Utah Aspire Plus
2024–2025 Technical Report

Prepared by Pearson for the Utah State Board of Education (USBE) | August 2025

Foreword

This technical report describes the technical characteristics of the Utah Aspire Plus assessments in grades 9 and 10 in reading, mathematics, and science. The intended audience of this report are those with a basic technical understanding of large-scale assessment systems and their uses. It assumes some technical knowledge of how score scales are developed and derived and how scores are intended to support valid interpretations of intended claims. The report is designed to reflect the Utah State Board of Education's (USBE's) commitment to high professional standards as outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) and federal peer review guidelines (U.S. Department of Education, 2015).

USBE's mission is to serve the public by providing measurable information about Utah students' core knowledge, skills, and abilities, acquired through high quality valid and reliable assessments. They strive to positively impact student learning and the public's understanding through quality assessment; provide meaningful assessment that is essential to assess the extent of student progress toward proficiency; provide accurate, understandable reporting that is essential so that all stakeholders in Utah education have the data needed for making effective decisions concerning school policies, programs and curricula; provide knowledge about use of accountability measures, resources/tools to support best practices in the area of assessment and support broad understandings; utilize innovative technologies that support valid and cost-effective indicators of student proficiency; and accomplish all tasks through positive collaborative partnerships with districts and state agencies.

Utah State Board of Education
250 East 500 South
Salt Lake City, UT 84111
(801) 538-7500
https://schools.utah.gov/

This report was created by Pearson under contract with USBE. The content and format of this report is determined by USBE.

Table of Contents

1. Introduction 7
    1.1. Assessment Overview 7
    1.2. Background 8
    1.3. Testing Requirements 8
2. Test Design 10
    2.1. Content Standards 10
    2.2. Claims and Subclaims 10
    2.3. Test Blueprints 12
    2.4. Item Types 12
    2.5. Cognitive Complexity 12
    2.6. Test Structure and Testing Time 13
3. Test Development 14
    3.1. Operational Forms Development 14
    3.2. Statistical Guidelines 15
    3.3. 2025 Match to Test Blueprint 15
4. Test Administration 17
    4.1. Test Security 17
    4.2. Remote Proctoring 18
    4.3. Test Accommodations and Supports 18
    4.4. Test-Taking Irregularities and Security Breaches 20
5. Scoring and Reporting 21
    5.1. IRT Pattern Scoring 21
    5.2. Interpretation of Test Scores 22
        5.2.1. Composite and Subject Area Scale Scores 22
        5.2.2. Performance Levels and PLDs 22
        5.2.3. ACT Predicted Scores 22
    5.3. Appropriate Uses for Scores and Reports 23
6. Standard Setting 24
7. Administration Results 25
    7.1. Test Taker Characteristics 25
    7.2. Testing Time 25
    7.3. Scale Score and Performance Level Distribution 26
8. Item Analyses 28
    8.1. Classical Item Analysis 28
        8.1.1. Item Difficulty (P-value and Item Mean Scores) 28
        8.1.2. Item-Total Correlations 29
    8.2. Differential Item Functioning 29
9. Calibration, Equating, and Scaling 32
    9.1. IRT Models 32
    9.2. Calibration 32
    9.3. Equating 33
        9.3.1. Drift Analysis 34
        9.3.2. Model Fit Evaluation Criteria 35
    9.4. Establishing the Reporting Scale 36
10. Quality Control Procedures 38
    10.1. Quality Control of Test Development 38
    10.2. Quality Control of Online Assessment Delivery 38

10.3. Quality Control of Production System Testing 39
10.4. Quality Control of Scoring and Reporting 40
10.5. Quality Control of Psychometric Processes 41
11. Reliability 42
11.1. Classical Reliability 42
11.1.1. Cronbach's Alpha 42
11.1.2. Standard Error of Measurement 43
11.2. IRT Reliability 43
11.3. Classification Accuracy and Consistency 43
11.3.1. Calculating Accuracy and Consistency 44
11.3.2. Calculating Kappa 45
11.3.3. Results 45
12. Validity 49
12.1. Evidence Based on Test Content 49
12.2. Evidence Based on Cognitive Process 50
12.3. Evidence Based on Internal Structure 51
12.4. Evidence Based on Different Student Populations 53
12.5. Summary 53
References 54
Appendix A: Test-Level Reporting Categories and Standards by Item Type and DOK 57
Appendix B: Student Testing Time Plots 60
Appendix C: Reliability Results by Subgroup 63
Appendix D: Conditional Standard Error of Scale Scores 66
Appendix E: Common Item Scatter Plots for 2025 Anchor Items 69
Appendix F: Scale Score Descriptive Statistics by Subgroup 72
Appendix G: Scale Score Distributions for Overall Testing Population 75
Appendix H: Performance Level Distributions by Subgroup 78
Appendix I: Principal Components Scree Plots 81
Appendix J: Subscore Correlations 84
Appendix K: Item Drift Plots 85

List of Tables

Table 2.1. Claims and Subclaims 11
Table 2.2. Test Form Composition 13
Table 3.1. Spring 2025 Field Test Forms 14
Table 3.2. Operational Test Blueprint Match—Reading 15
Table 3.3. Operational Test Blueprint Match—Mathematics 16
Table 3.4. Operational Test Blueprint Match—Science 16
Table 4.1. Testing Accommodations and Supports 19
Table 4.2. Test-Taking Irregularities and Security Breaches 20
Table 5.1. IRT Summary Parameter Estimates for Operational Items 21
Table 6.1. Scale Score Ranges and Cut Scores 24
Table 7.1. Spring 2025 Participation Rates 25
Table 7.2. Spring 2025 Student Testing Time (in minutes) 26
Table 7.3. Overall Scale Score Descriptive Statistics 27
Table 7.4. Overall Performance Level Distributions 27
Table 8.1. Item Difficulty for 1-Point Items 28

Table 8.2. Item Difficulty for 2-Point Items....28
Table 8.3. Item-Total Correlation for 1-Point Items....29
Table 8.4. Item-Total Correlation for 2-Point Items....29
Table 8.5. Item 2×2 Contingency Table for the k^{th} Score Level....30
Table 8.6. DIF Results: Number of Items by DIF Category....30
Table 9.1. 2025 Final Stocking and Lord Scaling Constants....33
Table 9.2. 2025 Items Showing Drift....34
Table 10.1. Quality Control of Production System Testing....39
Table 11.1. Example Accuracy Classification Table: True vs. Observed Scores....44
Table 11.2. Example Accuracy Classification Table for Proficient Cut Point: True vs. Observed Scores....44
Table 11.3. Example Consistency Classification Table: First vs. Second Form....45
Table 11.4. Classification Accuracy: True vs. Observed Scores....46
Table 11.5. Classification Accuracy at *Proficient* Cut Point: True vs. Observed Scores....47
Table 11.6. Classification Consistency: First vs. Alternate Form....48
Table 12.1. Model Fit Indices for Confirmatory Factor Analyses....52
Table A.1. Test-Level Reporting Categories and Standards—Reading Grade 9....57
Table A.2. Test-Level Reporting Categories and Standards—Reading Grade 10....57
Table A.3. Test-Level Reporting Categories and Standards—Mathematics Grade 9....58
Table A.4. Test-Level Reporting Categories and Standards—Mathematics Grade 10....59
Table C.1. Test Reliability by Subgroup and Reporting Category—Reading Grade 9....63
Table C.2. Test Reliability by Subgroup and Reporting Category—Reading Grade 10....63
Table C.3. Test Reliability by Subgroup and Reporting Category—Mathematics Grade 9....64
Table C.4. Test Reliability by Subgroup and Reporting Category—Mathematics Grade 10....64
Table C.5. Test Reliability by Subgroup and Reporting Category—Science Grade 9....65
Table C.6. Test Reliability by Subgroup and Reporting Category—Science Grade 10....65
Table F.1. Scale Score Descriptive Statistics by Subgroup—Reading Grade 9....72
Table F.2. Scale Score Descriptive Statistics by Subgroup—Reading Grade 10....72
Table F.3. Scale Score Descriptive Statistics by Subgroup—Mathematics Grade 9....73
Table F.4. Scale Score Descriptive Statistics by Subgroup—Mathematics Grade 10....73
Table F.5. Scale Score Descriptive Statistics by Subgroup—Science Grade 9....74
Table F.6. Scale Score Descriptive Statistics by Subgroup—Science Grade 10....74
Table H.1. Performance Level Distribution by Subgroup—Reading Grade 9....78
Table H.2. Performance Level Distribution by Subgroup—Reading Grade 10....78
Table H.3. Performance Level Distribution by Subgroup—Mathematics Grade 9....79
Table H.4. Performance Level Distribution by Subgroup—Mathematics Grade 10....79
Table H.5. Performance Level Distribution by Subgroup—Science Grade 9....80
Table H.6. Performance Level Distribution by Subgroup—Science Grade 10....80
Table J.1. Correlations of Total Score and Subscores....84

List of Figures

Figure B.1. Student Testing Time Plot—Reading Grade 9 60
Figure B.2. Student Testing Time Plot—Reading Grade 10 60
Figure B.3. Student Testing Time Plot—Mathematics Grade 9 61
Figure B.4. Student Testing Time Plot—Mathematics Grade 10 61
Figure B.5. Student Testing Time Plot—Science Grade 9 62
Figure B.6. Student Testing Time Plot—Science Grade 10 62
Figure D.1. CSEM of Scale Scores—Reading Grade 9 66
Figure D.2. CSEM of Scale Scores—Reading Grade 10 66
Figure D.3. CSEM of Scale Scores—Mathematics Grade 9 67
Figure D.4. CSEM of Scale Scores—Mathematics Grade 10 67
Figure D.5. CSEM of Scale Scores—Science Grade 9 68
Figure D.6. CSEM of Scale Scores—Science Grade 10 68
Figure E.1. IRT B Parameters for Operational Items—Reading Grade 9 69
Figure E.2. IRT B Parameters for Operational Items—Reading Grade 9 69
Figure E.3. IRT B Parameters for Operational Items—Mathematics Grade 9 70
Figure E.4. IRT B Parameters for Operational Items—Mathematics Grade 9 70
Figure E.5. IRT B Parameters for Operational Items—Science Grade 9 71
Figure E.6. IRT B Parameters for Operational Items—Science Grade 9 71
Figure G.1. Scale Score Distribution—Reading Grade 9 75
Figure G.2. Scale Score Distribution—Reading Grade 10 75
Figure G.3. Scale Score Distribution—Mathematics Grade 9 76
Figure G.4. Scale Score Distribution—Mathematics Grade 10 76
Figure G.5. Scale Score Distribution—Science Grade 9 77
Figure G.6. Scale Score Distribution—Science Grade 10 77
Figure I.1. Principal Component Scree Plot—Reading Grade 9 81
Figure I.2. Principal Component Scree Plot—Reading Grade 10 81
Figure I.3. Principal Component Scree Plot—Mathematics Grade 9 82
Figure I.4. Principal Component Scree Plot—Mathematics Grade 10 82
Figure I.5. Principal Component Scree Plot—Science Grade 9 83
Figure I.6. Principal Component Scree Plot—Science Grade 10 83
Figure K.1. Item Drift Plot—Reading Grade 9 85
Figure K.2. Item Drift Plot—Reading Grade 10 85
Figure K.3. Item Drift Plot—Mathematics Grade 9 86
Figure K.4. Item Drift Plot—Mathematics Grade 10 86
Figure K.5. Item Drift Plot—Science Grade 9 87
Figure K.6. Item Drift Plot—Science Grade 10 87

1. Introduction

This report provides details of the maintenance of the Utah Aspire Plus testing system at grades 9 and 10 for reading, mathematics, and science. In addition to a general overview that provides a frame of reference around key attributes of the assessments, the report provides details around development of items and test forms, the administration of operational tests, the scoring and reporting for all tests, and the maintenance of existing scales. Throughout the report, the narrative is intended to present an interpretive argument whereby the various claims of the assessment system are identified and described throughout the test development process from creation through administration and score reporting. Technical details address test design, development and implementation, test administration, test taker characteristics, classical item analyses, reliability analyses, item response theory (IRT) calibration, equating, and scaling, quality control procedures, and evidence of validity.

1.1. Assessment Overview

Utah Aspire Plus is a computer-based, fixed-form summative assessment that evaluates the knowledge and skills students should have by the end of grades 9 and 10 in reading, mathematics, and science. It is a hybrid assessment created through collaboration with Utah educators, the Utah State Board of Education (USBE), and Pearson using ACT Aspire and Utah test items to allow for alignment to the Utah Core Standards, calculation of student growth scores, and predictive scoring for ACT®, Utah's college and career readiness assessment. The assessment contains approximately 50% items from the Utah test bank and 50% from ACT Aspire. The Utah Aspire Plus assessments are designed for several purposes:

- Measure the breadth and depth of the Utah Core Standards and measure across all levels of student performance
- Provide awareness of individual achievement in relation to stated performance expectations
- Provide evidence of whether students are on track for college and career readiness
- Evaluate growth between grades 9 and 10

The reading assessment requires students to demonstrate comprehension skills, understand tone and point of view of texts, and evaluate texts with different types of text sources. The mathematics assessment assesses students in two general levels of content: Secondary Math I that extends the mathematics from the middle grades, particularly on linear and exponential relationships, and Secondary Math II that focuses on quadratic relationships and comparing them to the linear and exponential relationships from Secondary Math I. The science assessments are composed of test units designed to measure multi-dimensional knowledge and skill interactions across different scientific phenomena within core disciplines based on the Next Generation Science Standards (NGSS).

The Utah Aspire Plus assessments include multiple-choice, multiple-select, technology-enhanced, and evidence-based selected-response item types. Students receive predicted ACT score ranges for each subtest and a predicted composite ACT score range, as well as overall subject area scale scores for end-of-grade-level expectations that classify students into one of four performance levels: *Below Proficient*, *Approaching Proficient*, *Proficient*, and *Highly Proficient*.

1.2. Background

Spring 2019 marked the first administration of the Utah Aspire Plus assessments in English, reading, mathematics, and science and the creation of base reporting scales and performance level cut scores for each respective grade and subject assessment. The Utah Aspire Plus Science assessments aligned to the with Engineering Education (SEEd) standards were administered to Utah students for the first time in spring 2021 as an operational field test, meaning that items used to provide scores for students were identified after the administration. That identification activity was akin to the standard test construction process involving Pearson and USBE content experts and psychometricians working collaboratively to identify the best forms based on a match to the blueprints and statistical indices. After these forms were determined, they were used to set the cut scores during the standard setting in August 2021.

Prior to 2019, students were assessed on the Utah Core Standards through the Utah Student Assessment of Growth and Excellence (SAGE) assessment program. Utah Aspire Plus is an extension of Utah SAGE, still intended to measure student performance in relation to the Utah Core Standards but also to measure students' preparedness for meeting college readiness benchmarks. As such, the assessment content from Utah SAGE is used as one component of the Utah Aspire Plus assessments. Additional content from ACT Aspire is used to provide predictions of performance on the ACT. This content also aligns to the Utah Core Standards and is counted toward Utah Aspire Plus scores. The ACT is the primary college readiness assessment submitted to local universities in Utah. As such, the Utah Aspire Plus assessments incorporate items from the ACT Aspire assessments that contribute to students' overall test scores and are used to provide a predictive indicator of performance on the ACT®. Students receive predicted ACT® score ranges for each ACT® subtest (reading, mathematics, and science), as well as an overall predicted composite ACT® score range.

The spring 2020 administration was cancelled due to the COVID-19 pandemic and a waiver of the Utah Aspire Plus assessment requirements based on Senate Bill 3005 that was passed during the Utah Legislature's 3rd Special Session of 2020 and signed into law on April 22, 2020. Testing resumed in spring 2021 using the test forms developed for the 2020 administration, with the accountability, school identification, and related reporting requirements waived for the 2020–2021 school year based on approval from the U.S. Department of Education. Remote administration was permitted for qualifying students for the first time in spring 2024, which also marked the first time Pearson's Assessment Delivery and Management (ADAM) web application was used to manage online student testing and test data.

The English assessment was removed from the Utah Aspire Plus assessment program beginning in spring 2025 due to changes in the educational standards that resulted in a total test time reduction of 45 minutes. This change aligned the assessment more closely with the updated Utah Core Standards and aimed to streamline testing while maintaining alignment with college and career readiness benchmarks.

1.3. Testing Requirements

The Utah Aspire Plus assessments were created out of Utah Code 53E-4-304 that requires USBE to administer assessments that are predictive of college readiness at grades 9 and 10 in addition to providing overall performance scores and proficiency indicators for reading, mathematics, and science. The statute also requires the assessments to provide overall scores as indicators of end-of-grade-level expectations for students in grades 9 and 10 and performance level indicators (*Below Proficient*, *Approaching Proficient*, *Proficient*, and *Highly Proficient*) in each subject.

The Utah Aspire Plus assessments are designed for students completing their grades 9 and 10 courses in reading, mathematics, and science. The reading test is designed to assess the skills that students in grades 9 and 10 should have by the end of those respective years, the mathematics tests are designed to assess the skills that grade 9 (Secondary Math I) and grade 10 (Secondary Math II) students should have by the end of those respective years, and the science tests are designed to assess the skills that students in grades 9 and 10 taking Biology, Chemistry, Earth Science, or Physics should have by the end of instruction (regardless of the specific course).

All students are expected to participate in the state accountability system. This principle of full participation includes English learner (EL) students, students with an Individualized Education Program (IEP), and students with a Section 504 plan. Both state and federal laws require that all students be administered assessments intended to hold schools accountable for the academic performance of students, including state statutes that regulate Utah's accountability systems and federal laws including the Every Student Succeeds Act of 2015 (ESSA) and the Individuals with Disabilities Education Improvement Act of 2004 (IDEA). The Utah Aspire Plus tests are provided to account for a range of accessibility features for all testers and accommodations for students with disabilities as determined by an IEP, Section 504, or EL plan team.

2. Test Design

This section describes the framework guiding the development of the Utah Aspire Plus assessments, including the content standards, item types, test blueprints, and structure. This description of the intended construct and rationale behind the assessments supports the valid interpretations of the test scores and alignment with Utah's standards-based educational system.

2.1. Content Standards

The Utah Plus Aspire assessments are aligned to the Utah Core Standards, available online at https://www.uen.org/core/. The Utah Core Standards for reading in grades 9 and 10 aim to prepare students for college and career readiness by developing proficient reading skills across various disciplines, designed to enhance students' abilities to comprehend and analyze complex texts across various genres. The Utah Core Standards for mathematics in grades 9 and 10 are designed to promote procedural fluency and conceptual understanding and the ability to apply mathematics in real-world contexts. High school mathematics in Utah is structured into integrated courses. For grades 9 and 10, students typically engage in Secondary Mathematics I and II that encompass a blend of algebra, geometry, and statistics. Utah's high school science curriculum is guided by the SEEd standards that emphasize an integrated approach to science education structured around three dimensions:

- Science and Engineering Practices (SEPs): Engaging students in practices such as asking questions, developing models, planning and carrying out investigations, analyzing and interpreting data, constructing explanations, and designing solutions.
- Crosscutting Concepts (CCCs): Helping students explore concepts that bridge disciplinary boundaries, including patterns, cause and effect, scale, proportion and quantity, systems and system models, energy and matter, structure and function, and stability and change.
- Disciplinary Core Ideas (DCIs): Focusing on key concepts in Life Science, Physical Science, and Earth and Space Science

The SEEd standards were written by Utah educators and scientists using a wide array of resources and expertise including *A Framework for K–12 Science Education* (NRC, 2012), the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), and related works. These standards were written with students in mind, including developmentally appropriate progressions that foster learning that is age-appropriate and enduring. The aim was to address what an educated citizenry should know and understand to embrace the value of scientific thinking and make informed decisions. The SEEd standards are founded on what science is, how science is learned, and the multiple dimensions of scientific work.

2.2. Claims and Subclaims

The assessments are designed to measure student performance in categories called claims and subclaims that are based on the structure of the Utah Core Standards and frame the design and development of the summative tests at grades 9 and 10, as shown in Table 2.1. The primary claim reflects the overall performance goal for the subject assessments (e.g., overall reading performance) reported as an overall scale score and performance level, while the subclaims further explicate what is measured on the assessments based on the blueprints. For science, the subclaims are expressed across the Life Science, Earth and Space Science, and Physical Science DCIs. It is important to note that subclaims only provide structure within the respective blueprints and are not reported at the individual student level. Only the claims are reported on individual student reports (ISRs).

Table 2.1. Claims and Subclaims

Subject	Claims	Subclaims
Reading	1. Student performance reflects an indicator of career and college readiness as demonstrated through students' ability to read and comprehending complex informational and literary texts as expected to have been attained by the end of each respective year as a prediction of performance on the ACT® Reading test. 2. Overall performance reflects students' understanding of reading and comprehending complex informational and literary texts with respect to the breadth and depth of the Utah Core Standards and measures across all levels of student performance.	• Key Ideas • Craft and Structure • Integration of Knowledge and Ideas
Math	1. Student performance reflects an indicator of career and college readiness as demonstrated through students' ability to understand linear relationships, abstract and quantitative reasoning, and problem solving as expected to have been attained by the end of each respective year as a prediction of performance on the ACT® Math test. 2. Overall performance reflects students' understanding of linear relationships, abstract and quantitative reasoning, and problem solving with respect to the breadth and depth of the Utah Core Standards and measures across all levels of student performance.	• Math I (Grade 9) – Algebra – Functions – Geometry – Statistics and Probability • Math II (Grade 10) – Number and Quantity – Algebra – Functions – Geometry – Statistics and Probability
Science	1. Student performance reflects an indicator of career and college readiness as demonstrated through students' ability to understand and apply science as defined by the SEEd standards. Further, as expected to have been attained by the end of each respective year as a prediction of performance on the ACT® Science test. 2. Overall performance reflects students' understanding of science as defined by the SEEd standards with respect to the breadth and depth of the Utah Core Standards and measuring across all levels of student performance.	• Gathering and Investigating – SEPs: Asking questions and defining problems; Obtaining, evaluating, and communicating information; Planning and carrying out investigations – CCCs: Patterns; Cause and effect; Systems and system models; Energy and matter; Structure and function; Stability and change Use Science Process and Thinking Skills • Developing Models – SEPs: Developing and using models – CCCs: Patterns; Cause and effect; Scale, proportion and quantity; Systems and system models; Energy and matter; Stability and change • Using Mathematical Thinking – – SEPs: Analyzing and interpreting data; Using mathematics and computational thinking – CCCs: Patterns; Cause and effect; Scale, proportion, and quantity; Systems and system models; Energy and matter; Stability and change • Constructing Explanations – – SEPs: Constructing explanations and designing solutions; Engaging in argument from evidence – CCCs: Patterns; Cause and effect; Systems and system models; Energy and matter; Structure and function; Stability and change

2.3. Test Blueprints

The Utah Aspire Plus test blueprints, available online at https://utah.mypearsonsupport.com/admin-resources.html, define the components of the Utah Aspire Plus assessments that reflect the breadth of the Utah Core Standards across different levels of student understanding. The creation of test blueprints was driven by the assessments' intended purposes to support the respective claim structures. The blueprints are the distribution of item types across domains/reporting categories, level of cognitive demand, and the number of total points associated with each. For the science tests, the SEEds blueprints assume a design in which one of the three DCIs are assessed by two clusters and the other two DCIs with a single cluster. Coverage of the respective DCIs rotates across forms (either within a given year or across years) to ensure that the standards are fully represented over time.

The initial blueprints were developed in 2018 by Utah educators who participated in an educator blueprint review where they reviewed the content standards (including content breakout categories) in addition to the Utah SAGE and ACT Aspire test blueprints. The agenda, training presentation slides, and summary of the blueprint meeting are provided in the 2018–2019 technical report (Pearson, 2020). Panelists were chosen to reflect Utah's educator populations by subject according to characteristics such as grades and subjects taught, years of teaching, rural/suburban/urban district, experience with test development, regular/charter, special education experience, and English as a second language endorsement. During review and discussion of these materials, educators provided recommendations for creation of blueprints that would support the intended claims and appropriately sample content that covered the respective standards. They recommended content domain coverage, item type distribution, overall number of items and points for each test, and testing time. At the conclusion of the test blueprint workshops, Pearson and USBE reviewed the recommendations and finalized the test blueprints for each Utah Aspire Plus assessment.

2.4. Item Types

The Utah Aspire Plus tests are composed of several different types of items to measure student performance, including multiple choice, multiple select, evidence-based selected response, and technology enhanced. Multiple-choice items present students with four or five responses, of which there is one correct answer. Multiple-select items require students to select two or three correct choices from several presented choices. Evidence-based selected response items have two parts: Part A is designed as an *identification* component, where Part B is designed to elicit an *evidence*-based component. These types can also be designed as two multiple-choice items, or a combination of multiple-choice and technology-enhanced items. Technology-enhanced items require specialized interactions within the online presentation for capturing student responses.

2.5. Cognitive Complexity

Cognitive complexity refers to the cognitive demand associated with interacting with a given item/task. It is assessed for the Utah Aspire Plus assessments using Webb's Depth of Knowledge (DOK; Webb, 1997) framework for reading and mathematics that categorizes tasks based on the level of cognitive demand required, focusing on the type and level of thinking and reasoning required to answer a given item correctly or earn the most points and ranging from simple recall to complex reasoning and analysis across four levels: (1) Recall and Reproduction, (2) Skills and Concepts, and (3) Strategic Thinking and Reasoning, and (4) Extended Thinking. This framework ensures that test items assess a spectrum of thinking skills, from basic comprehension to strategic thinking.

During test development, each item is aligned with a specific DOK level to match the intended cognitive demand. This alignment is crucial for evaluating the depth of student understanding and ensuring that assessments accurately reflect the Utah Core Standards. The integration of DOK levels into the assessment design supports a comprehensive evaluation of student proficiency across various cognitive processes.

The science assessment does not use the DOK framework to measure cognitive complexity. Instead, it aligns with the SEEds standards based on the NGSS that emphasize a multidimensional approach to science education, integrating SEPs, CCCs, and DCIs. This multidimensional framework requires test items that capture the interplay between these components, which is not adequately addressed by the linear progression of cognitive demand in the DOK model. Therefore, the Utah Aspire Plus science assessment employs a structure that reflects the complexity and integrative nature of modern science education, focusing on students' ability to apply knowledge and skills across various contexts.

2.6. Test Structure and Testing Time

Table 2.2 presents the number of operational and embedded field test items on each Utah Aspire Plus test form administered in spring 2025, as well as the allotted testing time. The previous SAGE tests were untimed. To support the derivation of predictive scores on the ACT, the Utah Aspire Plus assessments follow the same fixed testing time conditions. Students whose IEP, Section 504, or EL plan specified an accommodation for extended time were able to use extended time accommodations as appropriate.

Table 2.2. Test Form Composition

Assessment	#OP Items	#FT Items	Testing Time
Reading 9	35	10	75 minutes
Reading 10	35	10	75 minutes
Mathematics 9	40	3	75 minutes
Mathematics 10	40	3	75 minutes
Science 9	23	5–6	60 minutes
Science 10	23	5–6	60 minutes

3. Test Development

This section describes the form construction processes for the spring 2025 Utah Aspire Plus test forms. All available content for creation of the assessments was based on the existing item banks with Utah SAGE and ACT Aspire content. For reading and mathematics, the ACT Aspire forms are alternated each year to help limit exposure of the ACT Aspire content that might otherwise negatively impact ACT predication score activities.

Each subject area assessment had one core operational form for regular online and text-to-speech (TTS) forms. Individual test scores are based on the operational items only. The science forms also consisted of a small set of embedded field test items, as shown in Table 3.1. In addition to the core forms used for TTS, three other accommodated forms were available: non-screen reader, screen reader, and Spanish. The science grade 10 core and accommodated forms were a reuse of the 2022 forms.

Table 3.1. Spring 2025 Field Test Forms

Assessment	#FT Forms
Science 9	12
Science 10	12

3.1. Operational Forms Development

Test form construction is the process of selecting and sequencing the operational and field test items for each test form, which is a complex, interactive task that requires both content and psychometric expertise. The construction of test forms was a coordinated effort between experts from USBE, Pearson, and ACT. This process required adhering to guidelines that promote fair and ethical testing practices. Using the content developed to measure the Utah Core Standards, specialists worked through an iterative process to evaluate the specific items, passages, and stimuli that best met the intended measurement targets and to support all stated claims.

