Illinois Science Assessment (ISA) Technical Report 2024–2025

Prepared by Pearson for the Illinois State Board of Education (ISBE)
December 1, 2025

Table of Contents

List of Tables

List of Figures

Section 1: Introduction

This technical report documents evidence of reliability and validity to support test users in evaluating the intended purposes, uses, and interpretations of the test scores for the spring 2025 administration of the Illinois Science Assessment (ISA). The evidence includes descriptions of the test design, development, and administration procedures; the student test results; and psychometric analyses including calibration, equating, and scaling to ensure that the test results are comparable across different test forms and administrations. The information provided herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

1.1. Assessment Overview

The ISA assessment are Illinois' statewide summative assessments administered each spring to measure student performance on the Illinois Learning Standards for Science incorporating the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) in grades 5 and 8.[1] The goals of the ISA are to measure student performance and monitor student progress toward the content standards; provide students, parents/guardians, and stakeholders with information regarding student score reports; and evaluate the performance and quality of educational programs in Illinois public schools. While ISA is designed as part of a school accountability assessment and the results are used to inform the calculations for the state's accountability, the results can also be used to inform instructional strategies, enrich curriculum, and remediate learning.

The assessments are administered online as fixed forms with paper accommodated forms available as needed, along with a wide range of accessibility features for all students and accommodations for students with disabilities and English learners (ELs), including screen readers, braille, and large print. Student results are reported as an overall scale score and performance level with domain-level scale scores for each of the three domains (Physical Science, Life Science, and Earth/Space Science). The four performance levels are Level 4: *Above Proficient*, Level 3: *Proficient*, Level 2: *Approaching Proficient*, and Level 1: *Below Proficient*. Students performing at Levels 3 and 4 are proficient or above proficient and have demonstrated readiness for the next grade level.

1.2. Background

Prior to the ISA, Illinois administered the Illinois Standards Achievement Test (ISAT) to students in grades 3–8 and the Prairie State Achievement Examination (PSAE) to high school students in reading, mathematics, and science. The last administration of the ISAT and PSAE occurred in 2014, with Illinois administering the Partnership for Assessment of Readiness for College and Careers (PARCC) summative assessments for English language arts/literacy (ELA/L) and mathematics beginning in 2015. To bring Illinois into compliance with the federal No Child Left Behind Act of 2001 (NCLB) that required a science assessment, the Illinois State Board of Education (ISBE) entered into an item-sharing agreement with the Office of the State Superintendent in the District of Columbia and administered the ISA for the first time in 2016 to students in grade 5, grade 8, and high school Biology.

[1] The ISA is not an alternate assessment. Students who participate in the Dynamic Learning Maps-Alternate Assessment (DLM-AA) will be assessed in science in grades 5 and 8 each year and will not take the ISA.

Due to resource constraints, scoring for the 2016 assessment did not take place until July 2017, followed by a data validation and quality review and standard setting, resulting in the 2016 ISA results being reported in January 2018. Results from the 2017 ISA were reported in February 2018, until the reporting schedule was finally established with the 2018 ISA results released to schools on time and officially reported along with all the other 2018 assessment data as part of the 2018 Illinois Report Card. Test development in 2018 consisted of the Southern Illinois University Carbondale (SIUC) coordinating multiple workshops to select NGSS-aligned items from the ISAT and PSAE to augment the 2019 ISA, with ISBE replacing roughly 50% of the 2018 items with these workshop items.

Peer review results from the U.S. Department of Education in January 2019 revealed that the 2015–2018 ISA did not meet federal assessment requirements, as expected, with the direction that "ISBE must have a new or substantially revised general science assessment in place to begin administering in the 2019–20 school year." In line with this feedback, development work for the ISA officially began in January 2019 with the convening of an Illinois Science Assessment Steering Committee (ISASC) to design a test blueprint for a fully redesigned ISA aligned to the Illinois Learning Standards for Science based on the three-dimensional NGSS.

In ISBE's response to the U.S. Department of Education describing their plans for coming into compliance, ISBE requested a waiver for the 2019–2020 assessment to conduct a census field test of multiple forms with all newly constructed items. The 2020 administration was cancelled due to the COVID-19 pandemic, but testing resumed in 2021 with the ISA census field test administered to all eligible students to gather data and validate the test items rather than for accountability purposes. The first operational administration of the redesigned ISA occurred in 2022, with ongoing item development and embedded field testing occurring each year.

ISBE transitioned to the ACT® as the high school accountability assessment for English language arts/literacy (ELA/L), mathematics, and science beginning in the 2024–2025 school year. As such, the ISA assessments include grades 5 and 8 only in 2025 and beyond. The performance level cut scores were established in 2025 during a standard setting to unify the performance levels across the ACT, Illinois Assessment of Readiness (IAR), and ISA. Technology-enhanced items were also field tested in the ISA assessments for the first time in spring 2025.

1.3. Student Participation

As stated in the *Accessibility Features and Accommodations Manual*, all students, including students with disabilities and English learners (ELs), are required to participate in statewide assessments and have their assessment results be part of the state's accountability systems, with narrow exceptions for students with disabilities who have been identified by their Individualized Education Program (IEP) team to take their state's alternate assessment. All other students participate in the statewide assessment. Federal laws governing student participation in statewide assessments include the Every Student Succeeds Act of 2015 (ESSA), the Individuals with Disabilities Education Improvement Act of 2004 (IDEA), Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008), and the Elementary and Secondary Education Act (ESEA) of 1965, as amended.

1.4. Organizations and Groups Involved

Table 1.1 presents the organizations and groups involved in ensuring the success of the ISAs, with each contributor playing a vital role in making sure the assessments yield valid and reliable test results. Input from Illinois educators in the test development and review process is vital to ensure that the assessments reflect the Illinois student population, and feedback from the Technical Advisory Committee (TAC) has been reviewed, addressed, and incorporated into the ISA program.

Table 1.1. Organizations and Groups Involved

Organization/Group	Roles and Responsibilities
Illinois State Board of Education (ISBE)	• Carries out the state and federal requirements for the implementation of the statewide assessments • Oversees the planning, scheduling, and implementation of all major assessment activities and supervises the current contract with partnering testing organizations • Conducts quality control activities for every aspect of the test development, administration, scoring, and reporting processes and monitors the security provisions of the assessment program
Pearson	• Responsible for all test development, administration, and psychometric analyses, including scoring all item responses and providing score reports • Develops future content for the assessment program, including locating and developing appropriate stimulus materials, creating standards-aligned items, coordinating and facilitating item review workshops, and selecting items for the operational and field test forms during form construction
Technical Advisory Committee (TAC)	• A group of psychometric, assessment design, and administration experts who provide consulting and advice for the assessment system • Provides input on the blueprint design, equating and scaling, peer review, standard setting, and reliability and validity issues
Illinois Educators	• Reviews test items and associated stimuli to ensure appropriate grade-level content, alignment to the content standards, and consistency with classroom instruction • Participates in bias and sensitivity reviews to ensure that items are fair and appropriate for all students • Reviews field tested items during data review to determine their eligibility to be included in the operational item pool • Participates in performance level descriptor (PLD) development and standard setting as needed

Section 2: Test Design

Section 2: Test Design provides a comprehensive overview of the processes, committees, and standards involved in developing and maintaining a high-quality assessment program. This section outlines the collaborative efforts among technical experts, educators, and stakeholders to ensure the integrity and relevance of test items, alignment to state standards, and fairness for all students. It also details the structure and content of the assessments, emphasizing the role of the Illinois Learning Standards for Science and the Next Generation Science Standards (NGSS) in shaping the test blueprint and performance expectations. Together, these elements form the foundation for a rigorous and effective assessment system that supports student learning and achievement across the state.

2.1. Content Standards

ISBE adopted the Illinois Learning Standards for Science based on the NGSS in 2014, available online at https://www.nextgenscience.org/search-standards. The NGSS are a set of three-dimensional K–12 science standards first released in 2013 that were informed by growing research in cognitive science and developmental psychology and guided by *A Framework for K–12 Science Education* (National Research Council, 2012). They are designed to reflect more recent research and thinking in science education and were developed to better prepare students with the science knowledge, skills, and habits of mind to be ready for college, career, and civic responsibilities.

The standards emphasize the integration of three dimensions (Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts (CCCs)) in the domains of Physical Science, Life Science, Earth and Space Science, and Engineering, Technology, and Applications of Science. Science instruction aligned to the NGSS requires students to engage in SEPs in the context of DCIs and to use CCCs to make connections across topics. In this way, the NGSS emphasize that science is not just a series of isolated facts. Instead, learning is structured so that students experience science more as an interrelated world of inquiry and phenomena rather than a static set of science disciplines.

The goals for science learning are outlined in the NGSS in the form of performance expectations (PEs) that define what students should be able to know and do at the end of instruction and are designed to help students not just memorize content but to apply knowledge in real-word contexts. Each PE integrates the three dimensions of learning (DCIs, SEPs, and CCCs) and is written to allow teachers to assess whether students can demonstrate their understanding and abilities.

The DCIs encompass the content that occurs at each grade and provides the background knowledge for students to develop sense-making around phenomena in the three domains of Physical Science, Life Science, and Earth and Space Science. The DCIs are as follows, formulating the three domains on the ISA:[2]

- Physical Science: Students know and understand common properties, forms, and changes in matter and energy.
    - PS1: Matter and its interactions
    - PS2: Motion and stability: Forces and interactions
    - PS3: Energy
    - PS4: Waves and their applications in technologies for information transfer

[2] Items on the test are aligned to the fourth standard Engineering Technology, and Application of Science (ETS), but this domain is not reported as a fourth domain subscore.

- Life Science: Students know and understand the characteristics and structure of living things, the processes of life, and how living things interact with each other and their environment.
    - LS1: From molecules to organisms: Structures and processes
    - LS2: Ecosystems: Interactions, energy, and dynamics
    - LS3: Heredity: Inheritance and variation of traits
    - LS4: Biological evolution: Unity and diversity

- Earth/Space Science: Students know and understand the processes and interactions of Earth's systems and the structure and dynamics of Earth and other objects in space.
    - ESS1: Earth's place in the universe
    - ESS2: Earth's systems
    - ESS3: Earth and human activity

The SEPs describe how scientists investigate and build models and theories of the natural world or how engineers design and build systems. They reflect science and engineering as they are practiced and experienced. There are eight SEPs:

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

CCCs cross boundaries between science disciplines and provide an organizational framework to connect knowledge from various disciplines into a coherent and scientifically based view of the world. They build bridges between science and other disciplines and connect the DCIs and SEPs throughout the fields of science and engineering. There are seven CCCs:

1. Patterns
2. Cause and Effect
3. Scale, Proportion, and Quantity
4. Systems and System Models
5. Energy and Matter
6. Structure and Function
7. Stability and Change

2.2. Test Blueprints

Each ISA assessment is based on the grade-level test blueprint that outlines the range and distribution of content and the distribution of points across the domains and item types to guide test construction. The ISA test blueprints were developed in 2019 by the Illinois Science Assessment Steering Committee (ISASC) comprised of 59 participants who met in a series of four all-day meetings on January 28, February 15, March 9, and March 29, 2019, to discuss the standards and determine the test design for the three-dimensional ISA. Table 2.1 presents the high-level blueprints resulting from this committee.

Table 2.1. High-Level Blueprints

Grade	Domain	#MC Items	#CR Items	Total #Items	Total #Points	%Score Points
5	Physical Science	24	1	25	27	33.3
	Life Science	24	1	25	27	33.3
	Earth/Space Science	24	1	25	27	33.3
	Total	72	3	75	81	100.0
8	Physical Science	24	1	25	27	33.3
	Life Science	24	1	25	27	33.3
	Earth/Space Science	24	1	25	27	33.3
	Total	72	3	75	81	100.0

Note. MC = multiple-choice, CR = constructed response

2.3. Test Sections

Table 2.2 presents the number of operational and field test items per test section of the ISA. Each grade-level assessment consists of three sections, each with 25 operational items (24 multiple-choice and 1 constructed response). Each section also contains seven field test items (6 items that are a mixture of multiple-choice and technology-enhanced items and 1 constructed-response item). Sections may be scheduled at any time during the testing window, but all sections must be completed within the window. A test section should take approximately 45 minutes to complete within a two-hour time limit (i.e., a student should not be allowed more than two hours per section).

Table 2.2. Test Sections

Section	#OP Items	#FT Items	Approx. Time
1	25	7	45 min.
2	25	7	45 min.
3	25	7	45 min.
Total	75	21	~135 min.

2.4. Test Structure

Table 2.3 presents an overview of the ISA design for each test form, including the number of item clusters, number of operational items by item type, and number of field test items. Most ISA items are two-dimensional at minimum (measuring both a DCI and SEP), and as many items as possible are three-dimensional (measuring a DCI, SEP, and CCC). Use of one-dimensional items is limited and used only if necessary to target a content standard. All ISA items (in item sets and standalone) involve phenomena and/or problems. Information related to the phenomenon provided by the scenario (e.g., graphs, data tables) is necessary to answer the items with a cluster. Each specific item cluster aligns to at least one domain topic but covers multiple SEPs, CCCs, and individual PEs.

The operational test form is the base form from which all the field test forms are created. Each field test form has the same operational forms but different field test items that do not count toward students' scores. The purpose of field testing is to administer newly developed items to generate item statistics and assess their quality to determine their eligibility to become operational. The goal is to maintain the item bank in terms of quality and quantity to allow for future operational form construction. The online test form is also used as the base for developing the accommodated forms, as described in Section 3.3.

Each ISA test form has 75 operational items on each assessment. They are fixed-form assessments, meaning all students receive the same set of operational items in a predetermined order. The

assessments are administered in three sections and contain multiple-choice and constructed-response items only. Each test form also contains three embedded field test items sets with seven items each that do not count toward a student's score. The multiple-choice items are machine-scored, while the constructed-response items are scored by human raters.

Table 2.3. Test Structure

Grade	Section	#Forms	#Item Clusters	#Items/ Cluster	Total #Items	#MC Items	#CR Items	#Raw Score Points
5	OP	1	12	6–7	75	72	3	81
	FT	12	3	7	21	6	1	0
8	OP	1	12	6–7	75	72	3	81
	FT	12	3	7	21	6	1	0

Note. OP = operational, FT = field test, MC = multiple-choice, CR = constructed-response

Section 3: Test Development

This section describes the process for developing new items for the ISAs to be embedded as field test items in the operational test administration, along with the process for constructing the test forms. ISBE developed the spring 2025 test forms in collaboration with Pearson. The test development process involves multiple steps, as outlined below:

- Update the test specifications for the 2024–2025 development cycle.
- Evaluate the strengths of the item bank and consider the needs for future tests to establish an asset development plan prior to the annual item development cycle.
- Conduct item development with Illinois educators to develop the item cluster sets, including the associated stimuli and scoring rubrics for the constructed-response items, based on the development needs in the asset development plan.
- Review the item cluster sets, both internally and externally by Illinois educator committees. Items that pass are eligible for field testing.
- Construct the test forms that include both operational and field test items.
- Conduct data review to review data about the performance of the field tested items and determine if the items should advance to the operational item bank.

3.1. Content Development and Management Tool

Pearson's Assessment Banking and Building solutions for Interoperable assessments tool (ABBI) is the content management tool, item bank, and publication system supporting both online and paper publication. The item development workflow moves items and assets from inception through a series of content, fairness, graphic, and other reviews to final publication. The system captures the outcomes at each review and maintains previous versions of each item. As items travel through the review process, every version of each asset is archived, along with each comment received in any review. Reviewers have immediate access to all older versions, providing version control throughout development.

Pearson's ABBI manages item content throughout the entire lifecycle of an item. It also manages item content beyond the operational life of the item, including items identified for use in practice tests or other training materials. ABBI provides on-demand reports of the content and item bank status. Each item is directed through a sequence of reviews and approvals by Pearson and ISBE before it is identified for field test or operational administration. ABBI allows remote internet access by item writers and reviewers while ensuring security with individualized passwords for all users, limited access for external users, and strong encryption of all information. Forms are also built in ABBI. After items are used, ABBI stores the resulting statistics, including exposure statistics and classical and IRT statistics.

3.2. Item Development

Individual item assignments were provided to the Illinois item writers to develop the stimuli, items, and rubrics designed to validly measure student understanding of the Illinois Learning Standards for Science. ISBE and Pearson content experts then review the products of the item cluster development for content prior to review by the external content and bias and sensitivity work group during which content and bias experts per review the item sets at each grade level to ensure that every stimulus, item, and rubric is scientifically accurate and gather appropriate evidence about student mastery of the content standards. Participants also review the items and stimuli for any bias and sensitivity issues. Items clusters are accepted as written, accepted as edited, or not accepted. Accepted item clusters continue in the development process.

3.3. Form Construction

Table 3.1 presents the number of test forms constructed for spring 2025. Each grade had one core operational form for regular online testing and 12 field test forms, along with another core operational form used for the various accommodated forms. Test form construction is the process of selecting and sequencing a set of operational and field test items for administration. It is a complex, interactive task that requires both content and psychometric expertise. The test forms were constructed to reflect the test blueprint in terms of content, item types, and test length, as well as expected difficulty and performance along the ability continuum. They were also constructed to adhere to the following goals outlined in the test construction specifications:

- Test forms are designed to appropriately measure the assessment PEs across the full range of ability.
- Scores are comparable across forms and administrations.
- Parallel forms are created among the ISA forms, as possible.
- Forms are developed to industry standards for validity, reliability, and fairness (AERA et al., 2014).

All students receive the same core operational items (either from the regular online or accommodated core form) but different field test items. Field test items were embedded on the test forms and administered in each test section. The test forms were randomly assigned at the student level by TestNav, Pearson's online test delivery platform.

