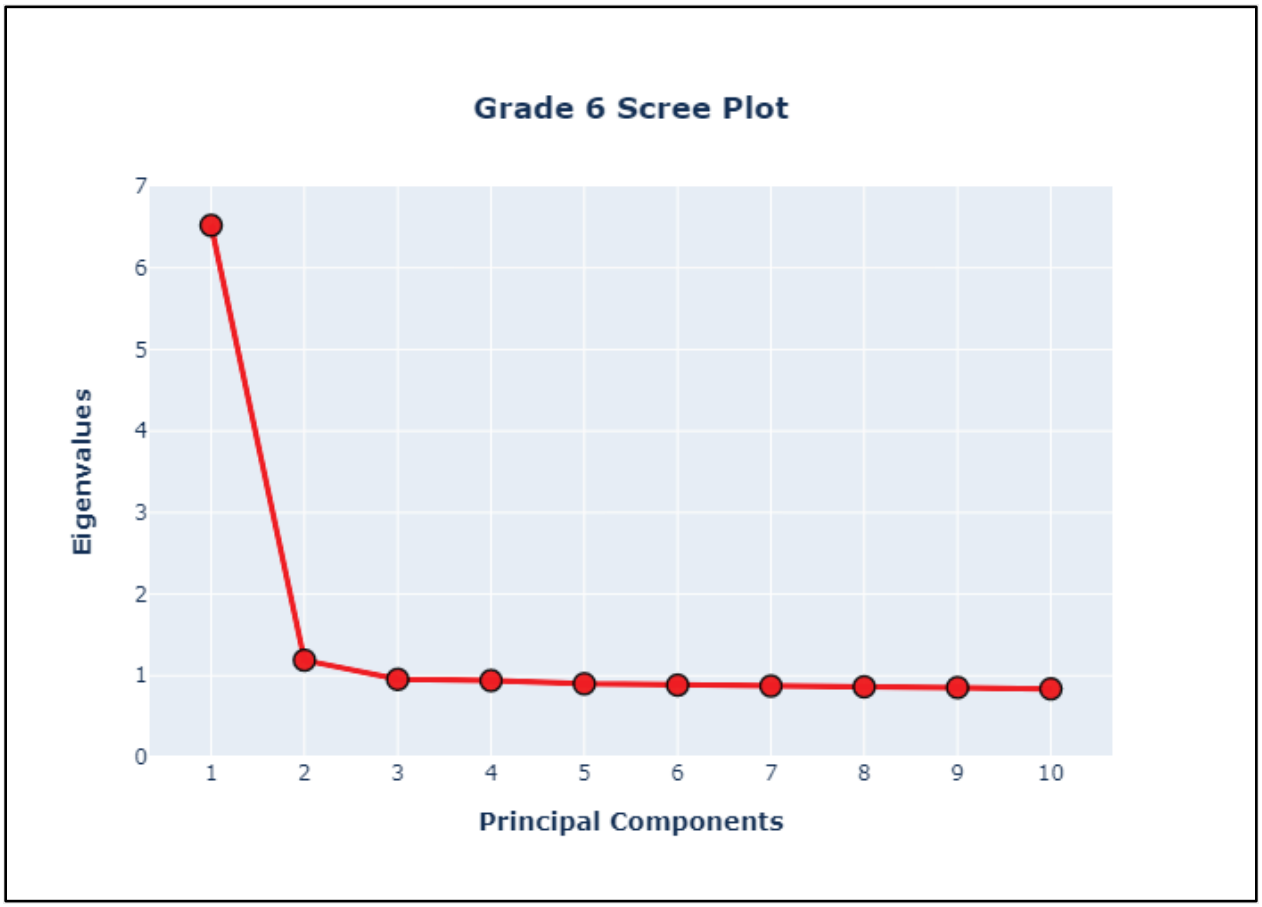

Figure 3.2.1. Grade 6 Scree Plot

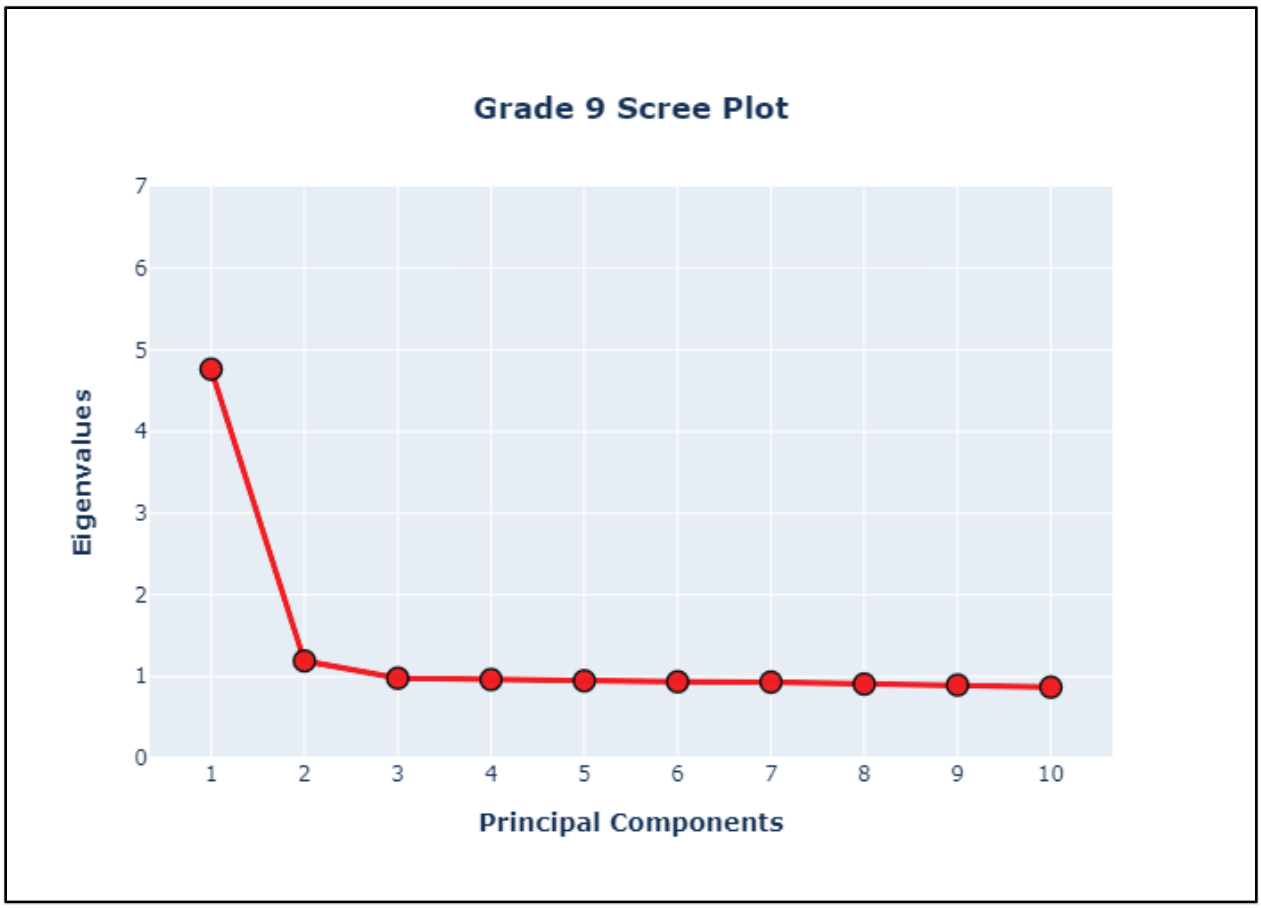

Figure 3.2.2. Grade 9 Scree Plot

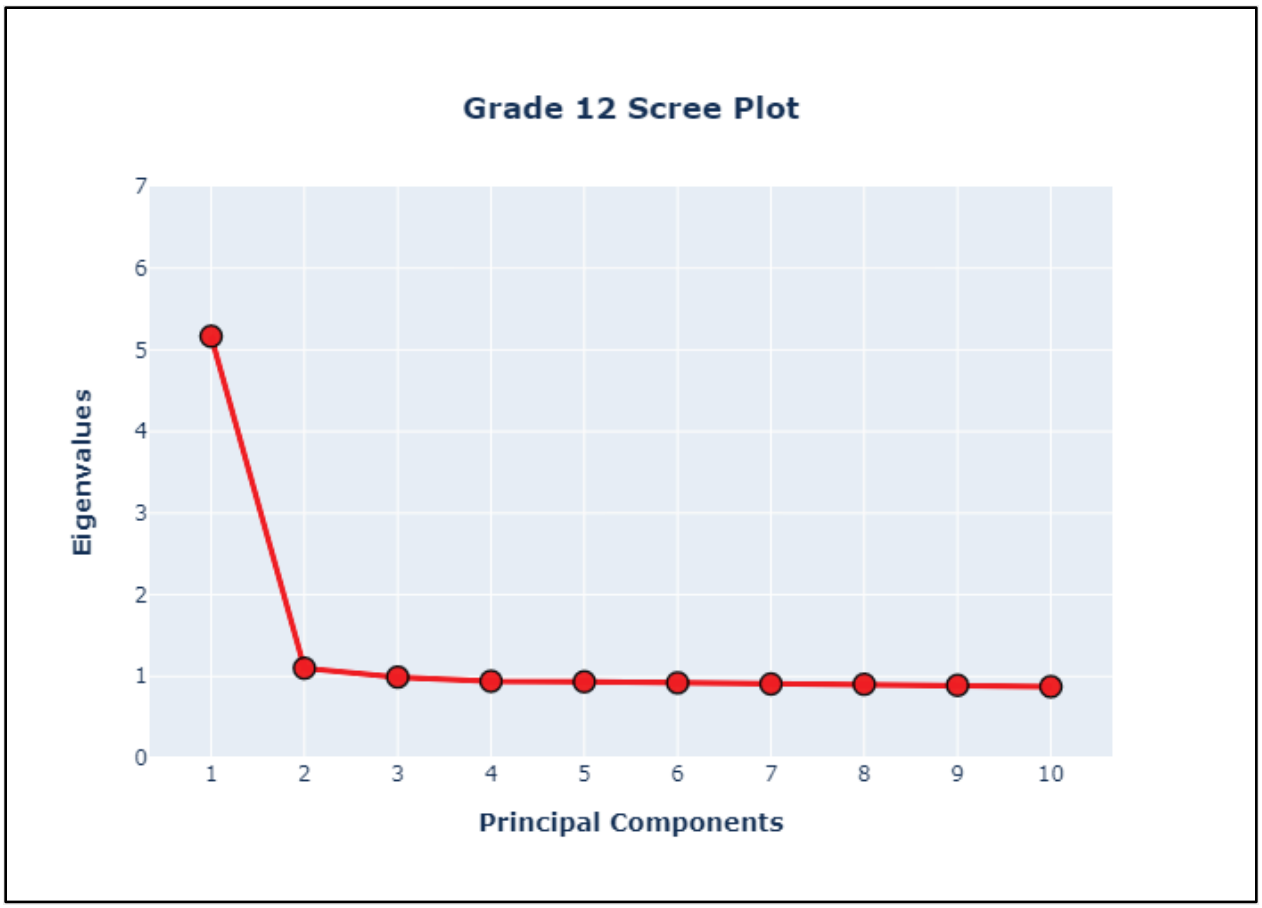

Figure 3.2.3. Grade 12 Scree Plot

3.2.2 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 2021 NJSSA–S IRT calibration processes via Winsteps 3.74 (Linacre, 2012). Detailed item parameter estimates and model fit statistics are presented in Appendix C. Overall, there is ample evidence that the items at all grades fit the assumptions of the PCM, as is described in the following sections.

3.2.2.1 Item infit and outfit. Rasch infit and outfit statistics range from 0 to infinity with 1 representing ideal model fit. For the NJSSA–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 one item across all grades was flagged for problematic infit statistics.

Outfit statistics are heavily influenced by unexpected student responses to items that are either relatively easy or relatively hard. The NJSSA–S outfit statistics were less positive; four Grade 9 items were flagged, whereas at Grade 6, two items were flagged, and at Grade 12, only one item was flagged. Flagged outfit statistics are less of a threat to the validity of test score interpretations than are problematic infit statistics, especially when the flagged items' outfit statistics are only slightly outside the flagging thresholds, as was the case for all flagged items. Thus, while there is clearly room for improving the item outfit, the infit and outfit statistics provide reasonable evidence that the assumptions of the PCM have been met. Table 3.2.4 provides a summary of item infit and outfit statistics at each grade level.

Table 3.2.4: Summary Infit and Outfit Statistics

Grade	Fit Statistics	Mean	Min	Max	Outside 0.7 to 1.3	% Flagged
6	Infit	0.97	0.76	1.19	0 out of 25	0.0%
6	Outfit	0.99	0.72	1.33	2 out of 25	8.0%
9	Infit	0.97	0.74	1.16	0 out of 24	0.0%
9	Outfit	0.98	0.53	1.34	4 out of 24	16.7%
12	Infit	1.01	0.83	1.32	1 out of 25	4.0%
12	Outfit	1.00	0.73	1.45	1 out of 25	4.0%

3.2.2.2 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 each grade, the Rasch discrimination statistics looked excellent, except for one Grade 12 item that was flagged for having a value outside the 0.5 to 1.5 threshold. Table 3.2.5 provides a summary of Rasch discrimination statistics at each grade level.

Table 3.2.5: Summary Rasch Discrimination Statistics

Grade	Fit Statistics	Mean	Min	Max	Outside 0.5 to 1.5	% Flagged
6	Discrimination	1.04	0.56	1.41	0 out of 25	0.0%
9	Discrimination	1.04	0.58	1.49	0 out of 24	0.0%
12	Discrimination	0.97	0.08	1.37	1 out of 25	4.0%

3.2.2.3 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 statistics. The 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 0.0 to 0.1.

At all grades, the lower asymptote statistics met the model assumptions. Only three items out of a total of 74 (4%) were flagged for having values outside the 0.1 threshold. The average lower asymptote statistics at each grade were close to the ideal value of 0.0. Grade 12 had the most items flagged with two out of 25. Table 3.2.6 provides a summary of the lower asymptote statistics at each grade level.

Table 3.2.6: Summary Rasch Lower Asymptote Statistics

Grade	Fit Statistics	Mean	Min	Max	Greater Than 0.1	% Flagged
6	Lower Asymptote	0.02	0.00	0.17	1 out of 25	4.0%
9	Lower Asymptote	0.02	0.00	0.09	0 out of 24	0.0%
12	Lower Asymptote	0.03	0.00	0.21	2 out of 25	8.0%

3.2.2.4 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 3.2.7 and 3.2.8 show person infit and outfit descriptive statistics by demographic variables. Tables 3.2.9 and 3.2.10 break down the person infit and outfit descriptive statistics by Computer-Based Testing (CBT), Paper-Based Testing (PBT), and Spanish forms. Figures 3.2.4 through 3.2.9 show two-dimensional density contours for each grade level for all students of both the person infit and outfit statistics. If

students are flagged for aberrant person-fit statistics, the two-dimensional density contours can help illuminate where along the scale the flags are occurring. Flags near the cut scores are more of a threat to the validity of test score interpretations than flags toward the lower or upper ends of the ability distributions.

Overall, there were very few students flagged for aberrant person infit statistics. Less than 5% of students were flagged for person infit statistics at all combinations of grade and demographic variables. At Grade 12, only three students out of 89,197 were flagged. As shown in Table 3.2.9, no form type at any grade level had more than 5% of students flagged.

The person outfit statistics were less positive, excepting Grade 12 where only 1.62% of students were flagged. At both Grades 6 and 9, large percentages of students were flagged for outfit across several demographic variables and form types. The two-dimensional density contour graphs for the Grades 6 and 9 person outfit statistics presented respectively in Figures 3.2.5 and 3.2.7 show where along the ability spectrum those students were being flagged. At Grade 6, a large proportion of students who received scores well above the Less Support threshold were flagged for having outfit statistics below 0.5. Given that the Grade 6 test did not have many items measuring above the Less Support cut score, this pattern is not unexpected. At Grade 9, the pattern was the opposite. There were few items measuring the lower part of the scale. Unsurprisingly, this is the same place along the student ability spectrum where students were likely flagged for person outfit. This incongruence between the student abilities and the item difficulties is revisited in Part 5.2, Item Response Theory Reliability.

Table 3.2.7: Person Infit Statistics by Demographic Group

Grade	Group	N	Mean Raw Score	Person Infit Mean	Person Infit Min	Person Infit Max	Flagged N	Flagged Percent	Flagged Mean Raw Score
6	NJSSA–S	95,332	13.33	0.98	0.57	1.72	69	0.07	15.71
6	Male	48,628	13.34	0.99	0.57	1.68	42	0.09	15.31
6	Female	46,688	13.32	0.97	0.57	1.72	27	0.06	16.33
6	Am. Indian	173	12.94	0.96	0.63	1.47	0	0.00	N/A
6	Asian	10,359	17.27	0.98	0.58	1.59	9	0.09	16.56
6	Black	13,904	10.49	0.98	0.57	1.64	10	0.07	15.60
6	Hispanic	29,691	11.09	0.98	0.57	1.68	18	0.06	14.06
6	Pacific Islander	186	14.22	0.96	0.58	1.40	0	0.00	N/A
6	White	38,314	14.94	0.97	0.57	1.72	32	0.08	16.44
6	EL – Yes	6,634	7.39	1.00	0.59	1.62	2	0.03	7.50
6	EL – No	88,698	13.77	0.97	0.57	1.72	67	0.08	15.96
6	EconDis – Yes	29,813	10.74	0.98	0.57	1.68	21	0.07	14.81
6	EconDis – No	65,516	14.51	0.98	0.57	1.72	48	0.07	16.10
6	SWD – Yes	19,310	9.88	0.99	0.57	1.61	7	0.04	15.86
6	SWD – No	76,022	14.21	0.97	0.57	1.72	62	0.08	15.69
9	NJSSA–S	100,693	10.16	0.99	0.58	2.43	1,572	1.56	5.84
9	Male	51,305	10.36	1.00	0.58	2.43	911	1.78	5.83
9	Female	49,317	9.95	0.98	0.58	2.38	661	1.34	5.86
9	Am. Indian	155	9.86	0.99	0.65	2.01	4	2.58	5.50
9	Asian	10,404	13.62	0.96	0.58	2.03	48	0.46	7.96
9	Black	14,488	8.05	1.02	0.58	2.36	375	2.59	5.52
9	Hispanic	31,932	8.33	1.02	0.58	2.38	739	2.31	5.46
9	Pacific Islander	207	11.21	0.96	0.64	1.56	2	0.97	9.00
9	White	41,148	11.39	0.97	0.58	2.43	373	0.91	6.67
9	EL – Yes	4,933	5.48	1.12	0.58	2.38	246	4.99	4.84
9	EL – No	95,760	10.41	0.98	0.58	2.43	1,326	1.38	6.03
9	EconDis – Yes	28,926	8.10	1.02	0.58	2.38	729	2.52	5.46
9	EconDis – No	71,765	10.99	0.98	0.58	2.43	843	1.17	6.17

Grade	Group	N	Mean Raw Score	Person Infit Mean	Person Infit Min	Person Infit Max	Flagged N	Flagged Percent	Flagged Mean Raw Score
9	SWD – Yes	18,953	7.89	1.05	0.58	2.43	640	3.38	5.33
9	SWD – No	81,740	10.69	0.98	0.58	2.38	932	1.14	6.19
12	NJSSA–S	89,197	12.76	1.01	0.64	1.54	3	0.00	12.33
12	Male	44,952	12.77	1.00	0.64	1.50	0	0.00	N/A
12	Female	44,205	12.75	1.02	0.66	1.54	3	0.01	12.33
12	Am. Indian	112	12.15	0.99	0.77	1.33	0	0.00	N/A
12	Asian	10,097	16.25	0.98	0.67	1.50	0	0.00	N/A
12	Black	11,404	10.33	1.02	0.67	1.48	0	0.00	N/A
12	Hispanic	24,823	10.71	1.02	0.65	1.53	1	0.00	11.00
12	Pacific Islander	209	13.35	1.01	0.75	1.33	0	0.00	N/A
12	White	41,038	13.77	1.01	0.64	1.54	2	0.00	13.00
12	EL – Yes	3,782	7.40	1.02	0.67	1.50	0	0.00	N/A
12	EL – No	85,415	12.99	1.01	0.64	1.54	3	0.00	12.33
12	EconDis – Yes	23,277	10.63	1.02	0.64	1.50	0	0.00	N/A
12	EconDis – No	65,918	13.51	1.01	0.66	1.54	3	0.00	12.33
12	SWD – Yes	16,927	10.34	1.02	0.65	1.52	1	0.01	14.00
12	SWD – No	72,270	13.32	1.01	0.64	1.54	2	0.00	11.50

Table 3.2.8: Person Outfit Statistics by Demographic Group

Grade	Group	N	Mean Raw Score	Person Outfit Mean	Person Outfit Min	Person Outfit Max	Flagged N	Flagged Percent	Flagged Mean Raw Score
6	NJSSA–S	95,332	13.33	0.99	0.23	9.73	8,909	9.35	13.79
6	Male	48,628	13.34	1.01	0.23	9.73	4,921	10.12	13.84
6	Female	46,688	13.32	0.98	0.23	9.73	3,987	8.54	13.71
6	Non-Binary	16	15.06	1.03	0.61	2.64	1	6.25	22.00
6	Am. Indian	173	12.94	0.99	0.29	3.47	23	13.29	14.74
6	Asian	10,359	17.27	0.98	0.23	5.69	1,305	12.60	20.52
6	Black	13,904	10.49	1.01	0.23	9.73	1,325	9.53	8.07
6	Hispanic	29,691	11.09	1.01	0.23	9.73	2,739	9.23	9.11
6	Pacific Islander	186	14.22	0.94	0.29	2.90	15	8.06	18.13
6	White	38,314	14.94	0.98	0.23	9.73	3,243	8.46	17.09
6	EL – Yes	6,634	7.39	1.06	0.23	9.73	913	13.76	5.18
6	EL – No	88,706	13.78	0.99	0.23	9.73	7,999	9.02	14.77
6	EconDis – Yes	29,813	10.74	1.01	0.23	9.73	2,718	9.12	8.39
6	EconDis – No	65,516	14.51	0.98	0.23	9.73	6,191	9.45	16.15
6	SWD – Yes	19,310	9.88	1.03	0.23	9.73	2,170	11.24	7.99
6	SWD – No	76,022	14.21	0.98	0.23	9.73	6,739	8.86	15.65
9	NJSSA–S	100,693	10.16	0.98	0.12	9.13	7,830	7.78	5.97
9	Male	51,305	10.36	0.99	0.12	9.13	4,326	8.43	6.29
9	Female	49,317	9.95	0.97	0.12	8.24	3,501	7.10	5.58
9	Non-Binary	71	12.90	0.96	0.46	2.22	3	4.23	10.33
9	Am. Indian	155	9.86	1.01	0.39	6.58	9	5.81	6.89
9	Asian	10,404	13.62	0.91	0.12	9.13	450	4.33	13.16
9	Black	14,488	8.05	1.04	0.12	8.24	1,686	11.64	4.83
9	Hispanic	31,932	8.33	1.03	0.12	8.24	3,484	10.91	4.69
9	Pacific Islander	207	11.21	0.91	0.42	2.38	9	4.35	7.11
9	White	41,148	11.39	0.94	0.12	8.24	2,062	5.01	7.42
9	EL – Yes	4,933	5.48	1.20	0.12	6.58	1,107	22.44	3.97
9	EL – No	95,760	10.41	0.97	0.12	9.13	6,724	7.02	6.30
9	EconDis – Yes	28,926	8.10	1.04	0.12	8.24	3,367	11.64	4.72
9	EconDis – No	71,765	10.99	0.96	0.12	9.13	4,463	6.22	6.92

