New York State Testing Program 2024: Elementary- and Intermediate-Level Science Grades 5 & 8

Technical Report

2024

Table of Contents

List of Tables

List of Figures

Section 1: Introduction and Overview

1.1. Introduction

This technical report provides detailed information regarding the technical, statistical, and measurement attributes of the New York State Testing Program (NYSTP) for the Elementary-Level Science (ELS) Grade 5 and Intermediate-Level Science (ILS) Grade 8 2024 Operational Tests. This report includes information about test content and test development, item (i.e., individual test question) and test statistics, validity and reliability, test administration, standard setting, scoring, scaling, and student performance.

1.2. Test Purpose

The 2024 ELS and ILS NYSTP has been designed to measure student knowledge and skills as defined in the New York State P-12 Science Learning Standards (NYSP12SLS). The 2024 tests were the first administration measuring these new learning standards. The tests are designed to allow the classification of student proficiency into four performance levels: Level 1, Level 2, Level 3, and Level 4. Likewise, the test provides opportunities for students at each of these performance levels to demonstrate their knowledge and skills in the NYSP12SLS. Details about the content standards for ELS and ILS are described in Section 2.4. Test Blueprints.

1.3. Expected Participants

Students in New York State (NYS) public schools in Grades 5 and 8 (and ungraded students of equivalent chronological ages) are the expected participants for the ELS and ILS assessments. Religious and independent schools may participate in the testing program, but their participation is not mandatory. In 2024, some religious and independent schools participated in the testing program across both grades. These schools were included in the data analyses. Public school and charter school students were required to take the science assessments administered at their grade, except for students who took a Regents-level course in science or a very small percentage of students with severe cognitive disabilities who took the New York State Alternate Assessment (NYSAA). For more detail on this exemption, please refer to the 2024 *NYSTP Grades 3–8 English Language Arts, Mathematics, and Science Tests School Administrator's Manual* (SAM), available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/sam-g3-8-2024.pdf.

1.4. Test Use and Decisions Based on Assessment

The NYSTP ELS and ILS Tests are used to measure science knowledge and skills as defined in the NYSP12SLS. The results are used to determine if schools, districts, and the State meet the required progress objectives specified in the NYS Every Student Succeeds Act (ESSA, 2015) accountability system. Several types of scores are available from the ELS and ILS Tests, which are discussed in this section.

1.4.1. Scale Scores

The scale scores are a quantification of the proficiency measured by the ELS and ILS Tests. Scale scores are comparable only within a given subject and grade. Scale scores are not comparable across grades nor across subjects. The scale scores are reported at the individual student level and can be aggregated. Detailed information on the derivation and properties of the scale scores, including the range of scale scores for each subject and grade, is provided in Section 6: IRT

Calibration. The ELS and ILS Tests' scale scores are the basis for placing students into performance levels, which can be used to determine student progress within schools and districts; support registration of schools and districts; determine eligibility of students for additional educational services; and provide educators with indicators of a student's need, or lack of need, for remediation in a specific content-area.

1.4.2. Performance Level Cut Scores and Classification

Student performance is classified as Level 1, Level 2, Level 3, or Level 4 for the ELS and ILS Tests. The definition of each performance level is as follows:

- **NYS Level 1:** Students performing at this level are below proficient in standards for their grade. They demonstrate limited knowledge, skills, and practices, as embodied by the Learning Standards, that are considered insufficient for the expectations at this grade.

- **NYS Level 2:** Students performing at this level are partially proficient in standards for their grade. They demonstrate knowledge, skills, and practices, as embodied by the Learning Standards, that are considered partial but insufficient for the expectations at this grade. Students performing at Level 2 are considered on track to meet current New York State high school graduation requirements but are not yet proficient in Learning Standards at this grade.

- **NYS Level 3:** Students performing at this level are proficient in standards for their grade. They demonstrate knowledge, skills, and practices, as embodied by the Learning Standards, that are considered sufficient for the expectations at this grade.

- **NYS Level 4:** Students performing at this level excel in standards for their grade. They demonstrate knowledge, skills, and practices, as embodied by the Learning Standards, that are considered more than sufficient for the expectations at this grade.

The performance level cut scores used to distinguish between Level 1, Level 2, Level 3, and Level 4 were established during the standard setting process in Summer 2024. This process is described in detail in Section 8: Standard Setting and Appendix N: Standard Setting Report.

1.4.3. Subscores

The ELS and ILS Tests have three major claims, or subscores: Life Science, Physical Science, and Earth and Space Sciences. Within the NYSP12SLS these assessment-based claims are the overarching statements that identify what student's should be able to do at the end of instruction and account for the majority of the ELS and ILS test items. Table 1.1 presents the reporting subscore categories and the point values that correspond to each on the 2024 tests. (The tables in Appendix A provide information on the numbers and types of items on the 2024 ELS and ILS Tests.)

Table 1.1. ELS and ILS Tests 2024 Subscore Categories and Total Possible Score Points

Grade	Reporting Subscores and Total Subscore Points		
	Life Science	Physical Science	Earth and Space Sciences
5	9	14	9
8	19	19	14

1.5. Testing Accommodations

In accordance with federal law under the Individuals with Disabilities Education Act (IDEA, 2004) and the "Fairness in Testing" section of the *Standards for Educational and Psychological Testing* (AERA et al., 2014, pp. 49–72), accommodations that do not alter the measurement of any construct being tested are allowed for test takers. This allowance is in accordance with a student's Individualized Education Program (IEP) or Section 504 Accommodation Plan (504 Plan). School principals are responsible for ensuring that proper accommodations are provided, when necessary, and that staff providing accommodations are properly trained. Details on testing accommodations can be found in the 2024 *NYSTP English Language Arts, Mathematics, and Science Tests School Administrator's Manual* (SAM), available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/sam-g3-8-2024.pdf.

1.6. Test Transcriptions

For visually impaired students, large type and braille editions of the test books are provided. In most cases, students dictate and/or record their responses, and teachers transcribe student responses to multiple-choice items onto scannable answer sheets and transcribe responses to constructed-response items onto the regular test books. Some of the students who use large type editions will fill in the answer sheets by themselves. The large type editions are created and printed by the New York State Education Department's (NYSED's) testing vendor, NWEA. SeeWriteHear, LLC, produces the braille editions. SeeWriteHear employs certified Library of Congress braille transcribers and delivers braille in accordance with the Braille Authority of North America (BANA) standards. Camera-ready versions of the regular test books are provided to the braille vendor, which then produces the braille editions.

1.7. Test Translations

The NYSTP ELS and ILS Tests are translated into eight languages: Arabic, Bengali, Chinese (Simplified), Chinese (Traditional), Haitian Creole, Korean, Russian, and Spanish. These tests are translated in order to provide students with the opportunity to demonstrate proficiency independent of their command of the English language. Translated tests in each language are available online at https://www.nysedregents.org/ei/ei-science-translations.html.

English Language Learners (ELLs) taking the ELS and ILS Tests may be provided with an oral translation of the test when a written translation is not available in the student's native language. The following testing accommodations are also made available to ELLs: separate testing location, bilingual glossaries, simultaneous use of English and alternative language editions, oral translation for lower incidence languages, and writing responses in the native language.

Section 2: Test Design and Development

2.1. Test Descriptions

The 2024 ELS and ILS Tests are criterion-referenced tests composed of multiple-choice (MC) and constructed-response (CR) items based on the New York State P-12 Science Learning Standards (NYSP12SLS). The tests were administered in NYS classrooms during a thirty-day period from April to May 2024. Details on the administration and scoring of these tests can be found in Section 4: Test Administration and Scoring. Additional information can be found in the 2024 *NYSTP English Language Arts, Mathematics, and Science Tests School Administrator's Manual* (SAM), available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/sam-g3-8-2024.pdf.

2.1.1. ELS Grade 5 and ILS Grade 8 Tests

The 2024 ELS and ILS Tests were designed to measure science knowledge and skills, as defined by the NYSP12SLS using a Principled Assessment Design. This approach uses claims and evidence to build tasks that allow students to provide/produce evidence to exemplify knowledge and skills across a range of performance. The tests assessed science standards by using 1-credit MC and 1-credit CR items, including Technology Enhanced Items (CR-TEIs). For MC questions, students select the response that best completes the statement or answers the question from four answer choices. For CR questions, students record their answer to an open-ended question. CR-TEIs are used to assess standards or parts of standards that cannot be adequately assessed via typical question types. CR-TEIs include four item types—graphing items, drag-and-drop items, multi-select items, and grid items. They allow students to show proficiency in skills such as completing models and graphing.

All questions on the ELS and ILS Tests are organized into clusters of questions, including a combination of MC, CR, and CR-TEI items, that follow an assessment storyline. Each assessment storyline provides a coherent path toward building Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts (CCCs) attached to a phenomenon. In question clusters, each question that is answered may add to the developing explanation, model, or design solution. The group of questions in a cluster follow a theme or storyline grounded in a phenomenon that is focused on an anchor Performance Expectation (PE). However, questions that address other related PEs are also included in the cluster.

Question clusters include an introduction (which informs students of how many questions are part of the cluster), multiple stimuli (reading passages, data tables, graphs, diagrams, photos, etc.), and questions that draw on one or more of the stimuli. These stimuli provide students with an interesting and relatable setting that drives the progression of the assessment storyline. Stimuli, derived from vetted sources that are appropriate to the grade level being tested, are scientifically accurate and use real data when applicable. There will be variation in the number of questions that make up each cluster depending upon the assessment storyline; as a result, there may be slight variation in the total number of test questions year-to-year.

2.2. Test Configuration

2.2.1. Test Design

The 2024 ELS and ILS Tests were one session each consisting of 1-point MC items, 1-point CR items (including CR-TEIs), and embedded field test items. Approximately 60% of the test is comprised of multiple-choice items while 40% is CR items. Schools were advised to allocate a minimum of 90 minutes for administration of the ELS test and 120 minutes for the administration of the ILS test.

The tables in Appendix A provide information on the numbers and types of items on the 2024 ELS and ILS Tests.

2.2.2. Embedded Field Test Items

In 2010, NYSED announced its commitment to work toward embedding items for field testing within the Grades 3–8 English Language Arts and Mathematics Operational Tests. This commitment was extended to include the ELS and ILS Tests. Embedding field test items allows for a better representation of student responses and provides more reliable field test data on which to build future operational tests. In other words, since the specific locations of the embedded field test items are not disclosed and they look the same as operational test items, students are unable to differentiate field test items from operational test items. Therefore, field test data derived from embedded items are free of the effects of differential student motivation that may characterize stand-alone field test designs.

For Spring 2024, all field test items for the ELS and ILS Tests were embedded in the operational test. Embedding field test items for the ELS and ILS Tests eliminated the need for stand-alone field test forms during Spring 2024. See Section 2.8: Field Testing for more information.

2.3. New York State Educators' Involvement in Test Development

New York State educators are actively involved in ELS and ILS test development. These educators provide critical input throughout all stages of the test development process, which include stimuli selection, item writing, educator item review, operational forms construction, a final eyes meeting (i.e., a final review of the test materials prior to printing), and rangefinding of field test items.

NYSED gathers a diverse group of educators to review all test materials to create fair and valid tests. The participants are selected for each testing activity based on:

- Certification and appropriate grade-level experience
- Special population experience
- Geographical region
- Gender
- Ethnicity
- Type of school (urban, suburban, or rural)

The selected participants must be certified and have both teaching and testing experience. Most of the participants are classroom teachers. Specialists such as science coaches and special

education and bilingual instructors may also participate. Some participants are recommended by principals, professional organizations, Big Five Cities (i.e., Buffalo, Rochester, Syracuse, Yonkers, and New York City), and/or the Staff and Curriculum Development Network (SCDN). A file of participants is maintained and routinely updated with current participant information, as well as the addition of possible future participants as recruitment forms are received. The process of continually updating and adding to this file contributes to NYSED's ability to include many educators in the test development process.

2.4. Test Blueprints

The NYSP12SLS for ELS and ILS are organized around Performance Expectations (PEs) that are connected to the Scientific and Engineering Practices (SEPs), Disciplinary Core Ideas (DEIs), and Crosscutting Concepts (CCCs). The assessments include questions that require students to connect all three dimensions (i.e., SEPs, DEIs, and CCCs). The ELS and ILS NYSTP has been designed to measure science knowledge and skills as defined by the NYSP12SLS. The ELS Test assesses science standards for Grades 3–5 (with a foundation of preK–2), and the ILS Test assesses science standards for Grades 6–8. All items on the ELS and ILS Tests are organized into clusters that include an introduction (which informs students of how many questions are a part of the cluster), multiple stimuli (reading passages, data tables, graphs, diagrams, photos, etc.), and questions that draw on one or more of the stimuli. The questions within the cluster will include MC and CR items (including CR-TEIs). Appendix B shows the test blueprints and the ranges of actual numbers of score points in the ELS and ILS Tests, including the ranges of allowable points for each area and the actual numbers of points on the 2024 test forms.

2.5. Item Review Criteria Documents

To guide test item development and to help ensure that NYS tests are measuring the NYSP12SLS with fidelity, criteria were established for selecting stimuli and writing test items.

Stimuli can include reading passages, data tables, graphs, diagrams, and photos, etc. Criteria documents were used to determine whether each stimuli suggested for testing use was grade appropriate, fair, and possessed the necessary characteristics to assess each standard.

Item review criteria for the ELS and ILS tests were used to ensure clarity, language and graphical appropriateness, fairness, freedom from bias, fidelity of measurement to the NYSP12SLS, and conformity to the expectations for specific item types and formats for each test item. Each section of the criteria includes pertinent questions that determine whether an item is of sufficient quality. The first two criteria, clarity and language and graphical appropriateness and fairness, identify the basic components of quality test items. The criteria for clarity and graphical appropriateness are used to help ensure that students understand what is being asked in each item and that the language in the item does not adversely affect a student's ability to perform the required task. For example, the criteria include checking to make sure that the visual load for any item containing a graphic is reasonable and that interpreting a graphic does not confuse the underlying construct. Likewise, the fairness criteria are used to evaluate whether items are unbiased, non-offensive, and not disadvantageous to any given subgroup. The criteria also require documentation of how each item measures the assigned science standard(s). Finally,

the criteria address the specific demands for different item types and formats. (See Appendix C for the Item Review Criteria.)

2.5.1. Principles of Universal Design

To create tests that are as equitable as possible for students, principles of Universal Design were employed during the creation of the tests and test items. In a report published by the National Council on Educational Outcomes, "'Universally designed assessments' are designed and developed from the beginning to allow participation of the widest possible range of students, and to result in valid inferences about performance for all students who participate in the assessment" (Thompson et al., 2002, p. 5). The report goes on to describe seven elements of a universally designed assessment. These elements are:

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

In accordance with these elements, the *Universal Design Item Checklist* in Appendix E was used during item development.

2.6. Item Development

Item development for the 2024 test forms was conducted during recent annual development cycles. The goal of item development was to develop a sufficient number of high-quality, NYSP12SLS-aligned clusters to populate the test forms. Using the criteria documents for ELS and ILS, and workshop presentations and activities, NYSED staff trained item writers. The item writers had teaching or assessment experience in elementary- and/or intermediate-level science; experience in writing for large-scale, high-stakes assessments; and, at a minimum, a bachelor's degree in either education or science. The item writers were given specific assignments, based on the test blueprints.

Item writers provided items to NYSED content specialists for review who retrieved and reviewed the items. If NYSED staff determined that an item did not meet the criteria, NYSED staff provided an explanation for rejection or revision. If NYSED staff determined that an item met the criteria but could be improved with editing, the NYSED staff recorded notes for the edits. Those notes were reviewed during meetings at which NYSED staff reviewed and edited all the items to ensure that they met the criteria. All items accepted were moved forward for educator item review.

2.7. Educator Item Review

After being reviewed by NYSED, the assessment clusters were presented to NYS educators. The reviews were facilitated by NYSED in conjunction with NWEA. The educators used the following checklist to review each cluster, including all items and stimuli within the cluster.

Science Cluster Checklist:

- Is the science accurate?
- Is the text clearly written and appropriate?
- Does the cluster follow a logical storyline?
- For items
    - Is the item aligned to the intended Performance Expectation (PE)?
    - Is the item aligned to the intended Performance Level Description (PLD)?
    - Is there one and only one key?
    - Are the distractors plausible?
    - Is the item free of bias and sensitivity concerns?
- For stimuli
    - Are the stimuli accurate and appropriate?
    - Are appropriate safety, data, and sensitivity issues addressed?

As the educators reviewed the clusters, they discussed their judgments about them. If the educators felt that an item or stimulus did not align to the standards, did not meet quality standards, or was not fair, they made recommendations for editing the item. NYSED staff later reviewed the recommendations and made the appropriate edits prior to field testing.

2.8. Field Testing

Once items have been developed and thoroughly reviewed by a variety of stakeholders, they must be field tested. Field testing is a critically important step in the test development process, as it is only through the gathering of actual student-response data that a variety of psychometric characteristics may be evaluated. More items are field tested than are needed for the operational forms because that enables tests to be constructed with items that include the best possible characteristics from both a content and psychometric perspective. All ELS and ILS field test items (MC, CR, and CR-TEIs) were embedded within the 2024 operational test forms.

A variety of analyses were conducted to better understand how the items field tested may perform on future operational forms including classical item analysis, inter-rater reliability for constructed-response items, differential item functioning (DIF), item response theory (IRT), item calibration, scaling, and fit evaluation. Many of these analyses are described at length in the *New York State Testing Program 2024: Elementary- and Intermediate-Level Science Grades 5 & 8 Field Test Technical Report*.

2.9. Rangefinding

NWEA conducts rangefinding after CR items have been field tested. The purpose of rangefinding was to have NYS educators review student responses and arrive at consensus scores based on the standards established by NYSED and the scoring rubrics. The consensus scores became the basis for operational rating guides and scoring ancillaries. To arrive at consensus, committees of NYS educators reviewed, discussed, and rated student responses to the constructed-response field test items. NYSED content experts and NWEA Scoring Directors oversaw this process.

Through rangefinding, CR field tested items (which included CR-TEIs) were reviewed to determine what level of knowledge and skills were necessary to be evident in the response to

receive credit. This determination was then used to score the remainder of the CR field tested items and to inform the development of scoring materials for the operational test.

After the committee reviewed the pre-approved grounding guide set, groups of committee members familiarized themselves with each item type, scoring a small number of responses representative of the different score points. After a group scoring exercise, committee members independently scored other student responses. The committee then reviewed and discussed their results and determined consensus scores for the responses. The rangefinding results were used to build training materials for NWEA scorers, who scored the field test responses to CR items.

2.10. Item Selection and Test Form Creation (Criteria and Process)

Test items for the 2024 operational tests were selected from the pools of available ELS and ILS items. These items were field tested by standalone field testing in 2022 or 2023.

The test construction process involved several iterative steps. Three criteria governed the item selection process:

- Meet content specifications
- Select items with the best psychometric characteristics from the item pools
- Combine psychometric characteristics of all selected items with the intended psychometric goals for each entire form

NYSED, with the help of NYS educators, used the test designs, blueprints, and psychometric guidelines for item selection. The psychometric guidelines are based on the classical and IRT statistics associated with the test items. Appendix F provides general psychometric guidelines for operational item selection. For example, one of the guidelines for building the ELS and ILS Tests was that the point-biserial correlation for MC items should be equal to or greater than 0.20, which would indicate that students who responded correctly to that item also tended to do well on the overall test. The few exceptions to this guideline were due to content considerations that required the inclusion of particular items. Decisions to use such items were made very carefully, and no item with a negative point-biserial correlation was allowed on the test.

Using the pool of items that were field tested, NYSED and NYS educators made preliminary selections for the ELS and ILS forms. The selections were then reviewed to make sure that the items conformed to the different criteria. If the content criteria were not met, new items were selected. After review, the item selections were reviewed by NYSED psychometricians. If items with undesirable statistics were selected, the psychometricians proposed items with more desirable statistics. Once NYSED and the psychometric team were satisfied that the content and statistics of the selected items and the proposed whole forms met the requirements, the items underwent a final review by NYSED and NYS educators during a meeting that took place in October 2023 in Albany, New York.

During the meeting, NYS educators worked with NYSED to review the content of the proposed ELS and ILS test forms. They looked at how those items combined to create entire operational forms and reviewed them for quality and appropriateness, using their subject-matter expertise. The goal was to ensure that the test items and forms were defensible from content and

psychometric perspectives. The outcome was ELS and ILS test forms that met psychometric parameters and contained items that met content criteria.

Educators participating in form construction received general information about the process and training. Once training was complete, participants began the form construction process by independently evaluating the items against the criteria on the provided checklists. Each participant completed their own checklist and had access to NWEA's Content Management System, which displayed the items corresponding to the order of items in the test. The educators initially reviewed single clusters of items and discussed each cluster as a group. Once they got used to the process, the educators reviewed the rest of the items followed by a discussion of each item. During this review, educators confirmed that there was only one correct answer for each multiple-choice item and that the items were aligned to the standard that it purported to address.

In both ELS and ILS, the educators, in consultation with NYSED, were permitted to recommend:

- revisions to the stated standard alignment,
- revisions to item sequencing to avoid cueing/clueing, and
- swapping any items and/or clusters of items that they judged as having problems flagged by the above reviews.

Given other constraints, it was not always possible to make every change that educators recommended, but they were given the opportunity to voice any and all concerns that they had. NYSED made the final decision about any educator recommendations.

The NYSED facilitators led a group discussion and helped the group reach consensus. Where time permitted, educators were presented with and approved the items that NYSED proposed for any necessary replacements. Following each session with educators, NYSED met to review the content and data of the proposed selections and explore alternate selections for consideration. NYSED then approved the item selections, including item positions within the test forms.

2.11. Test Form Production

Once the final forms were completed, the test forms were formatted for delivery via computer-based testing (CBT) by NWEA and were posted for NYSED to review. NYSED and NWEA reviewed the forms to look for any errors in formatting.

2.12. Final Eyes Committees

After NYSED and NWEA reviewed copies of the test forms, the test forms were reviewed by Final Eyes committees comprised of NYS educators. During that review, the educators were charged with taking the test to make sure that each MC item had a single correct answer and to look for errors in spelling, capitalization, punctuation, grammar, and formatting.

Following the Final Eyes review and after NYSED approved edits made as a result of the review, the tests were then considered final and produced for administration.

2.13. Standard Setting

The 2024 ELS and ILS Tests were the first administration based on the NYSP12SLS. In July 2024, after the operational administration of the 2024 tests, a standard setting meeting occurred in Albany, NY, where approximately 28 NYS educators (14 for ELS and 14 for ILS) went through a rigorous process (guided by the best practices indicated by this intensely studied process) to recommend updated performance standards for the NYSP12SLS. These recommendations were presented to the Commissioner of Education, who, in turn, adopted the recommended standards set forth by the committees. For additional details, see Section 8: Standard Setting and Appendix N: Standard Setting Report

Each test has four performance levels. Three cut points demarcate the performance levels needed to demonstrate each ascending level of performance. Section 6.4.5 contains the raw-to-scale score conversion tables, standard errors of measurement (SEMs), and detailed information related to the performance standards.