Table 3.1. Number of Test Forms Constructed in Spring 2025

Form Type	Grade 5	Grade 8
Online Operational	1	1
Online Field Test	12	12
Online Accommodated	6	6
English Text-to-Speech (TTS)	1	1
Spanish	1	1
Spanish TTS	1	1
Braille	1	1
American Sign Language	1	1
Human Reader	1	1
Paper Accommodated	5	5
English	1	1
English Large Print	1	1
Spanish	1	1
Spanish Large Print	1	1
Braille	1	1

3.4. Data Review

Following the spring 2025 test administration, an educator data review committee met in August 2025 to evaluate the flagged field tested items and associated performance data in terms of appropriateness, level of difficulty, and any potential differential item functioning (DIF) for groups of interest for grades 5 and 8 only. The committee recommended acceptance or rejection of each field tested item for inclusion in the operational item bank and made recommendations for some items to be revised and re-field tested. Items approved by the committee became eligible for use on either future operational ISAs or the formative assessment (QuISBE) item bank.

The meeting began with a training session that introduced the item review process, including an overview of the item statistics and how they should be used to evaluate items. Decisions about an item's quality cannot be made on statistics alone; the item itself and the content it measures should also be considered. Thus, the groups also reviewed the content of the items and how the items functioned according to the statistics before making a consensus decision about whether the item should be accepted or rejected for operational use. Revisions were recommended for the rejected items if applicable.

Table 3.2 presents the data review results based on the spring 2025 field test data. Committee members made these decisions based on the item content, using the item statistics to guide their discussion. Accepted items were added to the operational ISA item pool or formative (QuISBE) pools for future use.

Table 3.2. Data Review Results: Number of Field Tested Items

Grade	#Items	#Accepted	#Revise and Re-Field Tested	#Rejected
5	248	231	6	11
8	245	222	6	17

Section 4: Test Administration

The spring 2025 ISA testing window was from March 3 to April 30, 2025. Districts can schedule the administration of the test sections at any time during the testing window, but all sections must be completed within the window and in sequential order (unless a student misses a section due to illness, etc.). Each student is expected to complete a section in a single sitting, with limited exceptions due to illness during testing. The ISAs are administered online each spring, with paper accommodated forms available as needed. The online administration takes place in TestNav, Pearson's online testing platform. ADAM (Assessment Delivery and Management) is the student test management portal that Test Administrators use to manage student tests and registrations and order materials if needed. Table 4.1 presents the roles and responsibilities of the individuals involved with the administration of the assessments.

Table 4.1. Test Administration Roles and Responsibilities

Role	Description
District Test Coordinator (DTC)	The individual at the district level who is responsible for the overall coordination of test administration
School Test Coordinator (STC)	The individual at the school level who is responsible for the overall coordination of test administration (i.e., the principal or designee based on established criteria), including coordinating test administration, resolving testing issues at their school, and overseeing all post testing procedures (e.g., stopping all online test sessions).
Test Administrator	An individual at the school who is responsible for administering the assessment (e.g., individuals employed by the district as teachers, district- and school-level administrators, other certified educational professionals)
Proctor	An individual who may be called on to help a Test Administrator monitor a testing session under the supervision of the Test Administrator

4.1. Administration Manuals

To ensure a standardized administration for all students, School Test Coordinators and Test Administrators are instructed to follow the directions in the *Test Coordinator Manual* and *Test Administrator Manual* available online at https://il.mypearsonsupport.com/isa-summative-resources/ The *Test Coordinator Manual* provides instructions for the District and School Test Coordinators and Technology Coordinators, while the *Test Administration Manual* provides instructions for the Test Administrators, including the scripts they must follow during the test administration.

The standardization of directions, test administration conditions, and scoring procedures is necessary to support the comparability of test score interpretations both within and between administrations. When standardized procedures are not in place, differences in student performance cannot be clearly attributed to true differences in student ability because of the unknown effect of administration conditions on performance.

4.2. Administration Training

Administration training is intended to ensure that all individuals at the school and district levels involved in the ISA administration activities are well prepared to follow specific administration processes and procedures with confidence and fidelity, as well as support coherence to security procedures. Confidence in certain test conditions and fidelity to standardized test administration procedures help to ensure the scores are comparable and the interpretation of results accurate. Virtual test administrator and technology coordinator training were offered in January/February 2025 prior to the spring administration.

4.3. Practice Items

Students can find ISA practice test items online at https://il.mypearsonsupport.com/practice-tests/science/ to get familiar with the testing platform and types of questions they'll see during the actual assessment. Additional practice items on science content are also available through the QuISBE platform; these released summative items can be used by teachers in class as needed. However, these sets of practice tests do not cover every aligned content standard. Although students will receive scores on both platforms, these scores should not be used to evaluate a student's performance level. Students access the practice tests as guests, so no personal information is required.

4.4. Accessibility Features and Accommodations

The *ISA Accessibility Features and Accommodations Manual* provides guidance for ensuring that the ISAs provide valid results for all students. The ISA Accessibility and Accommodations manual relied on the manual developed for the Illinois Assessment of Readiness, as the accommodation rules are consistent across different testing subjects. It is important to ensure that performance on the ISAs is influenced minimally, if at all, by a student's disability or linguistic/cultural characteristics that may be unrelated to the content being assessed. Through a combination of Universal Design principles and accessibility features, accessibility was considered from the initial test design through item development, field testing, and implementation of the assessments for all students. As such, the ISAs include three levels of support for students: (a) features for all students, (b) accessibility features available to all participating students, and (c) accommodations for students with disabilities, ELs, and ELs with disabilities.

The accessibility features available to students should minimize the need for accommodations during testing and ensure the inclusive, accessible, and fair testing of the diverse students being assessed, but accommodations may still be needed for some SWDs and ELs to assist in demonstrating what they know and can do. While all students can receive accessibility features on the summative assessments, the following four distinct groups of students may receive accommodations on the ISA. Testing accommodations for SWDs or students who are ELs must be documented according to the guidelines and requirements outlined in the *ISA Accessibility Features and Accommodations Manual*.

- Students with disabilities who have an Individualized Education Program (IEP)
- Students with a Section 504 plan who have physical or mental impairment that substantially limits one or more major life activities, have a record of such an impairment, or are regarded as having such an impairment, but who do not qualify for special education services
- Students who are ELs
- Students who are ELs with disabilities who have an IEP or 504 plan

Accessibility features are tools or preferences that are either built into the assessment system or provided externally by Test Administrators. Accessibility features can be used by any student taking the ISA. Because the accessibility features are intended for all students, they are not classified as accommodations. A small selection of accessibility features available to all students needs to be identified in advance. Examples of accessibility features in TestNav include the line reader, answer eliminator, magnifier, highlighter, bookmark, pop-up glossary, and notepad. Students should have the opportunity to select and practice using these features prior to testing to determine which are appropriate for the assessment. Consideration should be given to the supports that a student finds helpful and consistently uses during instruction.

Accommodations are adjustments to the testing conditions, test format, or test administration that provide equitable access during assessments for SWDs and EL students. In general, the administration of the assessment should not be the first occasion on which an accommodation is introduced to the student. To the extent possible, accommodations should (a) provide equitable access during instruction and assessments, (b) mitigate the effects of a student's disability, (c) not reduce learning or performance expectations, (d) not change the construct being assessed, and (e) not compromise the integrity or validity of the assessment.

Accommodations are intended to reduce or eliminate the effects of a student's disability and/or English language proficiency level, but they should never reduce learning expectations by reducing the scope, complexity, or rigor of an assessment. Accommodations must also be consistent with those provided for classroom instruction and classroom assessments. Some accommodations may be used for instruction and for formative assessments that are not allowed for the summative assessment because they impact the validity of the assessment results (e.g., allowing a student to use a thesaurus or access the internet during an assessment). There may be consequences (e.g., excluding a student's test score) for the use of nonallowable accommodations during assessments. To the extent possible, accommodations should adhere to the following principles:

- Accommodations should enable students to participate more fully and fairly in instruction and assessments and to demonstrate their knowledge and skills.
- Accommodations should be based on an individual student's needs rather than on the category of a student's disability, level of English language proficiency alone, level of or access to grade-level instruction, amount of time spent in a general classroom, current program setting, or availability of staff.
- Accommodations should be based on a documented need in the instruction/assessment setting and should not be provided for the purpose of giving the student an enhancement that could be viewed as an unfair advantage.
- Accommodations for SWDs must be described and documented in the student's IEP or 504 plan and must be provided if they are listed.
- Accommodations for ELs should be described and documented.
- EL students with disabilities are eligible to receive accommodations for both SWDs and ELs.
- Accommodations should become part of the student's program of daily instruction as soon as possible after completion and approval of the appropriate plan.
- Accommodations should not be introduced for the first time during the testing of a student.
- Accommodations should be monitored for effectiveness.
- Accommodations used for instruction should also be used, if allowable, on local district assessments and state assessments.

Examples of accommodations provided for the ISA include assistive technology, screen reader version for a student who is blind or visually impaired, a braille edition, large print edition, a paper-based edition, American Sign Language (ASL) video, human signer for test directions, and a word-to-word dictionary for ELs. Please refer to the *ISA Accessibility Features and Accommodations Manual* for the full list of accommodations for students with disabilities and EL students. If a student refuses an accommodation listed in their IEP, 504 plan, or an EL plan, the school must document in writing that the student refused the accommodation. However, the accommodation must be offered and remain available to the student during the test administration.

4.5. Test Security

The ISA test administration is a secure testing event, and maintaining the security of test materials before, during, and after the test administration is crucial to obtaining valid and reliable results. The test security and administration policies are found in the *Test Coordinator Manual* and the *Test Administrator Manual* available online at https://il.mypearsonsupport.com/training-resources-sci/, including the responsibilities of the School Test Coordinators for ensuring that all personnel with authorized access to secure materials are trained in and subsequently act in accordance with all security requirements, as well as the responsibilities of the Test Administrator for maintaining test security before, during, and after the test administration.

Both administration manuals also provide information and directions regarding instances of testing irregularities and security breaches, as exemplified in Table 4.2, that must be reported to the School Test Coordinator immediately. The *Form to Report a Testing Irregularity or Security Breach* must be completed within two school days of the incident.

Table 4.2. Test Irregularity and Security Breach Examples

Topic	Examples
Electronic Devices	Using a cell phone or other prohibited handheld electronic device (e.g., smartphone, iPod, smart watch, personal scanner, eReader) while secure test materials are still distributed, while students are testing, after a student turns in their test materials, or during a break *Exception*: School Test Coordinators, Technology Coordinators, Test Administrators, and Proctors can use cell phones in the testing environment only in cases of emergencies or when timely administration assistance is needed.

Topic	Examples
Test Supervision	• Coaching students during testing, including giving students verbal or nonverbal cues, hints, suggestions, or paraphrasing or defining any part of the test • Engaging in activities (e.g., grading papers, reading a book, newspaper, or magazine) that prevent proper student supervision at all times while secure test materials are still distributed or while students are testing • Leaving students unattended without a Test Administrator for any period of time while secure test materials are still distributed or while students are testing (Proctors must be supervised by a Test Administrator at all times) • Allowing cheating of any kind • Providing unauthorized persons with access to secure materials • Administering a computer-based test in ADAM during non-testing times without state approval • Failing to provide a student with a documented accommodation or providing a student with an accommodation that is not documented and therefore is not appropriate • Allowing students to test before or after the test administration window without state approval
Test Materials	• Losing a student test booklet or human reader scripts • Leaving test materials unattended or failing to keep test materials secure at all times • Reading or viewing the test items before, during, or after testing • Copying or reproducing (e.g., taking a picture of) any part of the items or any secure test materials or online test forms • Revealing or discussing test items with anyone, including students and school staff, through verbal exchange, email, social media, or any other form of communication • Removing secure test materials from the school building or removing them from locked storage for any purpose other than administering the test
Testing Environment	• Failing to follow administration directions exactly as specified in the *Test Administrator Manual* • Displaying any resource (e.g., posters, models, displays, teaching aids) that defines, explains, illustrates terminology or concepts, or otherwise provides unauthorized assistance during testing • Allowing preventable disruptions such as talking, making noises, or excessive student movement around the classroom • Allowing unauthorized visitors in the testing environment

In addition to test security procedures required of all educators involved in the testing process, TestNav has built-in security features for the test content and personal data that relies on multiple levels of protection, including restricted user access, encryption of data in transit and at rest, systems monitoring for abnormal behavior, application, server, and network security testing, and qualified, verified and trusted support personnel.

Pearson uses Advanced Encryption Standard (AES) encryption for data at rest and Hypertext Transfer Protocol Secure (HTTPS) to provide encryption and data-in-motion security for online testing by creating a secure channel on the network with the Secure Socket Layer (SSL) /Transport Layer Security (TLS) protocols. Test content can only be viewed through a valid test registration and login, all of which are logged within the platform's audit trail system and cannot be deleted.

TestNav also locks down the student's desktop during testing to prevent students from accessing outside resources that could be used for cheating, such as email, instant messaging, or internet browsing. TestNav will stop students' tests if another background application attempts to interfere with or take "focus" away from the secure testing environment. These types of interruption cannot be blocked during testing and therefore could present additional opportunities for students to access unauthorized resources. However, TestNav also has a blocklist feature that prevents students from starting their test if certain applications that pose a threat to disrupt testing are running at the time TestNav is launched. In these situations, the student and/or proctor are prompted to shut down the offending application before attempting to start TestNav again.

Section 5: Scoring

Selected-response, technology-enabled, and technology-enhanced items are machine scored; constructed-response items are human scored using Pearson's scoring platform, OSCAR (Online Scoring and Reporting); and the Science CR items are primarily scored by Pearson's automated scoring engine known as the Intelligent Essay Assessor (IEA), with a 10% reliability score and some outlier scoring (where the IEA score and human score differ by more than 1 point) by human scorers (i.e., 10% of the CR item scores are also scored by humans in addition to IEA to compute the inter-rater agreement and monitor scoring).

5.1. Machine Scoring

Pearson performed a key check and adjudication near the end of the test administration and before reporting to verify that the answer keys were correct for each item. The keycheck process is a quality assurance step to ensure that multiple-choice or multi-select items are scoring correctly. This process involves a review of item statistics to confirm that the designated correct answer(s) are being awarded full credit and that the scoring logic is functioning as intended. Reviewers check the scoring data to verify that all correct responses receive full credit, and that incorrect or partially correct responses are scored appropriately according to the established rules.

If a test map contains an incorrect key, such as a misaligned correct answer or scoring rule, this issue is typically flagged during the statistical keycheck process, which uses item-level metrics and response patterns to detect anomalies. Once identified, the item is escalated for content review to confirm the error and determine its source. The test map is then corrected and re-published, ensuring alignment with scoring rules and platform logic. A re-analysis is conducted to validate the fix, and downstream systems are updated accordingly. If the correction affects scoring logic or item metadata, the Change Control Board (CCB) oversees the implementation to prevent disruptions. The corrected item is re-exported, rescored, and verified by psychometrics and content teams before finalization. This process ensures scoring integrity and prevents propagation of errors across systems.

The adjudication process is a specialized workflow designed to ensure the accuracy and fairness of scoring for technology-enhanced and free-response items, distinct from the standard key check. It provides a structured approach for addressing discrepancies and maintaining scoring integrity.

The adjudication workflow begins with the preparation phase. Relevant scoring materials—including item files, answer keys, scoring rubrics, and student responses—are gathered and made accessible within secure scoring platforms such as OSCAR. All reviewers involved in the adjudication process are granted appropriate access to these materials to facilitate a thorough review.

The initial review is conducted by a trained content specialist or designated scorer. This reviewer examines flagged items or student responses where scoring discrepancies have been identified, either through automated scoring algorithms or manual checks. The first reviewer documents the nature of the discrepancy, marks the issue in the scoring system, and provides an initial assessment or recommended action.

If the first reviewer is unable to resolve the discrepancy or if there is disagreement regarding the scoring decision, the issue is escalated to a senior content lead or adjudication manager. The second reviewer performs a detailed analysis of the flagged item or response, referencing scoring guidelines and consulting with subject matter experts as needed. This step ensures that all perspectives are considered and that the scoring aligns with established criteria.

For issues that remain unresolved after the second review, a third adjudication is conducted by the Test Development Manager (TDM) or a designated member of the Content Committee Board (CCB). This final review involves a comprehensive evaluation of all supporting documentation, prior reviewer notes, and scoring rubrics. The TDM or CCB approves the final scoring decision and oversees the implementation of any necessary changes to the scoring rules or answer keys.

After adjudication decisions are finalized, a series of post-adjudication checks are performed. These include rescoring affected responses, conducting a secondary review of item spreadsheets, and, if required, replacing or updating items within the scoring system. These steps confirm that adjudication outcomes have been correctly applied and that scoring reliability is maintained.

Throughout the adjudication workflow, all actions, findings, and decisions are logged in secure tracking systems. Clear communication is maintained among reviewers, scoring managers, and project stakeholders to ensure transparency. A summary of adjudication outcomes is provided to relevant parties, and procedural updates are incorporated into future scoring guidelines.

If discrepancies were identified during the adjudication process, a Pearson senior content specialist or content manager reviewed the flagged item(s) and worked to resolve the issue. Rule-based scoring refers to item types that use various scoring models, including choice interaction that presents a set of choices where one or more choices can be selected; text entry, where the response is entered in a text box; hot spot or text interaction, where an area in a graph or text in a paragraph can be highlighted; or match interaction, where an association can be made between pairs of choices in a set. These items include the scoring rules and correct responses as part of their item XML (markup language) coding. Following the initial development of the rule-based scoring rubrics, Pearson has continued to monitor and evaluate new item development to ensure that the scoring rules are maintained within all item types as approved.

