Utah Aspire Plus
2023–2024 Technical Report

2024

Table of Contents

1 Introduction 9

1.1 Background 9

1.2 Purpose of the Operational Tests 10

1.3 Prior Administrations 10

1.4 Spring 2024 Administration 11

1.5 Composition of the Operational Tests 11

1.6 Intended Population of the Operational Tests 12

1.7 Overview of the Technical Report 13

2 Test Development 14

2.1 Overview of the Utah Aspire Plus Assessments, Claims, and Blueprints 14

2.1.1 English Assessment Claims 15

2.1.2 Reading Assessment Claims 16

2.1.3 Mathematics Assessment Claims 17

2.1.4 Science Assessment Claims 18

2.2 Utah Aspire Plus Blueprints 19

2.3 Test Development Activities 20

2.3.1 Operational Forms Development 21

2.3.2 Statistical Guidelines 22

2.3.3 2024 Match to Test Blueprint 23

3 Operational Administration 27

3.1 Testing Window 27

3.2 Test Administration and Security Policies 27

3.2.1 Online Administration and Monitoring 28

3.3 Test Accommodations and Supports 29

3.4 Test Taking Irregularities and Security Breaches 31

3.4.1 Test Interruptions 32

3.4.2 Scoring of Interrupted Tests 32

3.4.3 Wrong Test Form/Accommodation 32

3.4.4 Extended Time Accommodation Issues 32

3.4.5 Test Invalidation 32

3.5 Test Taker Characteristics 33

3.6 Testing Time 34

4 Score Reporting 37

4.1 IRT Pattern Scoring 37

4.1.1 Quality Control of IRT Scoring 38

4.2 Appropriate Uses for Scores and Reports 38

4.3 Utah Aspire Plus Reporting Scale 39

4.4 Standard Setting 40

4.5 ACT Predicted Score Ranges 41
4.6 2023–2024 Utah Aspire Plus Performance Results 42
5 Classical Item Analyses 43
5.1 Item Analyses 43
5.1.1 *p*-Value and Item Mean Scores 43
5.1.2 Item-Test Score Correlations 43
5.1.3 Differential Item Functioning 44
5.2 Classical Item Summaries for Operational Administration 45
6 Reliability 46
6.1 Classical Definition of Reliability 46
6.2 Classical Test Theory Reliability Estimates 47
6.2.1 Cronbach’s Alpha 47
6.2.2 Standard Error of Measurement 47
6.3 IRT-Based Reliability 48
6.4 Reliability of Performance Level Categorization 48
6.4.1 Accuracy and Consistency 49
6.4.2 Calculating Accuracy 49
6.4.3 Calculating Consistency 50
6.4.4 Calculating Kappa 51
7 Field Test Calibration and Drift Analyses 52
7.1 IRT Overview 52
7.2 IRT Data Preparation 53
7.2.1 Student Inclusion/Exclusion Rules 53
7.2.2 Quality Control of the IRT Data Matrix Files 53
7.3 Description of the Calibration, Equating, and Scaling Process 54
7.3.1 IRT Models 54
7.3.2 IRTPRO Calibration Procedures and Convergence Criteria 55
7.3.3 Calibration Quality Control 55
7.3.4 Equating 56
7.3.5 Drift Analysis 57
7.4 Model Fit Evaluation Criteria 59
8 Quality Control 61
8.1 Online Assessment Delivery 61
8.1.1 Item Validation 61
8.1.2 Test Administration 62
8.1.3 Remote Proctoring 63
8.1.4 Operational Monitoring 64
8.2 Production System Testing 64
8.2.1 Functional Testing 64
8.2.2 Integration Testing 65

8.2.3 Program Validation End-to-End Testing 65
8.2.4 Load Testing 65
8.2.5 Performance Monitoring 66
8.2.6 Regression Testing 66
8.2.7 User Acceptance Testing 66
8.3 Reporting 67
8.4 Quality Control of Psychometric Processes 67
9 Validity 68
9.1 Evidence Based on Test Content 69
9.2 Evidence Based on Cognitive Process 71
9.3 Evidence Based on Internal Structure 72
9.3.1 Reliability 74
9.4 Evidence Based on Different Student Populations 75
9.5 Summary 75
10 References 76
Appendix A : Test-Level Reporting Categories and Standards by Item Type and DOK 78
Appendix B : Student Testing Time 84
Appendix C : Item Statistics Summaries 86
Appendix D : Reliability and Standard Error by Subgroup 88
Appendix E : Conditional Standard Error of Scale Scores 96
Appendix F : Accuracy and Consistency 100
Appendix G : Common Item Scatterplots for 2024 Anchor Items 108
Appendix H : Scale Score Descriptive Statistics by Subgroup 116
Appendix I : Scale Score Distributions for Overall Testing Population 124
Appendix J : Performance Level Distributions 128
Appendix K : Principal Components Scree Plot 136
Appendix L : Subscore Correlations 138
Appendix M : Item Drift 141

List of Tables

Table 2.1. Field Test Forms 20
Table 2.2. Utah Aspire Plus English Grade 9 Operational Test Blueprint Match 23
Table 2.3. Utah Aspire Plus English Grade 10 Operational Test Blueprint Match 23
Table 2.4. Utah Aspire Plus Reading Grade 9 Operational Test Blueprint Match 24
Table 2.5. Utah Aspire Plus Reading Grade 10 Operational Test Blueprint Match 24
Table 2.6. Utah Aspire Plus Mathematics Grade 9 Operational Test Blueprint Match 25
Table 2.7. Utah Aspire Plus Mathematics Grade 10 Operational Test Blueprint Match 25
Table 2.8. Utah Aspire Plus Science Grade 9 Operational Test Blueprint Match 26
Table 2.9. Utah Aspire Plus Science Grade 10 Operational Test Blueprint Match 26
Table 3.1. Spring 2024 Participation Rates for Utah Aspire Plus 33
Table 3.2. Student Testing Time for Spring 2024 Utah Aspire Plus: English and Reading 35
Table 3.3. Student Testing Time for Spring 2024 Utah Aspire Plus: Mathematics and Science 36
Table 4.1. IRT Summary Parameter Estimates for Utah Aspire Plus Operational Items 37
Table 4.2. Utah Aspire Plus Scale Score Cuts by Grade and Subject 40
Table 5.1. Item 2x2 Contingency Table for the kth Score Level 44
Table 6.1. Example Accuracy Classification Table 49
Table 6.2. Example Accuracy Classification Table for Proficient Cut Point 50
Table 6.3. Example Consistency Classification Table 51
Table 7.1. 2024 Final Stocking and Lord Scaling Constants 56
Table 7.2. 2024 Items Showing Drift 58
Table 9.1. Model Fit Indices for Confirmatory Factor Analyses 73
Table A.1. Test-Level Reporting Categories and Standards for English Grade 9 78
Table A.2. Test-Level Reporting Categories and Standards for English Grade 10 79
Table A.3. Test-Level Reporting Categories and Standards for Reading Grade 9 80
Table A.4. Test-Level Reporting Categories and Standards for Reading Grade 10 81
Table A.5. Test-Level Reporting Categories and Standards for Mathematics Grade 9 82
Table A.6. Test-Level Reporting Categories and Standards for Mathematics Grade 10 83
Table C.1. Item Mean for One-Point Items 86
Table C.2. Item Mean for Two-Point Items 86
Table C.3. Item Total Correlation for One-Point Items 86
Table C.4. Item Total Correlation for Two-Point Items 87
Table C.5. Differential Item Functioning 87
Table D.1. English Grade 9 Test Reliability 88
Table D.2. English Grade 10 Test Reliability 89
Table D.3. Reading Grade 9 Test Reliability 90
Table D.4. Reading Grade 10 Test Reliability 91
Table D.5. Mathematics Grade 9 Test Reliability 92
Table D.6. Mathematics Grade 10 Test Reliability 93
Table D.7. Science Grade 9 Test Reliability 94
Table D.8. Science Grade 10 Test Reliability 95

Table F.1. Accuracy Classification for English Grade 9 100
Table F.2. Accuracy Classification at Proficient Cut Point for English Grade 9 100
Table F.3. Consistency Classification for English Grade 9 100
Table F.4. Accuracy Classification for English Grade 10 101
Table F.5. Accuracy Classification at Proficient Cut Point for English Grade 10 101
Table F.6. Consistency Classification for English Grade 10 101
Table F.7. Accuracy Classification for Reading Grade 9 102
Table F.8. Accuracy Classification at Proficient Cut Point for Reading Grade 9 102
Table F.9. Consistency Classification for Reading Grade 9 102
Table F.10. Accuracy Classification for Reading Grade 10 103
Table F.11. Accuracy Classification at Proficient Cut Point for Reading Grade 10 103
Table F.12. Consistency Classification for Reading Grade 10 103
Table F.13. Accuracy Classification for Mathematics Grade 9 104
Table F.14. Accuracy Classification at Proficient Cut Point for Mathematics Grade 9 104
Table F.15. Consistency Classification for Mathematics Grade 9 104
Table F.16. Accuracy Classification for Mathematics Grade 10 105
Table F.17. Accuracy Classification at Proficient Cut Point for Mathematics Grade 10 105
Table F.18. Consistency Classification for Mathematics Grade 10 105
Table F.19. Accuracy Classification for Science Grade 9 106
Table F.20. Accuracy Classification at Proficient Cut Point for Science Grade 9 106
Table F.21. Consistency Classification for Science Grade 9 106
Table F.22. Accuracy Classification for Science Grade 10 107
Table F.23. Accuracy Classification at Proficient Cut Point for Science Grade 10 107
Table F.24. Consistency Classification for Science Grade 10 107
Table H.1. English Grade 9 Scale Score Descriptive Statistics 116
Table H.2. English Grade 10 Scale Score Descriptive Statistics 117
Table H.3. Reading Grade 9 Scale Score Descriptive Statistics 118
Table H.4. Reading Grade 10 Scale Score Descriptive Statistics 119
Table H.5. Mathematics Grade 9 Scale Score Descriptive Statistics 120
Table H.6. Mathematics Grade 10 Scale Score Descriptive Statistics 121
Table H.7. Science Grade 9 Scale Score Descriptive Statistics 122
Table H.8. Science Grade 10 Scale Score Descriptive Statistics 123
Table J.1. English Grade 9 Performance Level Distribution 128
Table J.2. English Grade 10 Performance Level Distribution 129
Table J.3. Reading Grade 9 Performance Level Distribution 130
Table J.4. Reading Grade 10 Performance Level Distribution 131
Table J.5. Mathematics Grade 9 Performance Level Distribution 132
Table J.6. Mathematics Grade 10 Performance Level Distribution 133
Table J.7. Science Grade 9 Performance Level Distribution 134
Table J.8. Science Grade 10 Performance Level Distribution 135
Table L.1. English Correlations of Total Score and Subscores 138
Table L.2. Reading Correlations of Total Score and Subscores 138
Table L.3. Mathematics Correlations of Total Score and Subscores 139

Table L.4. Science Correlations of Total Score and Subscores 140

List of Figures

Figure B.1. English Grade 9 Student Testing Time 84
Figure B.2. English Grade 10 Student Testing Time 84
Figure B.3. Reading Grade 9 Student Testing Time 84
Figure B.4. Reading Grade 10 Student Testing Time 84
Figure B.5. Mathematics Grade 9 Student Testing Time 85
Figure B.6. Mathematics Grade 10 Student Testing Time 85
Figure B.7. Science Grade 9 Student Testing Time 85
Figure B.8. Science Grade 10 Student Testing Time 85
Figure E.1. English Grade 9 Conditional Standard Error of Scale Scores 96
Figure E.2. English Grade 10 Conditional Standard Error of Scale Scores 96
Figure E.3. Reading Grade 9 Conditional Standard Error of Scale Scores 97
Figure E.4. Reading Grade 10 Conditional Standard Error of Scale Scores 97
Figure E.5. Mathematics Grade 9 Conditional Standard Error of Scale Scores 98
Figure E.6. Mathematics Grade 10 Conditional Standard Error of Scale Scores 98
Figure E.7. Science Grade 9 Conditional Standard Error of Scale Scores 99
Figure E.8. Science Grade 10 Conditional Standard Error of Scale Scores 99
Figure G.1. English Grade 9 IRT B Parameters for Operational Items 108
Figure G.2. English Grade 10 IRT B Parameters for Operational Items 109
Figure G.3. Reading Grade 9 IRT B Parameters for Operational Items 110
Figure G.4. Reading Grade 10 IRT B Parameters for Operational Items 111
Figure G.5. Mathematics Grade 9 IRT B Parameters for Operational Items 112
Figure G.6. Mathematics Grade 10 IRT B Parameters for Operational Items 113
Figure G.7. Science Grade 9 IRT B Parameters for Operational Items 114
Figure G.8. Science Grade 10 IRT B Parameters for Operational Items 115
Figure I.1. English Grade 9 Scale Score Distribution 124
Figure I.2. English Grade 10 Scale Score Distribution 124
Figure I.3. Reading Grade 9 Scale Score Distribution 125
Figure I.4. Reading Grade 10 Scale Score Distribution 125
Figure I.5. Mathematics Grade 9 Scale Score Distribution 126
Figure I.6. Mathematics Grade 10 Scale Score Distribution 126
Figure I.7. Science Grade 9 Scale Score Distribution 127
Figure I.8. Science Grade 10 Scale Score Distribution 127
Figure K.1. English Grade 9 Principal Components Scree Plot 136
Figure K.2. English Grade 10 Principal Components Scree Plot 136
Figure K.3. Reading Grade 9 Principal Components Scree Plot 136
Figure K.4. Reading Grade 10 Principal Components Scree Plot 136
Figure K.5. Mathematics Grade 9 Principal Components Scree Plot 137
Figure K.6. Mathematics Grade 10 Principal Components Scree Plot 137
Figure K.7. Science Grade 9 Principal Components Scree Plot 137
Figure K.8. Science Grade 10 Principal Components Scree Plot 137

Figure M.1. English Grade 9 Item Drift 141
Figure M.2. English Grade 10 Item Drift 142
Figure M.3. Reading Grade 9 Item Drift 143
Figure M.4. Reading Grade 10 Item Drift 144
Figure M.5. Mathematics Grade 9 Item Drift 145
Figure M.6. Mathematics Grade 10 Item Drift 146
Figure M.7. Science Grade 9 Item Drift 147
Figure M.8. Science Grade 10 Item Drift 148

1 Introduction

1.1 Background

The Utah Aspire Plus summative assessments were created out of Utah Statute 53E-4-304 (https://le.utah.gov/xcode/Title53E/Chapter4/53E-4-S304.html?v=C53E-4-S304_2019051420190514). The statute requires the Utah State Board of Education (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 English, Reading, Mathematics, and Science. The Utah Aspire Plus assessments are a hybrid of ACT Aspire and Utah Core test items. These are computer-based, fixed-length tests intended to measure end-of-grade-level high school knowledge and skills for students in grades 9 and 10. Spring 2019 marked the first administration of the Utah Aspire Plus assessments and the creation of base reporting scales for each respective grade and subject assessment.

Prior to 2019, students were assessed on the core standards through the Utah Student Assessment of Growth and Excellence (SAGE) assessment program. The Utah Aspire Plus assessment program is an extension of the Utah SAGE, still intended to measure student performance in relation to the Utah Core Standards (https://www.uen.org/core/), but also intending 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 too. The ACT® is the primary college readiness assessment submitted to local universities in Utah. As such, the Utah Aspire Plus assessments incorporate test questions from the ACT Aspire assessments that are used not only to contribute to student overall scores but also to provide a predictive indicator of performance on the ACT®. Students receive predicted ACT® score ranges for each ACT® subtest (English, Reading, Mathematics, and Science), as well as an overall predicted composite ACT® score range.

As required by the statute noted previously, the assessments also provide overall scores as indicators of end-of-grade-level expectations for 9th and 10th grade students and performance level indicators (*Below Proficient*, *Approaching Proficient*, *Proficient*, and *Highly Proficient*) for English, Reading, Mathematics, and Science.

1.2 Purpose of the Operational Tests

The Utah Aspire Plus assessments are designed for several purposes. First, the tests are intended to measure the breadth and depth of the Utah Core Standards and measure across all levels of student performance. Second, the tests are created to provide awareness of individual achievement in relation to stated performance expectations. Third, performance on the tests is intended to provide evidence of whether students are on track for college and career readiness. Finally, the tests are used to evaluate growth between 9th and 10th grade.

1.3 Prior Administrations

As stated, the first operational administration was conducted in the spring of 2019 at grades 9 and 10 for English, Reading, Mathematics, and Science. Data from that administration were used to establish the initial Utah Aspire reporting scales and the setting of performance levels. Technical details of these features and activities are presented in the *Utah Aspire Plus 2018–2019 Technical Report* (https://utah.mypearsonsupport.com/assets/pdf/UT1132740_UTPlusTechReportv4.3_WebTag.pdf).

Note that spring 2020 was intended to be the second operational administration of the Utah Aspire Plus tests. In spring of 2020, Senate Bill 3005, which included a waiver of the Utah Aspire Plus assessment requirements, was passed during the Utah Legislature’s 3rd Special Session of 2020 and signed into law on April 22, 2020. As a result, the spring testing of Utah Aspire Plus was cancelled. As a result, spring 2021 marked the second administration of the Utah Aspire Plus assessments. However, it should be noted that a waiver was sought and granted by the U.S. Department of Education (Department) to waive the accountability, school identification, and related reporting requirements for the 2020–2021 school year (https://www.schools.utah.gov/file/829f7300-020d-456e-85ac-49e85ef0795a).

English, Reading, and Mathematics summative assessments for the Utah Aspire Plus administration were created in 2019 for use in spring 2020. Given the cancellation of testing in spring 2020, the tests were instead rolled over and administered in spring 2021. Spring 2021 also marked the initial administration of new science tests. The Utah Aspire Plus Science with Engineering Education Standards (SEEds) summative assessments were administered to Utah students in spring 2021. These assessments are composed of test units that are designed to measure multi-dimensional knowledge and skill interactions across different scientific phenomena within core disciplines.

The tests were administered 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 to identify the best forms based on match to blueprint and statistical indices. After these forms were determined, they were then used to set performance standards in August of 2021.

1.4 Spring 2024 Administration

Spring of 2024 marked the fifth administration of the Utah Aspire Plus assessment for English, Reading and Mathematics and the fourth administration of the Science assessment (following the establishment of the base scale in 2021).

For the first time, remote administration was permitted for qualifying students. Additionally, this administration marked the first time Pearson's Assessment Delivery and Management (ADAM) web application was used to manage online student testing and test data.

1.5 Composition of the Operational Tests

Each operational Utah Aspire Plus test form was constructed to reflect the full test blueprint in terms of content, standards measured, and item types (Administration Resources | UT (mypearsonsupport.com)). All blueprints were designed to measure knowledge and skills described in the Utah Core Standards (https://www.uen.org/core/). For science, the operational assessments were created to measure the new Science with Engineering Education Standards (SEEds). The standards were derived from several research-based sources such as A Framework for K–12 Science Education and the Next Generation Science Standards (NGSS).

The Utah Aspire Plus tests are composed of several different types of items to measure student performance. These include multiple choice, multiple select, evidence-based selected response, and technology enhanced (TE). 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. Further, these types can be designed as two multiple-choice items, or a combination of multiple-choice and technology-enhanced (TE) items. Technology-enhanced (TE) items require specialized interactions within the online presentation for capturing student responses (e.g., drag and drop).

The Utah Aspire Plus English tests target language conventions and comprehension. Students should be able to demonstrate command of standard English grammar, usage, capitalization, punctuation, and spelling. In addition, students should be able to demonstrate vocabulary knowledge in comprehending complex texts.

The Utah Core Standards in Reading define expectations of comprehension skills, understanding tone and point of view of texts, and evaluating texts. On the Utah Aspire Plus Reading tests, students must demonstrate these skills with different types of text sources.

The assessment context for Utah Aspire Plus Mathematics is grounded in five conceptual categories from the Utah Core Standards: Number and Quantity, Algebra, Functions, Geometry, and Statistics and Probability. There are two general levels of math content for Utah Aspire Plus. The first level, referred to as Secondary Math I, extends the mathematics from the middle grades, particularly on linear and exponential relationships. The next level, Secondary Math II, focuses on quadratic relationships and comparing them to the linear and exponential relationships from Secondary Math I.

The primary emphasis of the new Utah Aspire Plus Science tests is on the multidimensional nature as expressed within the NGSS. Specific Science and Engineering Practices (SEP) and Cross-Cutting Concepts (CCC) are identified within four reporting targets (Gathering and Investigating, Developing Models, Using Mathematical Thinking, and Constructing Explanations). These are further represented within the Disciplinary Core Ideas (DCI) of Life Science, Physical Science, and Earth and Space Science.

1.6 Intended Population of the Operational Tests

The Utah Aspire Plus tests are designed for students completing their 9th and 10th grade courses in English Language Arts (ELA), mathematics, and science. The English and Reading tests are designed to assess the skills that 9th and 10th grade ELA students should have by the end of those respective years. The Mathematics tests are designed to assess the skills that 9th (Secondary Math I) and 10th grade (Secondary Math II) math students should have by the end of those respective years. The Science tests are designed to assess the skills that 9th and 10th grade students taking biology, chemistry, Earth science, or physics should have by the end of instruction (regardless of the specific course).

1.7 Overview of the Technical Report

The intended audience of the 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.

This report provides details of the maintenance of the Utah Aspire Plus testing system at grades 9 and 10 for English, 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 maintenance of existing scales, and of scoring and reporting for all tests. 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 are presented in the following chapters and address test design, development and implementation, test administration, test taker characteristics, classical item analyses, reliability analyses, item response theory (IRT) calibrations, equating, and scaling, quality control procedures, and evidence of validity.

2 Test Development

2.1 Overview of the Utah Aspire Plus Assessments, Claims, and Blueprints

The Utah Aspire Plus assessments are aligned to the Utah Core Standards and 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. Utah Aspire Plus content follows a rigorous development process that meets and often exceeds industry standards for best practices in assessment. Every item, written by Utah teachers, goes through an extensive review designed to ensure adherence to high quality and the principles of universal design.