Grade	Group	N	Mean Raw Score	Person Outfit Mean	Person Outfit Min	Person Outfit Max	Flagged N	Flagged Percent	Flagged Mean Raw Score
9	SWD – Yes	18,953	7.89	1.08	0.12	8.24	2,665	14.06	4.83
9	SWD – No	81,740	10.69	0.96	0.12	9.13	5,165	6.32	6.56
12	NJSSA–S	89,197	12.76	1.00	0.40	3.63	1,443	1.62	14.61
12	Male	44,952	12.77	0.99	0.40	3.63	768	1.71	15.15
12	Female	44,205	12.75	1.01	0.40	3.63	674	1.52	13.99
12	Non-Binary	40	15.55	1.06	0.65	1.54	1	2.50	18.00
12	Am. Indian	112	12.15	0.97	0.49	1.71	2	1.79	23.00
12	Asian	10,097	16.25	0.95	0.41	3.31	306	3.03	22.23
12	Black	11,404	10.33	1.02	0.40	2.34	141	1.24	5.50
12	Hispanic	24,823	10.71	1.01	0.40	3.31	317	1.28	7.98
12	Pacific Islander	209	13.35	1.00	0.49	1.85	2	0.96	13.50
12	White	41,038	13.77	1.00	0.40	3.63	643	1.57	16.06
12	EL – Yes	3,782	7.40	1.02	0.40	2.30	71	1.88	3.72
12	EL – No	85,430	13.00	1.00	0.40	3.63	1,372	1.61	15.17
12	EconDis – Yes	23,277	10.63	1.01	0.40	3.31	298	1.28	7.06
12	EconDis – No	65,918	13.51	0.99	0.40	3.63	1,145	1.74	16.57
12	SWD – Yes	16,927	10.34	1.02	0.40	3.31	283	1.67	7.73
12	SWD – No	72,270	13.32	0.99	0.40	3.63	1,160	1.61	16.29

Table 3.2.9: Person Infit Statistics by Form

Grade	Group	N	Mean Raw Score	Person Infit Mean	Person Infit Min	Person Infit Max	Flagged N	Flagged Percent	Flagged Mean Raw Score
6	CBT	94,070	13.42	0.98	0.57	1.72	69	0.07	15.71
6	PBT	174	8.44	1.03	0.66	1.43	0	0.00	N/A
6	SP	1068	6.70	0.99	0.67	1.47	0	0.00	N/A
9	CBT	99,547	10.22	0.99	0.58	2.43	1,518	1.52	5.89
9	PBT	40	10.03	1.01	0.71	1.55	1	2.50	3.00
9	SP	1,062	5.39	1.12	0.58	2.2	50	4.71	4.76
12	CBT	88,469	12.80	1.01	0.64	1.54	3	0.00	12.33
12	PBT	60	10.22	0.99	0.8	1.23	0	0.00	N/A
12	SP	561	7.71	1.05	0.75	1.5	0	0.00	N/A

Table 3.2.10: Person Outfit Statistics by Form

Grade	Group	N	Mean Raw Score	Person Outfit Mean	Person Outfit Min	Person Outfit Max	Flagged N	Flagged Percent	Flagged Mean Raw Score
6	CBT	94,070	13.42	0.99	0.23	9.73	8,744	9.30	13.96
6	PBT	174	8.44	1.12	0.54	6.21	21	12.07	5.86
6	SP	1,068	6.70	1.03	0.37	6.21	139	13.01	4.65
9	CBT	99,547	10.22	0.98	0.12	9.13	7,569	7.60	6.05
9	PBT	40	10.03	1.01	0.49	2.34	5	12.50	4.20
9	SP	1,062	5.39	1.20	0.12	6.58	244	22.98	3.79
12	CBT	88,469	12.80	1.00	0.4	3.63	1,421	1.61	14.78
12	PBT	60	10.22	0.98	0.69	1.56	1	1.67	4.00
12	SP	561	7.71	1.06	0.43	1.89	17	3.03	3.76

3.2.3 Local Independence

The PCM assumes that student responses to items are independent of responses to other items. In other words, student performance on one item does not affect performance on the other items on the test. If the assumption of local independence is violated, then 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 only one of 876 residual correlations between items displayed a Q3 value outside the acceptable threshold. Table 3.2.11 summarizes Yen's Q3 statistics at each grade level.

Table 3.2.11: Summary of the Yen's Q3 Statistics

Grade	Mean	Min	Max	Outside–0.2 to 0.2	% Flagged
6	–.04	–.12	.23	1 out of 300	.3%
9	–.04	–.12	.08	0 out of 276	.0%
12	–.04	–.15	.19	0 out of 300	.0%

3.2.4 Descriptive Statistics — Raw Score

This section contains descriptive statistics for raw scores and support-level distributions by form and links to NJDOE documentation detailing student support-level percentages by demographic group.

3.2.4.1 Raw score distributions by form. Descriptive statistics for raw scores and percentage distributions of students' support levels are summarized by form in Table 3.2.12. For all test forms, raw scores have a range of 0 to 25. The cut scores for each support level can be found in Part 5.3.1 of this report.

Table 3.2.12: Descriptive Statistics of Raw Scores and Students' Support Levels by Form

Grade	Form	N+	Mean	SD	Min	Max	%Strong Support	%Some Support	%Less Support
6	CBT	94,070	13.42	5.90	0	25	42.66	33.82	23.52
6	PBT	174	8.44	4.67	1	23	78.74	17.82	3.45
6	SP	1,068	6.70	3.87	0	21	90.92	8.15	0.94
9	CBT	99,547	10.22	4.82	0	25	41.63	42.22	16.15
9	PBT	40	10.03	4.90	0	20	35.00	55.00	10.00
9	SP	1,062	5.39	2.81	0	17	86.63	13.18	0.19
12	CBT	88,469	12.80	5.32	0	25	49.18	23.83	26.99
12	PBT	60	10.22	4.70	3	22	66.67	21.67	11.67
12	SP	561	7.71	3.22	1	20	92.34	6.77	0.89

* CBT: Computer-Based Test; PBT: Paper-Based Test; SP: Spanish

3.2.4.2 Raw score distributions by demographic group. Percentage distributions of students' support levels by demographic groups can be found on the New Jersey Statewide Assessment Reports webpage. Raw score cumulative frequency distributions are attached as Appendix B in this technical brief.

Part 4: Scale Stability

Standard 5.6 states that "Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported" (AERA, APA, NCME, 2014, p.103).Moreover, as described in Part 2 of this technical brief, the items had previously been administered on the 2019 NJSLA–S operational test forms, as well as the 2018 NJSLA–S field test. That means different cohorts of students in New Jersey have been tested via this set of items for four consecutive years. Thus, in order to ensure the comparability of NJSSA–S test scores, the stability of the underlying item difficulty parameters warranted checking.

Kolen and Brennan (2004) recommend that scale stability be inspected both statistically and visually. The methods used for testing the scale stability of the NJSSA–S are ideal for the types of evaluation they recommend. The first method was the Delta Plot method as described by Angoff (1982). The second method was the 0.3 Logit Absolute Difference criterion as described by Miller, Rotou, and Twing (2004). The former is a CTT-based method for detecting discrepancies between group item means and p-values. The latter is a Rasch, IRT-based method for detecting item parameter drift.

The following sections describe both methods and present the results. Both the results of the Delta Plot and 0.3 Logit Absolute Difference methods were generally positive. Moreover, the person and item fit statistics presented in Part 3.2 provide similarly positive evidence of scale stability. Overall, the NJSSA–S scale appears to have been stable.

4.1 Delta Plot Method

The Delta Plot method—also referred to as the Transformed Item Difficulty Index—was introduced by Angoff (1972). It was originally conceptualized as a method for identifying biased test items within the framework of CTT. Per the recommendations of the New Jersey Technical Advisory Committee, the Delta Plot method was added to the scale stability analysis for the purpose of complementing the IRT methods, including the fit statistics described in Part 3.2 of this technical brief, as well as the results of the 0.3 Logits Absolute Difference Method described in Part 4.2.

The Delta Plot method uses multiple transformations to place p-values onto the "delta" scale, which is a common scale used by Educational Testing Service (ETS) that has a mean of 13 and standard deviation of 4. Once the item p-values have been transformed, it is typical to plot the values onto a scatter plot and then create a trendline (Camilli & Shepard, 1994). The perpendicular distance (PD) of each item from the trendline is calculated; items are flagged if their perpendicular distance is two standard deviation units of PDs away from the trendline.

The results of the Delta Plot method were generally positive. Table 4.1.1 shows a summary of the results, including the percentage of items that were flagged at each grade level. The

"Flagging Distance" column shows the distance (i.e., two standard deviation units of PDs) from the trendline that served as the flagging threshold for each grade level. Figures 4.1.1 to 4.1.3 respectively show the delta plots for each grade level. At Grade 6, the two items that were flagged were slightly more challenging than anticipated, given that students generally did worse on all of the items on the 2021 NJSSA–S in comparison to the students who took them on the 2019 NJSLA–S. At Grade 9, one item was flagged. It is one of the several challenging items on the test. At Grade 12, zero items were flagged. A visual inspection of Grade 12 shows that the lack of flags is reasonable. Appendix D contains the detailed results of the Delta Plot method for each grade level. It should be noted that item means for 0–2 point TE items were converted to adjusted p-value (i.e., item mean divided by max score point) for running the Delta Plot method analyses.

Table 4.1.1: Summary of Delta Plot Method

Grade	Total Items	Flagged Items	Percent Flagged	Flagging Distance
6	25	2	8.0%	0.090
9	24	1	0.4%	0.086
12	25	0	0.0%	0.063

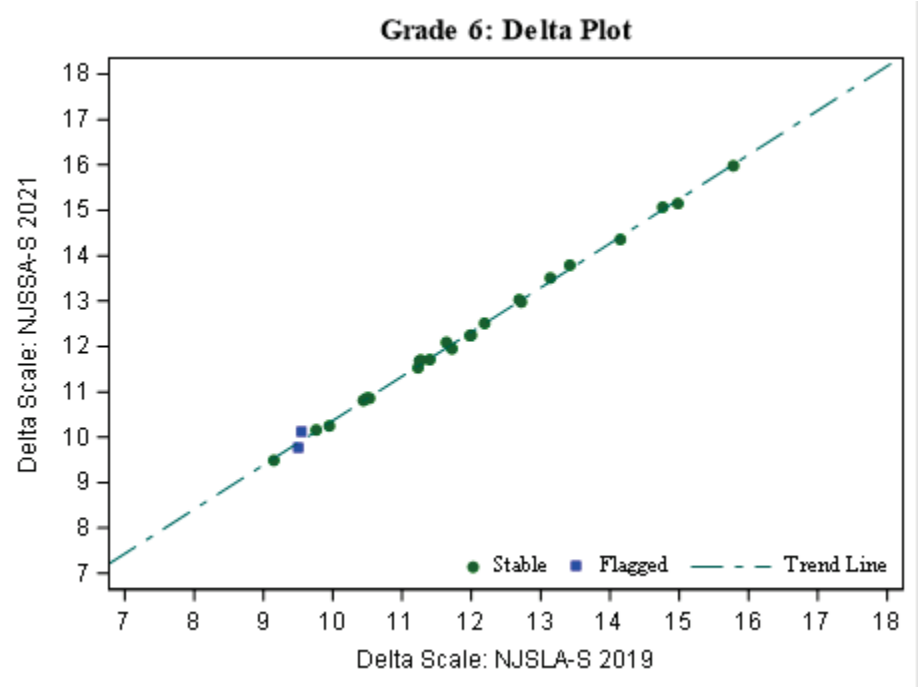

Figure 4.1.1. Grade 6 Delta Plot

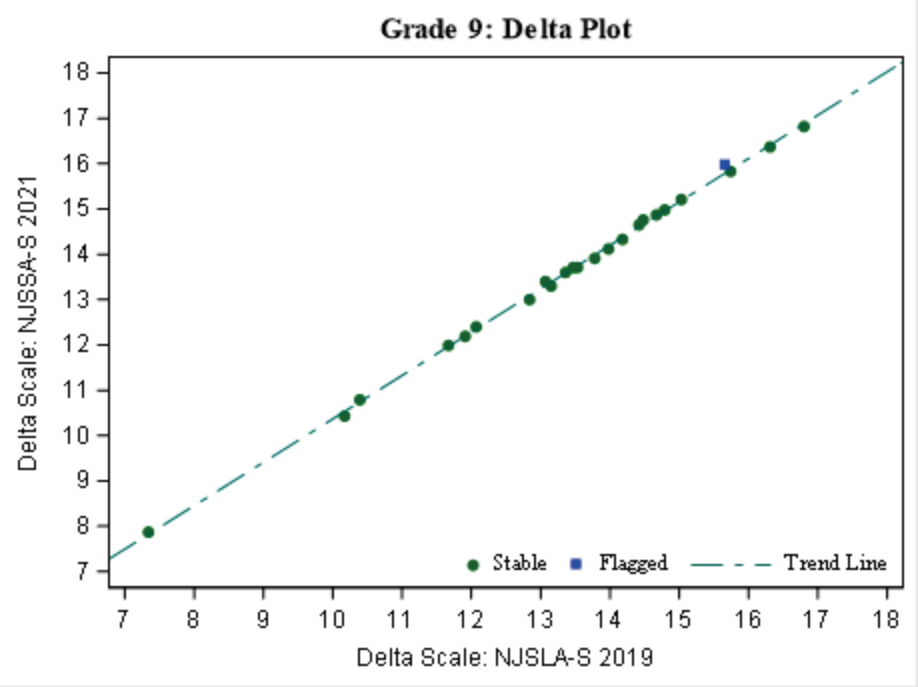

Figure 4.1.2. Grade 9 Delta Plot

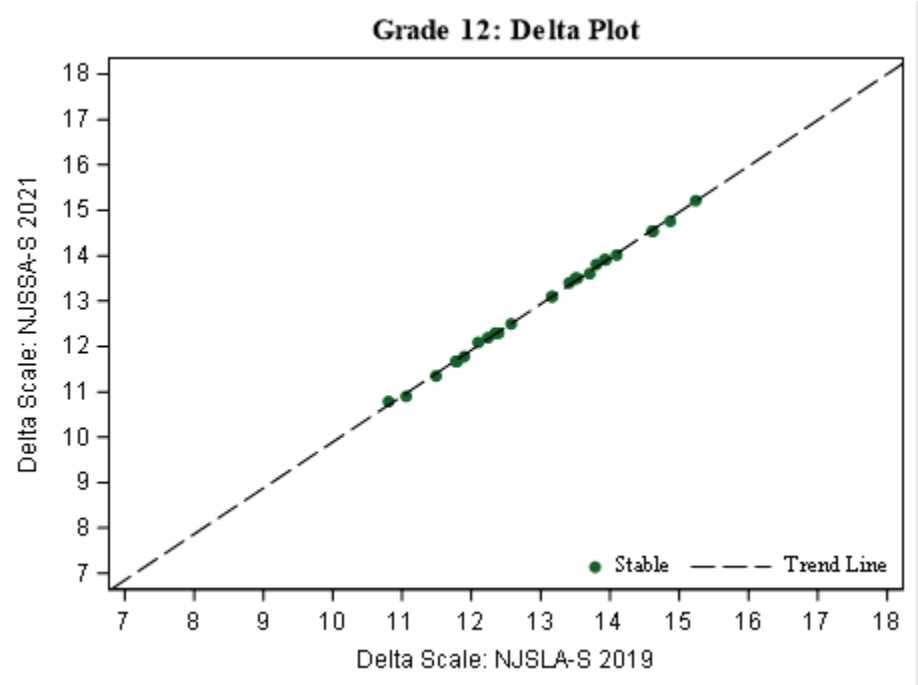

Figure 4.1.3. Grade 12 Delta Plot

4.2 0.3 Logits Absolute Difference Method

The 0.3 Logits Absolute Difference method was used to assess the stability of Rasch-based item difficulty parameters. The first step in the process was to recalibrate the NJSSA–S item difficulty parameters via Winsteps 3.74 (Linacre, 2012) using the 2021 test results. This calibration will hereafter be referenced as the unconstrained calibration, indicating that during the calibration process, the item difficulty parameters were allowed to freely converge regardless of previous results.

The second step was to use the 0.3 Logits Absolute Difference method to compare the results of the unconstrained calibration to the item difficulty parameters from the 2019 NJSLA–S. The latter serves as the basis for the NJSSA–S scale. To do so, an equating constant was calculated that represents the mean difference between the unconstrained item difficulty parameters and their 2019 NJSLA–S counterparts. Each unconstrained item difficulty parameter is then adjusted by adding to it the equating constant. If the absolute difference between an item's adjusted, unconstrained item difficulty parameter and its 2019 NJSLA–S counterpart is greater than 0.3 logits, then the item with the greatest absolute difference is flagged, and the process iterates until all items have adjusted, unconstrained item parameters within 0.3 logits of the constrained item difficulty parameters. Large percentages of items flagged would indicate that the item difficulty parameters were not stable from the 2019 NJSLA–S to the 2021 NJSSA–S administration.

The results of the 0.3 Logits Absolute Difference method were generally positive at all grade levels. Table 4.2.1 provides a summary of the results at each grade including the percentage of flagged items. Figures 4.2.1 to 4.2.3 are scatter plots with the constrained item difficulty parameters on the x-axis and the unconstrained item difficulty parameters on the y-axis; items are demarcated as either stable or flagged, depending on the results of the 0.3 Logits Absolute Difference procedure. At Grades 6 and 12, only two of 25 (8.0%) items were flagged. One of the Grade 6 items and both of the Grade 12 items were also flagged by the delta plot method. Grade 9 had three of 24 (12.5%) items flagged. However, all of the flags were only slightly above the 0.3 Logits Absolute Difference threshold. Moreover, as was described in the previous section, none of the items were flagged by the delta plot method. For all grades, almost all flagged items were due to the items being more difficult for the students than the model would have predicted. Appendix D contains the detailed results of the 0.3 Logits Absolute Difference for each grade level.

Table 4.2.1: Summary 0.3 Logits Absolute Difference Method

Grade	Total Items	Flagged Items	Percent Flagged	Average Constrained Item Difficulty	Average Unconstrained Item Difficulty	Equating Constant
6	25	2	8.0%	–.286	.000	–.286
9	24	3	12.5%	–.198	.000	–.198
12	25	2	8.0%	–.282	.000	–.282

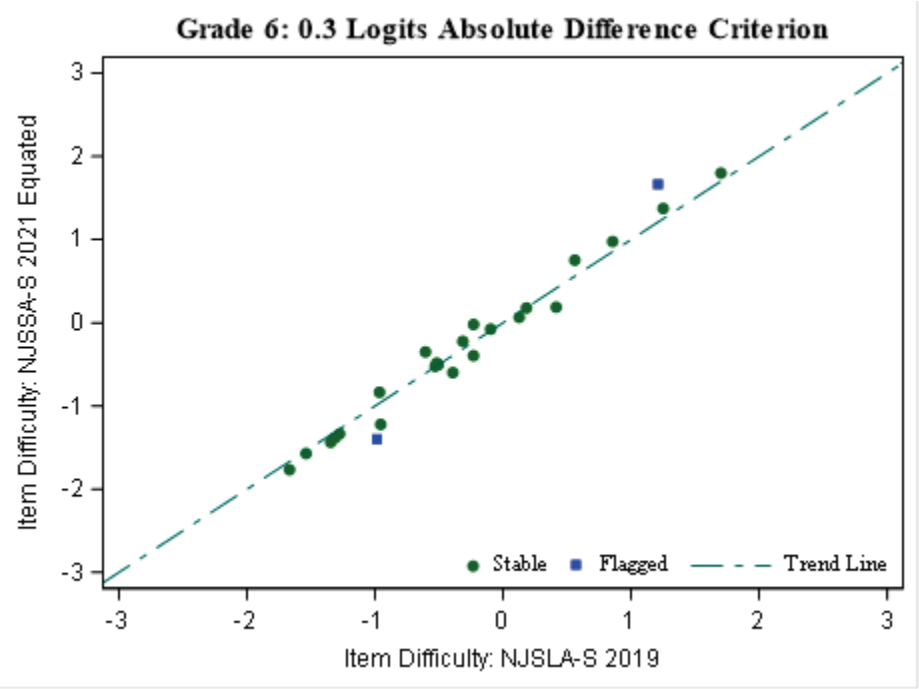

Figure 4.2.1. Grade 6 0.3 Logits Absolute Difference Criterion Item Difficulty Plot

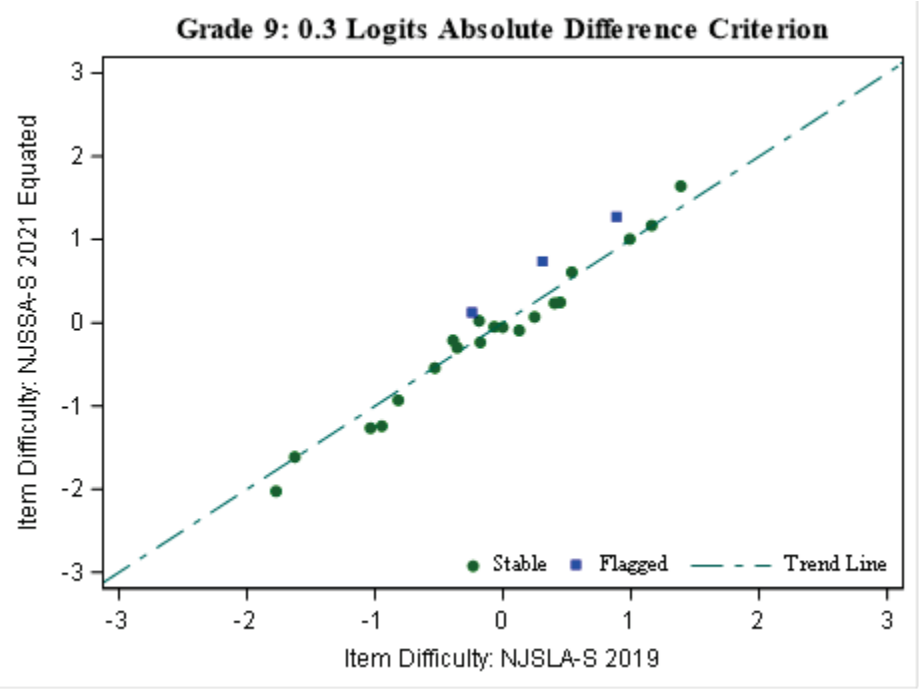

Figure 4.2.2. Grade 9 0.3 Logits Absolute Difference Criterion Item Difficulty Plot

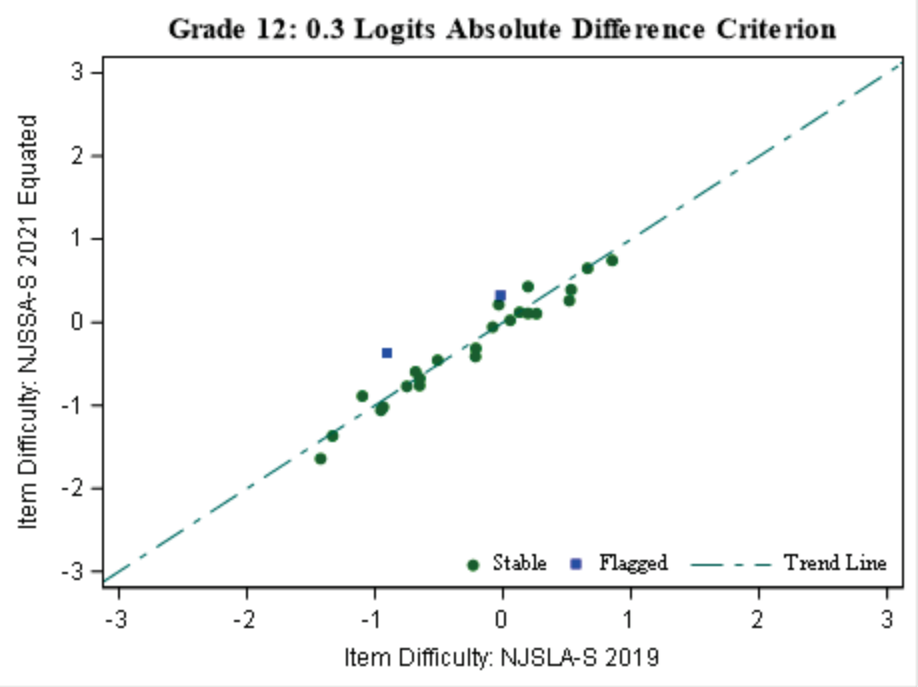

Figure 4.2.3. Grade 12 0.3 Logits Absolute Difference Criterion Item Difficulty Plot

Part 5: 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 NJSSA–S test scores and subscores.