The Utah Aspire Plus assessments are designed so that test scores can be linked to ACT scales to provide students with indicators of being prepared for meeting college readiness benchmark. To accomplish this, approximately 50% of the Utah Aspire Plus tests (less for mathematics) are composed of items from ACT Aspire. The general test development process was initiated with the selection of items from ACT based on a match to the blueprints and statistical indicators of item quality and fairness provided from the SAGE and ACT Aspire banks. ACT Aspire items were positioned within each form in the same locations as originally administered within the ACT Aspire forms to help facilitate the derivation of the predictive scores on Utah Aspire Plus.

The test construction procedure was an iterative process whereby the first proposed form was evaluated by each party (Pearson, USBE, and ACT) for content and psychometric quality, feedback provided, and revisions made until a final version was approved by all.

3.2. Statistical Guidelines

While the initial Utah Aspire Plus tests were primarily driven by content considerations, statistical indices were available based on use within the SAGE and ACT Aspire Plus assessments. For creation of the Utah Aspire Plus assessments, the following general guidelines were used to help support selection of a range of item difficulties and evaluate item quality to ensure the best overall test forms:

- **Target item difficulty range of between 0.30 and 0.85.** Based on *p*-values that reflect the proportion of students correctly responding to the item. Items awarding more than 1 point used the item mean divided by the maximum points possible to place on the *p*-value metric.
- **Target threshold for item discrimination of 0.20 and above.** Where item discrimination is defined by item-total score correlations.
- **Extreme differential item functioning (DIF) indices should be avoided.** A standard flagging convention indicates differences of magnitude and classifies the most extreme cases of DIF as "C," moderate DIF as "B," and minor to no DIF as "A." Items flagged "C" should be avoided and minimal use of items flagged "B" should be used and/or balanced within a form where possible.

Refer to Section 8 for details on these statistics. Item bank limitations meant there were instances where items with poorer statistical indices were included to meet the blueprint. These were infrequent and, in all cases, deemed reasonable in supporting the intended claims without negative impact. Moving forward, newly developed content will fill gaps and address such limitations as the assessments mature.

3.3. 2025 Match to Test Blueprint

Table 3.2 – Table 3.4 present the match between the final 2025 Utah Aspire Plus operational test forms and the test blueprints. All operational forms matched all targets by item type, DOK, and reporting category. For additional information on the 2025 operational forms, Appendix A contains a breakdown reporting categories and standards by item type and DOK (except for science that does not use DOK).

Table 3.2. Operational Test Blueprint Match—Reading

Grade	Component	#Items	Blueprint Target %	2025 Form %
9	Evidence-Based Selected Response	3–6	9–17%	14%
	Multiple Choice	22–30	63–86%	74%
	Technology Enhanced	2–7	6–20%	11%
	DOK Level 1	4–7	11–20%	17%
	DOK Level 2	14–20	40–57%	46%
	DOK Level 3	12–15	34–43%	37%
	Key Ideas	12–16	34–46%	46%
	Craft and Structure	12–18	34–51%	37%
	Integration of Knowledge and Ideas	3–7	9–20%	17%
10	Evidence-Based Selected Response	3–6	9–17%	14%
	Multiple Choice	22–30	63–86%	71%
	Technology Enhanced	2–7	6–20%	14%
	DOK Level 1	4–7	11–20%	17%
	DOK Level 2	14–20	40–57%	46%
	DOK Level 3	12–15	34–43%	37%
	Key Ideas	12–16	34–46%	43%
	Craft and Structure	12–18	34–51%	40%
	Integration of Knowledge and Ideas	3–7	9–20%	17%

Table 3.3. Operational Test Blueprint Match—Mathematics

Grade	Component	#Items	Blueprint Target %	2025 Form %
9	Multiple Choice	30–33	75–83%	78%
	Technology Enhanced	7–10	18–25%	23%
	DOK Level 1	8–12	20–30%	30%
	DOK Level 2	15–20	38–50%	48%
	DOK Level 3	9–13	23–33%	23%
	Algebra	9–11	23–28%	28%
	Functions	10–12	25–30%	28%
	Geometry	9–11	23–28%	25%
	Statistics and Probability	7–9	18–23%	20%
10	Multiple Choice	30–33	75–83%	75%
	Technology Enhanced	7–10	18–25%	25%
	DOK Level 1	8–12	20–30%	30%
	DOK Level 2	15–20	38–50%	48%
	DOK Level 3	9–13	23–33%	23%
	Number and Quantity	2–4	5–10%	10%
	Algebra	9–11	23–28%	25%
	Functions	10–12	25–30%	25%
	Geometry	11–13	28–33%	33%
	Statistics and Probability	2–4	5–10%	8%

Table 3.4. Operational Test Blueprint Match—Science

Grade	Component	#Items	Blueprint Target %	2025 Form %
9	Multiple Choice	18–21	78–91%	78%
	Technology Enhanced	3–6	13–26%	22%
	DCI: Life Science	4–8	17–35%	26%
	DCI: Earth and Space Science	4–8	17–35%	48%
	DCI: Physical Science	9–13	39–57%	26%
	Gathering & Investigating	4–8	17–35%	26%
	Developing Models	4–8	17–35%	26%
	Using Mathematical Thinking	5–9	22–39%	22%
	Construct Explanations	5–9	22–39%	26%
10	Multiple Choice	18–21	78–91%	87%
	Technology Enhanced	3–6	13–26%	13%
	DCI: Life Science	4–8	17–35%	52%
	DCI: Earth and Space Science	9–13	39–57%	22%
	DCI: Physical Science	4–8	17–35%	26%
	Gathering & Investigating	4–8	17–35%	30%
	Developing Models	4–8	17–35%	22%
	Using Mathematical Thinking	5–9	22–39%	17%
	Construct Explanations	5–9	22–39%	30%

4. Test Administration

The spring 2025 Utah Aspire Plus test administration window was March 3 – May 9, 2025. The assessments were administered online as fixed forms, with accommodated forms available as needed. The online administration took place in TestNav, Pearson's online testing platform. ADAM was the student test management portal that test administrators used to manage student registrations and order materials if needed.

Each local education agency (LEA) was responsible for determining school testing schedules. Subject tests did not have to be administered in any prescribed order but could *not* be divided into multiple sessions. Utah Aspire Plus could be administered on a subject-by-subject basis or as a complete battery with all tests administered in one sitting, but each subject test was to be administered in one sitting (i.e., once a subject test was started, it had to be completed within that sitting).

Comprehensive details of the Utah Aspire Plus test administration are provided on the Admin Resources page at https://utah.mypearsonsupport.com/admin-resources.html. These resources cover all policies, procedures, specifications, training, instructions, security, accommodations, and oversight for the Utah Aspire Plus test administration. These resources address those responsible for carrying out the administration for all students, provide tools for educators and students to become familiar with the assessments (e.g., via practice tests), and guide the interpretation of test scores.

4.1. Test Security

The Utah Aspire Plus assessments are secure tests that follow the test blueprints for each assessed subject area. All test items are secure and may not be reviewed with students, discussed as a class, or reviewed during instructional conversations. Discussing, reviewing, recording, or transcribing test items in any format is a violation of test security. All test security requirements of Utah Aspire Plus must be met, and personnel involved in test administration must complete testing ethics training. The *Standard Test Administration and Testing Ethics Policy for Utah Educators* is available online at https://schools.utah.gov/assessment/ under "Testing Ethics." The LEA Assessment Director is responsible for ensuring that each student has an appropriate opportunity to demonstrate knowledge, skills, and abilities related to the Utah Aspire Plus grade-based courses and assessments to ensure that each student has a standardized (similar and fair) testing experience for a given assessment.

During the online test administration, Pearson's operational monitoring practices and tools constantly verify that platforms remain available to users, performance stays within acceptable limits, and users do not encounter critical errors. The types of monitoring Pearson performs to help keep testing on time and reduce the chance of interruptions include the following:

- Site Availability Monitoring: Checking locations and providing alerts when response times or availability thresholds are crossed
- Synthetic User Monitoring: Simulating key end-user actions (e.g., launching a test, logging into the administrative site, viewing reports) and running from several locations on the public internet
- End User Monitoring: Analyzing page and click performance to verify that end users receive results in a reliable and timely manner
- Server Monitoring: Collecting detailed metrics on server performance to gauge health

- Application Performance Monitoring: Gathering detailed performance information about the health of Pearson's various assessment platforms
- Database Monitoring: Using a variety of tools to watch performance in real time
- Event Monitoring and Real-Time Security Auditing: Processing large volumes of machine-generated data in real time to look for trends, issues, or anomalies
- Systems Vulnerability Monitoring: Monitoring multiple sources for newly identified vulnerabilities in systems and applications Pearson uses

4.2. Remote Proctoring

Remote testing is subject to several guidelines. First, for a student to be eligible for remote testing, 100% of their learning needs to be online. Students requiring a paper test (i.e., large print, Braille, human reader) are not eligible for remote testing. If a group of students are testing together in one proctor group, the maximum allowed number of students is 10. USBE policy required two proctors for every 10 students in a proctor group. Both proctors had to be in the same physical location and able to converse with each other during the entire testing session.

Remote proctoring works much the same way as proctoring or taking the test in a brick-and-mortar building. The TestNav platform for taking the test is the same for students, except for additional system checks to ensure that the camera and microphone are on. Students can digitally raise their hand if assistance is needed, which alerts the proctors on the ADAM platform. The proctors can send a chat message to the student, call the student, or broadcast messages to the entire group of students testing remotely. Proctors can see the students through their cameras. The students can also see a proctor, but they are not able to see the other students in the proctor group. Proctors can also monitor student progress through the ADAM system. Should a student lose connection or turn off the camera, the proctor will notice that they can no longer see the student and can immediately exit the student from the test until they are able to regain connection. Once connected, the test can be resumed and the student can be allowed to continue where they left off.

4.3. Test Accommodations and Supports

Accommodated test forms for the Utah Aspire Plus assessments include Spanish-language forms and forms with assistive technology. These forms are modified reproductions of the original test forms. Modifications primarily involve incorporation of the accommodation with the intent of otherwise preserving the item content in its original form. Assistive technology within online test forms includes speech-to-text, magnification, and adaptive keyboard and mouse. Paper accommodations are also offered in the form of standard-print, large print, and Braille reproductions. Testing accommodations and supports are outlined in the test administration manual (TAM), available online at https://utah.mypearsonsupport.com/admin-resources.html under "Administration Guides > User Guides." A complete list of accessibility features and accommodations can be found at https://www.schools.utah.gov/specialeducation/programs/accessibilityaccommodationsassessment.

For students requiring Braille, paper versions of the original forms are created, and student responses are transcribed into one of the assistive technology test formats. For items that are not able to be adopted as is, some modification must occur to create the accommodated parallel version. These are referred to as "sister" items and are created directly from the original item to preserve every aspect of the item as it is used in the original form and capture of student responses such that item characteristics are directly comparable. While this typically involves only a few items on a given assessment, the Spanish-language forms must be fully transadapted. This process involves translating a test form's English text to Spanish, but also of adapting the content to account for the linguistic and cultural differences between speakers of the two different languages.

Creation of all transadapted and sister items for the Utah Aspire Plus assessments follows a similar process of creation and review as the original items, with an emphasis on fully matching to the original item in terms of content and function. Highly qualified item writers with extensive expert content experience are involved in the creation and review process of transadapted and/or sister item creation, and several reviews are held throughout the creative process involving Pearson and USBE content and psychometric experts to ensure match to source.

Table 4.1 presents the embedded and non-embedded supports that are generally available to all students, whether through the online system or locally arranged, as outlined in the TAM. It also presents the testing accommodations that require prior designation in a student's IEP, 504, or EL plan, in addition to the supports.

Table 4.1. Testing Accommodations and Supports

Embedded Supports	Non-Embedded Supports	Accommodations
• In browser/app zoom • Answer eliminator • Calculator – Desmos graphing and Desmos scientific • Bookmarking items for review • Line reader mask • Color contrast • Answer masking • Highlighter • Keyboard navigation • Text-to-speech (English) • Directions reread (text-to-speech) • Text-to-speech (Spanish) • Personalized visual modification of remaining time	• Word to word dictionary • Scratch paper • Line reader • Supervised breaks within each day • Special seating/grouping • Location for movement • Separate/alternate location • Minimized distractions • Food or medication for individuals with medical needs • Administration and optimum time of day • Special lighting • Adaptive equipment/furniture • Wheelchair-accessible room	• Assistive technology – screen reader • Speech to text – assistive technology scribe • Other assistive technology • Spanish transadaptation • Online test translation – other languages than Spanish or English • Standard print • Large print • Braille plus tactile graphics • Extra time • Personalized auditory notification of remaining time • Breaks: stop the clock • Breaks: extending over multiple days • Human scribe • Home administration • Human reader • Signed exact English (directions only) • Sign language interpretation • Cued speech • Alternate mouse pointer • Zoom percentage • Abacus

4.4. Test-Taking Irregularities and Security Breaches

Table 4.2 describes the processes taken to address test-taking irregularities during the test administration (i.e., non-standard situations that affect one or more students). This includes students experiencing computer problems, experiencing a sudden illness, having to leave the room, or becoming unduly disturbed by the testing situation. Testing staff are trained to become familiar with the policy around unexpected/unforeseen circumstances prior to testing. Some students may be unable to participate in regular testing schedules due to absence, technical difficulties, or other unforeseen circumstances. Opportunities for these students to complete each assessment are provided within the school's testing window. If there is an emergency that interrupts testing for an entire class or school, decisions about whether a test could be started again are made on a case-by-case basis by working with the USBE assessment team.

Table 4.2. Test-Taking Irregularities and Security Breaches

Event	Description
Test Interruptions	If a student gets sick, has to leave and cannot return during the test, or for any other reason does not complete a test that has already begun, the test is to be concluded and submitted immediately. To maintain the security of the test items, students are not allowed to restart or take a test over again.
Scoring of Interrupted Tests	If a student is interrupted and completes only part of a test before it is concluded and submitted, the student may not receive a score. A student must attempt 85% of the items to receive a score. If a student does not attempt at least 85% of the items, a score cannot be generated, and no test score is reported for that assessment. Overall composite scores are not available for students who have missing subject test scores because the composite score is calculated using all three subject tests.
Wrong Test Form/Accommodation	If a student begins a test using a test form or accommodation they are not supposed to have, the teacher/proctor should immediately stop the test. A new test assignment must be created, and a new test administration can proceed as normal from that point.
Extended Time Accommodation Issues	Extended time accommodations must be applied before applying any participation code and before starting sessions. If the accommodation is applied after the session has been prepared and started, students receive a time-expired warning that has a link for "proctor only." At that point, a proctor can confirm the student should have extended time and is able to set the student up to continue testing as per their accommodation.
Test Invalidation	Tests can be invalidated when a student's performance is not deemed an accurate measure of their ability (e.g., the student cheats, uses inappropriate materials). When a test is invalidated, the student is not given another opportunity to take the test. Invalidating a test must be completed by the district testing administrator.

5. Scoring and Reporting

This section details the scoring and score reporting processes for the spring 2025 Utah Aspire Plus assessments. All items on the Utah Aspire Plus assessments were machine-scored (i.e., scored by the online testing system based on an answer key), with scale scores produced via item response theory (IRT) pattern scoring. Final test scores were available on the individual student reports (ISRs) on May 23, 2025, and in the Family Portal for parents/guardians beginning June 3, 2025. Score interpretation guides were available online at https://utah.mypearsonsupport.com/admin-resources.html under "Reporting Resources." The guide for parents/guardians includes an example of an ISR.

5.1. IRT Pattern Scoring

Scale scores are computed using a method known as pattern scoring that considers not only the number of correct responses but also the specific items answered correctly. Pattern scoring applies IRT to analyze each response pattern in conjunction with the characteristics of all items, resulting in a score that better reflects the student's estimated ability. Item parameters derived from previous IRT calibrations were used to estimate student ability (theta) scores based on response patterns. (Refer to Section 9 for information on the IRT models and calibration process.)

The software package used to perform scoring was Operational Scoring: IRT Score Estimation (ISE V1.3.f; Chien & Shin, 2012). This application produced student scores on the IRT scale using scored student responses and calibrated item parameters. Two data-driven input files were required to execute the ISE software: a student response file and an item parameter file. The ISE algorithm combines the Newton-Raphson and Brute Force methods to generate maximum likelihood estimates (MLE) of theta values. Configuration settings included theta bounds of +4 and –4, the number of iterations for the Newton-Raphson estimation method (30), the grid length interval for the Brute Force algorithm, the number of checking points for which the first derivatives are computed (120), and theta estimates reported to four decimal places.

IRT parameters for all 2025 Utah Aspire Plus operational items were used for estimating individual student scores across all forms. Table 5.1 presents summary statistics for the IRT (*a*- and *b*-) parameter estimates, including the total number of items and the mean, standard deviation (SD), minimum, and maximum for each parameter.

Table 5.1. IRT Summary Parameter Estimates for Operational Items

Assessment	#Items	*a* Mean	*a* SD	*a* Min.	*a* Max.	*b* Mean	*b* SD	*b* Min.	*b* Max.
Reading 9	35	0.83	0.38	0.29	1.67	0.10	1.02	-1.87	2.17
Math 9	40	1.01	0.31	0.33	1.70	0.27	0.66	-1.19	1.45
Science 9	23	0.78	0.23	0.39	1.17	0.33	0.58	-0.83	1.42
Reading 10	35	1.09	0.52	0.17	2.09	-0.07	0.89	-2.12	2.38
Math 10	40	1.07	0.27	0.44	1.45	0.28	0.63	-1.09	1.43
Science 10	23	1.01	0.70	0.26	3.11	0.71	0.71	-0.02	2.65

5.2. Interpretation of Test Scores

Student performance is reported on the ISR using a composite scale score, predicted ACT scores, postsecondary readiness prediction, and subject area overall scale scores and performance levels. No subscores are reported on the ISRs.

5.2.1. Composite and Subject Area Scale Scores

For each subject area, student results are expressed as an overall scale score ranging from 100 to 300. Scale scores are used to report scores for all students on a scale that remains consistent across multiple years or forms and facilitate accurate comparison of test results over different administrations. A student's composite scale score is the average of all the Utah Aspire Plus assessments (reading, mathematics, and science) and is provided for students who take all three tests.

5.2.2. Performance Levels and PLDs

Based on a student's overall subject area scale score, they are classified into one of four performance levels that indicate how well a student has met the expectations or standards set for a particular grade or subject: Level 1: *Below Proficient*, Level 2: *Approaching Proficient*, Level 3: *Proficient*, and Level 4: *Highly Proficient*. The ranges for performance levels for each subject and grade were set during the standard setting process based on cut scores (as described in Section 6). Students are assigned a performance level when their scale scores fall within the scale score range of the associated performance level.

Each performance level has associated performance level descriptors (PLDs) that describe the knowledge and skills that students should know and be able to demonstrate at each performance level in each subject area as articulated in the Utah Core Standards (e.g., the set of statements describing what it means for a grade 9 student to have demonstrated proficiency in reading). PLDs can be used to articulate what students are expected to know and be able to do at each performance level; guide the development of cut scores during standard setting; inform item writers and test developers by clarifying the intended rigor and content coverage for each performance level, helping to ensure that test items align with the targeted performance expectations; and enhance score reporting by providing a qualitative description of what a test score means.

5.2.3. ACT Predicted Scores

A goal of the Utah Aspire Plus assessments is to be predictive of college readiness at grades 9 and 10. As such, students' test scores on Utah Aspire Plus are linked to ranges of predicted ACT scores for each subject area test (reading, mathematics, and science) and the composite score (the average of the three subject tests) and reported on the ISR. Students can use the predicted scores together with the ACT College Readiness Benchmarks to monitor their preparedness to be college-ready by the end of high school. Utah students take the ACT® during their junior year of high school.

The predicted ACT score ranges are determined through a statistical linking process. Predicted ranges of performance were originally determined between ACT Aspire and ACT scores, where for a given ACT Aspire score, there was a distribution of related ACT scores. The bounds of the range were denoted by the scores closest to the 25th and 75th percentiles of the ACT score distribution, conditional on ACT Aspire scores. For Utah Aspire Plus, an additional error term was added to account for error attributable to linking the Utah Aspire Plus scores.

To provide the predicted performance on the ACT tests in the first two administration years, a linking study was performed between scale scores of the Utah Aspire Plus assessments and established ACT score predictions for ACT Aspire tests, facilitated through common items between Utah Aspire Plus and ACT Aspire test forms (Pearson, 2020, Appendix J). The result of this linking study was a set of predicted ACT score ranges across the Utah Aspire Plus score scale (100–300) for each Utah Aspire Plus assessment. The predicted ACT score ranges have continued to be updated as longitudinal data become available to link the Utah Aspire Plus scores of grade 9 and 10 to ACT scores at grade 11, beginning with a second longitudinal study conducted in spring 2021 (Pearson, 2021, Appendix J) and a third longitudinal study conducted in 2022 to update the science grade 10 predictions (Pearson, 2022, Appendix H). These technical reports are available online at https://utah.mypearsonsupport.com/admin-resources.html under "Reporting Resources."

5.3. Appropriate Uses for Scores and Reports

Test forms constructed for Utah Aspire Plus cover a sampling of content as specified through test blueprints and reflective of the Utah Core Standards. The resulting scores reflect overall performance for each subject area based on expectations of students' knowledge at the end of grades 9 and 10. While each test covers the standards, there is a limit to incorporating everything (e.g., given test time limits). Test scores should only be interpreted and used in the context from which they are obtained. In other words, Utah Aspire Plus test scores should be used to describe student achievement on the content assessed (i.e., grade level) and not used to generalize achievement beyond the test. In addition, academic placement decisions and promotions should not be based solely on these test scores but should include other indicators of achievement.

The ISR communicates an individual student's test scores and interpretations of achievement based on those scores. The ISR provides a snapshot of achievement and explains the meaning of each piece of information provided, providing valuable information to students and parents. It is important that users of these reports do not extend the score information beyond the interpretations provided. A guide for understanding the ISR and its components can be found at https://utah.mypearsonsupport.com/admin-resources.html under "Score Interpretation Guide."

6. Standard Setting

Performance standards relate levels of performance on an assessment to what students are expected to learn by separating an assessment's score scale into performance levels. Standard setting is the process of establishing the cut scores that define the performance levels for an assessment.

Cut scores must be established following the first administration of a new assessment to ensure that student performance is properly categorized into performance levels. As such, a standard setting with Utah educators took place from August 6–9, 2019, to recommend cut scores for the Utah Aspire Plus assessments using the Extended Modified Yes/No Angoff method (Davis & Moyer, 2015; Plake et al., 2005). Another standard setting following the same method was conducted from August 9–12, 2021, for science following the first administration of the new assessment based on the SEEd standards. While there were some changes to reading standards in 2025, USBE determined that these changes were not significant enough to require a new standard setting, following a recommendation from the Technical Advisory Committee (TAC). For full details on the standard setting process, please refer to the standard setting reports (Pearson, 2019, 2021).

Table 6.1 presents the resulting Utah Aspire Plus subject area scale score cut scores (i.e., the minimum score students must receive to be classified into a certain performance level), as shown in bold.

Table 6.1. Scale Score Ranges and Cut Scores

Assessment	Level 1: *Below Proficient*	Level 2: *Approaching Proficient*	Level 3: *Proficient*	Level 4: *Highly Proficient*
Reading 9	100–165	**166**–203	**204**–230	**231**–300
Reading 10	100–174	**175**–203	**204**–234	**235**–300
Math 9	100–171	**172**–205	**206**–232	**233**–300
Math 10	100–180	**181**–209	**210**–235	**236**–300
Science 9	100–186	**187**–210	**211**–236	**237**–300
Science 10	100–186	**187**–209	**210**–239	**240**–300

7. Administration Results

This section presents the number of students who took the spring 2025 Utah Aspire Plus assessments, along with a summary of their results.

7.1. Test Taker Characteristics

Table 7.1 provides the participation rates for each Utah Aspire Plus test by subgroup. These are students that received a valid test score on a subject test. Cases that did not have a valid test score were excluded from being counted.

Table 7.1. Spring 2025 Participation Rates

Subgroup	Reading G9	Reading G10	Math G9	Math G10	Science G9	Science G10
Total #Students Scored	45,079	43,615	43,894	42,439	45,006	43,308
Female	48.48	47.72	48.17	47.46	48.45	47.67
Male	51.52	52.28	51.83	52.54	51.55	52.33
Hispanic or Latino Ethnicity	20.31	20.50	20.09	20.24	20.42	20.54
Asian	1.76	1.78	1.75	1.79	1.77	1.78
Native Hawaiian or Other Pacific Islander	1.43	1.32	1.40	1.36	1.43	1.31
Black or African American	1.35	1.34	1.34	1.29	1.37	1.30
American Indian or Alaska Native	0.93	0.92	0.92	0.91	0.93	0.90
White	70.64	70.68	70.90	70.95	70.46	70.70
Other	3.58	3.47	3.61	3.46	3.61	3.47
Limited English Proficient – No	92.08	92.18	92.04	92.22	91.97	92.12
Limited English Proficiency – Yes	7.92	7.82	7.96	7.78	8.03	7.88
Economic Disadvantaged – No	74.20	75.34	74.34	75.62	73.99	75.32
Economic Disadvantaged – Yes	25.80	24.66	25.66	24.38	26.01	24.68
Special Education – No	90.10	90.60	89.96	90.51	90.06	90.62
Special Education – Yes	9.90	9.40	10.04	9.49	9.94	9.38

7.2. Testing Time

After the test administration, student total testing time was analyzed for each test to gauge the extent to which the time allotted appears to be reasonable. Table 7.2 presents a breakdown of student testing time across the full range of testing times. (See Table 2.2 in Section 2.6 for the allotted testing times for the Utah Aspire Plus assessments.) Students needing extra time based on their IEP, Section 504, or EL plan fall into three categories: time and a half, double time, or triple time. The percentile rankings indicate the amount of time in minutes students took to complete the respective test. For example, the results for the 95th percentile (P95) for grade 9 reading students using regular time indicate that 95% of students finished the assessment in 72.5 minutes. Overall, students completed the assessments within the recommended testing times.

Appendix B presents a graphical display (box-and-whisker plot) of student testing time for each assessment. Box-and-whisker plots present the same information at each respective quartile, where the middle 50% of the given distribution is the box, and the whiskers represent the bottom 25% and top 25% of the distribution. Dots represent outliers and reflect very few overall cases. Most outliers for regular testers are still within the time allotment for the subject. For example, the outliers for grade 9 reading for regular testers are all below the 90-minute time threshold. Based on these data and plots, the evidence suggests that students in general had enough time to complete each respective test within the given allotments.

Table 7.2. Spring 2025 Student Testing Time (in minutes)

Assessment	Time	#Students	Min.	Max.	Mean	SD	P50	P75	P80	P85	P90	P95
Reading 9	Regular Time	40,446	0.9	74.7	48.2	15.6	49.2	59.9	62.4	65.3	68.7	72.5
	Time and a Half	4,043	1.6	112.0	50.0	24.8	47.2	65.5	70.1	76.5	85.9	97.6
	Double Time	504	3.1	148.8	52.7	28.8	48.4	68.8	74.3	81.9	92.1	110.2
	Triple Time	86	4.7	224.4	56.7	38.8	48.6	76.6	81.6	90.4	111.6	141.5
Reading 10	Regular Time	38,739	0.7	74.5	42.0	16.2	42.5	53.5	56.3	59.3	63.0	68.6
	Time and a Half	4,568	1.1	112.0	41.3	22.2	39.1	54.1	58.2	63.5	70.3	83.3
	Double Time	218	0.9	143.8	50.2	27.5	46.8	64.3	67.5	75.3	81.2	109.7
	Triple Time	90	2.0	223.7	44.4	33.6	39.6	57.0	59.6	67.5	85.8	89.1
Math 9	Regular Time	39,229	1.2	74.4	52.0	15.4	54.0	64.1	66.3	68.7	70.9	73.0
	Time and a Half	4,080	1.8	111.8	51.2	23.9	50.1	66.6	71.3	76.6	83.6	94.5
	Double Time	502	2.0	148.4	52.9	28.2	49.8	67.8	72.3	82.6	91.2	105.6
	Triple Time	83	7.0	149.8	59.7	29.1	59.4	73.3	84.0	87.4	91.1	102.1
Math 10	Regular Time	37,607	1.0	74.3	46.9	17.5	48.8	60.6	63.2	66.1	69.2	72.2
	Time and a Half	4,534	1.8	111.8	43.0	23.0	40.7	56.8	61.4	66.9	74.0	85.5
	Double Time	209	3.0	148.1	56.5	30.4	51.9	72.9	79.0	88.4	97.8	118.1
	Triple Time	89	4.7	175.0	54.2	30.5	49.3	64.1	68.8	78.4	98.1	106.4
Science 9	Regular Time	40,408	0.8	59.5	34.8	12.0	35.1	43.2	45.2	47.6	50.5	54.6
	Time and a Half	4,007	1.6	89.1	34.5	17.4	33.3	45.7	49.1	52.9	57.3	64.3
	Double Time	506	2.2	117.3	36.6	21.1	32.3	46.8	50.6	56.0	64.1	76.4
	Triple Time	85	3.1	175.4	43.1	29.1	35.2	53.7	63.3	66.6	84.5	93.0
Science 10	Regular Time	38,471	0.5	60.9	29.8	13.2	29.9	39.0	41.3	44.0	47.3	52.0
	Time and a Half	4,542	1.0	89.4	27.3	16.5	25.2	36.4	40.0	44.1	49.2	58.4
	Double Time	207	1.2	118.6	37.6	22.8	35.7	51.2	54.8	58.9	64.1	74.3
	Triple Time	88	1.9	143.5	32.8	26.2	26.2	43.8	52.7	59.1	67.3	78.0

7.3. Scale Score and Performance Level Distribution

Table 7.3 presents the overall scale score descriptive statistics across all students, including the mean; standard deviation (SD); scores at the 25th, median, and 75th percentiles; and skewness. Appendix F presents the results by demographic subgroup, and Appendix G presents the scale score distribution graphs for each subject area assessment for the overall testing population.