5.2. Human Scoring of Constructed-Response Items

Constructed-response items were handscored by human scorers who completed online training and qualification sets to demonstrate they could correctly score student responses based on the provided guidelines. Scorers who successfully completed the training and qualifying process were permitted to score student responses. All CBT and PBT responses were scored within the OSCAR system with monitoring conducted by Pearson. A handscoring specifications document detailed the handscoring schedule, customer requirements, quality management plans, item information, and staffing plans for each scoring administration. All Pearson employees involved in the scoring process possessed at least a four-year college degree. Roles and responsibilities were as follows:

- Scorers applied scores to student responses.
- Scoring supervisors monitored the work of a team of scorers through review of scorer statistics and backreading.
- Scoring directors managed the scoring quality of a subset of items and monitored the work of supervisors and scorers for their assigned items. Directors backread responses scored by supervisors and scorers as part of their quality-monitoring duties.
- Science content specialist managed the scoring quality and monitored the work of the directors.
- The project manager documented the procedures, identified risks, and managed day-to-day administrative matters.

5.2.1. OSCAR Scoring System

All human scoring of student responses was conducted through Pearson's scoring platform, OSCAR, that seamlessly integrates with Pearson's IEA automated scoring engine via Pearson's Continuous Flow platform. More than just a scoring system, OSCAR combines multiple processes—routing work, scoring responses, monitoring quality, providing feedback to scorers, and tracking progress and workflow—into a lean, streamlined design. Rather than pushing work into a scorer's queue, where work might sit idle if the scorer is away or has left the program, OSCAR allows each scorer to pull in work as needed. Automated functions like validity dispersion (i.e., the rate at which validity responses are dispersed to scores along with operational responses) and feedback enhance quality management and allow scoring leadership to identify problems before they affect scoring performance. Table 5.1 summarizes the important features and benefits of OSCAR. OSCAR facilitates early detection of issues, offers real-time feedback to scorers, and, ultimately, promotes accurate scoring for Illinois.

Table 5.1. Important Features and Benefits of OSCAR

Features	Function	Benefits
Scoring Broker	Student responses are routed for human, automated, and outlier scoring based on pre-defined parameters.	Human and automated scoring are seamlessly integrated, optimizing scoring performance and project completion.
Automatic Routing	Student responses are immediately routed to human scorer work queues based on project requirements.	Responses to each item are scored only by those scorers qualified for that particular item or task.
Automated Quality Monitoring Processes	Quality monitoring processes like validity distribution, or calibration are automated, random, and blind.	Blind, automated processes encourage consistency in scoring and ensure that the quality monitoring statistics are a true reflection of scorer accuracy.
Dynamic Feedback Tools	Supervisory staff use integrated tools to review and provide feedback on scorers' work.	Scoring experts continually monitor the accuracy of scoring on group and individual levels, intervening with scorers who need focused coaching and remediation.
Scalable Infrastructure	The cloud-based scoring platform can scale up and down to accommodate any program's scoring requirements.	The system has flexibility to handle high scoring volumes or quickly increase capacity to meet the program's schedule and scoring window requirements.
Scorer-Friendly User Interface	Intuitive navigation features allow scorers to easily scroll, zoom, quickly access the applicable rubric and anchor papers, and assign scores.	The easy-to-use scoring user interface optimizes scoring efficiency and accuracy, resulting in higher scorer productivity and lower costs.
Integrated Security Measures	Users' tasks are restricted by role, limiting exposure and preventing score manipulation.	Test materials and student data are protected throughout the scoring process.

5.2.2. Scorer Training and Qualification

The responses were provided by the customer and partial training materials were provided by ISBE, which were then further developed into full operational materials by the Pearson handscoring content specialist. When developing the scorer training materials, Pearson reviewed the prompt and rubric and field test training materials which were then supplemented with "live responses" from the current year to create complete training sets: an annotated anchor, two annotated practice sets and two qualification sets. During scorer training, Pearson used anchor, practice, and qualification sets, as described in Table 5.2. The anchor and practice sets included annotations for each student response (i.e., formal written explanations of the score).

To demonstrate that they could accurately apply the scoring methodology, scorers applied scores to two qualification sets consisting of 10 responses each and were required to match the approved score at a certain percentage to qualify. On at least one of the two qualifying sets for qualification, scorers must agree with at least 70% of the approved scores and attain at least 90% exact plus adjacent agreement. Scorers were required to complete both qualification sets.

Table 5.2. Scoring Training Materials

Training Material	Description	Specifications
Anchor Sets	Anchor sets consist of responses that are clear examples of student performance at each score point and are the primary reference for scorers as they internalize the rubric. The responses selected are representative of typical approaches to the task and are arranged to reflect a continuum of performance. All scorers have access to the anchor set when they are training and scoring and are directed to refer to it regularly.	3 annotated responses per score point
Practice Sets	Practice sets are used to help scorers practice applying the scoring guidelines. Scorers review the anchor sets, score the practice sets, and then compare their assigned scores for the practice sets to the actual assigned scores to help them learn. Some of these responses clearly reinforce the scoring guidelines presented in the anchor set, whereas others are more difficult to evaluate, fall near the boundary between two score categories, or represent unusual approaches to the task to provide guidance and practice in defining the line between score categories and applying the scoring criteria to a wider range of response types.	2 sets of 10 annotated responses
Qualification Sets	Qualification sets consist of student responses that are clear examples of score points to reinforce the application of the scoring criteria illustrated in the anchor set. These sets are used to confirm that scorers understand how to score the responses accurately. Scorers are required to meet specified agreement percentages on qualification sets to score student responses.	2 sets of 10 responses each (not annotated)

5.2.3. Scorer Monitoring

Score monitoring consisted of second scoring of at least 10% of the responses, backreading, the use of validity responses and calibration sets, and inter-rater reliability.

5.2.3.1. Second Scoring

During scoring, all online CR written responses are initially evaluated by the IEA engine, with outlier responses—those flagged as atypical or requiring further review—being human scored for the second read. In addition, the OSCAR scoring system automatically and randomly distributed a minimum of 20% of such student responses for second scoring, ensuring reliability through additional human review. Scorers had no indication whether a response had been previously scored. Furthermore, all reliability scores reflect human scoring. Notably, all Spanish responses are first and second scored by human raters. If the first and second human scores are nonadjacent, a third and occasionally fourth score are assigned to resolve scorer disagreements. When a resolution score (i.e., third score) is nonadjacent to one or both of the first two scores, the content specialist or scoring director applies an adjudication score (fourth score).

5.2.3.2. Backreading

Backreading required the scoring supervisor to review the scores applied by scorers to help them provide additional coaching or instruction and guard against scorer drift, where scorers score responses in comparison to one another instead of in comparison to the training responses. Scoring supervisors used the OSCAR backreading tool to review scores assigned to individual student responses by any given scorer to confirm that the scores were correctly assigned and to give feedback and remediation to individual scorers. Pearson backread approximately 5% of the handscored responses.

5.2.3.3. Validity Responses

Prescored validity responses were strategically interspersed in the pool of live responses and indistinguishable from any other responses so that scorers were unaware they were scoring validity responses rather than live responses to help ensure that scorers were applying the same standards throughout the project. Scorers had to meet the required validity agreement requirements in Table 5.3 to continue working on the project. Scorers who did not maintain the expected agreement statistics were given a series of interventions culminating in a targeted calibration set. Scorers who did not pass targeted calibration were removed from scoring the item, and all the scores they assigned were deleted.

Table 5.3. Scoring Validity Agreement Requirements

Grade	Score Point Range	Perfect Agreement	Within 1 Point*
5	0–3	70%	96%
8	0–3	70%	96%
11	0–3	70%	96%

*A numerical score compared to a blank or condition code score will have a disagreement greater than 1 point

In addition to the prescored validity responses, validity was at times shared with scorers in a process known as “validity as review” that provided scorers automated, immediate feedback, giving them a chance to review responses they mis-scored, with reference to the correct score and a brief explanation of that score. One validity response was sent to scorers for every 25 “live” responses scored. Selected validity responses are annotated by the scoring director and flagged for review. If a scorer incorrectly scores one of these responses, the paper will immediately appear on their screen with the true score, the score they assigned, and an annotation. This immediate feedback aids in preventing scorer drift before it occurs. Once a scorer has received feedback about a specific validity response, the response is flagged so the scorer does not receive it again. A scorer's validity agreement is recomputed after every validity response. For every 100 student responses scored, a scorer will also score 4 validity responses (i.e., there is a scoring validity check at a 4% interval). Validity responses are indistinguishable from “live” student responses.

5.2.3.4. Calibration Sets

Calibration sets were created by scoring directors to reinforce rangefinding standards, introduce scoring decisions, or address scoring issues and trends to help train scorers on areas of concern or focus. Calibration was used either to correct a scoring issue or trend or to continue scorer training by introducing a scoring decision. Calibration was administered regularly throughout scoring.

5.3. Automated Scoring

Pearson's IEA automated scoring engine is trained by field tested human-scored student responses. In some cases that do have sufficient responses to cover training at all score points, the field test items are supplemented with operational responses and human scores via Pearson's Continuous Flow scoring process that is used to improve and validate the scoring models. With Continuous Flow, responses flow between the engine and human scorers so the engine can learn from humans in real time. Once IEA obtains sufficient data to train or complete a scoring model (all score points can be scored), it can be used as the primary source of scoring. All sampling used for automated scoring model development is conducted using simple random sampling to ensure unbiased representation of the scored responses.

For each ISA constructed-response item and trait, responses used to train and evaluate the IEA models are selected using simple random sampling from the pool of available human-scored field test responses. The sampled responses are partitioned into training, validation, and test datasets to support model development and independent evaluation. During model development, Pearson applies text preprocessing and extracts large sets of features that reflect the scoring rubric, including writing-quality indicators such as mechanics, grammar, and organization, along with content-based semantic features aligned with the expectations of the item. Multiple supervised machine learning models are trained for each trait, and the model that shows the strongest performance on the validation and holdout sets is selected for operational use. Pearson also employs a Continuous Flow process whereby additional operational responses may be incorporated to update the models when appropriate, with all updates conducted under human oversight to maintain scoring consistency and quality.

When the engine is less confident in scoring a response, the response is marked with a low confidence flag that automatically routes it to human scorers (known as Smart Routing), and the IEA score is overridden. Smart Routing refers to the practice of using automated scoring results to detect responses that are likely to be challenging to score and applying automated routing rules to obtain one or more additional human scores. Low confidence occurs when responses produce feature measures that are statistical outliers or do not result in numerical IEA scores. In these cases, the model can still generate a score, but the modeling team does not have full confidence in it. Several factors can trigger a low confidence flag, most commonly in split-score models where certain categories are not scored due to limitations in pretraining, requiring human raters instead. Customer-defined rules may also dictate which aspects can and cannot be scored with high confidence. Low confidence flags may also appear in cases of cross-trait extreme scores (where a student receives both extremely high and low scores across multiple traits) or high-end boundary scores (where a score exceeds the upper boundary of the range).

IEA is used to score student responses to many different types of prompts, including argumentative, informational, narrative, and content-based prompts, in a variety of subject areas. Like human scorers, IEA evaluates content, grammar, style, and mechanics to score essays as well as short responses. IEA learns to score using a range of machine learning and natural language processing technologies. The engine is trained individually on each trait of each prompt based on hundreds to thousands of human-scored student responses, enabling IEA and human scorers to score alike. One of the hallmarks of IEA is its ability to score constructed responses in content areas beyond just ELA using Pearson's unique

implementation of Latent Semantic Analysis (LSA) that analyzes large bodies of relevant text to generate semantic similarity of words and passages. LSA can then "understand" the meaning of text much the same as a human scorer.

IEA's background knowledge of English is based on a collection of text of about 12 million words—roughly the amount of text a student will read over the course of their academic career. Because LSA operates over the semantic representation of texts, rather than at the individual word level, it can evaluate similarity even when texts have few or no words in common. For example, LSA finds the following two sentences to have a high semantic similarity:

- *Surgery is often performed by a team of doctors.*
- *On many occasions, several physicians are involved in an operation.*

IEA is trained to associate different aspects of student responses to scores assigned by human scorers. A machine learning-based approach determines the optimal set of features, and the weights for each of those features, to best model the scores for each essay. From these comparisons, a prompt- and trait-specific scoring model is derived to predict the scores the human scorers would assign to any new responses.

As evidence of Pearson's strong research base in automated scoring, Pearson has conducted wide-ranging studies for the PARCC consortium (Lochbaum et al., 2015; Way et al., 2016), including (a) a proof-of-concept study in 2014 using field test data showed encouraging results supporting the use of IEA as a 10% second score quality check on human scoring during the first operational assessment; (b) an evaluation of IEA's operational performance during the spring 2015 operational year showing that the automated scoring engine agreed with human scorers better than they agreed with each other and that IEA's exact agreement on validity responses was also higher than human exact agreement performance on those responses; and (c) a study to validate the use of Pearson's Continuous Flow approach to scoring during operational scoring starting in 2016 in which the combined human and automated scoring systems achieved over 72% exact agreement both years, far exceeding the 65% requirement.

Furthermore, Pearson understands the importance of ensuring that all students receive accurate scores, regardless of their gender, race, ethnicity, or other identity marker, and regardless of whether scores are assigned by human scorers or IEA. Human scorers are trained on unconscious bias and how to avoid bias during scoring. IEA learns from human scorers, who must successfully complete bias training prior to assigning scores, helping to ensure that bias is not introduced into the automated scoring engine. To validate the lack of bias, Pearson has conducted several studies looking at demographic subgroup performance. The automated scoring system receives only the response text and the associated human scores and does not receive additional student-level information unless specifically provided for the purposes of subgroup evaluation. As a result, the analyses presented in this report focus on evaluating the performance of the automated scoring models within the ISA scoring design and the available data. Evaluating alternative prompt designs or conducting specialized studies of subgroup performance would require additional targeted data collection and constitute separate research efforts beyond the scope of this technical report.

5.4. Inter-Rater Agreement

For 10% of all responses, a second reliability score was assigned to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. Inter-rater agreement is the agreement between the first and second scores assigned to student responses. Pearson used inter-rater agreement statistics as one factor in determining the needs for continuing training and intervention on

both individual and group levels. During handscoring, the OSCAR system included comprehensive inter-rater agreement reports that allowed supervisory personnel to monitor both individual and group performance. Based on reviews of these reports, scoring experts targeted individuals for increased backreading and feedback, and if necessary, retraining. Inter-rater agreement was also calculated for the items scored by IEA.

Automated scoring performance for ISA is evaluated using the same validity framework applied to human scoring. Following guidance from Williamson, Xi, and Breyer (2012), agreement between human and automated scores is summarized using quadratic weighted kappa (QWK), Pearson correlation, and standardized mean difference (SMD) between score distributions. For automated scoring models to be considered operationally comparable to human scoring, QWK values are expected to meet or exceed 0.70, to fall within approximately 0.10 of the corresponding human–human QWK values, and SMD values are expected to remain below 0.15 in absolute value. Percent exact agreement is also provided as a descriptive indicator of agreement but is treated as an ancillary metric rather than a primary indicator of scoring validity. This framework ensures that the ISA automated scoring models are evaluated against established standards that appropriately reflect the ordinal nature of the scoring scales and the expectations for operational comparability with human scoring. The agreement statistics presented in Table 5.5 therefore demonstrate that the ISA automated scoring models meet the expectations for operational comparability with human scoring when evaluated against these industry-standard criteria.

Table 5.4 presents the inter-rater agreement expectations and results for the CR items from the spring 2025 administration across all grades based on human scoring, and Table 5.5 presents the common agreement statistics across items for each grade, including the number of items included in the analyses, perfect agreement, kappa, quadratic weighted kappa (QWK), and Pearson correlation (*r*).

Table 5.4. Inter-Rater Agreement Expectations and Spring 2025 Results

Grade	#Items	Score Point Range	Perfect Agreement Expectation	Perfect Agreement 2025 Result	Within 1 Point Expectation*	Within 1 Point 2025 Result
5	4	0–3	70%	4 or 100%	96%	100%
8	3	0–3	70%	3 or 100%	96%	100%

*A numerical score compared to a blank or condition code score will have a disagreement greater than 1 point.

Table 5.5. Inter-Rater Agreement Spring 2025 Results

Grade	#Items	Score Point Range	Avg. Kappa	Avg. QWK	Avg. r
5	4	0-3	0.71	0.90	0.90
8	3	0-3	0.70	0.88	0.88

5.5. Hierarchy of Assigned Scores for Reporting

When multiple scores are assigned for a given response, the hierarchy rules in Table 5.6 determined which score was reported as the final operational score.

Table 5.6. Scoring Hierarchy Rules

Score Type	Rank	Final Score Calculation
Adjudication (fourth score)	1	If an adjudication score is assigned, this is the final score.
Resolution (third score)	2	If no adjudication score is assigned, this is the final score.
Backreading score	3	If no adjudication or resolution score is assigned, the latest backreading score is the final score.
Human first score	4	If no adjudication, resolution, or backreading score is assigned, this is the final score.
Human second score	5	If no adjudication, resolution, backreading, or human first score is assigned, this is the final score.

Section 6: Reporting

Student performance is reported on the Individual Student Report (ISR) using overall scale scores with associated performance levels and domain-level scale scores, as described in the ISA Score Interpretation Guide available online at https://il.mypearsonsupport.com/isa-summative-resources/ . The ISA student results are expressed as an overall scale score ranging from 700 to 900, along with associated performance levels to describe how well students met the academic standards for their grade level. Domain-level scale scores ranging from 300 to 500 are also provided for each of the three domains: Physical Science, Life Science, and Earth/Space Science.

Not all students respond to the same set of test items, so each student's raw score (actual points earned on the test) is converted onto a common scale to account for the differences in difficulty among the various forms and administrations of the test. The resulting scale score allows for an accurate comparison across test forms and administration years within a grade or course. For example, a student who receives a raw score of 55 on one form, meaning they answered 50 points correctly, might receive a scale score of 800. This scale score can then be compared to a different test form where a raw score of 60 translates into a scale score of 800. The scale scores, not the raw scores, reflect the same ability and knowledge levels.