This chapter describes the claims intended to support the purposes outlined in Chapter 1; the development of blueprints defining the components of the Utah Aspire Plus assessments that reflect the breadth of the Utah Core Standards across different levels of student understanding; and the development of tasks (items) intended to fulfill the respective blueprints and provide evidence of varying levels of performance reflective of each of the stated claims.

It should be noted that while both claims and sub claims are presented here for each subject, only the claims are reported on individual student reports (ISR). Sub claims currently only provide structure within the respective blueprints but are not reported at the individual student level.

2.1.1 English Assessment Claims

The Utah Aspire Plus English tests target language conventions and comprehension. Students should be able to demonstrate command of standard English grammar, usage, capitalization, punctuation, and spelling. In addition, students should be able to demonstrate vocabulary knowledge in comprehending complex texts.

The claim structure for the Utah Aspire Plus English tests is drawn from the Utah Core Standards and frames the design and development of the summative tests at grades 9 and 10.

Claims: The primary claims reflect the main goals for the use of the Utah Aspire Plus English tests. The first is that student performance reflects an indicator of career and college readiness as demonstrated through students' understanding of language conventions and comprehension as expected to have been attained by the end of each respective year as a prediction of performance on the ACT® English test. Second is that overall performance reflects students' understanding of language conventions and comprehension with respect to the breadth and depth of the Utah Core Standards and measures across all levels of student performance.

Sub Claims[*]: The sub claims further explicate what is measured on Utah Aspire Plus English tests and are grouped into the following categories:

- Production of Writing
- Knowledge of Language
- Conventions of Standard English

[*] It should be noted that sub claims are *not* reported on individual student reports but form an important structural element within the blueprints. They are included in this technical report for completeness.

2.1.2 Reading Assessment Claims

The Utah Aspire Plus Reading tests define expectations of comprehension skills, understanding tone and point of view of texts, and evaluating texts. On the Utah Aspire Plus Reading tests, students must demonstrate these skills with different types of text sources.

The claim structure for the Utah Aspire Plus Reading tests is drawn from the Utah Core Standards and frames the design and development of the summative tests at grades 9 and 10.

Claims: The primary claims reflect the main goals for the use of the Utah Aspire Plus Reading tests. The first is that 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. Second is that 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.

Sub Claims[*]: The sub claims further explicate what is measured on Utah Aspire Plus Reading tests and are grouped into the following categories:

- Key Ideas
- Craft and Structure
- Integration of Knowledge and Ideas

[*] It should be noted that sub claims are *not* reported on individual student reports but form an important structural element within the blueprints. They are included in this technical report for completeness.

2.1.3 Mathematics Assessment Claims

The Utah Aspire Plus Mathematics tests are grounded in five conceptual categories from the Utah Core Standards: Number and Quantity, Algebra, Functions, Geometry, and Statistics and Probability. There are two levels of math content for Utah Aspire Plus that reflect expectations at grades 9 and 10, respectively. The first level (grade 9), referred to as Secondary Math I, extends the mathematics from the middle grades, particularly on linear and exponential relationships. The next level, Secondary Math II (grade 10), focuses on quadratic relationships and comparing them to the linear and exponential relationships from Secondary Math I.

The claim structure for the Utah Aspire Plus Mathematics tests is drawn from the Utah Core Standards and frames the design and development of the summative tests at grades 9 and 10.

Claims: The primary claims reflect the main goals for the use of the Utah Aspire Plus Mathematics tests. The first is that 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. Second is that 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.

Sub Claims[*]: The sub claims further explicate what is measured on Utah Aspire Plus Mathematics tests and are grouped into the following categories:

[*] It should be noted that sub claims are *not* reported on individual student reports but form an important structural element within the blueprints. They are included in this technical report for completeness.

Math I (Grade 9)

- Algebra
- Functions
- Geometry
- Statistics and Probability

Math II (Grade 10)

- Number and Quantity
- Algebra
- Functions
- Geometry
- Statistics and Probability

2.1.4 Science Assessment Claims

The Utah Aspire Plus Science tests are developed around the Utah Core Standards for science as described in the Science with Engineering Education Standards (SEEds). These skills are applicable regardless of domain (Biology, Physics, Earth Science, and Chemistry). The claim structure for the Utah Aspire Plus Science tests is drawn from the Utah Core Standards as described in the SEEds and frames the design and development of the summative tests at grades 9 and 10.

Claims: The primary claims reflect the main goals for the use of the new Utah Aspire Plus Science tests. The first is that student performance reflects an indicator of career and college readiness as demonstrated through students' ability to understand and apply science as defined by the SEEds. Further, as expected to have been attained by the end of each respective year as a prediction of performance on the ACT® Science test. Second is that overall performance reflects students' understanding of science as defined by the SEEds with respect to the breadth and depth of the Utah Core Standards and measuring across all levels of student performance.

Sub Claims[*]: The sub claims further explicate what is measured on the new Utah Aspire Plus Science tests and are grouped into the following categories with respective SEP and CCC targets:

[*] It should be noted that sub claims are *not* reported on individual student reports but form an important structural element within the blueprints. They are included in this technical report for completeness.

- 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

These are expressed across the Life Science, Earth and Space Science, and Physical Science DCIs.

2.2 Utah Aspire Plus Blueprints

The Utah Aspire Plus tests are administered in English, Reading, Mathematics, and Science in grades 9 and 10 and are described in Section 1.5. For the Utah Aspire Plus tests, the creation of test blueprints was driven by the intended purposes detailed previously in order to support the respective claim structures. The blueprints for Utah Aspire Plus 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 will be 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 the standards are fully represented over time.

The 2024 Utah Aspire Plus blueprints can be found at: Administration Resources | UT (mypearsonsupport.com).

2.3 Test Development Activities

Prior to the creation of Utah Aspire Plus, students were tested on the Utah Core Standards through the Utah Student Assessment of Growth and Excellence (SAGE). The Utah Aspire Plus assessments were built from existing Utah SAGE banked content combined with items from ACT Aspire to allow for predictions of students' preparedness for meeting college readiness. All available content for creation of the 2024 Utah Aspire Plus tests was based on the existing item banks described in the *Utah Aspire Plus 2018–2019 Technical Report* (available under Reporting Resources at https://utah.mypearsonsupport.com/admin-resources.html).

For English, Mathematics, and Reading, the ACT Aspire forms that are used to source items are alternated each year. This helps limit exposure of the Aspire content that might otherwise negatively impact ACT predication score activities.

For 2024, there was one core operational form for regular online and text-to-speech forms. Mathematics and Reading forms consisted of operational items and a small set of field-test items. The number of field test forms for 2024 by grade and subject is shown in Table 2.1.

Table 2.1. Field Test Forms

Subject	Grade	Number of FT Versions
Reading	9	4
	10	8
Mathematics	9	20
	10	16

In addition to the ONEN forms, there are several accommodated forms. These include:

- Non-screen reader (NREN)
- Screen reader (SREN)
- Spanish (ONSP)

The grade 9 & 10 English, grade 9 and 10 Reading, and grade 9 mathematics accommodated forms were a reuse of the 2022 forms. Reading and English non-accommodated forms were also a reuse of the 2022 forms.

2.3.1 Operational Forms Development

The construction of test forms for the 2024 Utah Aspire Plus was a coordinated effort between experts from the Utah State Board of Education, 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 measure students' mastery of the Utah Core Standards or the Utah Aspire Plus Science with Engineering Education Standards. These standards are used to drive Utah instruction as well as developing the Utah Aspire Plus tests. As stated earlier, 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. In order to accomplish this, approximately 50% of the Utah Aspire Plus tests (less for Mathematics) are composed of items from ACT Aspire. As noted, these items serve multiple purposes, which include being used to derive prediction scores between the Utah Aspire Plus scales and ACT scales.

The general test development process for Utah Aspire Plus was initiated with the selection of items from ACT Aspire. Items were selected based on match to blueprint, as well as statistical indicators of item quality and fairness provided from the SAGE and ACT Aspire banks, respectively. ACT Aspire items were positioned within each form in the same locations as originally administered within 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 best final version was approved by all. It should be noted that bank limitations meant there were also 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.

2.3.2 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 Utah Aspire Plus tests, some 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. These indices are described in detail further on in the report.

The guidelines for creation of the Utah Aspire Plus forms were as follows:

- **Target item difficulty range of between 0.30 and 0.85.** Based on *p*-values, where the percentage reflects the percentage of students correctly responding to the item. Items awarding more than one 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.” As such, items flagged “C” should be avoided and minimal use of items flagged “B” should be used and/or balanced within a form where possible.

More detailed description of the statistical indices reflecting item functioning for the Utah Aspire Plus tests appears later in this report, and distributional results by grade and subject test from the 2024 operational administration are presented in Appendix C. It should be noted that Appendix C reflects post hoc calculations, not what was available within the context of test construction. It should further be noted that while most items selected to appear on the initial Utah Aspire Plus forms were within the guidelines described here, there were instances in which bank limitations meant some items did fall outside the thresholds.

2.3.3 2024 Match to Test Blueprint

Tables 2.2 through 2.9 present the match between the final 2024 operational forms of Utah Aspire Plus and the test blueprints. English, Reading, Mathematics, and Science final forms matched all targets by item type, depth of knowledge, and reporting category.

Table 2.2. Utah Aspire Plus English Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Multiple Choice	24–31	60%	89%	79%
Technology Enhanced	8–13	20%	37%	21%
Depth of Knowledge				
Level 1	15–20	38%	57%	47%
Level 2	8–16	20%	46%	21%
Level 3	12–15	30%	43%	32%
Reporting Categories				
Production of Writing	7–12	18%	34%	29%
Knowledge of Language	4–10	10%	29%	11%
Conventions of Standard English	20–30	50%	86%	61%

Table 2.3. Utah Aspire Plus English Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Multiple Choice	24–31	60%	89%	78%
Technology Enhanced	8–13	20%	37%	22%
Depth of Knowledge				
Level 1	15–20	38%	57%	41%
Level 2	8–16	20%	46%	24%
Level 3	12–15	30%	43%	35%
Reporting Categories				
Production of Writing	7–12	18%	34%	27%
Knowledge of Language	4–10	10%	29%	16%
Conventions of Standard English	20–30	50%	86%	57%

Table 2.4. Utah Aspire Plus Reading Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Evidence-Based Selected Response	3–6	9%	17%	14%
Multiple Choice	22–30	63%	86%	69%
Technology Enhanced	2–7	6%	20%	17%
Depth of Knowledge				
Level 1	4–7	11%	20%	14%
Level 2	14–20	40%	57%	51%
Level 3	12–15	34%	43%	34%
Reporting Categories				
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%

Table 2.5. Utah Aspire Plus Reading Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Evidence-Based Selected Response	3–6	9%	17%	14%
Multiple Choice	22–30	63%	86%	71%
Technology Enhanced	2–7	6%	20%	14%
Depth of Knowledge				
Level 1	4–7	11%	20%	14%
Level 2	14–20	40%	57%	49%
Level 3	12–15	34%	43%	37%
Reporting Categories				
Key Ideas	12–16	34%	46%	46%
Craft and Structure	12–18	34%	51%	40%
Integration of Knowledge and Ideas	3–7	9%	20%	14%

Table 2.6. Utah Aspire Plus Mathematics Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Multiple Choice	30–33	75%	83%	75%
Technology Enhanced	7–10	18%	25%	25%
Depth of Knowledge				
Level 1	8–12	20%	30%	30%
Level 2	15–20	38%	50%	48%
Level 3	9–13	23%	33%	23%
Reporting Categories				
Algebra	9–11	23%	28%	25%
Functions	10–12	25%	30%	28%
Geometry	9–11	23%	28%	25%
Statistics and Probability	7–9	18%	23%	23%

Table 2.7. Utah Aspire Plus Mathematics Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Multiple Choice	30–33	75%	83%	83%
Technology Enhanced	7–10	18%	25%	18%
Depth of Knowledge				
Level 1	8–12	20%	30%	28%
Level 2	15–20	38%	50%	48%
Level 3	9–13	23%	33%	25%
Reporting Categories				
Number and Quantity	2–4	5%	10%	8%
Algebra	9–11	23%	28%	28%
Functions	10–12	25%	30%	28%
Geometry	11–13	28%	33%	28%
Statistics and Probability	2–4	5%	10%	10%

Table 2.8. Utah Aspire Plus Science Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Multiple Choice	18–21	78%	91%	78%
Technology Enhanced	3–6	13%	26%	22%
DCI				
Life Science	4–8	17%	35%	26%
Earth and Space Science	4–8	17%	35%	26%
Physical Science	9–13	39%	57%	48%
Reporting Categories				
Gathering & Investigating	4–8	17%	35%	17%
Developing Models	4–8	17%	35%	17%
Using Mathematical Thinking	5–9	22%	39%	30%
Construct Explanations	5–9	22%	39%	35%

Table 2.9. Utah Aspire Plus Science Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2024 Form
Item Type				
Multiple Choice	18–21	78%	91%	87%
Technology Enhanced	3–6	13%	26%	13%
DCI				
Life Science	4–8	17%	35%	30%
Earth and Space Science	9–13	39%	57%	43%
Physical Science	4–8	17%	35%	26%
Reporting Categories				
Gathering & Investigating	4–8	17%	35%	22%
Developing Models	4–8	17%	35%	26%
Using Mathematical Thinking	5–9	22%	39%	26%
Construct Explanations	5–9	22%	39%	26%

For additional information on the 2024 operational forms, Appendix A contains a breakdown reporting categories and standards by item type and depth of knowledge (DOK), with the exception of science (which does not use DOK).

3 Operational Administration

3.1 Testing Window

The 2024 administration of the Utah Aspire Plus assessments was March 4–May 10, 2024. Utah Aspire Plus can be administered on a subject-by-subject basis or as a complete battery with all tests administered in one sitting. Each subject test, however, must be administered in one sitting. In other words, once a subject test is started, it must be completed within that sitting.

3.2 Test Administration and Security Policies

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 every aspect of the Utah Aspire Plus test administration. These resources are further presented in a manner that addresses those responsible for carrying out the administration for all students as well as for educators and students to become familiar with the tests themselves (e.g., via practice tests and such) and for interpretation of test scores.

The Utah Aspire Plus tests are secure tests that follow the Utah Aspire Plus blueprints for each assessed subject area. All test items are secured items and may not be reviewed with students, discussed as a class, or reviewed during instructional conversations. Discussing, reviewing, recording, or transcribing test questions in any format is a violation of test security. All test security requirements of Utah Aspire Plus must be met. Personnel involved in test administration must complete testing ethics training. The Utah Standard Test Administration and Testing Ethics policy can be found under Testing Ethics here: https://schools.utah.gov/assessment/.

The LEA Assessment Director was responsible for ensuring that each student had an appropriate opportunity to demonstrate knowledge, skills, and abilities related to Utah Aspire Plus grade-based courses and assessments. This ensures that each student had a standardized (similar and fair) testing experience for a given assessment. Each LEA was responsible for determining school testing schedules. Subject tests did not have to be administered in any prescribed order. Subject tests could *not* be divided into multiple sessions. Once a subject test session began, the subject test had to be completed within that sitting.

It should be noted that 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. For the 2023–2024 administration, the testing times were: 45 minutes for English, 75 minutes each for Reading and Mathematics, and 60 minutes for Science. It should be noted that students whose IEP, Section 504, or English Learner plan specified an accommodation for extended time were able to use extended time accommodations on Utah Aspire Plus as appropriate.

3.2.1 Online Administration and Monitoring

The Utah Aspire Plus tests are administered online via the Pearson assessment delivery and management systems. ADAM is the web application used by test staff (i.e., test proctors, teachers, and administrators) to manage online testing and start and monitor tests. TestNav is the test delivery engine used by examinees to take the tests.

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.

The spring 2024 administration was the first time students could take the assessment remotely. Remote testing is subject to the following guidelines:

- In order for a student to be eligible for remote testing, 100% of their learning needs to be online.
- Students requiring a paper test (i.e., LP, Braille, Human Reader) are not eligible for remote testing.

Remote proctoring is described in more detail in Section 8.1.3.

3.3 Test Accommodations and Supports

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

For Utah Aspire Plus, accommodated test forms 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.

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, to include 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 is not only a matter of directly 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 follow 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. That is, highly qualified item writers with extensive expert content experience are involved in the creation and review process of transadapted and/or sister item creation. Several reviews are held throughout the creative process involving Pearson and USBE content and psychometric experts to ensure match to source.

Testing accommodations and supports, including those mentioned above, are outlined in the TAM. (A complete list of accessibility and accommodation features for the Utah Aspire Plus assessments can be found in at https://www.schools.utah.gov/specialeducation/programs/accessibilityaccommodationsassessment.)

Embedded and non-embedded supports are generally available to all students, whether through the online system or locally arranged. The list below provides the embedded and non-embedded supports provided within Utah Aspire Plus, as outlined in the TAM:

- Embedded:
    - 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
- Non-embedded:
    - 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

Testing accommodations require prior designation in a student's Individualized Education Program (IEP), 504, or English Learner (EL) plan. The list below provides the test accommodations, in addition to those supports previously mentioned.

- 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

3.4 Test Taking Irregularities and Security Breaches

Test irregularities are non-standard situations that occur during test administration 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 were provided within the school's testing window. If there was an emergency that interrupted testing for an entire class or school, decisions about whether a test could be started again or not were to be made on a case-by-case basis by working with the Utah State Board of Education assessment team.

3.4.1 Test Interruptions

In the event that a student got sick, had to leave and could not return during the test, or for any other reason did not complete a test which had already begun, the test was to be concluded and submitted immediately. To maintain the security of the test questions, students were not allowed to restart or take a test over again.

3.4.2 Scoring of Interrupted Tests

If a student was interrupted and completed only part of a test before it was concluded and submitted, the student might not have received a score. A student must have attempted 85% of the questions to receive a score. If a student did not attempt at least 85% of the test questions, a score could not be generated, and no test score would be reported for that particular test. Overall composite scores would not be available for students who had missing subject test scores because the composite score is calculated using all four subject tests.

3.4.3 Wrong Test Form/Accommodation

If a student began a test using a test form or accommodation that they were not supposed to have, the teacher/proctor should have immediately stopped the test. In those instances, a new test assignment had to be created and a new test administration could proceed as normal from that point.

3.4.4 Extended Time Accommodation Issues

Extended time accommodations must be applied before applying any participation code and before starting sessions. In the event that 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.

3.4.5 Test Invalidation

Tests could be invalidated when a student’s performance was not deemed an accurate measure of their ability (e.g., the student cheated, used inappropriate materials, etc.). When a test is invalidated, the student is not given another opportunity to take the test. Invalidating a test has to be completed by the district testing administrator.

3.5 *Test Taker Characteristics*

Table 3.1 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 3.1. Spring 2024 Participation Rates for Utah Aspire Plus

Students	Subgroup	English		Reading		Mathematics		Science	
		Gr. 9	Gr.10	Gr. 9	Gr. 10	Gr. 9	Gr. 10	Gr. 9	Gr. 10
All	Students Scored	45,391	43,431	45,559	43,594	43,674	42,840	45,542	43,491
Sex	Female %	47.82	47.52	47.95	47.58	47.38	47.37	47.92	47.56
	Male %	52.06	52.39	51.93	52.32	52.51	52.54	51.95	52.34
	Unknown %	0.13	0.10	0.12	0.10	0.11	0.09	0.13	0.09
Ethnicity	Hispanic or Latino Ethnicity %	19.70	19.52	19.86	19.60	19.41	19.45	19.88	19.64
	Asian %	1.68	1.69	1.69	1.69	1.66	1.68	1.70	1.69
	Native Hawaiian or Other Pacific Islander %	1.29	1.39	1.29	1.43	1.28	1.42	1.31	1.41
	Black or African American %	1.30	1.34	1.30	1.35	1.27	1.35	1.31	1.37
	American Indian or Alaska Native %	0.93	0.95	0.93	0.96	0.90	0.93	0.93	0.95
	White %	71.54	71.55	71.37	71.41	71.94	71.63	71.31	71.39
	Other %	3.56	3.56	3.55	3.57	3.54	3.54	3.56	3.54
Limited English Proficiency	No %	91.59	91.93	91.37	91.85	91.49	91.85	91.37	91.79
	Yes %	8.41	8.07	8.63	8.15	8.51	8.15	8.63	8.21
Economic Disadvantage	No %	73.07	74.61	72.96	74.51	73.21	74.64	72.95	74.50
	Yes %	26.93	25.39	27.04	25.49	26.79	25.36	27.05	25.50
Special Education	No %	90.05	90.55	90.05	90.49	89.83	90.43	90.06	90.54
	Yes %	9.95	9.45	9.95	9.51	10.17	9.57	9.94	9.46

3.6 Testing Time

One of the key questions in moving from an untimed to a timed test administration (from SAGE to Utah Aspire Plus) is gauging the extent to which the time allotted appears to be reasonable. As mentioned in Section 3.2, the operational testing times for the Utah Aspire Plus tests are: 45 minutes for English, 75 minutes for Reading, 75 minutes for Mathematics, and 60 minutes for science. Students needing extra time fall into three categories: time and a half, double time, or triple time. After the spring 2024 test administration, student total testing time was analyzed for each test. Overall, students completed the assessments within the recommended testing times. Tables 3.2 and 3.3 provide breakdowns of student testing time across the full range of testing times. In other words, the percentile rankings are of the amount of time in minutes students took to complete the respective test. More specifically, with the Grade 9 English results for students testing using regular time (45 minutes), examination of the 95th percentile (P95) means that 95% of students finished the test in 41 minutes or less.

Additional information is presented in Appendix B, which provides a graphical display (box-and-whisker plot) of student testing time for each test. 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 students in general had enough time to complete each respective test within the given allotments.