5.1 Classical Test Theory Reliability Estimates

This section describes the Classical Test Theory (CTT) reliability estimates calculated for the 2021 NJSSA–S. Part 5.1.1 describes the concept of reliability in the CTT framework, and Part 5.1.2 displays the results.

5.1.1 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:

$$X = T + \text{error} \qquad \textit{Equation 5.1}$$

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 several 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

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2, \qquad \textit{Equation 5.2}$$

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 (x_1, x_2) on parallel forms, which is equal to true score variance (σ^2_t) divided by observed score variance (σ^2_x):

$$\rho_{x_1 x_2} = \sigma_t^2 \big/ \sigma_x^2 . \qquad \textit{Equation 5.3}$$

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

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{x_1 x_2}}. \qquad \textit{Equation 5.4}$$

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 NJSSA–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:

$$\alpha_{\text{Cronbach}} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_{Y_i}^2}{\sigma_X^2}\right], \qquad \textit{Equation 5.5}$$

where n is the number of items, σ^2_{Yi} is the variance of item i, and σ^2_X is the variance of observed total score, X. SEMs were calculated using the following formula:

$$SEM = S_X \sqrt{1 - \alpha_{\text{Cronbach}}}, \qquad \textit{Equation 5.6}$$

where S_X is the standard deviation of observed total scores.

5.1.2 Raw Score Internal Consistency

In order to accommodate the state's diverse testing population, the NJSSA–S was delivered in multiple formats. Most students received the online computer-based test (CBT), the Spanish (SP) test, and the paper-based test (PBT). It should be noted that reliability measures that are based on internal consistency, such as coefficient alpha, decrease when the students taking a given test form are more homogeneous in their test performance.

Table 5.1.1 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 6 stand out as particularly exceptional given that there are only 25 points on the NJSSA–S. The Grade 6 reliability coefficients ranged from .74 for the Spanish form to .88 for the CBT form. The most likely reason for the better results at Grade 6, the same test length as the other grades, is that the Grade 6 items were closely matched to the ability levels of the Grade 6 students, thereby increasing the variance among test scores. At Grade 9, where the distribution of test scores was heavily skewed toward the low end of the ability spectrum, reliability ranged from .54 on the Spanish form to .81 on the CBT form. The relatively low-reliability measures for the Spanish were due to that population doing very poorly on the test, which limited the amounts of variance in the Spanish form test scores. Similarly, the Grade 12 alpha coefficients ranged from .54 on the Spanish form to .83 on the CBT form.

Table 5.1.1: Coefficient Alpha and SEM, by Form

Grade	Form*	N-Count	Mean	SD	Alpha	SEM
6	CBT	94,070	13.42	5.90	.88	2.08
6	PBT	174	8.44	4.67	.80	2.09
6	SP	1,068	6.70	3.87	.74	1.98
9	CBT	99,547	10.22	4.82	.81	2.10
9	PBT	40	10.03	4.90	.81	2.11
9	SP	1,062	5.39	2.81	.54	1.90
12	CBT	88,469	12.80	5.32	.83	2.23
12	PBT	60	10.22	4.70	.78	2.20
12	SP	561	7.71	3.22	.54	2.18

* CBT: Computer-Based Test; PBT: Paper-Based Test; SP: Spanish

Table 5.1.2 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 on 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 highest subscore reliability of .79 was for Physical Science at Grade 6; whereas the lowest subscore reliability of .46 was for Critiquing at Grade 9. The reliability measures in Table 5.1.2 are based on all test takers at each grade level.

Table 5.1.2: Coefficient Alpha and SEM by Reporting Category

Grade	Reporting Category	Total Items	MC Items	TE Items	TE2 Items	Max Points	Alpha	SEM
6	Total	25	7	18	0	25	.88	2.08
6	Earth and Space	8	4	4	0	8	.64	1.21
6	Life	7	1	6	0	7	.67	1.12
6	Physical	10	2	8	0	10	.79	1.24
6	Investigating	9	2	7	0	9	.74	1.28
6	Sensemaking	10	4	6	0	10	.71	1.30
6	Critiquing	6	1	5	0	6	.65	0.98
9	Total	24	8	15	1	25	.81	2.10
9	Earth and Space	8	4	4	0	8	.59	1.22
9	Life	7	1	6	0	7	.62	1.06
9	Physical	9	3	5	1	10	.57	1.34
9	Investigating	9	6	3	0	9	.58	1.27
9	Sensemaking	10	1	9	0	10	.68	1.32
9	Critiquing	5	1	3	1	6	.46	1.03
12	Total	25	10	15	0	25	.83	2.23
12	Earth and Space	10	5	5	0	10	.72	1.36
12	Life	7	1	6	0	7	.60	1.16
12	Physical	8	4	4	0	8	.50	1.31
12	Investigating	8	4	4	0	8	.60	1.27
12	Sensemaking	11	5	6	0	11	.67	1.49
12	Critiquing	6	1	5	0	6	.54	1.06

Table 5.1.3 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 6, the lowest value was .80, for English learner (EL) students, which is still very strong, given that the NJSSA–S only consisted of 25 points worth of items. At Grade 9 the coefficient alphas were close to the .70 to .80 range, excepting the EL students (α_{EL-Yes} = .55). The same pattern was evident at Grade 12, where all the coefficient alphas hovered close to the .70 to .80 range, except for EL students (α_{EL-Yes} = .60)

Table 5.1.3: Coefficient Alpha and SEM by Demographic Group

Grade	Group	N	Mean	SD	Alpha	SEM
6	NJSSA–S	95,332	13.33	5.93	.88	2.08
6	Male	48,628	13.34	6.11	.89	2.07
6	Female	46,688	13.32	5.73	.87	2.08
6	Non-Binary	16	15.06	6.64	.91	1.99
6	Am. Indian	173	12.94	5.91	.88	2.07
6	Asian	10,359	17.27	4.99	.84	1.97
6	Black	13,904	10.49	5.56	.86	2.09
6	Hispanic	29,691	11.09	5.60	.86	2.10
6	Pacific Islander	186	14.22	5.90	.88	2.04
6	White	38,314	14.94	5.36	.85	2.07
6	EL - Yes	6,634	7.39	4.49	.80	2.01
6	EL - No	88,698	13.77	5.78	.87	2.08
6	EconDis - Yes	29,813	10.74	5.52	.86	2.10
6	EconDis - No	65,516	14.51	5.73	.87	2.06
6	SWD - Yes	19,310	9.88	5.81	.87	2.05
6	SWD - No	76,022	14.21	5.63	.86	2.08
9	NJSSA–S	100,693	10.16	4.83	.81	2.10
9	Male	51,305	10.36	5.07	.83	2.10
9	Female	49,317	9.95	4.55	.79	2.10
9	Non-Binary	71	12.90	4.58	.78	2.17
9	Am. Indian	155	9.86	4.86	.82	2.08
9	Asian	10,404	13.62	4.84	.81	2.11
9	Black	14,488	8.05	3.99	.73	2.06
9	Hispanic	31,932	8.33	4.17	.75	2.06
9	Pacific Islander	207	11.21	4.76	.81	2.10
9	White	41,148	11.39	4.66	.79	2.13
9	EL–Yes	4,933	5.48	2.85	.55	1.92
9	EL–No	95,760	10.41	4.78	.81	2.11
9	EconDis–Yes	28,926	8.10	4.02	.74	2.06
9	EconDis–No	71,765	10.99	4.87	.81	2.11
9	SWD–Yes	18,953	7.89	4.32	.78	2.04
9	SWD–No	81,740	10.69	4.78	.81	2.11
12	NJSSA–S	89,197	12.76	5.33	.83	2.23
12	Male	44,952	12.77	5.57	.84	2.20
12	Female	44,205	12.75	5.07	.80	2.25
12	Non-Binary	40	15.55	4.25	.72	2.25
12	Am. Indian	112	12.15	5.33	.83	2.21
12	Asian	10,097	16.25	5.17	.84	2.09
12	Black	11,404	10.33	4.49	.75	2.26
12	Hispanic	24,823	10.71	4.69	.77	2.25
12	Pacific Islander	209	13.35	5.05	.80	2.24

Grade	Group	N	Mean	SD	Alpha	SEM
12	White	41,038	13.77	5.16	.81	2.22
12	EL–Yes	3,782	7.40	3.38	.60	2.13
12	EL–No	85,415	12.99	5.27	.82	2.23
12	EconDis–Yes	23,277	10.63	4.64	.76	2.26
12	EconDis–No	65,918	13.51	5.36	.83	2.21
12	SWD–Yes	16,927	10.34	5.03	.81	2.22
12	SWD–No	72,270	13.32	5.24	.82	2.23

Table 5.1.4 displays coefficient alpha and SEM by the two main item types: multiple-choice (MC) and technology-enhanced (TE). Those item types are more thoroughly described in Part 2 of this technical brief. 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 from .73 to .84, and the MC items displayed alphas from .56 to .65.

Table 5.1.4: Coefficient Alpha and SEM by Item Type

Grade	Item Type	Items	Points	Mean	SD	Alpha	SEM
5	MC	7	7	4.62	1.85	.65	1.09
5	TE	18	18	8.71	4.44	.84	1.76
8	MC	8	8	3.78	2.02	.63	1.23
8	TE	16	17	6.38	3.24	.73	1.70
11	MC	10	10	5.37	2.18	.56	1.45
11	TE	15	15	7.39	3.62	.78	1.69

5.2 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) and item maps were evaluated at each grade level. Overall, the 2021 NJSSA–S was reliable from the perspective of IRT and the PCM.

5.2.1 Test Information Functions

In IRT, the reliability of an assessment is conceptualized via the test information function (TIF) (Hambleton & Swaminathan, 1985). Unlike coefficient alpha (Cronbach, 1951) the TIF is not uniform across the entire range of test scores. Instead, the TIF can assess test reliability across the full range of scores. This is particularly important to a criterion-referenced test such as the NJSSA–S because it allows for the reliability of the assessment to be evaluated specifically at the most important decision points (i.e., the “Some Support” and “Less Support” 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:

$$I_{ij}\left(\theta_i, \delta_j\right) = P_{ij} * \left(1 - P_{ij}\right) \qquad \textit{Equation 5.7}$$

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 5.2.1 to 5.2.3 illustrate, respectively, the TIFs for Grades 6, 9, and 12 at person ability estimates ranging from –5 to + 5. More information at a specific ability level implies less measurement error. Ideally, the peak of the information function would maximize information at both the "Some Support" and "Less Support" cut scores in order to minimize measurement error where the most important decisions are taking place. Within each figure, there are two vertical dashed lines representing the cut scores.

At Grade 6 the TIF peaked close to the "Some Support" cut score. There was a large drop in information at the "Less Support" cut. At Grade 9 the TIF peaked almost directly in between the "Some Support" and "Less Support" cut scores. Overall, the Grade 9 TIF is very close to being ideal. The Grade 12 TIF was similar to its Grade 6 counterpart. It peaked almost directly at the "Some Support" cut score, and there was a relatively large decrease in information at the "Less Support" cut score. Overall, the TIFs provide ample evidence that student ability estimates are reliable at the most important decision points. However, both Grades 6 and 12 could be improved to gather more information at the "Less Support" cut.

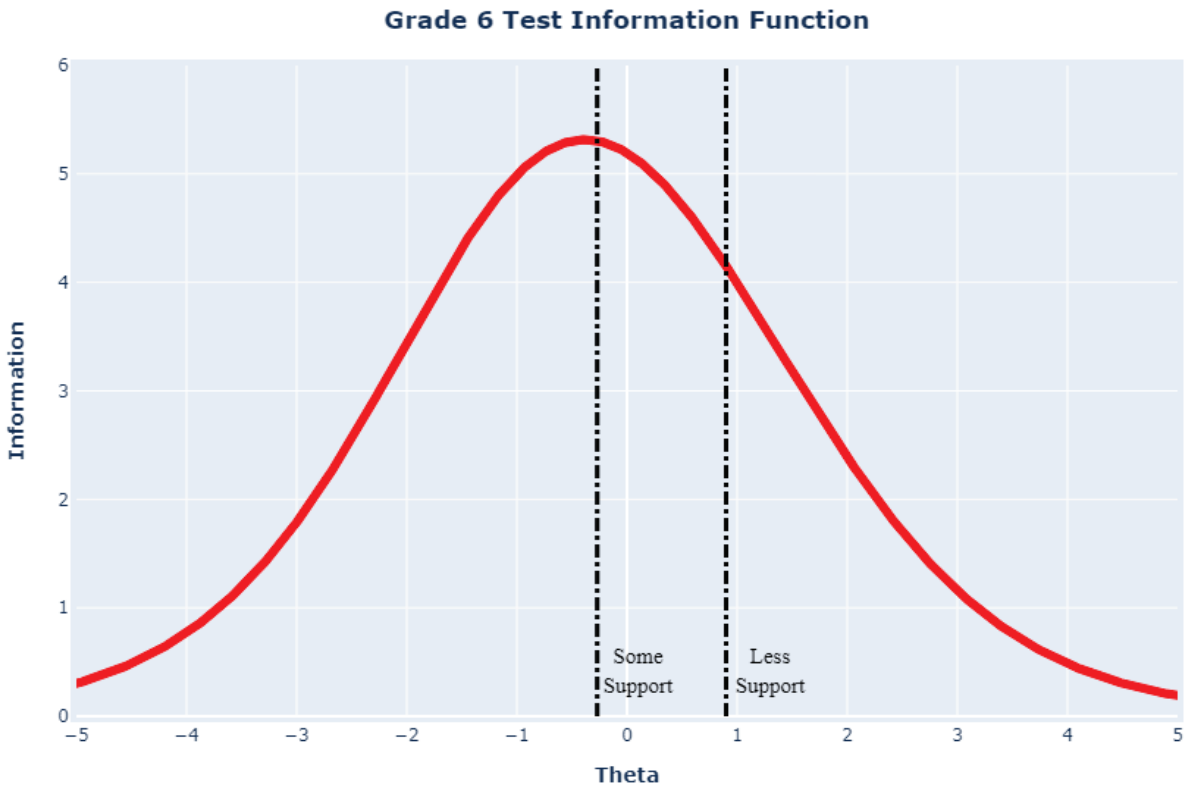

Figure 5.2.1. Grade 6 Test Information Function

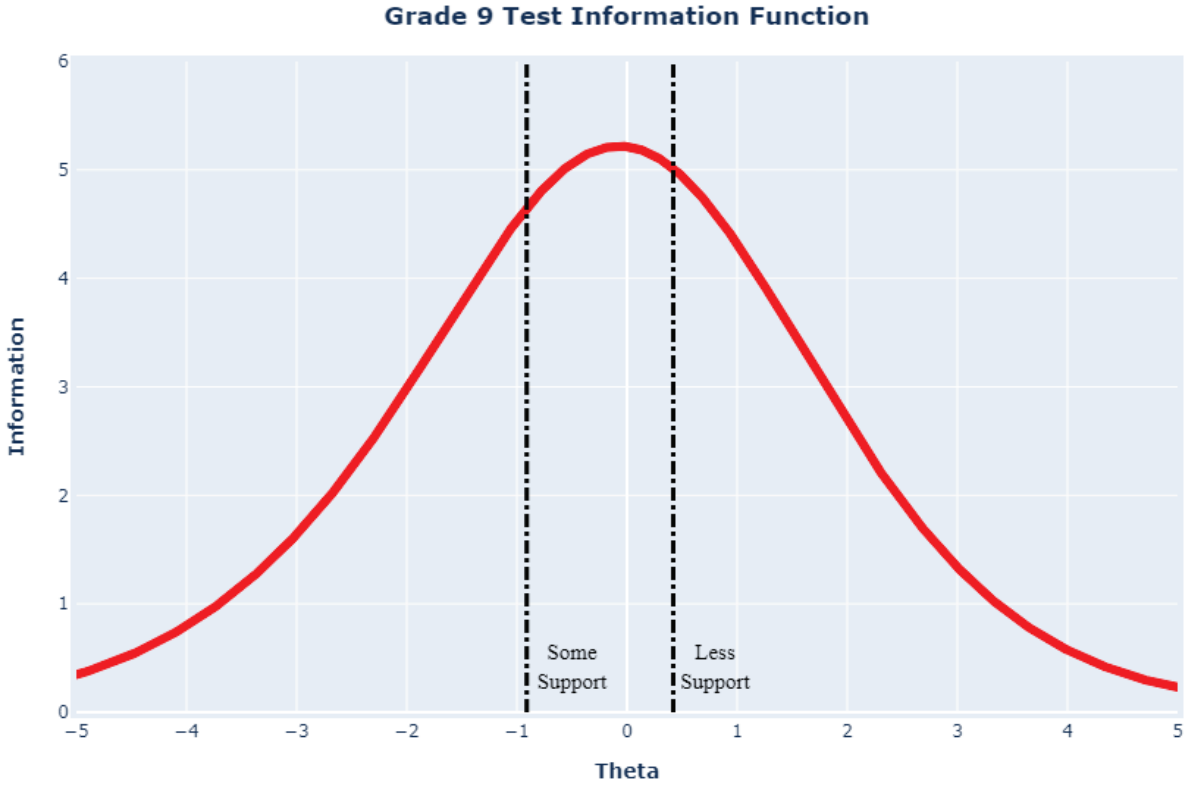

Figure 5.2.2. Grade 9 Test Information Function

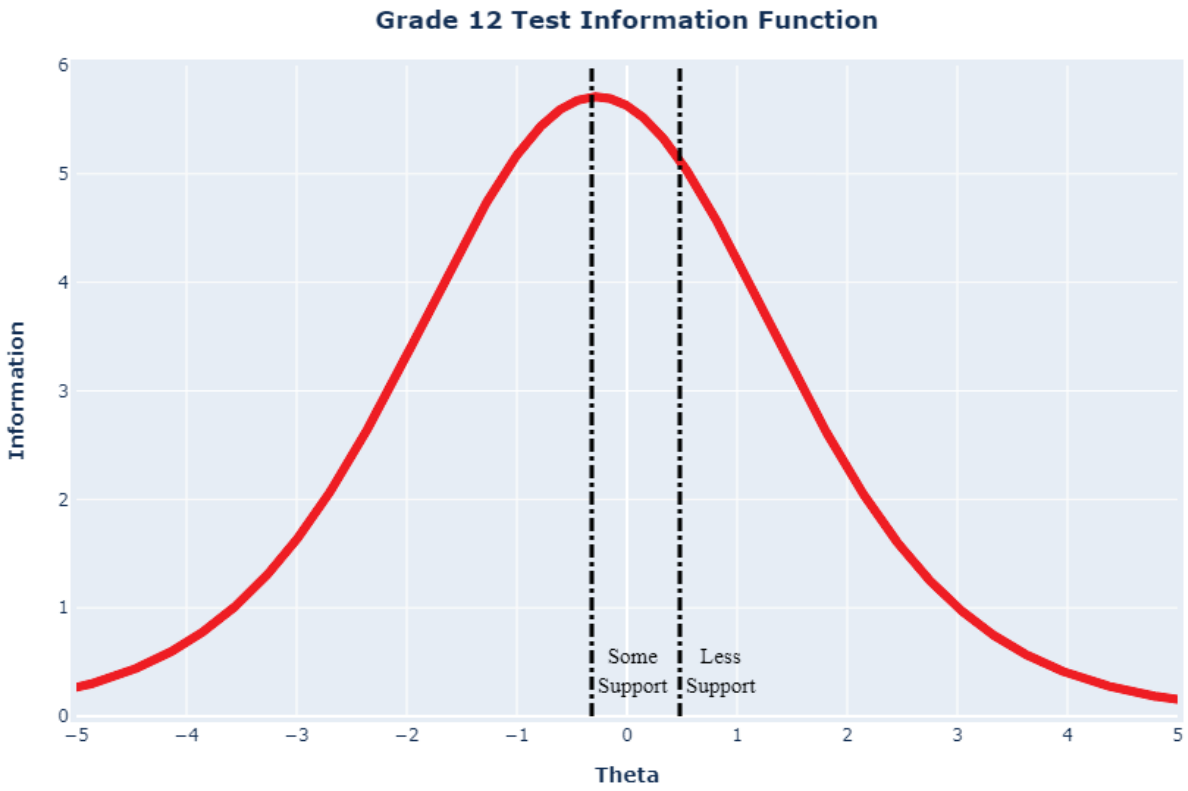

Figure 5.2.3. Grade 12 Test Information Function

5.2.2 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 5.2.4 through 5.2.6 show the 2021 NJSSA–S item maps. Unsurprisingly, given the TIFs, the Grades 6 and 12 item difficulty distributions peak slightly below the "Some Support" cut score, while the Grade 9 item difficulty distribution peaks directly in between the "Some Support" and "Less Support" cut scores. At Grade 6, the theta distribution was normally distributed with student ability peaking close to the "Some Support" cut score. The item distribution peaked below the "Some Support" cut score. The theta distribution at Grade 9 peaked in between the "Some Support" and "Less Support" cut scores; however, it was more skewed towards the lower end of the ability spectrum, with large peaks of students below the "Some Support" cut. The Grade 12 theta distribution peaked right at the "Some Support" cut score. The Grade 6 item distribution contained more items below the ability levels of the students than would be ideal. The Grade 9 item distribution, as the Grade 9 TIF showed, matched the decision points on the scale very well. However, there were many students below the "Some Support" cut score and very few items along that part of the scale. The extremely tight Grade 12 item distribution was lacking items at both the upper and lower parts of the scales in comparison to the ability levels of the students.