Section 3: Validity

The *Standards for Educational and Psychological Testing* refers to validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). Test validation is an ongoing process of gathering evidence from many sources to evaluate the soundness of the desired score interpretations or uses. This evidence is acquired from studies of the content of the test as well as studies involving scores produced by the test. Additionally, reliability must be taken into account before considerations of validity are made; a test cannot be valid if the test scores are not first reliable.

The *Standards for Educational and Psychological Testing* addresses the concept of validity in testing, which refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Validity is the most important consideration in test evaluation. Test validation is the process of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree to which evidence supports the inferences made from test scores.

3.1. Content Validity

Generally, achievement tests are used for student-level outcomes, either for making predictions about students or for describing students' performances (Mehrens & Lehmann, 1991). Tests are now also used for the purposes of accountability. Specific to student-level outcomes, the NYSTP documents student performance in science as defined by the New York State P-12 Science Learning Standards (NYSP12SLS).

For test score interpretations to be appropriate for this purpose, the content of the test must be carefully matched to the specified standards. The *Standards for Educational and Psychological Testing* states that content-related evidence of validity is a central concern during test development (AERA et al., 2014). Expert professional judgment should play an integral part in developing the definition of what is to be measured, such as describing the universe of the content, generating or selecting a content sample, and specifying the item format and scoring system.

Expert analysis of test content indicates the degree to which the content of a test covers the domain of content that the test is intended to measure. In the case of the New York State Testing Program (NYSTP), the content is defined by detailed blueprints that describe NYS content standards and define the skills that must be measured to assess these standards (see Appendix B). The NYSTP test development process requires specific attention to content representation and balance within each test form. NYS educators were involved in test construction at various development stages. For example, during the item review process, they reviewed field test items for alignment with the NYSP12SLS. They also participated in a process of establishing scoring rubrics for constructed-response items during rangefinding. Section 2: Test Design and Development contains more information specific to the item review process.

As a means of collecting further content validity evidence, a third-party alignment study was conducted by edCount, LLC in July 2024 to evaluate the degree to which the tests measure the content standards they are supposed to measure. See the *Alignment Evaluation for New York*

State Elementary- and Intermediate-level Science Tests for the full details of this alignment study.

3.2. Construct (Internal Structure) Validity

Construct validity (i.e., what scores mean and what kind of inferences they support) is often considered the most important type of test validity. Construct validity of the NYSTP Grades 5 and 8 Science Tests is supported by several types of evidence that can be obtained from the science test data.

3.2.1. Internal Consistency

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity because high coefficients imply that the test items measure the same domain of skill and are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section 7.1: Test Reliability. For the total population, the science reliability coefficients (Cronbach's alpha) ranged from 0.80 to 0.88. For all subgroups, the reliability coefficients were greater than or equal to 0.80, except for the English Language Learner (ELL) group. Overall, the high internal consistency of the NYSTP Grades 5 and 8 Science Tests provide sound evidence of construct validity.

3.2.2. Unidimensionality

Other validity evidence comes from analyses of the degree to which the test items conform to the requirements of the statistical models. These statistical models are used to scale and link the tests, as well as to generate student scores. The models require that the items fit the model well (item fit) and that the items in a test measure a single domain of skill (unidimensionality).

The first step is to assess the degree to which the items fit the item response theory (IRT) model. The item-model fit for the science tests was assessed using model-data fit plots, and the results are described in Section 6: IRT Calibration. Most items demonstrated sound fit across grades, except for one item in Grade 5 and two items in Grade 8. This provides solid evidence for the appropriateness of the IRT models used to calibrate and scale the test data.

Additional evidence for the efficacy of the model involves demonstrating that the items on the NYS tests are related to one another within their respective content areas. This relationship of the items within the science tests shows the common proficiency acquired by students studying the content area. This "common proficiency," or, more formally, underlying construct, could be labeled as science proficiency (using the science scores).

Factor analysis of the test data is one way of modeling the common construct. This analysis may show that there is a single, or main, factor that can account for much of the variability between responses to test items. A large first component in factor analysis would provide evidence of the latent proficiency that students have in common regarding the particular items. A large main factor found using this analysis would suggest a primary construct that may be related to what the items were designed to have in common (i.e., science proficiency).

To demonstrate the common factor underlying student responses to the science items, principal component factor analyses were conducted on a correlation matrix of individual items for the science tests. The study was conducted on NYS public, charter, and religious or independent school students for whom data were available. A large first principal component was evident in each analysis, demonstrating essential unidimensionality of the trait measured by each test. In other words, statistical evidence indicates that the science items are measuring one underlying construct: science proficiency.

The factor analyses conducted with the science data will show almost as many underlying constructs, or factors, as there are items on the test. Therefore, it is necessary to investigate the factor analysis results further to determine the number of "meaningful" factors. Specifically, more than one factor with an eigenvalue greater than 1.0 present in each dataset would suggest the presence of small additional factors (Kaiser, 1960). The magnitude of the ratio of the variance accounted for by the first factor compared with the remaining factors also provides evidence as to the number of meaningful factors (Cattell, 1966). In addition, the total amount of variance accounted for by the main factor was evaluated.

Factor analyses related to the Grades 5 and 8 Science Tests indicate that the ratio of the variance accounted for by the first factor to the remaining factors was sufficiently large to support the claim that the science test was essentially unidimensional. The science-related ratios show that the first eigenvalues were at least four times as large as the second eigenvalues for both grades.

Both the Grades 5 and 8 Science Tests exhibited the first principal component, accounting for more than 14.94% and 15.51% of the test variance, respectively. Table 3.1 presents the results of factor analyses, including eigenvalues greater than 1.0 and the proportions of variance explained by the extracted factors for science. The evidence in the table supports the claim that one single construct underlies the items/tasks in each science test and that scores from each test would represent performance primarily determined by that construct. Construct-irrelevant variance does not appear to create significant nuisance factors.

Table 3.1. Science Tests Factor Analysis

	Extracted Factor			
			Variance Accounted For	
Grade	N	Eigenvalue	%	Cumulative %
5	1	5.08	14.94	14.94
	2	1.21	3.55	18.49
	3	1.16	3.40	21.89
	4	1.05	3.09	24.98
	5	1.03	3.02	28.00
	6	1.01	2.97	30.97
8	1	8.22	15.51	15.51
	2	1.41	2.67	18.18
	3	1.16	2.18	20.36
	4	1.11	2.10	22.46
	5	1.05	1.98	24.44
	6	1.01	1.90	26.34

As additional evidence for construct validity, the same factor analysis procedure was employed to assess the dimensionality of the science construct for selected subgroups of students in each grade: ELLs, students with disabilities (SWD), and students using test accommodations (SUA). Appendix H provides the factor analysis results for these subgroup classifications. The results were comparable to those obtained from the total population data, except for Grade 5 ELLs who had a somewhat smaller first factor relative to the population.

Evaluation of the magnitude of the eigenvalue and proportion of variance explained by the main factor provide evidence of essential unidimensionality of the construct measured by the tests for these subgroups, with the Grade 5 ELLs having a somewhat smaller first factor.

3.2.3. Detection of Bias

Minimizing item bias means minimizing construct-irrelevant variance and helps establish a strong validity argument for the tests. Bias occurs if items function differentially for key pairs of groups, which may, in turn, cause a test to be differentially valid for certain groups of test takers. The statistical means for flagging items that may exhibit bias is referred to as differential item functioning (DIF). These statistical procedures were designed to be conservative (i.e., they were designed to flag more items for DIF rather than fewer). Therefore, it is rare in practice to observe a high-stakes test in which not a single item is flagged for DIF. Because these procedures tend to over-flag items, it is only through a review of those flagged items by experts that the items flagged for DIF may be judged to have or be free of bias. If the test involves irrelevant skills or knowledge, the possibility of bias is increased. Thus, preserving content validity is essential.

The developers of the NYSTP gave careful attention to items of possible ethnic, gender, socioeconomic status, and translation bias. All materials were written and reviewed to conform to the NYSED's editorial policies and guidelines for equitable assessment, as well as guidelines for item development. All materials were written to NYSED's specifications and carefully checked by groups of trained NYS educators during the item review process. These steps are essential in keeping bias to a minimum.

However, current evidence suggests that expertise in this area is no substitute for data; reviewers are sometimes wrong about which items work to the disadvantage of a group, apparently because some of their ideas about how students will react to items may be faulty (Jensen, 1980; Sandoval & Mille, 1980). Thus, empirical studies were conducted.

Statistical methods were employed to evaluate the amount of DIF in all test items. In 2024, all science test items were dichotomous and were therefore analyzed using Mantel-Haenszel (MH) method. In each grade, few items were flagged for DIF. See Section 5.4.3 for a summary of DIF results.

Section 4: Test Administration and Scoring

This section provides summaries of NYS test administration and scoring procedures. For further information, refer to the 2024 *NYSTP English Language Arts, Mathematics, and Science Tests School Administrator's Manual* (SAM) available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/sam-g3-8-2024.pdf and the 2024 NYSTP *Grades 3–8 English Language Arts, Mathematics, and Science Tests Scoring Leader Handbook* available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/3-8-scoring-leader-handbook-2024.pdf.

4.1. Test Administration

The ELS and ILS Tests were administered to students in a computer-based (CBT) testing mode. The CBT testing window for the ELS and ILS Tests was April 8–May 17, 2024.

4.2. Scoring Procedures of Operational Tests

Scoring of the 2024 ELS and ILS tests was conducted by NWEA in the ScorePoint system[1]. Operational tests contain multiple-choice (MC) and constructed-response (CR) items. All operational MC items were machine scored. This section describes the scoring of the operational CR items.

Qualified scoring directors oversaw the scoring process for the 2024 ELS and ILS Tests. Scoring directors are experts with years of experience managing professional scoring in the content areas. They provide leadership and management of the scoring process with special emphasis on training the scoring team leaders and scorers. Scoring team leaders, whose primary responsibility is to directly monitor the quality of scoring, have experience in scoring science content with at minimum a bachelor's degree in or related to the content being scored.

4.2.1. Scoring of Constructed-Response Items

The key resources used to train scorers on how to score student responses for CR items are scoring guides. These guides were created by NWEA from sets of actual field tested student responses that were consensus scored by NYSED and NYS teachers during rangefinding sessions. These materials are used to train NWEA scorers on the criteria for scoring CR items and rubric application. Additionally, the *Scoring Leader Handbook* for ELS and ILS provide guidelines, information, and procedures for both the NWEA scoring team leaders and scorers to facilitate scoring.

The CR items are divided into three groups for scoring, and a minimum of three separate scorers is necessary to score each CR item in the group they are assigned. After scoring is completed, the scoring director or scoring team leaders conduct read behinds for the scorers and items assigned to their scoring group.

[1] ScorePoint is NWEA's secure, online web-based scoring platform accessed through Google Chrome that allows scorers to access student constructed responses entered on the computer while protecting student data.

4.2.2. Scorer Qualifications and Training

Scoring guides are used to train scoring committee members on the criteria for scoring constructed-response items. Part of the training process is the administration of a consistency assurance set (CAS) that provides the scoring directors and team leaders with information regarding strengths and weaknesses of their scorers. This tool allows trainers to retrain their scorers, if necessary. The CAS also acknowledges those scorers who grasp all aspects of the content area being scored and are well prepared to score student responses.

4.2.3. Quality Control Process

Responses are randomly distributed throughout each scoring room so that completed tests from each region, district, school, or class are evenly dispersed. Scoring teams are divided into groups of three to ensure that a variety of scorers grade each test. If a scorer and a team leader cannot reach a decision after reviewing the scoring guides, they consult with a scoring director. If an issue is unable to be resolved, it is referred to NYSED for a scoring decision. A quality check is also performed to certify that all the items are scored and that the scores are appropriately entered into the system.

Section 5: Operational Test Data Collection and Classical Analysis

5.1. Data Collection

Test data were provided in a single phase. During this phase, the 100% student data file was provided to Pearson. The analyses described in Section 9: Summary of Operational Test Results were based on the data collected from the 100% student data file. Data collected from public, charter, and religious or independent schools were included in all data analyses.

5.2. Data Processing

Data processing refers to the cleaning and screening procedures used to identify errors (such as out-of-range data) and the decisions made to exclude student cases or to suppress particular items in certain analyses. Pearson's psychometric team performed data cleaning on the delivered data and excluded some student cases to obtain a sample of the utmost integrity. A student case being excluded from certain data analyses does not mean that the student record was invalidated. According to NYSED's specific instructions, additional procedures were taken to correct or recover these students' records so that their test results were scored properly. As mentioned above, their records were included in later analyses (see Section 9: Summary of Operational Test Results).

The largest group of cases excluded from the data set used for analyses (Sections 5, 6, and 7) was "Not Tested." These students were not tested for various reasons, including, for example, administrative error, not being enrolled at the time of the test, being medically excused, taking the New York State Alternate Assessment (NYSAA), being a first-year English Language Learner (ELL), or not attempting any test items. Other deleted cases included students with missing school type information, incorrect or incomplete grade information, duplicate records, no-response records, or mismatched form codes.

The data cleaning procedures and accompanying case counts are represented for science in Table 5.1 and Table 5.2.

Table 5.1. Science Grade 5 Data Cleaning

Exclusion Rule	#Deleted	#Cases Remaining
Initial Number of Cases	N/A	189,142
Missing Unique ID	0	189,142
Not Tested	32,587	156,555
Incorrectly Translated Forms	3,270	153,285
Duplicate Records	0	153,285
Missing Raw Score	0	153,285

Table 5.2. Science Grade 8 Data Cleaning

Exclusion Rule	#Deleted	#Cases Remaining
Initial Number of Cases	N/A	183,864
Missing Unique ID	0	183,864

Exclusion Rule	#Deleted	#Cases Remaining
Not Tested[2]	93,728	90,136
Incorrectly Translated Forms	210	89,926
Duplicate Records	0	89,926
Missing Raw Score	0	89,926

5.3. Classical Analysis and Calibration Sample Characteristics

The cleaned data were used for classical analyses and calibration. The demographic characteristics of students in these data sets are presented in Table 5.3 and Table 5.4, including gender, ethnicity, Needs Resource Capacity (NRC) category, ELL status, students with disabilities (SWDs), students using test accommodations (SUAs), SWD/SUA (includes students who are classified as having a disability and who use at least one disability-related accommodation), and ELLs using accommodations specific to their ELL status (ELL/SUA). The NRC category is assigned at the district level and is an indicator of district and school socioeconomic status. The ethnicity and gender designations are based on student-level information.

Table 5.3. Science Grade 5 Sample Characteristics

Demographic Category		N-Count	% of Total N-Count
Gender	Female	74,818	48.81
	Male	78,448	51.18
	Non-Binary	19	0.01
Ethnicity	American Indian or Alaska Native	1,232	0.81
	Asian	17,245	11.30
	Black or African American	24,779	16.23
	Hispanic or Latino	41,066	26.91
	Multiracial	5,702	3.74
	Native Hawaiian or Pacific Islander	369	0.24
	White	62,235	40.78
NRC	New York City	46,130	30.62
	Big 4 Cities	6,010	3.99
	Urban/Suburban	10,509	6.98
	Rural	8,942	5.94
	Average Needs	41,104	27.28
	Low Needs	19,089	12.67
	Charter	14,163	9.40
	Religious or Independent	4,715	3.13
SWD	No	127,570	83.22
	Yes	25,715	16.78
SUA	No	124,104	80.96
	Yes	29,181	19.04
ELL	No	143,543	93.64
	Yes	9,742	6.36
SWD/SUA	No	131,480	85.77

[2] The number of students "Not Tested" was larger here than for Grade 5 due to some students taking a Regents exam instead of the ILS test.

Demographic Category		N-Count	% of Total N-Count
	Yes	21,805	14.23
ELL/SUA	No	150,449	98.15
	Yes	2,836	1.85

Note. The total n-count was 153,285.

Table 5.4. Science Grade 8 Sample Characteristics

Demographic Category		N-Count	% of Total N-Count
Gender	Female	41,855	46.54
	Male	48,034	53.42
	Non-Binary	37	0.04
Ethnicity	American Indian or Alaska Native	693	0.78
	Asian	7,071	7.92
	Black or African American	16,499	18.49
	Hispanic or Latino	27,269	30.55
	Multiracial	2,702	3.03
	Native Hawaiian or Pacific Islander	187	0.21
	White	34,827	39.02
NRC	New York City	25,730	29.37
	Big 4 Cities	3,518	4.02
	Urban/Suburban	8,673	9.90
	Rural	7,221	8.24
	Average Needs	23,804	27.17
	Low Needs	6,875	7.85
	Charter	7,962	9.09
	Religious or Independent	3,813	4.35
SWD	No	71,905	79.96
	Yes	18,021	20.04
SUA	No	68,271	75.92
	Yes	21,655	24.08
ELL	No	83,048	92.35
	Yes	6,878	7.65
SWD/SUA	No	74,968	83.37
	Yes	14,958	16.63
ELL/SUA	No	86,621	96.32
	Yes	3,305	3.68

Note. The total n-count was 89,926.

5.4. Classical Data Analysis

Classical data analysis of the NYSTP Grades 5 and 8 Science Tests consists of several important elements. One element is the analysis of item-level statistical information about student performance. It is important to verify that the items and test forms function as intended. If any serious error were to occur with an item, errors should be flagged and evaluated for rectification (suppression, credit, or other acceptable solution) during item analysis. Analyses of test-level data comprise the second element of classical data analysis. These include examination of the raw score statistics (mean and standard deviation, or "SD") and the test reliability measures Cronbach's alpha (Cronbach, 1951) and the Feldt-Raju coefficient (Qualls, 1995). Classical differential item functioning (DIF) analysis is also conducted at this stage. DIF analysis includes the computation of Mantel-Haenszel statistics for NYS science items to identify potential item bias. All classical data analysis results contribute information on the validity and reliability of the tests (see also Section 3: Validity and Section 7: Reliability and Standard Error of Measurement).

5.4.1. Item Difficulty and Point-Biserial Correlation Coefficients

Item difficulty is classically measured by the *p*-value statistic. It assesses the proportion of students who responded correctly to each 1-point dichotomous item or the average proportion of the maximum score that students earned on each polytomous item. Point-biserial statistics are used to examine item-test correlations or item discrimination. Examining *p*-values and point-biserial correlations can identify item flaws such as wrong keys. This procedure was used to check the operational data. Items are flagged for review by a subject matter expert according to the criteria listed in Table 5.5. The number of 2024 operational items flagged for science in each grade is given in Table 5.6.

Table 5.5. Item Analysis Flagging Criteria

Dichotomous Items Only	• Low percentage receiving a score point (< 0.30) • Positive point-biserial (> 0) for 1 or more distractor(s) • Point-biserial correlation for distractor is greater than for key
Polytomous Items Only	• N/A—all items are dichotomous.

Table 5.6. Number of Flagged Items

			#Flagged Items	
Subject	**Grade**	**#Items**	***P*-Value**	**Point-Biserial**
Science	5	34	1	5
	8	53	4	9

If a multiple-choice (MC) item is flagged, a subject matter expert reviews the item and intended key to verify that the item was scored correctly. Choices are checked to verify that one and only one correct answer exists. If a constructed-response (CR) item is flagged, a subject matter expert reviews the item to ensure that all components are present (e.g., art was not omitted) and the item is clearly worded. If no defects are found in a flagged item, a subject matter expert may suggest a reason for the statistical flag, if apparent.

It is important to have a good range of *p*-values to increase test reliability and avoid floor or ceiling effects. *P*-values represent the overall degree of difficulty but do not account for

demonstrated student performance on other test items. Usually, *p*-value information is coupled with point-biserial correlations to verify that items are functioning as intended.

The summary statistics of the item difficulty (*p*-values) and item discrimination (point-biserial correlations) for the operational tests are shown in Table 5.7 and Table 5.8. The data show a reasonably wide range of item difficulties for each test. For the Grades 5 and 8 Science Tests, the mean item difficulties ranged from 0.33 to 0.37, and point-biserial correlations ranged from 0.00 to 0.56. The mean point-biserial correlations ranged from 0.36 to 0.37.

Table 5.7. Item Difficulty Distribution

Subject	Grade	N-Count	Mean	SD	Min.	Max.
Science	5	153,285	0.37	0.15	0.07	0.66
	8	89,926	0.33	0.18	0.05	0.80

Table 5.8. Item Discrimination Distribution

Subject	Grade	N-Count	Mean	SD	Min.	Max.
Science	5	153,285	0.36	0.11	0.04	0.53
	8	89,926	0.37	0.12	0.00	0.56

Appendix I provides classical test statistics for all items at each grade.

5.4.2. Omit Rates

Omit rates (i.e., the percentage of students not answering a given item) are routinely checked, based on test data, after each administration. Appendix I shows the omit rates for items on the Grades 5 and 8 Science Tests. The industry standard general rule is that omit rates for MC items should be less than 5%; omit rates for items on the Grades 5 and 8 Science Tests were less than 1.2%.

5.4.3. Differential Item Functioning (DIF)

Classical DIF analyses are statistical methods for identifying items that are estimated to have functioned differently for one group (i.e., the "focal" group) as compared with another group (i.e., the "reference" group). In other words, DIF analysis only *flags* items that may later be judged by content experts to exhibit bias rather than *directly detecting* bias. The psychometric phenomenon of DIF has been extensively investigated, and experts' judgments of bias were collected when items were field tested, which reduced the likelihood of including any differentially functioning items on the operational forms. DIF was evaluated for the science operational items using the Mantel-Haenszel Delta method (Dorans & Holland, 1992) for dichotomous items. Please refer to the *New York State Testing Program 2024: Elementary- and Intermediate-Level Science Grades 5 & 8 Field Test Technical Report* for details about these DIF methods and item-flagging criteria. Operational items flagged for DIF are given additional scrutiny by content specialists (above and beyond the existing rounds of reviews by NYS educators) to identify potential systematic issues that could be addressed in future item writing.

Section 6: IRT Calibration

6.1. IRT Models and Rationale for Use

Item response theory (IRT) allows for comparisons between items and scale scores, even those from different test forms, by using a common scale for all items and students (i.e., as if there were a hypothetical test that contained items from all forms).

IRT is a set of statistical models that attempt to relate observed responses to items on a test to latent traits. In the case of educational tests, the latent trait of interest is often students' mastery of a particular discipline, such as science. Computer programs that implement IRT models use student data to estimate the characteristics of the items on a test, called "parameters." The parameter estimation process is called "item calibration."

IRT models typically vary according to the number of parameters estimated. For the NYS tests, two types of item parameters are estimated: the discrimination parameter and the difficulty parameters. The discrimination parameter is an index of how well an item differentiates between high-performing and low-performing students. An item that cannot be answered correctly by low-performing students but can be answered correctly by high-performing students will have a high discrimination value. The difficulty parameter is an index of how easy or difficult an item is. The higher the difficulty parameter, the harder the item.