Table 7.3. Overall Scale Score Descriptive Statistics

Assessment	#Students	Mean	SD	P25	Median	P75	Skewness
Reading 9	45,079	198	28.18	180	199	217	-0.13
Reading 10	43,615	202	27.09	184	203	219	0.15
Math 9	43,894	194	32.20	177	198	215	-0.80
Math 10	42,439	189	34.23	174	193	212	-0.80
Science 9	45,006	206	32.80	185	208	227	-0.29
Science 10	43,308	195	33.79	176	198	217	-0.47

Table 7.4 presents the overall performance level distributions, or the percentage of students being classified into each performance level. Appendix H presents the performance level distributions by subgroup.

Table 7.4. Overall Performance Level Distributions

Assessment	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Reading 9	45,079	11.8	44.7	31.7	11.8
Reading 10	43,615	16.4	33.7	39.9	9.9
Math 9	43,894	20.4	40.3	31.6	7.7
Math 10	42,439	33.8	38.7	22.3	5.3
Science 9	45,006	26.2	27.5	31.3	15.1
Science 10	43,308	35.6	30.0	28.0	6.5

While results can be compared directly to previous years' performance within the same subject and grade, extra caution should be taken regarding the interpretations beyond high level due to impacts from the pandemic. These opportunity-to-learn (OTL) impacts are multi-faceted and differential across the state.

8. Item Analyses

To build the initial test forms for Utah Aspire Plus, item statistics based on use within the SAGE and ACT Aspire tests served to guide test construction activities, as described in Section 3.2. While the best possible initial forms were created, there were instances in which not all statistical targets were fully met. After the spring 2025 administration, the analyses were calculated again. This section presents the item analysis results for the operational items included on the spring 2025 test forms, including item difficulty, item-total correlation, and differential item functioning (DIF). The purpose of providing this information is to evaluate the quality and effectiveness of the test items, ensuring that they contribute to the assessment's overall measurement goals.

While some item statistics fell outside the guidelines for these analyses, their inclusion was necessary to meet blueprints given limitations to the available item banks. Overall, even where items fell outside the guidelines, they were still useful.

8.1. Classical Item Analysis

8.1.1. Item Difficulty (P-value and Item Mean Scores)

Item difficulty is measured by the *p*-value bounded by 0.0 and 1.0 that indicates how easy or hard an item is for students. The *p*-value for 1-point dichotomous items is the proportion of students who answered an item correctly. For multiple-point polytomous items, the *p*-value reflects the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) divided by the number of possible score points on the item. A high *p*-value indicates an easy item (high proportion of students answered it correctly), while a low *p*-value indicates a difficult item.

Table 8.1 and Table 8.2 present the item difficulty results across all 1-point and 2-point items from the spring 2025 administration. The p-value is presented for the 1-point items, while the item mean is presented for the 2-point items. Examination of the distribution of items by difficulty across each test shows that items do vary in difficulty across assessments, with most items between 0.30 and 0.75.

Table 8.1. Item Difficulty for 1-Point Items

Assessment	#Items	$p<0.30$	$0.30 \leq p< 0.55$	$0.55 \leq p< 0.75$	$0.75 \leq p< 0.95$	$p \geq 0.95$	Mean
Reading 9	27	2	8	11	6	0	0.59
Reading 10	27	0	7	14	6	0	0.65
Math 9	40	7	21	12	0	0	0.47
Math 10	40	6	22	12	0	0	0.45
Science 9	19	0	7	12	0	0	0.59
Science 10	17	0	14	3	0	0	0.46

Table 8.2. Item Difficulty for 2-Point Items

Assessment	#Items	Mean	Min.	Max.
Reading 9	8	0.72	0.46	1.02
Reading 10	8	0.91	0.48	1.66
Science 9	4	1.03	0.66	1.35
Science 10	6	0.69	0.44	0.94

Note. There were no 2-point mathematics items in spring 2025.

8.1.2. Item-Total Correlations

The item-total correlation is bounded by -1.0 and 1.0 and indicates how well an item distinguishes between low- and high-performing students. It is based on the relationship between student performance on a specific item and performance on the entire test based on their test score. An item with a high positive item-total correlation distinguishes between low- and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that lower-performing students did better on that item than higher-performing students.

Table 8.3 and Table 8.4 present the item-total correlation results across all 1-point and 2-point items from the spring 2025 administration.

Table 8.3. Item-Total Correlation for 1-Point Items

Assessment	#Items	$r<0.20$	$0.20 \leq r < 0.40$	$0.40 \leq r < 0.60$	$0.60 \leq r < 0.80$	$r \geq 0.80$	Median
Reading 9	27	1	11	15	0	0	0.42
Reading 10	27	0	4	20	3	0	0.51
Math 9	40	0	13	27	0	0	0.46
Math 10	40	0	12	28	0	0	0.45
Science 9	19	0	12	7	0	0	0.37
Science 10	17	0	9	8	0	0	0.39

Table 8.4. Item-Total Correlation for 2-Point Items

Assessment	#Items	Median	Min.	Max.
Reading 9	8	0.47	0.30	0.64
Reading 10	8	0.51	0.18	0.72
Science 9	4	0.58	0.47	0.69
Science 10	6	0.36	0.28	0.50

Note. There were no 2-point mathematics items in spring 2025.

8.2. Differential Item Functioning

Differential item functioning (DIF) exists when an item functions differentially across identifiable subgroups (e.g., sex or ethnicity) where students are matched on ability (meaning comparisons are made between students of the same ability, so differences are not attributable to overall group performance differences). In this context, DIF may indicate an issue with fairness or that the item may be measuring something other than the intended construct (i.e., possible evidence of DIF). However, it is important to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify potential biases. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

There are multiple statistical procedures for analyzing DIF, one of which is based on the Mantel-Haenszel chi-square statistic (M-H χ^2) for multiple-choice items (Holland & Thayer, 1988). The chi-square statistic determines whether the odds of a correct response on an item is the same for both focal and reference groups across all levels of proficiency. The Mantel-Haenszel odds ratio (α_{M-H}) is the odds of a correct response of the reference group divided by the odds of a correct response of the focal group. Data for these Mantel-Haenszel procedures are drawn from 2-by-2-by-*k* (score levels) contingency tables, for each item. As shown in Table 8.5, the number of focal and reference group members scoring in each possible item response is captured.

Table 8.5. Item 2×2 Contingency Table for the k^{th} Score Level

Group	Correct Score (1)	Incorrect Score (0)	Total
Focal (f)	n_{f1k}	n_{f0k}	n_{fk}
Reference (r)	n_{r1k}	n_{r0k}	n_{rk}
Total (t)	n_{t1k}	n_{t0k}	n_{tk}

For classifications of DIF, the Mantel-Haenszel Delta DIF statistic (Dorans & Holland, 1993) is computed from the Mantel-Haenszel odds ratio and used in conjunction with M-H χ^2 to classify items into three categories distinguishing magnitudes of DIF: negligible DIF (A), moderate DIF (B), and large DIF (C). Classification is based on the following guidelines:

- M-H χ^2 not significantly different from 0 or |MHD| less than 1 results in a classification of *A*.
- M-H χ^2 significantly different from 0 and |MHD| at least 1 but less than 1.5 ***or*** M-H χ^2 not significantly different from 0 and |MHD| greater than 1 results in a classification of *B*.
- M-H χ^2 significantly different from 0 and |MHD| at least 1.5 results in a classification of *C*.

In addition to these classifications, notation of DIF includes a positive (+) sign, indicating that the item favors the focal group, or a negative (–) sign, indicating that the item favors the reference group. Items that are designated with "B" or "C" DIF classifications are recommended for review before continued use on assessments.

The standardized mean difference (SMD; Zwick et al., 1993) procedure is also used for detecting DIF for items worth more than 1 point. SMD is a summary statistic used as an effect size estimate comparing the mean item score between the reference and focal groups (the two groups being compared). Although the numerical result of this statistical procedure is different from the M-H statistics, the classification of the results is the same—the results are classified into three categories indicating the magnitude of DIF with additional notation indicating the favored group.

Table 8.6 presents the DIF results for the items on the spring 2025 test forms.

Table 8.6. DIF Results: Number of Items by DIF Category

Assessment	DIF Comparison	Negligible DIF	Moderate DIF: Focal	Moderate DIF: Reference	Substantial DIF: Focal	Substantial DIF: Reference
Reading 9	Male-Female	35	0	0	0	0
	White-Black	34	0	1	0	0
	White-Hispanic	35	0	0	0	0
Reading 10	Male-Female	35	0	0	0	0
	White-Black	35	0	0	0	0
	White-Hispanic	34	0	0	0	1
Math 9	Male-Female	38	0	1	0	1
	White-Black	36	0	3	0	1
	White-Hispanic	40	0	0	0	0
Math 10	Male-Female	39	0	1	0	0
	White-Black	38	0	2	0	0
	White-Hispanic	40	0	0	0	0

Assessment	DIF Comparison	Negligible DIF	Moderate DIF: Focal	Moderate DIF: Reference	Substantial DIF: Focal	Substantial DIF: Reference
Science 9	Male-Female	23	0	0	0	0
	White-Black	23	0	0	0	0
	White-Hispanic	23	0	0	0	0
Science 10	Male-Female	23	0	0	0	0
	White-Black	22	0	1	0	0
	White-Hispanic	23	0	0	0	0

Note. "Focal" indicates DIF in favor of Female, Black, or Hispanic students; "Reference" indicates DIF in favor of Male or White students.

9. Calibration, Equating, and Scaling

Item response theory (IRT) analyses were used to create the base scales for the Utah Aspire Plus assessments. All Utah Aspire Plus assessments in spring 2025 were pre-equated, with item parameters estimated either from prior operational post-equating or field test calibration. (Refer to prior technical reports for details on these processes, available online at https://utah.mypearsonsupport.com/admin-resources.html under "Reporting Resources.") Student scores were estimated using IRT and then transformed to the final Utah Aspire Plus scale score reporting metric. Scores were reported on-demand. Following administration, a separate calibration and equating process was conducted. While these results did not affect student scores, they allowed for the calibration of field test items, identification of items with parameter drift, and an update of the bank parameters.

9.1. IRT Models

Multiple item types are used on Utah Aspire Plus assessments and require multiple measurement models. Traditional multiple-choice items, with one correct answer, are analyzed via the three-parameter logistic (3PL) model (Birnbaum, 1968), denoted as follows:

$$p_i(\theta_j) = c_i + \frac{1-c_i}{1+e^{-Da_i(\theta_j - b_i)}},$$

where $p_i(\theta_j)$ is the probability that student *j* would earn a score of 1 on item *i*, b_i is the difficulty parameter for item *i*, a_i is the slope (or discrimination) parameter for item *i*, c_i is the pseudo-chance (or guessing) parameter for item *i*, and *D* is the constant 1.7. Other selected-response items worth 1 point (e.g., technology-enhanced items) are analyzed via the two-parameter logistic (2PL) model (Birnbaum, 1968), which is a reduced model from 3PL, where the pseudo-chance parameter, *c*, is assumed zero. Items worth 2 points were analyzed via the generalized partial credit model (GPCM; Muraki, 1992), denoted as follows:

$$p_{im}(\theta_j) = \frac{\exp\left[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})\right]}{\sum_{v=0}^{M_i-1} \exp\left[Da_i(\theta_j - b_i + d_{iv})\right]},$$

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$, $p_{im}(\theta_j)$ is the probability of a student with θ_j getting score *m* on item *i*, and M_i is the number of score categories of item *i* with possible item scores as consecutive integers from 0 to $M_i - 1$. In the GPCM, the *d* parameters define the "category intersections" (i.e., the θ value at which examinees have the same probability of scoring 0 and 1, 1 and 2).

9.2. Calibration

The data preparation for the IRT calibration process began with all Utah students who were administered the base forms (i.e., the online, English-language forms). The samples for item parameter estimation included students from the online, English language test forms; with the same grade battery of tests; and with a valid test score status for a subject test. Students without a valid test score were excluded from calibration data. The primary goal of the IRT calibration was to place the operational and field test items from a given test onto a common scale. The additional step of equating was also completed to place these parameters onto the original Utah Aspire Plus base scales.

Large enough samples are necessary to sufficiently estimate IRT parameters for a given test and across the respective models (generally for state summative tests similar to Utah Aspire Plus, sample sizes of 2,000). IRTPRO (Scientific Software International, Inc., 2017) was used to obtain the IRT parameter estimates using the IRT measurement models. The software default estimation method, Bock-Aitkin (BAEM), was used for each calibration. The prior distributions for latent traits were set to a mean of zero and a standard deviation of one. The number of quadrature points used in the estimation was set to 49. For item parameters, a prior was placed on the lower asymptote (pseudo-chance) for 3PL: a normal distribution with a mean of -1.4 and a standard deviation of one. After calibration, convergence was checked.

To convert IRTPRO item parameters to the commonly used logistic parameter presentation, the *a*-parameter from the IRTPRO output needed to be converted since IRTPRO uses 1.0 for a scaling constant using the following formula:

$$a_{new} = \frac{a_{irtpro}}{1.7}.$$

9.3. Equating

A common item non-equivalent groups approach (Kolen & Brennan, 2014) was used for equating the 2025 forms to the base scales. The Stocking and Lord (1983) test characteristic curve (TCC) methodology was used to derive equating constants for each grade/subject area test. The operational items were used as the common-item linking set. The banked IRT item parameter estimates for all the Utah Aspire Plus operational items, and the respective item parameter estimates from the 2025 administration described in Section Calibration, were used to obtain transformation constants. This was conducted using the computer program STUIRT (Kim & Kolen, 2004).

Equating was carried out in conjunction with the drift analysis described in Section 9.3.1 that resulted in a final set of Stocking and Lord scaling constants. These constants were then applied to all 2025 calibrated items to obtain a set of parameters for the operational and field test items. Table 9.1 presents the final Stocking and Lord scaling constants used for placing tests onto the Utah Aspire Plus base scales.

Table 9.1. 2025 Final Stocking and Lord Scaling Constants

Assessment	Slope	Intercept
Reading 9	1.033	-0.047
Reading 10	0.987	0.058
Math 9	1.052	-0.165
Math 10	1.056	-0.311
Science 9	1.101	0.252
Science 10	1.037	-0.112

Final parameters were then updated in the item bank for items in the following categories:

- Item was field tested in 2025.
- Item was used operationally for the first time in 2025 (prior parameters were from field test administration).
- Item showed drift during the equating process.

9.3.1. Drift Analysis

A critical step in the equating process is to evaluate the anchor items for stability in relation to its banked item characteristics. Items that deviate substantively in relation to the entire set of anchor items may be removed from contributing to the final equating solution. For Utah Aspire Plus, the item parameter stability check for the operational items was conducted using classical item analyses, scatter plots of item parameter estimates, and item characteristic curve (ICC) comparison. For the ICC comparison, old and new ICCs were compared using the z-score approach based on D^2 (Wells et al., 2014) as outlined below:

1. Obtain the theoretically weighted estimated posterior theta distribution using 31 quadrature points (-5 to 5).
2. Compute the slope and intercept constants using Stocking and Lord in STUIRT with all operational items in the linking set.
3. Place the freely calibrated item parameter estimates onto the baseline scale by applying the constants obtained in Step 2.
4. For each operational item, calculate D^2 between the ICCs based on old (x) and new (y) parameters at each point in this theta distribution:

$$D_i^2 = \sum^{k} \left[P_{ix}(\theta_k) - P_{iy}(\theta_k)\right]^2 \bullet g(\theta_k)$$

   where i = item, x = old form, y = new form, k = theta quadrature point, and g = theoretically weighted posterior theta distribution.

5. Flag items with a D^2 that is greater than the mean D^2 value, and whose distance from the mean D^2 value is greater than twice the standard deviation of the D^2 values.
6. Examine the impact of removing a flagged item on the content representativeness of the resulting anchor set. A flag alone is not the sole criteria for removing an item from the anchor set. It is important to also make sure that the remaining anchor set continues to be representative of the overall content and structure of the test.

Table 9.2 presents the number of operational items showing drift from the spring 2025 administration. Appendix K presents the plots showing D^2 values following the initial equating.

Table 9.2. 2025 Items Showing Drift

Assessment	#Items Showing Drift
Reading 9	2
Reading 10	1
Math 9	2
Math 10	2
Science 9	1
Science 10	1

Following removal of items for drift, the STUIRT equating process was repeated with the updated anchor set to obtain a final set of Stocking and Lord scaling constants, which were applied to the freely calibrated item parameters to obtain a final set of parameters. Parameters in the item bank were updated to these parameters for items showing drift, as well as for field test items and items that were operational in 2025 for the first time.

Appendix E presents scatter plots of the operational items. Overall, item functioning of common items was typical and stable. No more than two items in any of the common item sets were removed from final linking solutions. Scatter plots and correlations of IRT difficulty and discrimination parameters showed strong correlations.

9.3.2. Model Fit Evaluation Criteria

The Q_1 statistic (Yen, 1981) was used as an index of correspondence between observed and expected performance. To compute Q_1, first the estimated item parameters and student response data (along with observed item scores) were used to estimate student ability ($\hat{\theta}$). Next, expected performance was computed for each item using students' ability estimates in combination with estimated item parameters. Differences between expected item performance and observed item performance were then compared at 10 intervals across the range of student achievement (with approximately the same number of students per interval). Q_1 was computed as a ratio involving expected and observed item performance. Q_1 is interpretable as a chi-squared (c^2) statistic, which can be compared to a critical chi-squared value to make a statistical inference about whether the data (observed item performance) were consistent with what might be observed if the IRT model was true (expected item performance). Q_1 is not directly comparable across different item types because items with different numbers of IRT parameters have different degrees of freedom (*df*). For that reason, a linear transformation (to a *Z*-score, Z_{Q_1}) was applied to Q_1. This transformation also made item fit results easier to interpret and addressed the sensitivity of Q_1 to sample size.

To evaluate item fit, Yen's Q_1 statistic was calculated for all items. Q_1 is a fit statistic that compares observed and expected item performance. For dichotomous items, Q_1 was computed as follows:

$$Q_{1i} = \sum_{j=1}^{J} \frac{N_{ij}\left(O_{ij} - E_{ij}\right)^2}{E_{ij}\left(1 - E_{ij}\right)}$$

where N_{ij} was the number of examinees in interval (or group) *j* for item *i*, O_{ij} was the observed proportion of the students for the same cell, and E_{ij} was the expected proportions of the students for the same interval. The expected proportion was computed as follows:

$$E_{ij} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_i\left(\hat{\theta}_a\right),$$

where $P_i(\hat{\theta}_a)$ was the item characteristic function for item *i* and students *a*. The summation is taken over students in interval *j*. The generalization of Q_1 for items with multiple response categories is as follows:

$$Gen\ Q_{1i} = \sum_{j=1}^{10} \sum_{k=1}^{m_i} \frac{N_{ij}\left(O_{ikj} - E_{ikj}\right)^2}{E_{ikj}},$$

where

$$E_{ikj} = \frac{1}{N_{ij}} \sum_{a \in j}^{N_{ij}} P_{ik}\left(\hat{\theta}_a\right).$$

Both Q_1 and generalized Q_1 results were transformed to ZQ_1 and were compared to a criterion $ZQ_{1,crit}$ to determine acceptable fit. The conversion formula was the following:

$$ZQ_1 = \frac{Q_1 - df}{\sqrt{2df}},$$

and

$$ZQ_{1,crit} = \frac{N}{1500} * 4,$$

where *df* is the number of degrees of freedom. The number of degrees of freedom is equal to the number of independent cells less the number of independent item parameters. For example, the degrees of freedom for polytomous items equals [10 × (number of score categories – 1) – number of independent item parameters]. For the GPCM, the number of independent item parameters equals 1 (for the *a*-parameter) plus the number of step values (e.g., an item scored 0, 1, 2 has two independent step values; the *b* parameter is simply the mean of the step values and is therefore not independent).

As all items were pre-equated, Q_1 statistics were calculated in previous administrations, along with item fit plots. All items included on previous forms showed adequate fit. Additionally, Q_1 and item fit plots were re-generated following the 2025 administration to assess pre-equating. Results were consistent with the drift analyses and did not suggest any concerns with model selection.

9.4. Establishing the Reporting Scale

Commonly derived scores based on IRT (as described in Section 5.1) are transformed to a reporting scale that is more consumable by users. The IRT metric being logit-based results in ability estimates typically ranging from –3.0 to 3.0 and to the second or third decimal. Interpreting differences across logits can be cumbersome, so scores are transformed to larger values without fractions. These are generally called scale scores. The purpose of scale scores is to facilitate interpretation and to report scores for all students on a scale that remains consistent across multiple years or forms, even if the overall difficulty of the test varies slightly. Scale scores ensure that the test results mean the same thing regardless of which year the test was administered. For the Utah Aspire Plus scales, the IRT metric uses a linear transformation to provide the final reporting scales as such:

$$SS = m\theta + b,$$

where *m* is the slope, and θ is the IRT person proficiency estimate obtained through pattern scoring. Using this equation, a scale scored is transformed to the final reporting scale. The scale score metric for Utah Aspire Plus was chosen to range from 100 to 300 for each test and composite score. This range allows for the assessment to differ from the previous and remaining scales, and the slope chosen to spread final scores enough to contain each respective score distribution without floor or ceiling effects and to be disperse enough to reasonably contain all transformed scores. The final transformation formula used for Utah Aspire Plus is as follows:

$$SS = 25 \times \theta + 200$$

This transformation provides the following characteristics: (a) the mean of the scale is 200, (b) the standard deviation of the scale is 25, (c) the lowest operating scale score (LOSS) is 100, and (d) the highest operating scale score (HOSS) is 300. Composite scores were also created for Utah Aspire Plus. A composite score representing Science, Technology, Engineering, and Mathematics (STEM) is the average of a student's mathematics and science scale scores.

10. Quality Control Procedures

Quality control is a critically important element of every phase of the Utah Aspire Plus development, administration, and score reporting in ensuring the accuracy of student-, school- and district-level data. Pearson has developed and refined a set of quality procedures to help ensure that all USBE's testing requirements are met or exceeded. These quality control procedures are detailed in the paragraphs that follow. In general, Pearson's commitment to quality is incorporated in both task-specific quality standards applied to processing functions and services as well as a network of systems and procedures that coordinate quality steps across functions and services.

10.1. Quality Control of Test Development

Test items for Utah Aspire Plus are housed in Pearson's Automated Banking and Building for Interoperability (ABBI) platform. ABBI supports building and publishing online and paper-based tests and drives creation of those forms to both Pearson's paper and online publishing systems. Through ABBI, item scoring configuration is validated during initial item review (i.e., at the time of item writing) and during forms development.

10.2. Quality Control of Online Assessment Delivery

Pearson's Assessment Delivery and Management (ADAM) was used for the Utah Aspire Plus assessment for the first time in spring 2024. This system provides seamless student rostering, streamlined test management, precise scoring, and insightful reporting. ADAM also provides comprehensive support for paper and online testing either through a single sign-on destination or by interfacing with other systems to provide a highly adaptable and configurable solution.

TestNav delivers online tests to the students. The core functionalities of TestNav include delivering tests to students, collecting student responses, and returning the responses to Pearson for scoring. TestNav provides advance warning of network issues that prevent sending student responses to the Pearson testing server. When the network is functioning normally, TestNav sends student responses to the Pearson testing server in real time, while the student is testing. If the student's device cannot connect to the Pearson servers, TestNav saves the response to an encrypted file and allows the student to continue testing. When the network connection is reestablished, the test proctor can upload a student's saved responses to Pearson's testing server, and then TestNav erases the encrypted response file from the student's device or local network. As part of test security, test administrators control individual student authorization by printing and distributing testing tickets with each student's identifying information and unique log-in credentials.

In the event of a non-network or non-internet issue, such as a power outage or student device shutdown, student responses are saved to the encrypted file. When the student resumes testing, the system uploads the data in the file to the servers, and the student continues at the point in the test when the issue occurred.

As part of test security, test administrators control individual student authorization by printing and distributing testing tickets with each student's identifying information and unique log-in credentials. The student enters their log-in and password on the testing workstation to gain access to the test. To further secure the testing environment, a non allowed list capability sends notifications when unapproved applications are running when the test is started. Once all non-allowed applications are shut down, TestNav starts in kiosk mode when a student signs into a secure test. Kiosk mode locks down the testing computer or device, so the student cannot print, cut, or copy test content. Students cannot visit websites or access other installed applications not approved for use during the test.

During the operational administration, Pearson's operational monitoring practices and tools constantly verify that platforms remain available to users; that performance stays within acceptable limits; and that users do not encounter critical errors. Additionally, monitoring includes real-time security auditing and systems vulnerability monitoring throughout a given testing window. Refer to Section 4.1 for more information.

10.3. Quality Control of Production System Testing

Table 10.1 describes the various steps involved in the production system testing process, which refers to the end-to-end testing of the administration platforms to ensure a seamless and secure test administration.

Table 10.1. Quality Control of Production System Testing

Testing	Description
Functional Testing	Well before testing the entire system, Pearson engineers develop tests for each discrete software unit, and for small groups of related units. Debugging code is emphasized in the earliest stages of development, so during unit testing, each developer creates unique tests for code that has been written.
Integration Testing	Digital and traditional paper solutions require testing that is specific to its unique interactions and specifications. After testing each piece of component code, the behavior of the integrated parts is tested. In the first stage of integration testing, the testing is done at the base system level to verify and validate that the system components function together. The second stage of integration testing examines accuracy of the unique configuration to each administration specified in the contract. Configuration requirements are the basis of Pearson's integration testing. This is documented, and test cases and results are maintained and verified prior to the final production scoring and reporting configuration, including item parameter files, keys, and cut scores.
Program Validation End-to-End Testing	After product testing approval, the Pearson program validation team uses a cross-system end-to-end approach to validate the user interface, scoring, data files, and reports. This testing confirms that all data are consistent with customer requirements by emulating the customer experience throughout the program lifecycle. The program validation team coordinates test-material processing (distribution and data collection) with the same operational areas that process live material during production. Where appropriate, there is a production sample verification process, which uses the first available student data as a final quality step before live production processing of materials to be distributed. An examination of the outputs verifies data are scored, aggregated, reported, and delivered accurately. After the program validation team approves, the delivery of code and configuration is moved to production.

Testing	Description
Load Testing	To examine the system's expected performance during peak usage days, Pearson engineers will assemble the components and test the system under load conditions. During load testing, a period of peak production is modeled to identify any issues within the application that might be triggered by maximum activity. Load testing is performed several times per year so that the system can be scaled to meet anticipated customer demand in advance of when it is needed.
Performance Monitoring	Systems are constantly monitored for anomalous system behavior, with special care being taken during student testing cycles to provide the highest possible levels of availability and performance. Monitors watch for anomalous activity throughout the entire system, not just at the application or network layers. If suspicious activity shows up, the system triggers alerts to technical support staff for investigation and handling. In addition to overall, system-wide monitoring for suspicious and anomalous system activity, systems are kept at current patch levels via a suite of tools to scan for vulnerabilities at the network, operating system, platform, and application layers.
Regression Testing	Core regression testing confirms that pre-existing functionality has not been adversely affected by changes introduced in a software update. The scope of regression testing is set up to match the changes that are being introduced into the systems by the implementation and testing teams. Regression testing is conducted for every release or patch that is created for our systems.
User Acceptance Testing	User acceptance testing is performed by states. Pearson maintains a testing platform so that states can review system functionality prior to a production release. The following steps are taken when designing the user acceptance testing plan: (a) create release notes for all new or modified functionality, (b) provide updated training and user documentation, (c) review checklist and ask questions, (d) provide user IDs and passwords to allow users to run tests on code along with associated documentation assisting users on the process and procedures, and (e) meet with users and share results to jointly establish appropriate action plans.