Based on a student's overall scale score, an inference is drawn about how much knowledge and skill in science the student has acquired. The overall scale scores also determine a student's performance level that classifies a student's competency based on their test performance as reflected by their test results, as provided below. Each performance level is defined by a range of overall scale scores for the assessment established during the standard setting (see Section 7 for more details). Students performing at Levels 3 and 4 are proficient or above proficient and have demonstrated readiness for the next grade level. The full PLDs are available online at https://www.isbe.net/Pages/Performance-Level-Descriptors.aspx.

- Level 4: *Above Proficient*
- Level 3: *Proficient*
- Level 2: *Approaching Proficient*
- Level 1: *Below Proficient*

Section 7: Standard Setting

Student performance on state assessments is required to be reported in performance levels that reflect how well students demonstrate the knowledge and skills expected at their grade level, as defined by a state's content standards. These performance levels are defined by cut scores, or points on the test's score scale (i.e., the full range of possible scores) that mark the minimum score a student must earn to be classified into a given performance level. During a standard setting meeting, committees of educators review test items and apply their content expertise and professional judgment to recommend these cut scores. This section summarizes the 2025 standard setting that took place from July 14–18, 2025, in Springfield, Illinois, to establish the most recent cut scores for the ACT, IAR, and ISA assessments, with the full details about the process provided in the standard setting report (Gardner, T. & Moore, J, 2025).

Illinois transitioned in 2025 to the ACT as the high school accountability assessment for ELA/L, mathematics, and science. In response to findings that Illinois' previous cut scores were among the highest in the nation, Illinois took advantage of the opportunity presented by the shift in high school assessment to unify the names, number, and definition of the performance levels across the ACT, IAR, and ISA assessments. This required a standard setting to recommend new cut scores that divide each assessment score scale into four levels: *Above Proficient*, *Proficient*, *Approaching Proficient*, and *Below Proficient*. Despite differences in test design, this unified approach aimed to maintain high expectations while better reflecting college and career readiness and establishing a unified, coherent reporting structure across the Illinois assessment system.

The standard setting used spring 2025 data and followed the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake et al., 2005) for all assessments in addition to the Modified Briefing Book approach (Camara et al., 2017; Haertel et al., 2012) for the ACT to establish the empirical link with college and career readiness benchmark data. Seventeen committees with 12 panelists each and two grade 4 committees with 18 panelists each for the grade 4 ELA and mathematics IAR assessments were convened, for a total of 155[3] panelists covering 240 slots (14 subject and grade-level committees for IAR and ISA and five committees for each ACT subject: English, reading, writing, mathematics, and science). A vertical articulation committee was then convened on the last day with 34 participants to ensure an appropriate progression of cuts across grades and subject areas. See the standard setting report for full details of the process (Gardner, T. & Moore, J, 2025)

7.1. Standard Setting Process

Each standard setting committee recommended three cut scores to divide the score scale into the four performance levels: the *Approaching Proficient* cut (between *Below Proficient* and *Approaching Proficient*), the *Proficient* cut (between *Approaching Proficient* and *Proficient*), and the *Above Proficient* cut (between *Proficient* and *Above Proficient*). Following the Extended Modified Yes/No Angoff method, panelists reviewed each item on one form of the spring 2025 operational assessment in test administration order and judged whether most borderline students (i.e., students scoring near the lower end of a performance level) would likely answer the item correctly (for single-point items) or how many points they would likely earn (for multi-point items). Because constructed-response items have a large impact on the IAR and ISA test scores, these judgments were supplemented by student score profiles illustrating how actual students performed across the score scale to ground panelists' decisions in actual student performance.

[3] Final participation was 147 of the 155 panelists recruited.

While the IAR and ISA are designed to measure mastery of the grade-level Illinois Learning Standards, the ACT has an additional purpose to predict college readiness. Therefore, in addition to the item-level Extended Modified Yes/No Angoff content judgments, ACT panelists received briefing books containing empirical data linking ACT scores to real-world student outcomes, such as the likelihood of earning a B or C in first-year college courses, high school GPA, and college enrollment rates, as well as performance on other assessments (e.g., IAR, ISA, NAEP). This information enabled panelists to interpret ACT scores in the context of college and career readiness and ensure that the recommended cut scores reflect meaningful readiness benchmarks in addition to content mastery.

Each standard setting meeting began with an overview of the purpose, test design, panelist role, and orientation to the materials, including the standard setting tool where the item judgments were made. Panelists first "experienced the assessment" by taking the spring 2025 operational assessment, followed by discussing the PLDs and exploring examples of performance that distinguishes students just entering the performance level from the full range of expected performance in the band. After additional system training and a practice activity, panelists confirmed their readiness before beginning the item-level judgments. Three rounds were conducted for IAR and ISA:

1. Round 1: Panelists reviewed each item on the test form and answered one of the following judgment questions for each performance level, beginning with the *Proficient* cut:
    a. Single-point items: *"Considering a variety of students at the lower end of the performance level, would most students get this item correct?"* Panelists answered "yes" or "no."
    b. Multi-point items: "*Considering a variety of students, which score point most likely represents the most common response for students at the lower end of this performance level?*" Panelists chose between 0–6 score points depending on the item type and maximum score.
2. Round 2: Panelists received the score profiles and discussed the Round 1 results before revising the initial judgments as needed following the same steps as Round 1.
3. Round 3: Final revisions were made after discussing the Round 2 results, including the impact data that showed the percentage of students who would be classified into each performance level based on the recommended cuts and performance of students on the spring 2025 assessment.

Four rounds were conducted for the ACT assessments using a hybrid approach:

1. Round 1 (Extended Yes/No Modified Angoff): Panelists followed the same judgment process as IAR/ISA.
2. Round 2 (Extended Yes/No Modified Angoff): Panelists discussed the Round 1 results, followed by a discussion of the impact and outcome data. Equipped with both the Round 1 results and an understanding of the relationships between performance on the ACT and performance in first-year college courses, panelists revised their initial judgments following the same steps as Round 1.
3. Round 3 (Modified Briefing Book): The English, reading, and writing committees were combined to form a 12-panelist ACT ELA committee to support coherence across the full ELA domain, ensuring that the cut scores for the three subtests reflected a unified definition of college and career readiness. After discussing the Round 2 results, panelists focused the discussion on the briefing books and used the empirical data to recommend ACT scores that define minimal performance for *Approaching Proficient*, *Proficient*, and *Above Proficient*.

4. Round 4 (Modified Briefing Book): Final cut recommendations were submitted after further discussion.

For all assessments, each panelist's judgments were summed across the items for each performance level to determine the test-level raw cut score, with "yes" = 1, "no" = 0, and score points used for constructed-response items. All raw scores were transformed to scale scores via a raw-to-scale score (RSS) conversion table, and all results were presented to the panelists on the scaled score metric. Final committee-level cut scores were the median of all the individual panelists' cut scores from the final round. Select panelists then participated in a vertical articulation meeting to refine the cut scores across grades 3–11 to ensure logical progression, including the statistically linked cut scores for the PreACT 9 Secure and PreACT Secure.

Panelists also completed three evaluation surveys throughout the meeting to determine their understanding of the process and their confidence in the results: after the practice activity, after Round 3, and after the vertical articulation. The ACT panelists completed an additional survey after Round 2 to capture input from the panelists who were not retained for the ELA panel. Overall, results indicated a strong understanding of the process and high confidence in the recommended cut scores.

7.2. Cut Scores

Error! Reference source not found. presents the resulting ISA scale score cut scores (i.e., the minimum s core students must receive to be classified into a certain performance level), as shown in bold.

Table 7.1. Scale Score Ranges and Cut Scores

Grade	Level 1: *Below Proficient*	Level 2: *Approaching Proficient*	Level 3: *Proficient*	Level 4: *Above Proficient*
5	700–769	**770**–811	**812**–855	**856**–900
8	700–769	**770**–811	**812**–855	**856**–900

Note. Cut scores used to identify Levels 2, 3, and 4 are shown in bold. Students with a score below the cut for Level 2 are placed in Level 1.

Section 8: Student Characteristics and Test Results

8.1. Student Participation

Table 8.1 presents the number and percentage of students who took the spring 2025 ISAs by administration mode (online vs. paper). The results include students taking the accommodated forms.

Table 8.1. Student Participation by Administration Mode

Grade	#Valid Cases	Online N	%	Paper N	%
5	129,936	129,875	99.95	61	0.05
8	133,395	133,362	99.98	33	0.02
Total	263,331	263,237	99.96	94	0.04

Table 8.2 shows the percentage of students with valid scores in each demographic subgroup, as recorded in ADAM through student data uploads by school districts during the assessment period. The demographic information was verified by ISBE before score reporting. Students who were missing data for any demographic variable were excluded from subgroup analyses. These results are based on demographic data from the Pearson score file, not updated information from SIS.

Table 8.2. Student Participation by Demographics

Subgroup	Grade 5		Grade 8	
	N	%	N	%
Economically Disadvantaged	66,578	51.2%	66,244	49.7%
Student with Disabilities (SWD)	25,415	19.6%	24,589	18.4%
English Learner (EL)	21,761	16.7%	21,251	15.9%
Male	66,346	51.1%	68,283	51.2%
Female	63,568	48.9%	65,063	48.8%
American Indian/Alaska Native	315	0.2%	285	0.2%
Asian	7,670	5.9%	7,757	5.8%
Black/African American	20,798	16.0%	21,126	15.8%
Hispanic/Latino	36,568	28.1%	38,629	29.0%
Middle Eastern or North African	279	0.2%	302	0.2%
Native Hawaiian or Other Pacific Islander	96	0.1%	127	0.1%
White/Caucasian	58,265	44.8%	59,653	44.7%
Two or More Races Reported	5,945	4.6%	5,516	4.1%

8.2. Scale Score Distributions

Figure 8.1 and Figure 8.2 present the spring 2025 ISA scale score distributions. The vertical *y*-axis labeled "Count" represents the number of students earning the scale score point indicated along the horizontal *x*-axis. The overall score scale ranges from 700 to 900. Appendix A presents the cumulative frequency distribution for the overall scale scores, and Appendix B presents the subgroup statistics for the overall scale scores.

Figure 8.1. Scale Score Distributions—Grade 5

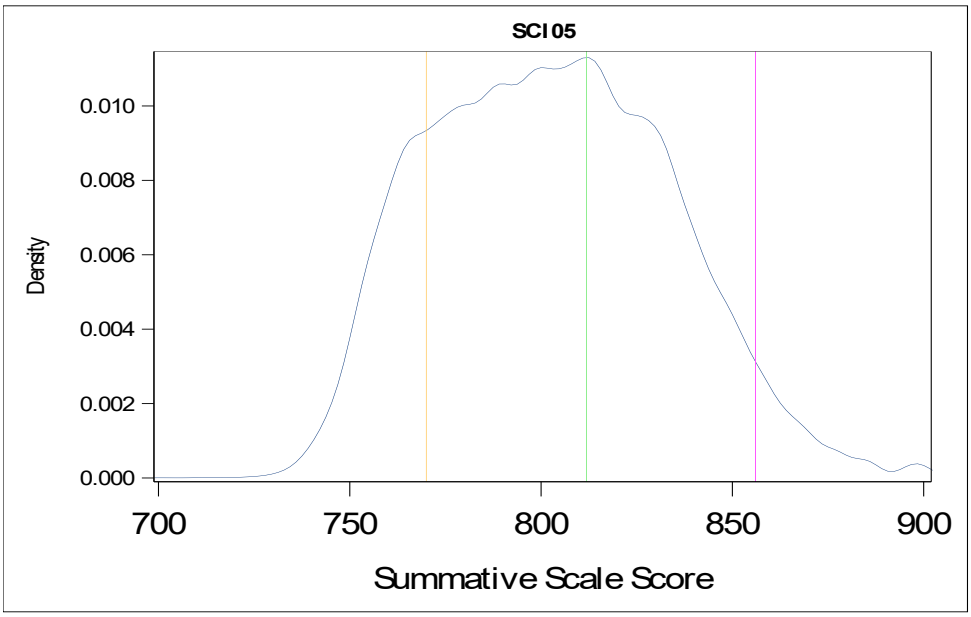
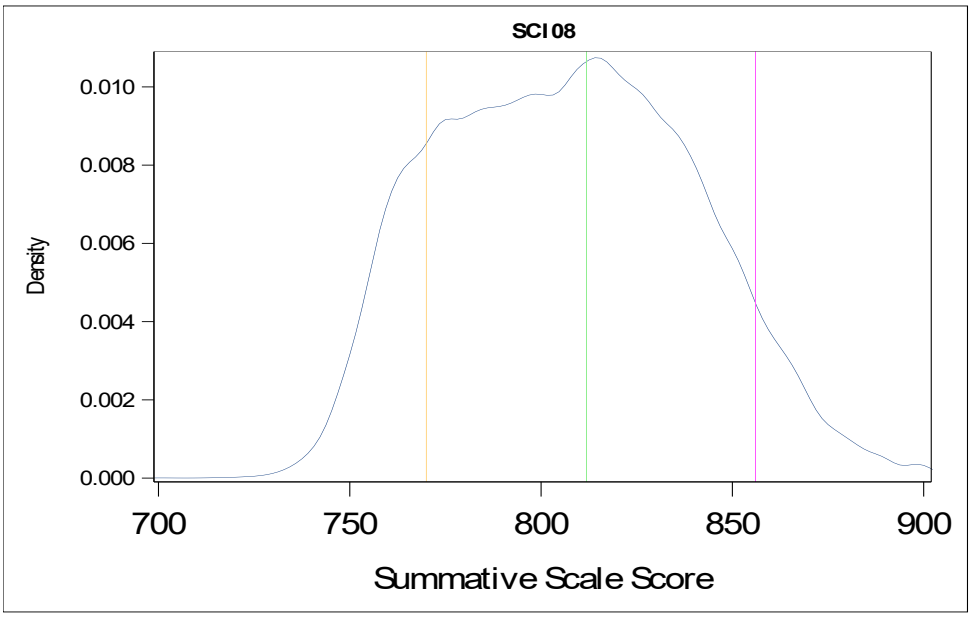


Figure 8.2. Scale Score Distributions—Grade 8



Section 9: Classical Item Analysis

This section presents item analysis results for the operational items included on the spring 2025 ISA test forms. All assessments were post-equated, meaning equating was conducted after the test administration using data from actual student results.

9.1. Data

In preparation for item analysis, student response files were processed to verify that the data were free of errors. Pearson Customer Data Quality staff ran predefined checks on all data files and verified that all fields and data needed to perform the statistical analyses were present and within expected ranges. Next, to produce higher-quality (albeit slightly smaller) datasets, Pearson psychometricians established the following criteria for including students in the operational analyses to determine which, if any, student records should be removed prior to conducting the analysis:

- Exclude all records with an invalid form number.
- Exclude all records flagged as "void."
- Exclude all records where the student attempted fewer than 25% of items.
- For students with more than one valid record, choose the record with the higher raw score.
- Exclude records for students with administration issues or anomalies.

The following factors were also considered during the analyses:

- An operational item may appear on multiple test forms. The item analysis results present unique item counts for an assessment, and the reported item statistics may be based on student responses across multiple occurrences of an item.
- Spoiled or "do not score" items were excluded from the total test score in the item analysis. These items were removed from scoring because of item performance, technical scoring issues, content concerns, or multiple/no correct answers.

9.2. Item Analyses

The following item-level analyses were calculated for the ISAs. Item difficulty and discrimination results are presented in this technical report, whereas the remaining analyses were conducted during key check and adjudication after the ISA test window. No issues were found for any of the operational items based on those statistics.

- Item difficulty (*p*-value)
- Item discrimination (item-total correlation)
- Distractor-total correlation for the multiple-choice items
- Percentage of students choosing each option for the multiple-choice items
- Percentage of students omitting or not reaching each item
- Distribution of item scores

9.2.1. Item Difficulty (P-value)

When constructing tests, a wide range of item difficulties is desired (from easy to hard items) so that students of all ability levels can be assessed with precision. Item difficulty is measured by the *p*-value statistic bounded by 0 and 1 that indicates how easy or hard an item is for students. The *p*-value for dichotomous items is based on the proportion of students who answered an item correctly and is derived by dividing the number of students who got the item correct by the total number of students who answered it. For polytomous items, the *p*-value is the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) divided by the number of possible score points on the item. A high *p*-value indicates that an item is easy (a high proportion of students answered it correctly), while a low *p*-value indicates that an item is difficult. For example, a *p*-value of 0.79 indicates that 79% of students answered the item correctly. Items were flagged for review if the *p*-value was above 0.95 (i.e., too easy) or below 0.25 (i.e., too difficult).

Table 9.1 presents the *p*-value summary statistics for the operational items. The average *p*-values varied across grades.

Table 9.1. Summary of *p*-Values

Grade	#Unique Items	Mean	SD	Min.	Max.	Median
5	75	0.60	0.14	0.25	0.83	0.61
8	75	0.60	0.14	0.25	0.87	0.63

Note. SD = standard deviation, Min. = minimum, Max. = maximum

9.2.2. Item Discrimination (Item-Total Correlation)

Item discrimination is represented by the item-total correlation bounded by -1 and 1 that describes the relationship between performance on a specific item and performance on the total test and indicates how well an item discriminates, or distinguishes, between low- and high-performing students. Students who do well on a test are expected to select the right answer to any given item, and students who do poorly are expected to select the wrong answer. This means that for a highly discriminating item, students who get the item correct will have a higher average test score than students who get the item incorrect. An item with a high positive item-total correlation discriminates between low- and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that low-performing students performed better on an item than high-performing students, an indication that the item may be flawed. The item-total correlation was calculated for both dichotomous and polytomous items as an estimate of the correlation between an observed continuous variable and an unobserved continuous variable hypothesized to underlie the variable with ordered categories (Olsson et al., 1982). Item-total correlations below 0.15 were flagged for review.

Table 9.2 presents the item-total correlation summary statistics for the operational items. The average item-total correlations varied across grades.