Table 3.2. Student Testing Time for Spring 2024 Utah Aspire Plus: English and Reading

				Testing Time (minutes)									
				Descriptive Statistics				Percentiles					
Subject	Grade	Group	N	Minimum	Maximum	Mean	St. Dev.	P50	P75	P80	P85	P90	P95
English	9	Regular Time	40936	0	45	27	8	27	33	34	36	38	41
		Time and a Half	3847	1	67	31	14	30	40	42	46	50	56
		Double Time	496	1	89	31	15	29	40	42	45	49	60
		Triple Time	112	6	134	37	19	34	44	47	52	57	66
	10	Regular Time	38864	0	45	26	9	26	32	34	36	38	41
		Time and a Half	4150	1	67	29	13	28	37	39	43	47	53
		Double Time	351	2	89	29	16	28	37	39	44	50	57
		Triple Time	66	6	108	36	22	31	39	43	51	59	98
Reading	9	Regular Time	41079	0	89	43	16	44	55	57	61	64	70
		Time and a Half	3864	1	112	42	23	40	56	61	65	71	83
		Double Time	504	2	149	46	24	44	59	63	69	75	89
		Triple Time	112	5	224	53	32	48	64	69	77	88	113
	10	Regular Time	38954	0	75	36	15	36	46	48	51	55	61
		Time and a Half	4210	1	112	36	21	33	47	51	55	62	74
		Double Time	360	2	149	38	24	36	49	54	58	63	79
		Triple Time	70	4	160	49	30	44	59	62	80	91	112

Table 3.3. Student Testing Time for Spring 2024 Utah Aspire Plus: Mathematics and Science

													Testing Time (minutes)
						Descriptive Statistics							Percentiles
Subject	Grade	Group	N	Minimum	Maximum	Mean	St. Dev.	P50	P75	P80	P85	P90	P95
Math	9	Regular Time	39217	1	89	52	16	54	64	67	69	71	73
		Time and a Half	3840	2	118	49	24	47	63	68	73	82	96
		Double Time	502	3	148	50	28	47	64	70	79	90	105
		Triple Time	115	10	223	64	34	60	77	79	87	107	144
	10	Regular Time	38195	1	89	45	18	47	60	62	65	69	72
		Time and a Half	4213	1	126	41	24	38	56	60	65	72	85
		Double Time	361	2	148	39	25	35	51	56	59	71	86
		Triple Time	71	6	132	50	26	46	63	69	76	81	98
Science	9	Regular Time	41076	0	59	31	12	32	40	42	44	47	52
		Time and a Half	3853	0	89	29	16	27	39	43	46	51	59
		Double Time	506	1	117	29	19	27	40	43	48	53	63
		Triple Time	107	1	85	37	17	37	46	48	51	58	72
	10	Regular Time	38861	0	59	27	13	27	36	38	40	44	49
		Time and a Half	4223	1	89	25	17	22	34	37	41	46	56
		Double Time	345	1	116	25	18	22	34	37	43	48	56
		Triple Time	62	3	86	32	19	31	41	43	47	51	74

4 Score Reporting

4.1 IRT Pattern Scoring

Item parameters derived from previous IRT calibrations were used to estimate student ability (“theta”) scores by item response patterns. This is commonly referred to as pattern scoring. Pattern scoring takes advantage of the fact that items differ in their item characteristics and that an estimate of a student’s ability is based on their specific pattern of responses in combination with the item characteristics across all items. See Chapter 7 for more discussion of the IRT model and calibration methods.

The software package Operational Scoring: IRT Score Estimation (ISE V1.3.f; Chien & Shin, 2012) was used to perform the pattern scoring process and provide student scores on the IRT metric, using the student scored responses and the item response theory (IRT) item parameters for the operational items.

Two data-driven input files are 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 algorithms to generate the maximum likelihood estimated (MLE) of *theta* values. Specific configuration details include setting the upper- and lower-bound theta estimates, in this case +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 the number of decimal places for theta estimates (4).

IRT parameters for all 2024 Utah Aspire Plus operational items were used for estimating individual student scores for all forms. Table 4.1 presents the summary statistics for the IRT (*a*-, and *b*-) parameter estimates. The summary statistics shown include the total number of items, along with the mean, standard deviation (SD), minimum, and maximum.

Table 4.1. IRT Summary Parameter Estimates for Utah Aspire Plus Operational Items

			Summary of *a* Estimates				Summary of *b* Estimates			
Grade	Subject	No. of Items	Mean	SD	Min	Max	Mean	SD	Min	Max
9	English	46	0.80	0.36	0.17	1.65	-0.36	1.29	-2.58	3.16
	Reading	35	0.71	0.35	0.19	1.40	0.38	1.29	-1.25	5.56
	Mathematics	40	1.11	0.34	0.57	2.09	0.27	0.86	-1.64	1.57
	Science	23	0.72	0.24	0.30	1.17	0.49	0.76	-0.83	2.38
10	English	44	0.93	0.42	0.28	2.03	-0.39	0.96	-1.79	2.06
	Reading	35	0.89	0.46	0.25	2.33	-0.19	0.95	-1.80	2.89
	Mathematics	40	1.13	0.39	0.48	2.84	0.53	0.70	-1.47	1.46
	Science	23	1.09	0.68	0.15	3.11	0.60	0.92	-1.64	3.31

4.1.1 Quality Control of IRT Scoring

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 two independent internal replicators. This scoring dry run was conducted at the overall test level as well as by reporting categories. Any differences were resolved and rerun until both parties' results were identical and deemed correct based on careful examination of output.

4.2 Appropriate Uses for Scores and Reports

As discussed, 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 content area based on expectations of students' knowledge at the end of grades 9 and 10. It should be noted that 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 Individual Student Report (ISR) communicates an individual student's test scores and interpretations of achievement based on those scores The ISR provides the "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 under "Score Interpretation Guide" here: https://utah.mypearsonsupport.com/admin-resources.html. For the Utah Aspire Plus tests, overall scale scores, performance level indicators, and predicted performance ranges for the ACT tests are provided. Note that no subscores are currently reported on student ISRs.

4.3 Utah Aspire Plus Reporting Scale

Commonly derived scores based on IRT 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 test-takers 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:

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

This transformation provides the following characteristics: 1) the mean of the scale is 200, 2) the standard deviation of the scale is 25, 3) the lowest operating scale score (LOSS) is 100, and 4) the highest operating scale score (HOSS) is 300. Composite scores were also created for Utah Aspire Plus. A composite score representing English Language Arts (ELA) is the average of a student’s English and Reading scale scores, whereas a composite score representing Science, Technology, Engineering, and Mathematics (STEM) is the average of a student’s Mathematics and Science scale scores.

4.4 Standard Setting

Descriptions of student performance are often used to help enhance the reporting of student scores beyond an overall reported score and references to other students or groups of students. Performance levels and descriptions of performance divide the test scores into meaningful categories and align to performance ranging from low to high. For Utah, these categories are called *Below Proficient*, *Approaching Proficient*, *Proficient*, and *Highly Proficient*. Performance level descriptions (PLDs) accompany these labels to describe typical performance of students within each group.

Standard settings were conducted in August of 2019 (for all subjects) and again for science in August of 2022 following the first administration of the new assessment based on the SEEds. PLDs are the core of all standard setting meetings. The PLDs for the Utah Aspire Plus assessments can be found online.

Utah educators were convened to operationalize the PLDs through standard setting, a process of determining test score thresholds, or "cut points," to divide the test scores into the four performance groups. Final scale score cuts for English, Reading, Mathematics, and Science are presented in Table 4.2.

Table 4.2. Utah Aspire Plus Scale Score Cuts by Grade and Subject

Grade	Subject	Scale Score Cut Points		
		Approaching Proficient	Proficient	Highly Proficient
9	English	165	202	242
	Reading	166	204	231
	Mathematics	172	206	233
	Science	187	211	237
10	English	161	200	245
	Reading	175	204	235
	Mathematics	181	210	236
	Science	187	210	240

4.5 ACT Predicted Score Ranges

As noted, one of the goals of the Utah Aspire Plus assessments is to be predictive of college readiness at grades 9 and 10, and the means of this is in terms of providing prediction score ranges of performance on the ACT for the four subject tests (English, Reading, Mathematics, and Science) and the Composite score (the average of the four subject tests). Predicted ranges of performance were determined originally between ACT Aspire scores 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.

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. Specific details from the original prediction score studies can be found in the *Utah Aspire Plus 2018–2019 Technical Report*.

In addition to relying on the relationship between the Utah Aspire Plus tests to the ACT Aspire scales for deriving the initial ACT prediction score ranges for the 2019 administration, the intention was to provide updated predictions based on longitudinal data as it becomes available. The updated ACT score ranges directly link the Utah Aspire Plus scores at grades 9 and 10 to ACT scores at grade 11. In spring 2020, the first longitudinal data was available for this purpose. The initial longitudinal Utah-to-ACT prediction studies were based on students who were in the 10th grade during the 2019 administration of the Utah Aspire Plus tests. The second longitudinal study was conducted in the spring of 2021. Details of this study can be found in Appendix J of the *Utah Aspire Plus 2020–2021 Technical Report*.

A third longitudinal study was conducted in 2022 to update the science grade 10 predictions. This study included students who were in 10th grade in 2021 and took the ACT as 11th grade students in spring 2022. Details of this study can be found in Appendix H of the *Utah Aspire Plus 2021–2022 Technical Report*.

All technical reports can be found under Reporting Resources at Administration Resources | UT (mypearsonsupport.com)

4.6 2023–2024 Utah Aspire Plus Performance Results

Descriptive statistics of the scale scores for each Utah Aspire Plus assessment are in Appendix H. The descriptive statistics are provided for the overall testing population, as well as by subgroups—sex, ethnicity, and special populations. Average scale scores as well as standard deviations, scores at the 25th, median, and 75th percentiles are also reported as well as skewness. Scale score distributions for each Utah Aspire Plus assessment are provided in Appendix I, for the overall testing population. Appendix J contains the performance level distributions of each Utah Aspire Plus title. The tables contain the percentages of students being classified into each respective performance level.

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

While comparatively, a similar number of students were tested in 2024 as compared to 2019, the percent of completed tests varied. In 2019 completion rates for registered testers was approximately 91–93%. In 2022, the completion rate ranged from 84–88%. In 2023, the completion rate ranged from 84–90%. In 2024, the completion rate ranged from 81%–86%. Overall performance was similar in 2024 to 2023.

5 Classical Item Analyses

5.1 Item Analyses

In Chapter 2, statistical indices used in the test construction process were introduced. 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 noted, while the best possible initial forms were created, there were instances in which not all statistical targets were fully met. This chapter describes in more detail those classical item statistics. Additionally, after the Utah Aspire Plus 2023–2024 operational administration, classical item statistics were also calculated. Results are presented in Appendix C.

5.1.1 *p*-Value and Item Mean Scores

Item difficulty offers an index of how easy or hard a given test question is to answer correctly or to earn a given score point for items scored according to a rubric. For dichotomously scored items (items scored correct or incorrect), item difficulty is indicated by its *p*-value, which is the proportion of test takers who answered that item correctly. The range for *p*-values is from 0 to 1.

For polytomously scored items (items scored according to a rubric with multiple points awarded), difficulty is indicated by the mean item score. Here the average ranges from 0 to the maximum total possible points for an item. To facilitate interpretation, the mean item values for polytomously scored items can also be expressed on the *p*-value metric as percentages of the maximum possible score.

5.1.2 Item-Test Score Correlations

Correlations between a given item score and total test score are used to evaluate how well items differentiate between “high” and “low” performing students. In general, the higher the correlation the better an item is at differentiating between high- and low-performing students. As this index is a correlation, it ranges from –1 to +1 (where +/– 1 reflects a perfect correlation and 0 reflects no correlation). When the correlation is negative, it means low-performing students on the test are answering the given question correctly more often than high-performing students, and this would be a reason to further investigate the item for potential flaws.

In addition to the correlation between item score and total test score, the same approach can be applied to each answer option of multiple-choice items. Although not provided in this report, this information is used within the context of data review and allows for further evaluation of the full functioning of multiple-choice items, as it focuses on the effective functioning of the options (distractors) which are other than the correct answer.

5.1.3 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). It is important, however, 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 and 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 5.1, the number of focal and reference group members scoring in each possible item response is captured.

Table 5.1. Item 2x2 Contingency Table for the kth Score Level

	Item Score		
Group	Correct (1)	Incorrect (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 (MHD: 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, Donoghue, and Grima, 1993) procedure is also used for detecting DIF, for items worth more than one 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.

5.2 Classical Item Summaries for Operational Administration

As noted, summaries of classical item statistics from the initial operational administration of Utah Aspire Plus are located in Appendix C. Examination of the distribution of items by difficulty across each test shows that items do vary in difficulty across each test, with most items between 0.30 and 0.75. There are items that did fall outside the guidelines outlined previously. Their inclusion was necessary to meet blueprints given limitations to the available item banks. The same can be said of the distributions of item-total correlations and DIF results, where there were items included in the tests that fell outside the guidelines but were ultimately included on final forms as the best available. Overall, even where items fell outside the guidelines, they were still useful. This was particularly true for the science assessments, where due to bank limitations and cluster design, some very difficult items and items with low discrimination were included on final operational forms to help hit blueprint targets.

6 Reliability

Estimation of reliability of a given assessment is critical in order to understand the precision of measurement for individual test scores. Test score reliability estimates are typically provided in both a classical as well as an item response theory (IRT) context. Classical reliability estimates such as standard error of measurement (SEM) or Cronbach's alpha 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. There are a number of different approaches available to estimate reliability of test scores. For Utah Aspire Plus tests, both classical reliability and reliability within an item response theory framework were computed.

6.1 Classical Definition of 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.

6.2 Classical Test Theory Reliability Estimates

6.2.1 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. Probably the most 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

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

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.

Coefficient alpha reliability estimates are provided in Appendix D for the overall testing population as well as by sex, ethnicity, and other student breakout groups. In addition, they are also provided by each reporting category (though again it should be noted that currently, only overall scores are reported on individual student reports, and no subscores are reported).

6.2.2 Standard Error of Measurement

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

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

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. Test standard errors of measurement are provided in Appendix D and are presented on the Utah Aspire Plus scale score metric (s_x = 25).

6.3 IRT-Based Reliability

Where estimation of reliability is within a classical test theory frame, it should be noted that such measures are sample specific. Additionally, error estimates such as the SEM are 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 examinees (identical for all).

For Utah Aspire Plus, student scores are derived within an item response theory framework (IRT) through pattern scoring based on the three-parameter logistic (3PL) and two-parameter logistic (2PL) measurement models (these are more thoroughly described later in this report). Under the IRT model, measurement precision is expressed as Conditional Standard Errors of Measurement (CSEM) and is equal to the inverse of the square root of the test information function across the ability continuum (see Hambleton and Swaminathan, 1985).

CSEMs depend upon 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.

The conditional standard errors for Utah Aspire Plus tests are provided in Appendix E, each including a line indicating the scale score cut score for Proficient. Ideally, the lowest value of conditional standard error of measurement occurs at the location of Proficient.

6.4 Reliability of Performance Level Categorization

Every test administration will result in some error in classifying examinees. 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 achievement levels. For example, some students may have a true achievement 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 achievement level. As discussed in Section 6.2.2, 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 achievement levels can be imperfect, especially for the borderline students whose true scores lie close to achievement-level cut scores.

For the Utah Aspire Plus assessment, the levels of achievement are *Below Proficient*, *Approaching Proficient, Proficient,* and *Highly Proficient.* A description and analysis of classification accuracy and consistency indices are provided below. All indices were calculated using the BB-CLASS software (Brennan, 2005).

6.4.1 Accuracy and Consistency

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.” Since 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.

6.4.2 Calculating Accuracy

To calculate accuracy, a 4 x 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 fall into performance level x and whose observed scores fall into performance level y. Table 6.1 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 6.1. Example Accuracy Classification Table

	Observed Score				
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 6.2 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 6.2. Example Accuracy Classification Table for Proficient Cut Point

	Observed Score				
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

6.4.3 Calculating Consistency

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 6.3 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 6.3. Example Consistency Classification Table

	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

6.4.4 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

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

where P is the proportion of consistent classifications and P_c is the proportion of consistent classification by chance. Using Table 6.3, P is the sum of the shaded cells whereas P_c is

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

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 6.3 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. Estimates of classification accuracy and consistency indices—including kappa coefficients—for overall performance level classification and at the Proficient cut point are provided in Appendix F.

7 Field Test Calibration and Drift Analyses

7.1 IRT Overview

Item response theory (IRT) was used to create the base scales for the Utah Aspire Plus assessments. All assessments were pre-equated. Item parameters were estimated either from prior operational post-equating, or field test calibration. See the *Utah Aspire Plus 2021–2022 Technical Report* (available under Reporting Resources at https://utah.mypearsonsupport.com/admin-resources.html) and prior technical reports for details on these processes. 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 served several purposes:

- Calibration of field test items
- Identification of items with parameter drift
- Update of bank parameters

In this section of the technical report, the following topics related to IRT calibration and equating are discussed:

- IRT Data Preparation
- Description of the Calibration Process
- Drift Analyses

7.2 IRT Data Preparation

7.2.1 Student Inclusion/Exclusion Rules

The data preparation for the IRT calibration process began with all Utah students who were administered the “base” forms (i.e., online, English-language forms).

The samples for item parameter estimation included the following:

- Students from the online, English language test forms,
- Students with the same grade battery of tests, and
- Students with a valid test score status for a subject test.

Students without a valid test score were excluded from calibration data.

7.2.2 Quality Control of the 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.

7.3 Description of the Calibration, Equating, and Scaling Process

7.3.1 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 model (3PLM; Birnbaum, 1968), denoted as

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

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 one point (e.g., technology-enhanced items) are analyzed via the two-parameter logistic model (2PLM; Birnbaum, 1968), which is a reduced model from the 3PLM, where the pseudo-chance parameter, *c*, is assumed zero. Items worth two points were analyzed via the generalized partial credit model (GPCM; Muraki, 1992), denoted as

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

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$, $p_{im}(\theta_j)$ is the probability of an examinee 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).

7.3.2 IRTPRO Calibration Procedures and Convergence Criteria

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.

Note that 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 on order of 2,000). IRTPRO (Scientific Software International, Inc., 2017) was used to obtain the IRT parameter estimates using the measurement models described in Section 7.3.1. 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 the 3PLM: 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. The formula for this conversion is:

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

7.3.3 Calibration Quality Control

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.

7.3.4 Equating

A common item non-equivalent groups approach (Kolen and Brennan, 2014) was used for equating the 2024 forms to the base scales.

The Stocking and Lord (1983) test characteristic curve methodology was used to derive equating constants for each grade-subject test. The operational items were used as the common-item linking set. The banked IRT item parameter estimates for all of the Utah Aspire Plus operational items, and the respective item parameter estimates from the 2024 administration described in Section 7.3.2, were used to obtain transformation constants. This was conducted using the computer program STUIRT (Kim & Kolen, 2004).

Equating was carried out in conjunction with a drift analysis procedure, described in Section 7.3.5, which resulted in a final set of Stocking and Lord scaling constants. These constants were then applied to all 2024 calibrated items to obtain a set of parameters for the operational and field test items. Final Stocking and Lord scaling constants used for placing tests onto the Utah Aspire Plus base scales are presented in Table 7.1.

Table 7.1. 2024 Final Stocking and Lord Scaling Constants

Subject	Grade	Slope	Intercept
English	9	1.070	-0.268
	10	1.058	-0.146
Reading	9	1.064	-0.204
	10	1.053	-0.203
Math	9	1.087	-0.249
	10	1.048	-0.320
Science	9	1.158	0.118
	10	1.031	-0.112

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

1. Item was field tested in 2024.
2. Item was used operationally for the first time in 2024 (prior parameters were from field test administration)
3. Item showed drift during the equating process, as described in Section 7.3.2.

7.3.5 Drift Analysis

A critical step in carrying out an equating 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, Hambleton, Kirkpatrick, & Meng, 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.

Plots showing D^2 values following the initial equating are given in Appendix M. Counts of operational items showing drift are given in Table 7.2.

Table 7.2. 2024 Items Showing Drift

Subject	Grade	Number of items showing drift
English	9	3
	10	1
Reading	9	1
	10	2
Math	9	1
	10	2
Science	9	1
	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 which were operational in 2024 for the first time.

Scatterplots of the operational items can be found in Appendix G. Overall, item functioning of common items can be described as typical and stable. No more than three items in any of the common item sets were removed from final linking solutions. Scatterplots and correlations of IRT difficulty and discrimination parameters showed strong correlations.

7.4 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

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

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

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

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

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

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 $ZQ_{1,\text{crit}}$ to determine acceptable fit. The conversion formula was

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

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., for an item scored 0, 1, 2: there are 2 independent step values—the *b* parameter is simply the mean of the step values and is not, therefore, 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 2024 administration to assess pre-equating. Results were consistent with the drift analyses and did not suggest any concerns with model selection.

8 Quality Control

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.

8.1 Online Assessment Delivery

8.1.1 Item Validation

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) as well as during forms development.

8.1.2 Test Administration

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.

8.1.3 Remote Proctoring

New in spring 2024 was the ability for ADAM to allow for remote proctoring and administration. 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. For students, the TestNav platform for taking the test is the same, with the exception that there are additional system checks making sure that camera and microphone are on. Students are able to digitally raise their hand if assistance is needed. This alerts the proctors on the ADAM platform. The proctors are able to send a chat message to the student or call the student. The proctor can also broadcast messages to the entire group of students testing remotely. Proctors are able to see the students through their cameras. The students are able to see a proctor, but students 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.

8.1.4 Operational Monitoring

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. The types of monitoring that 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 (launching a test, logging into the administrative site, viewing reports, etc.) 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

8.2 Production System Testing

8.2.1 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.

8.2.2 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 our 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.

8.2.3 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.

8.2.4 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.

8.2.5 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.

8.2.6 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.

8.2.7 User Acceptance Testing

One of the testing steps includes the user acceptance test, which 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:

1. Create release notes for all new or modified functionality.
2. Provide updated training and user documentation.
3. Review checklist and ask questions.
4. Provide user IDs and passwords to allow users to run tests on code along with associated documentation assisting users on the process and procedures.
5. Meet with users and share results to jointly establish appropriate action plans.

8.3 Reporting

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.

8.4 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, to include test construction, calibration, equating, scaling, field test analysis, data review, item bank creation and management, standard setting, and technical reporting, can be found in the *Utah Aspire Plus 2018–2019 Technical Report* (available under Reporting Resources at https://utah.mypearsonsupport.com/admin-resources.html).