10.4. Quality Control of Scoring and Reporting

Score tables used to estimate student scores on-demand were replicated independently through two parties internally. Additionally, a mock run of data was scored both using the on-demand process, and by internal replicators. This scoring dry run was conducted at the overall test level and by reporting categories. Any differences were resolved and rerun until both parties' results were identical and deemed correct based on careful examination of output.

From initial student data upload, through testing, data review, scoring, and reporting, Pearson completes multiple checks and confirms that all data are consistent with customer requirements. Quality assurance (QA) tasks are part of the project schedule, which is built by working backwards from the reporting dates, to allow for QA work to flow effectively. Solid requirements form the foundation of quality. USBE and Pearson collaborated to thoroughly and consistently document scoring and reporting requirements, so all involved have a clear understanding of desired results. Project management, product validation, reporting services, and Customer Data Quality (CDQ) teams also participated in requirements reviews to meet reporting requirements and provide accurate mockups.

All Utah Aspire Plus files go through a rigorous validation process as demonstrated by Pearson's comprehensive quality plan. The plan focuses on implementing test cases at the source of each activity, system, and process, thereby detecting defects at the earliest possible point. The impact, therefore, is minimized and resolution can be expedited. The mock data process has become a validation standard within Pearson. It demonstrates production readiness in advance of scoring and reporting actual student data. CDQ uses industry-standard validation tools focusing on SAS, which allows Pearson the breadth and depth needed for large-scale, high-stakes assessment validation. Pearson's test plans and individual test cases target areas of historical risk (based on the knowledge of Utah Aspire Plus requirements and file layouts) to provide quality results.

10.5. Quality Control of Psychometric Processes

For all psychometric tasks, quality management is central to ensuring on-time and error-free results. Details of Pearson's quality and control procedures for all psychometric tasks conducted including test construction, calibration, equating, scaling, field test analysis, data review, item bank creation and management, standard setting, and technical reporting can be found in Appendix O of the *2018–2019 technical report* (Pearson, 2020). Other quality control measures taken by the psychometrics team include the following:

- IRT Data Matrix Files: Student records in the calibration data files were ordered by ascending student identification number. In the case where field test forms are used, student records would first be sorted by form, then by student identification number. The array of item responses was presented in the order as administered in the test form, including items that are presented in field test slots. The IRT data matrices were created independently by two Pearson psychometric staff. The matrices were checked for accuracy by comparing numbers of students (counts) and the item response arrays. Any discrepancy found was resolved. Final calibration data files matched perfectly.
- Calibration: IRT calibrations were conducted independently by two Pearson psychometric staff using the same software program. All item parameters from both independent calibrations were compared. Item fit plots were generated as further analyses of reasonableness and support of decisions of items' future use.

11. Reliability

Estimation of reliability of a given assessment is critical to understand the precision of measurement for individual test scores. Test score reliability estimates are typically provided in both a classical and IRT context. Classical reliability estimates such as Cronbach's alpha or standard error of measurement (SEM) are reliability measures of internal consistency. Where classical approaches are generally single indicators for a given assessment, IRT reliability reflects precision across the ability spectrum. As such, reliability for the Utah Aspire Plus assessments was evaluated based on the following:

- Internal consistency (Cronbach's alpha)
- Standard error of measurement (SEM)
- Conditional standard error of measurement (CSEM)
- Classification accuracy and consistency

11.1. Classical Reliability

The basis of classical test theory is premised on the idea that a person's observed score is the sum of their true score (measured without error and not directly observable) plus error: *observed score = true score + error*. It provides a means of describing the quality of test scores through the interplay of these three elements. Arguably the most important descriptor is the concept of the reliability of test scores, where the reliability of observed scores is defined as follows:

$$Reliability = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2},$$

where σ_T^2 is the true score variance, σ_O^2 is the observed score variance, and σ_E^2 is the error variance. When there is no error, the reliability is the true score variance divided by true score variance, which is unity. However, as more error influences the measure, the error component in the denominator of the ratio increases and the reliability decreases.

11.1.1. Cronbach's Alpha

Internal consistency methods use a single administration to estimate test score reliability. For state assessments where student testing time is at a premium, internal consistency procedures have a practical advantage over reliability estimation procedures requiring multiple tests. A frequently used internal consistency reliability estimate is the coefficient alpha (Cronbach, 1951). Coefficient alpha assumes that inter-item covariance constitutes true-score variance and the fact that the average true score variance of items is greater than or equal to the average inter-item covariance. The formula for the coefficient alpha is as follows:

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} s_{Y_i}^2}{s_X^2}\right),$$

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the i^{th} item (or component), and s_X^2 is the observed score sample variance for the test.

Appendix C presents the coefficient alpha reliability estimates for the overall testing population and by demographic subgroup. Results are also provided by reporting category (although only overall scores are reported on ISRs, and no subscores are reported).

11.1.2. Standard Error of Measurement

A reliability coefficient expresses test score consistency in terms of variance ratios. In contrast, the SEM expresses score inconsistency (unreliability). The SEM is an estimate of how much error there is likely to be in a student's observed score, or alternately, how much score variation would be expected if the student were tested multiple times with equivalent forms of the test. The SEM is calculated as follows:

$$SEM = s_x\sqrt{1-\rho_{XX'}}$$

where s_x is the standard deviation of the total test (standard deviation of the raw scores), and $\rho_{xx'}$ is a reliability estimate for the set of test scores. Appendix C presents the SEMs on the Utah Aspire Plus scale score metric (s_x = 25).

11.2. IRT Reliability

Where estimation of reliability is within a classical test theory framework, such measures are sample specific. Error estimates such as the SEM are also group-level estimates that apply across test scores, and it is sometimes viewed as unrealistic that the size of errors would be unrelated to the "true scores" of students (identical for all). For Utah Aspire Plus, student scores are derived within an IRT framework through pattern scoring based on the 3PL and 2PL measurement models. Under the IRT model, measurement precision is expressed as conditional standard errors of measurement (CSEMs) and is equal to the inverse of the square root of the test information function across the ability continuum (Hambleton & Swaminathan, 1985).

CSEMs depend on both the unique set of items each student answers correctly and their estimated ability level (θ). Therefore, different students will likely have different CSEM values even if they have the same raw score and/or theta estimate. Each item contains a unique amount of information for a given ability level, which depends on each item's discrimination, difficulty, and pseudo-guessing parameters.

Appendix D presents the conditional standard errors for Utah Aspire Plus assessments, with each plot including a line indicating the scale score cut score for *Proficient*. Ideally, the lowest CSEM value occurs at the location of *Proficient*.

11.3. Classification Accuracy and Consistency

Every test administration will result in some error in classifying students. The concept of the SEM provides a mechanism for explaining how measurement error can lead to classification errors when cut scores are used to classify students into different performance levels. For example, some students may have a true performance level greater than a cut score. However, due to random variations (measurement error), their observed test score may be below the cut score. As a result, the students may be classified as having a lower performance level. A student's true score is most likely to fall into a standard error band around their observed score. Thus, the classification of students into different performance levels can be imperfect, especially for the borderline students whose true scores lie close to performance level cut scores.

For the Utah Aspire Plus assessment, the levels of achievement are *Below Proficient*, *Approaching Proficient, Proficient,* and *Highly Proficient.* Accuracy refers to the extent to which achievement decisions based on test scores match those that would be made if the scores did not contain any measurement error (i.e., "true scores"). Because true scores are not available, an estimate of the true score distribution must be determined for classification accuracy to be estimated. Consistency, on the other hand, refers to the extent to which achievement classification decisions based on test scores match the decisions based on a second, parallel form of the same test. This index assumes that two parallel forms of the same test are administered to the same group of students. In Utah, however, this is impractical. Livingston and Lewis (1995) developed techniques to estimate both accuracy and consistency that overcome the constraints of true scores and multiple test forms on the same students. These procedures are used to generate accuracy and consistency indices on the Utah Aspire Plus assessments. All indices were calculated using the BB-CLASS software (Brennan, 2005).

11.3.1. Calculating Accuracy and Consistency

To calculate accuracy, a 4×4 contingency table is created for each subject area and grade. The $[x, y]$ entry of an accuracy table represents the estimated proportion of students whose true score falls into performance level x and whose observed scores fall into performance level y. Table 11.1 is an example of an accuracy table where the columns represent test-based student achievement, and the rows represent true achievement-level decisions. In this example, the total accuracy is approximately 75%, the sum of the diagonal (shaded) cells.

Table 11.1. Example Accuracy Classification Table: True vs. Observed Scores

True Score	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*	Total
Below Proficient	0.117	0.034	0.000	0.001	0.152
Approaching Proficient	0.019	0.161	0.061	0.002	0.243
Proficient	0.000	0.034	0.294	0.061	0.389
Highly Proficient	0.000	0.000	0.036	0.179	0.215
Total	0.136	0.229	0.391	0.243	1.000

It is useful to consider decision accuracy based on a dichotomous classification of *Below Proficient* or *Approaching Proficient* versus *Proficient* or *Highly Proficient* because Utah uses *Proficient* and above as proficiency for accountability decision purposes as well as for an index tracking students' readiness to college and careers. To compute decision accuracy in this case, the table is dichotomized by combining cells associated with *Below Proficient* and *Approaching Proficient* and combining *Proficient* with *Highly Proficient*. The sum of the shaded cells in Table 11.2 indicates classification accuracy around the Proficient cut point of approximately 90%. The percentage of examinees incorrectly classified as *Approaching Proficient* or lower, when their true score indicates *Proficient* or above, is approximately 3%.

Table 11.2. Example Accuracy Classification Table for Proficient Cut Point: True vs. Observed Scores

True Score	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*	Total
Below Proficient	0.117	0.034	0.000	0.001	0.152
Approaching Proficient	0.019	0.161	0.061	0.002	0.243
Proficient	0.000	0.034	0.294	0.061	0.389
Highly Proficient	0.000	0.000	0.036	0.179	0.215
Total	0.136	0.229	0.391	0.243	1.000

Consistency can be calculated in the same manner, via 4 x 4 contingency table, albeit with data indicating an estimate of the joint distribution of classifications on (hypothetically) two independent, parallel test forms. Table 11.3 shows sample statistics of consistency classification. Based on this sample data, the overall consistency is approximately 67%. The consistency at *Proficient* is 87%. The agreement rates are lower than those for accuracy because both classifications contain measurement error; whereas in the accuracy table, true score classification is assumed to be without error.

Table 11.3. Example Consistency Classification Table: First vs. Second Form

First Form	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*	Total
Below Proficient	0.111	0.043	0.009	0.001	0.164
Approaching Proficient	0.019	0.147	0.073	0.004	0.243
Proficient	0.006	0.038	0.252	0.075	0.371
Highly Proficient	0.000	0.002	0.056	0.163	0.221
Total	0.136	0.230	0.390	0.243	1.000

11.3.2. Calculating Kappa

Another way to express overall consistency is to use Cohen's kappa (κ) coefficient (Cohen, 1960), which assesses the proportion of consistent classifications beyond chance. The coefficient is computed using the following:

$$\kappa = \frac{P - P_c}{1 - P_c},$$

where P is the proportion of consistent classifications and P_c is the proportion of consistent classification by chance. Using Table 11.3, P is the sum of the shaded cells whereas P_c is as follows:

$$\sum_x C_{x.} C_{.x},$$

where $C_{x.}$ is the proportion of students whose observed performance level would be x on the first form, and $C_{.x}$ is the proportion of students whose observed performance level would be x on the second form. Therefore, the kappa coefficient using the data from Table 11.3 is 0.548. Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

11.3.3. Results

Table 11.4 – Table 11.6 present the estimates of classification accuracy and consistency indices, including kappa coefficients, for overall performance level classification and at the *Proficient* cut point.

Table 11.4. Classification Accuracy: True vs. Observed Scores

Assessment	True Score	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*	Accuracy %
Reading 9	*Below Proficient*	0.085	0.022	0.000	0.000	77.44
	Approaching Proficient	0.034	0.373	0.061	0.000	
	Proficient	0.000	0.052	0.233	0.033	
	Highly Proficient	0.000	0.000	0.023	0.084	
Reading 10	*Below Proficient*	0.127	0.023	0.000	0.000	78.76
	Approaching Proficient	0.038	0.267	0.057	0.000	
	Proficient	0.000	0.047	0.321	0.026	
	Highly Proficient	0.000	0.000	0.021	0.073	
Math 9	*Below Proficient*	0.170	0.037	0.000	0.000	77.91
	Approaching Proficient	0.034	0.323	0.063	0.000	
	Proficient	0.000	0.044	0.228	0.019	
	Highly Proficient	0.000	0.000	0.024	0.058	
Math 10	*Below Proficient*	0.299	0.057	0.000	0.000	77.26
	Approaching Proficient	0.039	0.283	0.053	0.000	
	Proficient	0.000	0.046	0.152	0.014	
	Highly Proficient	0.000	0.000	0.018	0.039	
Science 9	*Below Proficient*	0.210	0.045	0.002	0.000	69.60
	Approaching Proficient	0.051	0.172	0.070	0.001	
	Proficient	0.002	0.057	0.205	0.040	
	Highly Proficient	0.000	0.001	0.035	0.109	
Science 10	*Below Proficient*	0.295	0.071	0.006	0.000	67.88
	Approaching Proficient	0.057	0.168	0.076	0.001	
	Proficient	0.004	0.060	0.178	0.026	
	Highly Proficient	0.000	0.000	0.021	0.038	

Table 11.5. Classification Accuracy at *Proficient* Cut Point: True vs. Observed Scores

Assessment	True Score	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*	Accuracy %
Reading 9	*Below Proficient*	0.085	0.022	0.000	0.000	88.63
	Approaching Proficient	0.034	0.373	0.061	0.000	
	Proficient	0.000	0.052	0.233	0.033	
	Highly Proficient	0.000	0.000	0.023	0.084	
Reading 10	*Below Proficient*	0.127	0.023	0.000	0.000	89.60
	Approaching Proficient	0.038	0.267	0.057	0.000	
	Proficient	0.000	0.047	0.321	0.026	
	Highly Proficient	0.000	0.000	0.021	0.073	
Math 9	*Below Proficient*	0.170	0.037	0.000	0.000	89.30
	Approaching Proficient	0.034	0.323	0.063	0.000	
	Proficient	0.000	0.044	0.228	0.019	
	Highly Proficient	0.000	0.000	0.024	0.058	
Math 10	*Below Proficient*	0.299	0.057	0.000	0.000	90.09
	Approaching Proficient	0.039	0.283	0.053	0.000	
	Proficient	0.000	0.046	0.152	0.014	
	Highly Proficient	0.000	0.000	0.018	0.039	
Science 9	*Below Proficient*	0.210	0.045	0.002	0.000	86.74
	Approaching Proficient	0.051	0.172	0.070	0.001	
	Proficient	0.002	0.057	0.205	0.040	
	Highly Proficient	0.000	0.001	0.035	0.109	
Science 10	*Below Proficient*	0.295	0.071	0.006	0.000	85.28
	Approaching Proficient	0.057	0.168	0.076	0.001	
	Proficient	0.004	0.060	0.178	0.026	
	Highly Proficient	0.000	0.000	0.021	0.038	

Table 11.6. Classification Consistency: First vs. Alternate Form

Assessment	First Form	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*	Consistency %	Kappa
Reading 9	*Below Proficient*	0.081	0.042	0.000	0.000	68.27	0.529
	Approaching Proficient	0.037	0.328	0.080	0.002		
	Proficient	0.000	0.075	0.193	0.036		
	Highly Proficient	0.000	0.002	0.043	0.080		
Reading 10	*Below Proficient*	0.120	0.041	0.001	0.000	70.03	0.568
	Approaching Proficient	0.043	0.230	0.078	0.000		
	Proficient	0.001	0.066	0.281	0.030		
	Highly Proficient	0.000	0.000	0.040	0.069		
Math 9	*Below Proficient*	0.161	0.059	0.001	0.000	68.63	0.550
	Approaching Proficient	0.042	0.279	0.081	0.001		
	Proficient	0.000	0.064	0.191	0.021		
	Highly Proficient	0.000	0.001	0.044	0.055		
Math 10	*Below Proficient*	0.286	0.082	0.002	0.000	68.10	0.538
	Approaching Proficient	0.051	0.234	0.063	0.001		
	Proficient	0.002	0.068	0.125	0.015		
	Highly Proficient	0.000	0.002	0.033	0.037		
Science 9	*Below Proficient*	0.198	0.066	0.011	0.000	59.34	0.450
	Approaching Proficient	0.056	0.131	0.081	0.005		
	Proficient	0.008	0.071	0.161	0.042		
	Highly Proficient	0.000	0.006	0.060	0.104		
Science 10	*Below Proficient*	0.276	0.092	0.020	0.000	57.70	0.399
	Approaching Proficient	0.064	0.123	0.076	0.003		
	Proficient	0.014	0.078	0.140	0.024		
	Highly Proficient	0.000	0.006	0.043	0.037		

12. Validity

As defined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" (p. 11). The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses.

The Utah Aspire Plus assessments are designed to measure the breadth and depth of the Utah Core Standards across all levels of student performance, to provide awareness of individual achievement in relation to stated performance expectations, and to provide evidence of whether students are on track for college and career readiness. The Utah Core Standards define what students should know and be able to do by the end of each respective school year.

Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of its validity, including test design, content specifications, item development, and psychometric characteristics. This technical report has detailed the processes implemented during the development, administration, and reporting cycles of the Utah Aspire Plus assessments. This section synthesizes the evidence using the validity framework outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) that organizes the evidence into five sources: evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing. While these sources highlight different facets of validity, they do not represent distinct types. Validity is a unified concept, reflecting the extent to which all accumulated evidence supports the intended interpretation and use of test scores (AERA et al., pp. 13–14).

12.1. Evidence Based on Test Content

Content validity evidence addresses whether a given assessment adequately samples from the full given domain. Where the assessment is determined to be representative in terms of the standards and in the manner intended, it is said to have high content validity. For the Utah Aspire Plus assessments, they are designed to measure the Utah Core Standards broadly.

The test design and blueprint specifications were developed in concert between USBE, Utah educators, and Pearson content experts well versed in the Utah Core Standards. Item and stimulus development targets focused on the measurement of the Utah Core Standards (SAGE) and on providing predictive measures of college and career readiness (ACT Aspire). Blueprints reflect a policy definition of how the makeup of a given assessment is intended to reflect an appropriate sampling of the standards necessary to meet the underlying reporting claims reliably, available online at https://utah.mypearsonsupport.com/admin-resources.html under "Reporting Resources."

All items were developed to measure the breadth of the Utah Core Standards or related standards. All items were rigorously scrutinized during the various expert content reviews, from initial creation through data review. These expert reviews check for the appropriateness of test items as aligned to the given standard. They also check that items are measuring intended targets of measurement, are clear and concise, and are appropriately aligned to a DOK level, as well as that vocabulary is appropriate for the given level, that the content is accurate and straightforward, and that supporting graphics or stimuli

are necessary to answer the question. Further reviews check for cluing within the context of an item set or test form. Every item is also evaluated for fairness by bias and sensitivity committees who review the items for language, or content, that may be inappropriate or offensive to students, parents, or community members, or that contain stereotypical or biased references to sex, ethnicity, or culture. As documented, USBE, Utah educators, Pearson, and the developers of the SAGE and ACT Aspire tests expended tremendous effort to ensure the Utah Aspire Plus tests are content-valid and support the intended claims detailed in this report. Additionally, evidence of the content coverage is presented in Appendix A.

ACT Aspire items are included on the Utah Aspire Plus assessments to provide Utah students a measure of college readiness, facilitate linking from Utah test scores to predicted ACT scores, and contribute to students' overall scores on the respective Utah Aspire Plus score scales. To ensure that specific items were aligned specifically with the Utah Core Standards, an alignment review meeting was conducted in July 2018. Experts from USBE, Pearson, and ACT initially matched items to their respective standards. Expert panels of Utah educators then reviewed the proposed item alignment designations for approval or suggest modification of a given alignment designation. The meeting agenda and training presentation are provided in Appendix B of the 2018–2019 technical report (Pearson, 2020). The result of the process was sufficient alignment of ACT Aspire items to the Utah Core Standards to fulfill the Utah Aspire Plus blueprints.

Lastly, Utah educators created and recommended PLDs for the Utah Aspire Plus tests that provide a description of typical end-of-grade performance expectations for each level of achievement in relation to the Utah Core Standards. The PLDs are descriptions of the knowledge and skills demonstrated by students in each performance category. Higher scores translate to a greater level of knowledge and skills demonstrated. There is a link between the PLDs and the knowledge and skills required to meet proficiency according to the standards. PLDs are used to relate performance on Utah Aspire Plus tests to the Utah Core Standards through the process of standard setting. Content experts and stakeholders participated in standard setting in August 2019 and in August 2022 to set the cut scores that delineate the four overall levels of achievement on the Utah Aspire Plus tests.

12.2. Evidence Based on Cognitive Process

Content comprising the Utah Aspire Plus assessments is specified by standard and DOK levels. Evidence related to DOK for items developed to measure the Utah Core Standards is provided in volume 4 (Validity) of the SAGE 2016–2017 technical report. The report notes that *the alignment of items by DOK also represents a structural model that can be evaluated using confirmatory factor analysis*. Further, they present a confirmatory factor analytic approach to evaluating DOK, where each item is an indicator of a DOK-level first-order factor, and each DOK is in turn an indicator of subject area achievement. They also describe evidence related to cognitive processes for SAGE content as being "highly similar" to content from the Smarter Balanced assessments and proceed to cite several formal cognitive lab studies that evaluated several facets of items by type and across subject areas.

ACT Aspire content also targets DOK within their development. The content reflects expectations that students need to think, reason, and analyze at high levels of cognitive complexity to be college- and career-ready, and that items and tasks require sampling different levels of cognitive complexity with most targeted at upper levels. ACT's definition of DOK is like Webb's, assigned to reflect complexity of the cognitive process required, not the psychometric "difficulty" of the item. Evidence of cognitive process is presented in Section 17.2.2 of their technical manual located online at https://actinc.my.salesforce.com/sfc/p/#300000000Wu5/a/4v0000005fHp/SLZ26Xzhfml8ibKP_Ca5G94_T3HuveFbNgFmfcRaHoY. The pilot of the ACT Aspire CR items used think-aloud tasks, surveys, and interviews to provide evidence of cognitive process.

12.3. Evidence Based on Internal Structure

Internal structure evidence shows the degree to which items and test components conform to the construct on which the proposed test score interpretations are based (AERA et al., 2014). For example, the Utah Aspire Plus tests report overall scale scores for individual students, as well as performance level indicators and ACT prediction ranges for reading, mathematics, and science at grades 9 and 10. Internal structure validity evidence identifies the degree to which the item relationships conform to the overall scores and individual subscales. While information is provided in the appendices examining the reporting categories as structural elements of design, the focus of evidence is intended to support the primary claim of each subject test as being unidimensional in nature and supportive of reporting a single overall scale score reflective of the given grade/subject Utah Aspire Plus assessment.

While individual items may each measure multiple elements of the standards and dimensions, they are crafted without dependencies on other items. As such, the tests are designed to be unidimensional and to measure the overall Utah Core Standards primarily. Assuming this holds true, it is appropriate to apply a unidimensional IRT model for calibrating and scaling the Utah Aspire Plus assessments. The IRT model application assumes that the domain being measured by the test is essentially unidimensional. To test this assumption, a principal components analysis is performed.

A general rule of thumb suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 in this analysis because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis within an IRT framework (Loehlin, 1987; Orlando, 2004). A scree plot is a convenient tool to examine results of factor analyses, as the resulting eigenvalues are plotted in order of magnitude. Appendix I presents the scree plots for the principal component analyses.

In addition to the principal components analyses, confirmatory factor analyses were also conducted to test the model of one factor construct within the Utah Aspire Plus assessments. Indices of model fit are used to determine how well this model fits the data. McDonald and Ho (2002) define absolute fit indices as determining how well an *a priori* model fits the sample data. The chi-square statistic assesses the magnitude of discrepancy between the sample and fitted covariance matrices (Hu and Bentler, 1999). However, this statistic is sensitive to sample size and often rejects the model when large samples are used (Bentler and Bonnet, 1980).

Alternatives to the chi-square, the goodness-of-fit statistic (GFI; Jöresky & Sörbom, 1993), and adjusted goodness-of-fit (AGFI; Tabachnick & Fidell, 2007) are also sensitive to sample size, which has led to researchers reporting them along with other fit indices (Hooper et al., 2008).

The root mean square error of approximation (RMSEA), a comparative fit index, tells how well the model would fit the population covariance matrix (Byrne, 1998). This fit index favors parsimony since it is sensitive to the number of estimated parameters in the model. There have been a few suggestions of index threshold cut-offs of good fit. The most stringent criterion is 0.06, as suggested in Hu and Bentler (1999). In addition, a confidence interval can be constructed for RMSEA, with a lower limit close to 0 signifying a well-fitting model as well as an upper limit less than 0.08.

The root mean square residual (RMR) and standardized root mean square residual (SRMR) are the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance model. The SRMR has a range of 0 to 1, with 0 indicating perfect fit. Byrne (1999) suggests well-fitting models having an SRMR less than 0.05. Hooper, Coughlan, and Mullen (2008) caution that SRMR will tend to be low with a high number of parameters and models with large sample sizes. Hu and Bentler (1999) suggested a two-index presentation when reporting model fit evaluation. One proposed combination is the RMSEA, with confidence interval, and the SRMR. Table 12.1 presents the estimates of these indices.

Table 12.1. Model Fit Indices for Confirmatory Factor Analyses

Assessment	SRMR	RMSEA Estimate	RMSEA Lower 90% CL	RMSEA Upper 90% CL
Reading 9	0.0211	0.0240	0.0236	0.0243
Reading 10	0.0229	0.0281	0.0278	0.0284
Mathematics 9	0.0210	0.0251	0.0248	0.0254
Mathematics 10	0.0193	0.0234	0.0231	0.0237
Science 9	0.0172	0.0234	0.0228	0.0239
Science 10	0.0207	0.0268	0.0263	0.0274

Note. CL = confidence limit

Model-data fit based on the IRT model calibrations are also indicators of unidimensionality. To the extent that indicators of fit suggest data do not appropriately fit the model as applied may be the result of multidimensionality. Discussion of model fit is presented in Section 9.3.2 in terms of Q_1 indices. These statistics support the overall fit of Utah Aspire Plus items to the respective IRT models.

In addition to evidence of essential unidimensionality described here, it should be acknowledged that tests are not designed to be *strictly* unidimensional. It is common to observe what might be considered transient factors common to one or more test items in the face of a dominant overall factor. As discussed in Section 2, the Utah Aspire Plus blueprints were designed to reflect the Utah Core Standards partly around reporting categories. Correlations among the Utah Aspire Plus overall test scores and reporting categories offer additional evidence of the internal structure of the Utah Aspire Plus tests. These correlations quantify the strength of the relationships across structural elements of the assessments. Results of these analyses are presented in Appendix J.

Additionally, the reliability analyses presented in Section 11 provide information about the internal consistency of the Utah Aspire Plus tests. Internal consistency is typically measured by correlations among the items on a test and provides an indication of how much the items measure the same general construct.

12.4. Evidence Based on Different Student Populations

Internal structure evidence should also show that individual items are functioning similarly for different demographic subgroups within the population being measured. The Utah Aspire Plus tests are developed to assess the Utah Core Standards and are administered to all students irrespective of any particular demographic characteristic. Great care has been taken to ensure the items on the Utah Aspire Plus tests are fair and representative of the content domains expressed in the standards. Special attention is given to finding evidence that construct-irrelevant content has not been inadvertently included in the test, as such content could result in an unfair advantage for one group versus another.

This begins with item writers trained on how to avoid economic, regional, cultural, and ethnic biases when writing items. After items have been written, they are reviewed by a bias and sensitivity committee, which evaluates each item to identify language or content that might be inappropriate or offensive to students, parents, or other community members or that contain stereotypical or biased references to sex, ethnic, or cultural groups. The bias and sensitivity committee accepts, edits, or rejects each item for use prior to the items' administration.