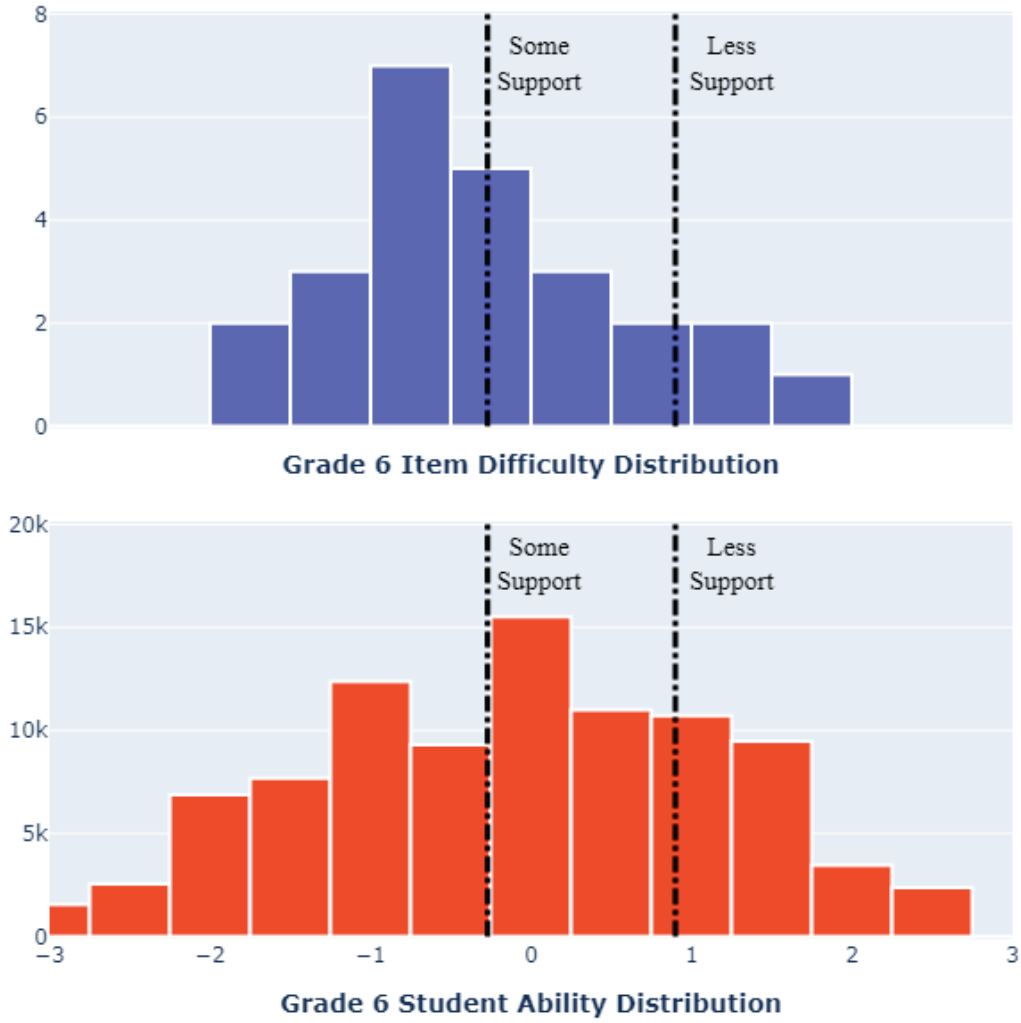

Figure 5.2.4. Grade 6 Item Difficulty and Student Ability Distributions

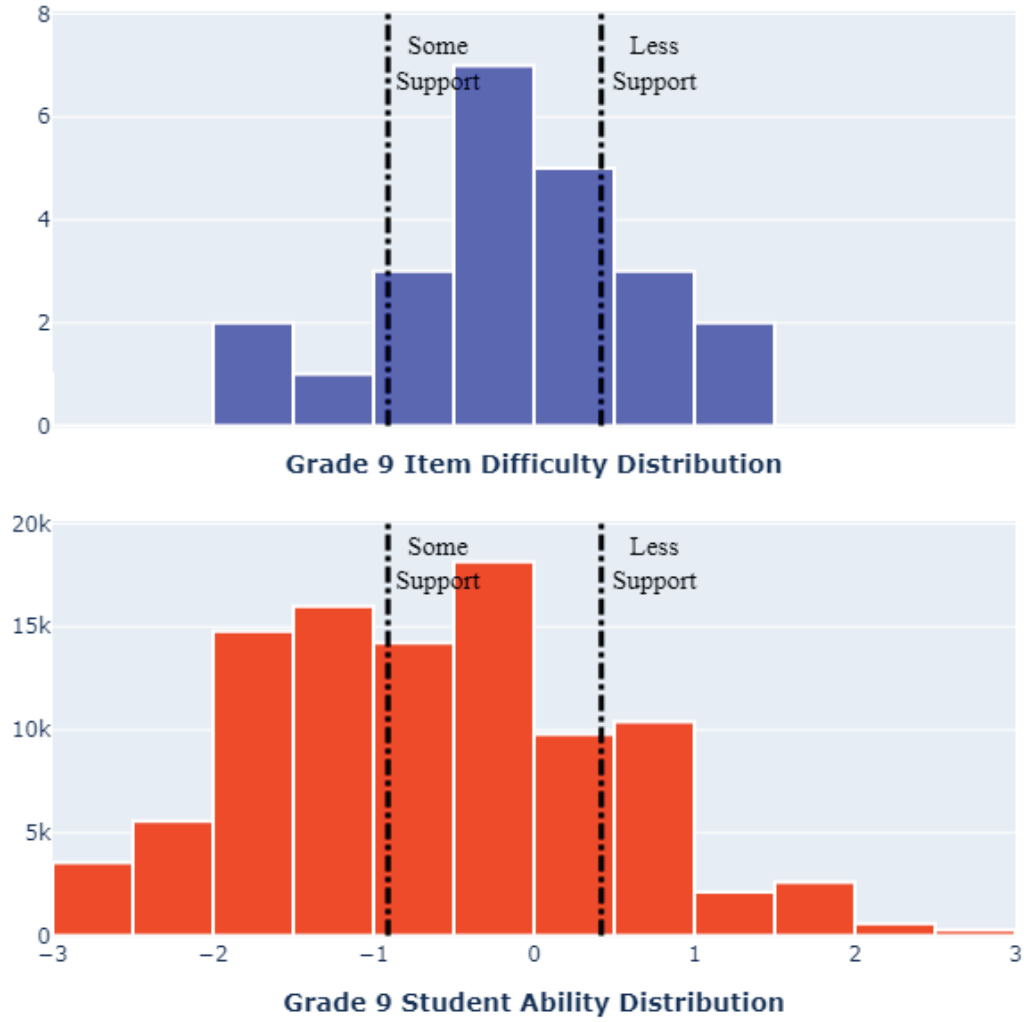

Figure 5.2.5. Grade 9 Item Difficulty and Student Ability Distributions

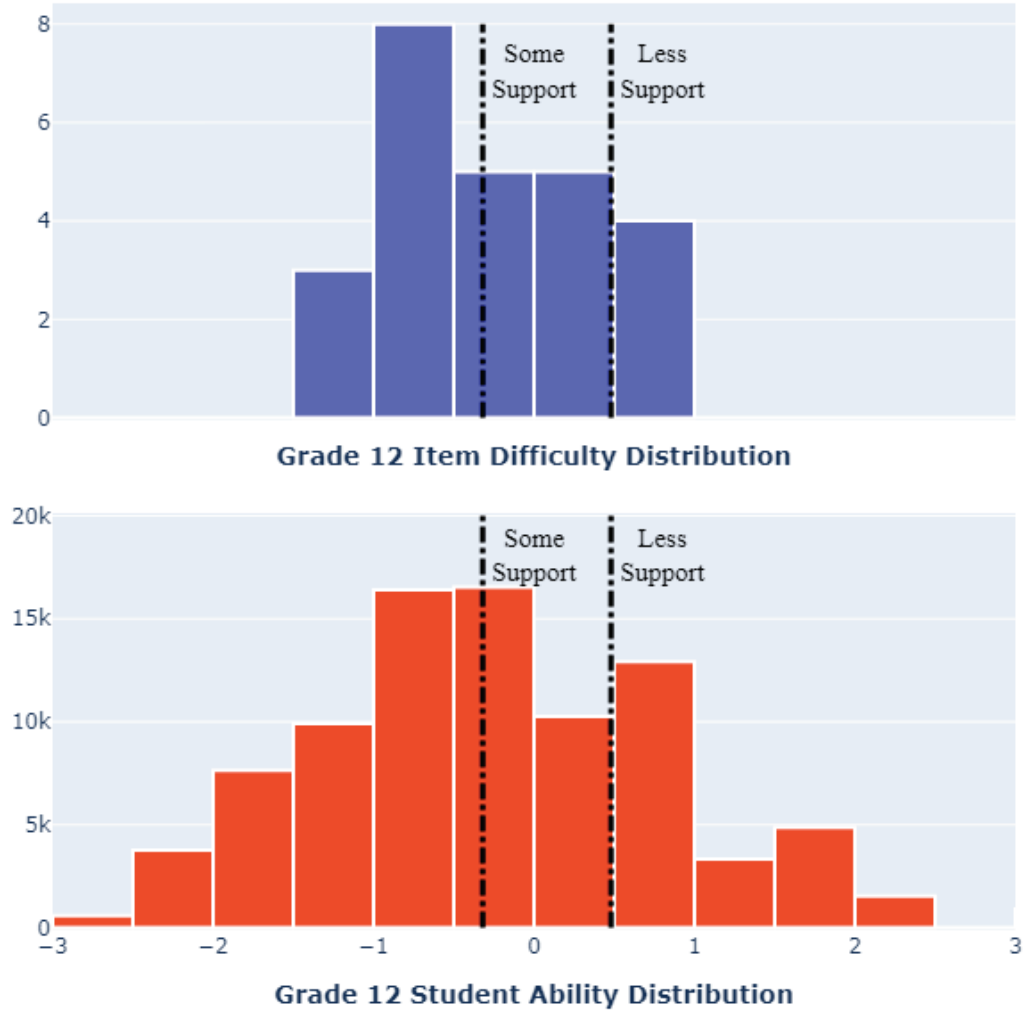

Figure 5.2.6. Grade 12 Item Difficulty and Student Ability Distributions

5.3 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 2021 NJSSA–S performance level classifications were reliable.

5.3.1 Conditional Standard Error of Measurement at Each Cut-Score

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

$$SE\left(\hat{\theta}\right) = \frac{1}{\sqrt{I\left(\theta\right)}} \qquad \textit{Equation 5.8}$$

where $I(\theta)$ is the information function for a test at a score point (theta).

The 2021 NJSSA–S Raw cut scores and the corresponding conditional standard error of measurement (CSEM) on the theta scale are summarized in Table 5.3.1. The theta scores and corresponding CSEMs for all raw scores are presented in Appendix E. The lower and upper bound values in Table 5.3.1 have been placed on the raw score scale. Given that the CSEMs are the inverse of the TIF, their interpretations are similar. The TCCs are graphical representations of the expected raw scores points a student would achieve at a given level of theta. The upper and lower bounds were defined by multiplying the theta cut score's CSEM by two and either adding it to or subtracting it from the cut score's theta value. Next, the upper and lower bound theta values were identified on the TCC curve to find their corresponding raw score point. Any overlap between the upper or lower bounds and one of the other cut scores could indicate reliability problems among the support level classifications. Appendix F contains both CSEM and TCC graphs for each grade level.

At Grade 6 and Grade 9, there was no overlap between the upper bound of the "Some Support" and the "Less Support" cut score. Nor was there overlap between the "Less Support" lower bound and the "Some Support" cut score. However, at Grade 12 both cut scores showed overlap between the relevant upper or lower bound. The upper bound of the "Some Support" cut score was 17.5, which overlapped with the "Less Support" cut score of 17. Similarly, the "Less Support" lower bound of 12.1 overlapped with the "Some Support" cut score of 13. The overlap at Grade 12 is due to the close proximity of the "Some Support" and "Less Support" cut scores on the theta scale and the relatively small number of items on the Start Strong in comparison to the NJSLA–S. If the Start Strong is continued with the same scale and cut scores, then a longer Grade 12 test might be appropriate to address the overlap among the cut scores and the upper and lower bounds.

Table 5.3.1: Raw Cut Scores with Conditional Standard Error of Measurement

Grade	Level	Raw Cut score	CSEM*	Lower Bound	Upper Bound
6	Some Support	13	.43	8.6	17.3
6	Less Support	19	.51	14.3	22.2
9	Some Support	9	.46	5.2	13.6
9	Less Support	16	.45	11.4	19.9
12	Some Support	13	.42	8.4	17.5
12	Less Support	17	.45	12.1	20.8

*CSEM placed on the theta scale

5.3.2 Classification Consistency Indices

The reliability index for proficiency classifications (kappa) is an estimate of how reliably the test classifies students into the support level categorizations (i.e., Strong Support, Some Support, Less Support). Kappa was computed with the BB-CLASS program (Brennan, 2004) based on the beta-binomial model. Coefficient kappa (*K)* is given by:

$$\kappa = \frac{\varphi - \varphi_c}{1 - \varphi_c}, \qquad \textit{Equation 5.9}$$

where φ is the probability of a consistent classification and φ_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 5.3.2 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 .68 to .74. Thus, if the NJSSA–S had been administered a second time, approximately 68–74% of the students would have been classified at exactly the same performance level. The classification consistency is similar to the much longer NJSLA–S (NJDOE, 2019). Overall, the NJSSA–S support level classifications should be interpreted as being consistent.

Table 5.3.2: Support Level Classification Reliability

Grade	Alpha	SEM	Some Support Cut	Less Support Cut	Kappa	Classification Consistency
6	.88	2.08	13	19	.59	.74
9	.81	2.10	9	16	.52	.68
12	.83	2.23	13	17	.53	.70

Part 6: 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, 2014, p. 14). If there is ample evidence to support reasonable interpretations and test uses, then they are considered to possess high validity (Kane, 2013).

The NJSSA-S' primary purpose is to provide instructional information for classroom teachers and school and district educators about student needs for additional support upon returning to school after the COVID pandemic hiatus. The NJSSA-S produces the resources used locally to evaluate the needs of students. The assessment provides an initial indication of conceptual or skill gaps that might exist in a student's understanding of NJSLS-S and the level of support students may need to inform instruction. The information provided by the assessment is only one piece of the puzzle used to holistically understand a student's academic performance. The data should be used with other supporting evidence (e.g., assessments, homework, etc.) in guiding instruction. It is important to note that the NJSSA-S does not assess all the learning standards on the summative assessment. The NJSSA-S is not a replacement for the NJSLA-S. Nonetheless, the NJSSA-S provides applicable information. This information can be evaluated in terms of its validity evidence.

Conceptually, Kane (2006) labeled the process of evaluating that evidence as validation. As such, test validation is an ongoing, ever-evolving process that extends through the duration of an assessment program. Every component within this technical brief, 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. Next, a section describes other validity evidence that was collected. Finally, the validity evidence pertaining to the intended interpretations is summarized. Overall, the evidence suggests that the 2021 NJSSA–S fosters valid interpretations and uses of test scores as they pertain to the overall classifications of students into support levels.

6.1 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, 2014, 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 NJSSA–S is critical because the tests purport to provide an indication of the support students need to progress toward achieving the knowledge, skills, and abilities (KSAs) identified in the NJSLS–S.

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 was taken into consideration in the writing of all NJSSA–S items. In accordance with the test blueprint, the test construction process attempts to balance the six reporting categories and to ensure that the NJSSA–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 and engineering practices (i.e., Investigating, Sensemaking, and Critiquing) were less balanced. At each grade level, the Sensemaking scientific practice was over-represented and Critiquing was under-represented.

On a more granular level the length of the test and the need for the items to be grouped as clusters made it impossible to effectively sample all the DCIs, SEPs, and CCCs. At Grade 6, only six out of eleven DCIs were tested, including only one of three Earth and Space Science DCIs. Grade 9 had all eleven DCIs multiple DCIs represented on the test; however, three DCIs were only aligned to one item. At Grade 12, three out of the eleven DCIs were not tested. The SEPs at both Grades 6 and 9 did not include OECI items. Moreover, some CCCs were not represented at a given grade level.

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 and engineering 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 levels.

6.2 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, 2014, p. 26). Evidence based on response processes is complementary to evidence based on test content; it can come from several sources including response times, eye-tracking, think-aloud protocols, interviews, and/or focus groups. This complementary evidence is different from content evidence because its source is not content experts or teachers, but rather the actual student test takers. Padilla and Benitez (2014) noted that "validation studies aimed at obtaining evidence from response processes are scant" (p. 139), and at present time the NJSSA–S evidence based on response processes is limited to judgments from the NJSAC and content specialists.

The alignment of each item to the NJSLA–S Range PLDs provides limited evidence of the cognitive processes theoretically being assessed by the NJSSA–S. The Support Level Descriptors were based on the NJSSA–S Range PLDs, which were created in a collaborative effort by NJDOE, the NJSAC, content specialists, and psychometricians; they are based upon the NJSLS–S content standards. A more detailed description of the NJSLA–S PLD development process can be found in the 2019 NJSLA–S Technical Report. The detailed test maps presented in Appendix A display the NJSLA–S Range PLD alignment for each NJSSA–S item.

The NJSSA–S program does not currently have its own Range PLDs. It instead uses NJSLA–S Range PLDs as the theoretical cognitive structure underlying all current NJSSA–S item and test development. The NJSLA–S Range PLDs contain detailed descriptions of the KSAs that a student needs to display in order to be classified at a given support level. Each item on the NJSSA–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 NJSSA–S's underlying cognitive structure to the student responses, which provides limited validity evidence based on response processes.

Table 6.2.1 shows the distributions of the support levels associated with each item by grade level and by DCI and SEP. The DCI distribution of items at Grade 6 had more items aligned to "Strong Support" than were necessary. Whereas the Grade 9 DCI distribution had too many items aligned to "Less Support," and too few items aligned to "Strong Support." Both grades had SEP distributions that were close to ideal. Grade 12's SEP alignment had too many items aligned to "Strong Support," but its DCI alignment distribution was close to ideal. These support level alignment distributions largely correspond to the item difficulty distributions illustrated in Figures 5.2.4 through 5.2.6.

Table 6.2.1: Support Level Alignment by DCI, SEP, and Grade Level

Grade	Domain/Practice	Strong Support	Some Support	Less Support
6	DCI	10	12	3
6	SEP	7	11	7
9	DCI	2	12	10

Grade	Domain/Practice	Strong Support	Some Support	Less Support
9	SEP	5	14	5
12	DCI	5	13	7
12	SEP	10	10	5

6.3 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, 2014, p. 16). The NJSSA–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 NJSSA–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.

6.3.1 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 3.2.1.1 of this document summarizes correlation coefficients among test content domains and clusters by grade level. The intercorrelations of the NJSSA–S provide clear evidence that the NJSSA–S is unidimensional. The lowest correlation among all subscores at all grade levels was .52 at Grade 12 between the Life and Physical content domain 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 correlated higher with certain practices. Ideally, at each grade level, the correlations among the content domains and scientific practices would be similar. However, at all three grade levels, certain scientific practices correlated higher with one or two content domains. At Grade 6, Investigating displayed a much higher correlation with Physical Science than Earth and Space Science, Sensemaking correlated to a much higher degree with Earth and Space Science than with Life Science, and Critiquing correlated higher with Life Science than Earth and Space Science. Similar patterns existed at both Grades 9 and 12. These types of possible dependencies are evidence that the internal structure of the NJSSA–S subscores is not performing as expected.

Table 6.3.1 shows a matrix of both the points available by content domain and scientific practice, as well as their subscore intercorrelations. It presents a likely explanation for the

unexpected correlational pattern. As an example, at Grade 6, the correlations for a given scientific practice and content domain were higher when more points were available. For Sensemaking at Grade 6, there were more Earth and Space Science points available than Life and Physical combined. Unsurprisingly, that dependency led to a much higher correlation between Sensemaking and Earth and Space Science than the other two content domains. A similar pattern existed at all the other combinations of grades, scientific practices, and content domains. When the domains and practices are too intertwined, that will lead to those subscores being overly dependent on each other. Test score interpretations of subscores that are so highly dependent upon each other could be misleading. For example, if a Grade 6 student has a low raw score in Earth and Space Science, it would be impossible to know whether that was due to a lack of Earth and Space Science skills or due to a lack of Sensemaking skills.

Table 6.3.1: Points Available and Intercorrelations by Domain and Practice

Grade	Practice	Earth	Life	Physical	Earth	Life	Physical
6	Investigating	1	3	4	.72	.81	.89
6	Sensemaking	6	1	4	.90	.71	.81
6	Critiquing	1	3	2	.68	.81	.76
9	Investigating	4	1	4	.77	.64	.75
9	Sensemaking	2	5	3	.71	.85	.73
9	Critiquing	2	1	3	.69	.60	.70
12	Investigating	3	0	5	.76	.59	.82
12	Sensemaking	5	4	2	.83	.79	.67
12	Critiquing	2	3	1	.71	.76	.59

6.3.2 Other Internal Structure Evidence

Evidence of the internal structure of the NJSSA–S was also presented via a principal component analysis (PCA). Its results are presented in Part 3.2.1.2. These scree plots show further evidence that the variability in the NJSSA–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 5 of this technical brief provides ample evidence to support NJSSA–S reliability. Reliability is a measure of internal consistency that provides a sign as to whether the internal structure of the NJSSA–S is unidimensional. The grade-level reliability coefficients presented in Part 5.1 were relatively strong, ranging from .81 to .88. At the subscore level, the reliability coefficients were adequate given the dearth of points available for many subscores, with only Grade 9 Critiquing falling below .50.

6.4 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. No evidence based on relationships of the NJSSA–S to other variables currently exists.

6.5 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 NJSSA–S validity evidence based on the consequences of testing exists at the moment.

6.6 Other Validity Evidence

Each section within this technical brief contributes evidence relevant to validity. The following is a summary of evidence within each section that is specific to the NJSSA–S:

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 NJSSA–S.

Part 2: Test Development—This section describes the processes used to design and develop the NJSSA–S including:

- The steps taken to link test development to the NJSSA–S' intended inferences and uses.
- The training and QC procedures implemented in the NJSLA–S item development process.
- The use of NJDOE, the NJSAC, and the Sensitivity committee during the initial creation of the items for the NJSLA–S to ensure the work of item writers and content specialists was aligned to the NJSLS–S.
- The steps taken to ensure the test construction process matched the NJSSA–S blueprint and statistical constraints.

Part 3: Item and Test Statistics—This section describes the battery of statistics that were used to evaluate the NJSSA–S at the test, item, and person levels 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 4: Scale Stability—This section describes the methods used to test the stability of the NJSSA–S scale including:

- Statistical and visual inspections of the stability of CTT and IRT item difficulty measures.
- Monitoring the IRT-based equating constant.

Part 5: Reliability—This section describes several reliability statistics that were calculated to verify the consistency of the NJSSA–S test scores including:

- Verifying the reliability at the total score, form, subscore, item type, and subgroup levels.
- Evaluating graphical displays of IRT reliability such as TIFs.
- Assessing the consistency of student support level classifications.

The following is a summary of validity evidence presented within this document that is specific to the entire NJSSA:

Part 2.5: Test Administration—This section describes the care that was taken to implement standardized test administration procedures including:

- Documents produced to communicate NJSSA–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 NJSSA–S test-takers with equal opportunities on the test.