Table 9.2. Summary of Item-Total Correlations

Grade	#Unique Items	Mean	SD	Min.	Max.	Median
5	75	0.45	0.11	0.24	0.75	0.46
8	75	0.45	0.12	0.18	0.77	0.46

Note. SD = standard deviation, Min. = minimum, Max. = maximum

The item-total correlation was also calculated for the distractors of selected-response items to describe the relationship between selecting an incorrect response (i.e., a distractor) for an item and performance on the total test. Items with distractor-total correlations above 0.0 were flagged for review as these items may have multiple correct answers, be miskeyed, or have other content issues.

9.2.3. Percentage of Students Choosing Each Answer Option

Selected-response items refer primarily to single-select multiple-choice scored items that require the student to select a response from several answer options. The percentage of students choosing each answer option for the single-select multiple-choice items is calculated, along with the percentages for the high-performing students who scored at the top 20% on the assessment. An item was flagged for review if more high-performing students chose an incorrect option than the correct response. Such a result could indicate that the item has multiple correct answers or is miskeyed.

9.2.4. Percentage of Students Omitting or Not Reaching Each Item

Calculating the percentage of students omitting or not reaching each item is useful for identifying problems with test features such as testing time and item/test layout. Typically, if students have an adequate amount of testing time, approximately 95% of students should attempt to answer each item on the test. A distinction is made between "omit" and "not reached" for items without responses: an item is considered "omit" if the student responded to subsequent items and "not reached" if the student did not respond to any subsequent items.

Patterns of high omit or not-reached rates for items located near the end of a test section may indicate that students did not have adequate time. Omit rates for polytomous items tend to be higher than for dichotomous items. Therefore, the omit rate for flagging individual items was 5% for dichotomous items and 15% for polytomous items. If a student omitted an item, they received a score of 0 for that item and was included in the n-count for that item. However, if an item was near the end of the test and classified as "not reached," the student did not receive a score and was not included in the n-count for that item.

9.2.5. Distribution of Item Scores

For constructed-response items, examination of the distribution of scores is helpful to identify how well the item is functioning. If no student responses are assigned the highest possible score point, this may indicate that the item is not functioning as expected (e.g., the item could be confusing, poorly worded, or unexpectedly difficult), the scoring rubric is flawed, and/or students did not have an opportunity to learn the content. If all or most students score at the extreme ends of the distribution (e.g., 0 and 2 for a three-category item), this may indicate that there are problems with the item or the rubric so that students can receive either full credit or no credit at all, but not partial credit.

The raw score frequency distributions for constructed-response items were computed to identify items with few or no observations at any score points. Items with no observations or a low percentage (i.e., less than 3%) of students obtaining any score point were flagged. Constructed-response items were also flagged if they had U-shaped distributions, with high frequencies for extreme scores and low frequencies for middle score categories.

9.3. Flagging Criteria

The review process for test items includes a systematic keycheck analysis of all operational items. During this process, operational items are evaluated for statistical flags that may indicate potential issues. If any operational items are flagged, they undergo a thorough internal review by Pearson's psychometrics team, followed by further examination by the Illinois State Board of Education (ISBE). For the 2025

administration, no operational items were flagged based on the keycheck analysis. Field test items are also subjected to statistical flagging, and any identified items are subsequently reviewed in a formal data review process involving Illinois educators. All flagged items, whether operational or field test, must be evaluated by Illinois educators to determine their suitability for inclusion in the item bank. These items are either accepted for future use, rejected, or, when appropriate, revised and refield tested.

- *P*-values below 0.15 that indicates too difficult items
- Item-total correlations below 0.20 indicate items that do not effectively distinguish between higher- and lower-performing students. Without approval from ISBE, items with item-total correlations near or below zero will not be used operationally.
- Distractor-total correlations above 0.05 as these items may have multiple correct answers, be miskeyed, or have other content issues
- 40% or more of students choosing a distractor over the keyed response, which indicates that the item may have multiple correct answers or is miskeyed
- High omit and not-reached rates above 5%, which may indicate that students did not have adequate time if patterns of high omit or not-reached rates for items are located near the end of a test section

Section 10: Differential Item Functioning (DIF)

Differential item functioning (DIF) is a statistical procedure used to flag items for potential bias when students from different demographic groups with the same overall ability have a different probability of getting an item correct (e.g., an item that seems easy for female students but not for male students). This section presents DIF results for the operational items included on the spring 2025 ISA test forms. It is important to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify *potential* item bias only. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

10.1. DIF Methods

DIF analyses were conducted for the operational items using the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) for the dichotomously scored items and the standardization DIF procedure for the polytomously scored items (Dorans, 2013; Dorans & Schmitt, 1991; Zwick et al., 1997) in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959). The group representing students in a specific demographic group is referred to as the focal group, and the group comprised of students from outside this group is referred to as the reference group.

In the MH method, students are classified into relevant subgroups of interest (e.g., gender or ethnicity). Using the raw score total as the criteria, students in a certain total score category in the focal group are compared with students in the same total score category in the reference group. For each item, students in the focal group are also compared to students in the reference group who performed equally well on the overall test. The common odds ratio is estimated across all categories of matched student ability using the following formula (Dorans & Holland, 1993), and the resulting estimate is interpreted as the relative likelihood of success on a particular item for members of two groups when matched on ability:

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

where S is = the number of score categories, R_{rs} is the number of students in the reference group who answer the item correctly, W_{fs} is the number of students in the focal group who answer the item incorrectly, R_{fs} is the number of students in the focal group who answer the item correctly, W_{rs} is the number of students in the reference group who answer the item incorrectly, and N_{ts} is the total number of students.

To facilitate the interpretation of the MH results, the common odds ratio is frequently transformed to the delta scale using the following formula (Holland & Thayer, 1988):

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

The standardization DIF procedure compares the item means of the two groups after adjusting for differences in the distribution of students across the values of the matching variable (i.e., total test score) and the standardized differences in expected item scores (i.e., STD-EISDIF) are calculated as follows:

$$STD - EISDIF = \frac{\sum_{s=1}^{S} N_{fs} \times E_f(Y|X=s)}{\sum_{s=1}^{S} N_{fs}} - \frac{\sum_{s=1}^{S} N_{fs} \times E_r(Y|X=s)}{\sum_{s=1}^{S} N_{fs}}$$

where *X* is the total score, *Y* is the item score, *S* is the number of score categories, N_{fs} is the number of students in the focal group in score category *s*, E_r is the expected item score for the reference group, and E_f is the expected item score for the focal group.

10.2. Classification

Based on the DIF statistics, items are classified into three categories (Zieky, 1993): Category A items contain negligible DIF, Category B items exhibit slight-to-moderate DIF, and Category C items possess moderate-to-large DIF values. Positive values indicate DIF in favor of the focal group (i.e., positive DIF items are differentially easier for the focal group), whereas negative values indicate DIF in favor of the reference group (i.e., negative DIF items are differentially easier for the reference group). Table 10.1 presents the flagging criteria for the dichotomous and polytomous items.

Table 10.1. DIF Categories

DIF Category	Dichotomous Items	Polytomous Items
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero or is less than one.	Mantel Chi-square *p*-value > 0.05 or |*STD-EISDIF/SD*| ≤ 0.17
B (slight to moderate)	1. Absolute value of the MH D-DIF is significantly different from zero but not from one and is at least one; or 2. Absolute value of the MH D-DIF is significantly different from one but is less than 1.5. Positive values are classified as "B+" and negative values as "B−."	Mantel Chi-square *p*-value < 0.05 and |*STD-EISDIF/SD*| > 0.17
C (moderate to large)	Absolute value of the MH D-DIF is significantly different from one and is at least 1.5. Positive values are classified as "C+" and negative values as "C−."	Mantel Chi-square *p*-value < 0.05 and |*STD-EISDIF/SD*| > 0.25

Note. STD-EISDIF = standardized DIF, SD = total group standard deviation of item score

10.3. Comparisons

DIF analyses were conducted on each test form for designated comparison groups based on demographic variables including gender, race/ethnicity, economic disadvantage, and special instructional needs such as students with disabilities or ELs, as shown in Table 10.2. DIF analyses were conducted when the following sample size requirements were met:

- The smaller group, reference or focal, had at least 100 students.
- The combined group, reference and focal, had at least 400 students.

Table 10.2. DIF Comparison Groups

Grouping Variable	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	American Indian/Alaska Native (AmerIndian)	White
	Asian	White
	Black or African American	White
	Hispanic/Latino	White
	Native Hawaiian or Pacific Islander	White
	Multiple Race Selected	White
Economic Status	Economically Disadvantaged (EcnDis)	Not Economically Disadvantaged (NoEcnDis)
Special Instructional Needs	English Learner (ELY)	Non-English Learner (ELN)
	Students with Disabilities (SWDY)	Students without Disabilities (SWDN)

Note. Economic status was based on participation in National School Lunch Program (receipt of free or reduced-price lunch).

10.4. Results

Table 10.3 and Table 10.4 present the DIF results for the spring 2025 ISA operational items. Spoiled or "do not score" items were excluded from the total test score for each form in the DIF analysis. These items were removed from scoring because of item performance, technical scoring issues, content concerns, multiple correct answers, or no correct answers. However, the results may include items for certain grade levels that were excluded from scoring based on later analyses. The column "DIF Comparisons" identifies the focal and reference groups for the analysis performed, and "Total #Unique Items" reports the number of unique items included in the analysis. "0" indicates that the DIF analysis did not classify any items in the particular DIF category, while "n/a" indicates that the DIF analysis was not performed due to insufficient sample sizes.

Items flagged with B or C DIF are reviewed by Illinois educators in a formal data review process to confirm their appropriateness to add to the ISA item bank for possible inclusion on a future test form (see Section 3.4 for information on data review). Data review participants review the items for content considerations and bias, fairness, and sensitivity considerations. If the item is approved at data review, Illinois educators shared a consensus that it was appropriate to include the item in the ISA item bank for potential use on a future assessment. When constructing forms, caution is taken to avoid the use of any C or B DIF items unless they are necessary to meet the test blueprint, and ISBE reviews the constructed test forms for approval.

Table 10.3. DIF Results—Grade 5

DIF Comparison[4]	Total #Unique Items	C-	%	B-	%	A	%	B+	%	C+	%
Female vs. Male	75			1	1	73	97	1	1		
White vs. Black/African American	75					75	100				
White vs. Hispanic/Latino	75					75	100				
White vs. Asian	75					75	100				
White vs. AI/AN	75					75	100				
White vs. NH/PI	75					75	100				
White vs. Multiracial	75					75	100				
NED vs. ED	75					75	100				
ELN vs. ELY	75					75	100				
SWDN vs. SWDY	75			1	1	74	99				

Table 10.4. DIF Results—Grade 8

DIF Comparison[5]	Total #Unique Items	C-	%	B-	%	A	%	B+	%	C+	%
Female vs. Male	75			2	3	70	93	2	3	1	1
White vs. Black/African American	75			2	3	72	96	1	1		
White vs. Hispanic/Latino	75					75	100				
White vs. Asian	75					75	100				
White vs. AI/AN	75					74	99	1	1		
White vs. NH/PI	75					75	100				
White vs. Multiracial	75					75	100				
NED vs. ED	75					75	100				
ELN vs. ELY	75			1	1	74	99				
SWDN vs. SWDY	75					75	100				

[4] AI/AN = American Indian/Alaska Native, NH/PI = Native Hawaiian or Pacific Islander, NED = not economically disadvantaged, ED = economically disadvantaged, ELN = not an English learner, ELY = English learner, SWDN = not student with disability, SWDY = student with disability

Section 11: Calibration, Equating, and Scaling

This section describes the item response theory (IRT) model used in this assessment program, provides descriptive statistics of the item parameters, and describes how the reporting scale was established. All ISA assessments in spring 2025 were post-equated.

11.1. IRT Models

Item response theory (IRT) models were used in the item calibration. The Rasch model (Rasch, 1960) was used for 1-point items, and the partial-credit model (Masters, 1982) was used for multiple-point items for calibration. Parameter estimation for items was implemented using Winsteps 4.8.1.0 (Linacre, 2022b) that uses joint maximum likelihood estimation (JMLE), as described by Wright & Masters (1982). All tests were calibrated separately by grade. If there was more than one operational form, all operational forms were calibrated concurrently. All calibration activities were replicated by two psychometricians independently as a quality control measure. The calibration results were also reviewed independently by a senior-level psychometrician at Pearson.

The Rasch model estimates item difficulty and student ability on the same scale. Under the Rasch model, the probability that student *j* with ability θ answers item *i* with difficulty of *b* correctly is as follows:

$$P_i(\theta_j) = \frac{exp(\theta_j - b_i)}{1 + exp(\theta_j - b_i)}$$

The partial-credit model is an extension of the Rasch model for items in which students may receive partial credit. Thus, the partial-credit model reduces to the Rasch model when items have only two response categories (i.e., 0 or 1). According to the partial-credit model, the probability that student *j* scores *x* on item *i*, which has a maximum possible score of *m* (*k* = *m*+1 possible response categories), is expressed as follows:

$$P_{ix}(\theta_j) = \frac{exp \sum_{l=0}^{x}(\theta_j - D_{il})}{\sum_{k=0}^{m_i}[exp \sum_{l=0}^{k}(\theta_j - D_{il})]}$$

where $x = 0, 1, \ldots, m_i$, D_{il} is a step difficulty for score *l* and by definition,

$$\sum_{l=0}^{0}(\theta_j - D_{il}) = 0$$

The step difficulty D_{il} can be decomposed such that

$$D_{il} = b_i + h_{il}$$

where b_i is an overall difficulty for item *i*, and h_{il} is a threshold for score *i* (Embretson & Reise, 2000; Linacre, 2022a). This parameterization allows b_i in the partial-credit model to be comparable to b_i in the Rasch model.

11.2. Checking Model Assumptions

It is important to evaluate how the IRT models applied for ISA fit the data because reported scale scores are derived from theta estimated under the IRT models. An assumption under the IRT models is unidimensionality, that there is exactly one latent variable (e.g., science proficiency) that an instrument intends to measure. However, while this is a more traditional and strict definition of the unidimensionality assumption, essential unidimensionality, in which there is one dominant latent variable with some minor latent variable(s), is a more practically applicable assumption (Stout, 1990). A factor analysis was performed on the item response data for the ISAs to analyze the number of dimensions the assessments appear to be measuring. Given that unidimensional IRT models are used for calibration and scaling, it is important that there be evidence to support their use.

Figure 11.1 and Figure 11.2 present the scree plots for the spring 2024 administration. For most assessments, one factor explained most of the variance, which supports the use of a unidimensional IRT model.

Figure 11.1. Scree Plot—Grade 5

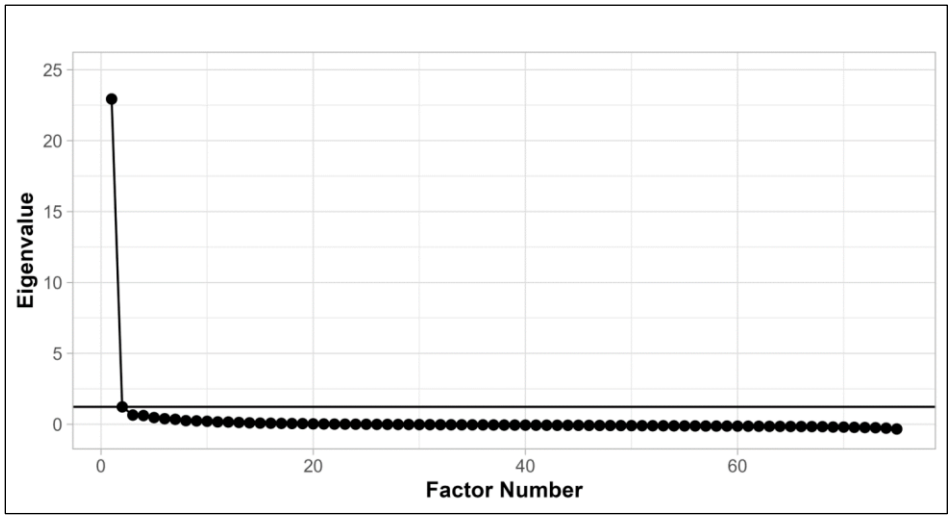
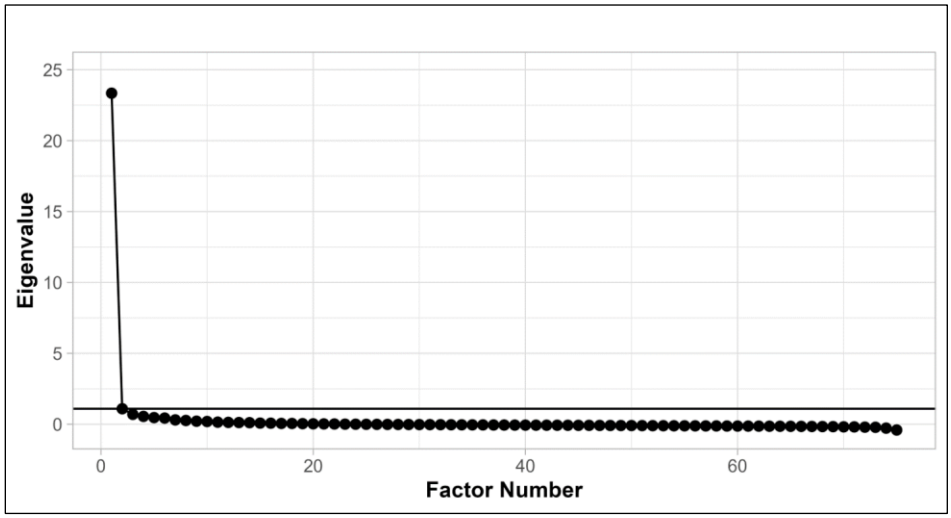


Figure 11.2. Scree Plot—Grade 8



In addition to the exploratory approach, a confirmatory approach was also adopted to ascertain unidimensionality by fitting a one-factor model. Due to the well-known mathematical equivalence between one-factor models with categorical indicators and unidimensional IRT models, a popular approach for unidimensionality assessment is to fit a one-factor model and check the goodness-of-fit indices of the fitted model, namely comparative fit index (CFI), Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA). Those Indices with values indicating satisfactory model fit constitute evidence for unidimensionality. According to Hu and Bentler (1999), CFI and TLI values greater than 0.95 and RMSEA values smaller than 0.06 indicate good model fit. As shown in Table 11.1, CFI, TLI, and RMSEA indicate excellent model fit for both grades, indicating that unidimensionality is a reasonable assumption for the ISA.