The Grades 5 and 8 Science Tests contain dichotomous items only. As such, all item parameters for science are estimated using the two-parameter logistic (2PL) model (Lord, 1980; Lord & Novick, 1968) that was adopted in 2024 for analyzing dichotomous items. In this model, the probability that a student with proficiency θ responds correctly to item i is:

$$P_i(\theta) = \frac{1}{1 + exp\left(-Da_i(\theta - b_i)\right)}$$

where D is a scaling constant of 1.7, and a_i and b_i are the discrimination and difficulty parameters of item i, respectively.

6.2. Calibration Sample

The cleaned data were used to calibrate the New York State Testing Program (NYSTP) 2024 Grades 5 and 8 Science Tests. Calibration sample sizes were adequate, as the calibration was performed using nearly all the NYS public and non-public school student population data in each grade. Table 6.1 shows the percentage of the 2024 operational test samples by demographic group for the Grades 5 and 8 Science Tests, respectively. The subgroups include gender, ethnicity, Needs Resource Capacity (NRC) category, English Language Learner (ELL) status, students with disabilities (SWDs), students using test accommodations (SUAs), SWD/SUA (includes students who are classified as having a disability and who use at least one disability-related accommodation), and ELLs using accommodations specific to their ELL status (ELL/SUA).

Table 6.1. Science Grades 5 and 8 Demographic Statistics

Demographic Category		Grade 5 2024 Sample	Grade 8 2024 Sample
Gender	Female	48.81	46.54
	Male	51.18	53.42
	Non-Binary	0.01	0.04
Ethnicity	Asian	0.81	0.78
	African American	11.30	7.92
	Hispanic	16.23	18.49
	American Indian	26.91	30.55
	Multiracial	3.74	3.03
	Pacific Islander	0.24	0.21
	White	40.78	39.02
NRC	New York City	30.62	29.37
	Big 4 Cities	3.99	4.02
	Urban/Suburban	6.98	9.90
	Rural	5.94	8.24
	Average Needs	27.28	27.17
	Low Needs	12.67	7.85
	Charter	9.40	9.09
	Religious or Independent	3.13	4.35
SWD	No	83.22	79.96
	Yes	16.78	20.04
SUA	No	80.96	75.92
	Yes	19.04	24.08
ELL	No	93.64	92.35
	Yes	6.36	7.65
SWD/SUA	No	85.77	83.37
	Yes	14.23	16.63
ELL/SUA	No	98.15	96.32
	Yes	1.85	3.68

6.2.1. Calibration Process

Item parameters were estimated using Scientific Software International (SSI) Inc.'s IRTPRO Version 6.0 (Cai et al., 2022) package. Dichotomous items were calibrated simultaneously using marginal maximum likelihood procedures.

The calibration of the NYSTP 2024 Grades 5 and 8 Science Tests did not exhibit any test-level issues. The estimated parameters were on the original theta scale, and all items were well within the prescribed parameter ranges except for a few, such as one in Grade 5 (a = -0.06) and one in Grade 8 (a = -0.09). Overall, all calibration estimation results were reasonable for the Grades 5 and 8 Science Tests. Table 6.2 presents the summaries of the calibration results for science. Additional details, including individual item parameter estimates, can be found in Appendix J. The parameter estimates are expressed on the theta metric and are defined as follows for the dichotomous items: a is a discrimination parameter and b is a difficulty parameter.

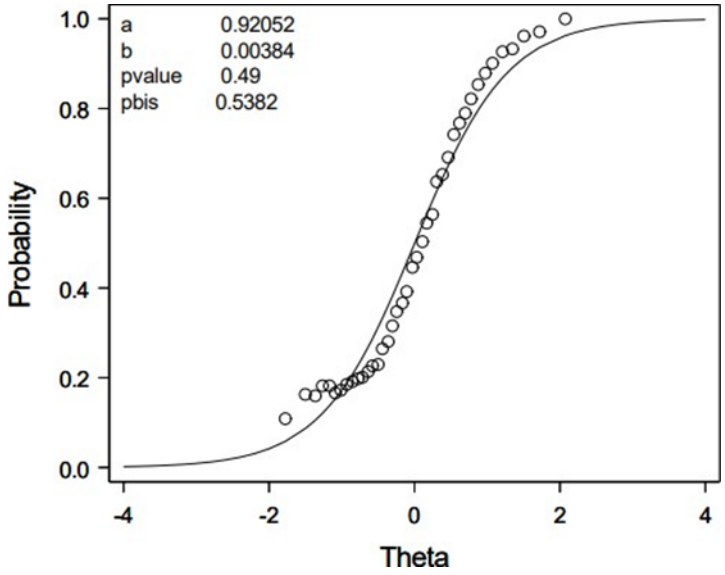
Table 6.2. Science Calibration Results

Grade	N-Count	Range of *a*-Parameters		Range of *b*-Parameters	
5	153,285	-0.06	1.15	-5.54	3.76
8	89,926	-0.09	1.13	-4.47	5.10

6.3. Item-Model Fit

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) suggest documenting evidence of model fit when model-based methods such as IRT are used to estimate item parameters in test development. The standard process of assessing the fit of an item under unidimensional IRT models involves steps, such as (a) defining a number of student groups ("buckets") and then (b) making an informed judgment by comparing the observed and model-predicted proportion-correct scores for the item by the students in different "buckets" (Sinharay, 2006). To make this judgment on each item, Hambleton and Swaminathan (1985) recommend the use of graphical plots comparing the estimated/predicted item response function to the empirical student-response data for an item. An example item fit plot is shown in Figure 6.1.

Figure 6.1. Example Item Fit Plot



Fit plots were produced and closely examined for all operational items to visually examine the model-data fit for each item. All items showed adequate model-data fit except for one item in Grade 5 and two items in Grade 8. This further supports the use of the chosen IRT models.

6.4. Scaling and Scoring Procedure

The 2024 Grades 5 and 8 Science Tests are new assessments developed based on the New York State P-12 Science Learning Standards (NYSP12SLS), which are different from previous content standards. Even though there is overlap between the old and new standards, there are significant content shifts and depth of learning changes. The 2024 Grades 5 and 8 Science Tests also have new item formats that led to changes in test specifications. The *Standards for Education and Psychological Testing* states that "When substantial changes in test specifications occur, scores should be reported on a new scale, or a clear statement should be provided to alert users that the scores are not directly comparable with those on earlier versions of the test" (AERA et al., 2014, p. 107). Being the first administration of the NYSTP tests to measure the NYSP12SLS, a new reporting scale was established following the standard setting meeting in Summer 2024. The reporting scale was developed to quantify the information captured by the assessment about what students know and can do. The reporting scale was developed to interpret changes, make comparisons, facilitate inferences, and inform educational decisions.

NYS student assessments were scored using the number correct (NC) scoring method. This method considers how many score points a student obtained on a test in determining their reported score, also called a scale score (i.e., two students with the same number of score points on the test will receive the same scale score, regardless of which items they answered correctly). In this method, the number correct (or "raw") score on the test is converted to a scale score by means of a conversion table.

6.4.1. Raw-Score-to-Theta-Score Conversion Tables

To create a raw-score-to-scale-score (RSSS) table, each raw score is first converted to a theta score that represents the student's proficiency under the IRT model. An inversed test characteristic curve (TCC) procedure is used to obtain the theta estimates. These estimates show negligible statistical bias (defined in statistics as the difference between an estimator's expected value and the true value of the parameter being estimated) for tests with maximum possible raw scores of at least 30 points. Both the Grade 5 and 8 NYSTP Science Tests have a maximum raw score higher than 30 points. In the inverse TCC method, a student's trait (i.e., proficiency) estimate is taken to be the trait value that has an expected raw score equal to the student's observed raw score. It was found that for tests containing only dichotomous items, the inverse of the TCC is an excellent first-order approximation of the number of correct maximum likelihood estimates (MLE), showing negligible bias for tests of at least 30 points (Yen, 1984).

The inverse TCC method relies on the following equation:

$$\sum_{i=1}^{n} v_i x_i = \sum_{i=1}^{n} v_i E\left(X_i|\tilde{\theta}\right)$$

where

- x_i is a student's observed raw score on item *i*,

- v_i is a non-optimal weight specified in a scoring process (v_i = 1 if no weights are specified), and
- $\tilde{\theta}$ is a trait estimate.

6.4.2. Theta Adjustments

With the adoption of the 2PL model, the θ scores can be obtained for all raw score points, except the zero, and perfect scores using the inverse TCC method. However, the θ scores at the two ends of the scale are much less reliable, as indicated by the large conditional standard errors of measurement (CSEMs). Therefore, an adjustment and interpolation were conducted to derive the adjusted theta scores following the rules, as outlined in Table 6.3.

Table 6.3. Smoothing Rules

Subject	Grade	Smoothing	
		Starting Point	Step Size
Science	5	CSEM > 0.56	0.16
	8	CSEM > 0.56	0.16

At both ends of the scale, for any theta estimates with CSEMs greater than 0.56 for science, 0.16 was subtracted (at the low end) or added (at the high end) from the preceding theta value. Table 6.4 shows an example of smoothing at the two ends of the science tests.

Table 6.4. Example of Smoothing in Raw-Score-to-Theta-Score Table

Science Grade 5				Science Grade 8			
Raw Score	Estimated Theta	CSEM of Theta	Adjusted Theta	Raw Score	Estimated Theta	CSEM of Theta	Adjusted Theta
0	–	–	-2.2422	0	–	–	-2.7999
1	-7.2486	3.9957	-2.0822	1	-8.9618	5.0931	-2.6399
2	-3.9256	1.5135	-1.9222	2	-4.4245	1.6352	-2.4799
3	-2.8426	1.0049	-1.7622	3	-3.2381	1.0040	-2.3199
4	-2.2103	0.7856	-1.6022	4	-2.5969	0.7477	-2.1599
5	-1.7633	0.6647	-1.4422	5	-2.1659	0.6144	-1.9999
6	-1.4136	0.5891	-1.2822	6	-1.8399	0.5350	-1.8399
7	-1.1222	0.5380	-1.1222	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	45	3.5504	0.5129	3.5504
27	2.8016	0.5381	2.8016	46	3.8350	0.5802	3.7104
28	3.1314	0.6146	2.9616	47	4.1887	0.6771	3.8704
29	3.5519	0.7363	3.1216	48	4.6526	0.8259	4.0304
30	4.1352	0.9460	3.2816	49	5.3105	1.0757	4.1904
31	5.0597	1.3641	3.4416	50	6.3635	1.5548	4.3504
32	6.9263	2.4372	3.6016	51	8.4643	2.7140	4.5104
33	20.5569	18.2879	3.7616	52	27.1730	35.7329	4.6704
34	–	–	3.9216	53	–	–	4.8304

Note. Theta and CSEM values are not shown for zero and perfect scores because these values cannot be obtained using the inverse TCC method.

6.4.3. Mean and Standard Deviation of Adjusted Theta Scores

The mean and standard deviation (SD) of the theta scores were computed from the 2024 Grades 5 and 8 Science calibration sample, as summarized in Table 6.5.

Table 6.5. Mean and Standard Deviation of Adjusted Theta Scores

Subject	Grade	Mean	SD
Science	5	0.0306185	1.0520847
	8	-0.0055407	1.0720957

6.4.4. Scaling Coefficients

The adjusted θ scores were converted to scale scores using a linear transformation by fixing two desired properties: the Level 3 cut score and the SD of scale scores (as shown in Table 6.6). The scale score of 450 was chosen as the desired Level 3 cut score so that the scale score ranges of the new 2024 scale would not overlap with previous Grades 5 and 8 Science Tests or other NYSTP tests. The desired SD of scale scores was set as 20 for both Grades 5 and 8 Science.

Table 6.6. Level 3 Cut Score and Standard Deviation of Scale Scores

		Scaling	
Subject	Grade	Level 3 Cut	Standard Deviation
Science	5	450	20
	8	450	20

The scaling slope and intercept are computed as follows:

$$Slope = \frac{\sigma(ScaleScore)}{\sigma(\theta)},$$

$$Intercept = cut(ScaleScore) - \frac{\sigma(ScaleScore)}{\sigma(\theta)} cut(\theta)$$

where $\sigma(ScaleScore)$ is the desired standard deviation of scale scores (20 for both Grades 5 and 8 Science); $\sigma(\theta)$ is the standard deviation of the adjusted theta scores based on the calibration sample; $cut(ScaleScore)$ is 450 for both Grades 5 and 8 Science; and $cut(\theta)$ is the theta score in the raw-to-theta conversion table that corresponds to the Level 3 cut score obtained from standard setting. Table 6.7 shows the resulting scaling coefficients for Grades 5 and 8 Science.

After smoothing the θ scores at the ends of the scale, the adjusted θ scores were obtained. The adjusted CSEMs were then computed. The scaling coefficients in Table 6.7 were then applied to the adjusted θ scores to obtain the corresponding scale scores using the equation below.

$$\text{Scale Score} = M_1^S \theta + M_2^S$$

The final RSSS tables could then be established. Scaling coefficients, M_1^S and M_2^S, were determined during the 2024 standard setting and will be used in subsequent administrations. Note that comparing scale scores across tests of different subjects or grades is not appropriate, as each test has different content specifications and does not use the same scale.

Table 6.7. Operational Scaling Coefficients

Subject	Grade	Slope (M_1^S)	Intercept (M_2^S)
Science	5	19.00988	440.96537
	8	18.65505	442.84896

6.4.5. RSSS Conversion Tables, TCCs, CSEMs, and Performance Levels

The scale score is the reported score for the NYSTP. The RSSS conversion tables are presented in Appendix L.

Test characteristic curves provide an overview of the tests in the IRT scale score metric. The 2024 TCCs were generated using final item parameters for all reporting test items administered in Spring 2024. TCCs are the summation of all the item characteristic curves (ICCs) contributing to the scale scores. The TCC plots for the science tests are presented in Appendix M.

The CSEM of a scale score is calculated as follows and is included in the RSSS table:

$$\text{CSEM}(\text{Scale Score}) = M_1^S \frac{1}{\sqrt{I(\hat{\theta})}}$$

where $\hat{\theta}$ theta estimate corresponding to the scale score, $I(\hat{\theta})$, is the value of the test information function (TIF) at $\hat{\theta}$, and M_1^S is the scaling coefficient in Table 6.7.

Scale score cuts were set in Summer 2024 through standard setting and can be applied to the future scale scores. See Section 8 for information on the standard setting process for Grades 5 and 8 Science.

The following procedure is conducted on an RSSS table to ensure that all cut scores are obtainable: If no rounded scale score matches a given scale score cut, the nearest available score below the cut is adjusted to match the cut score. For example, if the cut score of interest is 450 and only scale scores of 449 and 451 are obtainable (before adjustment), the scale score of 449 would be adjusted to 450 and the scale score of 451 would remain unaltered. The final element of the RSSS tables is the application of the performance level cut scores.

Table 6.8 presents scale score ranges associated with each performance level for science.

Table 6.8. Science Scale Score Ranges Associated with Each Performance Level

Grade	NYS Level 1	NYS Level 2	NYS Level 3	NYS Level 4
5	398–423	424–449	450–479	480–516
8	391–427	428–449	450–479	480–533

6.5. CSEMs

Conditional standard error of measurement curves graphically show the amount of measurement error at different ability levels. The CSEM curves are presented in Figure 6.2 and Figure 6.3.

Figure 6.2. Science Grade 5 CSEM Curve

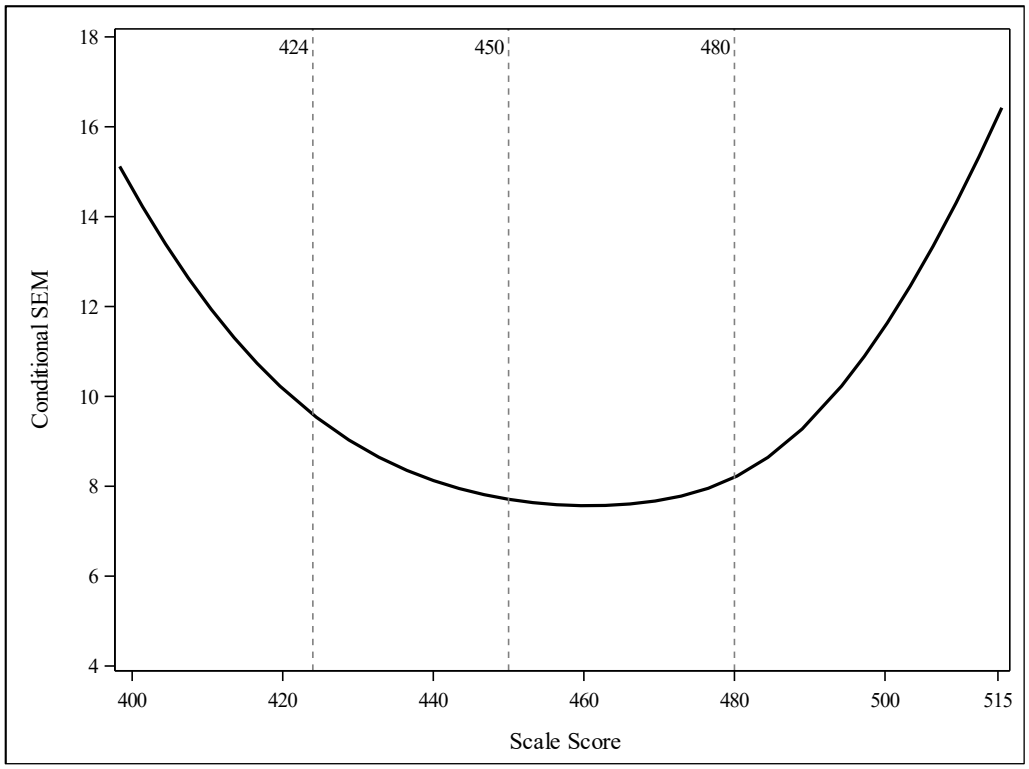
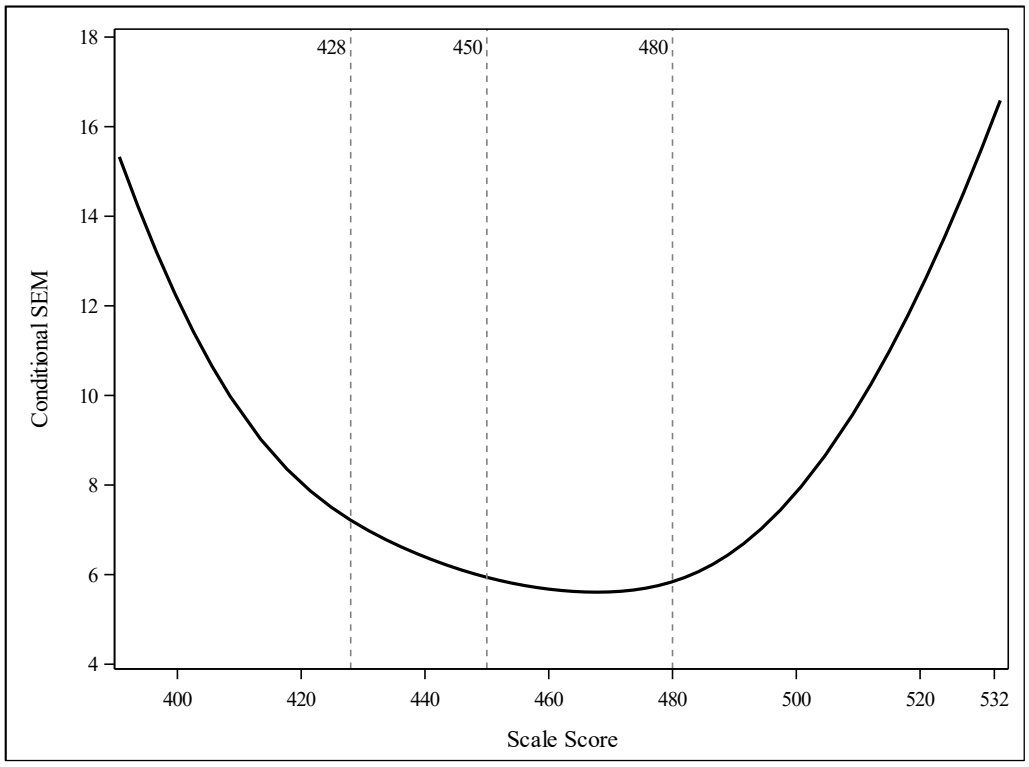


Figure 6.3. Science Grade 8 CSEM Curve



Section 7: Reliability and Standard Error of Measurement

This section presents information on various test reliability statistics, standard errors of measurement (SEMs), and the results of performance level classification accuracy and consistency analyses. The data set for these analyses includes NYS students who were tested and received valid scores.

7.1. Test Reliability

Test reliability is directly related to score stability and standard error and, as such, is an essential element of fairness and validity. Test reliability can be directly measured with an alpha statistic, and the alpha statistic can be used to derive the SEM. For the Grades 5 and 8 Science Tests, Pearson calculated two types of reliability statistics: Cronbach's alpha (Cronbach, 1951) and the Feldt-Raju coefficient (Qualls, 1995). These two measures are appropriate for assessing the internal consistency of a test when a single test is administered to a group of students on one occasion. The reliability of the test is then estimated by considering how well the items reflecting the same construct yield similar results (or how consistent the results are for different items that reflect the same construct measured by the test). Both Cronbach's alpha and the Feldt-Raju coefficient measures are appropriate for tests consisting of multiple item formats (MC and CR items).

7.1.1. Test Statistics and Reliability for Total Test

Table 7.1 presents the test statistics, including raw score means and raw score standard deviations (SDs) for the Grades 5 and 8 Science Tests. Table 7.2 presents the case counts ("N-Count"), number of test items ("#Items"), Cronbach's alpha and associated SEM, and the Feldt-Raju coefficient and associated SEM obtained for the total science tests. Reliability coefficients provide measures of internal consistency that range from 0 to 1. High reliability indicates that scores are consistent and not unduly influenced by random error. The total test reliability is a very good indication of each test's internal consistency.

Grades 5 and 8 Science reliability estimates (Cronbach's alpha and Feldt-Raju coefficient) ranged from 0.77 to 0.88 across both grades, which is a good indication that the New York State Testing Program (NYSTP) Grades 5 and 8 Science Tests are acceptably reliable.

Table 7.1. Science Test Form Statistics

	Item-Level			Student-Level			
	P-Value				Raw Score		
Grade	Mean	Min.	Max.	N-Count	Max	Mean	SD
5	0.37	0.07	0.66	153,285	33	12.74	5.67
8	0.33	0.05	0.80	89,926	51	17.65	8.50

Table 7.2. Science Test Reliability and Standard Error of Measurement

			Raw Score	Cronbach's Alpha		Feldt-Raju Coefficient	
Grade	N-Count	#Items	Points	Est.	SEM	Est.	SEM
5	153,285	34	34	0.80	2.55	0.77	2.70
8	89,926	53	53	0.88	2.99	0.86	3.18

7.1.2. Reliability by Item Type

In addition to overall test reliability, Cronbach's alpha and the Feldt-Raju coefficient were computed separately for MC and CR item sets. Reliability is directly affected by test length; therefore, reliability estimates for tests by item type will always be lower than reliability estimates for the overall test form. Table 7.3 presents reliabilities for the subsets of MC items, and Table 7.4 presents reliabilities for the subsets of CR items.