Part 2.9: Scores and Score Reports—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.
- Verifying that student raw scores and subscores were calculated accurately.

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 the Department prior to test administration. All student responses are machine-scored based on these prior approved keys. The data from the student responses are 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 Record File (SRF) 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.

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. The Content Development's Psychometric departments use the 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 five 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.

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, key changes are submitted to the Department, 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.

Reporting

The Start Strong Assessment's primary purpose was to provide instructional information to classroom teachers about students' needs for additional support upon returning to school in the fall of 2021. The information provided by this assessment is a snapshot of a student's understanding and should only be used with other supporting evidence (assignments, homework, etc.) when drawing conclusions about a student's overall academic performance. Examples and further documentation of each report are available on the New Jersey Assessments Resource Center and https://nj.mypearsonsupport.com/startstrong/ websites.

Student-Level Reports

Three student-level reports were produced for the Start Strong Assessment and are available via PearsonAccess^{next} (PAN).

- The **OnDemand Student Report** (ODR) is the first report for the Start Strong Assessments. It shows the student's support level and the scores on each reporting concept. Only students who received a score will receive an OnDemand Student Report.
- The **Student Performance Item Level Report** allows users to compare the support level assigned to individual students within a group, then drill down to an individual student's response to each item. This can be useful for understanding what misconceptions students may have.
- The **Individual Student Reports** (ISRs) are the last type of report to be released for the Start Strong Assessments. Users will be able to download PDFs of ISRs from PAN; school districts will also receive hard copies to distribute to students' parents or guardians. ISRs will be shipped for both testing sites and accountable schools if different schools are involved.

Classroom-, School-, District-Level Reports

In addition to the student-level reports described above, appropriate users will also have access to the Results by Question Reports and Support Level Reports via PearsonAccess^{next}.

- The **Results by Question Report** provides users with group-level information about student performance on specific items or standards. The Results by Question Report has two different ways to view information: the question list and the student list. The default view is the question list. You can switch between the two views by checking the "Show Students" checkbox at the bottom of the list.

  Drilling down to scores on individual test items enables the teacher to corroborate, verify, or otherwise build upon test information to identify instructional needs at the individual student or group level. This aids in the design and delivery of effective educational methods to meet these needs.
    - **Question List –** The question list shows items in numerical order, along with the standard(s) to which each item is aligned, the reporting concept(s) the item is associated with, and the number and percentage of students who answered the item correctly, incorrectly, and partially correctly (for those items that are worth more than 1 point).
    - **Show Students –** Users can view individual student results by question. Selected students are sorted by last name, first name, middle name, and then the Statewide Student Identifier (SSI). Questions for only a single standard can be displayed at one time, and a standard is automatically selected by default. You may select a different standard in the drop-down box above the student list.

- The **Support Level Reports** display the overall distribution of support levels for a group of students on a particular test which can be filtered by a school, grade, or demographic information (i.e., gender, ethnicity, students with disabilities, etc.). The groupings are completely flexible and can be defined to include any specific students of interest.

6.7 Summary

Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 13). Making an integrated evaluative judgment with such a diverse assortment of evidence is challenging given that the validity process is ongoing and exists throughout the duration of the testing program. Overall, there is ample evidence that the NJSSA–S will foster valid inferences and uses. However, the NJSSA–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. 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 NJSSA–S validity evidence by the primary inferences and uses of the test.

6.7.1 Student Support Level Classifications: Overall Raw Score

The most important inferences made from the NJSSA–S involve the student support level classifications. Students are classified as needing "Strong Support," "Some Support," or "Less Support." All interpretations based on NJSSA–S support level classifications should be validated for evaluating student support as it pertains to the KSAs defined in the NJSLS–S.

Validity evidence in support of the proposed support level classification interpretations has been presented throughout this document and within the validity section. The NJSSA–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 standardized processes and procedures. The accuracy of the scoring of all NJSLS–S items was verified. After the test administration, the items were statistically reviewed to ensure they met the assumptions of the proposed IRT model. Finally, both the overall scale and the support level classifications were verified as being internally consistent.

There are some areas in which the validity evidence in support of the support 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. 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 the difficulty of identifying the long-term consequences of a testing program after its first year of large-scale operational use. More pressingly, while there is ample validity evidence presented in both Part 5 of this document and in this validity section that the

challenge of integrating consequential validity evidence into a coherent validity argument (Cizek, 2016), as well as the difficulty of identifying the long-term consequences of a testing program after its first year of large-scale operational use. More pressingly, while there is ample validity evidence presented in both Part 5 of this document and in this validity section that the support level classifications were consistent, the creation of the support levels themselves does not conform to best practices. There was no standard-setting or standards validation for the NJSSA–S. The support levels map directly to cut scores associated with the NJSLA–S, and the Reporting PLDs used on the score reports are based on the KSAs detailed in the NJSLA–S Range PLDs. If the NJSSA–S is continued in the future, it would behoove NJDOE to validate the use of the NJSLA–S performance standards as the basis for their support levels.

Overall, the evidence in favor of the valid interpretations of support level classification outweighs the areas in which evidence is lacking or non-existent. The NJSSA–S is a standards-based assessment; thus, the content validity evidence linking the test scores and interpretations to the NJSLS–S and the test blueprint is of chief importance (Sireci et al., 2008). However, there is clearly a need for studying the issues noted above to enhance the validity evidence.

6.7.2 Domains and Practices Subscores

Inferences and uses of subscores are of secondary importance to the overall raw score and support level classifications. Students receive raw scores in the subscore categories. The validity evidence pertaining to interpretations based on NJSSA–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 raw subscore reporting procedures were approved by the NJTAC. Finally, the subscores generally displayed adequate reliability coefficients and CSEMs.

The intercorrelations presented in Part 3.2.1.1 and revisited in Part 6.3.1 of this technical brief 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.

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 NJSSA–S subscores to the NJSLS–S is unclear. The NJSLS–S emphasizes the SEPs, DCIs, and CCCs, whereas the NJSSA–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 NJSSA–S is to provide feedback to schools on their overall performance in the six subscore categories, but it is not clear how to use or interpret that information within the framework of the NJSLS–S. Constructing links between the NJSLS–S and the reporting categories of the NJSSA–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 NJSSA–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 NJSSA–S reporting categories, could provide insights into how to best interpret and use the subscores; as previously noted in Part 2.4, ongoing, iterative improvements to item development and test construction might alleviate the lack of balance between individual scientific practices and the three content domains.

Appendix A: Detailed Test Maps

Table A.1: Grade 6 Test Map — Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	NJSLA–S Range PLD Levels
518043_01	1	TE	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = B1; SEP = B3
518043_03	1	MC	CEDS	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = B1; SEP = B3
518043_05	1	MC	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = B1; SEP = B3
518008_01	1	MC	PACI	PS2	C and E	Physical Science	Investigating	DCI = B2; SEP = D2
518008_02	1	TE	CEDS	PS2	C and E	Physical Science	Sensemaking	DCI = B2; SEP = B1
518008_06	1	TE	EAE	PS2	C and E	Physical Science	Critiquing	DCI = B2; SEP = E2
518010_01	1	MC	AID	PS1	SC	Physical Science	Sensemaking	DCI = B2; SEP = B3
518010_03	1	TE	UMCT	PS1	PAT	Physical Science	Investigating	DCI = B2; SEP = B3
518010_05	1	TE	PACI	PS1	PAT	Physical Science	Investigating	DCI = B2; SEP = B3
518012_07	1	TE	EAE	LS4	SF	Life Science	Critiquing	DCI = C2; SEP = A2
518012_04	1	MC	EAE	LS4	SF	Life Science	Critiquing	DCI = A2; SEP = C2
518012_02	1	TE	EAE	LS4	SF	Life Science	Critiquing	DCI = A2; SEP = C2
519003_01a	1	TE	AQDP	LS1	SF	Life Science	Investigating	DCI = B1; SEP = A1
519003_02a	1	TE	PACI	LS1	PAT	Life Science	Investigating	DCI = B1; SEP = A1
519003_04a	1	TE	UMCT	LS1	S, P, and Q	Life Science	Investigating	DCI = A1; SEP = D1
519003_05a	1	TE	CEDS	LS1	C and E	Life Science	Sensemaking	DCI = B1; SEP = B2
518011_06	1	MC	UMCT	ESS2	S, P, and Q	Earth and Space Science	Investigating	DCI = A1; SEP = B2
518011_09	1	TE	EAE	ESS2	C and E	Earth and Space Science	Critiquing	DCI = A1; SEP = D2
518060_01	1	MC	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = C2; SEP = A1
518060_02	1	TE	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = C1; SEP = A1
518060_03	1	TE	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = C3; SEP = A1
519001_01a	1	TE	AQDP	PS3	E&M	Physical Science	Investigating	DCI = A2; SEP = C2
519001_07b	1	TE	DUM	PS3	E&M	Physical Science	Sensemaking	DCI = A3; SEP = E2
519001_08b	1	TE	EAE	PS3	E&M	Physical Science	Critiquing	DCI = B2; SEP = D2
519001_10b	1	TE	PACI	PS3	E&M	Physical Science	Investigating	DCI = A3; SEP = D3

Table A.2: Grade 6 Test Map — Item Statistics

UIN	Points	Item Type	Rasch	Mean	ITOTC	Median Time
518043_01	1	TE	–0.5302	.67	.51	174
518043_03	1	MC	–1.3127	.79	.38	79
518043_05	1	MC	–1.3463	.81	.45	54
518008_01	1	MC	–1.5379	.81	.31	67
518008_02	1	TE	–0.9556	.73	.44	32
518008_06	1	TE	1.7023	.24	.36	86
518010_01	1	MC	–0.9646	.73	.54	94
518010_03	1	TE	–0.3141	.63	.56	103
518010_05	1	TE	0.4156	.49	.58	73
518012_07	1	TE	0.1263	.53	.46	117
518012_04	1	MC	–0.9766	.74	.49	46
518012_02	1	TE	–0.5176	.67	.55	46
519003_01	1	TE	–0.6058	.67	.37	71
519003_02	1	TE	0.1824	.53	.42	45
519003_04	1	TE	–0.2313	.60	.25	68
519003_05	1	TE	–1.6682	.83	.31	40
518011_06	1	MC	–0.2299	.60	.28	43
518011_09	1	TE	–0.5062	.66	.37	82
518060_01	1	MC	–0.3923	.63	.30	90
518060_02	1	TE	1.2508	.31	.32	78
518060_03	1	TE	0.8560	.39	.30	97
519001_01	1	TE	–0.0962	.58	.47	78
519001_07	1	TE	–1.2766	.78	.38	40
519001_08	1	TE	1.2180	.33	.53	64
519001_10	1	TE	0.5602	.46	.59	53

Table A.3: Grade 9 Test Map — Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	NJSLA–S Range PLD Levels
818077	1	TE	EAE	ESS2	PAT	Earth and Space Science	Critiquing	DCI = A3; SEP = C1
818307_01	1	TE	CEDS	ESS3	SF	Earth and Space Science	Sensemaking	DCI = A3; SEP = A2
818283	1	MC	EAE	PS2	C and E	Physical Science	Critiquing	DCI = B2; SEP = C1
818033_02	1	MC	AQDP	PS4	E&M	Physical Science	Investigating	DCI = B2; SEP = A2
818055_02	1	TE	DUM	LS2	E&M	Life Science	Sensemaking	DCI = A2; SEP = E2
818055_01	1	TE	CEDS	LS2	C and E	Life Science	Sensemaking	DCI = C3; SEP = B2
818055_03	1	TE	DUM	LS2	SC	Life Science	Sensemaking	DCI = B3; SEP = E2
818095_01	1	MC	CEDS	ESS3	C and E	Earth and Space Science	Sensemaking	DCI = A2; SEP = B2
818300_01	1	TE	UMCT	ESS1	S, P, and Q	Earth and Space Science	Investigating	DCI = B2; SEP = C2
818306_01	1	MC	AQDP	ESS3	SF	Earth and Space Science	Investigating	DCI = A2; SEP = A1
818302	1	MC	AQDP	ESS1	S, P, and Q	Earth and Space Science	Investigating	DCI = A2; SEP = C2
818267	1	MC	AQDP	ESS2	S&SM	Earth and Space Science	Investigating	DCI = A3; SEP = B2
818271	1	TE	EAE	ESS2	PAT	Earth and Space Science	Critiquing	DCI = A3; SEP = C3
818003_02a	1	TE	AID	PS3	PAT	Physical Science	Sensemaking	DCI = A3; SEP = A2
818003_01a	1	TE	DUM	PS3	E&M	Physical Science	Sensemaking	DCI = B3; SEP = D2
818003_03a	2	TE	EAE	PS3	E&M	Physical Science	Critiquing	DCI = A3; SEP = B3
818109	1	TE	AID	LS4	C and E	Life Science	Sensemaking	DCI = B2; SEP = D2
818296_02	1	TE	DUM	LS3	PAT	Life Science	Sensemaking	DCI = A2; SEP = E2
818065	1	TE	EAE	LS1	SF	Life Science	Critiquing	DCI = C3; SEP = C3
818062	1	MC	PACI	LS1	E&M	Life Science	Investigating	DCI = A1; SEP = A3
818250	1	TE	PACI	PS4	SF	Physical Science	Investigating	DCI = B2; SEP = B4
818285	1	TE	UMCT	PS2	S&SM	Physical Science	Investigating	DCI = B2; SEP = D2
818089_01	1	MC	AQDP	PS2	C and E	Physical Science	Investigating	DCI = A1; SEP = A1
818028	1	TE	AID	PS1	SF	Physical Science	Sensemaking	DCI = C2; SEP = D1

Table A.4: Grade 9 Test Map — Item Statistics

UIN	Points	Item Type	Rasch	Mean	ITOTC	Median Time
818077	1	TE	–0.5325	.52	.30	100
818307_01	1	TE	–0.9462	.61	.41	85
818283	1	MC	–0.8181	.59	.54	55
818033_02	1	MC	0.4011	.34	.47	60
818055_02	1	TE	–3.1142	.92	.34	60
818055_01	1	TE	–0.1868	.45	.46	95
818055_03	1	TE	–0.3573	.49	.49	59
818095_01	1	MC	–1.7725	.76	.49	49
818300_01	1	TE	1.1610	.20	.25	105
818306_01	1	MC	–1.0357	.63	.47	59
818302	1	MC	0.1264	.38	.39	77
818267	1	MC	0.2461	.36	.41	53
818271	1	TE	0.4473	.33	.42	64
818003_02a	1	TE	–0.1778	.45	.35	82
818003_01a	1	TE	–0.0686	.42	.27	44
818003_03a	2	TE	0.9899	.49	.34	94
818109	1	TE	–1.6263	.74	.43	67
818296_02	1	TE	–0.2348	.46	.49	98
818065	1	TE	0.8934	.25	.44	53
818062	1	MC	–0.0026	.40	.26	80
818250	1	TE	1.3884	.17	.19	74
818285	1	TE	0.3165	.36	.56	95
818089_01	1	MC	–0.3900	.49	.31	54
818028	1	TE	0.5373	.31	.32	58

Table A.5: Grade 12 Test Map — Metadata

UIN	Points	Item Type	SEP	DCI	CCC	Domain	Practice	NJSLA–S Range PLD Levels
HS18038_02	1	TE	AID	LS4	S, P, and Q	Life Science	Sensemaking	DCI = A1; SEP = A3
HS18038_10	1	TE	AID	LS4	S&SM	Life Science	Sensemaking	DCI = A1; SEP = E2
HS18038_12	1	TE	EAE	LS4	S, P, and Q	Life Science	Critiquing	DCI = A2; SEP = B2
HS18038_16	1	TE	OECI	LS4	S&SM	Life Science	Critiquing	DCI = A2; SEP = B2
HS18004_01	1	MC	OECI	PS1	PAT	Physical Science	Critiquing	DCI = B2; SEP = A1
HS18004_04	1	TE	DUM	PS1	PAT	Physical Science	Sensemaking	DCI = B2; SEP = C2
HS18004_05	1	MC	DUM	PS1	S&SM	Physical Science	Sensemaking	DCI = B2; SEP = F1
HS18069_01	1	TE	AID	LS2	S, P, and Q	Life Science	Sensemaking	DCI = A1; SEP = A1
HS18069_04	1	MC	AID	LS2	SC	Life Science	Sensemaking	DCI = C1; SEP = A1
HS18069_07	1	TE	OECI	LS2	SC	Life Science	Critiquing	DCI = C2; SEP = A1
HS18013_01	1	MC	UMCT	ESS1	S&SM	Earth and Space Science	Investigating	DCI = A3; SEP = D3
HS18013_03	1	MC	UMCT	ESS1	S&SM	Earth and Space Science	Investigating	DCI = A4; SEP = F3
HS18013_05	1	MC	DUM	ESS1	S, P, and Q	Earth and Space Science	Sensemaking	DCI = B1; SEP = C2
HS18040_01	1	MC	AQDP	PS3	E&M	Physical Science	Investigating	DCI = C3; SEP = A2
HS18040_03	1	TE	PACI	PS3	C and E	Physical Science	Investigating	DCI = C4; SEP = E1
HS18040_04	1	TE	PACI	PS3	C and E	Physical Science	Investigating	DCI = C3; SEP = E1
HS19004_01a	1	TE	AQD	ESS3	S&SM	Earth and Space Science	Investigating	DCI = B2; SEP = D4
HS19004_03a	1	TE	CEDS	ESS3	S&SM	Earth and Space Science	Sensemaking	DCI = A2; SEP = E1
HS19004_06b	1	TE	OECI	ESS3	C and E	Earth and Space Science	Critiquing	DCI = A2; SEP = A1
HS19004_07a	1	TE	EAE	ESS3	S&SM	Earth and Space Science	Critiquing	DCI = B3; SEP = E3
HS18001_01	1	TE	UMCT	PS2	C and E	Physical Science	Investigating	DCI = A2; SEP = F2
HS18001_07	1	MC	UMCT	PS2	SC	Physical Science	Investigating	DCI = A3; SEP = F2
HS18071_01	1	MC	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = D2; SEP = A2
HS18071_04	1	MC	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = D2; SEP = A1
HS18071_05	1	TE	AID	ESS2	PAT	Earth and Space Science	Sensemaking	DCI = D2; SEP = A2

Table A.6: Grade 12 Test Map — Item Statistics

UIN	Points	Item Type	Rasch	Mean	ITOTC	Median Time
HS18038_02	1	TE	–0.6828	.58	.35	74
HS18038_10	1	TE	–0.2127	.48	.35	67
HS18038_12	1	TE	–1.4240	.71	.47	50
HS18038_16	1	TE	–0.7515	.59	.31	76
HS18004_01	1	MC	–0.0331	.45	.38	41
HS18004_04	1	TE	0.1956	.41	.25	39
HS18004_05	1	MC	–0.0117	.45	.27	38
HS18069_01	1	TE	–0.6516	.57	.55	59
HS18069_04	1	MC	–0.9354	.62	.42	36
HS18069_07	1	TE	0.2613	.39	.59	79
HS18013_01	1	MC	–1.0985	.65	.52	65
HS18013_03	1	MC	–1.3309	.69	.46	37
HS18013_05	1	MC	–0.5107	.54	.41	34
HS18040_01	1	MC	0.1944	.41	.27	32
HS18040_03	1	TE	0.5180	.34	.40	26
HS18040_04	1	TE	0.1295	.42	.59	35
HS19004_01a	1	TE	–0.2159	.48	.53	80
HS19004_03a	1	TE	0.5325	.34	.51	36
HS19004_06b	1	TE	0.6599	.32	.46	43
HS19004_07a	1	TE	0.8530	.29	.42	55
HS18001_01	1	TE	–0.8978	.61	.62	56
HS18001_07	1	MC	–0.0810	.46	.19	53
HS18071_01	1	MC	0.0555	.43	.43	47
HS18071_04	1	MC	–0.9538	.62	.52	38
HS18071_05	1	TE	–0.6498	.56	.55	21

Appendix B: Raw Score Cumulative Frequency Distributions

Table B.1: Grade 6 — Raw Score Cumulative Frequency Distribution — Gender

Raw Score	All Cum. #	All Cum. %	Female Cum. %	Male Cum. %	Non-Binary Cum. %
0	182	0.19	0.15	0.23	0.00
1	891	0.93	0.72	1.14	0.00
2	2,452	2.57	2.08	3.04	12.50
3	4,997	5.24	4.45	6.00	12.50
4	8,303	8.71	7.66	9.72	12.50
5	11,882	12.46	11.20	13.68	12.50
6	15,697	16.47	15.19	17.69	18.75
7	19,573	20.53	19.39	21.63	18.75
8	23,446	24.59	23.47	25.67	18.75
9	27,561	28.91	27.90	29.88	18.75
10	31,946	33.51	32.83	34.17	18.75
11	36,555	38.34	38.03	38.65	18.75
12	41,253	43.27	43.32	43.23	31.25
13	46,222	48.49	48.83	48.16	31.25
14	51,411	53.93	54.57	53.32	31.25
15	56,763	59.54	60.48	58.65	31.25
16	62,266	65.31	66.46	64.22	50.00
17	67,748	71.07	72.35	69.84	56.25
18	73,185	76.77	78.24	75.36	68.75
19	78,469	82.31	83.76	80.93	75.00
20	83,494	87.58	88.80	86.42	75.00
21	87,958	92.26	93.14	91.43	81.25
22	91,414	95.89	96.48	95.32	100.00
23	93,791	98.38	98.61	98.17	100.00
24	94,999	99.65	99.72	99.58	100.00
25	95,332	100.00	100.00	100.00	100.00

Table B.2: Grade 6 — Raw Score Cumulative Frequency Distribution — Ethnicity

Raw Score	All Cum. #	All Cum. %	Am. Indian Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	Pacific Islander Cum. %	White Cum. %
0	182	0.19	0.00	0.04	0.32	0.33	0.54	0.09
1	891	0.93	0.00	0.20	1.75	1.53	0.54	0.42
2	2,452	2.57	2.31	0.60	4.89	4.07	2.69	1.18
3	4,997	5.24	3.47	1.20	9.75	8.45	3.76	2.35
4	8,303	8.71	8.09	1.94	16.02	13.90	5.38	4.08
5	11,882	12.46	10.98	2.81	22.71	19.65	8.60	6.13
6	15,697	16.47	15.61	3.96	29.30	25.49	12.37	8.58
7	19,573	20.53	23.12	5.31	35.69	31.31	17.74	11.22
8	23,446	24.59	28.32	6.94	41.71	36.79	20.97	14.20
9	27,561	28.91	31.21	8.96	47.08	42.54	25.81	17.69
10	31,946	33.51	39.31	11.42	52.90	48.17	30.11	21.60
11	36,555	38.34	43.35	14.22	58.40	53.82	34.95	26.15
12	41,253	43.27	48.55	17.34	63.69	59.36	36.02	31.05
13	46,222	48.49	54.34	21.01	69.06	64.75	41.94	36.50
14	51,411	53.93	59.54	25.37	73.94	70.08	47.31	42.56
15	56,763	59.54	63.01	30.40	78.50	75.32	51.61	48.98
16	62,266	65.31	69.36	36.81	82.85	80.22	58.60	55.75
17	67,748	71.07	72.83	43.58	86.78	84.66	62.90	62.87
18	73,185	76.77	79.19	51.33	90.22	88.61	70.43	70.12
19	78,469	82.31	82.66	59.99	93.32	92.05	76.88	77.31
20	83,494	87.58	87.86	69.47	95.76	94.86	85.48	84.24
21	87,958	92.26	89.60	79.09	97.72	97.09	89.78	90.39
22	91,414	95.89	94.22	87.59	98.91	98.64	95.16	95.07
23	93,791	98.38	99.42	94.64	99.58	99.57	98.92	98.11
24	94,999	99.65	100.00	98.68	99.94	99.92	100.00	99.60
25	95,332	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table B.3: Grade 6 — Raw Score Cumulative Frequency Distribution — Other Demographics