Differential item functioning (DIF) analyses are conducted for the purpose of identifying items that are differentially difficult for different subpopulations of individuals. Section 8.2 details the methodology used to evaluate DIF for the Utah Aspire Plus items. Though DIF analyses flag items as being differentially difficult for one group as compared to another, it does not solely provide sufficient evidence for removing the item from use. Flagged items are re-examined post administration for any potentially overlooked biases attributable to the content of those items.

12.5. Summary

The process of validation involves accumulating relevant evidence to provide a sound scientific basis for stated score interpretations. Collection of validity evidence is an ongoing process and validity of interpretations are strengthened as positive evidence accrues. While this technical report reflects the continued administration of the Utah Aspire Plus assessments, sufficient evidence exists to support the primary claims detailed herein, including that test scores indicate the degree to which students achieved end-of-year expectations on the Utah Core Standards across subject tests in grades 9 and 10. Further, performance on the Utah Aspire Plus assessments could reasonably be linked to predictions of performance on the ACT college and career readiness benchmarks. These are supported by evidence of the content development processes that underpin the creation of assessments aligned to the Utah Core Standards and evidence that the internal structure aligns with the stated claims and is sound.

References

ACT Aspire. (2017). *Summative technical manual*. Version 3. ACT.

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

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming. Lawrence Erlbaum Associates.

Chien, M., & Shin, D. (2012). *IRT score estimation program*, V1.3 [computer program]. Pearson.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–47.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Davis, L. L., & Moyer, E. L. (2015, December). *PARCC performance level setting technical report*. https://eric.ed.gov/?q=source%3a%22Partnership+for+Assessment+of+Readiness+for+College+and+Careers%22&pg=2&id=ED599257

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

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

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*, 53–60.

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

Jöresky, K., & Sörbom, D. (1993). *LISREL 8: structural equation modeling with the SIMPLIS command language*. Scientific Software International Inc.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kim, S. and Kolen, M. (2004). STUIRT [computer program]. The University of Iowa.

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

Loehlin, J. C. (1987). *Latent variable models*. Lawrence Erlbaum Associates.

McDonald, R. P., & Ho, M.–H. R. (2002). Principles and practice in reporting statistical equation analyses. *Psychological Methods, 7*(1), 64–82.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement 16*, 159–176.

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

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

Orlando, M. (2004, June). *Critical issues to address when applying item response theory (IRT) models*. Paper presented at the Drug Information Association, Bethesda, M. D.

Pearson. (2019). *Utah Aspire Plus English, reading, science, and mathematics standard setting technical report*. Submitted to the Utah State Board of Education.

Pearson. (2021). *Utah Aspire Plus science grade 9 and 10 standard setting technical report*. Submitted to the Utah State Board of Education.

Pearson. (2020). *Utah Aspire Plus 2018–2019 technical report*. https://utah.mypearsonsupport.com/assets/pdf/UT1132740_UTPlusTechReportv4.3_WebTag.pdf

Pearson. (2021). *Utah Aspire Plus 2020–2021 technical report*. https://utah.mypearsonsupport.com/assets/pdf/UT1140119_UTPlusTechReport2022_WebTag.pdf

Pearson. (2022). *Utah Aspire Plus 2021–2022 technical report*. https://utah.mypearsonsupport.com/assets/pdf/UA+%202022%20Tech%20Report_v2_WEBTAG.pdf

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

Scientific Software International, Inc. (2017). IRTPRO. www.ssicentral.com

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn and Bacon.

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education*, 27, 214–231.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Appendix A: Test-Level Reporting Categories and Standards by Item Type and DOK

Table A.1. Test-Level Reporting Categories and Standards—Reading Grade 9

Reporting Category: Standard	MC DOK 1	MC DOK 2	MC DOK 3	TE DOK 1	TE DOK 2	TE DOK 3	EBSR DOK 1	EBSR DOK 2	EBSR DOK 3	Total
Key Ideas: 9-10.R.5	3	4	2	0	0	0	0	0	0	9
Key Ideas: 9-10.R.6	0	3	0	0	1	0	0	0	1	5
Key Ideas: 9-10.R.7	0	2	0	0	0	0	0	0	0	2
Craft and Structure: 9-10.R.10	0	3	2	0	0	0	0	0	0	5
Craft and Structure: 9-10.R.11	0	0	1	0	0	0	0	0	1	2
Craft and Structure: 9-10.R.8	0	0	0	0	0	1	0	0	1	2
Craft and Structure: 9-10.R.9	2	2	0	0	0	0	0	0	0	4
Integration of Knowledge and Ideas: 9-10.R.12	0	0	1	0	0	1	0	0	0	2
Integration of Knowledge and Ideas: 9-10.R.13	1	0	0	0	1	0	0	0	0	2
Integration of Knowledge and Ideas: 9-10.R.14	0	0	0	0	0	0	0	0	2	2
Total	6	14	6	0	2	2	0	0	5	35

Table A.2. Test-Level Reporting Categories and Standards—Reading Grade 10

Reporting Category: Standard	MC DOK 1	MC DOK 2	MC DOK 3	TE DOK 1	TE DOK 2	TE DOK 3	EBSR DOK 1	EBSR DOK 2	EBSR DOK 3	Total
Key Ideas: 9-10.R.5	3	4	2	1	0	0	0	0	0	10
Key Ideas: 9-10.R.6	0	1	0	0	0	0	0	0	0	1
Key Ideas: 9-10.R.7	0	1	1	0	0	0	0	0	2	4
Craft and Structure: 9-10.R.10	0	0	3	0	0	0	0	0	1	4
Craft and Structure: 9-10.R.11	0	0	1	0	0	0	0	0	0	1
Craft and Structure: 9-10.R.8	1	3	0	0	2	0	0	0	0	6
Craft and Structure: 9-10.R.9	0	2	0	0	0	0	0	0	1	3
Integration of Knowledge and Ideas: 9-10.R.13	1	1	1	0	0	0	0	0	0	3
Integration of Knowledge and Ideas: 9-10.R.14	0	0	0	0	2	0	0	0	1	3
Total	5	12	8	1	4	0	0	0	5	35

Table A.3. Test-Level Reporting Categories and Standards—Mathematics Grade 9

Reporting Category: Standard	MC DOK 1	MC DOK 2	MC DOK 3	TE DOK 1	TE DOK 2	TE DOK 3	Total
Algebra: MI.A.CED.1	0	1	0	0	0	0	1
Algebra: MI.A.CED.2	1	1	0	0	0	0	2
Algebra: MI.A.CED.4	0	1	0	0	0	0	1
Algebra: MI.A.REI.1	0	0	0	0	1	0	1
Algebra: MI.A.REI.12	1	0	0	0	0	0	1
Algebra: MI.A.REI.3	0	1	0	0	0	1	2
Algebra: MI.A.REI.6	0	1	0	0	0	1	2
Algebra: MI.A.SSE.1b	0	0	1	0	0	0	1
Functions: MI.F.BF.1a	1	0	0	0	0	0	1
Functions: MI.F.BF.2	1	0	0	0	0	0	1
Functions: MI.F.IF.1	1	0	0	0	0	0	1
Functions: MI.F.IF.2	1	0	0	0	0	0	1
Functions: MI.F.IF.4	0	0	0	0	0	1	1
Functions: MI.F.IF.6	1	0	0	0	0	0	1
Functions: MI.F.IF.7a	0	1	0	0	0	0	1
Functions: MI.F.LE.1b	1	0	0	0	0	0	1
Functions: MI.F.LE.1c	0	1	0	0	0	0	1
Functions: MI.F.LE.2	0	1	0	0	0	0	1
Functions: MI.F.LE.5	0	1	0	0	0	0	1
Geometry: MI.G.CO.3	0	0	1	0	0	1	2
Geometry: MI.G.CO.4	0	0	1	0	0	0	1
Geometry: MI.G.CO.5	0	0	0	1	0	0	1
Geometry: MI.G.CO.6	0	1	0	0	0	0	1
Geometry: MI.G.CO.7	0	1	0	0	0	0	1
Geometry: MI.G.CO.8	1	0	0	0	0	0	1
Geometry: MI.G.GPE.4	0	0	0	0	1	0	1
Geometry: MI.G.GPE.5	0	1	0	0	0	0	1
Geometry: MI.G.GPE.7	0	1	0	0	0	0	1
Statistics and Probability: MI.S.ID.1	0	0	0	0	1	1	2
Statistics and Probability: MI.S.ID.2	0	1	0	0	0	0	1
Statistics and Probability: MI.S.ID.3	0	0	1	0	0	0	1
Statistics and Probability: MI.S.ID.6	0	1	0	0	0	0	1
Statistics and Probability: MI.S.ID.6c	1	0	0	0	0	0	1
Statistics and Probability: MI.S.ID.7	1	0	0	0	0	0	1
Statistics and Probability: MI.S.ID.8	0	1	0	0	0	0	1
Total	11	16	4	1	3	5	40

Table A.4. Test-Level Reporting Categories and Standards—Mathematics Grade 10

Reporting Category: Standard	MC DOK 1	MC DOK 2	MC DOK 3	TE DOK 1	TE DOK 2	TE DOK 3	Total
Number and Quantity: MII.N.CN.1	0	0	0	0	0	1	1
Number and Quantity: MII.N.RN.1	1	0	0	0	0	0	1
Number and Quantity: MII.N.RN.2	0	2	0	0	0	0	2
Algebra: MII.A.APR.1	1	0	0	0	0	0	1
Algebra: MII.A.CED.1	0	1	0	0	0	0	1
Algebra: MII.A.CED.2	1	0	0	0	0	0	1
Algebra: MII.A.CED.4	1	0	0	0	0	0	1
Algebra: MII.A.REI.4b	0	1	0	0	1	0	2
Algebra: MII.A.REI.7	0	0	0	1	0	0	1
Algebra: MII.A.SSE.2	0	2	0	0	0	0	2
Algebra: MII.A.SSE.3a	0	1	0	0	0	0	1
Functions: MII.F.BF.1a	0	0	0	0	0	1	1
Functions: MII.F.BF.1b	0	1	0	0	0	0	1
Functions: MII.F.BF.3	0	1	0	0	0	0	1
Functions: MII.F.IF.4	0	1	1	0	0	0	2
Functions: MII.F.IF.5	1	0	0	0	0	0	1
Functions: MII.F.IF.7b	1	0	0	0	0	0	1
Functions: MII.F.IF.9	0	1	0	0	0	0	1
Functions: MII.F.LE.3	0	0	0	0	1	0	1
Functions: MII.F.TF.8	1	0	0	0	0	0	1
Geometry: MII.G.C.2	1	0	0	0	0	0	1
Geometry: MII.G.C.4	0	0	0	0	0	1	1
Geometry: MII.G.CO.10	0	0	1	0	0	0	1
Geometry: MII.G.CO.11	0	1	0	0	0	0	1
Geometry: MII.G.CO.9	0	2	0	0	0	0	2
Geometry: MII.G.GMD.3	0	1	0	0	0	0	1
Geometry: MII.G.GPE.1	1	0	0	0	0	0	1
Geometry: MII.G.GPE.6	0	0	0	0	1	0	1
Geometry: MII.G.SRT.2	0	0	1	0	0	0	1
Geometry: MII.G.SRT.3	1	0	0	0	0	0	1
Geometry: MII.G.SRT.4	0	0	0	0	0	1	1
Geometry: MII.G.SRT.6	0	1	0	0	0	0	1
Statistics and Probability: MII.S.CP.1	0	0	0	1	0	0	1
Statistics and Probability: MII.S.CP.4	0	0	0	0	0	1	1
Statistics and Probability: MII.S.CP.6	0	0	1	0	0	0	1
Total	10	16	4	2	3	5	40

Appendix B: Student Testing Time Plots

Figure B.1. Student Testing Time Plot—Reading Grade 9

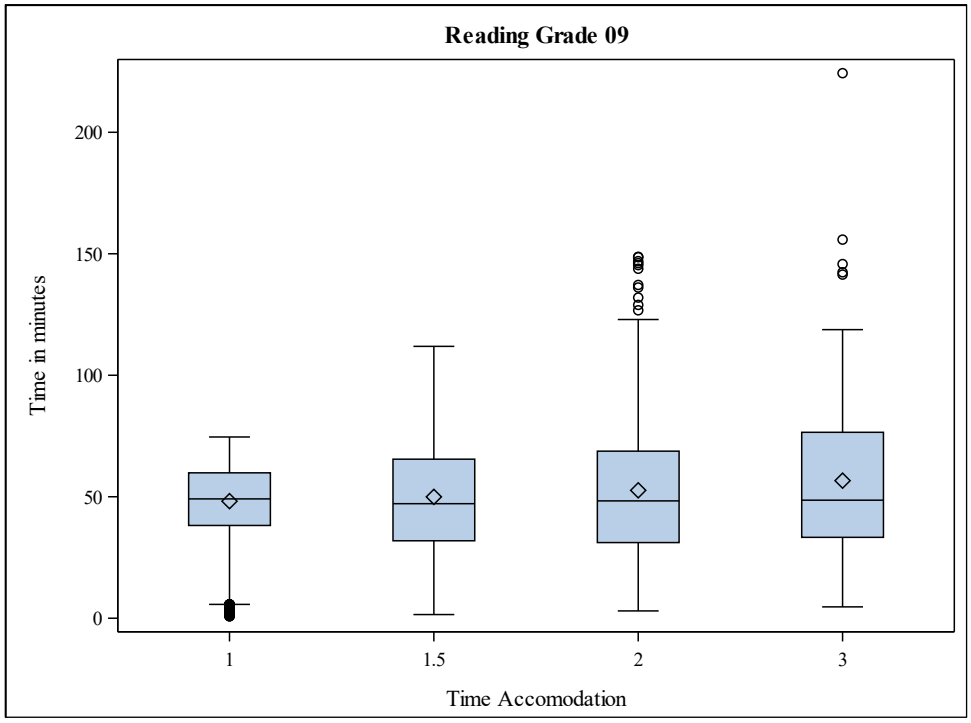


Figure B.2. Student Testing Time Plot—Reading Grade 10

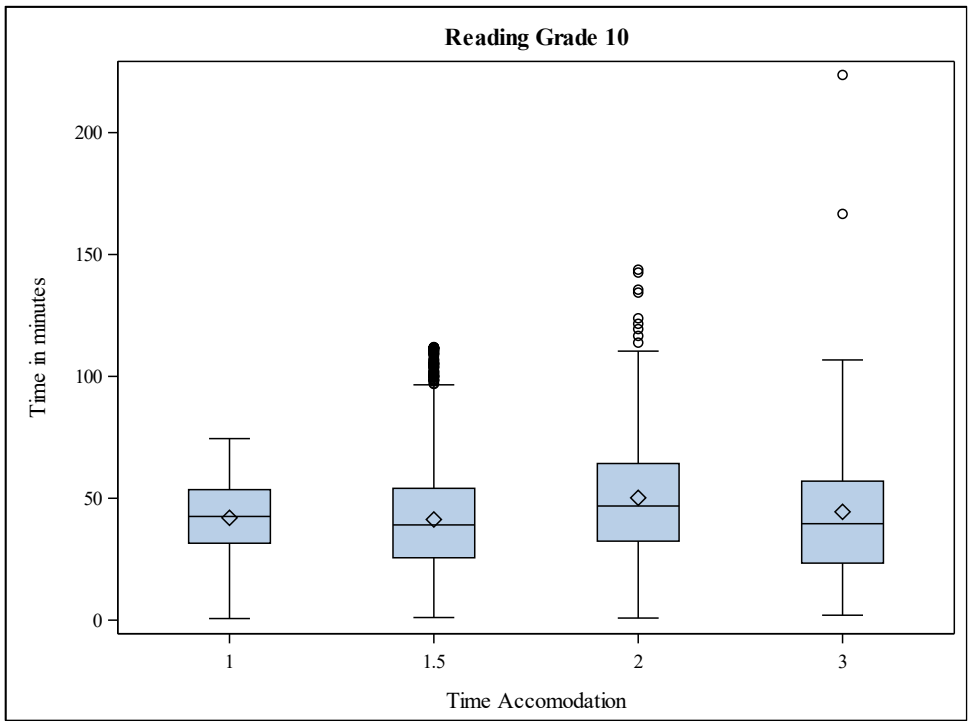


Figure B.3. Student Testing Time Plot—Mathematics Grade 9

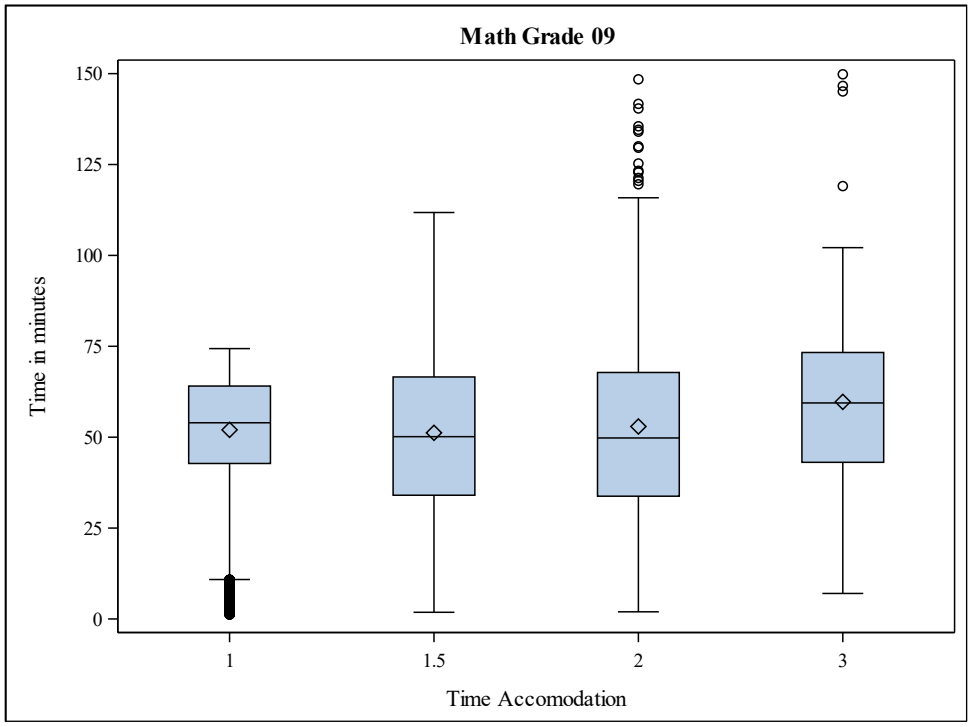


Figure B.4. Student Testing Time Plot—Mathematics Grade 10

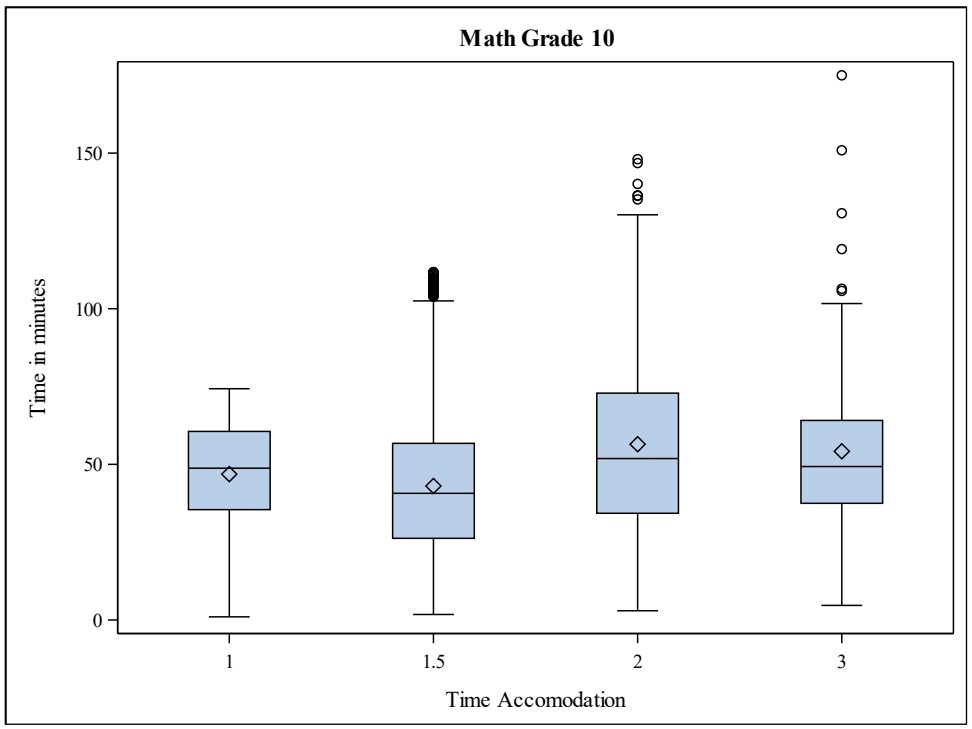


Figure B.5. Student Testing Time Plot—Science Grade 9

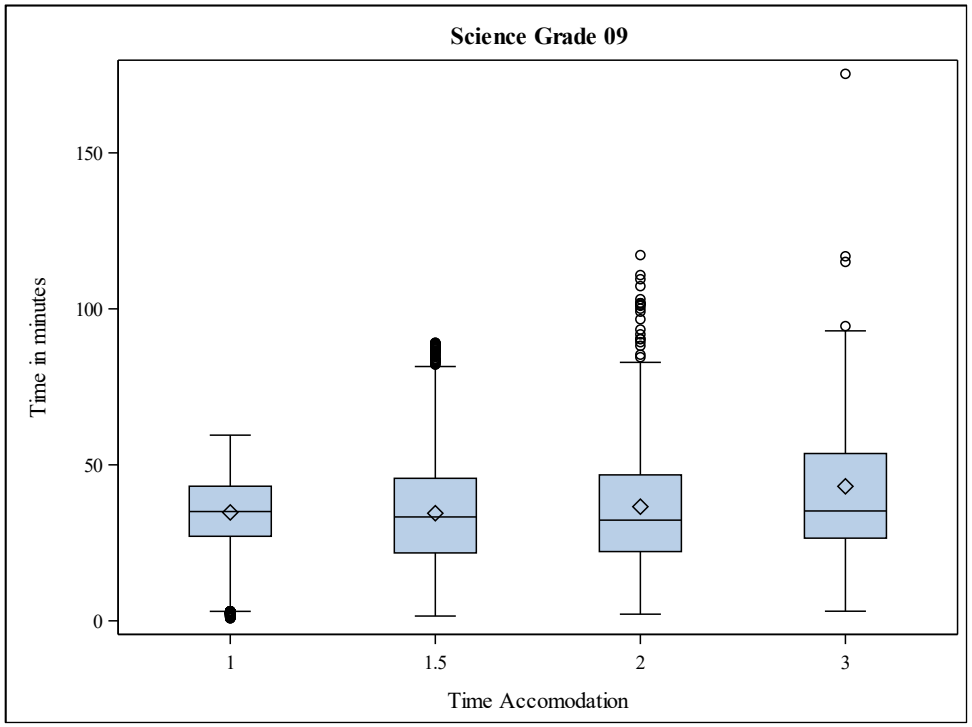
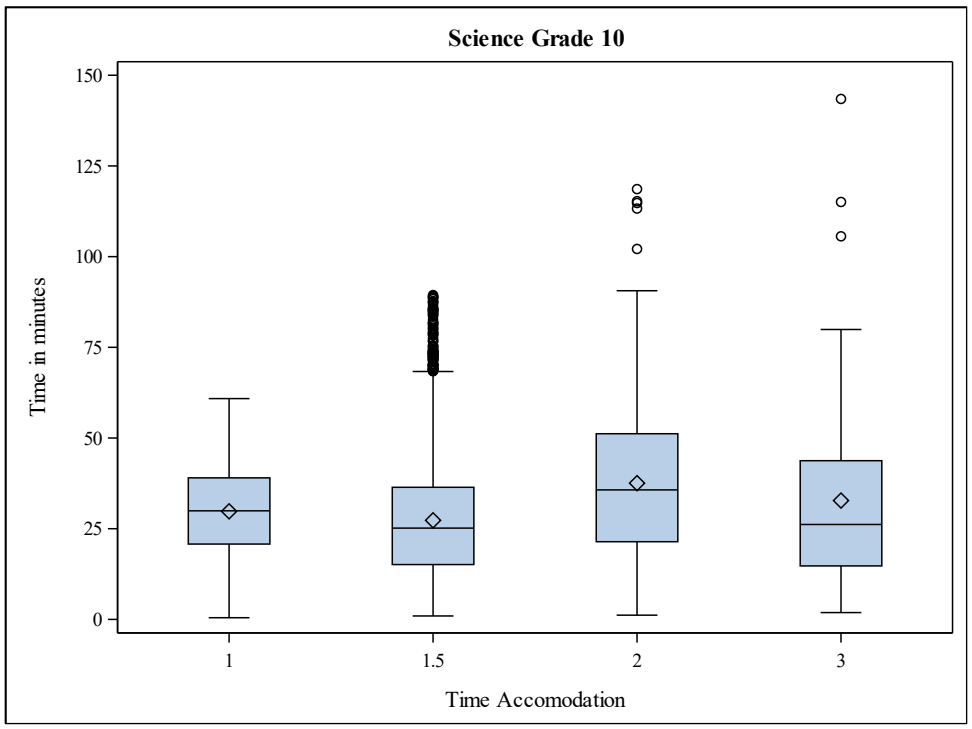


Figure B.6. Student Testing Time Plot—Science Grade 10



Appendix C: Reliability Results by Subgroup

Table C.1. Test Reliability by Subgroup and Reporting Category—Reading Grade 9

Subgroup	#Students	Alpha	SEM	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
Total #Students Tested	45,079	0.88	9.91	0.79	0.70	0.51
Female	21,840	0.87	9.81	0.78	0.69	0.49
Male	23,214	0.88	9.98	0.80	0.71	0.52
Hispanic or Latino Ethnicity	9,154	0.85	10.39	0.75	0.66	0.47
Asian	794	0.88	10.13	0.80	0.71	0.55
Native Hawaiian or Other Pacific Islander	645	0.84	10.21	0.73	0.61	0.48
Black or African American	607	0.85	10.49	0.74	0.67	0.52
American Indian or Alaska Native	418	0.82	10.03	0.70	0.66	0.36
White	31,846	0.87	9.75	0.78	0.68	0.49
Other	1,615	0.87	9.94	0.79	0.70	0.50
Limited English Proficient – No	41,510	0.87	9.76	0.78	0.69	0.50
Limited English Proficient – Yes	3,569	0.70	12.25	0.55	0.43	0.27
Economic Disadvantaged – No	33,448	0.87	9.80	0.78	0.69	0.50
Economic Disadvantaged – Yes	11,631	0.86	10.24	0.76	0.67	0.47
Special Education – No	40,616	0.87	9.83	0.78	0.69	0.50
Special Education – Yes	4,463	0.79	10.84	0.65	0.57	0.35

Table C.2. Test Reliability by Subgroup and Reporting Category—Reading Grade 10

Subgroup	#Students	Alpha	SEM	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
Total #Students Tested	43,615	0.90	8.43	0.80	0.82	0.55
Female	20,803	0.90	8.52	0.78	0.81	0.52
Male	22,787	0.91	8.36	0.80	0.83	0.57
Hispanic or Latino Ethnicity	8,939	0.89	8.11	0.77	0.80	0.52
Asian	776	0.90	8.51	0.79	0.83	0.58
Native Hawaiian or Other Pacific Islander	574	0.88	7.99	0.75	0.77	0.48
Black or African American	584	0.89	7.95	0.77	0.80	0.50
American Indian or Alaska Native	400	0.88	8.24	0.74	0.77	0.54
White	30,829	0.90	8.53	0.78	0.81	0.53
Other	1,513	0.90	8.37	0.79	0.81	0.52
Limited English Proficient – No	40,206	0.90	8.46	0.78	0.81	0.53
Limited English Proficient – Yes	3,409	0.78	8.42	0.54	0.66	0.32
Economic Disadvantaged – No	32,859	0.90	8.52	0.79	0.81	0.53
Economic Disadvantaged – Yes	10,756	0.90	8.18	0.78	0.81	0.53
Special Education – No	39,515	0.90	8.47	0.78	0.81	0.53
Special Education – Yes	4,100	0.86	8.29	0.71	0.75	0.47

Table C.3. Test Reliability by Subgroup and Reporting Category—Mathematics Grade 9