Table 7.3. Science MC Item Reliability and Standard Error of Measurement

Grade	N-Count	#Items	Raw Score Points	Cronbach's Alpha		Feldt-Raju Coefficient	
				Est.	SEM	Est.	SEM
5	153,285	19	19	0.60	2.05	0.57	2.13
8	89,926	29	29	0.75	2.43	0.73	2.56

Table 7.4. Science CR Item Reliability and Standard Error of Measurement

Grade	N-Count	#Items	Raw Score Points	Cronbach's Alpha		Feldt-Raju Coefficient	
				Est.	SEM	Est.	SEM
5	153,285	15	15	0.75	1.51	0.70	1.66
8	89,926	24	24	0.83	1.73	0.79	1.90

Note. Results should be interpreted with caution because the number of items is small.

7.1.3. Test Reliability for Subgroups

In this section, reliability coefficients that were estimated for the population and subgroups are presented. The subgroups include the following: gender, ethnicity, Needs Resource Capacity (NRC) category, English Language Learner (ELL) status, all students with disabilities (SWDs), all students using test accommodations (SUAs), SWD/SUA (includes students who are classified as having a disability and who use at least one disability-related accommodation), and ELLs using accommodations specific to their ELL status (ELL/SUA). Accommodations available to students include Flexibility in Scheduling/Timing, Flexibility in Setting, Method of Presentation (excluding braille), Method of Response, Braille and Large type, and others (IEP or 504 Plan). Accommodations available to ELLs are Separate Location and Bilingual Dictionary.

As shown in Table 7.5 and Table 7.6, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Except for the ELL group, Cronbach's alpha reliability coefficients were all at least 0.72, and the Feldt-Raju reliability coefficients were at least 0.70. These indicate a very good internal test consistency (reliability) for the analyzed subgroups of students.

Table 7.5. Science Grade 5 Test Reliability by Subgroup

Demographic Category		N-Count	Cronbach's Alpha		Feldt-Raju Coefficient	
			Est.	SEM	Est.	SEM
State	All Items	153,285	0.80	2.55	0.77	2.70
Gender	Female	74,818	0.79	2.56	0.76	2.69
	Male	78,448	0.81	2.54	0.78	2.69
	Non-Binary	19	0.81	2.56	0.79	2.72
Ethnicity	Asian	17,245	0.81	2.59	0.79	2.75
	African American	24,779	0.76	2.49	0.74	2.60
	Hispanic	41,066	0.75	2.51	0.73	2.62
	American Indian	1,232	0.79	2.51	0.77	2.65
	Multiracial	5,702	0.82	2.55	0.80	2.72
	Pacific Islander	369	0.80	2.55	0.78	2.70
	White	62,235	0.79	2.58	0.76	2.72
NRC	New York City	46,130	0.80	2.54	0.78	2.69
	Big 4 Cities	6,010	0.76	2.41	0.74	2.53
	Urban/Suburban	10,509	0.75	2.49	0.73	2.60
	Rural	8,973	0.77	2.53	0.74	2.65
	Average Needs	41,013	0.78	2.57	0.75	2.70
	Low Needs	19,089	0.79	2.62	0.76	2.75
	Charter	14,140	0.79	2.55	0.77	2.68
	Religious or Independent	4,720	0.76	2.60	0.74	2.72
SWD	All Codes	25,715	0.74	2.39	0.72	2.49
SUA	All Codes	26,199	0.74	2.40	0.72	2.50
ELL	ELL	9,742	0.59	2.34	0.58	2.39
SWD/SUA	SWD and SUA Codes	21,805	0.72	2.37	0.70	2.46
ELL/SUA	ELL and SUA Codes	2,836	0.59	2.36	0.57	2.41

Table 7.6. Science Grade 8 Test Reliability by Subgroup

Demographic Category		N-Count	Cronbach's Alpha		Feldt-Raju Coefficient	
			Est.	SEM	Est.	SEM
State	All Items	89,926	0.88	2.99	0.86	3.18
Gender	Female	41,855	0.87	3.01	0.85	3.19
	Male	48,034	0.88	2.97	0.87	3.18
	Non-Binary	37	0.88	3.14	0.86	3.34
Ethnicity	Asian	7,071	0.90	3.08	0.89	3.34
	African American	16,499	0.85	2.92	0.83	3.07
	Hispanic	27,269	0.85	2.95	0.83	3.10
	American Indian	693	0.85	2.93	0.84	3.09
	Multiracial	2,702	0.88	2.98	0.86	3.18
	Pacific Islander	187	0.90	3.01	0.88	3.24
	White	34,827	0.88	3.04	0.86	3.23
NRC	New York City	25,644	0.88	2.97	0.87	3.18
	Big 4 Cities	3,518	0.74	2.75	0.73	2.82
	Urban/Suburban	8,673	0.83	2.90	0.82	3.03
	Rural	7,221	0.86	2.98	0.84	3.15
	Average Needs	23,749	0.86	3.01	0.84	3.18
	Low Needs	6,875	0.89	3.09	0.87	3.31
	Charter	7,962	0.87	3.04	0.85	3.23
	Religious or Independent	3,815	0.89	3.12	0.87	3.34
SWD	All Codes	18,021	0.80	2.81	0.78	2.91
SUA	All Codes	18,205	0.82	2.83	0.80	2.94
ELL	ELL	6,878	0.65	2.73	0.63	2.77
SWD/SUA	SWD and SUA Codes	14,958	0.78	2.79	0.76	2.88
ELL/SUA	ELL and SUA Codes	3,305	0.63	2.73	0.61	2.77

7.2. Standard Error of Measurement (SEM)

Table 7.2 presented the SEMs computed from Cronbach's alpha and the Feldt-Raju reliability statistics for science. The SEMs ranged from 2.55 to 3.18 across grades and the two estimation methods, which were reasonable and small. The SEMs are directly related to reliability: the higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so the SEMs were expected to be low.

The SEMs for the subpopulations, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, were presented in Table 7.5 and Table 7.6. The SEMs associated with all reliability estimates across grades, estimation methods, and subpopulations, except for the ELL group, ranged from 2.34 to 3.34, which were close to those for the entire population. This narrow range indicates that all students' test scores are reasonably reliable across the Grades 5 and 8 Science Tests with minimal error.

7.3. Performance Level Classification Consistency and Accuracy

Classification consistency refers to the estimated degree of agreement between students' performance classification from two independent administrations of the same test (or from two parallel forms of the test). Because obtaining test scores from two independent administrations of NYS tests was not feasible due to item release after each administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Classification accuracy can be defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston & Lewis, 1995).

In conjunction with measures of internal consistency, classification consistency is an important type of reliability and is particularly relevant to high-stakes tests. As a form of reliability, classification consistency represents the extent to which a student's performance classification is expected to remain the same over repeated measurements.

Classification consistency is most relevant for students whose performance is near the proficiency cut score. For example, consider the cut score delineating Levels 2 and 3, or simply the "Level 3 cut." Students whose proficiency is far above or far below that cut score are unlikely to be misclassified because repeated administration of the test will nearly always result in the same classification. Students whose true scores are close to the cut score are a more serious concern. These students' true scores will likely lie within the SEM of the cut score. For this reason, the measurement error at the cut scores should be considered when evaluating the classification consistency of a test. Furthermore, the number of students near the cut scores should also be considered when evaluating classification consistency, as these numbers show the number of students who are at risk of being misclassified.

Scoring tables with SEMs and student scale score frequency distributions are located in Appendix L. Classification consistency and accuracy were estimated using the item response theory (IRT) procedure suggested by Lee et al. (2002) and Wang et al. (2000). Appendix K includes a description of the calculations and procedure based on the paper by Lee et al. (2002).

7.3.1. Consistency

The results for classifying students into four performance levels are separated from those based solely on the Level 3 cut. Table 7.7 and Table 7.8 include case counts ("N-Count"), classification consistency ("Agreement"), classification inconsistency ("Inconsistency"), and Cohen's kappa ("Kappa"). Consistency indicates the rate at which a second administration would yield the same performance category designation (or a different designation for the inconsistency rate). The agreement index is a sum of the diagonal elements in the contingency table. Kappa is a similar measure but corrects for chance agreement. The inconsistency index is equal to the "1-agreement index."

Table 7.7 depicts the consistency study results based on the range of performance levels for both grades. For science, 63–68% of students were estimated to be classified consistently into one of the four performance categories following a hypothetical second administration. Kappa coefficients, which correct for chance agreement, ranged from 0.44 to 0.52. These values are

between “moderate” and “substantial” agreement per Landis and Koch’s (1977) rules of thumb for kappa.

As mentioned above, all scores contain an acceptable measurement error for all tests. For example, by random chance, students testing twice may be classified first as Level 3 and second as Level 4. This is expected to occur more often for students scoring around a specific cut score and less often for students scoring closer to the middle of a performance level (i.e., close to the mid-point of two adjacent cut scores).

Table 7.7. Decision Consistency (All Cuts)

	Grade	N-Count	Agreement	Inconsistency	Kappa
Science	5	153,285	63%	37%	0.44
	8	89,926	68%	32%	0.52

Table 7.8 depicts the consistency study results based on two performance levels (NYS Level 2 and NYS Level 3) as defined by the Level 3 cut. For science, 83–88% of the classifications of individual students were estimated to remain stable with a second administration. Kappa coefficients for science classification consistency ranged from 0.62 to 0.73. These values are considered “substantial” agreement per Landis and Koch’s (1977) rules of thumb for kappa.

Table 7.8. Decision Consistency (Level 3 Cut)

	Grade	N-Count	Agreement	Inconsistency	Kappa
Science	5	153,285	83%	17%	0.62
	8	89,926	88%	12%	0.73

7.3.2. Accuracy

Table 7.9 presents the classification accuracy results for science across both grades. Included in the table are case counts (“N-Count”) and classification accuracy (“Accuracy”) for all performance levels (“All Cuts”) and for the Level 3 cut score. By definition, accuracy associated with the Level 3 cut is at least as great as that with the entire set of cut scores because there are only two categories for the former, as opposed to the four categories for the latter.

For science, the estimated accuracy rates indicate that the categorization of a student’s observed performance agrees with the location of their underlying proficiency 73% to 77% of the time across all performance levels and 88% to 91% of the time regarding the Level 3 cut score.

Table 7.9. Decision Agreement (Accuracy) Estimates

			Accuracy	
	Grade	N-Count	All Cuts	Level 3 Cut
Science	5	153,285	73%	88%
	8	89,926	77%	91%

Section 8: Standard Setting

Standard setting is the formal process by which panels of educators and subject matter experts recommend performance standards. These performance standards include cut points that divide the test scale into performance levels (i.e., Level 1, Level 2, Level 3, and Level 4). Students are placed into one of these performance levels based on their test results.

The adoption of the New York State P-12 Science Learning Standards (NYSP12SLS) in 2016 included the creation of new performance level descriptions for each standard in both grades. These new guiding documents informed the subsequent implementation for the Spring 2024 operational assessments. These changes compelled the establishment of new cut points for the Grades 5 and 8 Science Tests.

Standard setting was conducted in Summer 2024 to set performance standards for the new assessments. This section summarizes the background, methodology, and process of standard setting.

8.1. Goals of Standard Setting

The goals of standard setting were to:

- provide performance standards for the assessments in science and indicate the degree to which students have met the standards for their grades;
- recommend rigorous and attainable performance standards; and
- incorporate existing and future policy considerations relevant to NYS's educational system into the established performance standards.

8.2. Participants

The standard setting panelists were comprised of 28 qualified NYS educators who had knowledge of the current NYSED standards and were from diverse backgrounds regarding demographic characteristics and geographic locations within the State.

8.3. Methodology

The Modified Yes/No Angoff method was used in the standard setting process for setting the cut scores. This method requires panelists to work through each item in a test booklet and provide a "yes" or "no" judgment of whether a student with performance at the borderline of the performance level would get the item correct. The cut scores are derived based on the number of items with a "yes" judgment. The committee-level cut score recommendations are the median of the individual panelist cut scores from the final round.

8.4. Standard Setting Process

The following steps were used as the standard setting process:

1. Standards review committees are convened.
2. Panelists review the current performance level descriptors (PLDs) and develop threshold PLDs.

3. Panelists review and recommend cut score points following the Modified Yes/No Angoff standard setting methodology (three rounds of judgements).

8.5. Results

The cut score recommendations from Round 3 were affirmed during vertical articulation and then approved by the Commissioner of Education. The final raw score cuts are shown in Table 8.1 for both Grades 5 and 8, along with the corresponding scale score cuts.

Table 8.1. Science Performance Level Cut Scores

	Raw Score Cuts			Scale Score Cuts		
Grade	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
5	8	15	24	424	450	480
8	11	20	35	428	450	480

Appendix N: Standard Setting Report presents the full 2024 standard setting report that describes the general process, the composition of the committees, ratings from the various rounds, evaluation forms, and other materials.

Section 9: Summary of Operational Test Results

This section summarizes the distribution of scale score results on the New York State Testing Program (NYSTP) 2024 Grades 5 and 8 Science Tests. These include the scale score means, standard deviations, and performance level distributions for each grade's population and subgroups. Demographic subgroups include gender, ethnicity, Needs Resource Capacity (NRC) category, English Language Learner (ELL) status, students with disabilities (SWDs), and students using test accommodations (SUAs). Furthermore, the ELL/SUA subgroup is defined as ELLs who use one or more ELL-related accommodations, and the SWD/SUA subgroup is defined as SWDs who use one or more disability-related accommodations. The test translation language is also indicated. Science data include students with valid scores from all public, non-public, and charter schools. Complete scale score frequency distribution tables for science are located in Appendix L.

9.1. Scale Score Distribution Summary

The following subsections present science scale scores and subscore statistics by grade and selected subgroups. (Caution is advised when interpreting the statistics for subgroups with small n-counts.)

9.1.1. Science Scale Score and Subscore Distributions

Table 9.1 and Table 9.2 show the summary of scale scores and raw subscores, respectively, for each science grade. Table 9.3 and Table 9.4 show the summary of scale scores by subgroup. Some general observations from these tables include:

- Female and Male students performed comparably.
- Asian students scored considerably higher than other reported ethnic groups.
- Students from Low Needs districts (as identified by NRC category) outperformed students from other districts (New York City, Big 4 Cities, Urban/Suburban, Rural, Average Needs, and Charter).
- ELLs, SWDs, and/or SUAs tended to underperform the State population (All Students).

Table 9.1. Science Scale Score Distribution Summary

		Scale Score	
Grade	N-Count	Mean	SD
5	153,285	441.54	19.92
8	89,926	442.78	20.00

Table 9.2. Science Subscore Summary

			Subscore		
Grade	Subscore Category	N-Count	Max.	Mean	SD
5	ESS	153,285	9	3.00	1.80
	LS	153,285	9	3.10	1.90
	PS	153,285	14	5.79	2.65
8	ESS	89,926	14	3.95	2.52
	LS	89,926	19	6.89	3.81
	PS	89,926	19	6.27	3.01

9.1.1.1. Science Grade 5

Table 9.3 presents the Grade 5 scale score statistics and n-counts for key demographic subgroups. The population scale score mean was 441.54, with a standard deviation of 19.92. Female students tended to perform comparably to Male students. Asian, Multiracial, and White students' scale score means exceeded the State mean scale score, as did those of students from Low Needs districts and Religious or Independent schools. Across ethnic groups, Asian students earned the highest mean score (10 scale score points above the State population), and Black students earned the lowest mean score (7 scale score points below the State population). Across NRC subgroups, students from Low Needs districts earned the highest mean scale score (9 scale score points above the State population), and students from Big 4 Cities districts earned the lowest mean score (11 scale score points below the State population). The SWD, SUA, and ELL subgroups scored about 12–16 scale score points below the mean scale score for the tested population. ELLs were the lowest-performing subgroup analyzed, scoring 16 scale score points below the State mean.

Table 9.3. Science Grade 5 Scale Score Distribution by Subgroup

Demographic Category		N-Count	Scale Score	
			Mean	SD
State	All Items	153,285	441.54	19.92
Gender	Female	74,818	441.26	19.49
	Male	78,448	441.80	20.30
	Non-Binary	19	445.95	20.44
Ethnicity	American Indian or Alaska Native	1,232	438.04	19.24
	Asian	17,245	451.28	20.98
	Black or African American	24,779	434.76	18.12
	Hispanic or Latino	41,066	436.48	17.90
	Multiracial	5,702	443.61	21.27
	Native Hawaiian or Pacific Islander	369	441.99	20.07
	White	62,235	444.86	19.59
NRC	New York City	46,130	441.30	20.13
	Big 4 Cities	6,010	430.28	17.59
	Urban/Suburban	10,509	434.45	17.71
	Rural	8,942	438.15	18.43
	Average Needs	41,104	442.17	19.09
	Low Needs	19,089	450.99	19.68
	Charter	14,163	441.26	19.51
	Religious or Independent	4,715	445.47	18.63
SWD	Yes	25,715	429.04	16.62
SUA	Yes	29,181	429.73	16.77
ELL	Yes	9,742	425.50	13.24
SWD/SUA	Yes	21,805	427.85	15.90
ELL/SUA	Yes	2,836	426.06	13.27

9.1.1.2. Science Grade 8

Table 9.4 presents Grade 8 scale score statistics and n-counts for key demographic subgroups. The population scale score mean was 442.78, with a standard deviation of 20.00. Female students performed comparably to Male students. Asian, Native Hawaiian or Pacific Islander, and White students' scale score means exceeded the State mean scale score, as did those of students from New York City, Low Needs districts, Charter schools, and Religious or Independent schools. Across ethnic groups, Asian students earned the highest mean score (11 scale score points above the State population), and Black students earned the lowest mean score (5 scale score points below the State population). Across NRC subgroups, students from Low Needs districts earned the highest mean scale score (9 scale score points above the State population), and students from Big 4 Cities districts earned the lowest mean score (14 scale score points below the State population). The SWD, SUA, and ELL subgroups scored about 10–15 scale score points below the mean scale score for the tested population. ELLs tested under accommodations were the lowest-performing subgroup analyzed, scoring about 15 scale score points below the State mean.

Table 9.4. Science Grade 8 Scale Score Distribution by Subgroup

Demographic Category		N-Count	Scale Score	
			Mean	SD
State	All Items	89,926	442.78	20.00
Gender	Female	41,855	442.89	19.38
	Male	48,034	442.66	20.52
	Non-Binary	37	456.73	19.61
Ethnicity	American Indian or Alaska Native	693	438.59	18.46
	Asian	7,071	453.33	22.49
	Black or African American	16,499	437.36	18.37
	Hispanic or Latino	27,269	438.84	18.30
	Multiracial	2,702	442.27	20.39
	Native Hawaiian or Pacific Islander	187	444.84	21.93
	White	34,827	446.54	19.84
NRC	New York City	25,730	442.31	20.38
	Big 4 Cities	3,518	428.61	14.71
	Urban/Suburban	8,673	436.37	17.46
	Rural	7,221	441.83	18.58
	Average Needs	23,804	443.79	18.86
	Low Needs	6,875	451.99	20.71
	Charter	7,962	445.87	19.56
	Religious or Independent	3,813	454.21	20.59
SWD	Yes	18,021	431.57	16.08
SUA	Yes	21,655	432.45	16.50
ELL	Yes	6,878	427.45	12.88
SWD/SUA	Yes	14,958	430.63	15.55
ELL/SUA	Yes	3,305	427.40	12.63

9.2. Performance Level Distribution Summary

Students under the NYSTP are classified into performance levels as Level 1, Level 2, Level 3, or Level 4. The cut scores for these performance levels were established during the standard setting in Summer 2024. The very nature of grade-specific content, differing performance expectations, and panel-set cut scores result in cut score differences across grades. Students are considered proficient if they are classified as Level 3 or Level 4.

9.2.1. Science Test Performance Level Distributions

Table 9.5 shows the performance level distributions for all students for each science grade. Table 9.6 and Table 9.7 show the performance level distributions by subgroup for each grade. The percentage of proficient students at a subgroup level reflected the mean scale score distributions for the subgroup. Therefore, similar achievement trends were observed for the percentage of proficient students:

- Male students performed slightly better than Female students.
- Asian students outperformed other reported ethnic groups.
- Students from Low Needs districts (as identified by NRC category) outperformed students from other districts (New York City, Big 4 Cities, Urban/Suburban, Rural, Average Needs, and Charter).
- ELLs, SWDs, and/or SUAs tended to underperform the State population (All Students).

Table 9.5. Science Test Performance Level Distributions

		Performance Levels				
Grade	N-Count	Level 1	Level 2	Level 3	Level 4	Levels 3 & 4
5	153,285	19.81	43.87	32.35	3.96	36.32
8	89,926	21.65	42.72	30.94	4.69	35.63

9.2.1.1. Science Grade 5

Table 9.6 presents the Science Grade 5 performance level distributions and n-counts for key demographic subgroups. The percentage of proficient students was 36.32% for the State population. That percentage was 2% higher for Male students than for Female students. Compared with the State population, the percentages of proficient students were higher for Asian, Multiracial, and White students; the same is true for students enrolled in New York City, Low Needs districts, or Religious or Independent schools. Across ethnic groups, the percentage of proficient students was the highest for Asian students (19% above the State population) and the lowest for Black students (12% below the State population). Across NRC subgroups, the percentage of proficient students was the highest for Low Needs districts (20% above the State population) and the lowest for Big 4 Cities districts (20% below the State population). The percentages of proficient students for SWD, SUA, and ELL subgroups were about 21–30% below that for the State population. ELLs had the lowest percentage of proficient students, 30% below that for the State population.