9 Validity

The *Standards for Educational and Psychological Testing*, issued jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014), reports:

> 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 is not to validate the test itself but to validate interpretations of the test scores for specific uses. In that sense, then, test validation is not quantifiable but an ongoing process of evidence gathering beginning at initial conceptualization and continuing throughout the full cycle of an assessment. Every component of an assessment provides evidence in support of its validity, including design, content specifications, item development, and psychometric characteristics.

For the Utah Aspire Plus assessment, operational test development and administration provided the chance to collect initial validity evidence based on test content and internal structure of the tests. Validation is the process of collecting evidence to support inferences from assessment results. As noted, 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.

9.1 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.

For the Utah Aspire Plus tests, design and blueprint specifications were developed in concert between USBE, Utah educators, and Pearson content experts well versed in the Utah Core Standards. As described in Chapter 2, 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. USBE has published the Utah Aspire Plus blueprints publicly (available at Administration Resources | UT (mypearsonsupport.com)).

As described in the respective SAGE and ACT Aspire technical manuals noted in Chapter 2, 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 depth of knowledge (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 noted, details of these procedures can be found in the respective technical manuals for SAGE and ACT Aspire referenced in Chapter 2 (see Volumes 2 and 4 of the 2016–2017 SAGE Technical Report and Chapter 2 of the ACT Aspire technical manual).

The process of developing the Utah Aspire Plus test design, development, and test construction is described in Chapter 2 of this report, and includes expert evaluation of the alignment of all content to the Utah Core Standards. 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.

As also described in Chapter 2, Utah educators created and recommended performance level descriptors for the Utah Aspire Plus tests, which 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. As described, content experts and stakeholders participated in standard setting in August 2019 for English, Reading, and Mathematics. In August 2022, similar meetings were conducted in support of the new Utah Aspire Plus SEEds Science tests. These committees set the cut scores that delineate the four overall levels of achievement on the Utah Aspire Plus tests. Evidence of these activities is presented in the context of student performance on the Utah Aspire Plus tests described in Chapter 4.

9.2 Evidence Based on Cognitive Process

Content comprising the Utah Aspire Plus assessments is specified by standard as well as DOK levels. “Depth of knowledge” (DOK), or cognitive complexity, refers to the cognitive demand associated with interacting with a given item/task. *Levels* of cognitive demand generally focus on the type and level of thinking and reasoning required to answer a given question correctly or earn the most points. For Utah Aspire Plus content, Webb’s definitions of levels of cognitive demand (Webb, 2002) were used to define the 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. In Section 2.3.4 of that report, it is noted 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. Further, in Section 2.4, they 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 as well as across content area.

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: 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.

9.3 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, APA, and the NCME, 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 English, reading, math, 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. It should be noted that, 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. The scree plots for the principal component analyses for each subject and grade are provided in Appendix K.

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 and Sörbom, 1993), and adjusted goodness-of-fit (AGFI: Tabachnick and Fidell, 2007) are also sensitive to sample size, which has led to researchers reporting them along with other fit indices (Hooper, Coughlan, and Mullen, 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. The estimates of these indices are presented in Table 9.1.

Table 9.1. Model Fit Indices for Confirmatory Factor Analyses

Subject	Grade	SRMR	RMSEA	RMSEA 90% Lower CL	RMSEA 90% Upper CL
English	9	0.0298	0.0326	0.0324	0.0328
	10	0.0306	0.0350	0.0347	0.0352
Reading	9	0.0187	0.0217	0.0214	0.0221
	10	0.0223	0.0272	0.0269	0.0276
Mathematics	9	0.0272	0.0293	0.0290	0.0296
	10	0.0241	0.0271	0.0268	0.0274
Science	9	0.0183	0.0235	0.0230	0.0240
	10	0.0231	0.0300	0.0295	0.0305

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 7.4 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 Chapter 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 L.

9.3.1 Reliability

Additionally, the reliability analyses presented in Chapter 6 of this technical report 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.

9.4 Evidence Based on Different Student Populations

In addition, internal structure evidence should 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 (as described in Chapter 2). 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 5.1.3 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.

9.5 Summary

As noted, 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.

10 References

ACT Aspire. (2017). *Summative Technical Manual*. Version 3. Iowa City, IA: ACT.

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. Joint Technical Committee. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

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). Reading, MA: Addison-Wesley.

Byrne, B. M. (1998). Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming. Mahwah, NJ: Lawrence Erlbaum Associates.

Chien, M. and Shin, D. (2012). *IRT Score Estimation Program*, V1.3 [computer program]. Iowa City, IA: 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). Hillsdale, NJ: Lawrence Erlbaum.

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). Hillsdale, NJ: 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). Hillsdale, NJ: 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*. Chicago, IL: Scientific Software International Inc.

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

Kim, S. and Kolen, M. (2004). STUIRT [computer program]. Iowa City, IA: 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*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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. 2012. *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

Next Generation Science Standards (NGSS Lead States. 2013. *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press) http://www.nextgenscience.org

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.

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

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). New York: 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 for English Grade 9

Reporting Category: Standard	Multiple Choice			Technology Enhanced		
	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
Production of Writing: W.9-10.4	1	0	8	1	0	0
Production of Writing: W.9-10.5	0	0	0	0	1	0
Knowledge of Language: L.9-10.3	0	0	2	0	0	0
Knowledge of Language: L.9-10.4a	2	0	0	0	0	0
Conventions of Standard English: L.9-10.1	5	3	1	0	0	0
Conventions of Standard English: L.9-10.1a	0	0	0	2	1	0
Conventions of Standard English: L.9-10.1b	0	0	0	1	2	0
Conventions of Standard English: L.9-10.2	5	0	0	0	0	0
Conventions of Standard English: L.9-10.2a	1	0	0	2	0	0
Conventions of Standard English: L.9-10.2b	1	0	0	1	0	0
Conventions of Standard English: L.9-10.2c	0	0	0	2	0	0
Conventions of Standard English: L.9-10.5b	0	0	0	1	1	0
Conventions of Standard English: L.9-10.6	0	0	1	1	0	0
Total						46

Table A.2. Test-Level Reporting Categories and Standards for English Grade 10

Reporting Category: Standard	Multiple Choice			Technology Enhanced		
	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
Production of Writing: W.9-10.4	0	1	8	0	1	0
Knowledge of Language: L.9-10.3	0	0	5	0	0	0
Knowledge of Language: L.9-10.4b	0	0	0	0	1	0
Conventions of Standard English: L.9-10.1	7	1	0	0	0	0
Conventions of Standard English: L.9-10.1a	0	0	0	2	0	0
Conventions of Standard English: L.9-10.1b	0	1	0	2	0	0
Conventions of Standard English: L.9-10.2	4	0	0	0	0	0
Conventions of Standard English: L.9-10.2a	1	0	0	3	0	0
Conventions of Standard English: L.9-10.2b	0	0	0	1	0	0
Conventions of Standard English: L.9-10.2c	0	0	0	1	0	0
Conventions of Standard English: L.9-10.5a	0	0	0	0	1	0
Conventions of Standard English: L.9-10.5b	0	1	0	1	2	0
Total						44

Table A.3. Test-Level Reporting Categories and Standards for Reading Grade 9

Reporting Category: Standard	Multiple Choice			Technology Enhanced			Evidence-Based Selected Response		
	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
Key Ideas: RI.9-10.1	2	0	0	0	1	0	0	0	0
Key Ideas: RI.9-10.2	0	3	0	0	0	1	0	0	1
Key Ideas: RL.9-10.1	0	2	0	0	0	0	0	0	0
Key Ideas: RL.9-10.2	0	1	0	0	0	0	0	0	1
Key Ideas: RL.9-10.3	0	3	0	0	0	0	0	0	0
Key Ideas: RL.9-10.7	0	0	0	0	0	1	0	0	0
Craft and Structure: RI.9-10.4	1	1	0	0	0	0	0	0	1
Craft and Structure: RI.9-10.6	0	0	1	0	1	0	0	0	1
Craft and Structure: RL.9-10.4	1	0	1	0	0	0	0	0	1
Craft and Structure: RL.9-10.5	0	1	1	0	0	0	0	0	0
Craft and Structure: RL.9-10.6	0	0	1	0	1	0	0	0	0
Integration of Knowledge and Ideas: RI.9-10.5	0	1	0	0	0	0	0	0	0
Integration of Knowledge and Ideas: RI.9-10.8	1	2	1	0	1	0	0	0	0
Total									35

Table A.4. Test-Level Reporting Categories and Standards for Reading Grade 10

Reporting Category: Standard	Multiple Choice			Technology Enhanced			Evidence-Based Selected Response		
	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
Key Ideas: RI.9-10.1	2	2	1	0	0	0	0	0	0
Key Ideas: RI.9-10.2	0	1	1	0	0	0	0	0	0
Key Ideas: RI.9-10.3	0	0	0	0	2	0	0	0	0
Key Ideas: RL.9-10.1	0	2	0	0	0	0	0	0	1
Key Ideas: RL.9-10.2	0	0	0	1	0	0	0	0	1
Key Ideas: RL.9-10.3	0	2	0	0	0	0	0	0	0
Craft and Structure: L.9-10.4a	0	1	0	0	0	0	0	0	0
Craft and Structure: RI.9-10.4	0	2	0	0	0	1	0	1	0
Craft and Structure: RI.9-10.5	0	0	1	0	0	0	0	0	1
Craft and Structure: RI.9-10.6	0	0	2	0	0	0	0	0	0
Craft and Structure: RL.9-10.3	0	0	0	0	1	0	0	0	0
Craft and Structure: RL.9-10.4	1	0	1	0	0	0	0	0	0
Craft and Structure: RL.9-10.5	0	0	1	0	0	0	0	0	1
Integration of Knowledge and Ideas: RI.9-10.7	0	1	0	0	0	0	0	0	0
Integration of Knowledge and Ideas: RI.9-10.8	1	2	0	0	0	0	0	0	0
Integration of Knowledge and Ideas: RL.9-10.7	0	0	1	0	0	0	0	0	0
Total									35

Table A.5. Test-Level Reporting Categories and Standards for Mathematics Grade 9

Reporting Category: Standard	Multiple Choice			Technology Enhanced		
	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
Algebra: MI.A.CED.1	0	1	0	0	0	1
Algebra: MI.A.CED.3	0	0	0	0	2	0
Algebra: MI.A.REI.1	1	0	0	0	0	0
Algebra: MI.A.REI.3b	0	1	0	0	0	0
Algebra: MI.A.REI.6	0	0	0	0	0	1
Algebra: MI.A.REI.10	1	0	0	0	0	0
Algebra: MI.A.REI.11	0	0	0	0	0	1
Algebra: MI.A.REI.12	0	1	0	0	0	0
Functions: MI.F.IF.2	1	0	0	0	0	0
Functions: MI.F.IF.4	1	0	0	0	0	1
Functions: MI.F.IF.7a	1	0	0	0	0	0
Functions: MI.F.IF.7e	0	1	0	0	0	0
Functions: MI.F.LE.1b	0	1	0	0	0	1
Functions: MI.F.LE.2	1	1	0	0	0	0
Functions: MI.F.LE.3	0	1	0	0	0	0
Functions: MI.F.LE.5	0	1	0	0	0	0
Geometry: MI.G.CO.1	1	0	0	0	0	0
Geometry: MI.G.CO.2	0	1	0	0	0	0
Geometry: MI.G.CO.4	0	0	1	0	0	0
Geometry: MI.G.CO.5	1	1	0	0	0	0
Geometry: MI.G.CO.6	1	0	0	0	0	0
Geometry: MI.G.CO.7	0	0	0	0	1	0
Geometry: MI.G.CO.12	0	0	0	0	0	1
Geometry: MI.G.GPE.5	0	1	0	0	0	0
Geometry: MI.G.GPE.7	0	1	0	0	0	0
Statistics and Probability: MI.S.ID.1	1	0	0	0	0	1
Statistics and Probability: MI.S.ID.2	1	0	0	0	0	0
Statistics and Probability: MI.S.ID.3	0	1	1	0	0	0
Statistics and Probability: MI.S.ID.6a	0	1	0	0	0	0
Statistics and Probability: MI.S.ID.7	1	1	0	0	0	0
Statistics and Probability: MI.S.ID.8	0	1	0	0	0	0
Total						40

Table A.6. Test-Level Reporting Categories and Standards for Mathematics Grade 10

Reporting Category: Standard	Multiple Choice			Technology Enhanced		
	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
Number and Quantity: MII.N.RN.2	0	1	0	0	0	0
Number and Quantity: MII.N.RN.3	1	0	0	0	1	0
Algebra: MII.A.APR.1	0	1	0	0	0	0
Algebra: MII.A.CED.1	0	1	1	0	0	0
Algebra: MII.A.CED.2	1	0	0	0	0	0
Algebra: MII.A.CED.4	0	1	0	0	0	0
Algebra: MII.A.REI.4a	0	1	0	0	0	0
Algebra: MII.A.REI.4b	0	0	1	0	0	0
Algebra: MII.A.REI.7	0	0	0	0	0	1
Algebra: MII.A.SSE.3a	0	1	0	0	0	0
Algebra: MII.A.SSE.3b	0	1	0	0	0	0
Algebra: MII.A.SSE.3c	1	0	0	0	0	0
Functions: MII.F.BF.1a	1	0	0	0	0	1
Functions: MII.F.BF.1b	0	1	0	0	0	0
Functions: MII.F.BF.3	1	0	0	0	0	0
Functions: MII.F.IF.4	0	1	1	0	0	0
Functions: MII.F.IF.5	1	0	0	0	0	0
Functions: MII.F.IF.7b	1	0	0	0	0	0
Functions: MII.F.IF.8a	0	1	0	0	0	0
Functions: MII.F.IF.8b	0	0	1	0	0	0
Functions: MII.F.LE.3	0	0	0	0	1	0
Geometry: MII.G.C.3	0	1	0	0	0	0
Geometry: MII.G.C.4	0	0	0	0	0	1
Geometry: MII.G.CO.9	1	0	0	0	0	0
Geometry: MII.G.CO.10	0	1	1	0	0	0
Geometry: MII.G.GPE.4	0	1	0	0	0	0
Geometry: MII.G.GPE.6	1	0	0	0	0	0
Geometry: MII.G.SRT.1a	0	0	0	0	1	0
Geometry: MII.G.SRT.2	0	0	1	0	0	0
Geometry: MII.G.SRT.5	0	1	0	0	0	0
Geometry: MII.G.SRT.6	1	0	0	0	0	0
Statistics and Probability: MII.S.CP.1	1	0	0	0	0	0
Statistics and Probability: MII.S.CP.4	0	0	0	0	0	1
Statistics and Probability: MII.S.CP.5	0	1	0	0	0	0
Statistics and Probability: MII.S.ID.5	0	1	0	0	0	0
Total						40

Appendix B: Student Testing Time

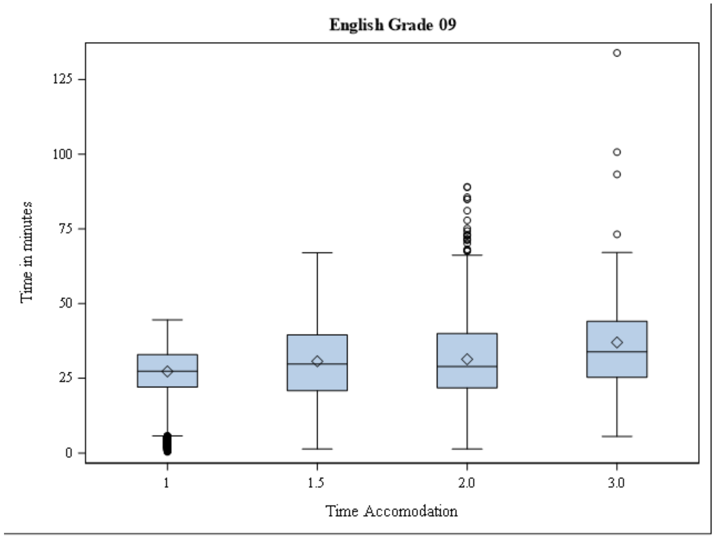

Figure B.1. English Grade 9 Student Testing Time

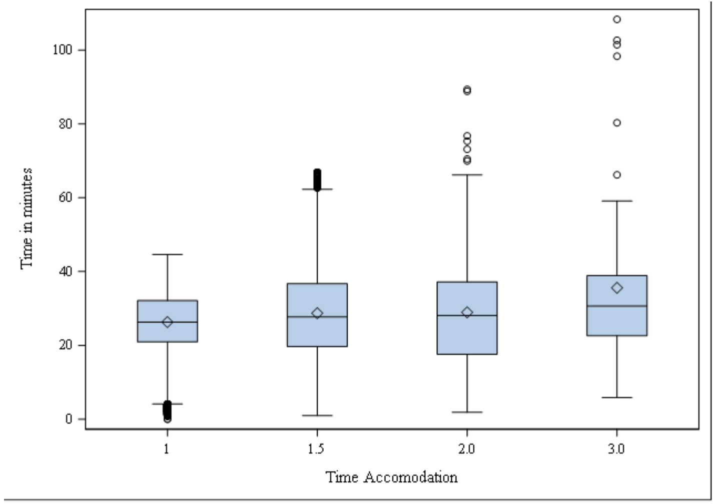

Figure B.2. English Grade 10 Student Testing Time

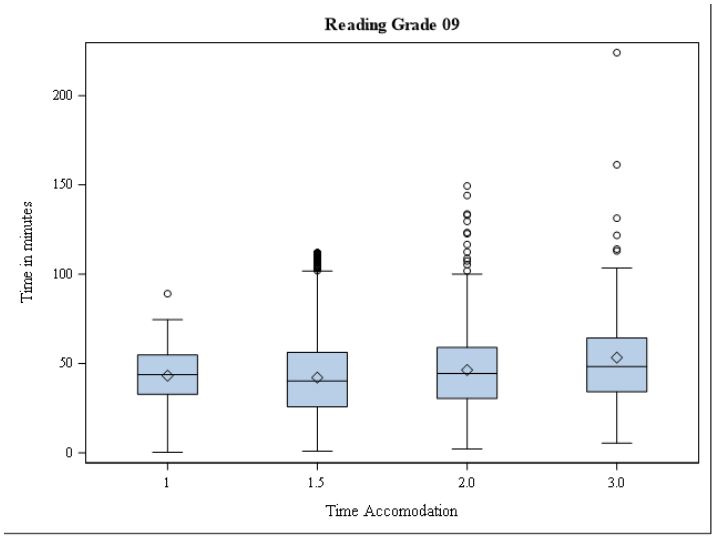

Figure B.3. Reading Grade 9 Student Testing Time

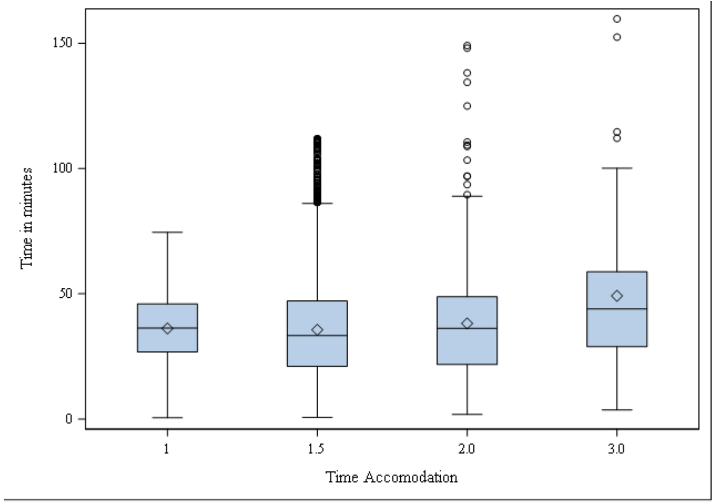

Figure B.4. Reading Grade 10 Student Testing Time

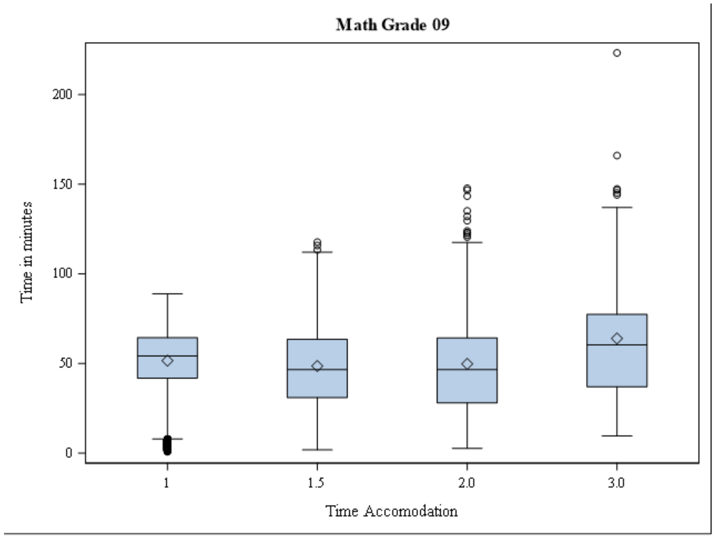

Figure B.5. Mathematics Grade 9 Student Testing Time

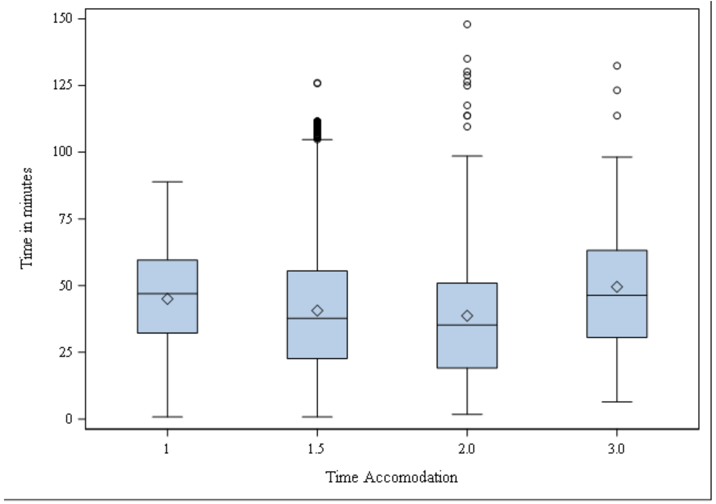

Figure B.6. Mathematics Grade 10 Student Testing Time

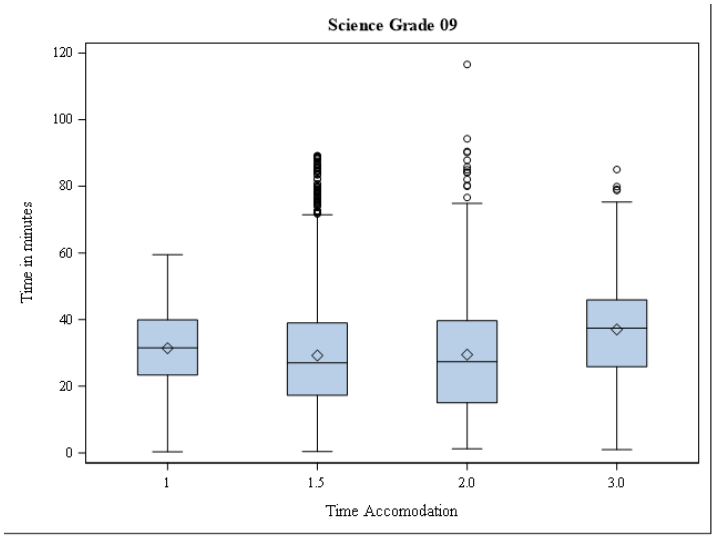

Figure B.7. Science Grade 9 Student Testing Time

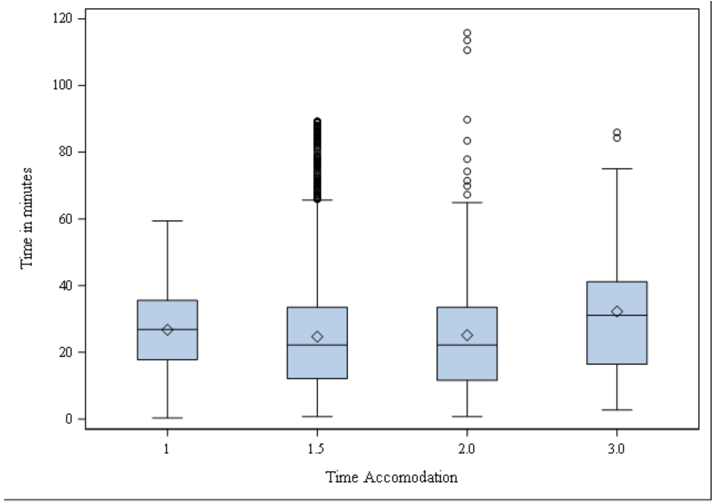

Figure B.8. Science Grade 10 Student Testing Time

Appendix C: Item Statistics Summaries

Table C.1. Item Mean for One-Point Items

Subject	Grade	N	*p*<0.30	0.30≤*p*<0.55	0.55≤*p*<0.75	0.75≤*p*<0.95	*p*≥0.95	Mean *p*
English	9	42	2	13	14	13	0	0.60
	10	38	2	9	16	11	0	0.63
Reading	9	27	4	12	10	1	0	0.51
	10	28	1	10	13	4	0	0.58
Mathematics	9	40	8	19	11	2	0	0.45
	10	40	7	27	4	2	0	0.42
Science	9	18	0	7	11	0	0	0.57
	10	18	2	11	5	0	0	0.44