Subgroup	#Students	Alpha	SEM	Algebra	Functions	Geometry	Statistics and Probability
Total #Students Tested	43,894	0.91	9.58	0.78	0.72	0.74	0.62
Female	21,132	0.90	9.46	0.76	0.69	0.71	0.58
Male	22,739	0.92	9.64	0.79	0.75	0.76	0.65
Hispanic or Latino Ethnicity	8,818	0.88	11.48	0.73	0.63	0.67	0.48
Asian	769	0.92	9.05	0.80	0.76	0.73	0.66
Native Hawaiian or Other Pacific Islander	613	0.85	11.61	0.68	0.59	0.65	0.44
Black or African American	588	0.87	12.36	0.70	0.62	0.69	0.43
American Indian or Alaska Native	402	0.86	11.91	0.72	0.57	0.63	0.42
White	31,119	0.91	9.05	0.77	0.72	0.72	0.62
Other	1,585	0.91	9.78	0.78	0.73	0.72	0.63
Limited English Proficient – No	40,400	0.91	9.30	0.77	0.72	0.73	0.62
Limited English Proficient – Yes	3,494	0.75	15.04	0.58	0.42	0.48	0.26
Economic Disadvantaged – No	32,629	0.91	9.08	0.77	0.72	0.72	0.62
Economic Disadvantaged – Yes	11,265	0.89	11.11	0.75	0.67	0.71	0.54
Special Education – No	39,489	0.91	9.18	0.77	0.72	0.73	0.61
Special Education – Yes	4,405	0.83	13.71	0.65	0.52	0.60	0.39

Table C.4. Test Reliability by Subgroup and Reporting Category—Mathematics Grade 10

Subgroup	#Students	Alpha	SEM	Number and Quantity	Algebra	Functions	Geometry	Statistics and Probability
Total #Students Tested	42,439	0.91	10.07	0.51	0.75	0.68	0.80	0.41
Female	20,132	0.90	9.80	0.46	0.72	0.64	0.79	0.35
Male	22,285	0.92	10.22	0.55	0.77	0.72	0.82	0.46
Hispanic or Latino Ethnicity	8,589	0.87	12.50	0.39	0.67	0.54	0.71	0.29
Asian	759	0.93	9.32	0.62	0.79	0.77	0.83	0.42
Native Hawaiian or Other Pacific Islander	576	0.85	13.29	0.39	0.66	0.51	0.67	0.33
Black or African American	547	0.86	13.86	0.37	0.68	0.55	0.68	0.12
American Indian or Alaska Native	388	0.84	13.81	0.33	0.62	0.43	0.68	0.22
White	30,111	0.91	9.47	0.51	0.74	0.69	0.80	0.39
Other	1,469	0.91	9.88	0.51	0.75	0.69	0.80	0.44
Limited English Proficient – No	39,139	0.91	9.71	0.51	0.74	0.69	0.80	0.40
Limited English Proficient – Yes	3,300	0.71	18.12	0.19	0.48	0.30	0.44	0.11
Economic Disadvantaged – No	32,092	0.91	9.56	0.51	0.74	0.69	0.80	0.40
Economic Disadvantaged – Yes	10,347	0.88	12.03	0.43	0.69	0.58	0.75	0.33
Special Education – No	38,411	0.91	9.67	0.51	0.74	0.69	0.80	0.40
Special Education – Yes	4,028	0.77	16.38	0.22	0.52	0.35	0.56	0.22

Table C.5. Test Reliability by Subgroup and Reporting Category—Science Grade 9

Subgroup	#Students	Alpha	SEM	Gathering & Investigating	Developing Models	Using Mathematical Thinking	Construct Explanations
Total #Students Tested	45,006	0.86	12.48	0.64	0.58	0.62	0.62
Female	21,792	0.84	12.35	0.61	0.55	0.58	0.60
Male	23,189	0.87	12.55	0.66	0.60	0.65	0.64
Hispanic or Latino Ethnicity	9,191	0.83	12.99	0.59	0.53	0.57	0.55
Asian	797	0.87	12.84	0.66	0.58	0.61	0.66
Native Hawaiian or Other Pacific Islander	644	0.80	13.10	0.54	0.54	0.52	0.51
Black or African American	618	0.81	13.02	0.55	0.51	0.53	0.53
American Indian or Alaska Native	419	0.78	14.32	0.54	0.45	0.49	0.50
White	31,712	0.85	12.29	0.62	0.56	0.59	0.61
Other	1,625	0.85	12.29	0.62	0.55	0.60	0.60
Limited English Proficient – No	41,390	0.85	12.37	0.63	0.56	0.60	0.61
Limited English Proficient – Yes	3,616	0.68	14.94	0.38	0.36	0.38	0.35
Economic Disadvantaged – No	33,302	0.85	12.29	0.62	0.56	0.60	0.61
Economic Disadvantaged – Yes	11,704	0.84	13.00	0.60	0.55	0.60	0.58
Special Education – No	40,531	0.85	12.33	0.62	0.56	0.60	0.61
Special Education – Yes	4,475	0.77	14.18	0.43	0.46	0.51	0.49

Table C.6. Test Reliability by Subgroup and Reporting Category—Science Grade 10

Subgroup	#Students	Alpha	SEM	Gathering & Investigating	Developing Models	Using Mathematical Thinking	Construct Explanations
Total #Students Tested	43,308	0.82	14.37	0.71	0.52	0.42	0.52
Female	20,635	0.79	14.67	0.70	0.44	0.36	0.49
Male	22,648	0.84	14.12	0.72	0.57	0.46	0.55
Hispanic or Latino Ethnicity	8,896	0.73	16.73	0.61	0.39	0.33	0.35
Asian	769	0.82	13.24	0.72	0.53	0.42	0.55
Native Hawaiian or Other Pacific Islander	567	0.63	18.32	0.54	0.26	0.21	0.19
Black or African American	564	0.69	17.24	0.58	0.34	0.33	0.24
American Indian or Alaska Native	390	0.64	16.86	0.49	0.32	0.25	0.26
White	30,618	0.82	13.83	0.72	0.53	0.42	0.55
Other	1,504	0.82	14.39	0.71	0.54	0.41	0.52
Limited English Proficient – No	39,895	0.82	14.09	0.71	0.52	0.41	0.53
Limited English Proficient – Yes	3,413	0.42	21.61	0.23	0.20	0.16	0.05
Economic Disadvantaged – No	32,619	0.82	13.96	0.72	0.53	0.42	0.55
Economic Disadvantaged – Yes	10,689	0.76	16.01	0.65	0.42	0.35	0.40
Special Education – No	39,246	0.82	14.07	0.71	0.52	0.42	0.53
Special Education – Yes	4,062	0.63	18.84	0.45	0.32	0.24	0.22

Appendix D: Conditional Standard Error of Scale Scores

Figure D.1. CSEM of Scale Scores—Reading Grade 9

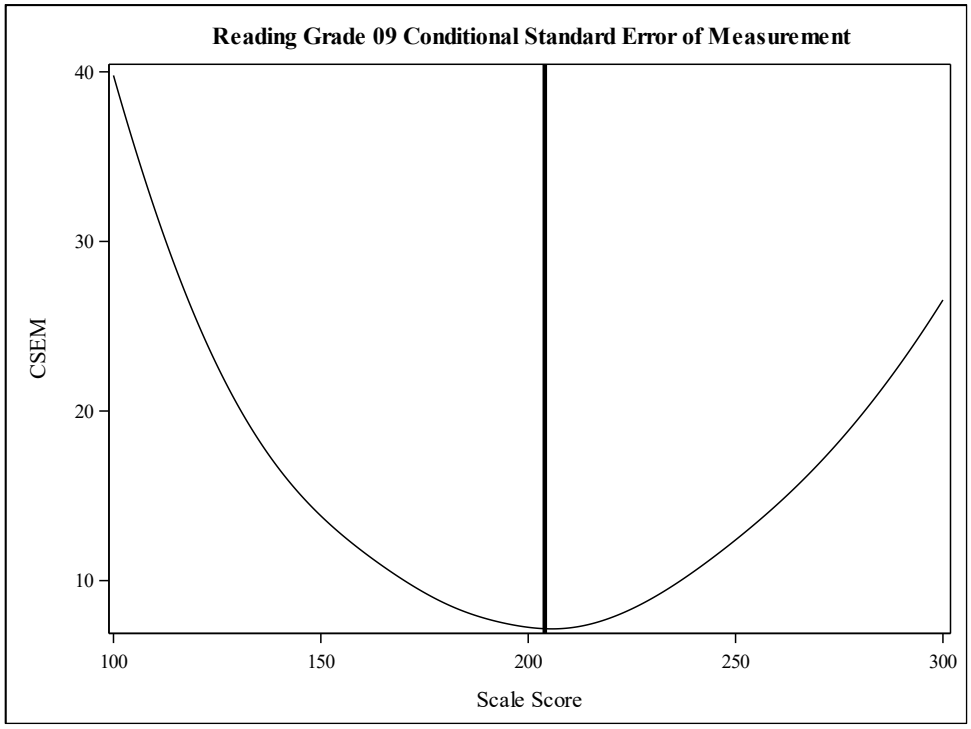


Figure D.2. CSEM of Scale Scores—Reading Grade 10

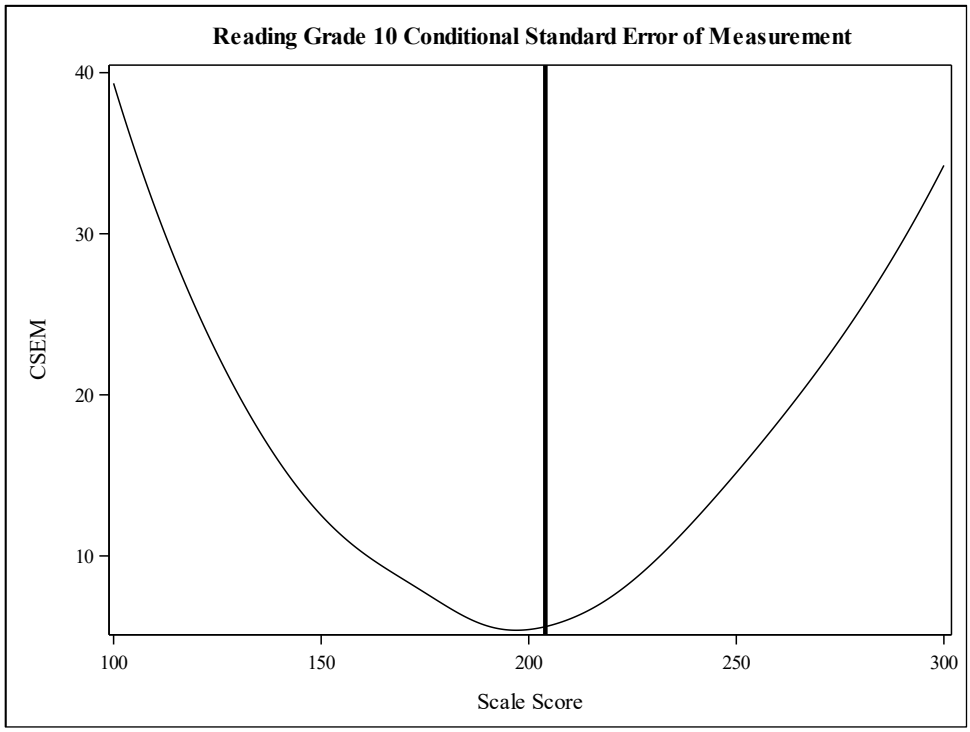


Figure D.3. CSEM of Scale Scores—Mathematics Grade 9

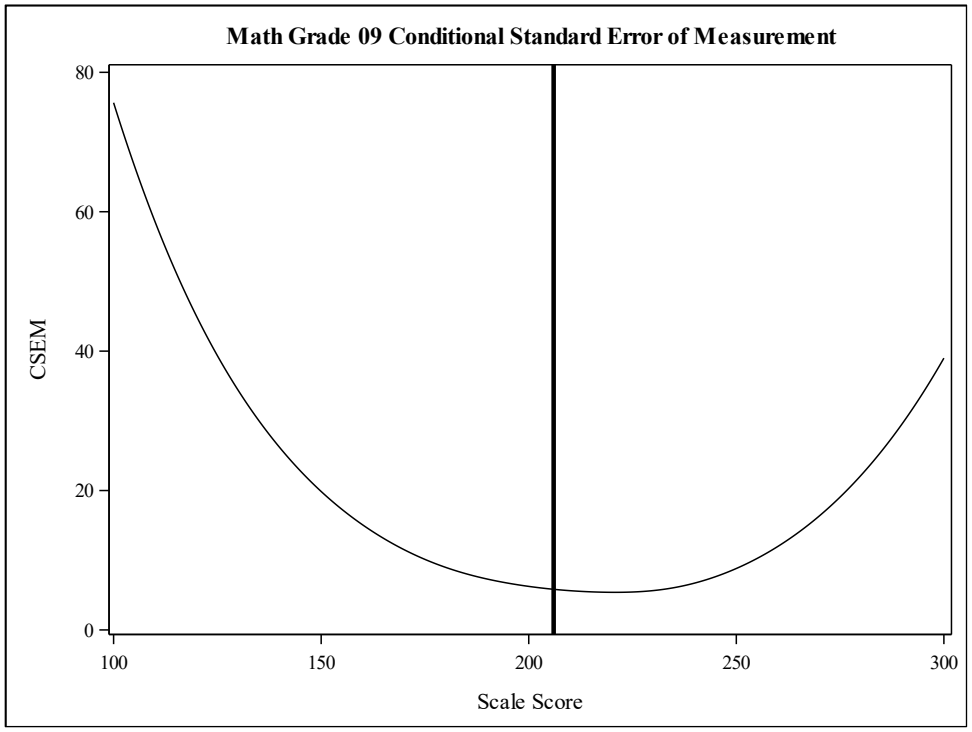


Figure D.4. CSEM of Scale Scores—Mathematics Grade 10

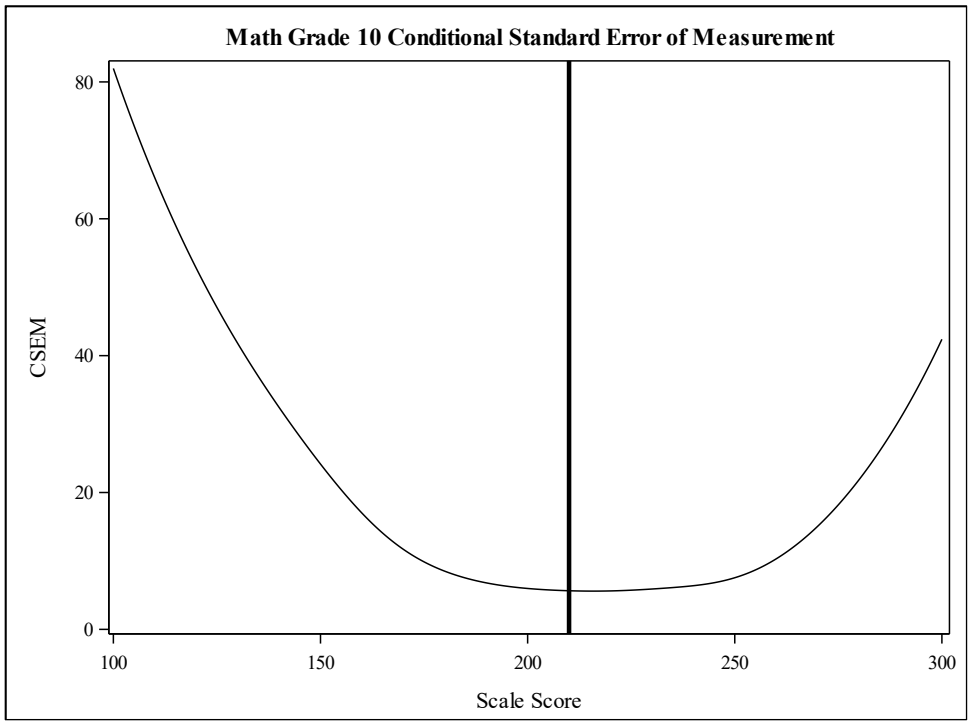


Figure D.5. CSEM of Scale Scores—Science Grade 9

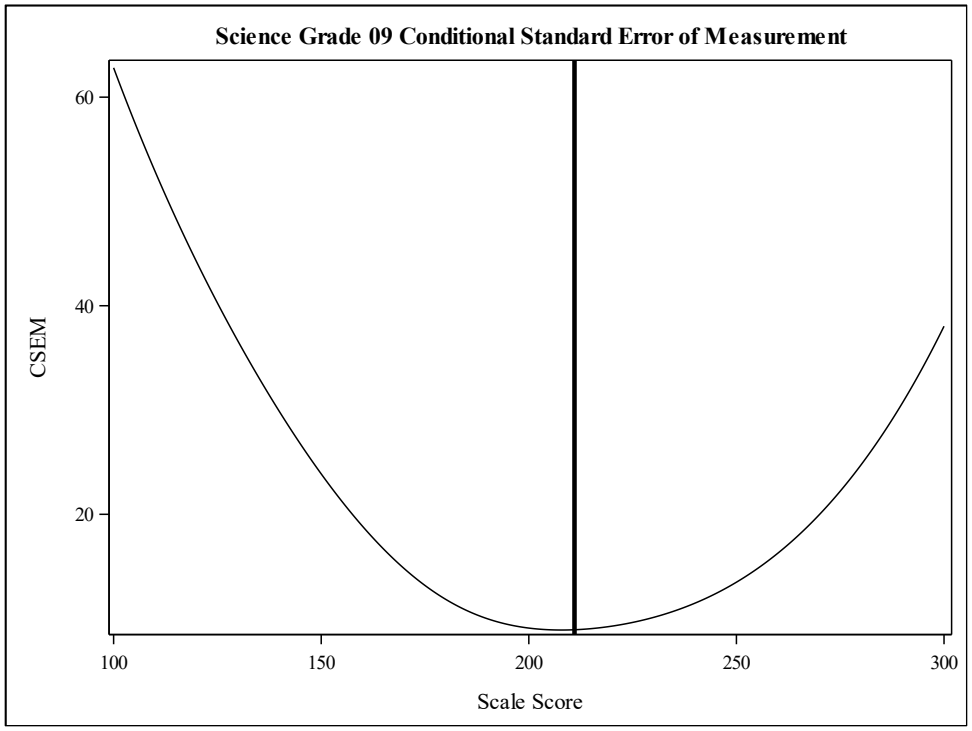
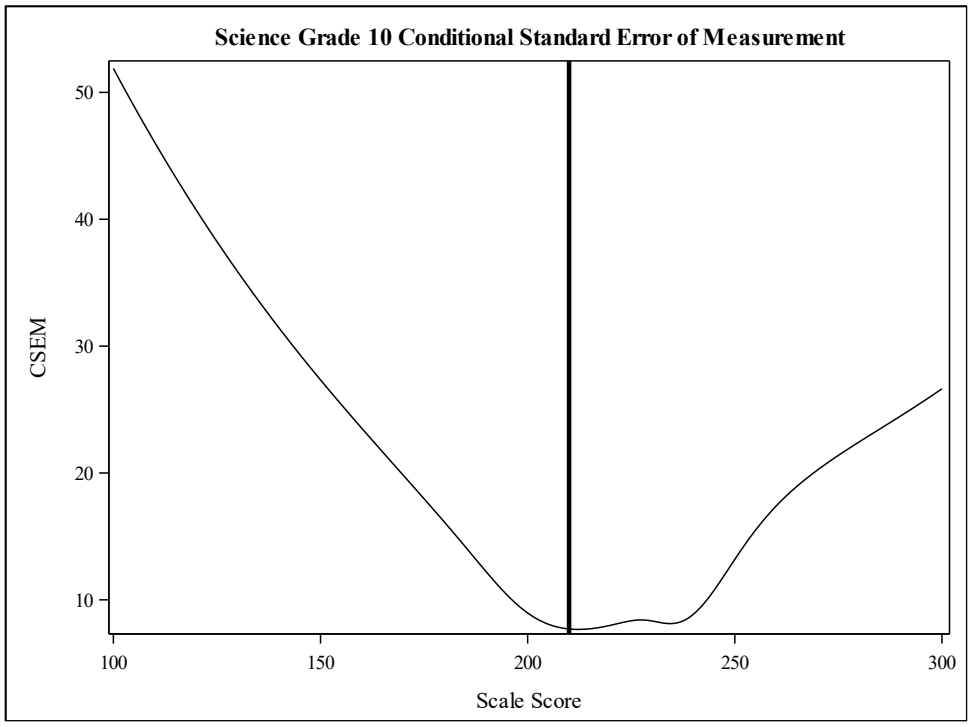