Table 9.6. Science Grade 5 Performance Level Distribution by Subgroup

Demographic Category		N-Count	Performance Levels				
			Level 1	Level 2	Level 3	Level 4	Levels 3 & 4
State	All Items	153,285	19.81	43.87	32.35	3.96	36.32
Gender	Female	74,818	19.38	45.32	31.69	3.62	35.30
	Male	78,448	20.23	42.49	32.99	4.29	37.28
	Non-Binary	19	15.79	42.11	42.11	0.00	42.11
Ethnicity	American Indian or Alaska Native	1,232	24.51	46.27	26.95	2.27	29.22
	Asian	17,245	9.96	34.15	45.68	10.21	55.89
	Black or African American	24,779	29.45	47.77	21.26	1.52	22.79
	Hispanic or Latino	41,066	25.32	49.30	23.76	1.62	25.38
	Multiracial	5,702	18.91	41.02	33.83	6.24	40.07
	Native Hawaiian or Pacific Islander	369	17.62	47.15	30.08	5.15	35.23
	White	62,235	14.86	41.67	38.89	4.59	43.48
NRC	New York City	46,130	20.12	44.72	30.75	4.41	35.16
	Big 4 Cities	6,010	40.58	42.95	15.26	1.21	16.47
	Urban/Suburban	10,509	29.89	47.86	21.11	1.14	22.25
	Rural	8,942	23.19	47.45	27.48	1.88	29.36
	Average Needs	41,104	17.61	44.79	34.36	3.24	37.60
	Low Needs	19,089	8.64	35.45	47.57	8.34	55.91
	Charter	14,163	19.05	46.13	30.96	3.86	34.82
	Religious or Independent	4,715	12.66	42.76	40.74	3.84	44.58
SWD	Yes	25,715	42.02	44.50	12.63	0.85	13.48
SUA	Yes	29,181	40.46	44.89	13.81	0.85	14.66
ELL	Yes	9,742	47.66	46.23	5.88	0.23	6.11
SWD/SUA	Yes	21,805	44.37	44.06	10.98	0.59	11.57
ELL/SUA	Yes	2,836	46.86	46.54	6.38	0.21	6.59

9.2.1.2. Science Grade 8

Table 9.7 presents the Science Grade 8 performance level distributions and n-counts for key demographic subgroups. The percentage of proficient students was 35.63% for the State population. That percentage was comparable for Female students and Male students. Compared with the State population, the percentages of proficient students were higher for Asian, Native Hawaiian or Pacific Islander, and White students; the same is true for students enrolled in Low Needs districts, Charter schools, or Religious or Independent schools. Across ethnic groups, the percentage of proficient students was the highest for Asian students (20% above the State population) and the lowest for Black students (10% below the State population). Across NRC subgroups, the percentage of proficient students was the highest for Low Needs districts (20% above the State population) and the lowest for Big 4 Cities districts (26% below the State population). The percentages of proficient students for SWD, SUA, and ELL subgroups were about 20–30% below that for the State population. ELLs tested under accommodations had the lowest percentage of proficient students, 30% below that for the State population.

Table 9.7. Science Grade 8 Performance Level Distribution by Subgroup

Demographic Category		N-Count	Performance Levels				
			Level 1	Level 2	Level 3	Level 4	Levels 3 & 4
State	All Items	89,926	21.65	42.72	30.94	4.69	35.63
Gender	Female	41,855	20.35	44.13	31.26	4.27	35.53
	Male	48,034	22.80	41.51	30.64	5.05	35.69
	Non-Binary	37	5.41	29.73	48.65	16.22	64.86
Ethnicity	American Indian or Alaska Native	693	25.69	48.48	23.38	2.45	25.83
	Asian	7,071	11.47	32.77	42.14	13.62	55.76
	Black or African American	16,499	29.51	45.36	22.98	2.15	25.13
	Hispanic or Latino	27,269	26.08	47.06	24.32	2.54	26.86
	Multiracial	2,702	23.72	41.30	30.35	4.63	34.97
	Native Hawaiian or Pacific Islander	187	21.39	40.11	30.48	8.02	38.50
	White	34,827	16.01	40.14	38.03	5.82	43.85
NRC	New York City	25,730	22.68	43.56	28.44	5.32	33.75
	Big 4 Cities	3,518	46.30	44.34	9.27	0.09	9.35
	Urban/Suburban	8,673	30.31	47.48	20.53	1.67	22.21
	Rural	7,221	21.34	44.99	30.59	3.07	33.67
	Average Needs	23,804	18.37	43.81	33.90	3.92	37.82
	Low Needs	6,875	10.78	33.92	45.56	9.75	55.30
	Charter	7,962	16.19	41.80	36.82	5.19	42.01
	Religious or Independent	3,813	8.71	32.31	47.23	11.75	58.98
SWD	Yes	18,021	40.52	45.51	13.17	0.79	13.96
SUA	Yes	21,655	38.77	45.68	14.56	1.00	15.55
ELL	Yes	6,878	48.40	45.78	5.71	0.10	5.82
SWD/SUA	Yes	14,958	42.62	45.02	11.79	0.57	12.36
ELL/SUA	Yes	3,305	48.84	45.48	5.66	0.03	5.69

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Cai, L., Thissen, D. J., & du Toit, S. (2022). *IRTPRO* (version 6.0) [Computer software]. Vector Psychometric Group (VPG). https://store.vpgcentral.com/

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 33–66). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Dorans, N. J., Schmitt, A. P. & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309–319. http://www.jstor.org/stable/1435087

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. https://doi.org/10.1007/978-94-017-1988-9

Individuals with Disabilities Education Act (IDEA). Pub L. No. 108–446 (2004). https://www.congress.gov/108/plaws/publ446/PLAW-108publ446.pdf

Jensen, A. R. (1980). *Bias in mental testing*. Free Press.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kim S. & Kolen M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models* (Version 1.0). [Computer software]. Iowa Testing Programs, University of Iowa.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*(1), 1–17. https://doi.org/10.1111/j.1745-3984.2009.00096.x

Lee, W.-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*(4), 412–432. https://doi.org/10.1177/014662102237797

Lee, W.-C., & Kolen, M. J. (2006, Revised 2008). *IRT-CLASS: A computer program for item response theory classification consistency and accuracy* (version 2.0) [Computer software]. Center for Advanced Studies in Measurement and Assessment, The University of Iowa. http://www.education.uiowa.edu/casma

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. http://www.jstor.org/stable/1435147

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

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

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (3rd Ed.). Holt, Rinehart, and Winston.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8*(2), 111–120. https://doi.org/10.1207/s15324818ame0802_1

Sandoval, J., & Mille, M. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology*, *48*(2), 249–253. https://doi.org/10.1037/0022-006X.48.2.249

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 429–449. https://doi.org/10.1348/000711005X66888

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessment using polytomous IRT. *Journal of Educational Measurement*, *37*(2), 141–162. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2000.tb01080.x

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*(2), 93–111. https://www.jstor.org/stable/1434536

Appendix A: 2024 Elementary-Level Grade 5 and Intermediate-Level Grade 8 Science Test Configurations

Table A1. Elementary-Level Science Grade 5 Test Configuration

Grade	Number of Items				
	Multiple-Choice		Constructed-Response		
	Operational	Embedded	Operational	Embedded	Total
5	19	2–4	15	2–3	38–39

Table A2. Intermediate-Level Science Grade 8 Test Configuration

Grade	Number of Items				
	Multiple-Choice		Constructed-Response		
	Operational	Embedded	Operational	Embedded	Total
8	29	2–4	24	2–4	58–59

Additional details on security, scheduling, classroom organization and preparation, test materials, and administration can be found on NYSED's website. The 2024 Teacher's Directions manuals are available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/cbt-td-math-science-g3-5-2024.pdf and https://www.nysed.gov/sites/default/files/programs/state-assessment/cbt-td-math-science-g6-8-2024.pdf. The 2024 *NYSTP Grades 3–8 English Language Arts, Mathematics, and Science Tests School Administrator's Manual* (SAM) is available online at https://www.nysed.gov/sites/default/files/programs/state-assessment/sam-g3-8-2024.pdf.

Appendix B: 2024 Elementary-Level Grade 5 and Intermediate-Level Grade 8 Science Test Blueprints

Grade	Total Points on OP Test	Strand	Point Range		% of Test	
			Target	Actual	Target	Actual
5	34	Life Science	8–10	9	23–29%	26.5%
		Physical Science	5–14	14	34–40%	41%
		Earth and Space Sciences	9–11	9	27–33%	26.5%
		Engineering, Technology, and the Applications of Science[1]	1–2	2	3–7%	6%
8	53	Life Science	16–20	19	31–37%	36%
		Physical Science	17–20	19	32–38%	36%
		Earth and Space Sciences	11–14	14	21–27%	26%
		Engineering, Technology, and the Applications of Science	1–3	1	2–6%	2%

[1]In addition to questions directly aligned to the Engineering, Technology, and the Applications of Science (ETS) domain, ETS skills and concepts can also be assessed through questions aligned to Physical Science, Life Science, and Earth and Space Sciences.

Appendix C: Item Review Criteria

Master #: _____ **Date:** _____ **Initials:** _____

Science Item Review Criteria for _____

Review the following items to identify any major red flags (⚑). If you find one or more red flags, consider the purpose of the task and the evidence gathered to determine whether the item warrants further review.

Also consider any support materials, such as information about the item and answer keys or rubrics that are provided to students or teachers.

Question	Yes	No
1. Does the task require students to perform the action(s) required in the specified PLD?		⚑
2. Does the task follow the format of the specified task model?		⚑
3. Can the specified disciplinary core idea (DCI) be linked back to a foundational phenomenon?		⚑
4. If a stimulus is provided, does it support the task (as opposed to seeming dropped in)?		⚑
5. If a stimulus is provided, is it real-world and, if taken from a source, appropriately cited?		⚑
6. Can significant portions of the task be completed successfully by using rote knowledge (e.g., definitions, prescriptive or memorized procedures only)?	⚑	
7. Do students need to use scientific reasoning to complete the task?		⚑
8. Does the task require students to use some understanding of the specified disciplinary core idea (DCI) to complete the task?		⚑
9. Do students have to use the specified science and engineering practice(s) (SEP) to successfully complete the task?		⚑
10. Do students have to use the specified crosscutting concept(s) (CCC) to successfully complete the task?		⚑
11. Are the dimensions integrated in the task the student must perform?		⚑
12. Is the task clear and understandable from the student perspective for all students at this grade level?		⚑
13. Are all aspects of the item scientifically accurate?		⚑
14. [MC Only] Does the item have one and only one correct answer?		⚑

Appendix D: Criteria for Item Acceptability

The following criteria represent best practices in item development and were implemented during the creation and review of the NYS ELS Grade 5 and ILS Grade 8 Test items.

For Multiple-Choice Items:

Check that the content of each item:

- is targeted to assess only one objective or skill (unless specifications indicate otherwise)
- deals with material that is important in testing the targeted performance indicator
- uses grade-appropriate content and thinking skills
- is presented at a reading level suitable for the grade being tested
- has a stem that facilitates answering the question or completing the statement without looking at the answer choices
- has a stem that does **not** present clues to the correct answer choice
- has answer choices that are plausible and attractive to the student who has not mastered the objective or skill
- has mutually exclusive distractors
- has one and only one correct answer choice
- is free of cultural, racial, ethnic, age, gender, disability, regional, or other apparent bias

Check that the format of each item:

- is worded in the positive unless it is absolutely necessary to use the negative form
- is free of extraneous words or expressions in both the stem and the answer choices (e.g., the same word or phrase does not begin each answer choice)
- indicates emphasis on key words, such as “best,” “first,” “least,” “not,” and others that are important and might be overlooked
- places the interrogative word at the **beginning** of a stem in the form of a question or places the omitted portion of an incomplete statement at the **end** of the statement
- indicates the correct answer choice
- provides the rationale for all distractors
- is conceptually, grammatically, and syntactically consistent—between the stem and answer choices and among the answer choices
- has answer choices balanced in length or contains two long and two short answer choices
- clearly identifies the passage or other stimulus material associated with the item
- clearly identifies a need for art, if applicable, and the art is conceptualized and sketched, with important considerations explicated

Also check that:

- one item does not present clues to the correct answer choice for any other item
- there is a balance of reasonable, non-stereotypical representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art

For Constructed-Response Items:

Check that the content of each item is:

- designed to assess the targeted performance expectation
- appropriate for the grade being tested
- presented at a reading level suitable for the grade being tested
- appropriate in context
- written so that a student possessing the knowledge or skill being tested can construct a response that can be scored with the specified rubric or scoring tool; that is, the range of possible correct responses must be wide enough to allow for a diversity of responses but narrow enough so that students who do not clearly show their grasp of the objective or skill being assessed cannot obtain the maximum score
- presented without clues to the correct response
- checked for accuracy and documented against reliable, up-to-date sources (including rubrics)
- free of cultural, racial, ethnic, age, gender, disability, or other apparent bias

Check that the format of each item is:

- appropriate for the question being asked and the intended response
- worded clearly and concisely, using simple vocabulary and sentence structure
- precise and unambiguous in its directions for the desired response
- free of extraneous words or expressions
- worded in the positive form rather than in the negative form
- conceptually, grammatically, and syntactically consistent
- marked with emphasis on key words, such as “best,” “first,” “least,” and others that are important and might be overlooked
- clearly identified as needing art, if applicable, and the art is conceptualized and sketched, with important considerations explicated

Also check that:

- one item does not present clues to the correct response to any other item
- there is a balance of reasonable, non-stereotypical representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art

Appendix E: Universal Design Item Checklist

Universal Design Item Checklist	
A.	**Precisely Designed Constructs**
Definition	The item construct is clearly defined so that all irrelevant cognitive, sensory, emotional, and physical barriers are removed.
√	The item does not add skills to those being measured (no extraneous skills tested).
B.	**Language Appropriateness**
Definition	The item avoids words or phrases that are sexist, racist, or otherwise offensive, inappropriate, or negative to any subgroup. Language should be simple and clear.
√	The item uses commonly used words—simpler is better.
√	The item uses vocabulary appropriate for the grade.
√	Idiomatic speech and figurative language are avoided unless being measured.
√	The item avoids technical terms unrelated to the content.
√	The item contains no unnecessary words.
√	The sentence complexity contained in the item is appropriate for the grade.
√	The item avoids ambiguous or multiple-meaning words (e.g., crane—the bird—can easily be confused with crane—heavy machinery).
√	All pronouns have clear referents.
√	The item avoids the use of proper names. (Such names may be unfamiliar or difficult for cultural subgroups.)
√	The item avoids irregularly spelled words.
C.	**Gender Stereotypes**
Definition	The item avoids stereotyping as results of associating genders with certain professions or activities. All groups of society should be portrayed accurately and fairly regarding gender.
√	The item is free of content that might offend a gender subgroup.
√	The item is free of content that might unfairly advantage or disadvantage a gender subgroup.
D.	**Ethnic Stereotypes**
Definition	The item avoids unnecessary references to and uses the proper reference for ethnic, racial, or cultural groups.
√	The item is free of content that might offend an ethnic subgroup.
√	The item is free of content that might unfairly advantage or disadvantage an ethnic subgroup.
√	The artwork included in an item adequately reflects the diversity of the student population.
E.	**Cultural Familiarity**
Definition	Does not rely on an assumed shared experience that is class oriented or native-English-speaking oriented. Presentations of cultural or ethnic differences should neither explicitly nor implicitly rely on stereotypes nor make moral judgments.
√	The item does not rely on an assumed shared experience that is class oriented or native-English-speaking oriented.
√	The item is free from content that might offend a socioeconomic subgroup.
√	The item is free of content that might unfairly advantage or disadvantage a socioeconomic subgroup.

Universal Design Item Checklist	
√	The item is free from unnecessary cultural references.
√	The item is free from religious references.
F.	**Geographic Bias**
Definition	All groups of society should be portrayed accurately and fairly regarding geographic setting. A particular geographic setting shouldn't be used repeatedly, and urban, suburban, and rural settings should be represented across items.
√	The item is free of content that might offend a geographic subgroup.
√	The item is free of content that might unfairly advantage or disadvantage a geographic subgroup.
G.	**Disability Bias**
Definition	All groups of society should be portrayed accurately and fairly regarding disability. Stereotypes related to any particular disability should be avoided. No undue restrictions should exist in the item that would interfere with the ability of a student to comprehend or respond to the item.
√	The item is free of content that might offend a disability subgroup.
√	The item is free of content that might unfairly advantage or disadvantage a disability subgroup.
√	A graphic representation is used in the items, as appropriate. The complexity of the graphic is appropriate to the purpose—simpler is better.
√	The item avoids content that depends on sensory knowledge (such as references to movement, sound, smell, etc.) unless this is crucial to the overall item.
√	The item could be put into braille.
√	The item avoids using both O and Q.
√	Letter pairs can be easily distinguished when read. (S and T are okay; S and X are not).
H.	**Art Supports Text**
Definition	The art is related to the item and supports the reader when possible. The item text and art are legible and accessible, and the art is appropriately placed in the item to support the reader. The art does not distract the test taker but instead provides a scaffold to overall comprehension.
√	All pictures relate to items.
√	The item is free from pictorial clutter: All pictures are needed to answer the item.
√	Graphics are clear and non-fuzzy.
√	Any symbols used are highly distinguishable.
√	Visual load requirements are reasonable for the grade.
√	Multi-dimensional graphics and complex shading are avoided.
√	Tables have replaced any cluttered graphs.
√	Labels read clockwise (as is easier for braille readers).
I.	**Special Populations Considerations**
Definition	Consideration must be given for maximum accessibility to all students, including, but not limited to, English Language Learners/Multilingual Learners, limited sight, hearing impaired, cognitively challenged, etc. These considerations will assist all students.
√	The item contains scaffolding techniques to support student understanding of what is being asked in the item.
√	Text is replaced with graphic representations, when appropriate.

Universal Design Item Checklist	
√	The item is written with simplified text load.
√	The item is written with simplified sentences.
√	The item has as little extraneous information as possible.
√	The item provides context, but it is simplified.
√	The item uses smaller or less-complicated numbers or expressions where not otherwise required.
√	The item avoids negative phrasing or questions; for example, questions are not asked in the negative.

Appendix F: Psychometric Guidelines for Operational Item Selection

It is primarily up to the content development department to select items for the 2024 Operational Test. The psychometrics department provides support, as necessary, and reviews the final item selection. The psychometrics department provides data files with parameters for all field test (FT) items eligible for the item pool. The pools of items eligible for 2024 item selection included 2022–2023 embedded and stand-alone FT items.

Here are the general guidelines for item selection:

- Satisfy the content specifications in terms of objective coverage and the number and percentage of multiple-choice (MC) and constructed-response (CR) items on the test. An often-used criterion for objective coverage is within 5% of the percentages of score points and items per objective.
- To the extent possible, select both easy and difficult items to provide good measurement information at both ends of the performance scale.
- Avoid selecting items with too high/low *p* values, items with flagged point-biserials, and poorly fitting items.
- Minimize the number of items flagged for differential item functioning (DIF) (gender, ethnic, and High/Low Needs schools). Flagged items should be reviewed for content again. Keep in mind that some items may be flagged for DIF by chance only and that their content may not necessarily be biased against any of the analyzed subgroups. The psychometrics department provides DIF information for each item. It is also possible to get "significant" DIF but not bias if the content is a necessary part of the construct that is measured; that is, there may be some flagged DIF items that do not exhibit bias.
- Consideration of the following summary information:
    - Overview of the statistical properties of the tests
    - Blueprint comparison between the test build and the target—the focus is on the total number of points on the test

Appendix G: Operational Item Maps

The following tables show the operational item maps for the 2024 New York State Testing Program (NYSTP) Grades 5 and 8 Science Tests. Field test items that do not contribute to students' scores have been omitted. Additional details on the standards to which these items align are available online at https://www.nysed.gov/sites/default/files/programs/curriculum-instruction/p-12-science-learning-standards.pdf.

Table G1. Science Grade 5 Operational Test Map

Item	Type	Points	Standard	Strand	Subscore Category
1	CR	1	4-LS1-2	LS1.D	LS
2	MC	1	4-LS1-2	LS1.D	LS
3	MC	1	4-LS1-2	LS1.D	LS
4	MC	1	4-PS4-2	PS4.B	PS
5	CR	1	4-PS3-1	PS3.A	PS
6	MC	1	3-PS2-1	PS2.A	PS
7	MC	1	4-PS3-2	PS3.B	PS
8	CR	1	3-5ETS1-2	ETS1.B	—
9	MC	1	4-PS3-3	PS3.C	PS
10	MC	1	3-ESS2-1	ESS2.D	ESS
11	CR	1	3-ESS2-1	ESS2.D	ESS
12	CR	1	3-ESS2-2	ESS2.D	ESS
13	MC	1	3-ESS3-1	ESS3.B	ESS
14	MC	1	4-ESS2-1	ESS2.A	ESS
15	MC	1	5-ESS3-1	ESS3.C	ESS
16	CR	1	5-ESS3-1	ESS3.C	ESS
17	CR	1	5-ESS3-1	ESS3.C	ESS
18	MC	1	4-ESS2-2	ESS2.B	ESS
19	CR	1	5-PS1-3	PS1.A	PS
20	MC	1	5-PS1-1	PS1.A	PS
21	CR-TEI	1	5-PS1-3	PS1.A	PS
22	MC	1	5-PS1-4	PS1.B	PS
23	CR	1	5-PS1-3	PS1.A	PS
24	MC	1	4-LS1-1	LS1.A	LS
25	CR-TEI	1	5-LS2-1	LS2.A	LS
26	CR	1	3-LS4-2	LS4.B	LS
27	CR	1	3-LS2-1	LS2.D	LS
28	CR	1	3-LS3-2	LS3.A	LS
29	MC	1	3-LS4-4	LS2.C	LS
30	MC	1	3-PS2-3	PS2.B	PS
31	MC	1	3-PS2-3	PS2.B	PS
32	CR	1	3-PS2-4	PS2.B	PS
33	MC	1	3-PS2-3	PS2.B	PS
34	MC	1	3-5ETS1-3	ETS1.B	—

Table G2. Science Grade 8 Operational Test Map

Item	Type	Points	Standard	Strand	Subscore Category
1	MC	1	MS-PS4-1	PS4.A	PS
2	MC	1	MS-PS4-2	PS4.B	PS
3	MC	1	MS-PS4-2	PS4.B	PS
4	CR-TEI	1	MS-PS4-2	PS4.B	PS
5	CR	1	MS-PS4-1	PS4.A	PS
6	MC	1	MS-LS4-5	LS4.B	LS
7	CR	1	MS-LS4-5	LS4.B	LS
8	CR	1	MS-LS3-2	LS3.A	LS
9	MC	1	MS-LS4-5	LS4.B	LS
10	MC	1	MS-LS3-1	LS3.A	LS
11	MC	1	MS-ETS1-2	ETS1.B	—
12	CR	1	MS-PS3-1	PS3.A	PS
13	CR-TEI	1	MS-PS3-1	PS3.A	PS
14	MC	1	MS-PS3-1	PS3.A	PS
15	CR	1	MS-PS3-1	PS3.A	PS
16	MC	1	MS-PS3-2	PS3.A	PS
17	CR	1	MS-ESS3-3	ESS3.C	ESS
18	MC	1	MS-ESS3-1	ESS3.A	ESS
19	MC	1	MS-ESS3-4	ESS3.C	ESS
20	CR	1	MS-ESS3-2	ESS3.B	ESS
21	MC	1	MS-ESS3-2	ESS3.B	ESS
22	CR	1	MS-ESS3-4	ESS3.C	ESS
23	MC	1	MS-LS4-3	LS4.A	LS
24	CR-TEI	1	MS-LS4-3	LS4.A	LS
25	CR	1	MS-LS4-2	LS4.A	LS
26	CR	1	MS-LS1-4	LS1.B	LS
27	MC	1	MS-LS4-1	LS4.A	LS
28	MC	1	MS-PS2-4	PS2.B	PS
29	MC	1	MS-PS2-4	PS2.B	PS
30	CR-TEI	1	MS-PS2-5	PS2.B	PS
31	MC	1	MS-PS2-2	PS2.A	PS
32	CR	1	MS-ESS1-2	ESS1.B	ESS
33	MC	1	MS-ESS2-1	ESS2.A	ESS
34	MC	1	MS-ESS2-1	ESS2.A	ESS
35	MC	1	MS-ESS2-3	ESS2.B	ESS
36	CR	1	MS-ESS2-4	ESS2.C	ESS
37	CR	1	MS-ESS3-2	ESS3.B	ESS
38	MC	1	MS-ESS3-2	ESS3.B	ESS
39	CR	1	MS-LS1-7	LS1.C	LS
40	CR	1	MS-LS1-3	LS1.A	LS
41	MC	1	MS-LS1-2	LS1.A	LS
42	MC	1	MS-LS2-4	LS2.C	LS
43	MC	1	MS-LS4-4	LS4.B	LS
44	MC	1	MS-LS2-2	LS2.A	LS

Item	Type	Points	Standard	Strand	Subscore Category
45	CR	1	MS-LS2-2	LS2.A	LS
46	CR	1	MS-LS2-2	LS2.A	LS
47	MC	1	MS-LS2-4	LS2.C	LS
48	MC	1	MS-ESS3-3	ESS3.C	ESS
49	MC	1	MS-PS1-1	PS1.A	PS
50	CR	1	MS-PS1-2	PS1.B	PS
51	MC	1	MS-PS1-5	PS1.B	PS
52	CR-TEI	1	MS-PS1-1	PS1.A	PS
53	CR	1	MS-PS1-4	PS1.A	PS

Appendix H: Factor Analysis Results for Selected Subgroups

As described in Section 3: Validity, a principal component factor analysis was conducted on the 2024 Grades 5 and 8 Science Tests data. The analyses were conducted for the total population of students and select subgroups: English Language Learners (ELLs), students with disabilities (SWDs), and students using test accommodations (SUAs). Table H1 and Table H2 present the results of the factor analysis on the subpopulation data for the Grades 5 and 8 Science Tests, respectively.