Table 11.1. One-Factor Model Goodness of Fit

Grade	CFI	TLI	RMSEA
5	0.993	0.993	0.017
8	0.994	0.994	0.016

11.3. Equating

For the accommodated test forms that were administered to less than 5% of the students, a pre-equating design was adopted. This section focuses on the procedures in the post-equating design used for the non-accommodated test forms.

For the non-accommodated test forms that were administered to the majority of the students, the spring 2025 ISA tests were post-equated and placed on the operational ISA scale using a non-equivalent groups anchor item (NEAT) design. All 75 operational items were considered as potential anchor items that went through an anchor item stability check discussed in the next paragraph. A fixed anchor parameter equating was implemented within Winsteps to place the tests on the operational reporting scale. This was implemented by constraining the parameter estimates in the existing item bank for the anchor items to equal the final parameter estimates obtained in the original calibration analyses. The displacement statistic, which estimates the difference between the fixed parameter and the estimate had the item parameter not been constrained, was evaluated for each anchor item.

Items with a displacement statistic greater than 0.30 or less than -0.30 were reiteratively removed from the anchor set. The criterion of 0.30 has been used to flag displaced anchor items under a common item, non-equivalent group equating design for many state programs (Miller et al., 2004). If more than one anchor item was flagged, the item with the largest magnitude of displacement value was dropped from the anchor set. The displacement values of the remaining anchor items were then re-estimated by implementing the fixed anchor parameter equating with the remaining anchor items. This process was repeated until all the anchor items had displacement values of a magnitude smaller than 0.30 and greater than -0.30.

Table 11.2 presents the number of items for the initial anchor set of each grade and the number of items dropped from each initial anchor set.

Table 11.2. Summary of Anchor Items

Grade	#Items in Initial Anchor Set	#Items Dropped from Anchor	#Final Anchor Items	%Final Anchor Items	#Total Operational Items
5	75	8	67	89%	75
8	75	6	69	92%	75

11.4. IRT Analysis Results

All items converged during calibration using typical procedures for Winsteps software. Standard error of estimates for the Rasch difficulty measures indicated that the parameters were well-estimated. Table 11.3 presents a summary of the IRT statistics that include all the operational items administered in the spring 2025 administration.

Table 11.3. IRT *b* Parameter Estimates Summary

Grade	Form	#Points	#Items	*b* Mean	*b* SD	*b* Min.	*b* Max.
5	ACC	81	75	0.19	0.77	-1.61	1.70
	ONL	81	75	0.16	0.75	-1.29	2.05
8	ACC	81	75	-0.13	0.77	-1.88	1.82
	ONL	81	75	-0.09	0.80	-1.77	1.82

Note. SD = standard deviation, Min. = minimum, Max. = maximum

Three graphs are presented to summarize the characteristics of each operational assessment in Appendix C. The test characteristic curve (TCC) shows an expected total raw score across different student abilities, whereas the test information function (function) shows the amount of test information, and CSEM curve presents an amount of standard error across different student abilities. The CSEM has an inverse relationship with the test information function (TIF) as follows:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}}$$

where *SE(*θ*)* is the conditional standard error of measurement at a given theta value and *TI(*θ*)* is the test information at the theta value.

11.5. Establishing the Reporting Scale

Reporting scales designate student performance into one of four performance levels, with Level 1 indicating the lowest level of performance and Level 4 indicating the highest level of performance. Threshold or cut scores associated with performance levels were initially expressed as raw scores during the standard setting. The raw scores were then transformed into scale scores for reporting purposes.

Derived scores are scores transformed from the theta metric onto a metric that can be reported to and understood by a wide audience. Typically, these are "scale scores." The transformation involves one or more numeral constants that are mathematically combined to the theta ("ability") scores to produce a new metric. The scale score metric for ISA ranges from 700–900. The scale scores for each grade are derived using the following transformation:

$$SS = (m * \theta + b)$$

where *m* is the slope, *b* is the intercept, and θ is the person ability estimate from the Winsteps. Table 11.4 presents the scaling constants for the three grades. Pearson and HumRRO applied the transformation and scaling constants to the theta scores from Winsteps to obtain the scale scores for the overall tests. Scale scores were rounded to the nearest integer. Appendix D presents the resulting raw-to-scale score (RSS) conversion tables for each grade. A student's performance level for the overall test is based on the individual scale score and the scale score cut points as presented in Section 7.

Table 11.4. Scaling Constants

Grade	Slope	Intercept
5	27.7915	783.7664
8	29.2258	792.9673

In addition to the total scale score, the scale score for each domain is reported individually (i.e., Physical Science, Earth/Space Science, and Life Science). These scale scores are generated by including the items associated with each domain and using the item parameter estimates from the concurrent calibration across all domains. The domain-level scale scores (range = 300–500) are also created in a similar linear transformation process. In each domain, the domain-level θ is obtained from Winsteps, and the slope (*m*) and intercept (*b*) listed in Table 11.5 were used to obtain the domain-level scale scores.

Table 11.5. Scaling Constants at the Domain Level

Domain	Grade	Slope	Intercept
Physical Science	5	26.8380	395.8264
	8	26.6146	395.6916
Life Science	5	23.9588	387.3656
	8	25.5938	389.6757
Earth/Space Science	5	24.9408	378.6043
	8	25.5163	401.2682

Section 12: Quality Control Procedures

Quality control in a testing program is a comprehensive and ongoing process. This section describes procedures put into place to monitor the quality of the item bank, test form, and ancillary material development. Additional quality information can be found in the Program Quality Plan document.

12.1. Quality Control of the Item Bank

The ISA item bank consists of test stimuli and items, their metadata, and status (e.g., operational ready, field test ready, released). The ISA items were put in Pearson's Assessment Banking for Building and Interoperability (ABBI) bank houses the items, art, associated metadata, rubrics, and alternate text for use on accommodated forms. It provides an item previewer that allows items to be viewed and interacted with in the same way students see and interact with them, and it manages versioning of items with a date/time stamp. Reviewers can vote on item acceptance and record and retain their review notes for later reconciliation and reference. Item and passage review participants conduct their review in the item banking system and also view the items as the student would, voting to edit, accept, or reject the item and record their comments in the system.

12.2. Quality Control of Test Form Development

The operational test forms were built based on targets and the established blueprints set, and items were pulled into forms based on the criteria approved in the test specifications. The forms then went through an internal review process to ensure content accuracy, completeness, style guide conformity, and tools function. Revisions were incorporated into the forms before final review and approval. The forms quality assurance was performed by Pearson's Assessment and Information Quality (AIQ) organization. AIQ completed a comprehensive review of all online forms for the administration cycle. This group is part of Pearson's larger Organizational Quality group and operates exclusively to validate form operability. The group verifies that the functionality of every online form is working to specifications. The overall functionality and maneuverability of each form is checked, and the behavior of each item within the form is verified.

The items within each form were tested to verify that they operated as expected for students. As a further aspect of the testing process, AIQ confirmed that forms were loaded correctly and that the audio was correct when compared to text. Sections and overviews were reviewed. Technology-enhanced items also were tested as an additional measure. As enumerated in the *Technology Guidelines for Assessments*, user interfaces were compatible with a range of common computer devices, operating systems, and browsers.

Pearson also performed quality control tests to verify that a standard set of responses was outputted to XML as expected after the final version of the form was approved. These responses were based on the keys provided in the test map or a standard open-ended responses string that contained a valid range of characters. As part of these tests, the test maps also were validated against the form layout and item types for correctness. Pearson conducted a multifaceted validation of all item layout, rendering, and functionality. Reviewers conducted comparisons between the approved item and the item as it appeared in the field test form or how it previously appeared; verified that tools and functions in the test delivery system, TestNav, were accurately applied; and verified that the style and layout met all requirements. Answer keys were also validated through a formal key review process.

12.3. Quality Control of Test Materials

Pearson provided high-quality materials in a timely and efficient manner to meet the test administration needs. Because most printing work was done in-house, it was possible to fully control the production environment, press schedule, and quality process for print materials. Strict security requirements were also employed to protect secure materials production. Materials were produced according to the style guide and to the detailed specifications supplied in the materials list.

Pearson Print Service operates within the sanctions of an ISO 9001:2008 Quality Management System, and practices process improvement through Lean principles and employee involvement. Raw materials (paper and ink) used for scannable forms production were manufactured exclusively for Pearson Print Service using specifications created by Pearson Print Service. Samples of ink and paper were tested by Pearson prior to use in production. Project specialists were the point of contact for incoming production.

Purchase orders and other order information were assessed against manufacturing capabilities and assigned to the optimal production methodology. Expectations, quality requirements, and cost considerations were foremost in these decisions. Prior to release for manufacture, order information was checked against specifications, technical requirements, and other communication that includes expected outcomes. Records of these checks were maintained.

Files for image creation flow through one of two file preparation functions: digital pre-press for digital print methodology, or plateroom for offset print methodology. Both the digital prepress and plateroom functions verify content, file naming, imposition, pagination, numbering stream, registration of technical components, color mapping, workflow, and file integrity. Records of these checks are created and saved.

Offset production requires printing that uses a lithographic process. Offline finishing activities are required to create books and package offset output. Digital output may flow through an inkjet digital production line or a sheet-fed toner application process in the Xpress Center. A battery of quality checks was performed in these areas. The checks included color match, correct file selection, content match to proof, litho-code to serial number synchronization, registration of technical components, ink density controlled by densitometry, inspection for print flaws, perforations, punching, pagination, scanning requirements, and any unique features specified for the order. Records of these checks and samples pulled from planned production points were maintained. Offline finishing included cutting, shrink-wrapping, folding, and collating. The collation process has three robust inline detection systems that inspected each book for the following:

- Caliper validation that detects too few or too many pages. This detector will stop the collator if an incorrect caliper reading is registered.
- An optical reader that will only accept one sheet. Two or zero sheets will result in a collator stoppage.
- The correct bar code for the signature being assembled. An incorrect or upside down signature will be rejected by the bar code scanner and will result in a collator stoppage.

Pearson's Quality Assurance department personnel inspected print output prior to collation and shipment. Quality Assurance also supported process improvement, work area documentation, audited process adherence, and established training programs for employees.

12.4. Quality Control of Scoring

12.4.1. Quality Control of Scanning

Establishing and maintaining the accuracy of scanning, editing, and imaging processes is a cornerstone of the Pearson scoring process. While the scanners are designed to perform with great precision, Pearson implements other quality assurance processes to confirm that the data captured from scan processing produces a complete and accurate map to the expected results.

Pearson pioneered optical mark reading and image scanning and continues to improve in-house scanners for this purpose. Software programs drive the capture of student demographic data and student responses from the test materials during scan processing. Routinely scheduled maintenance and adjustments to the scanner components (e.g., camera) maintain scanner calibration. Test sheets inserted into every batch test scanner accuracy and calibration. Controlled processes for developing and testing software specifications included a series of validation and verification procedures to confirm the captured data can be mapped accurately and completely to the expected results and that editing application rules are properly applied.

12.4.2. Quality Control of Image Editing

The final step in producing accurate data for scoring is the editing process. Once information from the documents was captured in the scanning process, the scan program file was executed, comparing the data captured from the student documents to the project specifications. The result of the comparison was a report (or edit listing) of documents needing corrections or validation. Image Editing Services performed the tasks necessary to correct and verify the student data prior to scoring. Using the report, editors verified that all unscanned documents were scanned, or the data were imported into the system through some other method such as flatbed scan or key entry. Documents with missing or suspect data were pulled and verified, and corrections or additional data were entered. Standard edits included

- Incorrect or double gridding
- Incorrect dates (including birth year)
- Mismatches between pre-ID label and gridded information
- Incomplete names

When all edits were resolved, corrections were incorporated into the document file containing student records. Additional quality checks were also performed, including student n-count checks to ensure that

- students were placed under the correct header,
- all sheets belonged to the appropriate document,
- documents were not scanned twice, and
- no blank documents existed.

Finally, accuracy checks were performed by checking random documents against scanned data to verify the accuracy of the scanning process. Once all corrections were made, the scan program was tested a second time to verify all data were valid. When the resulting output showed that no fields were flagged as suspect, the file was considered clean and scoring began. Once all scanning was completed, the right/wrong response data were securely handed off.

12.4.3. Quality Control of Answer Document and Data

Quality control of answer document processing and scoring involves all aspects of the scoring procedures, including key-based and rule-based machine scoring and handscoring for constructed-response items and performance tasks. Based on lessons learned from previous administrations, the following quality steps were implemented:

- Raw score validation (e.g., score key validation; evidence statement, field test non-score; double-grid combinations; possible correct combination, if applicable; out-of-range/negative test cases)
- Matching (e.g., validation of high-confidence criteria, low-confidence criteria, cross document, external or forced matching by customer; prior to and after data updates; extract file of matched and unmatched documents)
- Demographic update tests (e.g., verification of data extract against corresponding layout; valid values for updatable fields; invalid values for updatable/non-updatable fields; negative test for non-existing record or empty file)

The following components were also included in the quality control process:

- XML Validation: A combination of automated validation against 100% of item XMLs and human inspection of XML from selected difficult item types or composite items
- Administration/End-to-End Data Validation: An automated generation of response data from approved test maps that have known conditions against the operational scoring systems and data generation systems to verify scoring accuracy
- Psychometric Validation: Verification of data integrity using criteria typically used in psychometric processes (e.g., statistical key checks) and categorization of identified issues to help inform investigation by other groups
- Content Validation: An examination, by subject matter experts, of all items using a combination of automated tools to generate response and scoring data

The following quality control process for answer keys and scoring was also implemented:

- Pearson's psychometrics team conducted empirical analyses based on preliminary data files and flagged items based on statistical criteria.
- Pearson content team reviewed the flagged items and provided feedback on the accuracy of content, answer keys, and scoring.
- Items potentially requiring changes were added to the product validation log for further investigation by other Pearson teams.
- Staff was notified of items for which keys or scoring changes were recommended.
- Illinois approved/rejected scoring changes.
- All approved scoring changes were implemented and validated prior to the generation of the data files used for psychometric processing.

12.5. Quality Control of Psychometric Processes

High-quality psychometric work for the operational administrations was necessary to provide accurate and reliable results of student performance. The psychometric analyses were all conducted according to well-defined specifications, and data cleaning rules were clearly articulated and applied consistently throughout the process. Results from all analyses underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were used by members of the team for each statistical procedure.

Quality control steps performed at different stages of the psychometric analyses including data screening, classical item analysis, and the creation of conversion tables. Data screening is an important first step to ensure quality data input for meaningful analysis. The Pearson Customer Data Quality team validated all student data files used in the operational psychometric analyses. The data validation for the student data files and item response files included the following steps:

1. Validated variables in the data file for values in acceptable ranges
2. Validated that the test form ID, unique item numbers, and item sequence on the data file were consistent with the test form values on the corresponding test map
3. Computed the composite raw score, and domain raw scores, given the item scores in the student data file
4. Compared computed raw scores to the raw scores in the student data file
5. Compared the student item response block to the item scores
6. Flagged student records with inconsistencies for further investigation

All classical item analysis results were reviewed by Pearson psychometricians, and items flagged for unusual statistical properties were reviewed by the content team. Refer to Section 9.3 for the classical item analysis item flagging criteria.

Pearson also brought in an external third-party replicator, HumRRO, to provide three key psychometric services under this contract: item calibrations, test equating, and score verifications. In the calibration process, HumRRO psychometricians independently calibrated test items based on the provided item data and test maps from Pearson, following the Rasch model as specified in the psychometric specifications. The calibrated item parameters were compared with Pearson's results, and any discrepancies were identified and resolved collaboratively. For test equating, HumRRO used the same operational procedures as Pearson to independently equate forms and items, ensuring consistency in the equating method used, whether pre- or post-equating. The linking transformation coefficients and item parameters were compared, and discrepancies were addressed iteratively with Pearson. Finally, in score verification, HumRRO generated raw score-to-scale score (RSS) conversion tables and student ability estimates, comparing them to Pearson's results to ensure accuracy. Any discrepancies identified were investigated and resolved through ongoing collaboration. This collaborative approach, including regular team meetings and a shared process, ensured consistency and accuracy throughout all stages of the psychometric replication process.

Finally, conversion tables are used to generate reported scores for students and must be accurate. Comprehensive records were maintained on item-level decisions, and thorough checks were made to ensure that the correct items were included in the final score. Post-equated conversion tables were developed independently by two psychometricians and completely matched. A reasonableness check was also conducted by psychometricians for each grade level to make sure the results were in alignment with observations during the analyses prior to conversion table creation.

Section 13: Reliability

Reliability focuses on the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations due to chance. Thus, reliability measures the consistency of the scores across conditions that can be assumed to differ at random. In statistical terms, the variance in the distribution of test scores (i.e., the differences among individuals) is partly due to real differences in the knowledge, skill, or ability being tested (true variance) and partly due to random errors in the measurement process (error variance). Reliability is an estimate of the proportion of the total variance that is true variance. Reliability for the ISAs was evaluated based on the following analyses for both raw and scale scores:

- Internal consistency
- Standard error of measurement (SEM)
- Decision accuracy and consistency
- Inter-rater agreement (see Section 5.4)

13.1. Internal Consistency and SEM

Reliability coefficients for both raw and scale scores range from 0.0 to 1.0. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain similar scores upon repeated testing occasions if the students do not change in their level of the knowledge or skills measured by the test. The reliability estimates attempt to answer the question, "How consistent would the scores of these students be over replications of the entire testing process?" Raw score reliability estimates are an internal-consistency measure derived from analysis of the consistency of the performance of individuals across items within a test. It serves as a good estimate of alternate forms reliability but does not consider form-to-form variation due to lack of test form parallelism, nor is it responsive to day-to-day variation due to, for example, the student's state of health or the testing environment. The scale score reliability results use a modified measure of internal consistency that accounts for the conversions between raw scores and scale scores.