Table C.2. Item Mean for Two-Point Items

Subject	Grade	N	Mean	Min	Max
English	9	4	1.21	0.82	1.55
	10	6	1.06	0.44	1.40
Reading	9	8	0.74	0.42	1.42
	10	7	1.12	0.75	1.41
Science	9	5	0.77	0.42	1.05
	10	5	0.91	0.47	1.23

Note: There were no 2-point mathematics items in Spring 2024.

Table C.3. Item Total Correlation for One-Point Items

Subject	Grade	N	*r*<0.20	0.20≤*r*<0.40	0.40≤*r*<0.60	0.60≤*r*<0.80	*r*≥0.80	Median ITC
English	9	42	5	14	23	0	0	0.41
	10	38	0	13	23	2	0	0.47
Reading	9	27	3	9	15	0	0	0.40
	10	28	1	8	18	1	0	0.43
Mathematics	9	40	0	14	25	1	0	0.44
	10	40	1	18	21	0	0	0.41
Science	9	18	0	13	5	0	0	0.36
	10	18	1	7	10	0	0	0.41

Note: ITC=Item Total Correlation

Table C.4. Item Total Correlation for Two-Point Items

Subject	Grade	N	Median *r*	Min *r*	Max *r*
English	9	4	0.47	0.45	0.63
	10	6	0.42	0.32	0.60
Reading	9	8	0.46	0.29	0.63
	10	7	0.64	0.35	0.68
Science	9	5	0.48	0.35	0.57
	10	5	0.23	0.18	0.67

Note: There were no 2-point mathematics items in Spring 2024.

Table C.5. Differential Item Functioning

			DIF Categories				
				Moderate DIF		Substantial DIF	
Subject	Grade	Subgroups	Negligible DIF	Focal	Reference	Focal	Reference
English	9	Male-Female	46	0	0	0	0
		White-Black	46	0	0	0	0
		White-Hispanic	46	0	0	0	0
	10	Male-Female	43	0	1	0	0
		White-Black	44	0	0	0	0
		White-Hispanic	43	0	1	0	0
Reading	9	Male-Female	35	0	0	0	0
		White-Black	34	0	1	0	0
		White-Hispanic	35	0	0	0	0
	10	Male-Female	32	2	1	0	0
		White-Black	33	0	2	0	0
		White-Hispanic	34	0	1	0	0
Mathematics	9	Male-Female	37	0	3	0	0
		White-Black	37	0	3	0	0
		White-Hispanic	40	0	0	0	0
	10	Male-Female	40	0	0	0	0
		White-Black	37	0	3	0	0
		White-Hispanic	39	0	1	0	0
Science	9	Male-Female	22	0	1	0	0
		White-Black	23	0	0	0	0
		White-Hispanic	23	0	0	0	0
	10	Male-Female	21	0	1	0	1
		White-Black	23	0	0	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.

Appendix D: Reliability and Standard Error by Subgroup

Table D.1. English Grade 9 Test Reliability

	Test Group	N	Alpha	SEM	Production of Writing	Knowledge of Language	Conventions of Standard English
All	Students Tested	45,391	0.90	9.21	0.72	0.57	0.84
Sex	Female	21,705	0.89	9.19	0.71	0.53	0.83
	Male	23,629	0.90	9.26	0.73	0.60	0.84
	Unknown	57	0.93	9.51	0.73	0.70	0.90
Ethnicity	Hispanic or Latino Ethnicity	8,943	0.88	9.28	0.69	0.55	0.82
	Asian	764	0.91	9.59	0.74	0.59	0.85
	Native Hawaiian or Other Pacific Islander	584	0.87	9.38	0.66	0.52	0.79
	Black or African American	590	0.89	9.63	0.66	0.60	0.83
	American Indian or Alaska Native	422	0.86	9.34	0.67	0.53	0.78
	White	32,473	0.89	9.18	0.71	0.54	0.83
	Other	1,615	0.90	9.19	0.73	0.56	0.84
Limited English Proficiency	No	41,574	0.89	9.18	0.71	0.54	0.83
	Yes	3,817	0.80	10.00	0.53	0.46	0.70
Economic Disadvantage	No	33,168	0.89	9.20	0.71	0.55	0.83
	Yes	12,223	0.89	9.28	0.70	0.57	0.83
Special Education	No	40,876	0.89	9.19	0.71	0.54	0.83
	Yes	4,515	0.84	9.81	0.60	0.50	0.74

Table D.2. English Grade 10 Test Reliability

	Test Group	N	Alpha	SEM	Production of Writing	Knowledge of Language	Conventions of Standard English
All	Students Tested	43,431	0.92	8.28	0.77	0.58	0.87
Sex	Female	20,637	0.91	8.32	0.76	0.56	0.86
	Male	22,752	0.92	8.24	0.77	0.58	0.88
	Unknown	42	0.94	7.98	0.81	0.65	0.90
Ethnicity	Hispanic or Latino Ethnicity	8,476	0.90	8.13	0.73	0.54	0.85
	Asian	734	0.92	8.40	0.78	0.57	0.87
	Native Hawaiian or Other Pacific Islander	603	0.88	7.78	0.70	0.51	0.81
	Black or African American	584	0.90	8.26	0.72	0.52	0.84
	American Indian or Alaska Native	411	0.88	7.90	0.68	0.46	0.82
	White	31,077	0.91	8.33	0.76	0.56	0.86
	Other	1,546	0.91	8.26	0.76	0.56	0.87
Limited English Proficiency	No	39,926	0.91	8.31	0.75	0.56	0.86
	Yes	3,505	0.83	8.38	0.56	0.40	0.76
Economic Disadvantage	No	32,405	0.91	8.33	0.76	0.57	0.87
	Yes	11,026	0.91	8.16	0.75	0.56	0.86
Special Education	No	39,328	0.91	8.31	0.75	0.56	0.86
	Yes	4,103	0.87	8.29	0.65	0.46	0.80

Table D.3. Reading Grade 9 Test Reliability

	Test Group	N	Alpha	SEM	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
All	Students Tested	45,559	0.87	10.64	0.79	0.71	0.33
Sex	Female	21,844	0.86	10.47	0.78	0.70	0.32
	Male	23,661	0.87	10.77	0.79	0.72	0.34
	Unknown	54	0.90	11.20	0.82	0.78	0.42
Ethnicity	Hispanic or Latino Ethnicity	9,050	0.84	10.88	0.74	0.67	0.29
	Asian	770	0.87	10.65	0.79	0.72	0.36
	Native Hawaiian or Other Pacific Islander	588	0.82	10.67	0.71	0.64	0.19
	Black or African American	593	0.86	10.92	0.75	0.70	0.34
	American Indian or Alaska Native	423	0.79	10.57	0.68	0.60	0.16
	White	32,516	0.86	10.60	0.78	0.70	0.32
	Other	1,619	0.86	10.71	0.77	0.72	0.29
Limited English Proficiency	No	41,626	0.86	10.57	0.78	0.70	0.32
	Yes	3,933	0.69	12.19	0.53	0.48	0.12
Economic Disadvantage	No	33,239	0.86	10.57	0.78	0.70	0.32
	Yes	12,320	0.85	10.84	0.76	0.69	0.29
Special Education	No	41,026	0.86	10.56	0.78	0.70	0.32
	Yes	4,533	0.78	11.54	0.65	0.58	0.18

Table D.4. Reading Grade 10 Test Reliability

	Test Group	N	Alpha	SEM	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
All	Students Tested	43,594	0.90	9.27	0.83	0.75	0.46
Sex	Female	20,741	0.89	9.11	0.81	0.73	0.44
	Male	22,810	0.90	9.38	0.84	0.77	0.48
	Unknown	43	0.93	9.59	0.89	0.82	0.62
Ethnicity	Hispanic or Latino Ethnicity	8,544	0.88	9.25	0.82	0.72	0.38
	Asian	735	0.91	9.56	0.84	0.78	0.51
	Native Hawaiian or Other Pacific Islander	624	0.87	9.29	0.80	0.68	0.36
	Black or African American	590	0.87	9.46	0.81	0.69	0.36
	American Indian or Alaska Native	417	0.86	9.23	0.78	0.69	0.26
	White	31,129	0.89	9.25	0.82	0.74	0.46
	Other	1,555	0.89	9.30	0.83	0.75	0.44
Limited English Proficiency	No	40,039	0.89	9.24	0.82	0.74	0.46
	Yes	3,555	0.78	9.87	0.72	0.54	0.14
Economic Disadvantage	No	32,483	0.89	9.27	0.82	0.75	0.46
	Yes	11,111	0.89	9.25	0.83	0.74	0.41
Special Education	No	39,450	0.89	9.24	0.82	0.75	0.46
	Yes	4,144	0.84	9.62	0.76	0.64	0.28

Table D.5. Mathematics Grade 9 Test Reliability

	Test Group	N	Alpha	SEM	Algebra	Functions	Geometry	Statistics and Probability
All	Students Tested	43,674	0.91	9.47	0.75	0.70	0.74	0.68
Sex	Female	20,694	0.89	9.44	0.72	0.64	0.71	0.64
	Male	22,934	0.92	9.48	0.78	0.73	0.76	0.71
	Unknown	46	0.83	10.91	0.61	0.40	0.68	0.53
Ethnicity	Hispanic or Latino Ethnicity	8,477	0.86	11.62	0.65	0.56	0.66	0.56
	Asian	725	0.92	9.33	0.80	0.75	0.74	0.71
	Native Hawaiian or Other Pacific Islander	560	0.84	12.16	0.64	0.47	0.62	0.51
	Black or African American	555	0.85	12.24	0.64	0.52	0.66	0.54
	American Indian or Alaska Native	393	0.80	12.67	0.47	0.48	0.63	0.43
	White	31,417	0.91	8.89	0.75	0.70	0.72	0.67
	Other	1,547	0.91	9.55	0.77	0.71	0.74	0.68
Limited English Proficiency	No	39,958	0.91	9.13	0.75	0.70	0.72	0.67
	Yes	3,716	0.70	15.70	0.40	0.31	0.45	0.30
Economic Disadvantage	No	31,974	0.91	9.00	0.75	0.70	0.73	0.68
	Yes	11,700	0.88	10.93	0.69	0.61	0.70	0.61
Special Education	No	39,233	0.91	9.15	0.75	0.70	0.72	0.67
	Yes	4,441	0.80	13.12	0.54	0.43	0.59	0.45

Table D.6. Mathematics Grade 10 Test Reliability

	Test Group	N	Alpha	SEM	Number and Quantity	Algebra	Functions	Geometry	Statistics and Probability
All	Students Tested	42,840	0.89	11.16	0.55	0.61	0.70	0.74	0.43
Sex	Female	20,294	0.87	11.30	0.51	0.57	0.65	0.72	0.42
	Male	22,507	0.90	11.04	0.59	0.64	0.73	0.76	0.45
	Unknown	39	0.87	13.19	0.54	0.47	0.69	0.75	0.64
Ethnicity	Hispanic or Latino Ethnicity	8,332	0.81	14.17	0.45	0.44	0.52	0.62	0.36
	Asian	720	0.92	10.12	0.58	0.75	0.78	0.78	0.47
	Native Hawaiian or Other Pacific Islander	610	0.77	14.97	0.46	0.33	0.42	0.58	0.33
	Black or African American	577	0.80	15.34	0.45	0.42	0.48	0.60	0.33
	American Indian or Alaska Native	398	0.78	15.87	0.41	0.37	0.45	0.56	0.28
	White	30,685	0.89	10.49	0.55	0.62	0.71	0.74	0.42
	Other	1,518	0.89	10.74	0.56	0.63	0.72	0.73	0.41
Limited English Proficiency	No	39,349	0.89	10.74	0.54	0.62	0.70	0.74	0.42
	Yes	3,491	0.61	19.56	0.29	0.23	0.23	0.39	0.19
Economic Disadvantage	No	31,974	0.89	10.57	0.55	0.63	0.72	0.74	0.42
	Yes	10,866	0.84	13.36	0.50	0.48	0.57	0.67	0.40
Special Education	No	38,741	0.89	10.67	0.55	0.62	0.70	0.73	0.42
	Yes	4,099	0.69	17.94	0.34	0.27	0.34	0.46	0.22

Table D.7. Science Grade 9 Test Reliability

	Test Group	N	Alpha	SEM	Gathering & Investigating	Developing Models	Using Mathematical Thinking	Construct Explanations
All	Students Tested	45,542	0.83	13.99	0.45	0.49	0.63	0.64
Sex	Female	21,825	0.81	14.00	0.42	0.46	0.59	0.60
	Male	23,659	0.85	13.95	0.48	0.51	0.66	0.67
	Unknown	58	0.79	13.98	0.49	0.18	0.64	0.64
Ethnicity	Hispanic or Latino Ethnicity	9,053	0.78	14.80	0.36	0.42	0.57	0.54
	Asian	772	0.85	14.09	0.52	0.56	0.63	0.66
	Native Hawaiian or Other Pacific Islander	598	0.75	14.93	0.36	0.37	0.52	0.49
	Black or African American	597	0.76	15.52	0.30	0.49	0.48	0.53
	American Indian or Alaska Native	422	0.73	15.05	0.24	0.34	0.55	0.51
	White	32,478	0.83	13.78	0.46	0.48	0.61	0.64
	Other	1,622	0.83	14.18	0.43	0.49	0.61	0.66
Limited English Proficiency	No	41,611	0.83	13.85	0.45	0.48	0.61	0.64
	Yes	3,931	0.62	17.01	0.21	0.33	0.39	0.32
Economic Disadvantage	No	33,225	0.83	13.79	0.46	0.48	0.62	0.65
	Yes	12,317	0.80	14.68	0.39	0.45	0.59	0.58
Special Education	No	41,014	0.82	13.82	0.45	0.48	0.61	0.64
	Yes	4,528	0.72	16.03	0.27	0.34	0.52	0.46

Table D.8. Science Grade 10 Test Reliability

	Test Group	N	Alpha	SEM	Gathering & Investigating	Developing Models	Using Mathematical Thinking	Construct Explanations
All	Students Tested	43,491	0.82	14.05	0.56	0.44	0.60	0.60
Sex	Female	20,686	0.79	14.27	0.55	0.39	0.54	0.56
	Male	22,764	0.84	13.86	0.58	0.48	0.63	0.63
	Unknown	41	0.84	13.56	0.51	0.43	0.74	0.71
Ethnicity	Hispanic or Latino Ethnicity	8,542	0.72	16.49	0.45	0.35	0.46	0.44
	Asian	736	0.84	13.64	0.60	0.51	0.63	0.64
	Native Hawaiian or Other Pacific Islander	614	0.63	17.64	0.35	0.29	0.39	0.33
	Black or African American	596	0.69	17.83	0.43	0.29	0.33	0.36
	American Indian or Alaska Native	413	0.72	16.47	0.46	0.32	0.40	0.42
	White	31,050	0.82	13.55	0.56	0.44	0.61	0.61
	Other	1,540	0.83	13.79	0.57	0.45	0.59	0.60
Limited English Proficiency	No	39,921	0.82	13.78	0.56	0.43	0.60	0.60
	Yes	3,570	0.42	21.59	0.27	0.21	0.20	0.12
Economic Disadvantage	No	32,402	0.82	13.64	0.56	0.44	0.61	0.61
	Yes	11,089	0.77	15.61	0.52	0.40	0.50	0.51
Special Education	No	39,376	0.82	13.77	0.56	0.44	0.60	0.60
	Yes	4,115	0.62	18.77	0.40	0.30	0.32	0.29