Table H1. Science Grade 5 Test Factor Analysis by Subgroup

Demographic Category	Extracted Factor			
			Variance Accounted For	
	N	Eigenvalue	%	Cumulative %
ELL	1	2.98	8.76	8.76
	2	1.18	3.47	12.23
	3	1.14	3.36	15.59
	4	1.10	3.24	18.83
	5	1.08	3.18	22.01
	6	1.07	3.16	25.17
	7	1.06	3.11	28.28
	8	1.04	3.06	31.34
	9	1.02	2.99	34.33
	10	1.01	2.97	37.30
SWD	1	4.30	12.64	12.64
	2	1.23	3.60	16.24
	3	1.15	3.38	19.62
	4	1.10	3.22	22.84
	5	1.07	3.14	25.98
	6	1.03	3.04	29.02
	7	1.02	3.01	32.03
SUA	1	4.33	12.74	12.74
	2	1.25	3.68	16.42
	3	1.15	3.40	19.82
	4	1.10	3.22	23.04
	5	1.07	3.15	26.19
	6	1.04	3.05	29.24
	7	1.02	3.00	32.24

Table H2. Science Grade 8 Test Factor Analysis by Subgroup

Demographic Category	Extracted Factor			
			Variance Accounted For	
	N	Eigenvalue	%	Cumulative %
ELL	1	4.05	7.64	7.64
	2	1.35	2.55	10.19
	3	1.30	2.45	12.64
	4	1.22	2.30	14.94
	5	1.20	2.26	17.20
	6	1.13	2.13	19.33
	7	1.10	2.08	21.41
	8	1.09	2.06	23.47
	9	1.08	2.03	25.50
	10	1.07	2.01	27.51
	11	1.05	1.97	29.48
	12	1.04	1.96	31.44
	13	1.03	1.95	33.39
	14	1.02	1.93	35.32
	15	1.02	1.92	37.24
	16	1.01	1.91	39.15
	17	1.01	1.90	41.05
	18	1.00	1.89	42.94
SWD	1	5.96	11.24	11.24
	2	1.36	2.57	13.81
	3	1.28	2.41	16.22
	4	1.20	2.27	18.49
	5	1.14	2.15	20.64
	6	1.05	1.98	22.62
	7	1.04	1.96	24.58
	8	1.03	1.95	26.53
	9	1.02	1.93	28.46
	10	1.01	1.91	30.37
	11	1.00	1.89	32.26
SUA	1	6.44	12.16	12.16
	2	1.37	2.59	14.75
	3	1.27	2.40	17.15
	4	1.19	2.25	19.40
	5	1.14	2.15	21.55
	6	1.05	1.98	23.53
	7	1.03	1.94	25.47
	8	1.02	1.93	27.40
	9	1.02	1.92	29.32
	10	1.01	1.90	31.22

Appendix I: Classical Test Theory Statistics

These tables support the classical test theory analyses described in Section 5: Operational Test Data Collection and Classical Analysis. They include item type, sample size, percent of omitted responses, *p* value, and the point-biserial correlations (PBis). Field test items that do not contribute to students' scores have been omitted.

Table I1. Science Grade 5 Classical Item Analysis

Item	Type	N-Count	%Omit	*P*-Value	PBis
1	CR	153,285	0.00	0.53	0.42
2	MC	153,285	0.03	0.48	0.45
3	MC	153,285	0.04	0.46	0.28
4	MC	153,285	0.02	0.37	0.27
5	CR	153,285	0.00	0.15	0.36
6	MC	153,285	0.08	0.27	0.23
7	MC	153,285	0.07	0.45	0.46
8	CR	153,285	0.00	0.51	0.43
9	MC	153,285	0.08	0.54	0.42
10	MC	153,285	0.07	0.52	0.39
11	CR	153,285	0.00	0.39	0.48
12	CR	153,285	0.00	0.18	0.36
13	MC	153,285	0.17	0.37	0.30
14	MC	153,285	0.14	0.44	0.33
15	MC	153,285	0.16	0.39	0.36
16	CR	153,285	0.00	0.25	0.38
17	CR	153,285	0.00	0.07	0.31
18	MC	153,285	0.38	0.38	0.22
19	CR	153,285	0.00	0.53	0.49
20	MC	153,285	0.33	0.34	0.15
21	CR-TEI	153,285	0.00	0.66	0.48
22	MC	153,285	0.41	0.47	0.21
23	CR	153,285	0.00	0.44	0.52
24	MC	153,285	0.49	0.53	0.46
25	CR-TEI	153,285	0.00	0.40	0.36
26	CR	153,285	0.00	0.07	0.38
27	CR	153,285	0.00	0.09	0.39
28	CR	153,285	0.00	0.23	0.53
29	MC	153,285	0.90	0.31	0.33
30	MC	153,285	0.91	0.47	0.27
31	MC	153,285	0.97	0.46	0.51
32	CR	153,285	0.00	0.29	0.47
33	MC	153,285	1.16	0.36	0.04
34	MC	153,285	1.20	0.34	0.34

Table I2. Science Grade 8 Classical Item Analysis

Item	Type	N-Count	%Omit	*P*-Value	PBis
1	MC	89,926	0.05	0.39	0.38
2	MC	89,926	0.07	0.37	0.16
3	MC	89,926	0.06	0.60	0.40
4	CR-TEI	89,926	0.00	0.37	0.51
5	CR	89,926	0.00	0.55	0.18
6	MC	89,926	0.10	0.64	0.43
7	CR	89,926	0.00	0.18	0.48
8	CR	89,926	0.00	0.17	0.44
9	MC	89,926	0.13	0.73	0.45
10	MC	89,926	0.16	0.23	0.19
11	MC	89,926	0.18	0.55	0.49
12	CR	89,926	0.00	0.11	0.43
13	CR-TEI	89,926	0.00	0.14	0.36
14	MC	89,926	0.24	0.80	0.41
15	CR	89,926	0.00	0.11	0.46
16	MC	89,926	0.25	0.44	0.16
17	CR	89,926	0.00	0.15	0.27
18	MC	89,926	0.31	0.46	0.26
19	MC	89,926	0.30	0.37	0.42
20	CR	89,926	0.00	0.07	0.33
21	MC	89,926	0.38	0.38	0.36
22	CR	89,926	0.00	0.27	0.52
23	MC	89,926	0.37	0.60	0.44
24	CR-TEI	89,926	0.00	0.35	0.53
25	CR	89,926	0.00	0.14	0.29
26	CR	89,926	0.00	0.31	0.37
27	MC	89,926	0.50	0.53	0.38
28	MC	89,926	0.57	0.37	0.26
29	MC	89,926	0.58	0.42	0.38
30	CR-TEI	89,926	0.00	0.07	0.29
31	MC	89,926	0.61	0.29	0.28
32	CR	89,926	0.00	0.13	0.47
33	MC	89,926	0.63	0.38	0.25
34	MC	89,926	0.63	0.30	0.30
35	MC	89,926	0.66	0.29	0.32
36	CR	89,926	0.00	0.05	0.33
37	CR	89,926	0.00	0.20	0.47
38	MC	89,926	0.73	0.40	0.31
39	CR	89,926	0.00	0.39	0.46
40	CR	89,926	0.00	0.06	0.40
41	MC	89,926	0.77	0.33	0.29
42	MC	89,926	0.83	0.35	0.44
43	MC	89,926	0.85	0.39	0.50
44	MC	89,926	0.78	0.61	0.50
45	CR	89,926	0.00	0.30	0.54
46	CR	89,926	0.00	0.20	0.44

Item	Type	N-Count	%Omit	*P*-Value	PBis
47	MC	89,926	0.92	0.39	0.35
48	MC	89,926	0.91	0.48	0.47
49	MC	89,926	0.89	0.20	0.17
50	CR	89,926	0.00	0.18	0.45
51	MC	89,926	0.97	0.34	0.00
52	CR-TEI	89,926	0.00	0.13	0.40
53	CR	89,926	0.00	0.39	0.56

Appendix J: IRT Statistics

Table J1 and Table J2 present the item-calibration results for the operational (OP) items.

Table J1. Science Grade 5 OP Item Parameter Estimates

Item	Max. Pts.	*a*	*b*
1	1	0.559	-0.177
2	1	0.589	0.107
3	1	0.265	0.329
4	1	0.251	1.345
5	1	0.623	1.941
6	1	0.213	2.791
7	1	0.616	0.249
8	1	0.566	-0.042
9	1	0.527	-0.218
10	1	0.463	-0.144
11	1	0.680	0.502
12	1	0.586	1.824
13	1	0.300	1.076
14	1	0.347	0.411
15	1	0.411	0.686
16	1	0.543	1.382
17	1	0.767	2.485
18	1	0.171	1.760
19	1	0.721	-0.142
20	1	0.102	3.759
21	1	0.769	-0.669
22	1	0.179	0.455
23	1	0.803	0.237
24	1	0.639	-0.145
25	1	0.401	0.667
26	1	1.153	2.007
27	1	0.990	1.941
28	1	1.064	1.039
29	1	0.391	1.354
30	1	0.254	0.318
31	1	0.741	0.182
32	1	0.728	0.953
33	1	-0.062	-5.541
34	1	0.396	1.068

Table J2. Science Grade 8 OP Item Parameter Estimates

Item	Max. Pts.	*a*	*b*
1	1	0.453	0.666
2	1	0.131	2.485
3	1	0.551	-0.549
4	1	0.770	0.512
5	1	0.179	-0.649

Item	Max. Pts.	*a*	*b*
6	1	0.661	-0.679
7	1	0.840	1.457
8	1	0.750	1.587
9	1	0.908	-0.933
10	1	0.208	3.542
11	1	0.748	-0.217
12	1	0.905	1.917
13	1	0.613	2.105
14	1	0.964	-1.206
15	1	0.973	1.811
16	1	0.133	1.081
17	1	0.422	2.644
18	1	0.254	0.357
19	1	0.524	0.667
20	1	0.763	2.536
21	1	0.446	0.742
22	1	0.837	0.944
23	1	0.679	-0.454
24	1	0.850	0.594
25	1	0.473	2.533
26	1	0.481	1.120
27	1	0.462	-0.169
28	1	0.248	1.292
29	1	0.427	0.473
30	1	0.642	2.810
31	1	0.327	1.741
32	1	0.950	1.656
33	1	0.240	1.236
34	1	0.345	1.534
35	1	0.367	1.534
36	1	0.881	2.564
37	1	0.791	1.340
38	1	0.325	0.780
39	1	0.612	0.518
40	1	1.134	2.168
41	1	0.296	1.457
42	1	0.560	0.788
43	1	0.705	0.475
44	1	0.832	-0.444
45	1	0.889	0.782
46	1	0.690	1.458
47	1	0.391	0.763
48	1	0.650	0.098
49	1	0.162	5.101
50	1	0.747	1.520
51	1	-0.086	-4.468
52	1	0.710	1.964
53	1	0.930	0.402

Appendix K: Derivation and Estimation of Classification Consistency and Accuracy

Classification Consistency

Assume that θ is a single latent trait measured by a test and denote Φ as a latent random variable. When a test, X, consists of K items and its maximum number correct score is N, the marginal probability of the number correct (NC) score x is

$$P(X = x) = \int P(X = x|\Phi = \theta)g(\theta)d(\theta), x = 0,1, \dots, N$$

where $g(\theta)$ is the density of θ.

In this report, the marginal distribution, $P(X = x)$, is denoted as $f(x)$, and the conditional error distribution, $P(X = x|\Phi = \theta)$, is denoted as $f(x|\theta)$. It is assumed that students are classified into one of H mutually exclusive categories on the basis of predetermined $H - 1$ observed score cutoffs, $C_1, C_2, \dots, C_{H-1}$. Let L_h represent the h^{th} category into which students with $C_{h-1} \leq X < C_h$ are classified. $C_0 = 0$ and $C_H =$ the maximum number correct score plus one. Then, the conditional marginal probabilities of each category classification are as follows:

$$P(X \in L_h|\theta) = \sum_{x=C_{h-1}}^{C_h-1} f(x|\theta), h = 1, 2, \dots, H$$

$$P(X \in L_h) = \int \sum_{x=C_{h-1}}^{C_h-1} f(x|\theta)g(\theta)d\,\theta, h = 1, 2, \dots, H$$

Because obtaining test scores from two independent administrations of NYS tests was not feasible due to item release after each operational (OP) administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Based on the psychometric model, a symmetric H-by-H contingency table can be constructed. The elements of the H-by-H contingency table consist of the joint probabilities of the row and column observed category classifications.

That two administrations are independent implies that if X_1 and X_2 represent the raw score random variables on the two administrations, then, conditioned on θ, X_1 and X_2 are independent and identically distributed. Consequently, the conditional bivariate distribution of X_1 and X_2 is

$$f(x_1, x_2|\theta) = f(x_1|\theta)f(x_2|\theta)$$

The marginal bivariate distribution of X_1 and X_2 can be expressed as follows:

$$f(x_1, x_2) = \int f(x_1, x_2|\theta)f(\theta)d\theta$$

Consistent classification means that both X_1 and X_2 fall in the same category. The conditional probability of falling in the same category for the two administrations is

$$P(X_1 \in L_h, X_2 \in L_h|\theta) = \left[\sum_{x_1=C_{h-1}}^{C_{h-1}} f(x_1|\theta)\right]^2, h = 1, 2, \ldots, H$$

The agreement index, P, conditional on theta, is obtained by

$$P(\theta) = \sum_{h=1}^{H} P(X_1 \in L_h, X_2 \in L_h|\theta)$$

The agreement index (classification consistency) can be computed as

$$P = \int P(\theta)g(\theta)d(\theta)$$

The probability of consistent classification by chance, P_c, is the sum of squared marginal probabilities of each category classification.

$$P_C = \sum_{h=1}^{H} P(X_1 \in L_h)P(X_2 \in L_h) = \sum_{h=1}^{H} [P(X_1 \in L_h)]^2$$

Then, kappa (Cohen, 1960) is

$$k = \frac{P - P_C}{1 - P_C}$$

Classification Accuracy

Let Γ_w denote true category. When a student has an observed score, $x \in L_h(h = 1, 2, \ldots, H)$, and a latent score, $\theta, \in \Gamma_w$ (w = 1, 2, …, H) an accurate classification is made when $h = w$. The conditional probability of accurate classification is

$$\gamma(\theta) = P(X \in L_w|\theta)$$

where w is the category such that $\theta \in \Gamma_w$.

Lee (2010) thoroughly discusses this item response theory (IRT) method for estimating decision indices, including the computational method used to estimate the results when integrating across the latent variable, θ.

Estimating Classification Indices

The classification consistency and accuracy estimates were obtained using an open-source software program, IRT-CLASS v2.0 (Lee & Kolen, 2006). Below is a brief description of the

files that are used and their purpose. (See the IRT-CLASS v2.0 manual for complete instructions.)

Files needed:

- Raw-to-scale score conversion file
    a. Contains the raw-to-scale score conversions
    b. This is used to provide both raw and scale score classification estimates, which is useful when the raw-to-scale score transformation is not one-to-one.

- Cut score file
    a. Contains the cut scores to be used
    b. Results are provided for all cut scores simultaneously (all performance levels), as well as the estimates based on each of the cut scores separately (Level 3 only).

- Item parameter file
    a. This contains the IRT model used and item parameter estimates.
    b. This information is used when calculating the classification indices.

- Theta file
    a. Contains the theta distribution in terms of quadrature points
    b. The theta and the item parameter files are used to solve the integrals mentioned above.

- Control card
    a. This is used to run the program.
    b. Identifies the names of the four files above and gives a name to the output file.

Appendix L: RSSS and Scale Score Frequency Tables

Table L1 and Table L2 show the raw-score-to-scale-score (RSSS) conversion tables. Table L3 and Table L4 show the scale score distributions that include all students with valid scores by frequency (n-count), percent, cumulative frequency, and cumulative percent.

Table L1. Science Grade 5 RSSS Table

Raw Score	Scale Score	Standard Error
0	398	15
1	401	14
2	404	13
3	407	13
4	411	12
5	414	11
6	417	11
7	420	10
8	424	10
9	429	9
10	433	9
11	436	8
12	440	8
13	443	8
14	447	8
15	450	8
16	453	8
17	456	8
18	460	8
19	463	8
20	466	8
21	469	8
22	473	8
23	477	8
24	480	8
25	484	9
26	489	9
27	494	10
28	497	11
29	500	12
30	503	12
31	506	13
32	509	14
33	512	15
34	516	16

Table L2. Science Grade 8 RSSS Table

Raw Score	Scale Score	Standard Error
0	391	15
1	394	14
2	397	13
3	400	12
4	403	11
5	406	11
6	409	10
7	413	9
8	418	8
9	421	8
10	425	8
11	428	7
12	431	7
13	434	7
14	436	7
15	439	6
16	441	6
17	443	6
18	446	6
19	448	6
20	450	6
21	452	6
22	454	6
23	456	6
24	458	6
25	460	6
26	462	6
27	464	6
28	466	6
29	468	6
30	470	6
31	472	6
32	474	6
33	476	6
34	478	6
35	480	6
36	482	6
37	484	6
38	486	6
39	489	6
40	492	7
41	494	7
42	497	7
43	501	8
44	505	9
45	509	10

Raw Score	Scale Score	Standard Error
46	512	10
47	515	11
48	518	12
49	521	13
50	524	14
51	527	15
52	530	16
53	533	17

Table L3. Science Grade 5 Scale Score Frequency Distribution

			Cumulative	
Scale Score	Freq.	%	Freq.	%
398	73	0.05	73	0.05
401	350	0.23	423	0.28
404	968	0.63	1,391	0.91
407	2,298	1.50	3,689	2.41
411	3,948	2.58	7,637	4.98
414	5,885	3.84	13,522	8.82
417	7,778	5.07	21,300	13.90
420	9,071	5.92	30,371	19.81
424	9,730	6.35	40,101	26.16
429	10,148	6.62	50,249	32.78
433	10,129	6.61	60,378	39.39
436	9,899	6.46	70,277	45.85
440	9,633	6.28	79,910	52.13
443	9,047	5.90	88,957	58.03
447	8,660	5.65	97,617	63.68
450	8,302	5.42	105,919	69.10
453	7,552	4.93	113,471	74.03
456	7,077	4.62	120,548	78.64
460	6,321	4.12	126,869	82.77
463	5,563	3.63	132,432	86.40
466	4,855	3.17	137,287	89.56
469	4,078	2.66	141,365	92.22
473	3,266	2.13	144,631	94.35
477	2,579	1.68	147,210	96.04
480	2,009	1.31	149,219	97.35
484	1,453	0.95	150,672	98.30
489	1,066	0.70	151,738	98.99
494	700	0.46	152,438	99.45
497	427	0.28	152,865	99.73
500	241	0.16	153,106	99.88
503	111	0.07	153,217	99.96
506	47	0.03	153,264	99.99
509	18	0.01	153,282	100.00
512	3	0.00	153,285	100.00

Table L4. Science Grade 8 Scale Score Frequency Distribution

			Cumulative	
Scale Score	Freq.	%	Freq.	%
391	31	0.03	31	0.03
394	52	0.06	83	0.09
397	98	0.11	181	0.20
400	206	0.23	387	0.43
403	478	0.53	865	0.96
406	990	1.10	1,855	2.06
409	1,707	1.90	3,562	3.96
413	2,694	3.00	6,256	6.96
418	3,745	4.16	10,001	11.12
421	4,536	5.04	14,537	16.17
425	4,932	5.48	19,469	21.65
428	5,207	5.79	24,676	27.44
431	5,175	5.75	29,851	33.20
434	4,960	5.52	34,811	38.71
436	4,653	5.17	39,464	43.88
439	4,258	4.74	43,722	48.62
441	4,088	4.55	47,810	53.17
443	3,635	4.04	51,445	57.21
446	3,292	3.66	54,737	60.87
448	3,151	3.50	57,888	64.37
450	3,031	3.37	60,919	67.74
452	2,742	3.05	63,661	70.79
454	2,585	2.87	66,246	73.67
456	2,377	2.64	68,623	76.31
458	2,333	2.59	70,956	78.90
460	2,101	2.34	73,057	81.24
462	1,989	2.21	75,046	83.45
464	1,821	2.02	76,867	85.48
466	1,627	1.81	78,494	87.29
468	1,540	1.71	80,034	89.00
470	1,394	1.55	81,428	90.55
472	1,299	1.44	82,727	91.99
474	1,203	1.34	83,930	93.33
476	939	1.04	84,869	94.38
478	840	0.93	85,709	95.31
480	788	0.88	86,497	96.19
482	658	0.73	87,155	96.92
484	587	0.65	87,742	97.57
486	497	0.55	88,239	98.12
489	392	0.44	88,631	98.56
492	361	0.40	88,992	98.96
494	246	0.27	89,238	99.23
497	215	0.24	89,453	99.47
501	151	0.17	89,604	99.64
505	132	0.15	89,736	99.79