Raw Score	All Cum. #	All Cum. %	EL – Yes Cum. %	EL – No Cum. %	Econ. Dis. – Yes Cum. %	Econ. Dis. – No Cum. %	SWD – Yes Cum. %	SWD – No Cum. %
0	182	0.19	0.80	0.15	0.34	0.12	0.55	0.10
1	891	0.93	3.65	0.73	1.54	0.66	2.60	0.51
2	2,452	2.57	9.83	2.03	4.36	1.76	6.84	1.49
3	4,997	5.24	19.16	4.20	8.99	3.54	13.16	3.23
4	8,303	8.71	30.31	7.09	14.70	5.98	20.70	5.66
5	11,882	12.46	41.38	10.30	21.00	8.58	28.36	8.43
6	15,697	16.47	51.09	13.88	27.29	11.54	35.64	11.59
7	19,573	20.53	59.92	17.59	33.34	14.70	42.00	15.08
8	23,446	24.59	66.85	21.43	39.16	17.97	47.83	18.69
9	27,561	28.91	72.44	25.65	44.97	21.60	53.41	22.69
10	31,946	33.51	77.30	30.24	50.70	25.69	58.88	27.07
11	36,555	38.34	81.90	35.09	56.48	30.09	63.54	31.94
12	41,253	43.27	85.66	40.10	62.01	34.75	68.01	36.99
13	46,222	48.49	88.66	45.48	67.47	39.84	72.29	42.44
14	51,411	53.93	91.42	51.12	72.77	45.35	76.41	48.22
15	56,763	59.54	93.43	57.01	77.71	51.27	80.05	54.33
16	62,266	65.31	95.36	63.07	82.28	57.59	83.63	60.66
17	67,748	71.07	96.71	69.15	86.45	64.06	86.85	67.06
18	73,185	76.77	97.74	75.20	90.04	70.73	89.69	73.49
19	78,469	82.31	98.61	81.09	93.03	77.43	92.34	79.77
20	83,494	87.58	99.23	86.71	95.52	83.97	94.77	85.76
21	87,958	92.26	99.55	91.72	97.46	89.90	96.86	91.10
22	91,414	95.89	99.76	95.60	98.79	94.57	98.48	95.23
23	93,791	98.38	99.91	98.27	99.58	97.84	99.41	98.12
24	94,999	99.65	99.98	99.63	99.92	99.53	99.87	99.60
25	95,332	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table B.4: Grade 9 — Raw Score Cumulative Frequency Distribution — Gender

Raw Score	All Cum. #	All Cum. %	Female Cum. %	Male Cum. %	Non-Binary Cum. %
0	170	0.17	0.13	0.20	0.00
1	801	0.80	0.71	0.88	0.00
2	2,498	2.48	2.18	2.77	1.41
3	6,047	6.01	5.41	6.58	1.41
4	11,619	11.54	10.88	12.19	4.23
5	18,535	18.41	17.80	19.01	4.23
6	26,407	26.23	26.01	26.46	7.04
7	34,438	34.20	34.32	34.12	12.68
8	42,411	42.12	42.78	41.51	18.31
9	49,692	49.35	50.39	48.39	25.35
10	56,642	56.25	57.64	54.96	32.39
11	63,093	62.66	64.52	60.91	38.03
12	69,196	68.72	70.84	66.71	49.30
13	74,811	74.30	76.53	72.17	56.34
14	79,934	79.38	81.84	77.05	57.75
15	84,605	84.02	86.57	81.59	67.61
16	88,738	88.13	90.55	85.81	77.46
17	92,137	91.50	93.65	89.46	83.10
18	94,984	94.33	96.08	92.66	87.32
19	97,109	96.44	97.71	95.22	91.55
20	98,708	98.03	98.92	97.17	95.77
21	99,715	99.03	99.50	98.58	98.59
22	100,294	99.60	99.82	99.40	100.00
23	100,584	99.89	99.95	99.83	100.00
24	100,678	99.99	99.99	99.98	100.00
25	100,693	100.00	100.00	100.00	100.00

Table B.5: Grade 9 — Raw Score Cumulative Frequency Distribution — Ethnicity

Raw Score	All Cum. #	All Cum. %	Am. Indian Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	Pacific Islander Cum. %	White Cum. %
0	170	0.17	0.00	0.03	0.31	0.29	0.00	0.07
1	801	0.80	1.94	0.12	1.35	1.42	0.97	0.32
2	2,498	2.48	2.58	0.46	4.18	4.29	1.93	1.07
3	6,047	6.01	5.81	1.49	9.89	9.76	2.90	3.02
4	11,619	11.54	12.90	3.09	18.97	18.19	6.76	6.14
5	18,535	18.41	21.29	5.74	29.49	27.97	11.11	10.67
6	26,407	26.23	29.03	9.08	40.60	38.74	15.94	16.25
7	34,438	34.20	40.00	12.80	51.36	48.68	24.15	22.88
8	42,411	42.12	43.23	17.20	60.86	57.75	32.37	30.33
9	49,692	49.35	53.55	21.75	68.82	65.52	39.13	37.59
10	56,642	56.25	58.71	27.31	75.50	72.32	46.86	44.95
11	63,093	62.66	65.81	32.97	80.79	78.16	55.07	52.29
12	69,196	68.72	69.68	39.48	85.37	83.02	62.80	59.65
13	74,811	74.30	75.48	46.73	89.29	87.03	70.53	66.55
14	79,934	79.38	79.35	53.88	92.21	90.63	75.36	72.94
15	84,605	84.02	83.87	61.85	94.55	93.36	81.64	78.94
16	88,738	88.13	89.03	69.25	96.46	95.56	84.54	84.41
17	92,137	91.50	91.61	76.18	97.70	97.18	87.92	88.97
18	94,984	94.33	93.55	83.07	98.72	98.32	90.34	92.68
19	97,109	96.44	96.77	88.41	99.26	99.09	93.72	95.53
20	98,708	98.03	99.35	92.88	99.70	99.55	96.14	97.61
21	99,715	99.03	100.00	96.20	99.89	99.80	99.52	98.84
22	100,294	99.60	100.00	98.36	99.96	99.94	100.00	99.54
23	100,584	99.89	100.00	99.57	99.99	99.98	100.00	99.87
24	100,678	99.99	100.00	99.97	100.00	100.00	100.00	99.98
25	100,693	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table B.6: Grade 9 — Raw Score Cumulative Frequency Distribution — Other Demographics

Raw Score	All Cum. #	All Cum. %	EL – Yes Cum. %	EL – No Cum. %	Econ. Dis. – Yes Cum. %	Econ. Dis. – No Cum. %	SWD – Yes Cum. %	SWD – No Cum. %
0	170	0.17	0.85	0.13	0.22	0.15	0.32	0.13
1	801	0.80	4.50	0.60	1.23	0.62	1.57	0.62
2	2,498	2.48	11.94	1.99	4.11	1.82	5.05	1.89
3	6,047	6.01	24.67	5.04	9.84	4.46	12.27	4.55
4	11,619	11.54	40.48	10.05	18.77	8.62	22.58	8.98
5	18,535	18.41	55.97	16.47	29.18	14.07	34.25	14.73
6	26,407	26.23	70.34	23.95	40.38	20.52	45.43	21.77
7	34,438	34.20	80.28	31.83	50.65	27.57	55.34	29.30
8	42,411	42.12	87.13	39.80	60.27	34.80	63.88	37.07
9	49,692	49.35	91.63	47.17	68.16	41.77	70.56	44.43
10	56,642	56.25	94.73	54.27	75.05	48.68	76.18	51.63
11	63,093	62.66	96.31	60.93	80.37	55.52	80.80	58.45
12	69,196	68.72	97.49	67.24	84.98	62.17	84.89	64.97
13	74,811	74.30	98.32	73.06	88.83	68.44	88.11	71.09
14	79,934	79.38	98.82	78.38	91.88	74.35	90.83	76.73
15	84,605	84.02	99.41	83.23	94.26	79.89	92.95	81.95
16	88,738	88.13	99.66	87.53	96.36	84.81	94.84	86.57
17	92,137	91.50	99.80	91.08	97.70	89.00	96.23	90.41
18	94,984	94.33	99.90	94.04	98.67	92.58	97.49	93.60
19	97,109	96.44	99.94	96.26	99.22	95.32	98.52	95.96
20	98,708	98.03	99.98	97.93	99.62	97.39	99.19	97.76
21	99,715	99.03	100.00	98.98	99.86	98.69	99.58	98.90
22	100,294	99.60	100.00	99.58	99.98	99.45	99.82	99.55
23	100,584	99.89	100.00	99.89	99.99	99.85	99.95	99.88
24	100,678	99.99	100.00	99.98	100.00	99.98	99.99	99.98
25	100,693	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table B.7: Grade 12 — Raw Score Cumulative Frequency Distribution — Gender

Raw Score	All Cum. #	All Cum. %	Female Cum. %	Male Cum. %	Non-Binary Cum. %
0	41	0.05	0.02	0.07	0.00
1	211	0.24	0.15	0.32	0.00
2	792	0.89	0.67	1.10	0.00
3	2,190	2.46	1.96	2.95	0.00
4	4,561	5.11	4.12	6.09	0.00
5	7,978	8.94	7.40	10.46	2.50
6	12,197	13.67	11.77	15.56	2.50
7	16,881	18.93	16.90	20.94	2.50
8	22,118	24.80	22.97	26.61	5.00
9	27,471	30.80	29.36	32.23	5.00
10	32,920	36.91	35.93	37.89	7.50
11	38,519	43.18	42.79	43.59	17.50
12	44,160	49.51	49.61	49.43	25.00
13	49,673	55.69	56.24	55.17	30.00
14	55,047	61.71	62.64	60.82	40.00
15	60,222	67.52	68.87	66.20	50.00
16	65,304	73.21	74.81	71.65	62.50
17	70,031	78.51	80.19	76.87	70.00
18	74,339	83.34	84.95	81.76	80.00
19	78,222	87.70	89.34	86.08	82.50
20	81,560	91.44	92.73	90.18	85.00
21	84,301	94.51	95.46	93.58	90.00
22	86,426	96.89	97.62	96.18	90.00
23	87,956	98.61	98.98	98.24	97.50
24	88,864	99.63	99.77	99.49	100.00
25	89,197	100.00	100.00	100.00	100.00

Table B.8: Grade 12 — Raw Score Cumulative Frequency Distribution — Ethnicity

Raw Score	All Cum. #	All Cum. %	Am. Indian Cum. %	Asian Cum. %	Black Cum. %	Hisp. Cum. %	Pacific Islander Cum. %	White Cum. %
0	41	0.05	0.00	0.01	0.04	0.08	0.00	0.03
1	211	0.24	0.89	0.05	0.32	0.38	0.00	0.18
2	792	0.89	0.89	0.17	1.44	1.37	0.96	0.63
3	2,190	2.46	1.79	0.59	3.94	3.77	2.39	1.73
4	4,561	5.11	6.25	1.48	8.05	7.75	2.87	3.66
5	7,978	8.94	14.29	2.98	14.05	13.48	6.22	6.34
6	12,197	13.67	17.86	4.68	21.81	20.49	8.61	9.61
7	16,881	18.93	21.43	6.94	30.14	28.07	13.40	13.37
8	22,118	24.80	28.57	9.44	38.58	36.45	20.10	17.89
9	27,471	30.80	33.93	12.61	46.92	44.38	25.36	22.82
10	32,920	36.91	38.39	15.74	55.32	51.81	32.06	28.23
11	38,519	43.18	43.75	19.88	62.94	59.42	39.23	33.86
12	44,160	49.51	52.68	24.19	69.84	66.39	44.50	40.14
13	49,673	55.69	60.71	29.16	75.87	72.52	49.76	46.72
14	55,047	61.71	66.96	34.68	81.30	77.88	56.46	53.46
15	60,222	67.52	74.11	40.49	85.95	82.80	63.16	60.13
16	65,304	73.21	79.46	46.72	89.59	87.17	70.81	67.05
17	70,031	78.51	84.82	53.65	92.72	90.65	75.60	73.61
18	74,339	83.34	87.50	60.92	95.06	93.43	81.34	79.71
19	78,222	87.70	91.07	68.25	96.89	95.69	86.60	85.31
20	81,560	91.44	91.96	76.09	98.11	97.46	92.82	89.91
21	84,301	94.51	93.75	83.31	99.05	98.54	95.69	93.74
22	86,426	96.89	95.54	89.80	99.47	99.34	98.09	96.56
23	87,956	98.61	98.21	94.98	99.82	99.74	99.52	98.54
24	88,864	99.63	100.00	98.52	99.96	99.93	99.52	99.64
25	89,197	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table B.9: Grade 12 — Raw Score Cumulative Frequency Distribution — Other Demographics

Raw Score	All Cum. #	All Cum. %	EL – Yes Cum. %	EL – No Cum. %	Econ. Dis. – Yes Cum. %	Econ. Dis. – No Cum. %	SWD – Yes Cum. %	SWD – No Cum. %
0	41	0.05	0.13	0.04	0.07	0.04	0.09	0.03
1	211	0.24	0.93	0.21	0.32	0.21	0.48	0.18
2	792	0.89	3.60	0.77	1.38	0.71	1.91	0.65
3	2,190	2.46	9.55	2.14	3.85	1.96	5.12	1.83
4	4,561	5.11	18.85	4.51	7.84	4.15	10.24	3.91
5	7,978	8.94	30.70	7.98	13.65	7.28	17.52	6.94
6	12,197	13.67	44.66	12.30	20.58	11.24	25.67	10.86
7	16,881	18.93	57.48	17.22	28.16	15.67	33.72	15.46
8	22,118	24.80	68.43	22.86	36.39	20.70	42.29	20.70
9	27,471	30.80	76.60	28.77	44.80	25.85	50.06	26.29
10	32,920	36.91	83.66	34.84	52.45	31.42	57.19	32.16
11	38,519	43.18	88.58	41.17	60.04	37.23	63.57	38.41
12	44,160	49.51	92.09	47.62	67.04	43.32	69.36	44.86
13	49,673	55.69	94.63	53.96	73.31	49.46	74.33	51.32
14	55,047	61.71	96.14	60.19	78.84	55.66	78.72	57.73
15	60,222	67.52	97.25	66.20	83.64	61.82	82.58	63.99
16	65,304	73.21	98.41	72.10	87.66	68.11	86.06	70.20
17	70,031	78.51	99.10	77.60	91.14	74.05	89.28	75.99
18	74,339	83.34	99.50	82.63	93.89	79.62	91.77	81.37
19	78,222	87.70	99.71	87.16	96.11	84.72	94.26	86.16
20	81,560	91.44	99.81	91.07	97.66	89.24	96.13	90.34
21	84,301	94.51	99.84	94.28	98.63	93.06	97.74	93.76
22	86,426	96.89	99.92	96.76	99.35	96.03	98.75	96.46
23	87,956	98.61	99.97	98.55	99.76	98.20	99.62	98.37
24	88,864	99.63	99.97	99.61	99.97	99.51	99.91	99.56
25	89,197	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Appendix C: Item Parameter Estimates and Model Fit Tables

Table C.1: Grade 6 — IRT Item Parameter Estimates and Fit Statistics

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
518008_01	–1.53790	1.08	1.16	.35	0.88	.05	.75
518008_02	–0.95565	0.96	0.89	.44	1.09	.10	.69
518008_06	1.70235	1.00	1.18	.32	0.97	.00	.20
518010_01	–0.96460	0.89	0.81	.55	1.20	.00	.63
518010_03	–0.31411	0.81	0.74	.61	1.41	.00	.52
518010_05	0.41557	0.88	0.83	.56	1.25	.00	.45
518011_06	–0.22986	1.07	1.15	.41	0.82	.00	.49
518011_09	–0.50623	1.05	1.07	.41	0.89	.01	.57
518012_02	–0.51762	0.89	0.85	.54	1.23	.00	.57
518012_04	–0.97657	0.86	0.77	.49	1.26	.05	.72
518012_07	0.12630	0.96	0.95	.48	1.08	.00	.47
518043_01	–0.53023	0.88	0.81	.55	1.26	.00	.58
518043_03	–1.31270	1.05	1.08	.38	0.93	.07	.72
518043_05	–1.34630	0.86	0.79	.51	1.22	.00	.73
518060_01	–0.39235	1.19	1.33	.30	0.57	.17	.59
518060_02	1.25079	1.07	1.30	.29	0.84	.03	.25
518060_03	0.85600	1.12	1.29	.30	0.75	.04	.31
519001_01a	–0.09625	0.90	0.86	.53	1.22	.00	.50
519001_07b	–1.27660	0.94	0.88	.47	1.10	.00	.71
519001_08b	1.21804	0.76	0.72	.46	1.35	.00	.22
519001_10b	0.56024	0.80	0.78	.55	1.39	.00	.35
519003_01a	–0.60579	1.10	1.12	.41	0.80	.00	.55
519003_02a	0.18245	1.02	1.01	.44	0.97	.01	.45
519003_04a	–0.23129	1.19	1.32	.31	0.56	.08	.55
519003_05a	–1.66820	1.00	1.09	.37	0.98	.00	.77

Table C.2: Grade 9 — IRT Item Parameter Estimates and Fit Statistics

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
818003_01a	–0.06862	1.15	1.22	.21	0.63	.07	.38
818003_02a	–0.17777	1.07	1.09	.30	0.82	.05	.42
818003_03a	0.98992	1.16	1.18	.31	0.83	.03	.45
818028	0.53728	1.01	1.04	.31	0.98	.00	.27
818033_02	0.40108	0.99	0.99	.41	1.02	.01	.33
818055_01	–0.18681	0.94	0.94	.39	1.14	.00	.37
818055_02	–3.11420	0.77	0.53	.33	1.20	.00	.90
818055_03	–0.35735	0.90	0.86	.45	1.28	.00	.43
818062	–0.00256	1.11	1.21	.26	0.71	.06	.38
818065	0.89336	0.80	0.73	.37	1.26	.00	.17
818077	–0.53246	1.13	1.19	.24	0.62	.08	.48
818089_01	–0.39005	1.15	1.23	.21	0.58	.09	.41
818095_01	–1.77250	0.77	0.65	.45	1.42	.00	.75
818109	–1.62630	0.86	0.81	.45	1.28	.00	.68
818250	1.38845	1.01	1.34	.12	0.93	.01	.13
818267	0.24607	1.02	1.09	.38	0.94	.03	.36
818271	0.44726	1.03	1.01	.39	0.96	.02	.33
818283	–0.81809	0.85	0.80	.49	1.43	.00	.55
818285	0.31647	0.74	0.66	.51	1.49	.00	.25
818296_02	–0.23481	0.85	0.81	.47	1.38	.00	.35
818300_01	1.16099	1.08	1.30	.22	0.88	.02	.19
818302	0.12636	1.07	1.10	.35	0.85	.04	.39
818306_01	–1.03570	0.89	0.84	.44	1.31	.02	.62
818307_01	–0.94620	0.97	0.95	.36	1.10	.05	.61

Table C.3: Grade 12 — IRT Item Parameter Estimates and Fit Statistics

Item	Rasch	Infit	Outfit	Corr.	Discrim.	Lower	Item Mean
HS18001_01	–0.89782	0.93	0.88	.55	1.19	.00	.53
HS18001_07	–0.08102	1.32	1.45	.07	0.08	.21	.47
HS18004_01	–0.03307	1.06	1.07	.30	0.84	.01	.41
HS18004_04	0.19559	1.16	1.24	.19	0.57	.07	.37
HS18004_05	–0.01170	1.16	1.23	.20	0.54	.07	.40
HS18013_01	–1.09850	1.06	1.03	.36	0.90	.00	.63
HS18013_03	–1.33090	0.94	0.93	.37	1.11	.00	.72
HS18013_05	–0.51066	1.00	0.97	.38	1.02	.02	.55
HS18038_02	–0.68282	1.08	1.09	.30	0.79	.05	.57
HS18038_10	–0.21273	1.06	1.08	.32	0.84	.09	.52
HS18038_12	–1.42390	0.83	0.73	.42	1.31	.00	.76
HS18038_16	–0.75146	1.09	1.13	.27	0.77	.12	.61
HS18040_01	0.19436	1.16	1.23	.24	0.59	.10	.43
HS18040_03	0.51802	1.11	1.10	.34	0.78	.06	.40
HS18040_04	0.12950	0.86	0.81	.51	1.37	.00	.43
HS18069_01	–0.65162	0.88	0.84	.46	1.32	.00	.61
HS18069_04	–0.93538	0.96	0.95	.37	1.09	.00	.65
HS18069_07	0.26127	0.90	0.85	.49	1.25	.00	.43
HS18071_01	0.05550	1.00	1.01	.37	1.00	.02	.45
HS18071_04	–0.95385	0.88	0.82	.44	1.29	.00	.66
HS18071_05	–0.64984	0.88	0.83	.47	1.33	.00	.59
HS19004_01a	–0.21588	0.90	0.86	.47	1.28	.00	.54
HS19004_03a	0.53253	0.94	0.89	.46	1.13	.00	.38
HS19004_06b	0.65995	0.98	0.95	.39	1.05	.00	.33
HS19004_07a	0.85302	1.06	1.01	.36	0.93	.01	.32