Appendix E: Conditional Standard Error of Scale Scores

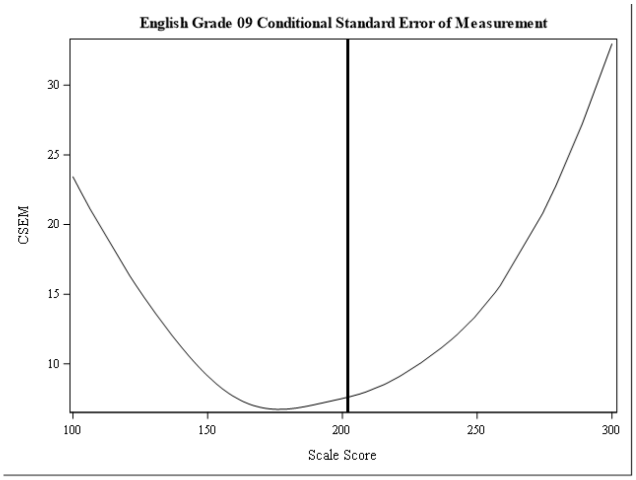

Figure E.1. English Grade 9 Conditional Standard Error of Scale Scores

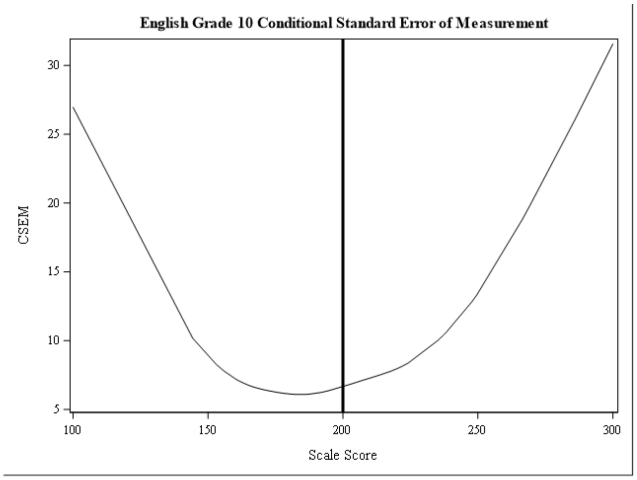

Figure E.2. English Grade 10 Conditional Standard Error of Scale Scores

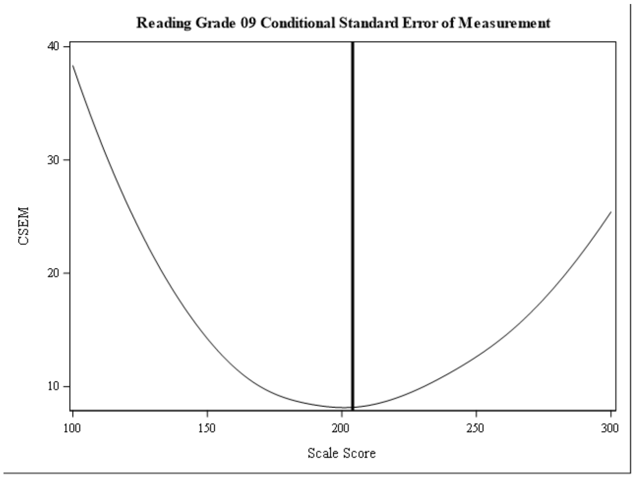

Figure E.3. Reading Grade 9 Conditional Standard Error of Scale Scores

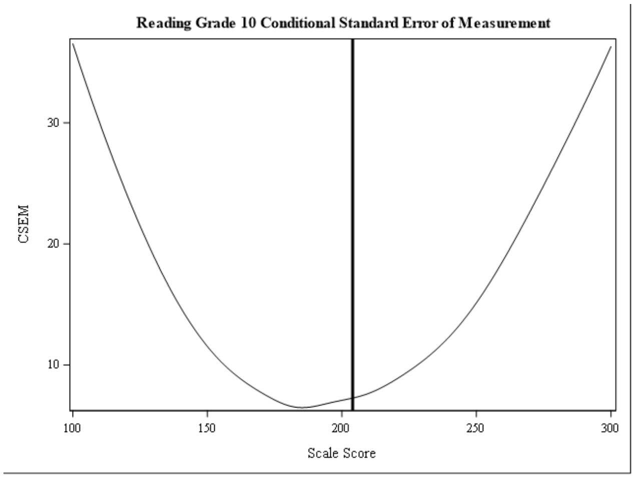

Figure E.4. Reading Grade 10 Conditional Standard Error of Scale Scores

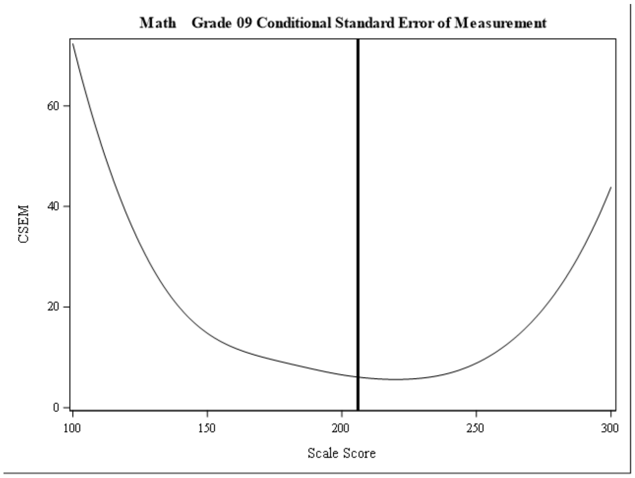

Figure E.5. Mathematics Grade 9 Conditional Standard Error of Scale Scores

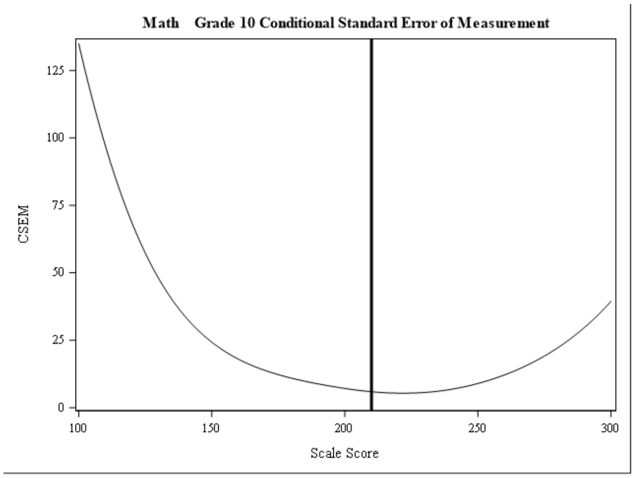

Figure E.6. Mathematics Grade 10 Conditional Standard Error of Scale Scores

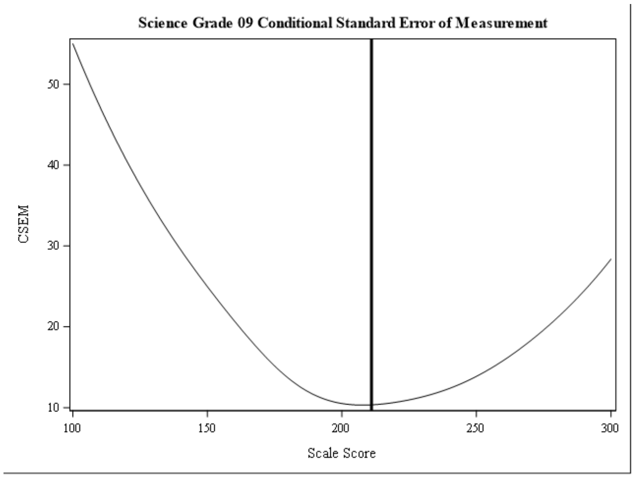

Figure E.7. Science Grade 9 Conditional Standard Error of Scale Scores

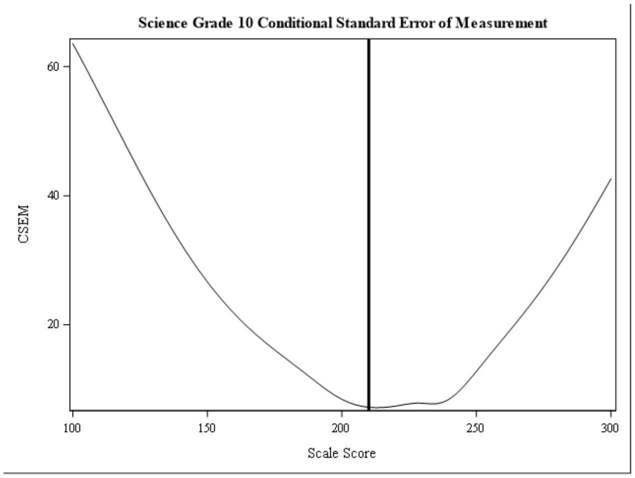

Figure E.8. Science Grade 10 Conditional Standard Error of Scale Scores

Appendix F: Accuracy and Consistency

Table F.1. Accuracy Classification for English Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.118	0.027	0.000	0.000	81.43
Approaching Proficient	0.035	0.378	0.054	0.000	
Proficient	0.000	0.047	0.291	0.015	
Highly Proficient	0.000	0.000	0.009	0.027	

Table F.2. Accuracy Classification at Proficient Cut Point for English Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.118	0.027	0.000	0.000	89.95
Approaching Proficient	0.035	0.378	0.054	0.000	
Proficient	0.000	0.047	0.291	0.015	
Highly Proficient	0.000	0.000	0.009	0.027	

Table F.3. Consistency Classification for English Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.112	0.048	0.000	0.000	73.55	0.593
Approaching Proficient	0.040	0.334	0.071	0.000		
Proficient	0.000	0.069	0.263	0.016		
Highly Proficient	0.000	0.000	0.020	0.026		

Table F.4. Accuracy Classification for English Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.077	0.016	0.000	0.000	84.64
Approaching Proficient	0.025	0.369	0.047	0.000	
Proficient	0.000	0.044	0.369	0.014	
Highly Proficient	0.000	0.000	0.008	0.031	

Table F.5. Accuracy Classification at Proficient Cut Point for English Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.077	0.016	0.000	0.000	90.94
Approaching Proficient	0.025	0.369	0.047	0.000	
Proficient	0.000	0.044	0.369	0.014	
Highly Proficient	0.000	0.000	0.008	0.031	

Table F.6. Consistency Classification for English Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.074	0.030	0.000	0.000	78.20	0.651
Approaching Proficient	0.028	0.336	0.065	0.000		
Proficient	0.000	0.063	0.343	0.015		
Highly Proficient	0.000	0.000	0.017	0.030		

Table F.7. Accuracy Classification for Reading Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.124	0.027	0.000	0.000	75.79
Approaching Proficient	0.045	0.355	0.065	0.000	
Proficient	0.000	0.049	0.205	0.033	
Highly Proficient	0.000	0.000	0.022	0.074	

Table F.8. Accuracy Classification at Proficient Cut Point for Reading Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.124	0.027	0.000	0.000	88.56
Approaching Proficient	0.045	0.355	0.065	0.000	
Proficient	0.000	0.049	0.205	0.033	
Highly Proficient	0.000	0.000	0.022	0.074	

Table F.9. Consistency Classification for Reading Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.118	0.051	0.001	0.000	66.34	0.511
Approaching Proficient	0.051	0.307	0.082	0.002		
Proficient	0.000	0.072	0.167	0.035		
Highly Proficient	0.000	0.002	0.041	0.071		

Table F.10. Accuracy Classification for Reading Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.206	0.032	0.000	0.000	78.27
Approaching Proficient	0.046	0.260	0.059	0.000	
Proficient	0.000	0.042	0.264	0.022	
Highly Proficient	0.000	0.000	0.017	0.054	

Table F.11. Accuracy Classification at Proficient Cut Point for Reading Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.206	0.032	0.000	0.000	89.98
Approaching Proficient	0.046	0.260	0.059	0.000	
Proficient	0.000	0.042	0.264	0.022	
Highly Proficient	0.000	0.000	0.017	0.054	

Table F.12. Consistency Classification for Reading Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.195	0.051	0.001	0.000	69.54	0.569
Approaching Proficient	0.056	0.220	0.076	0.000		
Proficient	0.001	0.061	0.229	0.024		
Highly Proficient	0.000	0.000	0.033	0.052		

Table F.13. Accuracy Classification for Mathematics Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.189	0.038	0.000	0.000	78.57
Approaching Proficient	0.037	0.336	0.059	0.000	
Proficient	0.000	0.042	0.208	0.018	
Highly Proficient	0.000	0.000	0.020	0.052	

Table F.14. Accuracy Classification at Proficient Cut Point for Mathematics Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.189	0.038	0.000	0.000	89.88
Approaching Proficient	0.037	0.336	0.059	0.000	
Proficient	0.000	0.042	0.208	0.018	
Highly Proficient	0.000	0.000	0.020	0.052	

Table F.15. Consistency Classification for Mathematics Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.180	0.060	0.001	0.000	69.60	0.562
Approaching Proficient	0.046	0.292	0.074	0.001		
Proficient	0.000	0.064	0.175	0.020		
Highly Proficient	0.000	0.001	0.037	0.050		

Table F.16. Accuracy Classification for Mathematics Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.300	0.061	0.000	0.000	75.42
Approaching Proficient	0.045	0.271	0.059	0.000	
Proficient	0.000	0.047	0.148	0.016	
Highly Proficient	0.000	0.000	0.018	0.035	

Table F.17. Accuracy Classification at Proficient Cut Point for Mathematics Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.300	0.061	0.000	0.000	89.29
Approaching Proficient	0.045	0.271	0.059	0.000	
Proficient	0.000	0.047	0.148	0.016	
Highly Proficient	0.000	0.000	0.018	0.035	

Table F.18. Consistency Classification for Mathematics Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.285	0.086	0.004	0.000	65.85	0.504
Approaching Proficient	0.057	0.219	0.068	0.002		
Proficient	0.002	0.069	0.120	0.016		
Highly Proficient	0.000	0.003	0.034	0.034		

Table F.19. Accuracy Classification for Science Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.242	0.054	0.003	0.000	67.07
Approaching Proficient	0.058	0.160	0.072	0.002	
Proficient	0.003	0.059	0.169	0.044	
Highly Proficient	0.000	0.001	0.032	0.099	

Table F.20. Accuracy Classification at Proficient Cut Point for Science Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.242	0.054	0.003	0.000	85.90
Approaching Proficient	0.058	0.160	0.072	0.002	
Proficient	0.003	0.059	0.169	0.044	
Highly Proficient	0.000	0.001	0.032	0.099	

Table F.21. Consistency Classification for Science Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.227	0.075	0.016	0.001	57.00	0.416
Approaching Proficient	0.063	0.119	0.075	0.008		
Proficient	0.013	0.071	0.129	0.043		
Highly Proficient	0.001	0.009	0.056	0.095		

Table F.22. Accuracy Classification for Science Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.272	0.073	0.005	0.000	67.79
Approaching Proficient	0.056	0.190	0.074	0.001	
Proficient	0.003	0.066	0.180	0.025	
Highly Proficient	0.000	0.000	0.020	0.037	

Table F.23. Accuracy Classification at Proficient Cut Point for Science Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.272	0.073	0.005	0.000	85.13
Approaching Proficient	0.056	0.190	0.074	0.001	
Proficient	0.003	0.066	0.180	0.025	
Highly Proficient	0.000	0.000	0.020	0.037	

Table F.24. Consistency Classification for Science Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.255	0.098	0.018	0.000	57.30	0.395
Approaching Proficient	0.062	0.139	0.075	0.003		
Proficient	0.013	0.086	0.143	0.023		
Highly Proficient	0.000	0.006	0.042	0.036		