Scale Score	Freq.	%	Cumulative Freq.	Cumulative %
509	62	0.07	89,798	99.86
512	66	0.07	89,864	99.93
515	24	0.03	89,888	99.96
518	18	0.02	89,906	99.98
521	13	0.01	89,919	99.99
524	5	0.01	89,924	100.00
527	2	0.00	89,926	100.00

Appendix M: Test Characteristic Curves

Figure M1. Science Grade 5 TCC

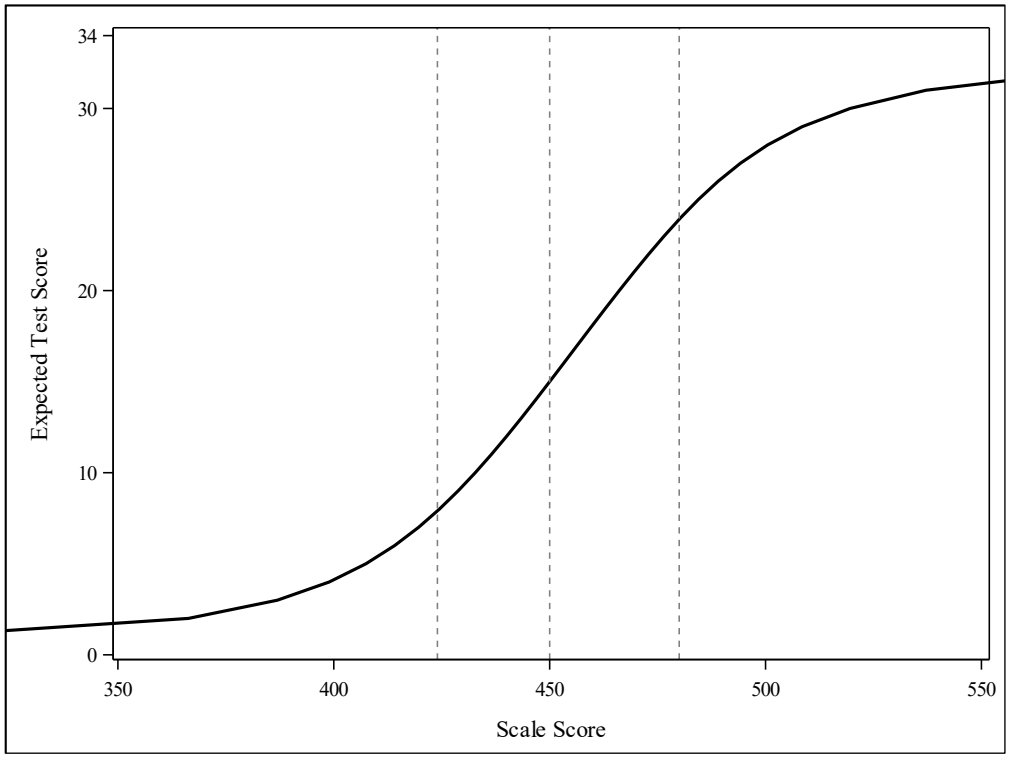
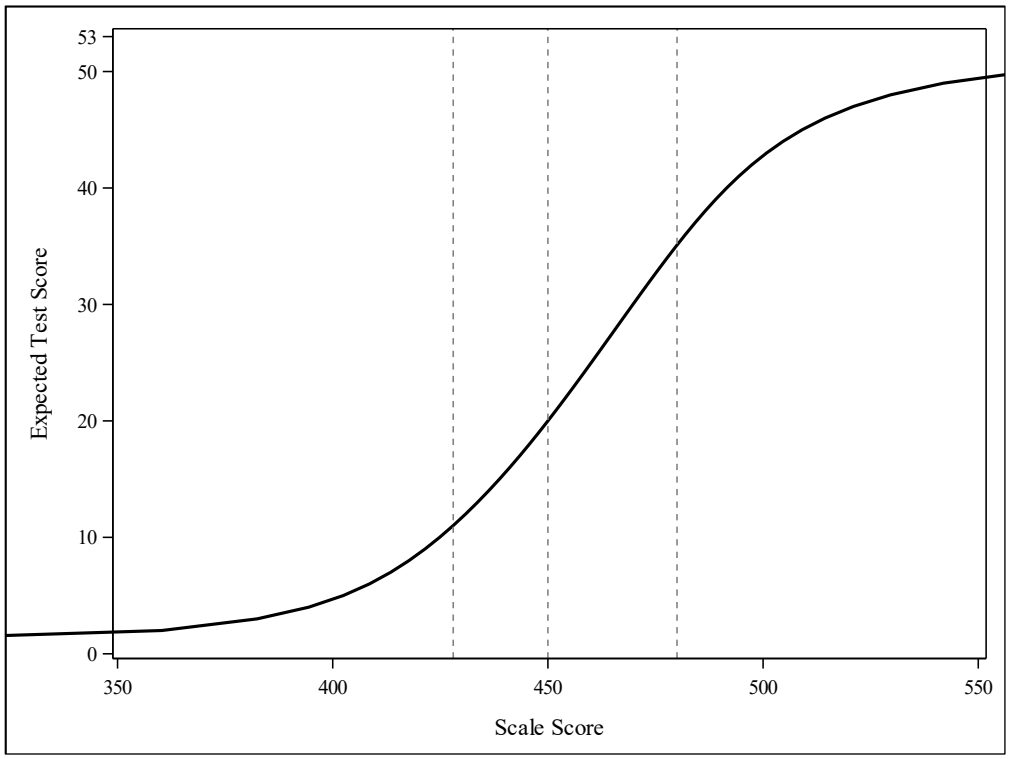


Figure M2. Science Grade 8 TCC



Appendix N: Standard Setting Report

New York State Elementary-level (Grade 5) and Intermediate-level (Grade 8) Science Tests

Standard Setting Report

Prepared for the New York State Education Department
by Pearson

October 2024

Copyright

Contents

List of Tables

List of Figures

Executive Summary

A standard setting meeting was conducted for the New York State Elementary-level (Grade 5) and Intermediate-level (Grade 8) Science Tests. The primary goal for this standard setting was to recommend cut scores that operationally define four performance levels: Level 1, Level 2, Level 3, and Level 4. The performance level designations are used by local, state, and federal accountability programs and are central to communicating with parents, teachers, and the public. This document provides a detailed description of the activities held at the meeting.

The standard setting meeting was held July 10–11, 2024, in Troy, New York. Panelists were trained in and followed the Modified Yes/No Angoff standard setting procedure, resulting in cut score recommendations that were brought to the New York State Education Department (NYSED).

In this report, panelists, materials, methodologies, and results are presented for the New York State Grade 5 and Grade 8 Science Tests standard setting.

1. Grades 5 and 8 Science Tests

The Office of State Assessment (OSA) at NYSED worked with NYS educators to develop the Grade 5 and Grade 8 Science Tests. The tests are designed to measure students' knowledge and understanding of the *NYS Grades 3–8 Science Learning Standards*, first adopted by NYS in 2016, which is part of the transition to the *Next Generation Science Standards* (NGSS) nationally. The Grade 5 Science Test assesses science standards for Grades 3–5, and the Grade 8 Science Test assesses science standards for Grades 6–8.

The new Grade 5 and Grade 8 Science Tests were first administered in spring 2024, and the standard setting activities used the test forms and data from this administration. Both tests are organized through four scientific domains that define the content to be covered on the exams. Table 1.1 presents the four domains along with the estimated percent of points for each domain. The tests are comprised of 1-point multiple-choice items along with 1-point constructed-response and 1-point technology enhanced items (TEIs). The TEIs include some graphing items, drag-and-drop items, multiple-select items, and grid items.

Table 1.1. Domain-level Operational Test Blueprint—Percent Ranges

Grade	Physical Science	Life Science	Earth and Space Science	Engineering, Technology and Applications of Science
5	34–40%	23–29%	27–33%	3–7%
8	32–38%	31–37%	21–27%	2–6%

All questions on the Grade 5 and Grade 8 Science Tests are organized into clusters of questions that follow an assessment storyline. An assessment storyline provides a coherent path toward building Science and Engineering Practices, Disciplinary Core Ideas, and Crosscutting Concepts attached to a phenomenon. In question clusters, each question that is answered may add to the developing explanation, model, or design solution. The group of questions in a cluster follow a theme or storyline grounded in a phenomenon that is focused on an anchor Performance Expectation. However, questions that address other related Performance Expectations can also be included in the cluster. Table 1.2 presents the test designs for the 2024 Grades 5 and 8 Science Tests.

Table 1.2. Grades 5 and 8 Science Test Designs

Grade	Number of Question Clusters	Total Number of Questions
5	7–9	36–43
8	10–12	56–62

2. Performance Level Descriptions

Performance level descriptions (PLDs) are the foundation of standard setting activities because they provide the explanation of how student performance differs from one performance level to the next (Perie, 2008). PLDs are of such influence that, in a well-run standard setting workshop, they determine the rigor of the performance and thus the decisions made about placement of the cut score (Perie et al., 2008). PLDs also serve multiple purposes in terms of communicating policy, facilitating test development, guiding standard setting, and providing score interpretation. Three types of PLDs (Egan et al., 2012) are used as an organizing framework for developing PLDs for the Science examinations:

- Policy PLD statements are designed to capture the vision an agency has for its performance levels. They specify the number of levels and the names for each level and summarize the expectations of student performance for a testing program, including any policy decisions being made at particular levels. Table 2.1 presents the Policy PLDs for the Grade 5 and Grade 8 Science Tests.
- Range PLDs are designed to describe the full range of performance for students at a given performance level. In other words, Range PLDs describe the aspects of test content or specific items that are indicative of a range of students at a specific performance level. Range PLDs can be informative in guiding item and test development as a testing program evolves. They are critical in that they are used to articulate the borderline descriptions, which are a key component for standard setting.
- Borderline descriptions (also known as threshold PLDs) are designed to articulate the transition points between the different ranges of performance defined by the Range PLDs. Specifically, they describe the knowledge and skills a student at the border between performance levels should know and be able to do. Because they articulate the specific performance that distinguishes levels of performance, borderline descriptions are typically used in standard setting activities. Range PLDs and borderline descriptions are interdependent, which necessitates that they be developed in conjunction with each other.

Table 2.1. New York State Science Tests Policy PLDs

Performance Level	Policy PLD
Level 4	Students performing at this level **excel** in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the Learning Standards that are considered **more than sufficient** for the expectations at this grade.
Level 3	Students performing at this level are **proficient** in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the Learning Standards that are considered **sufficient** for the expectations at this grade.
Level 2	Students performing at this level are **partially proficient** in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the Learning Standards that are considered partial but insufficient for the expectations at this grade. Students performing at Level 2 are considered on track to meet current New York high school graduation requirements but are **not yet proficient** in Learning Standards at this grade.
Level 1	Students performing at this level are **below proficient** in standards for their grade. They may demonstrate **limited** knowledge, skills, and practices embodied by the Learning Standards that are considered **insufficient** for the expectations at this grade.

Ultimately, the three types of PLDs are designed to describe the competencies of each performance level in relation to grade-level content standards while addressing their different functions. PLDs play a critical role in the standard setting process.

3. Standard Setting

Standard setting is the process whereby a group of educators is convened to recommend the cut scores (also known as performance or achievement standards) that separate an assessment's score scale into performance levels (i.e., a cut score is the minimum score students must receive to be classified into a certain performance level). Cut scores for the Grade 5 and Grade 8 Science Tests were recommended by two panels of 14 NYS educators each over a two-day standard setting meeting. The Modified Yes/No Angoff procedure (Impara & Plake, 1997; Plake, Ferdous, Impara, & Buckendahl, 2005) of determining cut scores was used in a multi-round process of performance judgments, feedback data, and discussions.

3.1. PANELISTS

The panelists, recruited by NYSED, represented the major geographic regions of NYS, as shown in Table 3.1. As shown in Table 3.2, a high percent of the panelists were classroom teachers, with those not serving as teachers indicating roles such as Curriculum Instruction or Academic Coordinator. In Table 3.3, the variety of settings for the panelists can be observed, with panelists coming from across Rural, Suburban, and Urban settings. Appendix B presents additional details on the demographic characteristics of the panelists.

Table 3.1. Number of Panelists by Geographic Location

Geographic Location	Grade 5	Grade 8
Capital District	3	3
Central	1	2
Long Island	2	1
Lower Hudson	2	1
Mid-Hudson	1	0
North Country/Adirondacks	1	1
NYC	3	3
Southern Tier	1	1
Western	0	2

Table 3.2. Number of Panelists by Current Role

Role	Grade 5	Grade 8
Classroom Teacher	12	10
Other (e.g. Curriculum/Learning Director)	2	4

Table 3.3. Number of Panelists by Current Setting

Setting	Grade 5	Grade 8
Rural	5	3
Suburban	6	4
Urban	3	7

3.2. METHODOLOGY

The Modified Yes/No Angoff standard setting method was used for the standard setting meeting. This is a content- and item-based method that leads participants through a standardized process through which they consider student expectations, as defined by PLDs, and the individual items that could be administered to students to recommend cut scores for each performance level. The process that was followed by the panel to establish their cut score recommendations involved the following steps:

- Review and familiarize themselves with the test form
- Review the current PLDs and develop borderline PLDs for each cut score
- Review and receive training on the Modified Yes/No Angoff methodology
- Complete independent Round 1 ratings and discuss with group after receiving feedback
- Complete independent Round 2 ratings and discuss with group after receiving feedback
- Complete independent Round 3 ratings

Once all three rounds of ratings were completed, the panelists completed an evaluation survey and concluded the meeting activities.

The standard setting process focused on students *just barely* at each performance level, or *threshold* (borderline) students. Therefore, the judgments provided by the panelists for each item and performance level were considered in terms of the success of borderline students. For example, *"Would a student with knowledge and skills at the borderline of the performance level be likely to answer the item correctly?"*

3.3. PRE-WORKSHOP

To engage in the judgment process of standard setting, there must be an understanding of content expectations for each performance level. Prior to the standard setting workshop, panelists were provided some pre-workshop tasks through the Pearson standard setting website, including an introductory standard setting training video, and copies of the Policy and Range PLDs. These tasks were provided ahead of the workshop to set the context for standard setting. Panelists were also asked to review the Educator Guide that includes some sample test items—items available to the public as practice items—to understand what students had to do on the test. Panelists were also asked to review and sign a non-disclosure agreement prior to the workshop and complete a brief demographic survey.

3.4. WORKSHOP

The standard setting workshop was held in Troy, New York, from July 10–11, 2024. Appendix A presents the workshop agenda. The workshop began with a welcome from NYSED, introductory remarks about the Grade 5 and Grade 8 Science Tests, and the goals for setting performance standards on the tests. The lead facilitator provided an overview of the standard setting process, explaining the different types of contextual information used (e.g., PLDs, test content), the standard setting judgment process, and the different types of feedback data that would be presented throughout the workshop. After the general orientation, including workshop logistics, the panelists split into their separate panels and began their work by first reviewing an online version of an operational test form for their grade level.

3.5. PEARSON STANDARD SETTING WEBSITE

The Pearson standard setting website (Moodle) was used as the online platform for meeting pre-work, facilitating the standard setting meeting, and collecting panelist judgments throughout the standard setting process. Each panelist was provided a unique user identification and password that provided secure access to the site. Panelist access was restricted to the section of the site associated with the specific exam assigned to their panel. The standard setting website provided panelists the opportunity to access all resource materials within a secure environment. The website also allowed for streamlining of the data collection from the individual judgment process.

3.6. TEST REVIEW

The panelists were provided access to the spring 2024 computer-based tests that included the full operational test. This provided them with an opportunity to review the multiple-choice items, constructed-response items, and technology-enhanced items to better understand what students were asked to do on the tests. The Rating Guide was provided via the standard setting website to provide the key idea assessed for each multiple-choice item, the answer key for the multiple-choice items, and the scoring rubrics for constructed-response or technology-enhanced items.

3.7. PERFORMANCE LEVEL DESCRIPTIONS

After the test review, the facilitator discussed the Range PLDs and their use during the standard setting process. Panelists were given 15 minutes to discuss the Range PLDs in their table groups, focusing on key differences between the performance levels. The facilitator then provided an explanation for how to derive borderline descriptions from the Range PLDs. Prior to the standard setting, the PLDs were unpacked to highlight the multi-dimensional nature of the standards. For each PLD, the Crosscutting Concepts (CCCs), the Disciplinary Core Ideas (DCIs), and Scientific and Engineering Principles (SEPs) were included within the PLD statements. Using the unpacked PLDs, the facilitator led the full panel through the process of creating borderline descriptions for a small set of PLDs. Following the initial development, panelists split into smaller table discussions and proceeded with the development of the remaining PLDs. To complete the work, the panels first focused on the development of borderline descriptions for the Level 3 cut. After completing the Level 3 borderline descriptions, panelists proceeded to complete the Level 2 and Level 4 descriptions.

After the panelists drafted the borderline descriptions within their table, the facilitator organized the draft descriptions from each table group into a master Google doc. The facilitator then led the whole group through a review of the descriptions and captured any group-approved edits into the master document. The borderline descriptions were printed and shared with the panelists to reference during the judgment activities.

3.8. MODIFIED YES/NO ANGOFF JUDGMENT TRAINING

The panelists were provided thorough training on how to make their recommendations as part of the standard setting meeting. They were instructed on using the Modified Yes/No Angoff method. All items on the test were scored dichotomously. Because all items were scored dichotomously, the essential question that panelists were asked to address was, "*Would a student with knowledge and skills at the borderline of the performance level be likely to answer the item correctly?*" Panelists were instructed to review this question for each of the three cut scores for each item. Significant time was spent on describing the thought process the panelists should go through using parts of the question:

- "*Would…*"— When considering the expected student response to an item, the panelists needed to consider how a student would respond rather than how they should respond. Where "should" is an aspirational expectation, "would" is a more realistic expectation of a student response to the item.
- "*...a student with knowledge and skills at the borderline of the performance level…*"— Panelists should reference the borderline descriptions for each performance level to determine how a student with knowledge and skills at the borderline would be expected to respond.
- "*...be likely answer the question correctly?*"—The panelists will review the knowledge and skills necessary to provide a correct response to the item compared to the expected PLDs for the borderline performance level student. In this context, "likely" is defined as 2 out of 3 times, or 67%. To make this concrete for panelists, facilitators asked them to think about three students at the borderline of a performance level.

Panelists were then instructed to answer the judgment question using the thought process and determine a Yes or No answer for each of the three cut scores for each item. An illustration of the rating form is shown in Figure 3.1.

Figure 3.1. Available Response Options to Judgment Question for Multiple-Choice, CR, and TE Items

	A	B	C	E	F	G
1	**New York State**	**Science Grade 8**				
2	**Round 1 Rating Sheet**					
3	Test	Sequence#	Item Type	Level 2	Level 3	Level 4
4	Grade 8 Science	1	MC			
5	Grade 8 Science	2	MC			
6	Grade 8 Science	3	MC			
7	Grade 8 Science	4	CR (TEI-match)			
8	Grade 8 Science	5	CR			
9	Grade 8 Science	6	MC			

Another step in the standard setting process is a practice judgment task to give the panelists the opportunity to practice making judgments prior to beginning the actual judgment rounds. A set of five practice items was selected from the NYS Question Sampler for use in this activity. However, this activity did not take place during the actual standard setting, given that the borderline description development activity took more time than anticipated. As a result, NYSED and Pearson made the decision to forgo the practice activity and move directly to making the actual judgments in the standard setting rounds in an effort to manage the panelists' time as effectively as possible.

3.9. STANDARD SETTING ROUNDS

Prior to starting each judgment round, panelists were asked a series of readiness questions (via a survey on the website, as shown in Appendix C) to verify that they understood their task and were ready to begin:

- Do you understand your task for the item judgment activity?
- Are you ready to begin the item judgment activity?

Following the readiness survey, the facilitator reviewed the responses. If a panelist were to have responded "no" to either of the questions in the readiness survey, the facilitator would have provided additional training and support as needed to the panelist. Once the facilitator ensured that all panelists were ready to proceed, panelists were asked to make judgments for the first item starting at the lowest performance level based on the borderline descriptions and the knowledge and skills required by the item. The panelists then made judgments for the same item for the rest of the performance levels before proceeding to the next item. Judgments were recorded in an Excel rating form available through the Pearson standard setting website. Once the panelists completed making judgments for all items, they notified their facilitator, who then aggregated all ratings for all panelists. After all panelists completed each judgment activity, the facilitators gathered the item judgments, performed the necessary analysis of the data, and created feedback data that were provided to the panelists.

For the purposes of this workshop, the ratings for all panelists were determined by summing up the number of items that panelists indicated "Yes" for each performance level. This score represented the raw score recommendation for each panelist. However, within the test form being reviewed and rated, there were some field test, or pretest items, that were not used in the calculation of scores for candidates. In order to keep the location of the pretest items confidential, feedback to panelists was not provided at a raw score level. Instead, a pseudo scale score was created for each exam. All feedback to panelists was provided using the pseudo scale score.

Using the pseudo scale score prevented the panelists from easily calculating their personal cut scores, which could have alerted the panelists to the location of the field test items. Instead, a linear transformation of the raw scores was completed to arrive at the pseudo scale score for each cut score recommendation. The linear transformation was designed to create a unique scale used only for the standard setting meeting with a minimum score of 570 and a maximum score of 790. Panelists were provided the guideposts provided in Table 3.4 to aid in their interpretation of the pseudo scale. A copy of the raw score to pseudo scale score is included in Appendix C.

Table 3.4. Guidance Provided to Panelists on the Interpretation of the Pseudo Scale

Minimum	25% Correct	50% Correct	75% Correct	Maximum
570	640	680	720	790

After Round 1, the facilitator provided cut scores generated from the panelists' item-level judgments. Each panelist was able to see their recommended cut score in the Excel rating form. The facilitator then presented a summary of the overall ratings. These feedback in the minimum and maximum values received for each cut score, along with the mean and median across all panelists. Panelists were also shown a histogram that indicated the number of people who provided each cut score recommendation. Using this information, panelists could compare their own cut scores to those from the overall panel and consider if their cut scores matched their level of expectations. The facilitator then led a discussion with the panel regarding their ratings and how they fit within the overall distribution and if panelists felt comfortable with their overall ratings.

After this review, the facilitator led a discussion of the ratings for specific items. Using the panels' Round 1 judgments, items were flagged that witnessed significant disagreements for any of the cut scores. These disagreements could be reflected with a wide range of ratings, with some panelists rating an item fairly easy (the borderline level 2 would get the item correct) and still others rating it fairly difficult (the borderline level 4 would not get the item correct). The facilitator led the panel in a discussion of the items and panelists discussed what characteristics or features

of the items moved them to rate as they did. During this discussion, the facilitator also had available estimates for item difficulty. The item difficulty estimates were not shared directly with the panelists, but the facilitator did share difficulty estimates at a broad level (easy item, hard item, medium difficulty).