Appendix D: Scale Stability Results Tables

Table D.1: Grade 6 — Delta Plot Results Table

UIN	Item Type	Points	P-Value 2019	P-Value 2021	Delta 2019	Delta 2021	Distance	Distance Limit	Decision
818003_01a	TE	1	0.423	0.410	13.778	13.910	0.056	0.086	Stable
818003_02a	TE	1	0.448	0.430	13.528	13.705	0.021	0.086	Stable
818008_01	MC	1	0.808	0.792	9.315	9.730	0.101	0.090	Flagged
818003_03a	TE	2	0.247	0.240	15.740	15.825	0.029	0.086	Stable
818008_02	TE	1	0.731	0.704	10.532	10.860	0.018	0.090	Stable
818028	TE	1	0.252	0.230	15.669	15.955	0.113	0.086	Flagged
818008_06	TE	1	0.244	0.228	15.773	15.979	0.019	0.090	Stable
818033_02	MC	1	0.328	0.320	14.670	14.871	0.021	0.086	Stable
818010_01	MC	1	0.735	0.704	10.490	10.859	0.011	0.090	Stable
818055_01	TE	1	0.453	0.430	13.473	13.705	0.007	0.086	Stable
818010_03	TE	1	0.633	0.590	11.639	12.090	0.089	0.090	Stable
818055_02	TE	1	0.921	0.900	7.342	7.874	0.036	0.086	Stable
818010_05	TE	1	0.486	0.450	13.142	13.506	0.031	0.090	Stable
818055_03	TE	1	0.493	0.460	13.069	13.402	0.067	0.086	Stable
818011_06	MC	1	0.601	0.575	11.973	12.244	0.036	0.090	Stable
818062	MC	1	0.171	0.170	16.708	16.817	0.045	0.086	Stable
818011_09	TE	1	0.655	0.626	11.403	11.711	0.018	0.090	Stable
818065	TE	1	0.403	0.390	13.982	14.117	0.048	0.086	Stable
818012_02	TE	1	0.667	0.628	11.272	11.696	0.063	0.090	Stable
818077	TE	1	0.516	0.500	12.842	13.000	0.066	0.086	Stable
818012_04	MC	1	0.739	0.708	10.440	10.808	0.009	0.090	Stable
818089_01	MC	1	0.306	0.290	15.031	15.214	0.019	0.086	Stable
818012_07	TE	1	0.528	0.502	12.715	12.983	0.024	0.090	Stable
818095_01	MC	1	0.761	0.740	10.168	10.427	0.075	0.086	Stable
818043_01	TE	1	0.669	0.629	11.250	11.680	0.067	0.090	Stable
818109	TE	1	0.743	0.710	10.386	10.786	0.034	0.086	Stable
818043_03	MC	1	0.791	0.761	9.762	10.158	0.017	0.090	Stable
818250	TE	1	0.356	0.330	14.481	14.760	0.071	0.086	Stable
818043_05	MC	1	0.805	0.766	9.562	10.102	0.118	0.090	Flagged
818267	MC	1	0.362	0.340	14.417	14.650	0.037	0.086	Stable
818060_01	MC	1	0.625	0.604	11.722	11.949	0.070	0.090	Stable
818271	TE	1	0.328	0.310	14.785	14.983	0.023	0.086	Stable
818060_02	TE	1	0.310	0.295	14.980	15.150	0.056	0.090	Stable
818283	MC	1	0.592	0.560	12.073	12.396	0.020	0.086	Stable
818060_03	TE	1	0.386	0.367	14.153	14.381	0.044	0.090	Stable
818285	TE	1	0.485	0.470	13.150	13.301	0.062	0.086	Stable
819001_01a	TE	1	0.580	0.549	12.194	12.512	0.002	0.090	Stable
818296_02	TE	1	0.464	0.440	13.357	13.604	0.014	0.086	Stable
819001_07b	TE	1	0.777	0.754	9.946	10.247	0.047	0.090	Stable
818300_01	TE	1	0.204	0.200	16.311	16.366	0.034	0.086	Stable
819001_08b	TE	1	0.330	0.303	14.764	15.068	0.036	0.090	Stable
818302	MC	1	0.384	0.370	14.182	14.327	0.024	0.086	Stable
819001_10b	TE	1	0.459	0.422	13.417	13.788	0.061	0.090	Stable
818306_01	MC	1	0.630	0.600	11.672	11.987	0.013	0.086	Stable
819003_01a	TE	1	0.671	0.643	11.233	11.531	0.027	0.090	Stable
818307_01	TE	1	0.607	0.580	11.910	12.192	0.004	0.086	Stable
819003_02a	TE	1	0.531	0.497	12.694	13.030	0.024	0.090	Stable
519003_04a	TE	1	0.598	0.574	12.004	12.250	0.052	0.090	Stable
519003_05a	TE	1	0.832	0.810	9.155	9.492	0.035	0.090	Stable

Table D.2: Grade 9 — Delta Plot Results Table

UIN	Item Type	Points	P-Value 2019	P-Value 2021	Delta 2019	Delta 2021	Distance	Distance Limit	Decision
818003_01a	TE	1	0.448	0.430	13.528	13.705	0.021	0.086	Stable
818008_01	MC	1	0.808	0.792	9.315	9.730	0.101	0.090	Flagged
818003_03a	TE	2	0.247	0.240	15.740	15.825	0.029	0.086	Stable
818008_02	TE	1	0.731	0.704	10.532	10.860	0.018	0.090	Stable
818028	TE	1	0.252	0.230	15.669	15.955	0.113	0.086	Flagged
818008_06	TE	1	0.244	0.228	15.773	15.979	0.019	0.090	Stable
818033_02	MC	1	0.328	0.320	14.670	14.871	0.021	0.086	Stable
818010_01	MC	1	0.735	0.704	10.490	10.859	0.011	0.090	Stable
818055_01	TE	1	0.453	0.430	13.473	13.705	0.007	0.086	Stable
818010_03	TE	1	0.633	0.590	11.639	12.090	0.089	0.090	Stable
818055_02	TE	1	0.921	0.900	7.342	7.874	0.036	0.086	Stable
818010_05	TE	1	0.486	0.450	13.142	13.506	0.031	0.090	Stable
818055_03	TE	1	0.493	0.460	13.069	13.402	0.067	0.086	Stable
818011_06	MC	1	0.601	0.575	11.973	12.244	0.036	0.090	Stable
818062	MC	1	0.171	0.170	16.708	16.817	0.045	0.086	Stable
818011_09	TE	1	0.655	0.626	11.403	11.711	0.018	0.090	Stable
818065	TE	1	0.403	0.390	13.982	14.117	0.048	0.086	Stable
818012_02	TE	1	0.667	0.628	11.272	11.696	0.063	0.090	Stable
818077	TE	1	0.516	0.500	12.842	13.000	0.066	0.086	Stable
818012_04	MC	1	0.739	0.708	10.440	10.808	0.009	0.090	Stable
818089_01	MC	1	0.306	0.290	15.031	15.214	0.019	0.086	Stable
818012_07	TE	1	0.528	0.502	12.715	12.983	0.024	0.090	Stable
818095_01	MC	1	0.761	0.740	10.168	10.427	0.075	0.086	Stable
818043_01	TE	1	0.669	0.629	11.250	11.680	0.067	0.090	Stable
818109	TE	1	0.743	0.710	10.386	10.786	0.034	0.086	Stable
818043_03	MC	1	0.791	0.761	9.762	10.158	0.017	0.090	Stable
818250	TE	1	0.356	0.330	14.481	14.760	0.071	0.086	Stable
818043_05	MC	1	0.805	0.766	9.562	10.102	0.118	0.090	Flagged
818267	MC	1	0.362	0.340	14.417	14.650	0.037	0.086	Stable
818060_01	MC	1	0.625	0.604	11.722	11.949	0.070	0.090	Stable
818271	TE	1	0.328	0.310	14.785	14.983	0.023	0.086	Stable
818060_02	TE	1	0.310	0.295	14.980	15.150	0.056	0.090	Stable
818283	MC	1	0.592	0.560	12.073	12.396	0.020	0.086	Stable
818060_03	TE	1	0.386	0.367	14.153	14.381	0.044	0.090	Stable
818285	TE	1	0.485	0.470	13.150	13.301	0.062	0.086	Stable
819001_01a	TE	1	0.580	0.549	12.194	12.512	0.002	0.090	Stable
818296_02	TE	1	0.464	0.440	13.357	13.604	0.014	0.086	Stable
819001_07b	TE	1	0.777	0.754	9.946	10.247	0.047	0.090	Stable
818300_01	TE	1	0.204	0.200	16.311	16.366	0.034	0.086	Stable
819001_08b	TE	1	0.330	0.303	14.764	15.068	0.036	0.090	Stable
818302	MC	1	0.384	0.370	14.182	14.327	0.024	0.086	Stable
819001_10b	TE	1	0.459	0.422	13.417	13.788	0.061	0.090	Stable
818306_01	MC	1	0.630	0.600	11.672	11.987	0.013	0.086	Stable
819003_01a	TE	1	0.671	0.643	11.233	11.531	0.027	0.090	Stable
818307_01	TE	1	0.607	0.580	11.910	12.192	0.004	0.086	Stable
819003_02a	TE	1	0.531	0.497	12.694	13.030	0.024	0.090	Stable
519003_04a	TE	1	0.598	0.574	12.004	12.250	0.052	0.090	Stable
519003_05a	TE	1	0.832	0.810	9.155	9.492	0.035	0.090	Stable

Table D.3: Grade 12 — Delta Plot Results Table

UIN	Item Type	Points	P-Value 2019	P-Value 2021	Delta 2019	Delta 2021	Distance	Distance Limit	Decision
HS18001_01	TE	1	0.609	0.620	11.897	11.778	0.028	0.063	Stable
HS18001_07	MC	1	0.459	0.460	13.412	13.402	0.034	0.063	Stable
HS18004_01	MC	1	0.450	0.450	13.501	13.503	0.042	0.063	Stable
HS18004_04	TE	1	0.407	0.410	13.940	13.910	0.016	0.063	Stable
HS18004_05	MC	1	0.447	0.450	13.529	13.503	0.022	0.063	Stable
HS18013_01	MC	1	0.647	0.660	11.491	11.350	0.039	0.063	Stable
HS18013_03	MC	1	0.686	0.700	11.062	10.902	0.048	0.063	Stable
HS18013_05	MC	1	0.542	0.550	12.575	12.497	0.005	0.063	Stable
HS18038_02	TE	1	0.575	0.580	12.239	12.192	0.020	0.063	Stable
HS18038_10	TE	1	0.484	0.490	13.160	13.100	0.002	0.063	Stable
HS18038_12	TE	1	0.709	0.710	10.804	10.786	0.054	0.063	Stable
HS18038_16	TE	1	0.589	0.590	12.098	12.090	0.048	0.063	Stable
HS18040_01	MC	1	0.409	0.410	13.919	13.910	0.030	0.063	Stable
HS18040_03	TE	1	0.344	0.350	14.611	14.541	0.019	0.063	Stable
HS18040_04	TE	1	0.421	0.420	13.802	13.808	0.041	0.063	Stable
HS18069_01	TE	1	0.565	0.570	12.343	12.295	0.018	0.063	Stable
HS18069_04	MC	1	0.620	0.630	11.776	11.673	0.015	0.063	Stable
HS18069_07	TE	1	0.392	0.400	14.096	14.013	0.023	0.063	Stable
HS18071_01	MC	1	0.430	0.440	13.708	13.604	0.034	0.063	Stable
HS18071_04	MC	1	0.618	0.630	11.795	11.673	0.029	0.063	Stable
HS18071_05	TE	1	0.560	0.570	12.392	12.295	0.017	0.063	Stable
HS19004_01a	TE	1	0.483	0.490	13.167	13.100	0.003	0.063	Stable
HS19004_03a	TE	1	0.343	0.350	14.622	14.541	0.026	0.063	Stable
HS19004_06b	TE	1	0.320	0.330	14.869	14.760	0.049	0.063	Stable
HS19004_07a	TE	1	0.288	0.290	15.237	15.214	0.008	0.063	Stable

Table D.4: Grade 6 — 0.3 Logits Absolute Difference Results Table

UIN	Item Type	Points	Rasch 2019	Rasch 2021	EQK	Adjusted Difficulty	Difficulty Difference	Absolute Difference	Decision
518008_01	MC	1	–1.538	–1.285	–.286	–1.571	.033	.033	Stable
518008_02	TE	1	–0.956	–0.935	–.286	–1.221	.266	.266	Stable
518008_06	TE	1	1.702	2.085	–.286	1.799	–.097	.097	Stable
518010_01	MC	1	–0.965	–0.549	–.286	–0.835	–.129	.129	Stable
518010_03	TE	1	–0.314	0.063	–.286	–0.223	–.091	.091	Stable
518010_05	TE	1	0.416	0.474	–.286	0.188	.228	.228	Stable
518011_06	MC	1	–0.230	0.266	–.286	–0.020	–.210	.210	Stable
518011_09	TE	1	–0.506	–0.218	–.286	–0.504	–.003	.003	Stable
518012_02	TE	1	–0.518	–0.199	–.286	–0.485	–.033	.033	Stable
518012_04	MC	1	–0.977	–1.126	–.286	–1.412	.435	.435	Flagged
518012_07	TE	1	0.126	0.352	–.286	0.066	.060	.060	Stable
518043_01	TE	1	–0.530	–0.237	–.286	–0.523	–.007	.007	Stable
518043_03	MC	1	–1.313	–1.097	–.286	–1.383	.070	.070	Stable
518043_05	MC	1	–1.346	–1.150	–.286	–1.436	.090	.090	Stable
518060_01	MC	1	–0.392	–0.313	–.286	–0.599	.207	.207	Stable
518060_02	TE	1	1.251	1.660	–.286	1.374	–.123	.123	Stable
518060_03	TE	1	0.856	1.263	–.286	0.976	–.120	.120	Stable
519001_01a	TE	1	–0.096	0.210	–.286	–0.076	–.020	.020	Stable
519001_07b	TE	1	–1.277	–1.049	–.286	–1.335	.058	.058	Stable
519001_08b	TE	1	1.218	1.936	–.286	1.650	–.432	.432	Flagged
519001_10b	TE	1	0.560	1.039	–.286	0.753	–.192	.192	Stable
519003_01a	TE	1	–0.606	–0.064	–.286	–0.350	–.256	.256	Stable
519003_02a	TE	1	0.182	0.464	–.286	0.178	.004	.004	Stable
519003_04a	TE	1	–0.231	–0.109	–.286	–0.395	.164	.164	Stable
519003_05a	TE	1	–1.668	–1.480	–.286	–1.766	.098	.098	Stable

Table D.5: Grade 9 — 0.3 Logits Absolute Difference Results Table

UIN	Item Type	Points	Rasch 2019	Rasch 2021	EQK	Adjusted Difficulty	Difficulty Difference	Absolute Difference	Decision
818003_01a	TE	1	–0.069	0.150	–.198	–0.048	–.020	.020	Stable
818003_02a	TE	1	–0.178	–0.037	–.198	–0.236	.058	.058	Stable
818003_03a	TE	2	0.990	1.205	–.198	1.007	–.017	.017	Stable
818028	TE	1	0.537	0.807	–.198	0.609	–.072	.072	Stable
818033_02	MC	1	0.401	0.434	–.198	0.236	.165	.165	Stable
818055_01	TE	1	–0.187	0.222	–.198	0.024	–.211	.211	Stable
818055_02	TE	1	–3.114	–3.113	–.198	–3.311	.197	.197	Stable
818055_03	TE	1	–0.357	–0.101	–.198	–0.299	–.058	.058	Stable
818062	MC	1	–0.003	0.145	–.198	–0.053	.051	.051	Stable
818065	TE	1	0.893	1.459	–.198	1.261	–.367	.367	Flagged
818077	TE	1	–0.532	–0.344	–.198	–0.542	.010	.010	Stable
818089_01	MC	1	–0.390	–0.013	–.198	–0.211	–.179	.179	Stable
818095_01	MC	1	–1.773	–1.826	–.198	–2.024	.251	.251	Stable
818109	TE	1	–1.626	–1.415	–.198	–1.613	–.014	.014	Stable
818250	TE	1	1.388	1.841	–.198	1.643	–.254	.254	Stable
818267	MC	1	0.246	0.268	–.198	0.070	.176	.176	Stable
818271	TE	1	0.447	0.444	–.198	0.246	.202	.202	Stable
818283	MC	1	–0.818	–0.732	–.198	–0.930	.112	.112	Stable
818285	TE	1	0.316	0.927	–.198	0.728	–.412	.412	Flagged
818296_02	TE	1	–0.235	0.313	–.198	0.115	–.350	.350	Flagged
818300_01	TE	1	1.161	1.368	–.198	1.170	–.009	.009	Stable
818302	MC	1	0.126	0.107	–.198	–0.092	.218	.218	Stable
818306_01	MC	1	–1.036	–1.067	–.198	–1.265	.229	.229	Stable
818307_01	TE	1	–0.946	–1.043	–.198	–1.241	.295	.295	Stable

Table D.6: Grade 12 — 0.3 Logits Absolute Difference Results Table

UIN	Item Type	Points	Rasch 2019	Rasch 2021	EQK	Adjusted Difficulty	Difficulty Difference	Absolute Difference	Decision
HS18001_01	TE	1	–0.898	–0.102	–.282	–0.384	–.514	.514	Flagged
HS18001_07	MC	1	–0.081	0.224	–.282	–0.058	–.023	.023	Stable
HS18004_01	MC	1	–0.033	0.494	–.282	0.213	–.246	.246	Stable
HS18004_04	TE	1	0.196	0.711	–.282	0.430	–.234	.234	Stable
HS18004_05	MC	1	–0.012	0.591	–.282	0.309	–.321	.321	Flagged
HS18013_01	MC	1	–1.099	–0.606	–.282	–0.887	–.211	.211	Stable
HS18013_03	MC	1	–1.331	–1.086	–.282	–1.368	.037	.037	Stable
HS18013_05	MC	1	–0.511	–0.175	–.282	–0.457	–.054	.054	Stable
HS18038_02	TE	1	–0.683	–0.314	–.282	–0.596	–.087	.087	Stable
HS18038_10	TE	1	–0.213	–0.032	–.282	–0.313	.101	.101	Stable
HS18038_12	TE	1	–1.424	–1.360	–.282	–1.642	.218	.218	Stable
HS18038_16	TE	1	–0.751	–0.490	–.282	–0.772	.020	.020	Stable
HS18040_01	MC	1	0.194	0.389	–.282	0.107	.087	.087	Stable
HS18040_03	TE	1	0.518	0.543	–.282	0.262	.256	.256	Stable
HS18040_04	TE	1	0.130	0.403	–.282	0.121	.008	.008	Stable
HS18069_01	TE	1	–0.652	–0.479	–.282	–0.760	.109	.109	Stable
HS18069_04	MC	1	–0.935	–0.738	–.282	–1.020	.084	.084	Stable
HS18069_07	TE	1	0.261	0.386	–.282	0.105	.157	.157	Stable
HS18071_01	MC	1	0.056	0.306	–.282	0.024	.032	.032	Stable
HS18071_04	MC	1	–0.954	–0.777	–.282	–1.058	.104	.104	Stable
HS18071_05	TE	1	–0.650	–0.394	–.282	–0.675	.025	.025	Stable
HS19004_01a	TE	1	–0.216	–0.130	–.282	–0.411	.195	.195	Stable
HS19004_03a	TE	1	0.533	0.675	–.282	0.394	.139	.139	Stable
HS19004_06b	TE	1	0.660	0.932	–.282	0.651	.009	.009	Stable
HS19004_07a	TE	1	0.853	1.027	–.282	0.745	.108	.108	Stable

Appendix E: Raw-to-Theta Score Tables

Table E.1: Grade 6 — Raw-to-Theta Score Table

Raw Score	Theta	CSEM	Information	Support Level
0	–5.0236	1.8429	0.29	Strong Support
1	–3.7750	1.0317	0.94	Strong Support
2	–3.0164	0.7527	1.77	Strong Support
3	–2.5432	0.6342	2.49	Strong Support
4	–2.1855	0.5669	3.11	Strong Support
5	–1.8897	0.5235	3.65	Strong Support
6	–1.6318	0.4937	4.10	Strong Support
7	–1.3989	0.4725	4.48	Strong Support
8	–1.1831	0.4574	4.78	Strong Support
9	–0.9791	0.4467	5.01	Strong Support
10	–0.7830	0.4395	5.18	Strong Support
11	–0.5919	0.4353	5.28	Strong Support
12	–0.4033	0.4337	5.32	Strong Support
13	–0.2149	0.4347	5.29	Some Support
14	–0.0246	0.4382	5.21	Some Support
15	0.1697	0.4442	5.07	Some Support
16	0.3709	0.4533	4.87	Some Support
17	0.5818	0.4658	4.61	Some Support
18	0.8063	0.4826	4.29	Some Support
19	1.0498	0.5053	3.92	Less Support
20	1.3202	0.5364	3.48	Less Support
21	1.6307	0.5806	2.97	Less Support
22	2.0053	0.6481	2.38	Less Support
23	2.4974	0.7659	1.70	Less Support
24	3.2772	1.0424	0.92	Less Support
25	4.5419	1.8494	0.29	Less Support

Table E.2: Grade 9 — Raw-to-Theta Score Table

Raw Score	Theta	CSEM	Information	Support Level
0	–5.2107	1.8725	0.29	Strong Support
1	–3.8920	1.0747	0.87	Strong Support
2	–3.0554	0.7957	1.58	Strong Support
3	–2.5249	0.6719	2.22	Strong Support
4	–2.1244	0.5987	2.79	Strong Support
5	–1.7961	0.5500	3.31	Strong Support
6	–1.5131	0.5156	3.76	Strong Support
7	–1.2607	0.4904	4.16	Strong Support
8	–1.0297	0.4718	4.49	Strong Support
9	–0.8138	0.4582	4.76	Some Support
10	–0.6086	0.4484	4.97	Some Support
11	–0.4106	0.4420	5.12	Some Support
12	–0.2170	0.4385	5.20	Some Support
13	–0.0252	0.4378	5.22	Some Support
14	0.1670	0.4398	5.17	Some Support
15	0.3624	0.4447	5.06	Some Support
16	0.5634	0.4527	4.88	Less Support
17	0.7734	0.4644	4.64	Less Support
18	0.9964	0.4808	4.33	Less Support
19	1.2380	0.5033	3.95	Less Support
20	1.5064	0.5345	3.50	Less Support
21	1.8150	0.5791	2.98	Less Support
22	2.1882	0.6475	2.39	Less Support
23	2.6801	0.7663	1.70	Less Support
24	3.4614	1.0437	0.92	Less Support
25	4.7286	1.8506	0.29	Less Support

Table E.3: Grade 12 — Raw-to-Theta Score Table

Raw Score	Theta	CSEM	Information	Support Level
0	–4.8788	1.8410	0.30	Strong Support
1	–3.6351	1.0282	0.95	Strong Support
2	–2.8839	0.7477	1.79	Strong Support
3	–2.4184	0.6279	2.54	Strong Support
4	–2.0687	0.5596	3.19	Strong Support
5	–1.7813	0.5152	3.77	Strong Support
6	–1.5322	0.4845	4.26	Strong Support
7	–1.3086	0.4624	4.68	Strong Support
8	–1.1026	0.4463	5.02	Strong Support
9	–0.9089	0.4346	5.29	Strong Support
10	–0.7237	0.4265	5.50	Strong Support
11	–0.5442	0.4213	5.63	Strong Support
12	–0.3679	0.4188	5.70	Strong Support
13	–0.1927	0.4187	5.70	Some Support
14	–0.0165	0.4211	5.64	Some Support
15	0.1627	0.4262	5.51	Some Support
16	0.3475	0.4342	5.30	Some Support
17	0.5408	0.4457	5.03	Less Support
18	0.7463	0.4618	4.69	Less Support
19	0.9693	0.4838	4.27	Less Support
20	1.2177	0.5145	3.78	Less Support
21	1.5043	0.5588	3.20	Less Support
22	1.8531	0.6272	2.54	Less Support
23	2.3177	0.7471	1.79	Less Support
24	3.0680	1.0277	0.95	Less Support
25	4.3109	1.8408	0.30	Less Support

Appendix F: Conditional Standard Error of Measurement and Test Characteristic Curve Graphs