Appendix G: Common Item Scatterplots for 2024 Anchor Items

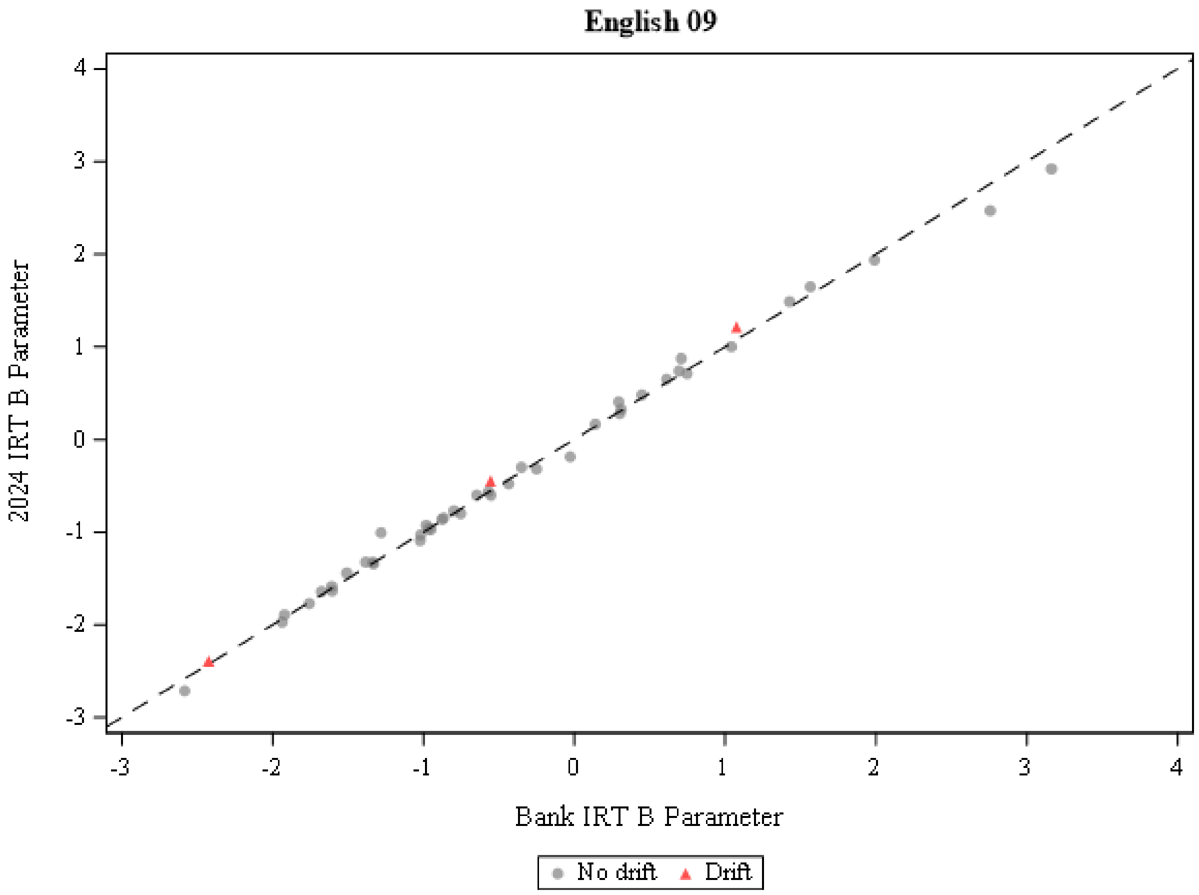

Figure G.1. English Grade 9 IRT B Parameters for Operational Items

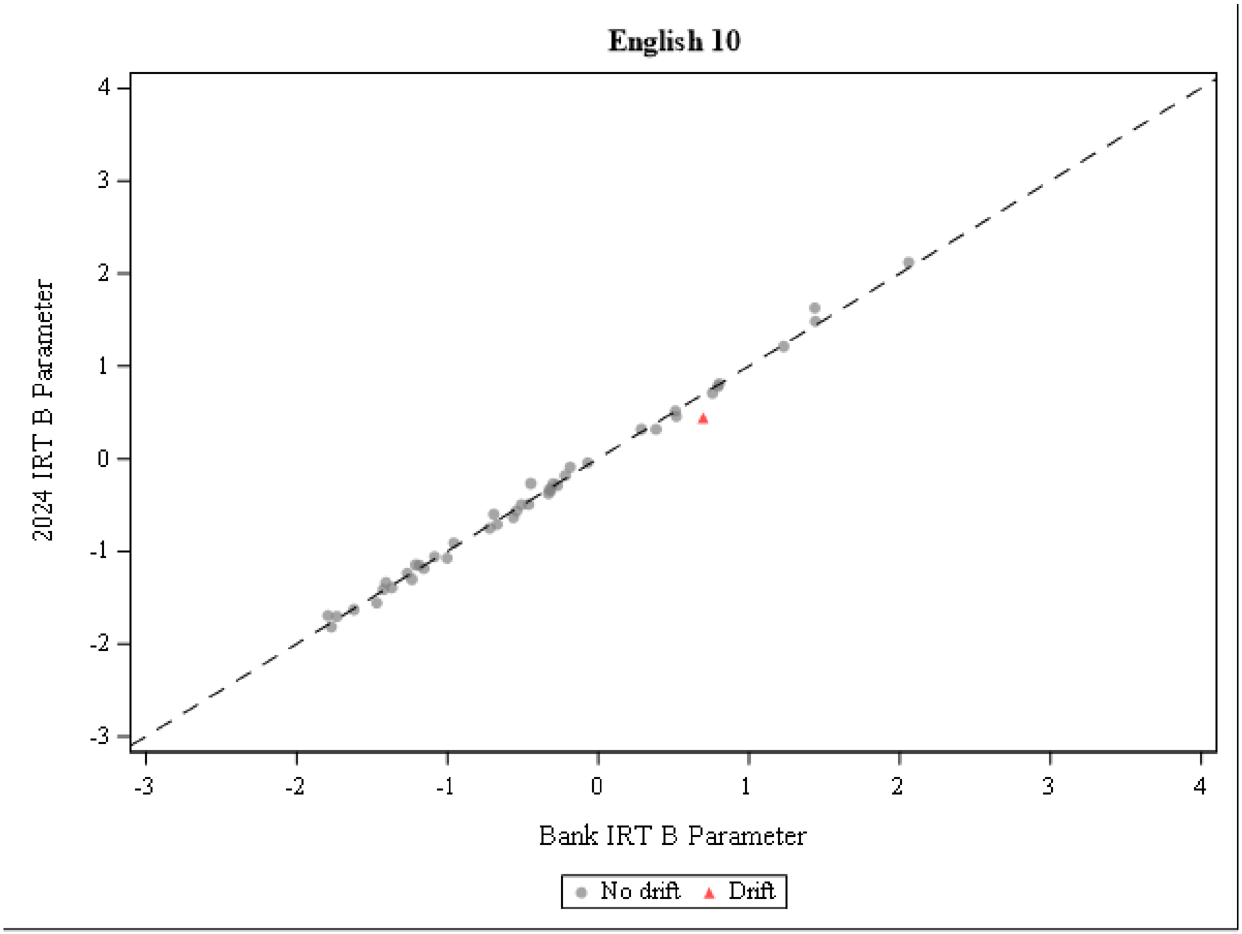

Figure G.2. English Grade 10 IRT B Parameters for Operational Items

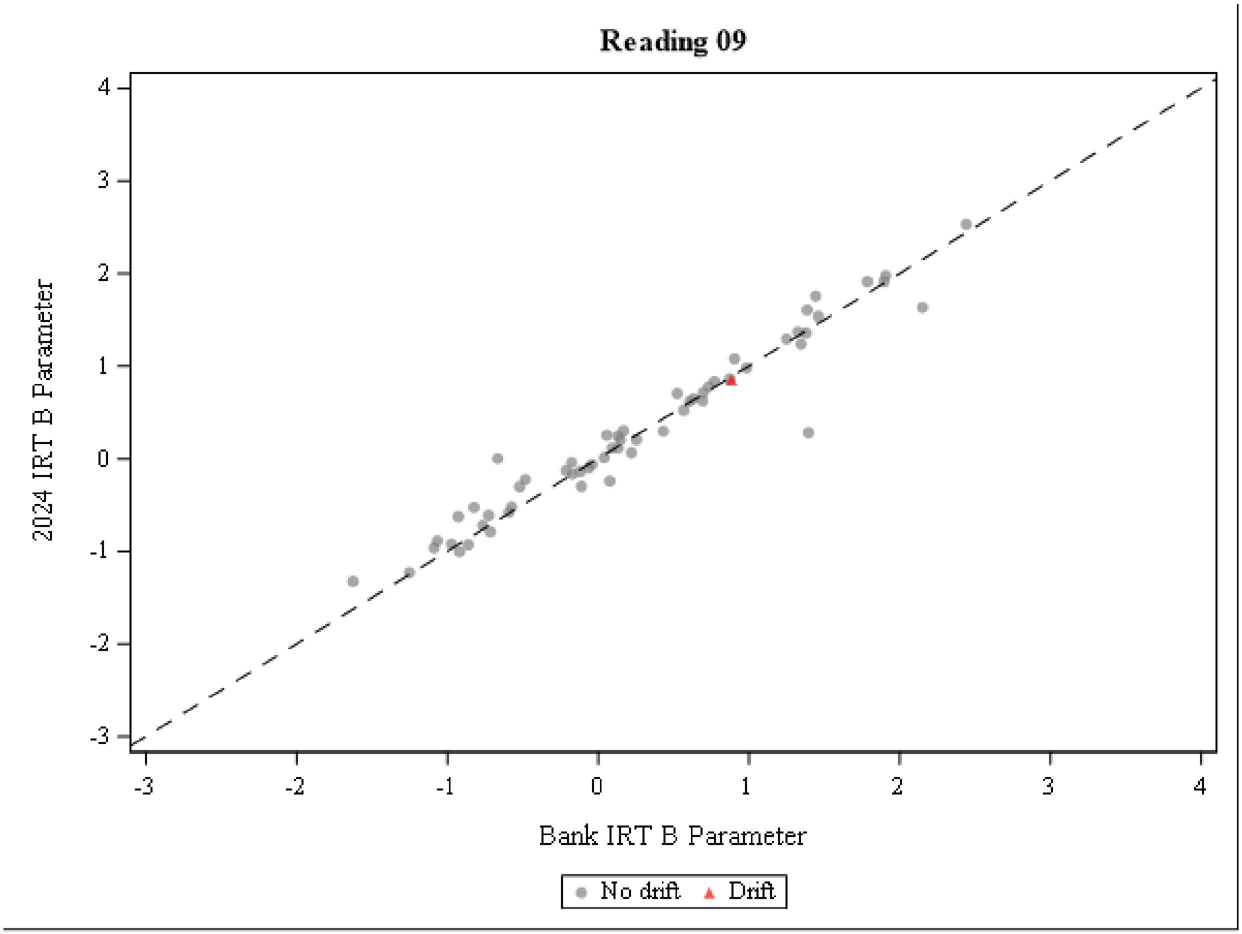

Figure G.3. Reading Grade 9 IRT B Parameters for Operational Items

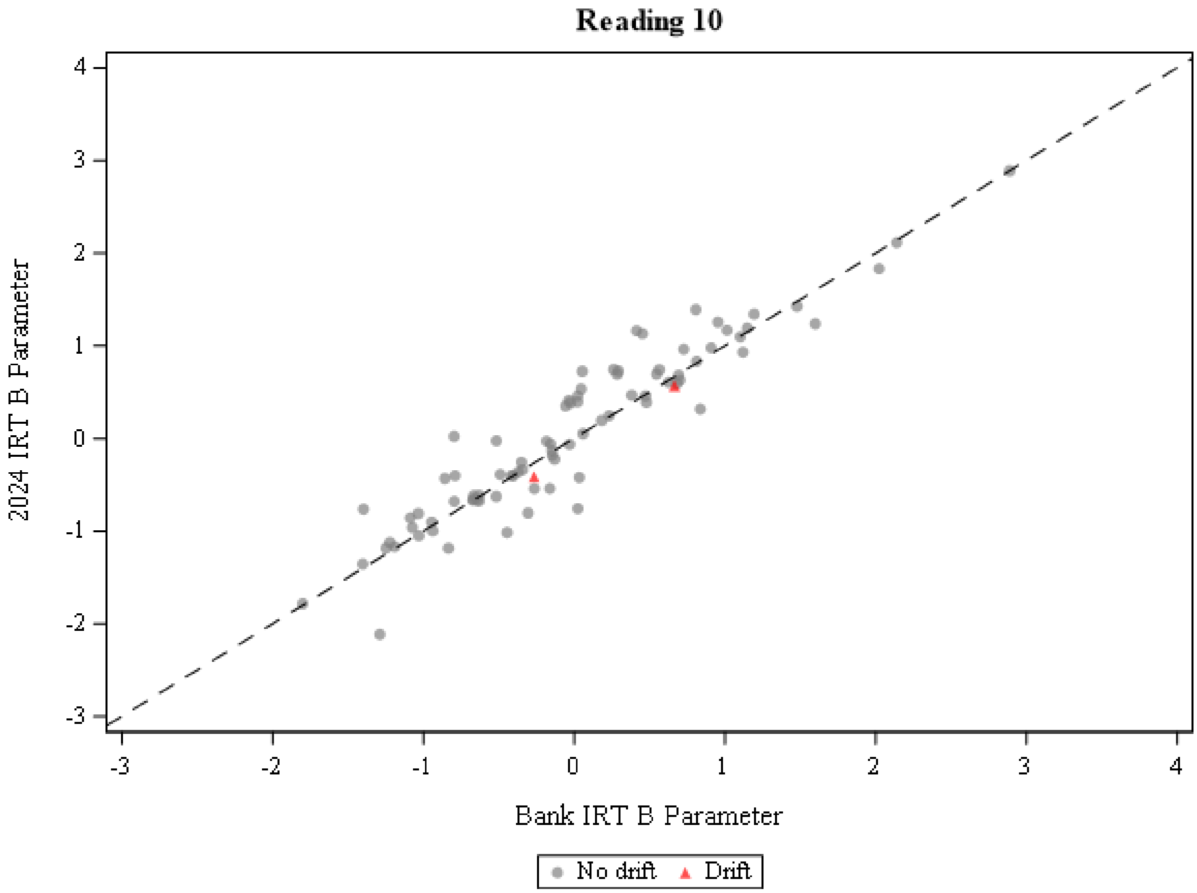

Figure G.4. Reading Grade 10 IRT B Parameters for Operational Items

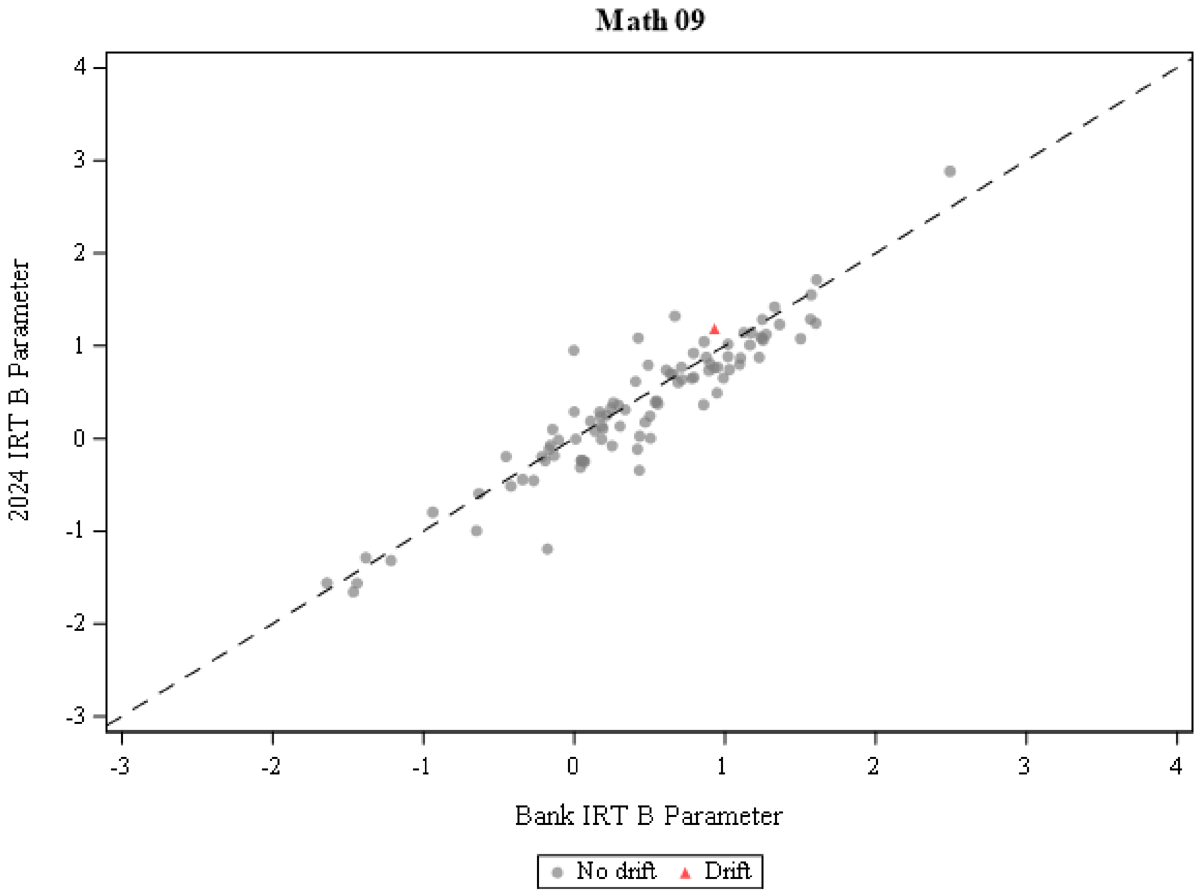

Figure G.5. Mathematics Grade 9 IRT B Parameters for Operational Items

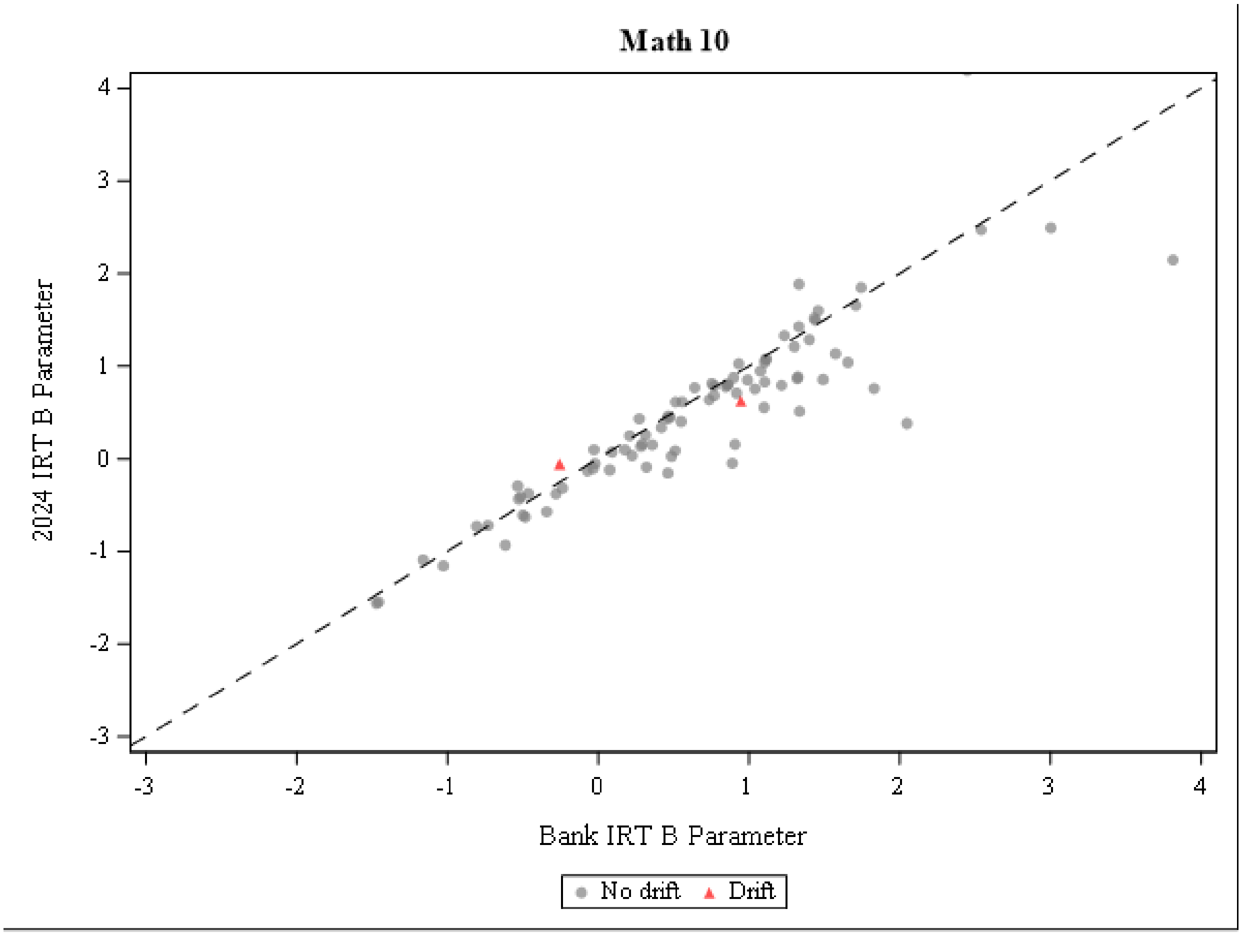

Figure G.6. Mathematics Grade 10 IRT B Parameters for Operational Items

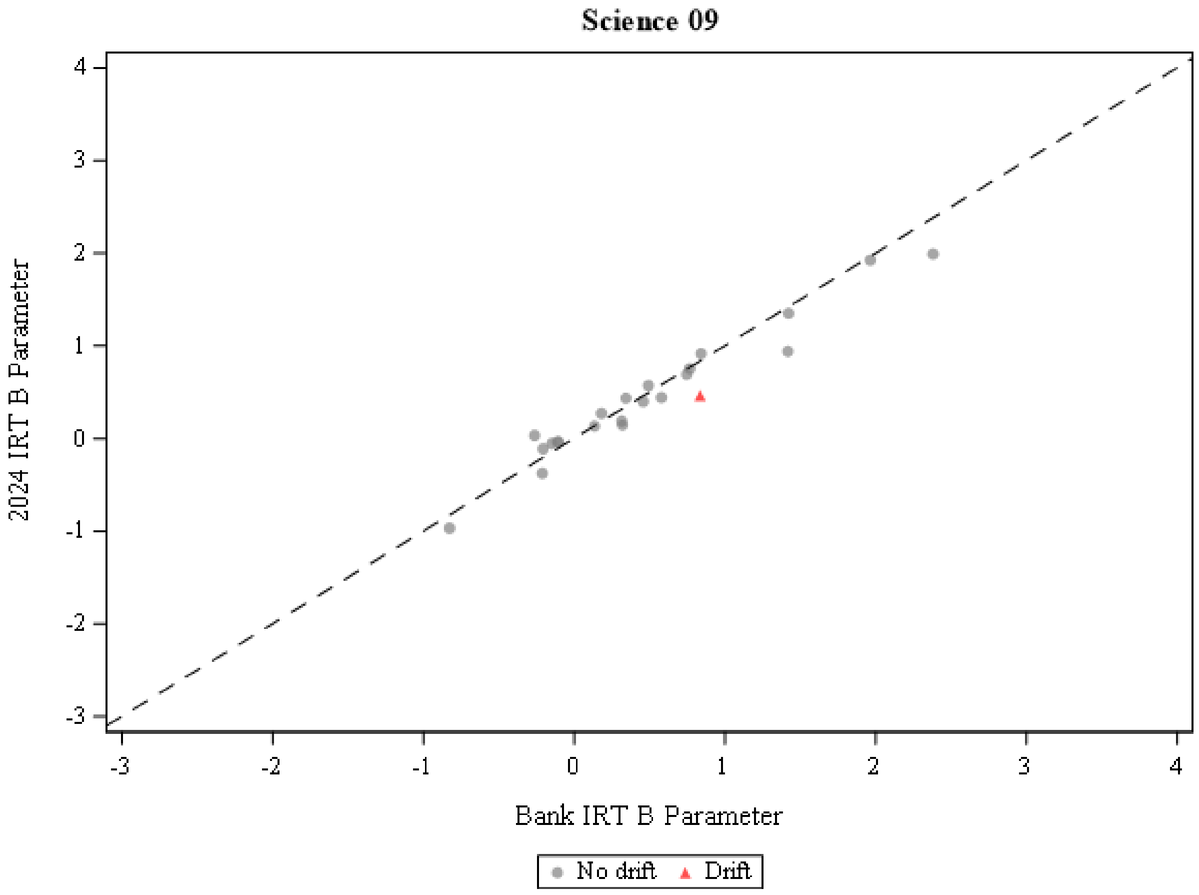

Figure G.7. Science Grade 9 IRT B Parameters for Operational Items

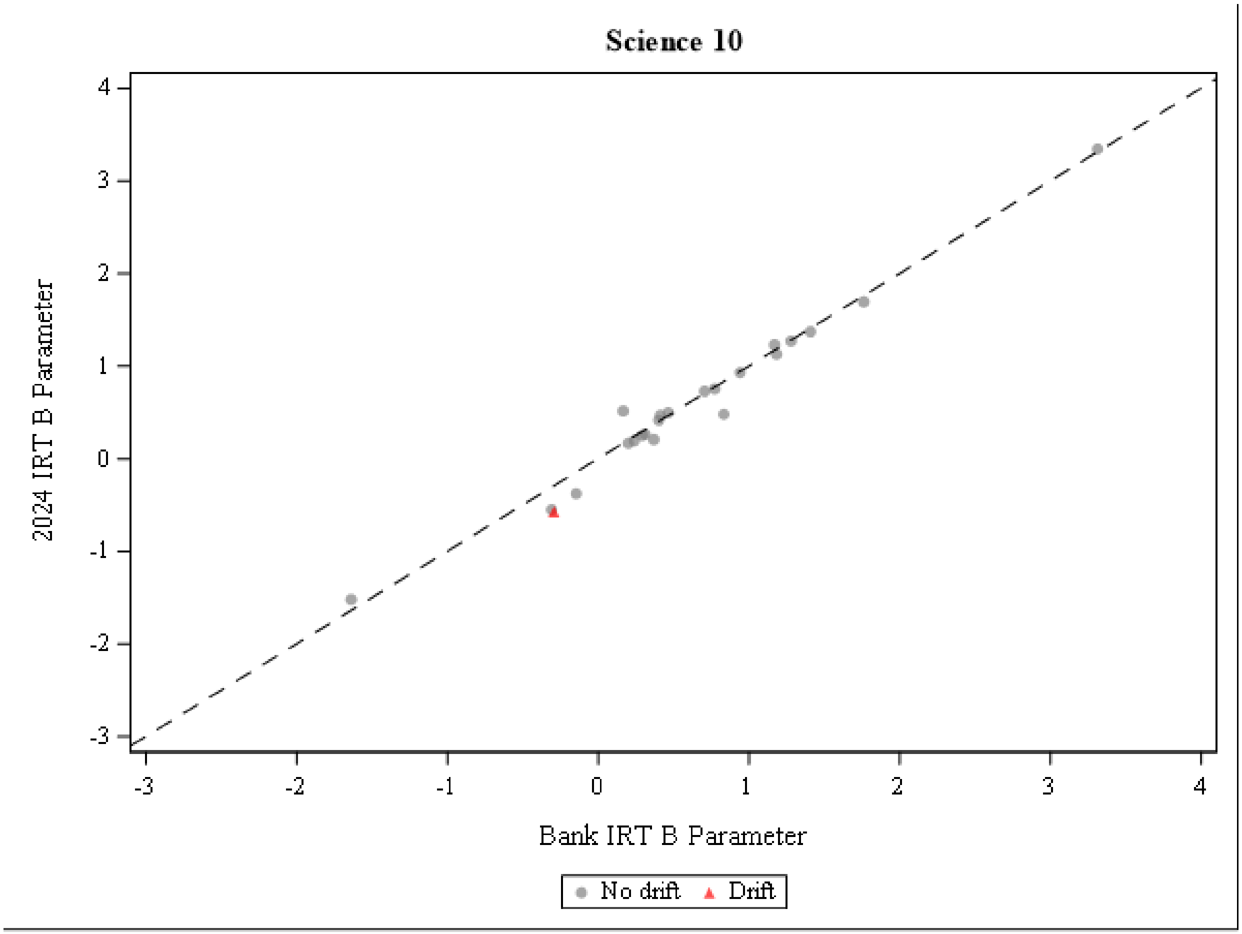

Figure G.8. Science Grade 10 IRT B Parameters for Operational Items

Appendix H: Scale Score Descriptive Statistics by Subgroup

Table H.1. English Grade 9 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	45,391	193	29.17	176	195	212	-0.13
Sex	Female	21,705	198	28.22	180	198	215	-0.07
	Male	23,629	190	29.48	171	191	209	-0.16
	Unknown	57	180	36.44	155	179	212	-0.12
Ethnicity	Hispanic or Latino Ethnicity	8,943	178	27.34	161	179	196	-0.06
	Asian	764	196	31.43	178	196	216	-0.14
	Native Hawaiian or Other Pacific Islander	584	177	25.96	162	179	195	-0.10
	Black or African American	590	173	29.31	151	176	193	-0.24
	American Indian or Alaska Native	422	175	25.34	158	175	191	0.05
	White	32,473	198	27.87	182	199	216	-0.15
	Other	1,615	194	28.96	177	195	214	-0.03
Limited English Proficiency	No	41,574	196	27.90	180	197	214	-0.12
	Yes	3,817	161	22.61	147	163	177	-0.25
Economic Disadvantage	No	33,168	198	28.29	181	199	216	-0.16
	Yes	12,223	181	28.11	163	182	200	-0.04
Special Education	No	40,876	197	27.87	180	197	214	-0.14
	Yes	4,515	165	24.55	149	164	179	0.18

Table H.2. English Grade 10 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,431	197	28.69	178	197	216	0.06
Sex	Female	20,637	201	27.61	183	200	219	0.13
	Male	22,752	194	29.22	174	195	213	0.05
	Unknown	42	190	32.48	161	186.5	213	0.25
Ethnicity	Hispanic or Latino Ethnicity	8,476	183	25.88	166	183	200	0.16
	Asian	734	201	29.70	181	202	221	0.07
	Native Hawaiian or Other Pacific Islander	603	183	22.34	168	183	197	-0.02
	Black or African American	584	177	25.76	160	176	194	0.10
	American Indian or Alaska Native	411	180	22.77	165	178	194	0.31
	White	31,077	202	28.11	184	202	219	0.00
	Other	1,546	198	28.22	179	198	217	0.03
Limited English Proficiency	No	39,926	200	27.88	182	200	217	0.05
	Yes	3,505	168	20.28	155	169	180	-0.11
Economic Disadvantage	No	32,405	201	28.28	183	201	219	0.02
	Yes	11,026	186	27.15	168	186	204	0.18
Special Education	No	39,328	200	27.81	182	200	218	0.05
	Yes	4,103	170	22.70	156	170	183	0.34

Table H.3. Reading Grade 9 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	45,559	195	29.39	175	196	215	-0.15
Sex	Female	21,844	198	28.25	180	199	217	-0.13
	Male	23,661	192	30.07	170	193	213	-0.13
	Unknown	54	181	35.01	160	178.5	209	-0.17
Ethnicity	Hispanic or Latino Ethnicity	9,050	181	27.24	162	181	199	0.01
	Asian	770	197	29.93	178	197.5	218	-0.14
	Native Hawaiian or Other Pacific Islander	588	180	24.99	163	179	196	0.14
	Black or African American	593	176	28.69	156	174	196	0.24
	American Indian or Alaska Native	423	179	23.33	164	178	195	0.02
	White	32,516	199	28.59	181	201	219	-0.24
	Other	1,619	196	29.12	176	198	216	-0.13
Limited English Proficiency	No	41,626	198	28.46	179	199	217	-0.19
	Yes	3,933	165	21.98	152	165	180	-0.12
Economic Disadvantage	No	33,239	199	28.61	180	201	218	-0.21
	Yes	12,320	183	28.41	163	183	203	0.03
Special Education	No	41,026	198	28.38	179	199	217	-0.19
	Yes	4,533	168	24.56	153	167	183	0.22