The SEM quantifies the amount of error in the test scores. SEM is the extent by which students' scores tend to differ from the scores they would receive if the test were perfectly reliable. As the SEM increases, the variability of students' observed scores is likely to increase across repeated testing. Observed scores with large SEMs pose a challenge to the valid interpretation of a single test score.

Reliability estimates are influenced by test length, test characteristics, and sample characteristics (Lord & Novick, 1968; Tavakol & Dennick, 2011; Cortina, 1993). As test length decreases and samples become smaller and more homogeneous, lower estimates of alpha are obtained (Tavakol & Dennick, 2011; Pike & Hudson, 1998). Moderate to acceptable ranges of reliability tend to exceed 0.5 (Cortina, 1993; Schmitt, 1996). Estimates lower than 0.5 may indicate a lack of internal consistency. Additional analyses investigate whether lower estimates of alpha are due to a restriction in range of the sample. In these cases, the alpha estimates are not appropriate measures of internal consistency. As a result, sample-free reliability estimates are also provided, such as scale score reliability (Kolen et al., 1996).

13.1.1. Raw Score Estimation

Coefficient alpha (Cronbach, 1951), the most used measure of reliability, is an internal consistency measure derived from analysis of the consistency of the performance of students across items within a test. It is estimated by substituting sample estimates for the parameters as follows:

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

where *n* is the number of items, σ_i^2 is the variance of scores on the *i*th item, and σ_X^2 is the variance of the total score (sum of scores on the individual items).

However, because the test forms have mixed item types (dichotomous and polytomous items), it is more appropriate to report stratified alpha (Feldt & Brennan, 1989), which is a weighted average of coefficient alphas for item sets with different maximum score points or "strata." Stratified alpha is a reliability estimate computed by dividing the test into parts (strata), computing alpha separately for each part, and using the results to estimate a reliability coefficient for the total score. Stratified alpha is used here because different parts of the test consist of different item types and may measure different skills. The formula for the stratified alpha is as follows:

$$\rho_{strata} = 1 - \frac{\sum_{h=1}^{H}\sigma_{x_h}^2(1-\alpha_h)}{\sigma_X^2}$$

where $\sigma_{X_h}^2$ is the variance for part h of the test, σ_X^2 is the variance of the total scores, and α_h is coefficient alpha for part *h* of the test. Estimates of stratified alpha are computed by substituting sample estimates for the parameters in the formula. The average stratified alpha is a weighted average of the stratified alphas across the test forms. The formula for the SEM is as follows:

$$\sigma_E = \sigma_X\sqrt{1-\rho_{XX'}}$$

where σ_E is the standard error of measurement, σ_X is the standard deviation of the test raw score, and $\rho_{XX'}$ is the reliability estimated by substitution of appropriate statistics for the parameters.

13.1.2. Scale Score Estimation

Like the stratified alpha coefficients, scale score reliability coefficients range from 0.0 to 1.0. The higher the reliability coefficient for a set of scores, the more likely students would be to obtain similar scores upon repeated testing occasions if they do not change in their level of the knowledge or skills measured by the test. Because the scale scores are computed from a total score and do not have an item-level component, a stratified alpha coefficient cannot be computed for scale scores. Instead, Kolen et al.'s (1996) method for scale score reliability was used. The general formula for a reliability coefficient,

$$\rho = 1 - \frac{\sigma^2(E)}{\sigma^2(X)}$$

involves the error variance, $\sigma^2(E)$, and the total score variance, $\sigma^2(X)$. Using Kolen et al.'s (1996) method, conditional raw score distributions are estimated using Lord and Wingersky's (1984) recursion formula. The conditional raw score distributions are transformed into conditional scale score distributions. Denote *X* as the raw sum score ranging from 0 to *X*, and *s* as a resulting scale score after transformation. The conditional distribution of scale scores is written as $P(X = \chi|\theta)$. The mean and variance, $\sigma^2[s(X)]$, of this distribution can be computed using these scores and their associated probabilities.

The average error variance of the scale scores is computed as follows:

$$\sigma^2{}_{Error\_scale} = \int_{\theta} \sigma^2(s(X)|\theta)\, g(\theta) d\theta$$

where $g(\theta)$ is the ability distribution. The square root of the error variance is the conditional standard error of measurement of the scale scores. Just as the reliability of raw scores is one minus the ratio of error variance to total variance, the reliability of scale scores is one minus the ratio of the average variance of measurement error for scale scores to the total variance of scale scores:

$$\rho_{scale} = 1 - \frac{\sigma^2{}_{Error\_scale}}{\sigma^2[s(X)]}$$

The Windows program POLYCSEM (Kolen, 2004) was used to estimate scale score error variance and reliability.

13.1.3. Results

Table 13.1 and Table 13.2 present the overall raw and scale score test reliability estimates, including the average reliability that is estimated by averaging the internal consistency estimates computed for all the individual forms of the test. The tables also present the number of forms, the sample size of the minimum and maximum reliability, and the average maximum possible score for each set of tests. Estimates were calculated only for groups of 100 or more students administered a specific test form.

Grade	#Forms	Avg. Scale Score SEM	Avg. Reliability	Online1 Reliability	ACC1 Reliability
5	2	7.44	0.93	0.93	0.93
8	2	7.66	0.93	0.93	0.93

Table 13.3 presents the reliability estimates for each domain.

Table 13.1. Summary of Raw Score Test Reliability for Total Group

Grade	#Forms	Max. Possible Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online1 Alpha	ACC1 N	ACC1 Alpha
5	2	81	4.06	0.91	125,215	0.94	4,796	0.87
8	2	81	3.98	0.91	130,098	0.94	3,313	0.87

Note. The reported n-counts represent the number of students taking the non-TTS form.

Table 13.2. Summary of Scale Score Test Reliability for Total Group

Grade	#Forms	Avg. Scale Score SEM	Avg. Reliability	Online1 Reliability	ACC1 Reliability
5	2	7.44	0.93	0.93	0.93
8	2	7.66	0.93	0.93	0.93

Table 13.3. Average Reliability Estimates by Domain

Grade	Domain	Max. Possible Score	Avg. Reliability
5	Physical Science	27	0.74
	Life Science	27	0.79
	Earth/Space Science	27	0.78
8	Physical Science	27	0.74
	Life Science	27	0.78
	Earth/Space Science	27	0.78

Note. RS = raw score, Avg. = average

Table 13.4 and Table 13.5 present the raw score reliability and SEM for various demographic subgroups with sufficiently large samples (i.e., 100 or more for a given test form). Reliability estimates depend on score variance, and subgroups with smaller variance are likely to have lower reliability estimates than the total group. Overall, the reliability estimates for the subgroups of interest were close to the reliability estimates of the total group.

Table 13.4. Reliability by Subgroup—Grade 5

Subgroup	Max. Possible Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	ACC1 N	ACC1 Alpha
Total Group	81	4.06	0.91	125,215	0.94	4,796	0.87
Male	81	4.03	0.91	63,899	0.95	2,481	0.88
Female	81	4.08	0.90	61,297	0.94	2,312	0.86
American Indian/Alaska Native	81	n/a	n/a	277	0.94	n/a	n/a
Asian	81	n/a	n/a	7,627	0.94	n/a	n/a
Black/African American	81	4.09	0.91	20,579	0.92	222	0.89
Hispanic/Latino	81	4.11	0.90	32,652	0.93	3,972	0.86
Middle Eastern or North African	81	n/a	n/a	276	0.94	n/a	n/a
Native Hawaiian/Pacific Islander	81	n/a	n/a	n/a	n/a	n/a	n/a
Two or More Races	81	n/a	n/a	5,901	0.95	n/a	n/a
White	81	4.00	0.92	57,811	0.94	470	0.90
Economically Disadvantaged	81	4.11	0.90	62,792	0.93	3,840	0.86
Not Economically Disadvantaged	81	3.98	0.91	62,423	0.94	956	0.89
English Learner (EL)	81	4.14	0.88	17,888	0.90	3,924	0.86
Non-EL	81	4.02	0.93	107,327	0.94	872	0.91
Students with Disabilities (SWD)	81	4.08	0.91	24,261	0.94	1,201	0.89
Students without Disabilities	81	4.04	0.90	100,954	0.94	3,595	0.86
American Sign Language (ASL)	81	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	81	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	81	n/a	n/a	n/a	n/a	209	0.90
Text-to-Speech (TTS)	81	n/a	n/a	n/a	n/a	3,762	0.86

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

Table 13.5. Reliability by Subgroup—Grade 8

Subgroup	Max. Possible Score	Avg. Raw Score SEM	Avg. Reliability	Online1 N	Online 1 Alpha	ACC1 N	ACC1 Alpha
Total Group	81	3.98	0.91	130,098	0.94	3,313	0.87
Male	81	3.95	0.91	66,478	0.95	1,814	0.88
Female	81	3.99	0.90	63,571	0.94	1,499	0.86
American Indian/Alaska Native	81	n/a	n/a	268	0.95	n/a	n/a
Asian	81	n/a	n/a	7,747	0.94	n/a	n/a
Black/African American	81	4.02	0.88	20,959	0.92	169	0.83
Hispanic/Latino	81	4.03	0.89	35,882	0.93	2,753	0.86
Middle Eastern or North African	81	n/a	n/a	301	0.94	n/a	n/a
Native Hawaiian/Pacific Islander	81	n/a	n/a	125	0.94	n/a	n/a
Two or More Races	81	n/a	n/a	5,477	0.95	n/a	n/a
White	81	3.90	0.93	59,339	0.94	323	0.92
Economically Disadvantaged	81	4.04	0.89	63,599	0.93	2,654	0.86
Not Economically Disadvantaged	81	3.90	0.92	66,499	0.94	659	0.90
English Learner (EL)	81	4.07	0.87	18,506	0.89	2,751	0.85
Non-EL	81	3.93	0.93	111,592	0.94	562	0.91
Students with Disabilities (SWD)	81	3.99	0.91	23,842	0.94	750	0.89
Students without Disabilities	81	3.96	0.90	106,256	0.94	2,563	0.86
American Sign Language (ASL)	81	n/a	n/a	n/a	n/a	n/a	n/a
Closed-Caption	81	n/a	n/a	n/a	n/a	n/a	n/a
Screen Reader	81	n/a	n/a	n/a	n/a	253	0.87
Text-to-Speech (TTS)	81	n/a	n/a	n/a	n/a	2,586	0.86

Note. SEM = standard error of measurement, n/a = not applicable. ASL, closed-caption, screen reader, and TTS present the results for students taking the accommodated forms.

13.2. Decision Accuracy and Consistency

The reliability of the classifications for the students was calculated using the computer program BB-CLASS (Brennan, 2004), which operationalizes a statistical method developed by Livingston and Lewis (1993, 1995). As Livingston and Lewis (1993, 1995) explain, this method uses information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficient) to estimate two kinds of statistics, decision accuracy and decision consistency. Decision accuracy refers to the extent to which the classifications of students based on their scores on the test form agree with the classifications made based on the classifications that would be made if the test scores were perfectly reliable. Decision consistency refers to the agreement between these classifications based on two nonoverlapping, equally difficult forms of the test.

Decision consistency values are always lower than the corresponding decision accuracy values because both classifications are subject to measurement error in decision consistency. In decision accuracy, only one of the classifications is based on a score that contains an error(s). It is not possible to know which students were accurately classified, but it is possible to estimate the proportion of the students who were accurately classified. Similarly, it is not possible to know which students would be consistently classified if they were retested with another form, but it is possible to estimate the proportion of the students who would be consistently classified.

Table 13.6 presents decision accuracy and consistency results based on the summative scale. “Exact Level” presents the estimates of the indices based on classifications of students into one of the four performance levels, and “Level 2 or Higher vs. 2 or Lower” presents the estimates of the indices based on classifications of students as being either in one of the upper two levels (Levels 3 and 4) or in one of the lower three levels (Levels 1 and 2). Level 3 is considered the college and career readiness standard on the ISAs.

Table 13.6. Decision Accuracy and Consistency Summary

Statistic	Grade	Exact Level	Level 3 or Higher vs. 2 or Lower
Accuracy	5	0.82	0.92
	8	0.82	0.91
Consistency	5	0.75	0.88
	8	0.74	0.88

Table 13.7 and Table 13.8 present more detailed information about the accuracy and the consistency of the classification of students into performance levels by grade. Each cell in the tables shows the estimated proportion of students who would be classified into a particular combination of performance levels. The sum of the bold values on the diagonal is approximately equal to the level of decision accuracy or consistency presented in Table 13.6. For “Level 3 and Higher vs. 2 and Lower” in the summary tables, the sum of the shaded values in these tables is approximately equal to the level of decision accuracy or consistency presented in Table 13.6. The sums based on values in these tables may not match exactly to the values in the summary table due to truncation and rounding.

Table 13.7. Decision Accuracy and Consistency by Performance Level—Grade 5

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	700–769	**0.12**	0.03	0.00	0.00	0.15
	770–811	0.04	**0.40**	0.05	0.00	0.49
	812–855	0.00	0.04	**0.27**	0.02	0.32
	856–900	0.00	0.00	0.01	**0.04**	0.05
Consistency	700–769	**0.11**	0.05	0.00	0.00	0.16
	770–811	0.05	**0.36**	0.06	0.00	0.47
	812–855	0.00	0.06	**0.25**	0.02	0.32
	856–900	0.00	0.00	0.02	**0.04**	0.05

Table 13.8. Decision Accuracy and Consistency by Performance Level—Grade 8

Statistic	Scale Score Range	Level 1	Level 2	Level 3	Level 4	Category Total
Accuracy	700–769	**0.10**	0.03	0.00	0.00	0.13
	770–811	0.04	**0.36**	0.05	0.00	0.45
	812–855	0.00	0.04	**0.29**	0.02	0.36
	856–900	0.00	0.00	0.01	**0.06**	0.07
Consistency	700–769	**0.10**	0.05	0.00	0.00	0.14
	770–811	0.05	**0.32**	0.06	0.00	0.43
	812–855	0.00	0.06	**0.27**	0.02	0.35
	856–900	0.00	0.00	0.02	**0.06**	0.08

Section 14: Validity

As stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" (p. 11). The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity (or evidence of lack of validity), including design, content specifications, and psychometric characteristics. This section summarizes the various sources of validity evidence for the assessments based on the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

14.1. Evidence Based on Test Content

Content validity addresses whether the test adequately samples the relevant material it purports to cover. Evidence based on content of achievement tests is supported by the degree of correspondence between test items and content standards, and the degree to which the test measures what it claims to measure is known as construct validity. The ISAs adhere to the principles of evidence-centered design, in which the standards to be measured (the Illinois Learning Standards for Science) are identified, and the performance a student needs to achieve to meet those standards is delineated in the performance expectations. Test items are reviewed for adherence to Universal Design principles, which maximize the participation of the widest possible range of students. Accommodations were also made available based on individual student need documented in the student's IEP, 504 Plan, or an EL Plan.

Gathering construct validity evidence for the assessments is embedded in the process by which the test content is developed and validated. See Sections 2 and 3 for an overview of the content development process. The items and tasks were then field tested prior to their operational use. Finally, an important consideration when constructing test forms is recognition of items that may introduce construct-irrelevant variance. Such items should not be included on test forms to help ensure fairness to all subgroups of students. Bias and sensitivity committees were convened to review all newly developed items, in addition to a review by content experts.

14.2. Evidence Based on Internal Structure

Internal structure refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test interpretations are based" (AERA et al., 2014, p. 16). If an item has poor internal structure, it may not be measuring the intended construct accurately, which can lead to invalid or unreliable results. Evidence for the ISAs includes (a) intercorrelations between an assessment's domains to examine how they relate to each other and verify the unidimensionality of the assessment (i.e., measuring only one construct); (b) reliability correlation coefficients that measure a test's internal consistency, or the extent to which the items in an assessment are measuring the same underlying construct; and (c) local item independence, an assumption under the IRT model that assumes any item pair is uncorrelated, conditioned on the latent trait an instrument is intended to measure (e.g., science proficiency).

14.2.1. Intercorrelations

The ISAs have three domain scale scores (Physical Science, Life Science, and Earth/Space Science). Table 14.1 presents the weighted average Pearson intercorrelations between domains by averaging the intercorrelations computed for all the core operational forms of each assessment. The shaded values along the diagonal are the reliabilities from Section 13.1.3. The average intercorrelations are provided in the lower portion of the tables, and the total sample sizes are provided in the upper portion of the tables.

Table 14.1. Average Interrelations and Reliability between Subclaims

Grade	Domain	Physical Science	Life Science	Earth/Space Science
5	Physical Science	0.74	130,011	130,011
	Life Science	0.82	0.79	130,011
	Earth/Space Science	0.83	0.86	0.78
8	Physical Science	0.74	133,411	133,411
	Life Science	0.83	0.78	133,411
	Earth/Space Science	0.83	0.86	0.78

14.2.2. Reliability

Internal consistency is typically measured via correlations among the items on an assessment and provides an indication of how much the items measure the same general construct. As shown in Section 13.1, the reliability estimates computed using coefficient alpha (Cronbach, 1951) indicate an acceptable level of reliability for the ISAs. Overall, the reliability estimates indicate that the items within each assessment measure a similar construct.

14.3. Evidence Based on Responses Processes

Evidence based on response processes refers to the cognitive process engaged in by students when answering test items, or the "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA et al., 2014, p. 15). A census field test was administered in spring 2022 by the previous vendor to field test all newly developed items and explore their functioning and performance. Future operational assessments will continue to include embedded field test items, with any item that does not pass the item reviews or data review being removed from the item pool.