Figure D.6. CSEM of Scale Scores—Science Grade 10



Appendix E: Common Item Scatter Plots for 2025 Anchor Items

Figure E.1. IRT B Parameters for Operational Items—Reading Grade 9

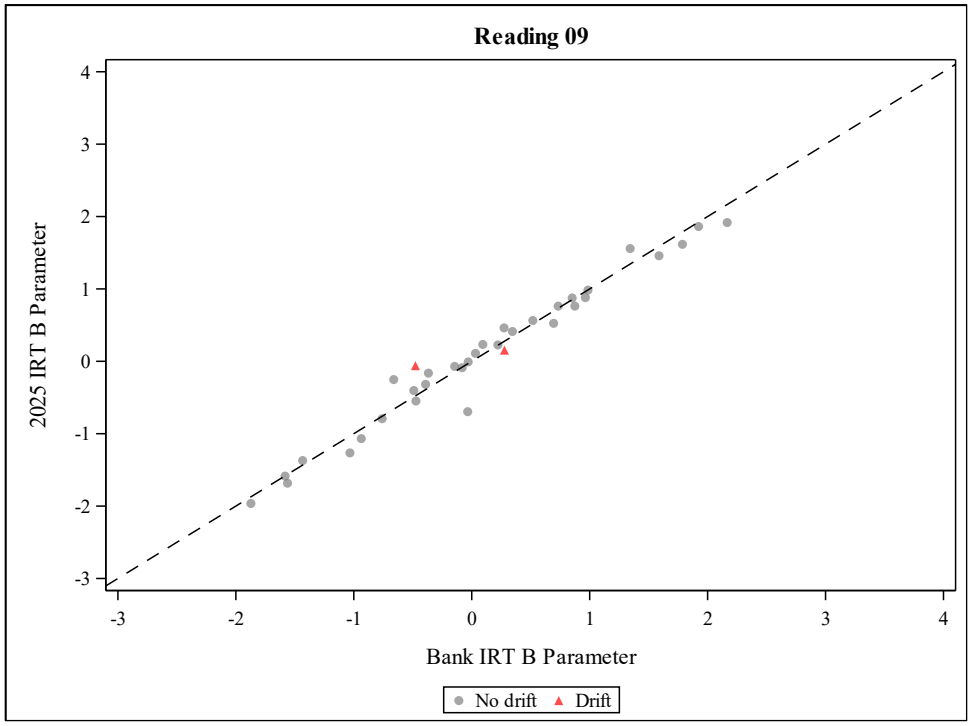


Figure E.2. IRT B Parameters for Operational Items—Reading Grade 10

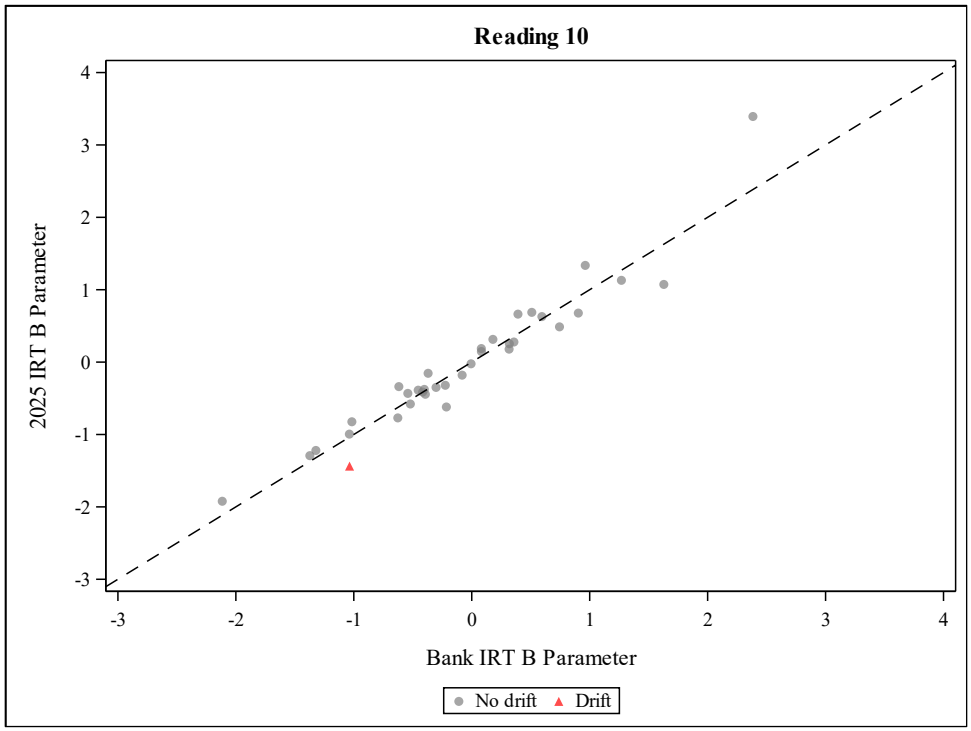


Figure E.3. IRT B Parameters for Operational Items—Mathematics Grade 9

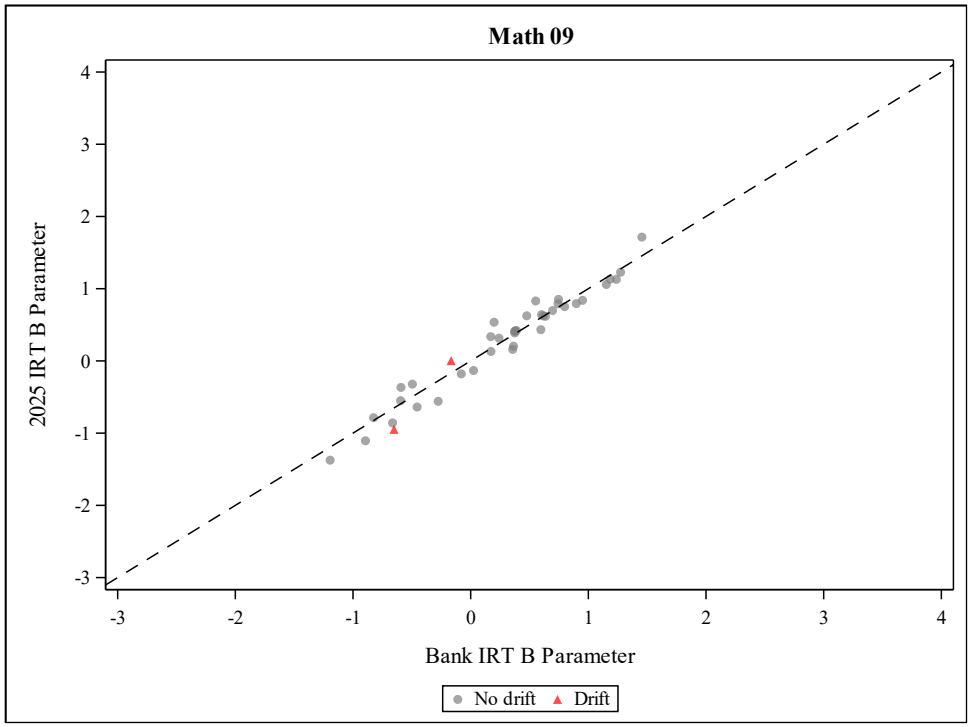


Figure E.4. IRT B Parameters for Operational Items—Mathematics Grade 10

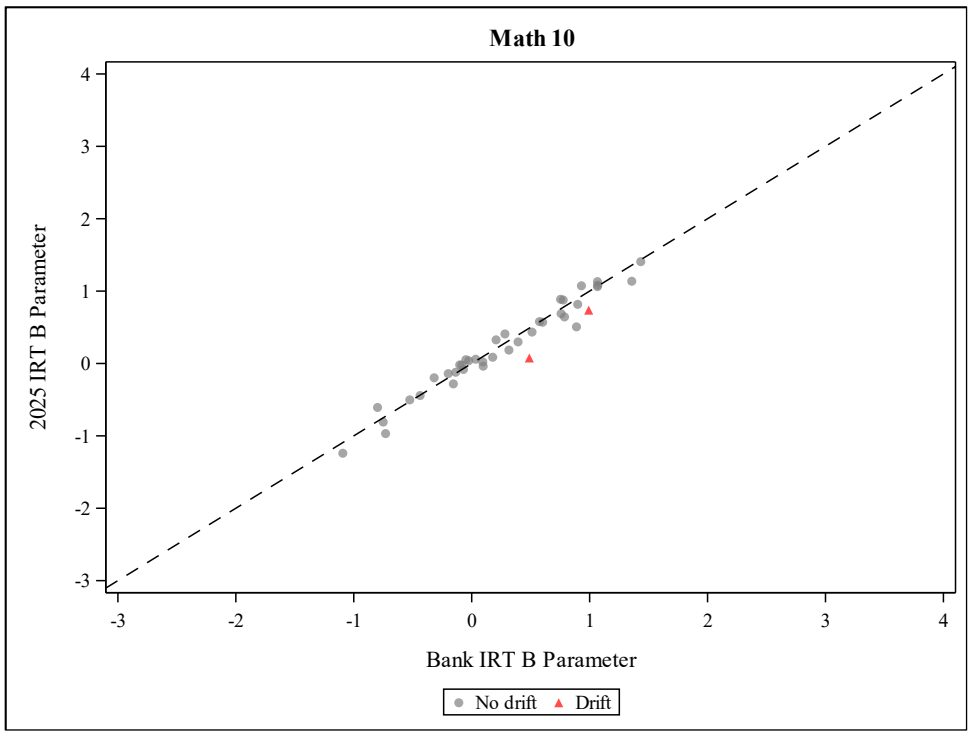


Figure E.5. IRT B Parameters for Operational Items—Science Grade 9

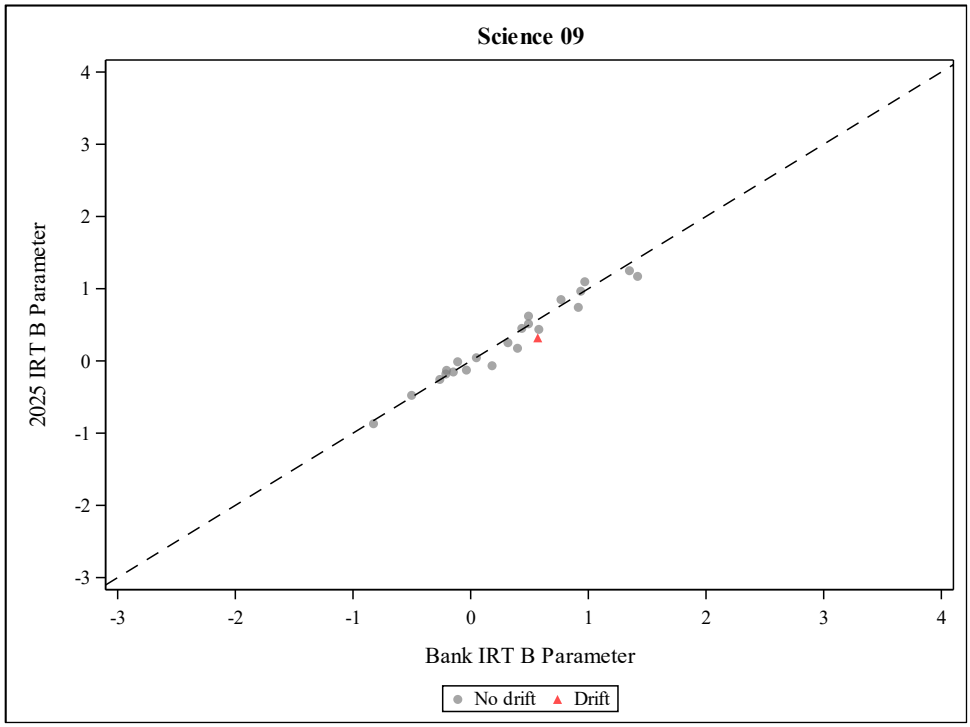
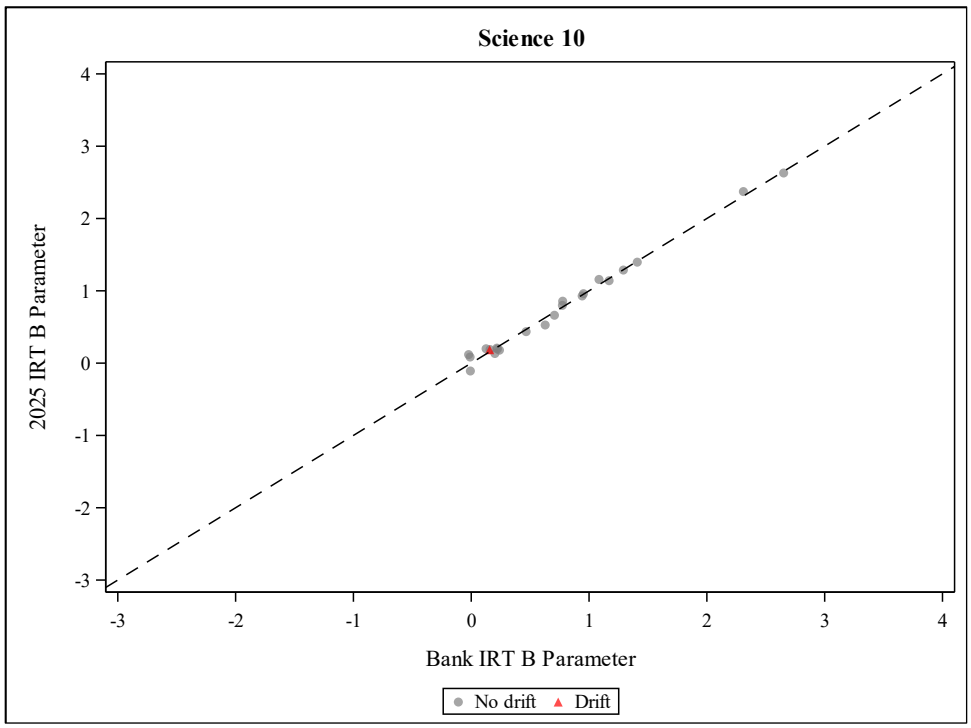


Figure E.6. IRT B Parameters for Operational Items—Science Grade 10



Appendix F: Scale Score Descriptive Statistics by Subgroup

Table F.1. Scale Score Descriptive Statistics by Subgroup—Reading Grade 9

Subgroup	#Students	Mean	SD	P25	Median	P75	Skew
Total #Students Scored	45,079	198	28.18	180	199	217	-0.13
Female	21,840	201	27.17	183	201	219	-0.04
Male	23,214	196	28.83	177	197	215	-0.18
Hispanic or Latino Ethnicity	9,154	185	27.19	167	184	203	-0.03
Asian	794	203	29.29	185	204	222	-0.29
Native Hawaiian or Other Pacific Islander	645	183	25.42	166	182	199	0.08
Black or African American	607	182	27.45	164	183	201	-0.14
American Indian or Alaska Native	418	182	23.77	165	180	199	0.08
White	31,846	203	26.99	186	203	220	-0.15
Other	1,615	200	27.93	182	201	218	-0.14
Limited English Proficient – No	41,510	201	27.08	184	201	219	-0.12
Limited English Proficient – Yes	3,569	168	22.50	156	169	182	-0.34
Economic Disadvantaged – No	33,448	203	27.24	186	203	220	-0.16
Economic Disadvantaged – Yes	11,631	186	27.17	168	185	204	-0.03
Special Education – No	40,616	201	27.29	184	202	219	-0.16
Special Education – Yes	4,463	173	23.51	159	173	187	0.02

Table F.2. Scale Score Descriptive Statistics by Subgroup—Reading Grade 10

Subgroup	#Students	Mean	SD	P25	Median	P75	Skew
Total #Students Scored	43,615	202	27.09	184	203	219	0.15
Female	20,803	204	26.34	188	205	220	0.20
Male	22,787	200	27.64	181	202	218	0.13
Hispanic or Latino Ethnicity	8,939	189	24.67	170	189	206	0.25
Asian	776	205	27.56	188	205	221	0.19
Native Hawaiian or Other Pacific Islander	574	189	22.93	172	188	205	0.27
Black or African American	584	187	24.35	168	185.5	203.5	0.45
American Indian or Alaska Native	400	188	23.42	170.5	188	202	0.46
White	30,829	207	26.41	191	208	222	0.11
Other	1,513	203	26.14	187	204	219	-0.02
Limited English Proficient – No	40,206	205	26.27	189	206	220	0.14
Limited English Proficient – Yes	3,409	173	17.81	161	172	184	0.14
Economic Disadvantaged – No	32,859	206	26.65	190	207	222	0.10
Economic Disadvantaged – Yes	10,756	192	25.62	172	192	208	0.33
Special Education – No	39,515	205	26.30	189	206	220	0.14
Special Education – Yes	4,100	178	22.33	163	176	191	0.59

Table F.3. Scale Score Descriptive Statistics by Subgroup—Mathematics Grade 9

Subgroup	#Students	Mean	SD	P25	Median	P75	Skew
Total #Students Scored	43,894	194	32.20	177	198	215	-0.80
Female	21,132	193	29.81	178	198	213	-0.92
Male	22,739	194	34.27	176	199	218	-0.72
Hispanic or Latino Ethnicity	8,818	176	32.64	159	180	199	-0.61
Asian	769	200	32.13	184	202	221	-0.62
Native Hawaiian or Other Pacific Islander	613	178	30.43	165	182	199	-0.86
Black or African American	588	171	34.13	151	175	195	-0.48
American Indian or Alaska Native	402	173	31.50	157	175.5	195	-0.54
White	31,119	200	29.75	185	203	219	-0.90
Other	1,585	193	32.64	178	198	215	-0.83
Limited English Proficient – No	40,400	197	30.74	181	201	217	-0.86
Limited English Proficient – Yes	3,494	161	30.22	146	165	182	-0.53
Economic Disadvantaged – No	32,629	199	29.88	184	203	218	-0.88
Economic Disadvantaged – Yes	11,265	178	33.66	160	182	202	-0.56
Special Education – No	39,489	197	30.13	182	201	217	-0.84
Special Education – Yes	4,405	162	32.93	145	166	184	-0.33

Table F.4. Scale Score Descriptive Statistics by Subgroup—Mathematics Grade 10

Subgroup	#Students	Mean	SD	P25	Median	P75	Skew
Total #Students Scored	42,439	189	34.23	174	193	212	-0.80
Female	20,132	190	30.89	175	193	210	-0.91
Male	22,285	189	36.98	172	193	213	-0.72
Hispanic or Latino Ethnicity	8,589	172	34.23	160	177	193	-0.74
Asian	759	198	36.28	180	201	221	-0.65
Native Hawaiian or Other Pacific Islander	576	172	34.62	163	178	194	-0.92
Black or African American	547	167	37.15	153	175	192	-0.65
American Indian or Alaska Native	388	169	34.27	159	174.5	191.5	-0.79
White	30,111	195	32.04	179	198	215	-0.87
Other	1,469	190	33.74	174	192	211	-0.72
Limited English Proficient – No	39,139	192	32.75	177	195	213	-0.83
Limited English Proficient – Yes	3,300	156	33.59	143	166	178	-0.63
Economic Disadvantaged – No	32,092	194	32.61	178	197	215	-0.85
Economic Disadvantaged – Yes	10,347	175	35.03	163	179	197	-0.71
Special Education – No	38,411	193	32.51	177	196	213	-0.85
Special Education – Yes	4,028	158	33.87	146	166	178	-0.53

Table F.5. Scale Score Descriptive Statistics by Subgroup—Science Grade 9

Subgroup	#Students	Mean	SD	P25	Median	P75	Skew
Total #Students Scored	45,006	206	32.80	185	208	227	-0.29
Female	21,792	205	30.86	187	207	225	-0.37
Male	23,189	206	34.52	183	209	229	-0.24
Hispanic or Latino Ethnicity	9,191	189	31.21	170	189	209	-0.25
Asian	797	211	35.12	191	212	231	-0.30
Native Hawaiian or Other Pacific Islander	644	187	28.96	171	189	206	-0.48
Black or African American	618	185	29.53	168	184	206	-0.17
American Indian or Alaska Native	419	185	30.74	169	189	203	-0.66
White	31,712	211	31.31	193	213	231	-0.32
Other	1,625	206	31.41	186	208	226	-0.29
Limited English Proficient – No	41,390	209	31.72	189	210	228	-0.30
Limited English Proficient – Yes	3,616	173	26.50	160	175	189	-0.50
Economic Disadvantaged – No	33,302	211	31.41	192	213	230	-0.31
Economic Disadvantaged – Yes	11,704	191	32.42	172	191	213	-0.19
Special Education – No	40,531	209	31.60	190	211	229	-0.29
Special Education – Yes	4,475	177	29.63	162	177	194	-0.20

Table F.6. Scale Score Descriptive Statistics by Subgroup—Science Grade 10

Subgroup	#Students	Mean	SD	P25	Median	P75	Skew
Total #Students Scored	43,308	195	33.79	176	198	217	-0.47
Female	20,635	195	32.16	177	199	216	-0.66
Male	22,648	195	35.22	174	198	218	-0.33
Hispanic or Latino Ethnicity	8,896	180	32.41	162	183	202	-0.45
Asian	769	202	31.57	182	204	223	-0.21
Native Hawaiian or Other Pacific Islander	567	178	29.99	163	183	197	-0.73
Black or African American	564	176	31.16	159	179	199	-0.50
American Indian or Alaska Native	390	179	28.12	164	182	199	-0.59
White	30,618	200	32.88	182	203	221	-0.54
Other	1,504	196	34.07	178	199.5	217	-0.39
Limited English Proficient – No	39,895	197	33.07	179	201	219	-0.51
Limited English Proficient – Yes	3,413	166	28.47	150	171	187	-0.62
Economic Disadvantaged – No	32,619	199	33.23	180	202	220	-0.52
Economic Disadvantaged – Yes	10,689	183	32.92	165	186	205	-0.40
Special Education – No	39,246	197	33.01	179	201	219	-0.51
Special Education – Yes	4,062	170	30.92	152	173	190	-0.36

Appendix G: Scale Score Distributions for Overall Testing Population

Figure G.1. Scale Score Distribution—Reading Grade 9

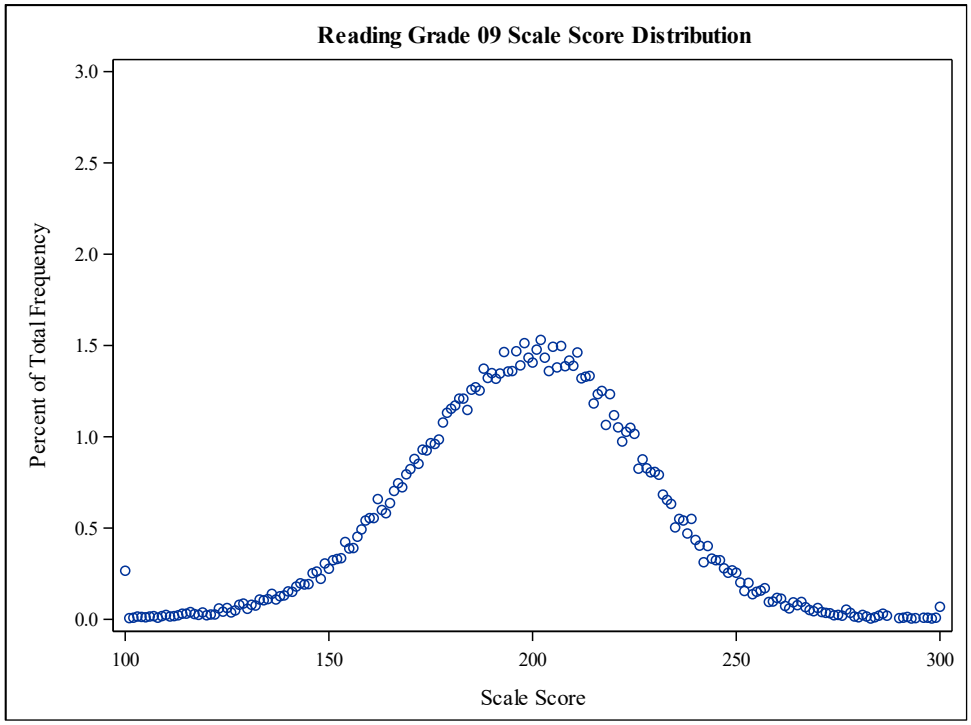


Figure G.2. Scale Score Distribution—Reading Grade 10

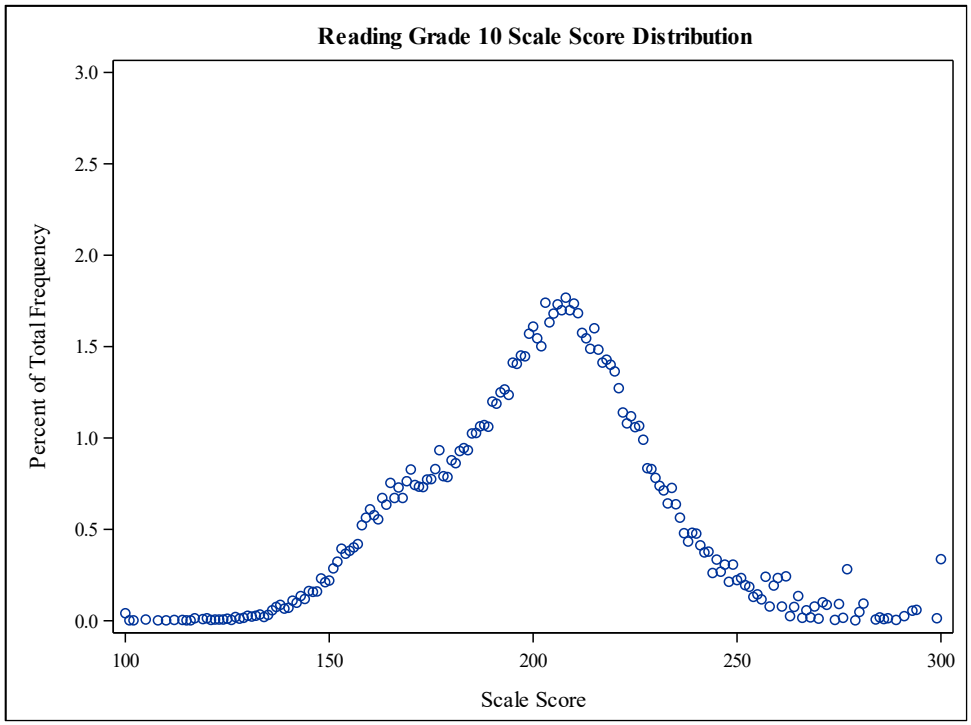


Figure G.3. Scale Score Distribution—Mathematics Grade 9

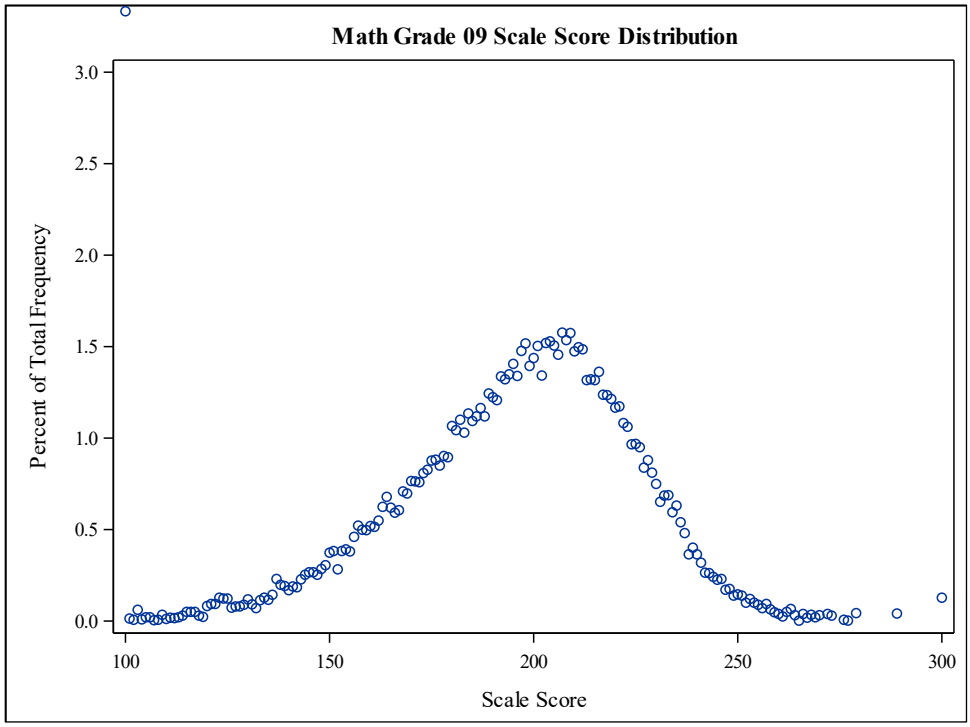


Figure G.4. Scale Score Distribution—Mathematics Grade 10

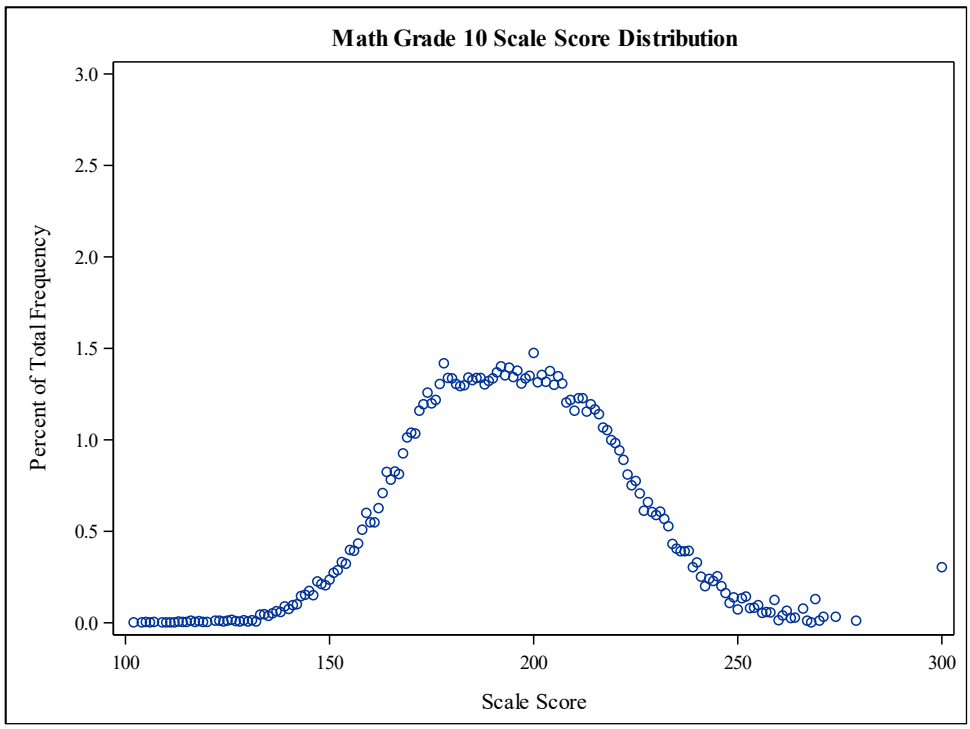


Figure G.5. Scale Score Distribution—Science Grade 9

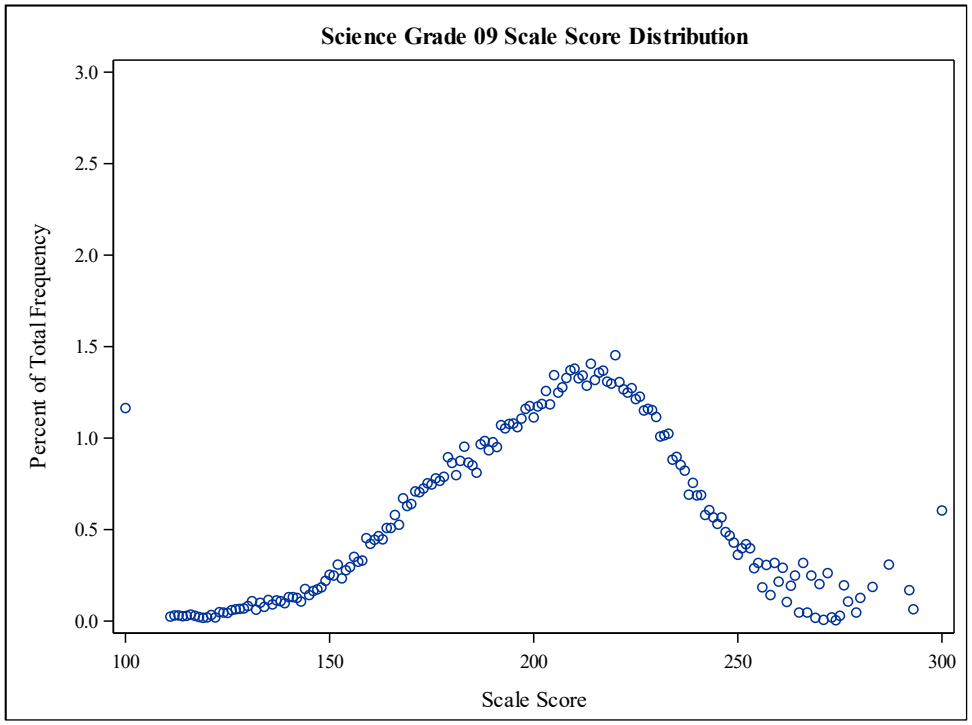
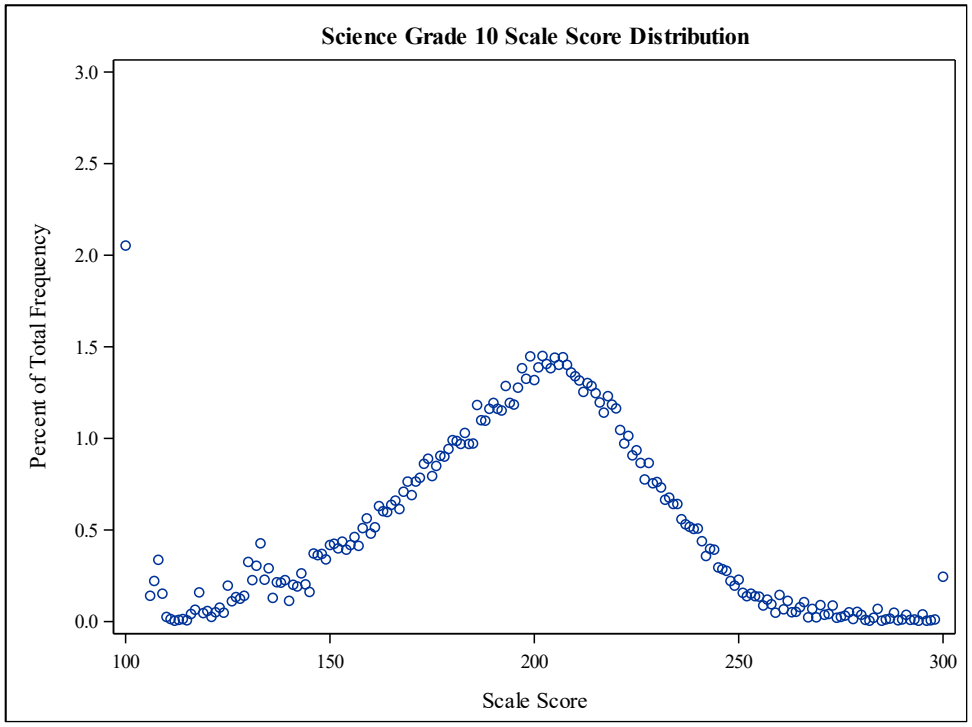


Figure G.6. Scale Score Distribution—Science Grade 10



Appendix H: Performance Level Distributions by Subgroup

Table H.1. Performance Level Distribution by Subgroup—Reading Grade 9

Subgroup	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Total #Students Scored	45,079	11.8	44.7	31.7	11.8
Female	21,840	9.0	44.3	33.5	13.2
Male	23,214	14.5	45.1	30.0	10.3
Hispanic or Latino Ethnicity	9,154	23.2	52.2	20.1	4.5
Asian	794	10.6	38.5	34.3	16.6
Native Hawaiian or Other Pacific Islander	645	24.3	55.7	16.9	3.1
Black or African American	607	26.9	51.9	17.3	4.0
American Indian or Alaska Native	418	25.4	55.7	16.5	2.4
White	31,846	7.9	42.3	35.7	14.1
Other	1,615	10.7	43.4	32.4	13.5
Limited English Proficient – No	41,510	9.1	44.1	34.0	12.7
Limited English Proficient – Yes	3,569	43.4	51.5	4.8	0.3
Economic Disadvantaged – No	33,448	8.4	42.0	35.5	14.1
Economic Disadvantaged – Yes	11,631	21.7	52.5	20.8	4.9
Special Education – No	40,616	9.3	43.5	34.3	12.9
Special Education – Yes	4,463	35.2	55.4	8.0	1.3