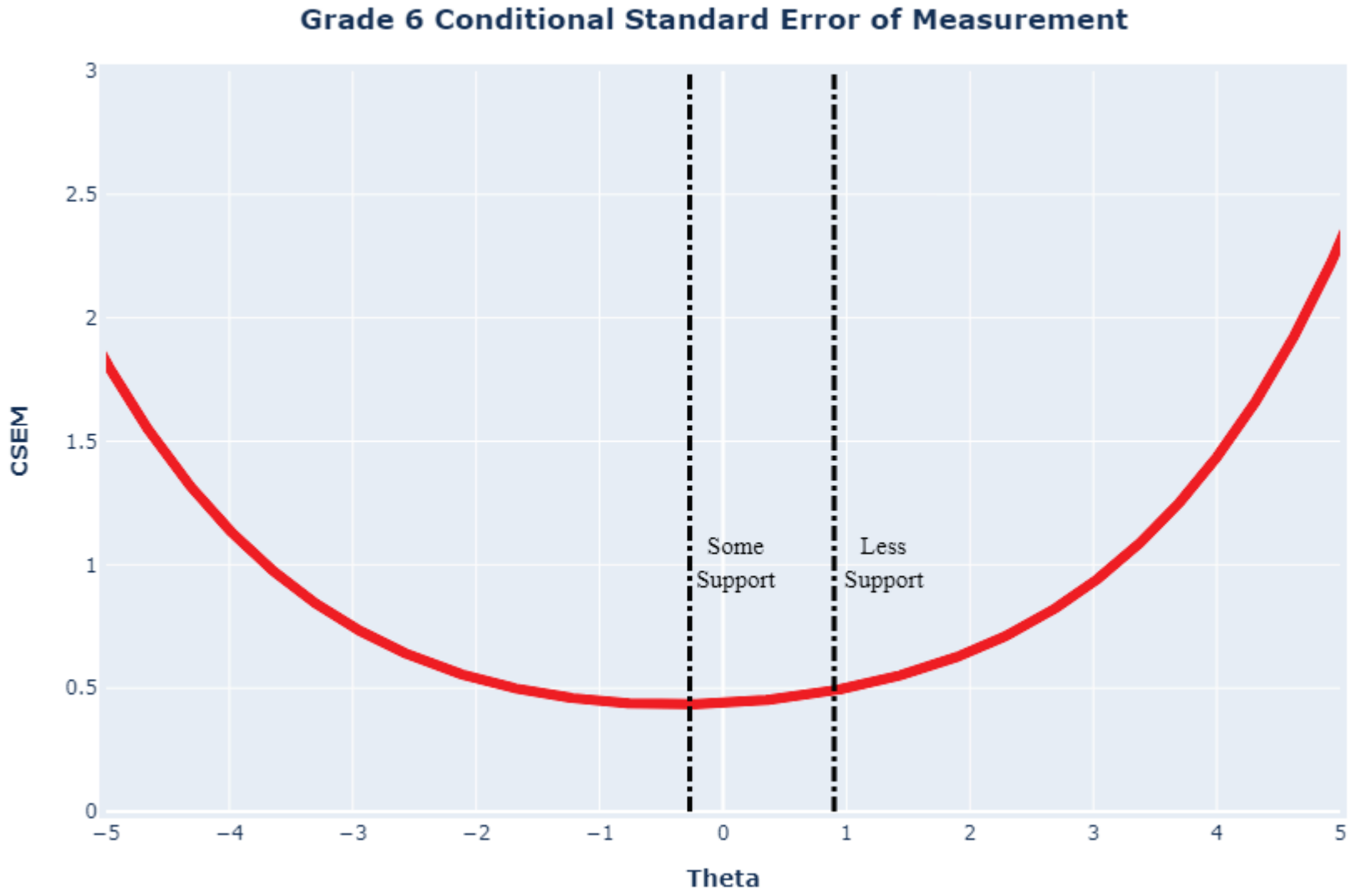

Figure F.1. Grade 6 Conditional Standard Error of Measurement

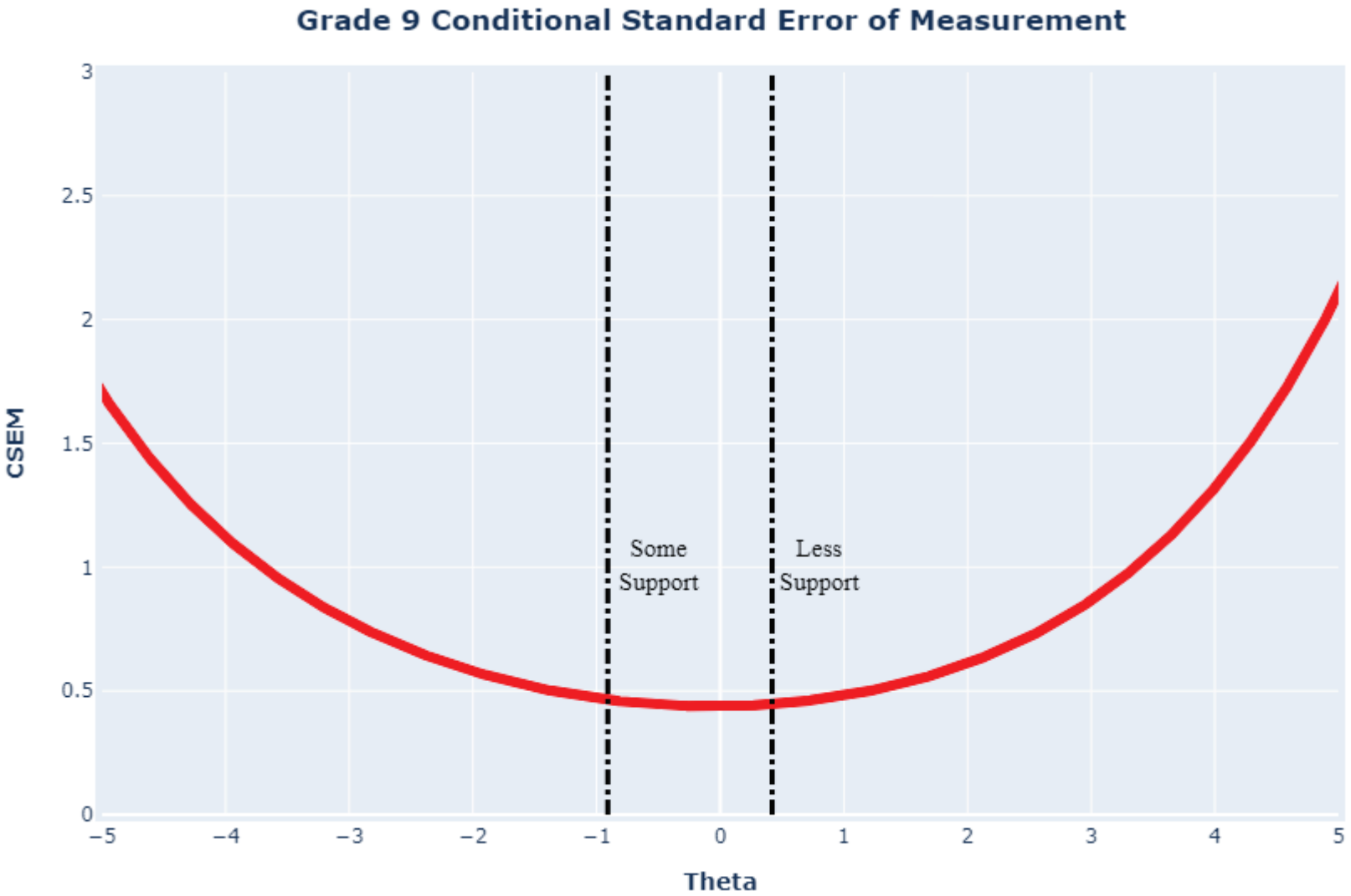

Figure F.2. Grade 9 Conditional Standard Error of Measurement

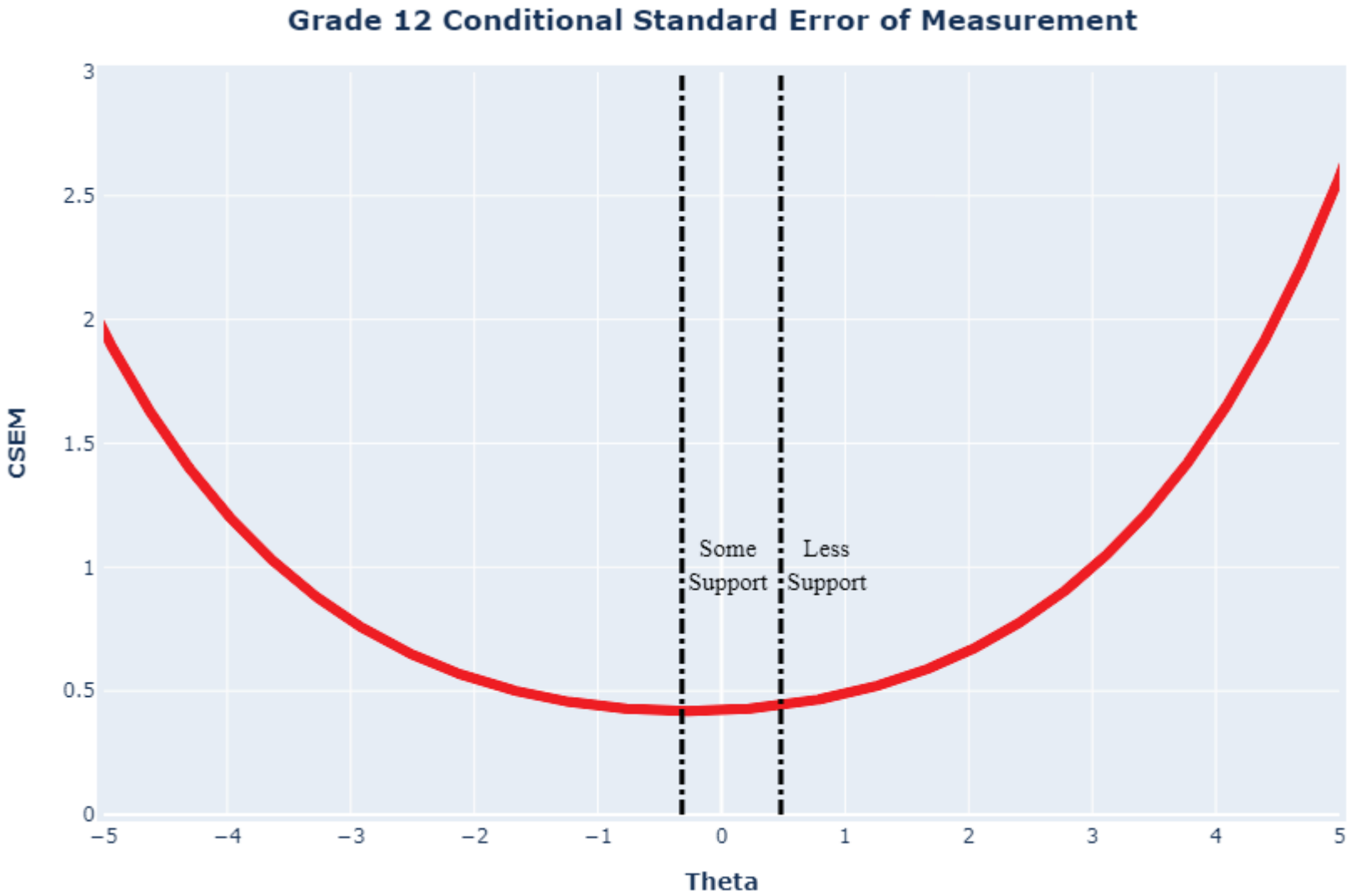

Figure F.3. Grade 12 Conditional Standard Error of Measurement

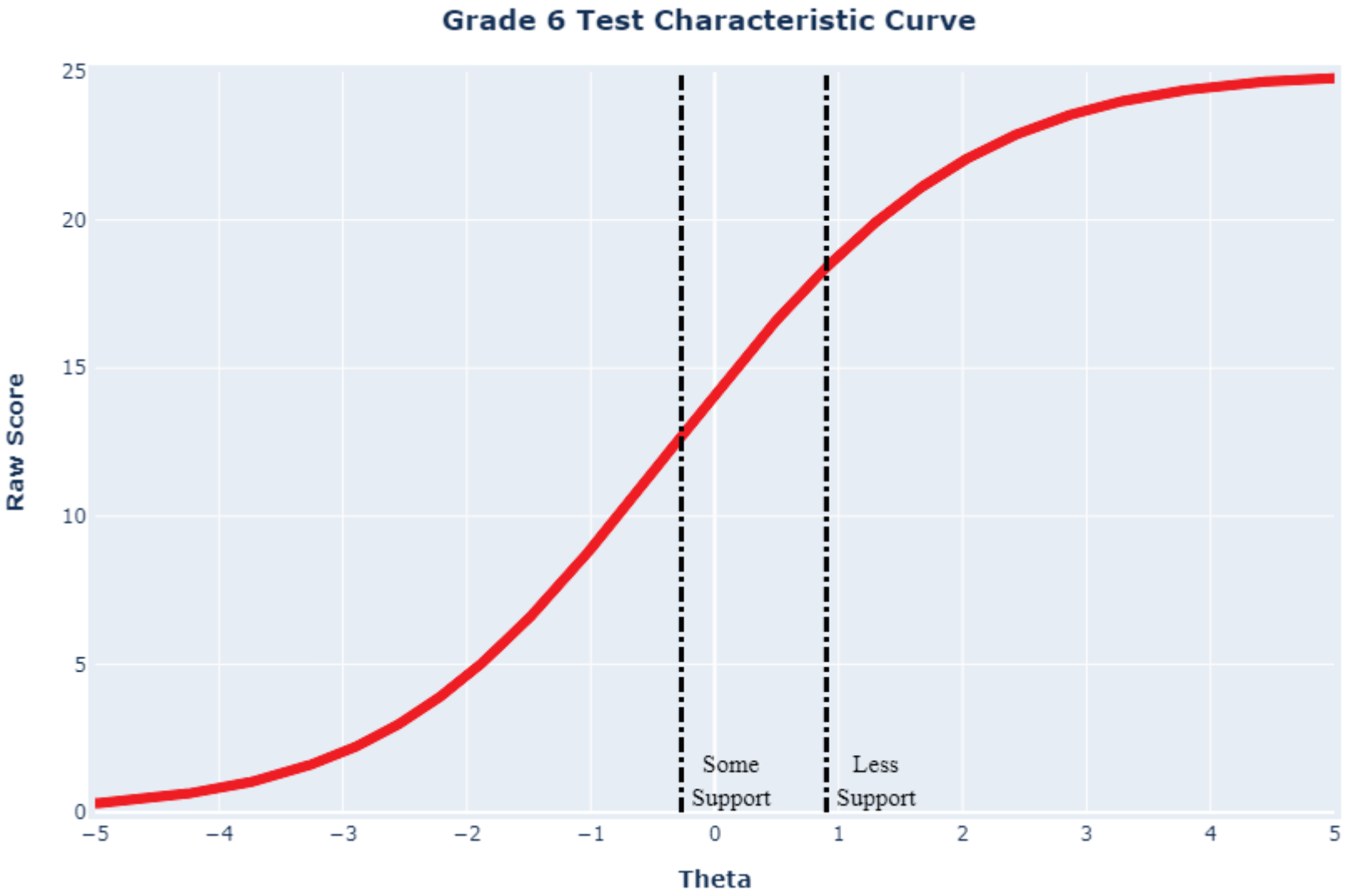

Figure F.4. Grade 6 Test Characteristic Curve

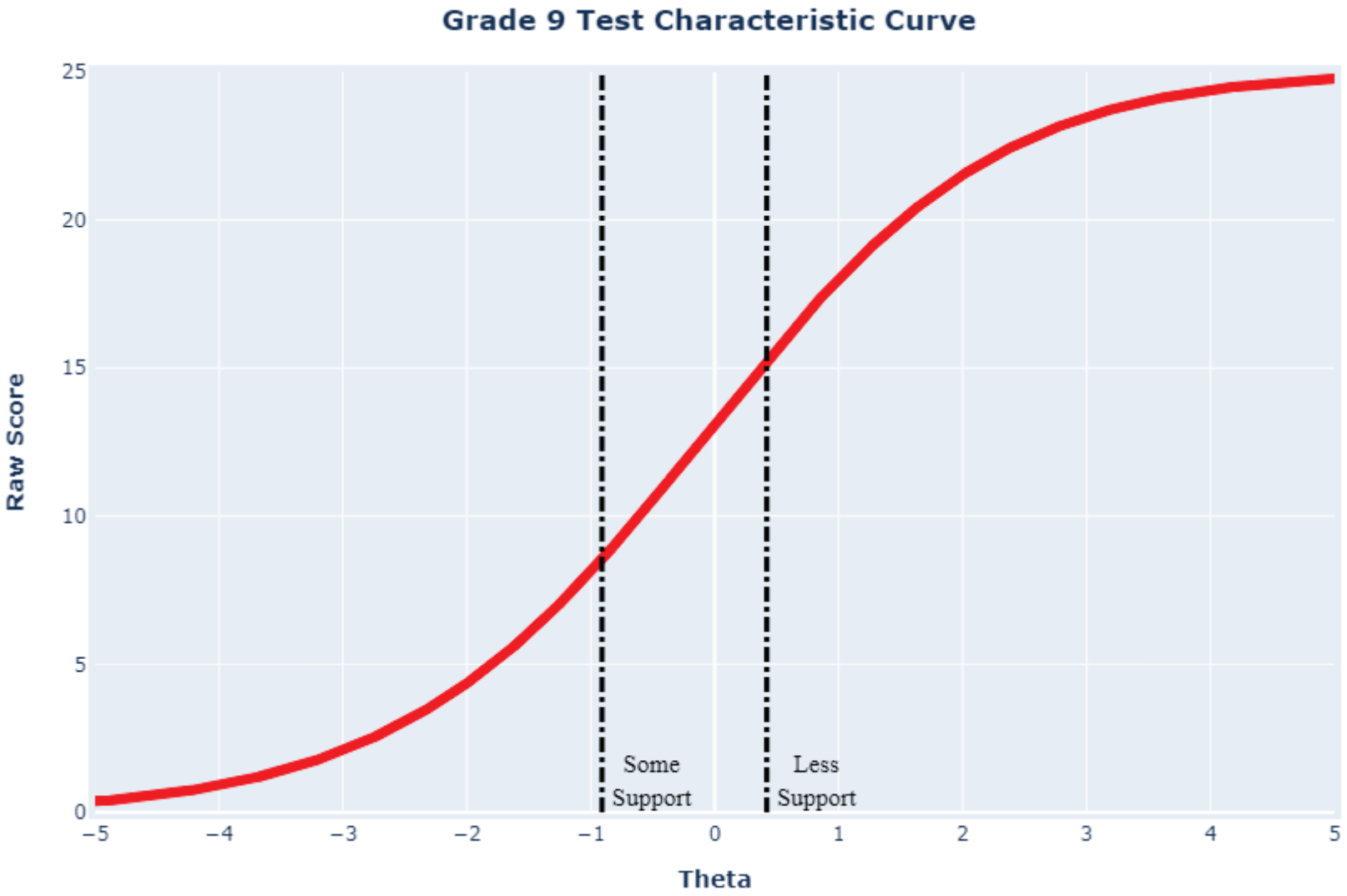

Figure F.5. Grade 9 Test Characteristic Curve

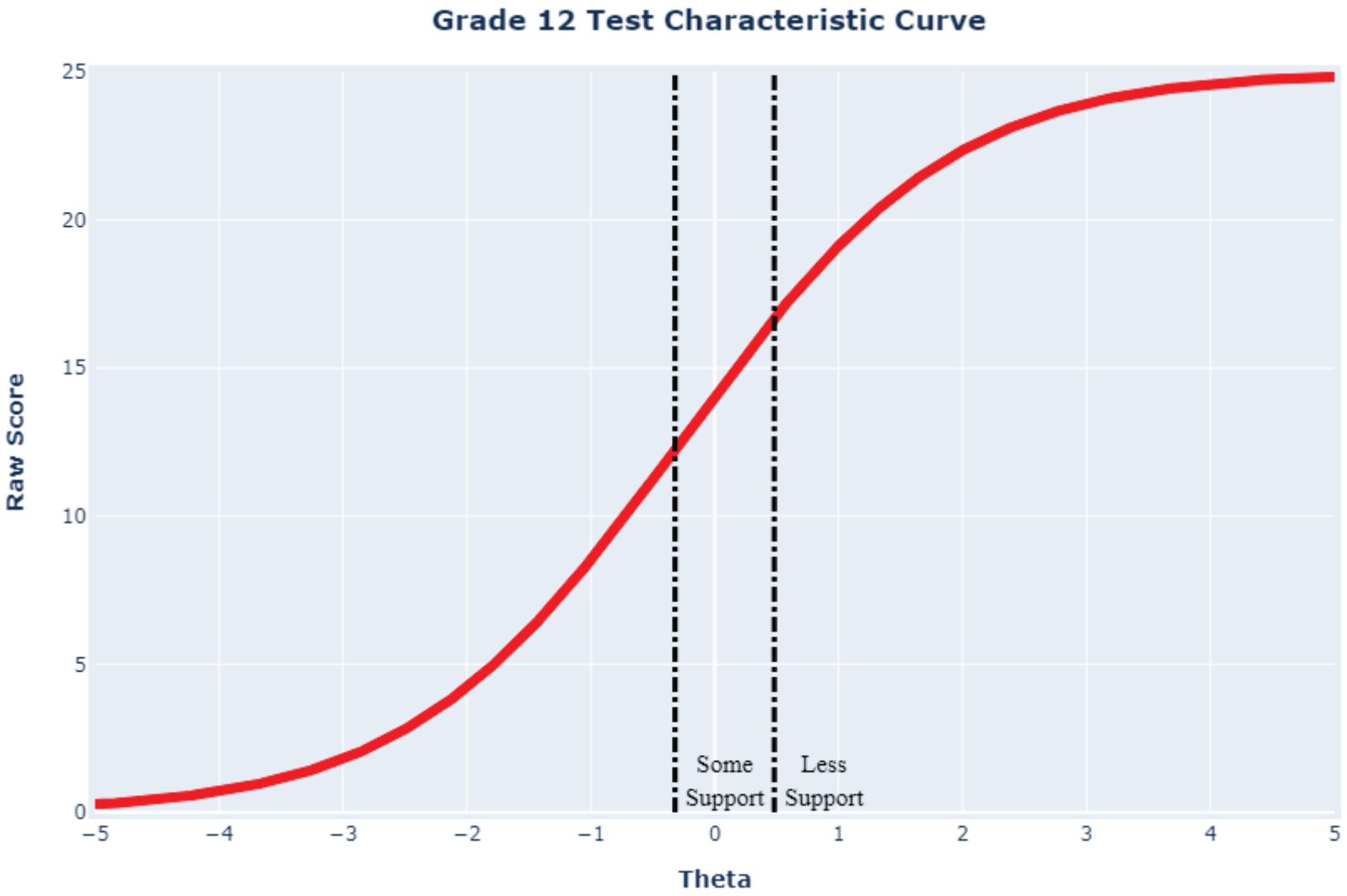

Figure F.6. Grade 12 Test Characteristic Curve

References

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

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069 686)

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore: Johns Hopkins University Press.

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.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

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.

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

Egan, K. L., Schneider, M. C., & 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., & 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, W. J. (1982). Advanced in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*, 373–378.

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.

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

Linacre, J. M. (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-Wesley.

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

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

Miller, G. E., Rotou, O., & Twing, J. S. (2004). Evaluation of the 0.3 logits screening criterion in common item equating. *Journal of Applied Measurement, 5*(2), 172–177.

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). 2014 NJASK Technical Report. Trenton, NJ.

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

New Jersey Department of Education (2019). 2019 NJSLA–S Technical Report. 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.

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.

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

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.

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

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

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.