14.4. Evidence Based on Relationships to Other Variables

Empirical results concerning the relationships between test scores and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. As indicated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), the variables investigated can include other tests that measure the same construct and different constructs, criterion measures that scores on the test are expected to predict, and demographic characteristics of students that are expected to be related and unrelated to test performance. For example, when a test's scores are highly correlated with scores from a different, external assessment, it provides evidence that the tests measure the same or similar construct. Such a study is currently out of scope for the ISA, but a future study may be conducted as the program evolves.

14.5. Evidence for Validity and Consequences of Testing

Because state tests are administered "with the expectation that some benefit will be realized from the intended use of the scores" (AERA et al., 2014, p. 19), validity evidence supporting the use and interpretation of the through year test results may be investigated as a consequence of testing. However, as the *Standards* note, "validation is the joint responsibility of the test developer and the test user…the test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used" (AERA et al., 2014, p. 13). As such, this technical report provides information about test content and technical quality but does not interfere in the use of scores. Ultimate use of test scores is determined by Illinois educators.

14.6. Summary

The goal of providing validity evidence is to demonstrate that the assessment is accurately measuring the intended construct. The item development process involved educators, assessment experts, and bias and sensitivity experts in review of item sets for accuracy, appropriateness, and freedom from bias. Items were then field tested prior to the initial operational administration. Psychometric analyses further provided evidence that the assessments measure what is intended. For example, the intercorrelations of the domains and the reliability analyses indicate that the ISAs are unidimensional. In addition to the validity information presented in this section of the technical report, other information in support of the uses and interpretations of the scores appear in the following sections:

- Section 8 presents information regarding student characteristics and test results for the spring administration.
- Section 9 provides information concerning the test characteristics based on classical test theory.
- Section 10 provides information regarding the DIF analyses.
- Section 13 provides information on the test reliability.

References

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

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0). (CASMA Research Report No. 9). Center for Advanced Studies in Measurement, University of Iowa.

Camara, W. J., Allen, J. M., & Moore, J. L. (2017). Empirically-based college and career readiness cut scores and performance standards. In K. L. McClarty, K. D. Mattern & M. N. Gaertner (Eds.), *Preparing students for college and careers: Theory, measurement, and educational practice*. Routledge.

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

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

Davis, L. L., & Moyer, E. L. (2015). *PARCC performance level setting technical report*. Partnership for Assessment of Readiness for College and Careers (PARCC).

Dorans, N. J. (2013). *ETS contributions to the quantitative assessment of item, test and score fairness* (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04). Educational Testing Service.

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

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. RR-91-47). Educational Testing Service.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Erlbaum.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.

Haertel, E. H., Beimers, J. N., & Miles, J. A. (2012). The briefing book method. In Cizek G. J. (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 283–299). Routledge.

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

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

Kolen, M. J. (2004). *POLYCSEM windows console version* [Computer software]. The Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129–140.

Linacre, J. M. (2022a). *Winsteps® Rasch measurement computer program user's guide, Version 4.8.1.0.* Winsteps.com.

Linacre, J. M. (2022b). *Winsteps® (Version 4.8.1.0)* [Computer Software]. http://www.winsteps.com/

Livingston, S. A., & Lewis, C. (1993). *Estimating the consistency and accuracy of classifications based on test scores* (ETS Research Report No. RR-93-48). Educational Testing Service.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.

Lochbaum, K. E., Way, D., & Song, T. (2015). *Phase I research results of PARCC automated scoring operational study*.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*(4), 453–461.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748.

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

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

National Research Council (NRC). (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academic Press. https://nap.nationalacademies.org/catalog/18290/next-generation-science-standards-for-states-by-states

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Biometrika, 47*, 337–347.

Pike, C. K., & Hudson, W. W. (1998). Reliability and measurement error in the presence of homogeneity. *Journal of Social Service Research, 24*(1–2), 149–163.

Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). *Setting multiple performance standards using the Yes/No method: An alternative item mapping method.* Meeting of the NCME*,* Montreal, Canada.

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

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

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Way, D., Lochbaum, K. E., & Song, T. (2016). *Phase II research results of PARCC automated scoring operational study*.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Lawrence Erlbaum

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items* (ETS Research Report RR-97-05). Educational Testing Service.

Appendix A: Scale Score Cumulative Frequencies

Table A.1. Scale Score Cumulative Frequencies—Grade 5

Score Band	N	%	Cumulative N	Cumulative %
700–704	4	0.00	4	0.00
705–709	1	0.00	5	0.00
710–714	8	0.01	13	0.01
715–719	3	0.00	16	0.01
720–724	11	0.01	27	0.02
725–729	20	0.02	47	0.04
730–734	106	0.08	153	0.12
735–739	327	0.25	480	0.37
740–744	806	0.62	1,286	0.99
745–749	1,622	1.25	2,908	2.24
750–754	2,531	1.95	5,439	4.19
755–759	5,189	3.99	10,628	8.18
760–764	4,254	3.27	14,882	11.45
765–769	6,370	4.90	21,252	16.36
770–774	6,348	4.89	27,600	21.24
775–779	6,114	4.71	33,714	25.95
780–784	6,318	4.86	40,032	30.81
785–789	8,321	6.40	48,353	37.21
790–794	6,520	5.02	54,873	42.23
795–799	6,786	5.22	61,659	47.45
800–804	7,055	5.43	68,714	52.88
805–809	7,104	5.47	75,818	58.35
810–814	7,551	5.81	83,369	64.16
815–819	7,577	5.83	90,946	69.99
820–824	5,003	3.85	95,949	73.84
825–829	7,641	5.88	103,590	79.72
830–834	4,923	3.79	108,513	83.51
835–839	4,819	3.71	113,332	87.22
840–844	4,415	3.40	117,747	90.62
845–849	3,794	2.92	121,541	93.54
850–854	1,759	1.35	123,300	94.89
855–859	1,597	1.23	124,897	96.12
860–864	1,374	1.06	126,271	97.18
865–869	1,087	0.84	127,358	98.02
870–874	896	0.69	128,254	98.71
875–879	702	0.54	128,956	99.25
880–884	–	–	–	–
885–889	503	0.39	129,459	99.63
890–894	–	–	–	–
895–899	298	0.23	129,757	99.86
900	179	0.14	129,936	100.00

Table A.2. Scale Score Cumulative Frequencies—Grade 8

Score Band	N	%	Cumulative N	Cumulative %
700–704	5	0.00	5	0.00
705–709	1	0.00	6	0.00
710–714	2	0.00	8	0.01
715–719	6	0.00	14	0.01
720–724	16	0.01	30	0.02
725–729	50	0.04	80	0.06
730–734	58	0.04	138	0.10
735–739	298	0.22	436	0.33
740–744	721	0.54	1,157	0.87
745–749	1,502	1.13	2,659	1.99
750–754	2,455	1.84	5,114	3.83
755–759	3,445	2.58	8,559	6.42
760–764	6,139	4.60	14,698	11.02
765–769	4,535	3.40	19,233	14.42
770–774	6,448	4.83	25,681	19.25
775–779	6,378	4.78	32,059	24.03
780–784	6,355	4.76	38,414	28.80
785–789	6,304	4.73	44,718	33.52
790–794	6,358	4.77	51,076	38.29
795–799	6,584	4.94	57,660	43.23
800–804	6,667	5.00	64,327	48.22
805–809	7,111	5.33	71,438	53.55
810–814	5,110	3.83	76,548	57.38
815–819	7,608	5.70	84,156	63.09
820–824	8,068	6.05	92,224	69.14
825–829	5,429	4.07	97,653	73.21
830–834	5,784	4.34	103,437	77.54
835–839	5,454	4.09	108,891	81.63
840–844	5,419	4.06	114,310	85.69
845–849	4,968	3.72	119,278	89.42
850–854	4,714	3.53	123,992	92.95
855–859	2,031	1.52	126,023	94.47
860–864	1,870	1.40	127,893	95.88
865–869	1,523	1.14	129,416	97.02
870–874	1,245	0.93	130,661	97.95
875–879	1,026	0.77	131,687	98.72
880–884	735	0.55	132,422	99.27
885–889	542	0.41	132,964	99.68
890–894	–	–	–	–
895–899	264	0.20	133,228	99.87
900	167	0.13	133,395	100.00

Appendix B: Scale Score Performance by Demographic Subgroup

Table B.1. Scale Score Performance by Demographic Subgroup—Grade 5

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Scale Score	**129,936**	**100.0%**	**802.62**	**31.00**	**700**	**900**
Female	63,568	48.9%	802.06	30.02	700	900
Male	66,346	51.1%	803.14	31.90	700	900
American Indian/Alaska Native	315	0.2%	795.66	29.09	727	897
Asian	7,670	5.9%	820.37	32.04	733	900
Black or African American	20,798	16.0%	784.76	25.78	700	900
Hispanic/Latino	36,568	28.1%	792.75	27.31	713	900
Middle Eastern or North African	279	0.2%	793.58	29.92	741	870
Native Hawaiian or Pacific Islander	96	0.1%	808.07	32.15	733	897
Two or More Races	5,945	4.6%	806.66	32.08	726	900
White	58,265	44.8%	812.50	29.50	700	900
Not Economically Disadvantaged	63,358	48.8%	814.97	29.86	700	900
Economically Disadvantaged	66,578	51.2%	790.86	27.27	700	900
Non-English Learner (EL)	108,175	83.3%	807.03	30.73	700	900
English Learner (EL)	21,761	16.7%	780.67	21.57	722	900
Students without Disabilities	104,521	80.4%	806.79	29.79	701	900
Student with Disability (SWD)	25,415	19.6%	785.46	29.96	700	900
Spanish	3,797	2.9%	779.64	18.96	727	867

Note. SD = standard deviation. Economic status was based on participation in the National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table B.2. Scale Score Performance by Demographic Subgroup—Grade 8

Subgroup	N	%	Mean	SD	Min.	Max.
Overall Scale Score	**133,395**	**100.0%**	**806.96**	**32.68**	**700**	**900**
Female	65,063	48.8%	806.92	31.23	700	900
Male	68,283	51.2%	806.97	34.00	700	900
American Indian/Alaska Native	285	0.2%	798.69	33.18	700	889
Asian	7,757	5.8%	829.21	32.88	700	900
Black or African American	21,126	15.8%	788.33	26.62	700	900
Hispanic/Latino	38,629	29.0%	796.53	29.02	700	900
Middle Eastern or North African	302	0.2%	803.13	31.67	741	882
Native Hawaiian or Pacific Islander	127	0.1%	810.54	32.58	751	876
Two or More Races	5,516	4.1%	810.04	34.01	711	900
White	59,653	44.7%	817.17	31.12	700	900
Not Economically Disadvantaged	67,151	50.3%	819.30	31.53	700	900
Economically Disadvantaged	66,244	49.7%	794.44	28.84	700	900
Non-English Learner (EL)	112,144	84.1%	811.51	32.31	700	900
English Learner (EL)	21,251	15.9%	782.92	22.53	700	882
Students without Disabilities	108,806	81.6%	811.11	31.60	700	900
Student with Disability (SWD)	24,589	18.4%	788.58	30.99	700	900
Spanish	2,618	2.0%	776.38	19.40	720	864

Note. SD = standard deviation. Economic status was based on participation in the National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Appendix C: TCCs, CSEM Curves, and TIF Curves

This appendix presents the IRT test characteristic curves (TCCs), conditional standard error of measurement (CSEM) curves, and test information function (TIF) curves by grade.

Figure C.1. TCCs, CSEM Curves, and TIF Curves—Grade 5

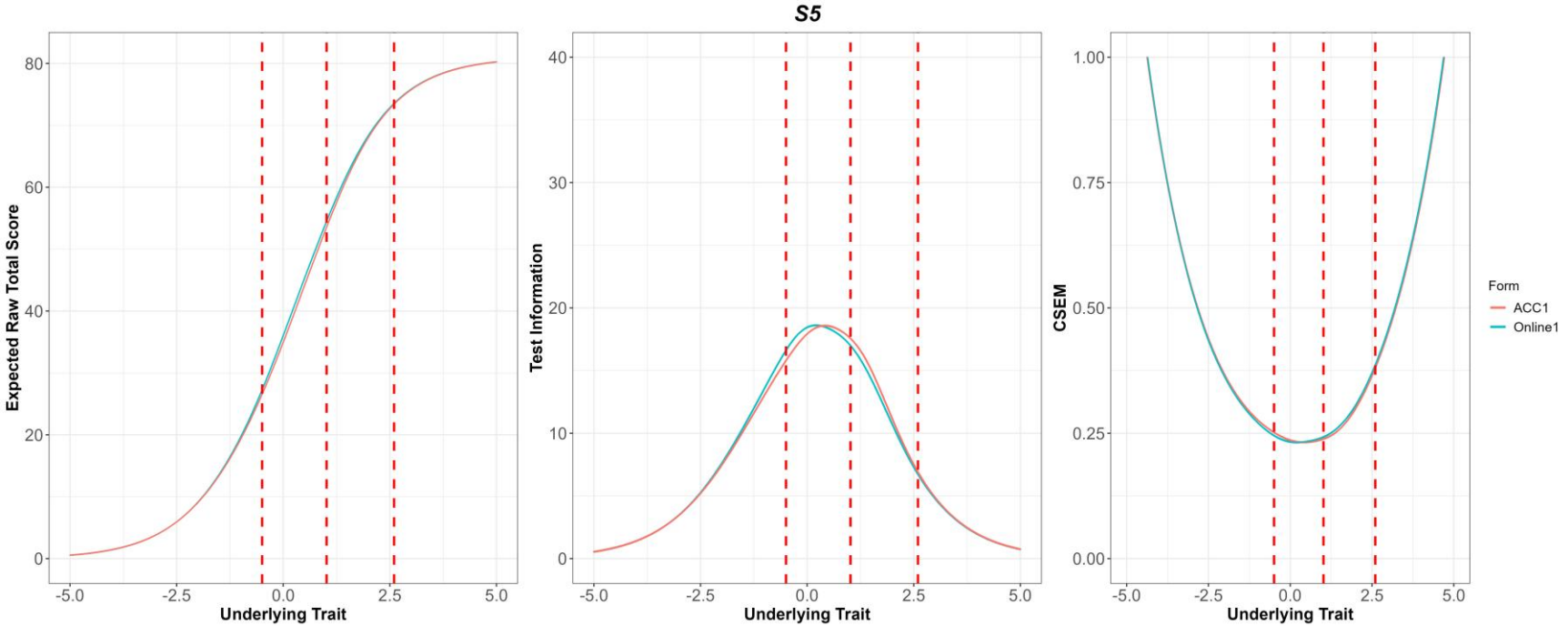
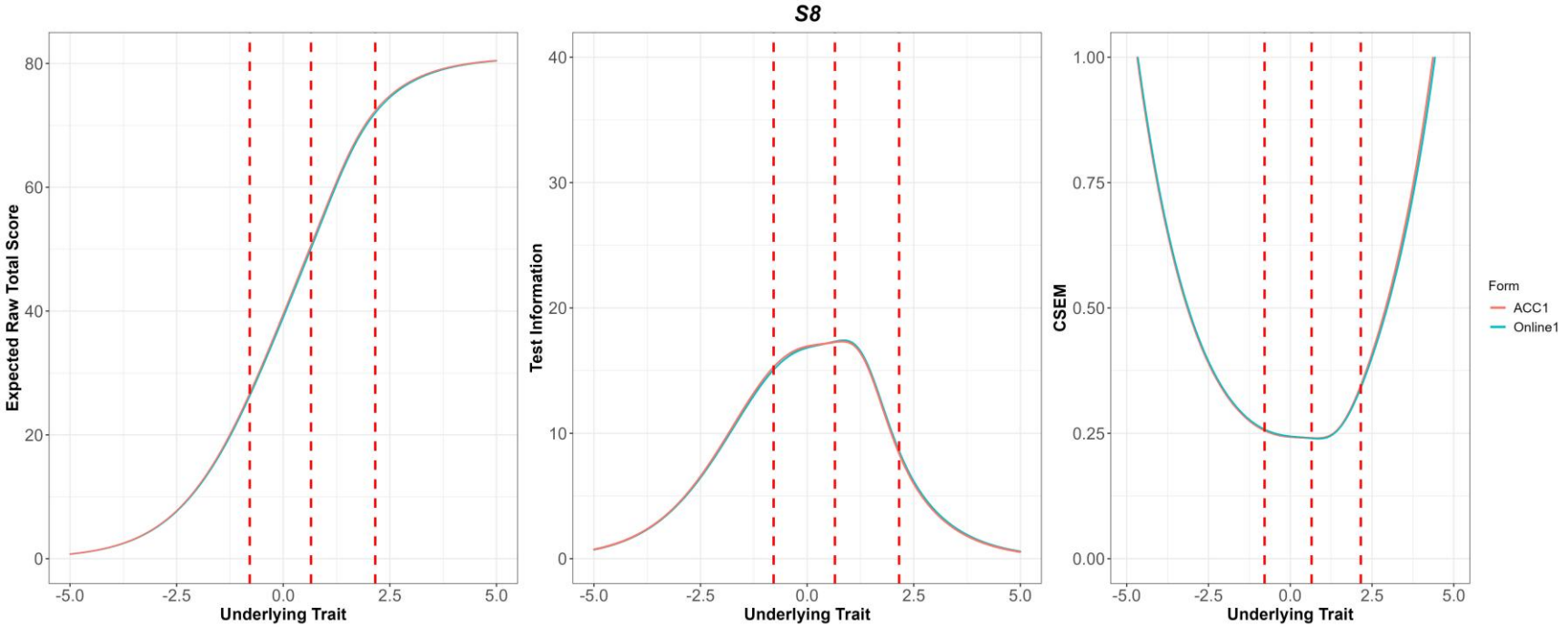


Figure C.2. TCCs, CSEM Curves, and TIF Curves—Grade 8



Pearson