Table H.2. Performance Level Distribution by Subgroup—Reading Grade 10

Subgroup	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Total #Students Scored	43,615	16.4	33.7	39.9	9.9
Female	20,803	13.5	34.2	41.9	10.5
Male	22,787	19.2	33.3	38.1	9.4
Hispanic or Latino Ethnicity	8,939	30.6	41.2	24.9	3.3
Asian	776	13.8	33.9	40.1	12.2
Native Hawaiian or Other Pacific Islander	574	28.2	45.6	23.7	2.4
Black or African American	584	35.1	39.9	22.1	2.9
American Indian or Alaska Native	400	27.5	49.3	20.5	2.8
White	30,829	11.7	31.0	45.1	12.2
Other	1,513	15.2	34.4	41.5	8.9
Limited English Proficient – No	40,206	13.1	33.3	42.9	10.7
Limited English Proficient – Yes	3,409	56.3	38.6	5.1	0.1
Economic Disadvantaged – No	32,859	12.8	31.7	43.9	11.7
Economic Disadvantaged – Yes	10,756	27.6	40.1	27.8	4.5
Special Education – No	39,515	13.2	33.2	42.9	10.8
Special Education – Yes	4,100	48.0	39.1	11.5	1.4

Table H.3. Performance Level Distribution by Subgroup—Mathematics Grade 9

Subgroup	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Total #Students Scored	43,894	20.4	40.3	31.6	7.7
Female	21,132	18.8	43.5	32.4	5.3
Male	22,739	21.8	37.4	30.9	9.9
Hispanic or Latino Ethnicity	8,818	39.0	43.6	15.4	2.0
Asian	769	15.9	39.5	31.7	12.9
Native Hawaiian or Other Pacific Islander	613	33.4	49.8	16.0	0.8
Black or African American	588	46.8	38.6	12.9	1.7
American Indian or Alaska Native	402	42.0	44.0	12.7	1.2
White	31,119	14.2	39.2	37.1	9.6
Other	1,585	20.2	40.8	31.8	7.3
Limited English Proficient – No	40,400	16.9	40.8	34.0	8.4
Limited English Proficient – Yes	3,494	60.4	35.3	4.2	0.1
Economic Disadvantaged – No	32,629	14.8	39.6	36.2	9.4
Economic Disadvantaged – Yes	11,265	36.5	42.5	18.3	2.7
Special Education – No	39,489	16.3	40.9	34.4	8.5
Special Education – Yes	4,405	57.3	35.4	6.6	0.7

Table H.4. Performance Level Distribution by Subgroup—Mathematics Grade 10

Subgroup	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Total #Students Scored	42,439	33.8	38.7	22.3	5.3
Female	20,132	32.3	42.3	21.8	3.7
Male	22,285	35.2	35.4	22.7	6.7
Hispanic or Latino Ethnicity	8,589	56.9	32.8	9.2	1.2
Asian	759	25.4	34.3	27.9	12.4
Native Hawaiian or Other Pacific Islander	576	53.5	37.2	8.5	0.9
Black or African American	547	60.0	30.0	9.5	0.6
American Indian or Alaska Native	388	59.3	32.7	7.7	0.3
White	30,111	26.2	40.7	26.6	6.5
Other	1,469	34.5	39.6	20.2	5.8
Limited English Proficient – No	39,139	29.9	40.4	24.0	5.7
Limited English Proficient – Yes	3,300	79.6	18.5	1.8	0.2
Economic Disadvantaged – No	32,092	27.7	40.1	25.9	6.4
Economic Disadvantaged – Yes	10,347	52.9	34.3	11.0	1.8
Special Education – No	38,411	29.1	40.8	24.3	5.8
Special Education – Yes	4,028	78.3	18.6	2.5	0.7

Table H.5. Performance Level Distribution by Subgroup—Science Grade 9

Subgroup	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Total #Students Scored	45,006	26.2	27.5	31.3	15.1
Female	21,792	24.9	30.1	32.2	12.9
Male	23,189	27.5	25.0	30.4	17.1
Hispanic or Latino Ethnicity	9,191	46.3	29.9	18.3	5.5
Asian	797	21.3	26.5	32.8	19.5
Native Hawaiian or Other Pacific Islander	644	46.7	32.3	18.3	2.6
Black or African American	618	53.9	25.9	17.5	2.8
American Indian or Alaska Native	419	45.8	37.2	14.8	2.2
White	31,712	19.4	26.5	35.7	18.5
Other	1,625	25.1	28.3	32.8	13.9
Limited English Proficient – No	41,390	22.3	27.8	33.5	16.4
Limited English Proficient – Yes	3,616	70.9	23.2	5.6	0.4
Economic Disadvantaged – No	33,302	20.0	26.9	35.1	18.0
Economic Disadvantaged – Yes	11,704	43.7	29.1	20.3	6.8
Special Education – No	40,531	22.0	27.7	33.8	16.5
Special Education – Yes	4,475	64.4	25.0	8.5	2.2

Table H.6. Performance Level Distribution by Subgroup—Science Grade 10

Subgroup	#Students	*Below Proficient*	*Approaching Proficient*	*Proficient*	*Highly Proficient*
Total #Students Scored	43,308	35.6	30.0	28.0	6.5
Female	20,635	33.8	32.0	29.3	4.9
Male	22,648	37.2	28.1	26.9	7.9
Hispanic or Latino Ethnicity	8,896	54.1	29.3	14.9	1.8
Asian	769	29.0	29.0	32.9	9.1
Native Hawaiian or Other Pacific Islander	567	57.0	31.2	11.3	0.5
Black or African American	564	61.2	25.5	12.6	0.7
American Indian or Alaska Native	390	57.4	30.0	12.6	0.0
White	30,618	29.3	30.2	32.6	8.0
Other	1,504	33.7	31.7	27.4	7.2
Limited English Proficient – No	39,895	32.2	30.6	30.2	7.0
Limited English Proficient – Yes	3,413	74.8	22.3	2.9	0.0
Economic Disadvantaged – No	32,619	30.6	30.2	31.4	7.7
Economic Disadvantaged – Yes	10,689	50.5	29.1	17.6	2.7
Special Education – No	39,246	32.0	30.8	30.2	7.1
Special Education – Yes	4,062	70.1	22.2	7.0	0.8

Appendix I: Principal Components Scree Plots

Figure I.1. Principal Component Scree Plot—Reading Grade 9

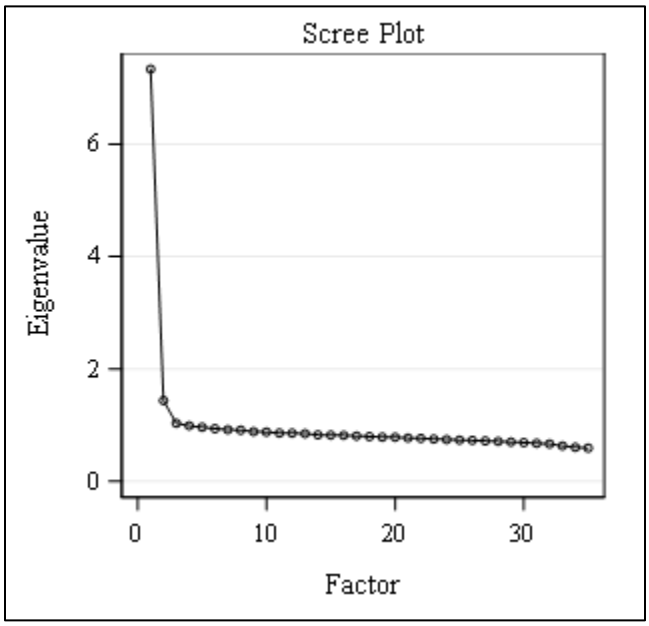


Figure I.2. Principal Component Scree Plot—Reading Grade 10

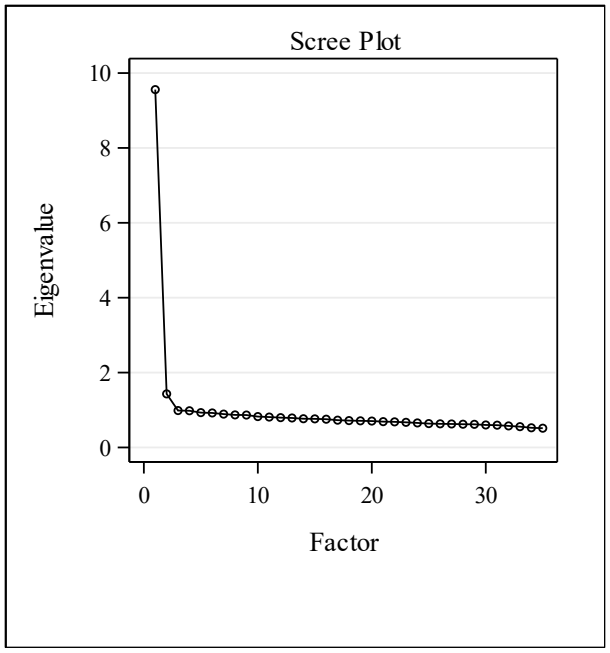


Figure I.3. Principal Component Scree Plot—Mathematics Grade 9

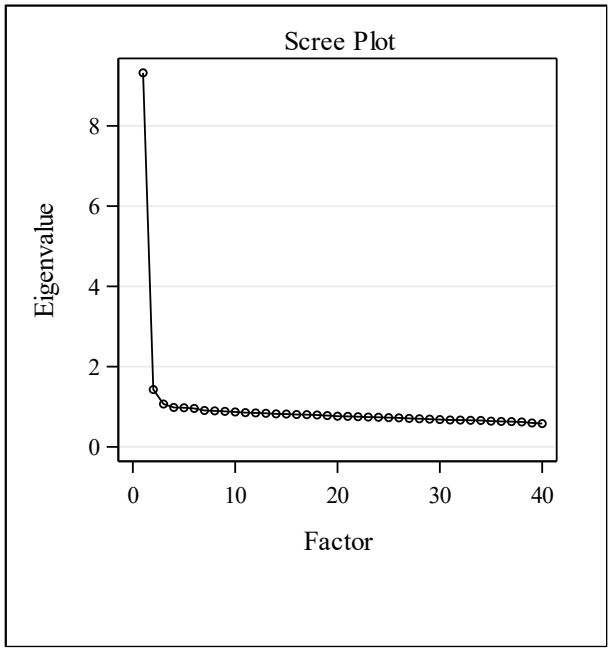


Figure I.4. Principal Component Scree Plot—Mathematics Grade 10

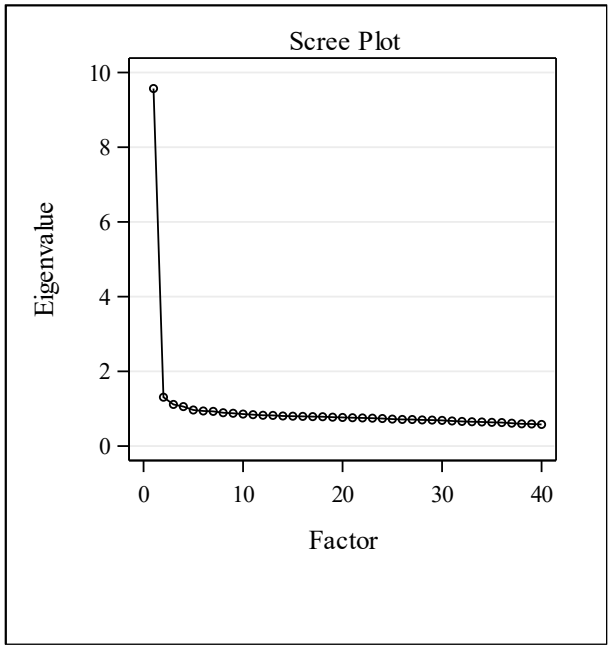


Figure I.5. Principal Component Scree Plot—Science Grade 9

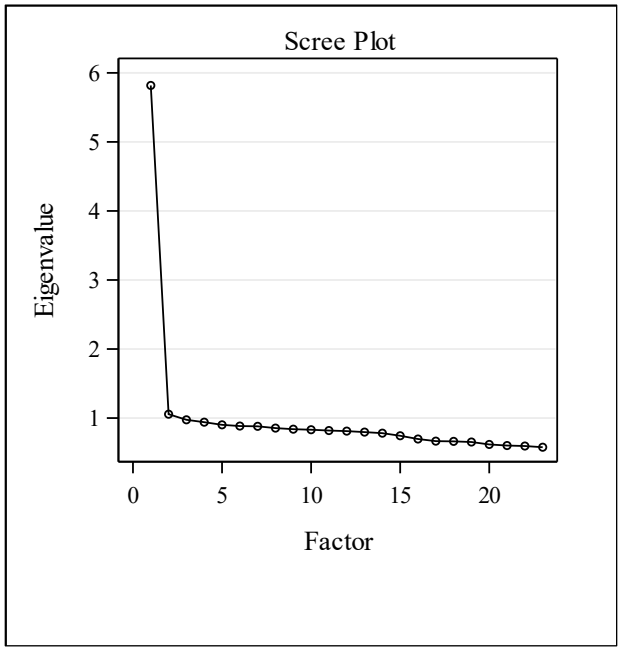
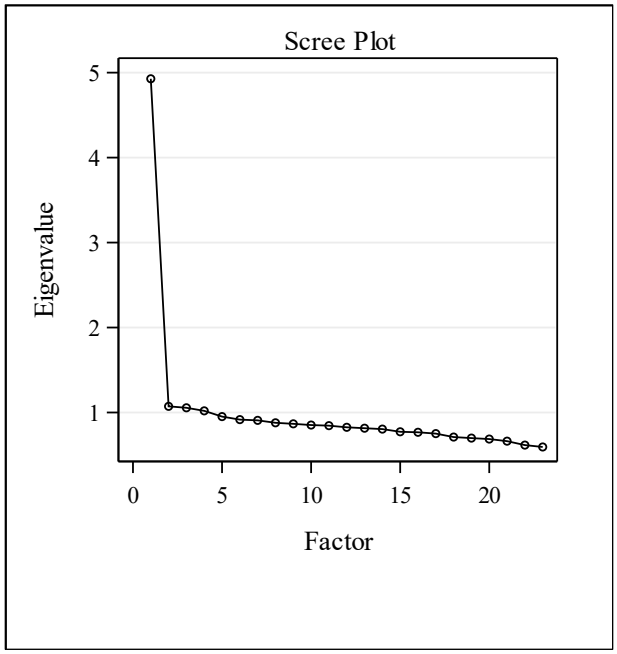


Figure I.6. Principal Component Scree Plot—Science Grade 10



Appendix J: Subscore Correlations

Table J.1. Correlations of Total Score and Subscores

Assessment	Subdomain	Total	Subdomain 1	Subdomain 2	Subdomain 3	Subdomain 4	Subdomain 5
Reading 9	Total	1.00	0.91	0.88	0.75	–	–
	Key Ideas	0.91	1.00	0.74	0.65	–	–
	Craft and Structure	0.88	0.74	1.00	0.61	–	–
	Integration of Knowledge and Ideas	0.75	0.65	0.61	1.00	–	–
Reading 10	Total	1.00	0.89	0.90	0.77	–	–
	Key Ideas	0.89	1.00	0.80	0.64	–	–
	Craft and Structure	0.90	0.80	1.00	0.65	–	–
	Integration of Knowledge and Ideas	0.77	0.64	0.65	1.00	–	–
Mathematics 9	Total	1.00	0.83	0.82	0.83	0.75	–
	Algebra	0.83	1.00	0.76	0.73	0.67	–
	Functions	0.82	0.76	1.00	0.71	0.66	–
	Geometry	0.83	0.73	0.71	1.00	0.65	–
	Statistics and Probability	0.75	0.67	0.66	0.65	1.00	–
Mathematics 10	Total	1.00	0.63	0.83	0.72	0.82	0.66
	Number and Quantity	0.63	1.00	0.61	0.58	0.61	0.46
	Algebra	0.83	0.61	1.00	0.68	0.75	0.57
	Functions	0.72	0.58	0.68	1.00	0.70	0.53
	Geometry	0.82	0.61	0.75	0.70	1.00	0.59
	Statistics and Probability	0.66	0.46	0.57	0.53	0.59	1.00
Science 9	Total	1.00	0.80	0.78	0.81	0.81	–
	Gathering & Investigating	0.80	1.00	0.59	0.62	0.60	–
	Developing Models	0.78	0.59	1.00	0.58	0.57	–
	Using Mathematical Thinking	0.81	0.62	0.58	1.00	0.61	–
	Construct Explanation	0.81	0.60	0.57	0.61	1.00	–
Science 10	Total	1.00	0.81	0.73	0.68	0.63	–
	Gathering & Investigating	0.81	1.00	0.52	0.52	0.57	–
	Developing Models	0.73	0.52	1.00	0.43	0.44	–
	Using Mathematical Thinking	0.68	0.52	0.43	1.00	0.43	–
	Construct Explanation	0.63	0.57	0.44	0.43	1.00	–

Appendix K: Item Drift Plots

Figure K.1. Item Drift Plot—Reading Grade 9

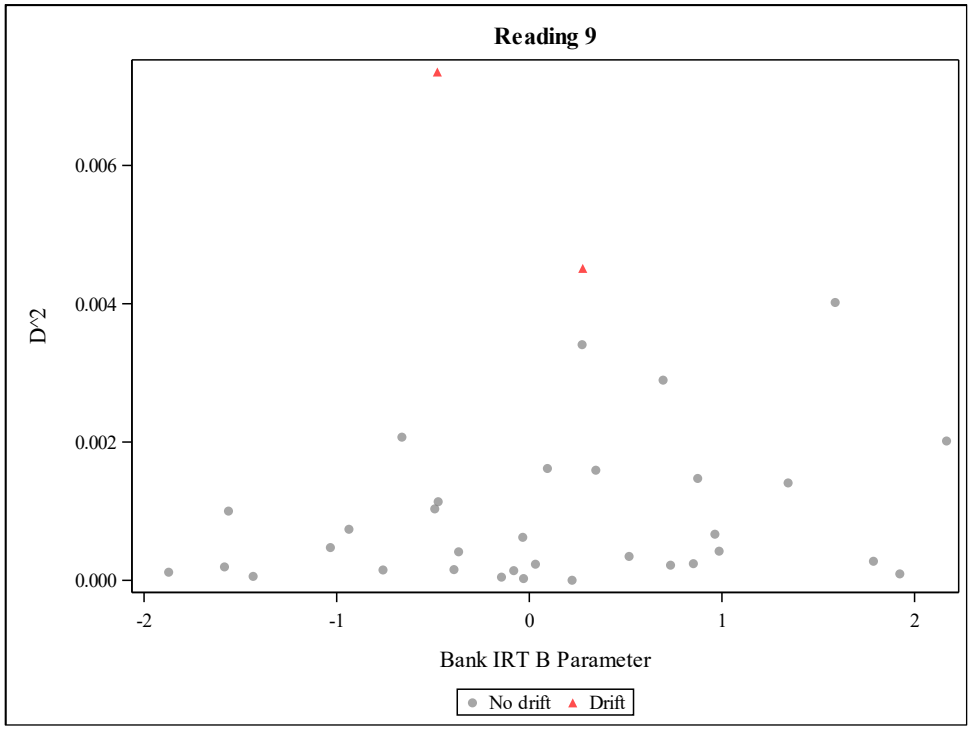


Figure K.2. Item Drift Plot—Reading Grade 10

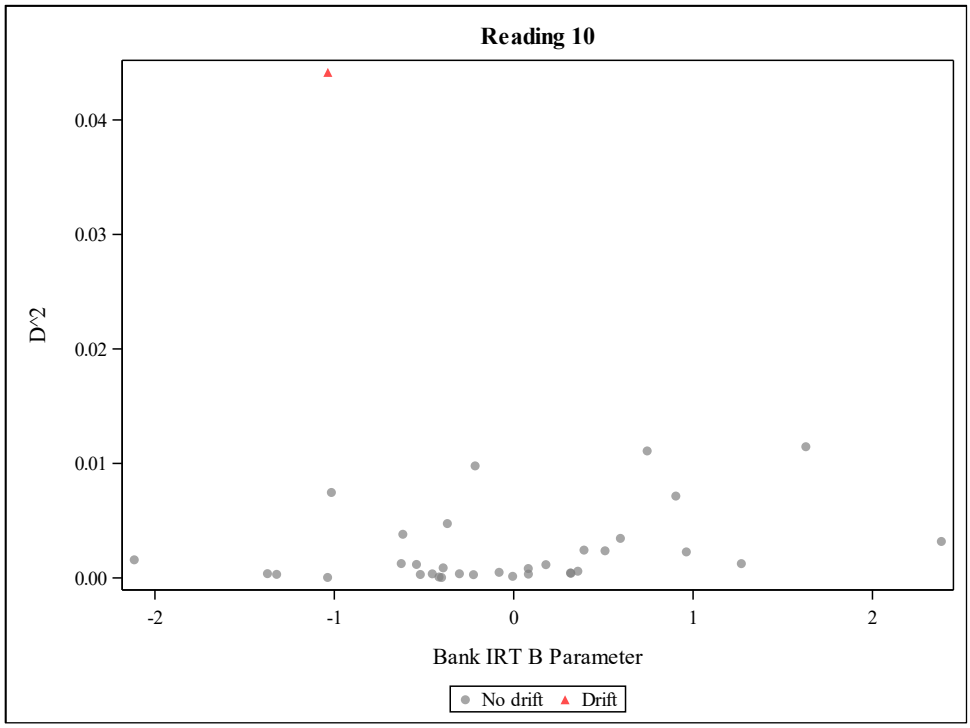


Figure K.3. Item Drift Plot—Mathematics Grade 9

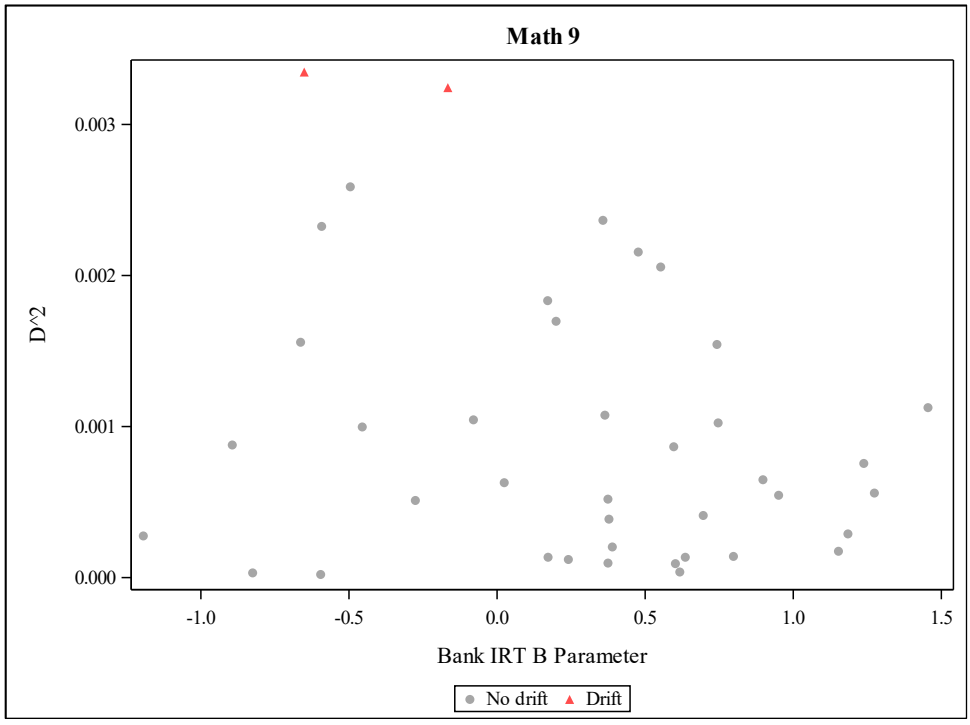


Figure K.4. Item Drift Plot—Mathematics Grade 10

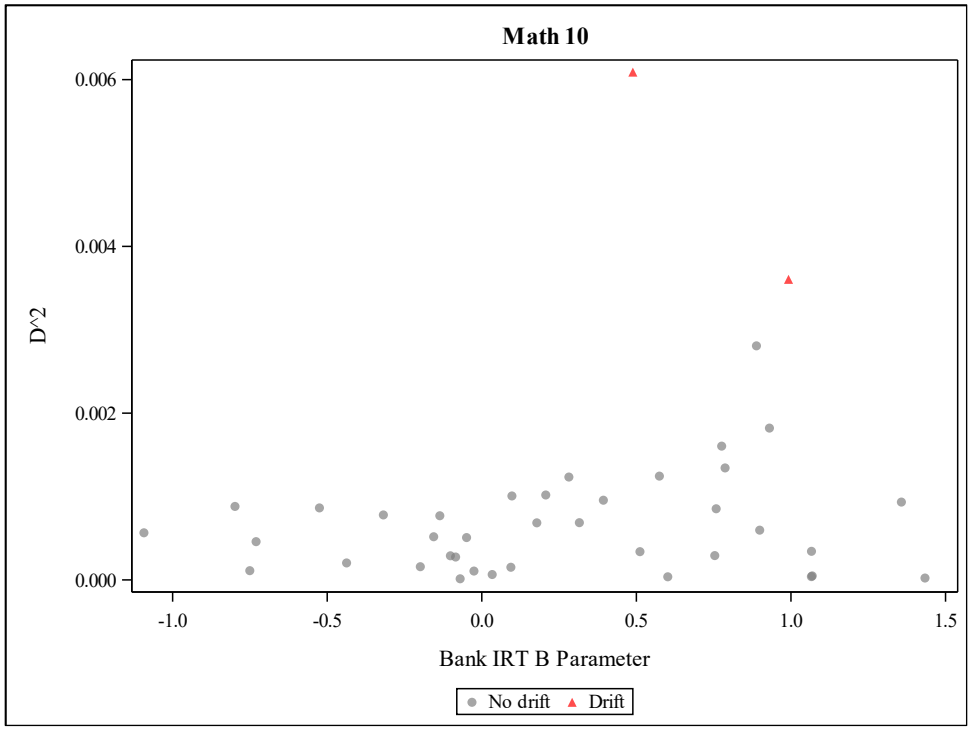


Figure K.5. Item Drift Plot—Science Grade 9

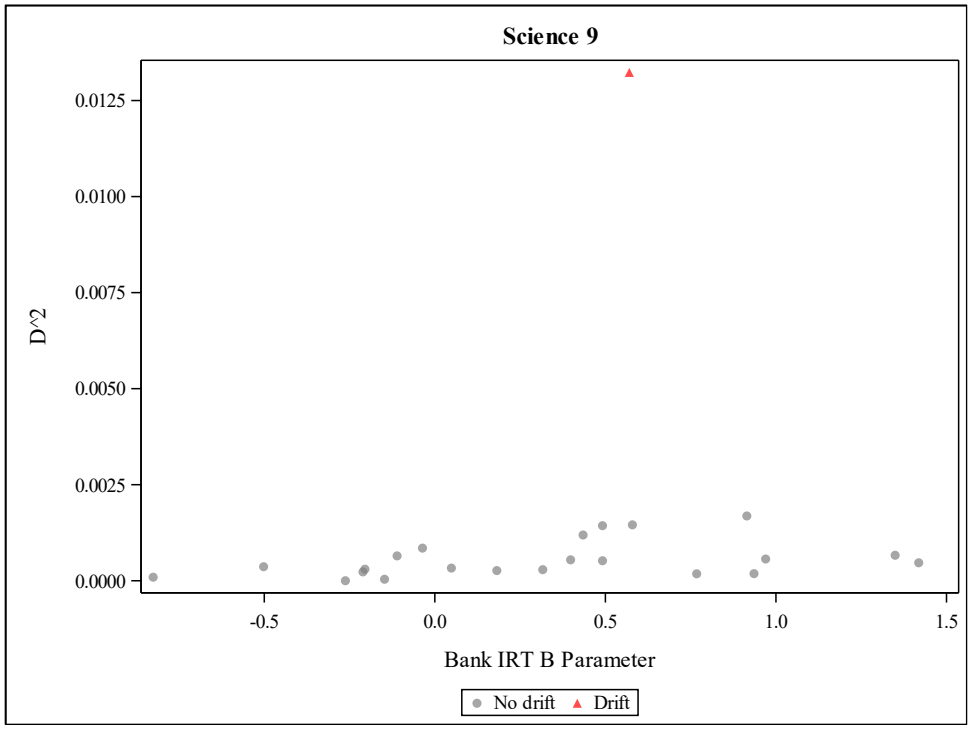


Figure K.6. Item Drift Plot—Science Grade 10