After this discussion concluded, panelists completed their Round 2 ratings. Round 2 of standard setting was performed just as Round 1 had been. Panelists were instructed to revisit their judgments from Round 1 and make a new set of judgments, keeping their judgments from Round 1 or making revisions as they felt necessary. After Round 2 judgments, panelists were provided with another set of individual and panel-level cut score information. The facilitator also led a discussion of another set of items where significant disagreement on the ratings were observed. The facilitator led the discussion for both the feedback and the specific items reviewed.

The facilitator also displayed impact data, or the distribution of students among performance levels based on the panel's overall cut scores. Presenting these data during the standard setting process gave the panelists the opportunity to see the consequences of their judgments and whether these consequences fit their expectations. The panelists were reminded that the data should not drive their judgments; rather, their judgments should be driven by content expectations. A discussion was led by the facilitator to discuss whether the impact data aligned with their content expectations.

Following the discussion of the Round 2 feedback data, the panelists provided one final round of judgments. This round was performed just as the previous two rounds. Once the results for Round 3 were complete, panelists were shown the final recommended cut scores and corresponding impact data. As a final task, the panelists completed a workshop evaluation that asked questions ranging from how comfortable they were with specific workshop activities to how comfortable they were with the final recommended cut scores. Table 3.5 presents the types of feedback data and at what round they were provided to the panelist. Appendix D presents examples of the feedback.

Table 3.5. Feedback Data by Judgment Round

Level	Feedback	Round 1	Round 2	Round 3
Item Level	Panelist Agreement Data	✓	✓	
	Score Point Distributions	✓		
Test Level	Individual Cut Scores	✓	✓	
	Committee Scores	✓	✓	✓
	Panelist Agreement Data	✓	✓	
	Impact Data		✓	✓

3.10. CUT SCORES AND IMPACT DATA

Cut scores were generated after each round of judgments. The median value of the individual panelists' cut scores, per performance level, was used as the recommended cut score of the standard setting panel. The standard error of judgment (SEJ) was also calculated for the final recommended cut scores to serve as additional information. Table 3.6 and Table 3.7 present a summary of the cut scores for all three rounds. Figure 3.2 presents the impact data for the third and final round of ratings from the panelists.

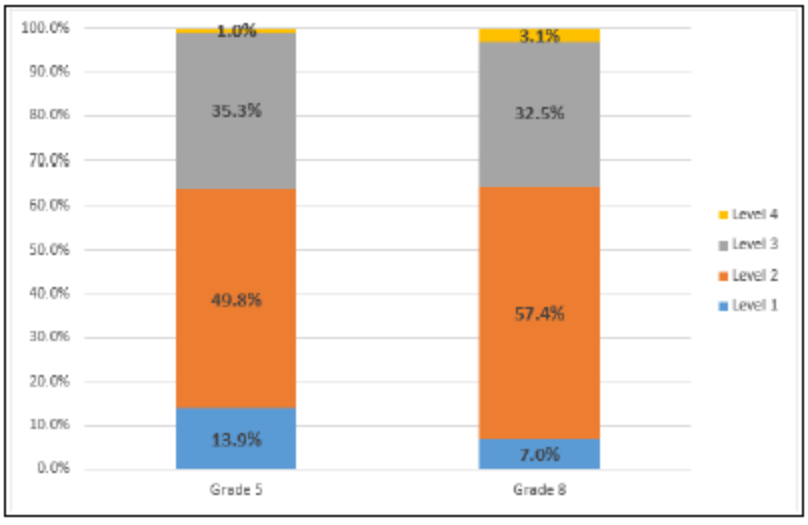
Table 3.6. Recommended Cut Scores Across Rounds—Grade 5

Round	Performance Level	Min.	Max.	Average	Median	SE	25th	75th	% Impact
Round 1	Level 1	--	--	--	--	--	--	--	52.1
	Level 2	9	21	14.6	13.5	1.1	11.25	18.75	47.6
	Level 3	21	34	28.6	29.5	0.9	26.25	31	0.3
	Level 4	31	34	33.6	34.0	0.2	34	34	0.0
Round 2	Level 1	--	--	--	--	--	--	--	45.9
	Level 2	3	22	12.4	12.0	1.4	9	16	50.2
	Level 3	12	31	23.9	24.5	1.3	23	26	3.8
	Level 4	26	32	30.0	30.5	0.5	29.25	31	0.1
Round 3	Level 1	--	--	--	--	--	--	--	13.9
	Level 2	0	20	7.7	7.0	1.3	5.25	9.75	49.8
	Level 3	11	28	16.6	15.0	1.4	12.5	19.25	35.3
	Level 4	21	32	27.2	27.0	0.9	25.25	30.5	0.1

Table 3.7. Recommended Cut Scores Across Rounds—Grade 8

Round	Performance Level	Min.	Max.	Average	Median	SE	25th	75th	% Impact
Round 1	Level 1	--	--	--	--	--	--	--	81.2
	Level 2	4	42	26.1	26.0	2.4	21.5	32	18.2
	Level 3	12	51	42.7	43.5	2.6	42.25	48	0.5
	Level 4	29	53	50.7	53.0	1.7	52	53	0.0
Round 2	Level 1	--	--	--	--	--	--	--	48.6
	Level 2	0	22	12.7	16.0	2.0	6.25	17.75	43.4
	Level 3	6	38	26.6	32.0	3.2	17.75	34	7.5
	Level 4	27	48	40.9	43.5	1.8	37.5	46	0.5
Round 3	Level 1	--	--	--	--	--	--	--	7.0
	Level 2	0	16	7.7	8.5	1.3	5.25	9.75	57.4
	Level 3	7	53	21.8	20.5	3.0	13.5	24.75	32.5
	Level 4	20	47	36.0	37.0	1.9	33.25	41.25	3.1

Figure 3.2. Impact Data Based on Round 3 Ratings



3.11. WORKSHOP EVALUATION

Once the standard setting process was complete and the final recommended cut scores were shown, panelists completed a workshop evaluation on the various materials and activities of the standard setting process and the final recommended cut scores. The intent of this survey was to gather how well panelists understood the process and the materials used and how comfortable they felt about the final recommended cut scores. For the survey questions covering recommended cut scores, panelists were able to express how they would modify the percent of students classified into each performance level if they were somewhat uncomfortable with the overall final recommendation. Most survey questions used a Likert scale, with different scales of affect (e.g., not confident to very confident, not adequate to very adequate, not useful to very useful) across the evaluation.

A complete summary of the evaluation results for both grade levels can be found in Appendix E. One question assessed panelists' confidence in the final cut scores. More specifically, the panelists were asked to rate:

- Please indicate your opinion regarding whether you feel the group's final recommended cut scores were too low, about right, or too high for each cut score. Please bubble *only* one of the three options for each cut score.

As shown in Table 3.8 and Table 3.9, the panelists generally felt comfortable with the cut score recommendations they had developed. There were some panelists in both panels that felt the Level 4 cut score was too high, but a clear majority still rated it as "About Right."

Table 3.8. Ratings from Grade 5 Panel

Performance Level	Too Low	About Right	Too High
Level 2	--	13	1
Level 3	--	14	--
Level 4	1	9	4

Table 3.9. Ratings from Grade 8 Panel

Performance Level	Too Low	About Right	Too High
Level 2	5	8	1
Level 3	1	11	2
Level 4	1	9	4

3.12. FINAL RECOMMENDATIONS

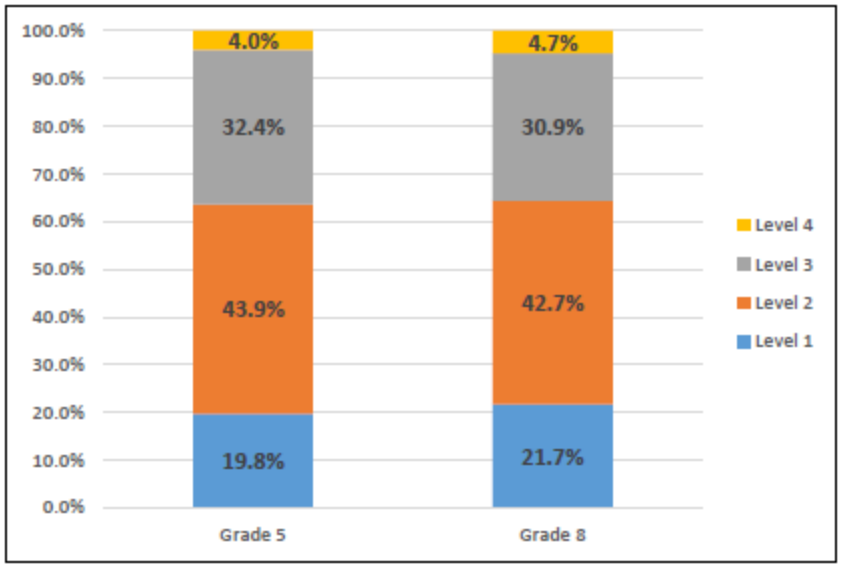
The goal of the standard setting meeting was to identify performance level cut scores consistent with the PLDs and state policy directives using a standardized procedure called the Modified Yes/No Angoff method. The meeting reflected best practice as articulated in the *Standards for Educational and Psychological Measurement* (AERA et al., 2014) and proceeded according to plans reviewed by the New York State Technical Advisory Committee. The panelists were diverse and representative of the state, and the group followed, without incident, instructions delivered by the standard setting facilitator. All activities were formally overseen by the OSA senior management and psychometric staff.

After careful consideration of the nature of the new examination, the rigor of the new curricula, the transitional and aspirational aspects of the NYSED policy directives, and the role of the assessment in student learning, the standard setting committee made recommendations on the cut scores to the Commissioner of Education. The Commissioner of Education subsequently made adjustments to the recommendations based on the committee feedback from the survey, standard errors of judgement, and historical data. The final approved cut scores were implemented within the scale of measurement used to report student performance on the New York State Grade 5 and Grade 8 Science Tests. Table 3.10 presents the approved cuts scores, with subsequent impact data provided in Figure 3.3.

Table 3.10. Final Approved Cut Scores

Grade	Level 2 Cut	Level 3 Cut	Level 4 Cut
5	8	15	24
8	11	20	35

Figure 3.3. Impact Data Based on Final Approved Cut Scores



4. 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.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp.508–600). American Council on Education.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29.

Perie, M., Hess, K., & Gong, B. (2008). Writing performance level descriptors: Applying lessons learned from the general assessment to alternate assessments based on alternate and modified achievement standards. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Appendix A: Standard Setting Agenda

Standard Setting Meeting

New York State Elementary-level Science (Grade 5)
and Intermediate-level Science (Grade 8)
Agenda

Day 1- July 10, 2024

7:30 – 8:00am	*Breakfast*
8:00 – 8:30am	Welcome and Standard Setting Overview
****	Break into Grade-level panels ****
8:30 – 8:45am	Introductions, logins, material orientation, meeting security
8:45 – 9:30am	Experience the Assessment
9:30 – 9:45am	*Break*
9:45 – 10:15am	Review and Discuss Performance Level Descriptions
10:15 – 10:45am	Borderline Performance Level Descriptors Training [Includes modeling]
10:45 – 11:45am	Borderline PLD Level 3 Creation
	Table Discussion
	Group Discussion
11:45 – 12:30pm	*Lunch*
12:30 – 2:00pm	Borderline PLD Levels 2 and 4 Creation
	Table Discussion
	Group Discussion
2:00 – 2:30pm	Standard Setting Training
2:30 – 3:00pm	Practice Judgment Activity and Discussion
3:00 – 3:15pm	Break
3:15 – 4:30pm	Round 1 Judgments

Day 2 - July 11, 2024

7:30 – 8:30am	*Breakfast*
****	Break into Grade-level panels ****
8:30 – 8:45am	Round 1 Judgment Feedback
	Item Level - Item means and distributions
	Test Level - Cut score recommendations; Panelist agreement
8:45 – 9:30am	Table Discussion - Round 1 Feedback
	Panelists discuss feedback data at their tables
9:30 – 9:45am	Whole Group Discussion – Item Disagreement Data
9:45 – 10:45am	Round 2 Judgments
	Round 2 Readiness form
	Panelists work independently to make Round 2 judgments
10:45 – 11:00am	*Break*
11:00 – 11:15am	Round 2 Judgment Feedback
	Item Level - Item means and distributions
	Test Level - Cut score recommendations, Panelist agreement
11:15 – 11:45am	Table Discussion - Round 2 Feedback
11:45 – 12:30pm	*Lunch*
12:30 – 1:30pm	Whole Group Discussion - Round 2 Feedback
	Impact Data
1:30 – 2:15pm	Round 3 Judgments
	Round 3 Readiness form
	Panelists work independently to make Round 3 judgments
2:15 – 2:45pm	*Break*
2:45 – 3:15pm	Round 3 Feedback, Evaluation, and Workshop Wrap-up

Appendix B: Panelist Demographics

Panelists responded to an information survey to provide demographic and other pertinent information for validity evidence of the standard setting. A total of 28 panelists participated in the standard setting. The survey results have been tabulated below.

Table B.1. What is your current position?

Answer Option	Grade 5	Grade 8
Classroom Teacher	12	10
Other (e.g. Curriculum/Learning Director)	2	4

Table B.2. How many years have you been in the education field?

Answer Option	Grade 5	Grade 8
1–5 years	--	1
6–10 years	--	2
11–15 years	1	3
16–20 years	4	4
More than 20 years	9	4

Table B.3. What is the highest educational degree you have earned?

Answer Option	Grade 5	Grade 8
Master's degree (M.A., M.S.)	13	14
Doctoral degree (Ph.D., Ed.D.)	1	--

Table B.4. What is your gender?

Answer Option	Grade 5	Grade 8
Female	13	10
Male	--	4
No response	1	--

Table B.5. What is your race/ethnicity?

Answer Option	Grade 5	Grade 8
Asian	1	1
Black or African American	--	2
Hispanic or Latino	--	1
Multi-racial	2	--
White	10	8
No response	1	2

Table B.6. In what type of school district do you work?

Answer Option	Grade 5	Grade 8
Rural	5	3
Metropolitan/Urban	6	4
Suburban	3	7

Appendix C: Raw Score to Pseudo Scale for the Workshop

Table C.1. Grade 5 Raw Score to Pseudo Scale Score

Raw Score	Pseudo Scale
0	571
1	576
2	580
3	584
4	588
5	592
6	596
7	600
8	606
9	612
10	618
11	623
12	627
13	632
14	636
15	641
16	645
17	649
18	654
19	658
20	662
21	667
22	671
23	676
24	681
25	687
26	693
27	700
28	704
29	708
30	712
31	716
32	720
33	724
34	729

Table C.2. Grade 8 Raw Score to Pseudo Scale Score

Raw Score	Pseudo Scale
0	573
1	578
2	582
3	586
4	591
5	595
6	600
7	607
8	614
9	619
10	624
11	629
12	633
13	637
14	641
15	645
16	649
17	652
18	655
19	659
20	662
21	665
22	668
23	671
24	674
25	677
26	680
27	683
28	686
29	688
30	691
31	694
32	697
33	700
34	703
35	706
36	710
37	713
38	716
39	720
40	724
41	728
42	733
43	738
44	743
45	750
46	755
47	759
48	763
49	768
50	772
51	777
52	781
53	786

Appendix D: Sample Feedback

Table D.1. Rating Summary (provided all rounds)

	Panelists	Average Rating	Median Rating	Min	Max
Level 1	--	--	--	--	--
Level 2	14	605.5	600.0	571	662
Level 3	14	648.3	641.0	623	704
Level 4	14	698.3	700.0	667	720

Figure D.1. Cut Score Rating Distribution (provided all rounds)

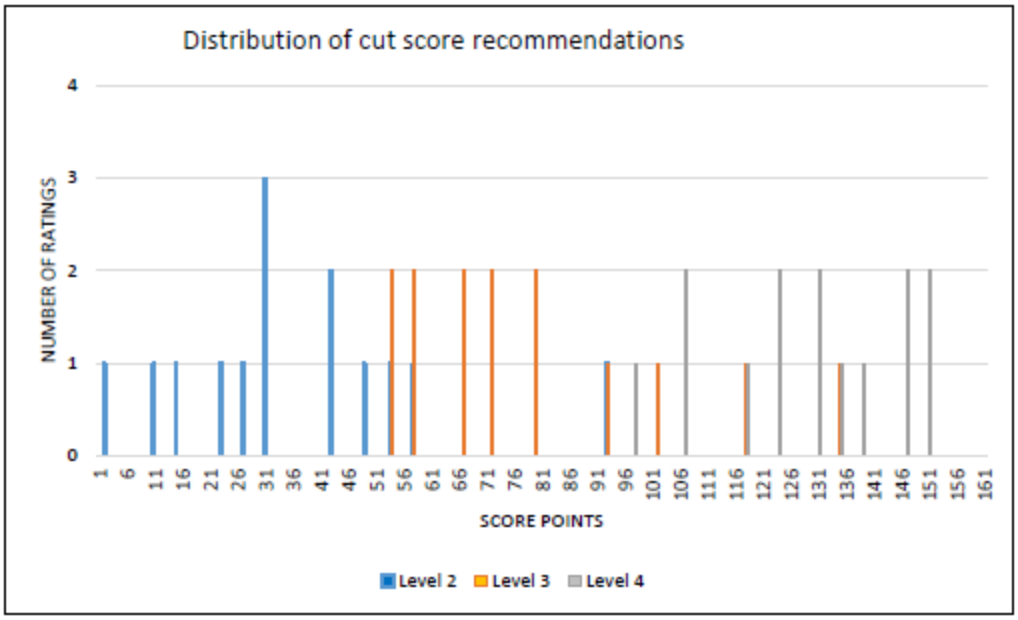
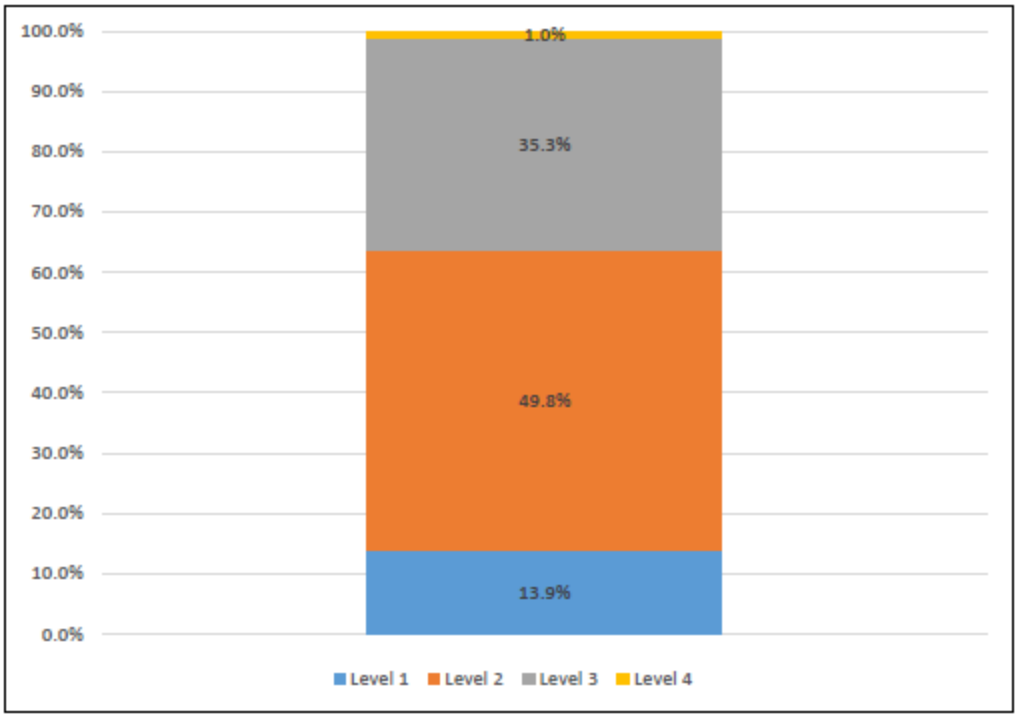


Figure D.2. Impact Data (Provided after Round 2 and Round 3)



3695831.401

Appendix E: Workshop Evaluation Results

The purpose of this evaluation is to help document the process used to recommend cut scores for New York State's Grades 5 and 8 Science Tests.

GRADE 5

Table E.1. Training Process—Grade 5

Response Option	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Before Round 1 began, I was comfortable with the item rating procedure.	2	4	6	3	--
I understood the cut-score summary data that was presented between the rounds.	--	1	--	10	3
I understood the impact data that were presented after Round 2.	--	--	--	9	5
By the end of Round 3, I was comfortable with the item rating procedure.	--	--	--	4	10
Overall, I believe my opinions were considered and valued by my group.	--	--	--	3	11

Table E.2. Influence—Grade 5

Response Option	Not Influential	Somewhat Influential	Influential	Very Influential
The Performance Level Descriptions (PLDs)	--	1	4	9
The descriptions of students demonstrating borderline performance.	--	--	8	6
My perception of the difficulty of the items	--	--	8	6
My experiences with students	--	1	7	6
Discussion within my group	--	--	5	9
The item ratings of other participants	1	6	5	2
The percent of students in each performance level (the impact data)	1	4	8	1
My sense of what a student needs to know to be identified at Level 2.	--	1	5	8
My sense of what a student needs to know to be identified at Level 3	--	1	5	8
My sense of what a student needs to know to be identified at Level 4	--	1	5	8

Table E.3. Cut Scores—Grade 5

Response Option	Too Low	About Right	Too High
Level 2 cut score	--	13	1
Level 3 cut score	--	14	--
Level 4 cut score	1	9	4

GRADE 8

Table E.4. Training Process—Grade 8

Response Option	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Before Round 1 began, I was comfortable with the item rating procedure.	1	5	3	4	1
I understood the cut-score summary data that was presented between the rounds.	--	--	1	9	4
I understood the impact data that were presented after Round 2.	--	--	--	4	10
By the end of Round 3, I was comfortable with the item rating procedure.	--	--	1	3	10
Overall, I believe my opinions were considered and valued by my group.	--	--	--	4	10

Table E.5. Influence—Grade 8

Response Option	Not Influential	Somewhat Influential	Influential	Very Influential
The Performance Level Descriptions (PLDs)	--	2	5	7
The descriptions of students demonstrating borderline performance.	1	1	5	7
My perception of the difficulty of the items	--	--	7	7
My experiences with students	--	--	4	10
Discussion within my group	--	1	6	7
The item ratings of other participants	--	4	8	2
The percent of students in each performance level (the impact data)	--	4	4	6
My sense of what a student needs to know to be identified at Level 2.	--	2	5	7
My sense of what a student needs to know to be identified at Level 3	--	1	6	7
My sense of what a student needs to know to be identified at Level 4	--	1	6	7

Table E.6. Cut Scores—Grade 8

Response Option	Too Low	About Right	Too High
Level 2 cut score	5	8	1
Level 3 cut score	1	11	2
Level 4 cut score	1	9	4