Table H.4. Reading Grade 10 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,594	195	28.85	174	197	216	-0.06
Sex	Female	20,741	199	26.93	182	201	217	-0.14
	Male	22,810	192	30.15	168	193	215	0.06
	Unknown	43	188	36.55	158	194	215	-0.13
Ethnicity	Hispanic or Latino Ethnicity	8,544	182	27.01	162	181	202	0.15
	Asian	735	201	31.18	179	202	222	-0.18
	Native Hawaiian or Other Pacific Islander	624	180	25.50	161	180	197	0.11
	Black or African American	590	177	26.39	158	175.5	194	0.19
	American Indian or Alaska Native	417	182	24.47	164	181	198	0.02
	White	31,129	200	28.05	181	202	219	-0.13
	Other	1,555	197	28.56	176	199	216	-0.03
Limited English Proficiency	No	40,039	198	27.98	179	200	217	-0.09
	Yes	3,555	166	20.97	153	164	180	0.09
Economic Disadvantage	No	32,483	199	28.37	179	201	218	-0.12
	Yes	11,111	185	27.83	164	185	205	0.12
Special Education	No	39,450	198	28.11	179	200	217	-0.10
	Yes	4,144	171	23.83	155	168	185	0.45

Table H.5. Mathematics Grade 9 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,674	192	31.37	174	196	213	-0.67
Sex	Female	20,694	191	28.85	175	195	211	-0.78
	Male	22,934	193	33.47	173	197	216	-0.62
	Unknown	46	172	26.61	151	169	193	-0.15
Ethnicity	Hispanic or Latino Ethnicity	8,477	174	30.98	157	176	195	-0.53
	Asian	725	198	33.36	180	201	219	-0.62
	Native Hawaiian or Other Pacific Islander	560	174	30.47	158	177	193	-0.64
	Black or African American	555	168	31.94	149	170	191	-0.37
	American Indian or Alaska Native	393	172	28.08	158	176	190	-0.76
	White	31,417	198	28.99	182	201	217	-0.76
	Other	1,547	191	32.34	173	195	213	-0.67
Limited English Proficiency	No	39,958	195	29.76	179	198	215	-0.71
	Yes	3,716	158	28.44	146	162	178	-0.60
Economic Disadvantage	No	31,974	197	29.58	182	201	217	-0.77
	Yes	11,700	178	31.70	160	180	199	-0.49
Special Education	No	39,233	195	29.73	179	199	215	-0.74
	Yes	4,441	162	29.40	148	164	180	-0.33

Table H.6. Mathematics Grade 10 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	42,840	189	33.40	172	192	212	-0.74
Sex	Female	20,294	189	31.46	173	192	210	-0.84
	Male	22,507	190	35.03	171	193	214	-0.67
	Unknown	39	174	37.28	155	178	200	-0.46
Ethnicity	Hispanic or Latino Ethnicity	8,332	173	32.66	158	176	194	-0.59
	Asian	720	198	36.66	178	201	221.5	-0.54
	Native Hawaiian or Other Pacific Islander	610	171	31.22	155	175	191	-0.65
	Black or African American	577	164	34.24	150	168	186	-0.42
	American Indian or Alaska Native	398	169	33.48	153	176	190	-0.60
	White	30,685	194	31.65	179	198	215	-0.87
	Other	1,518	190	32.85	174	192	212	-0.67
Limited English Proficiency	No	39,349	192	32.11	175	195	213	-0.79
	Yes	3,491	157	31.23	142	164	178	-0.48
Economic Disadvantage	No	31,974	194	32.21	177	197	215	-0.81
	Yes	10,866	176	33.42	160	179	198	-0.62
Special Education	No	38,741	192	31.84	176	195	213	-0.79
	Yes	4,099	159	32.36	143	165	179	-0.40

Table H.7. Science Grade 9 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	45,542	202	34.02	181	204	225	-0.29
Sex	Female	21,825	202	31.99	182	204	223	-0.35
	Male	23,659	203	35.79	180	205	227	-0.26
	Unknown	58	190	30.19	172	187.5	215	-0.34
Ethnicity	Hispanic or Latino Ethnicity	9,053	186	31.62	168	187	207	-0.17
	Asian	772	207	35.84	184	209	230	-0.35
	Native Hawaiian or Other Pacific Islander	598	184	29.73	167	185	204	-0.29
	Black or African American	597	182	31.86	164	183	202	-0.39
	American Indian or Alaska Native	422	185	29.17	169	184.5	204	-0.33
	White	32,478	208	33.01	188	210	229	-0.37
	Other	1,622	202	34.80	180	204	225	-0.19
Limited English Proficiency	No	41,611	205	33.14	185	207	227	-0.32
	Yes	3,931	172	27.54	158	175	189	-0.48
Economic Disadvantage	No	33,225	207	33.30	187	209	229	-0.34
	Yes	12,317	190	32.86	171	190	212	-0.19
Special Education	No	41,014	205	33.02	185	207	227	-0.32
	Yes	4,528	175	30.52	159	175	193	-0.08

Table H.8. Science Grade 10 Scale Score Descriptive Statistics

	Test Group	N	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,491	196	32.89	180	199	216	-0.48
Sex	Female	20,686	195	31.32	180	199	215	-0.72
	Male	22,764	197	34.24	179	199	218	-0.33
	Unknown	41	192	34.43	179	188	214	-0.32
Ethnicity	Hispanic or Latino Ethnicity	8,542	182	31.15	169	186	202	-0.60
	Asian	736	203	34.44	185	204	223	-0.53
	Native Hawaiian or Other Pacific Islander	614	180	29.12	167	185	197	-0.92
	Black or African American	596	176	31.84	161.5	181	195.5	-0.48
	American Indian or Alaska Native	413	182	30.89	168	186	201	-0.67
	White	31,050	201	32.07	185	203	220	-0.51
	Other	1,540	197	32.99	179	199	217	-0.39
Limited English Proficiency	No	39,921	199	32.24	182	201	218	-0.51
	Yes	3,570	170	28.39	159	176	189	-0.88
Economic Disadvantage	No	32,402	200	32.32	183	202	219	-0.49
	Yes	11,089	186	32.42	171	189	206	-0.54
Special Education	No	39,376	199	32.26	182	201	218	-0.51
	Yes	4,115	174	30.39	162	179	192	-0.60

Appendix I: Scale Score Distributions for Overall Testing Population

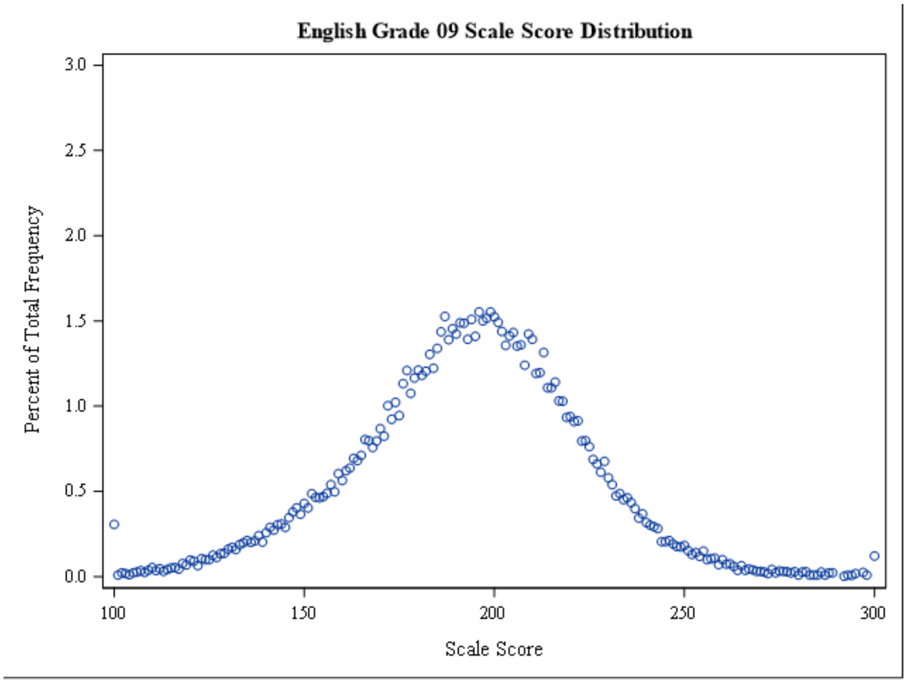

Figure I.1. English Grade 9 Scale Score Distribution

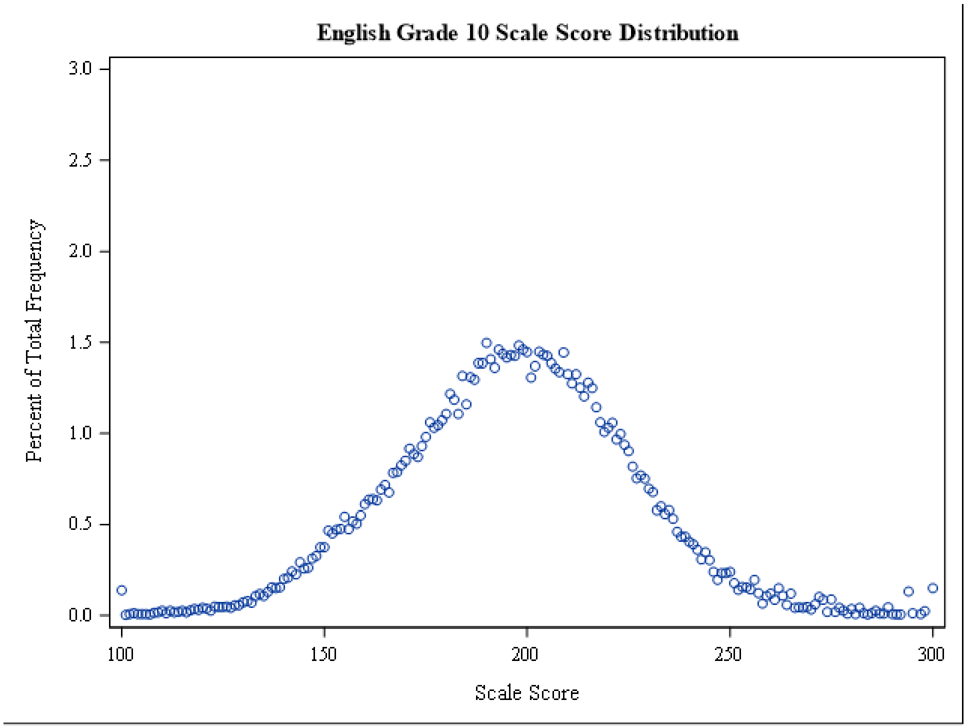

Figure I.2. English Grade 10 Scale Score Distribution

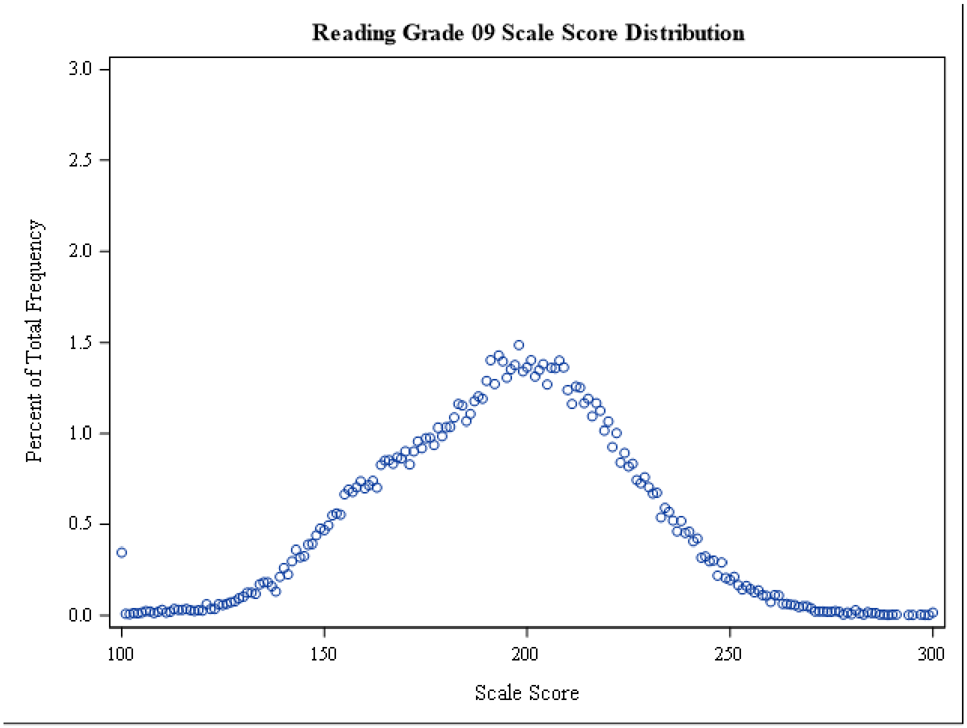

Figure I.3. Reading Grade 9 Scale Score Distribution

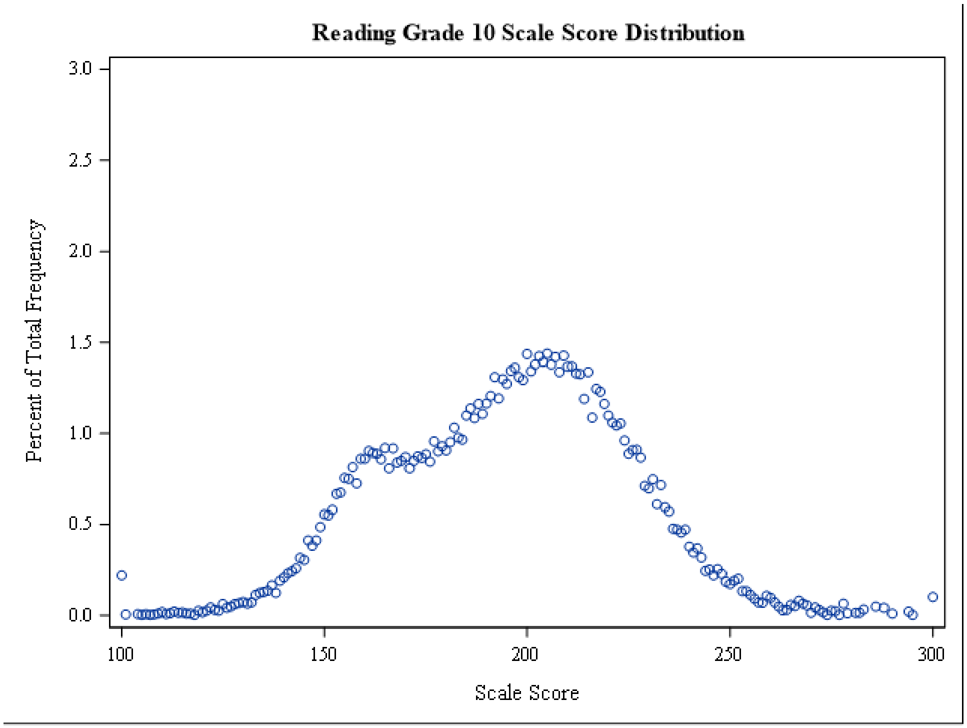

Figure I.4. Reading Grade 10 Scale Score Distribution

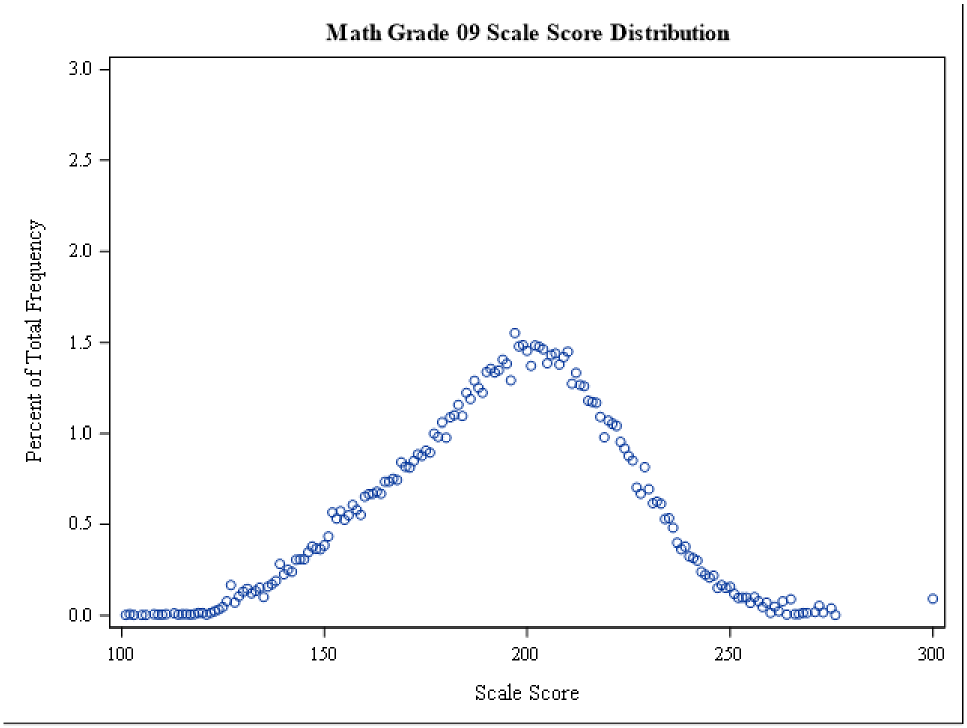

Figure I.5. Mathematics Grade 9 Scale Score Distribution

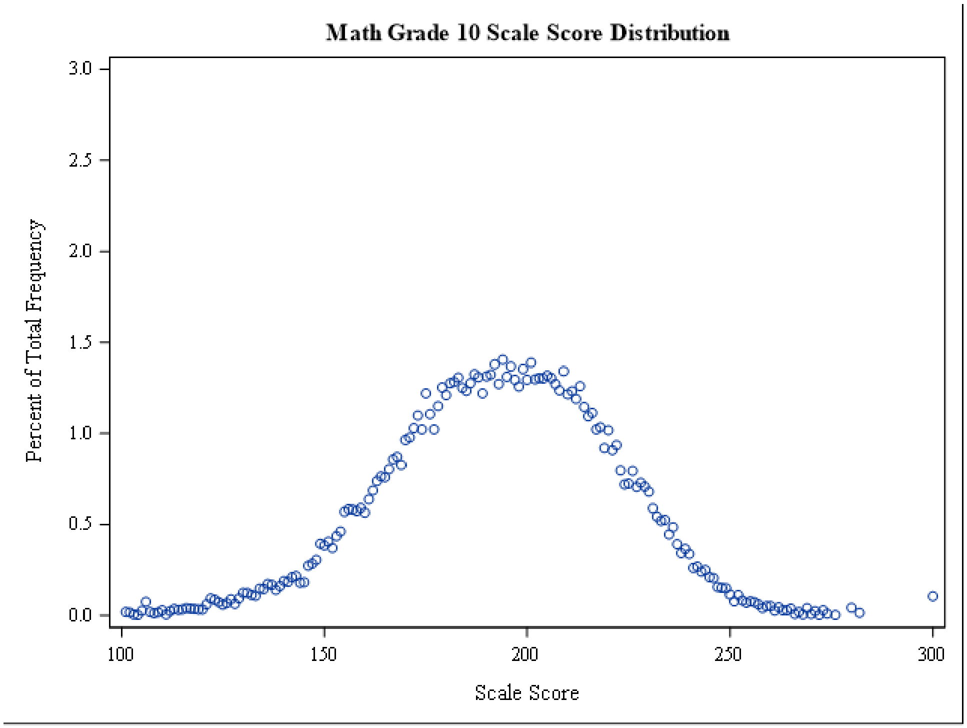

Figure I.6. Mathematics Grade 10 Scale Score Distribution

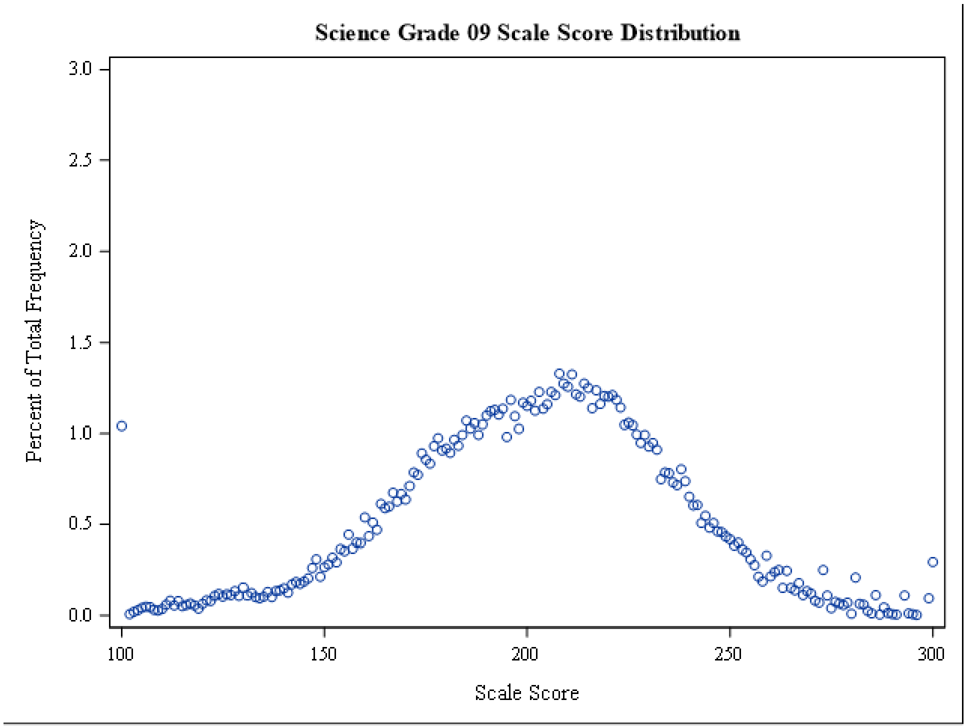

Figure I.7. Science Grade 9 Scale Score Distribution

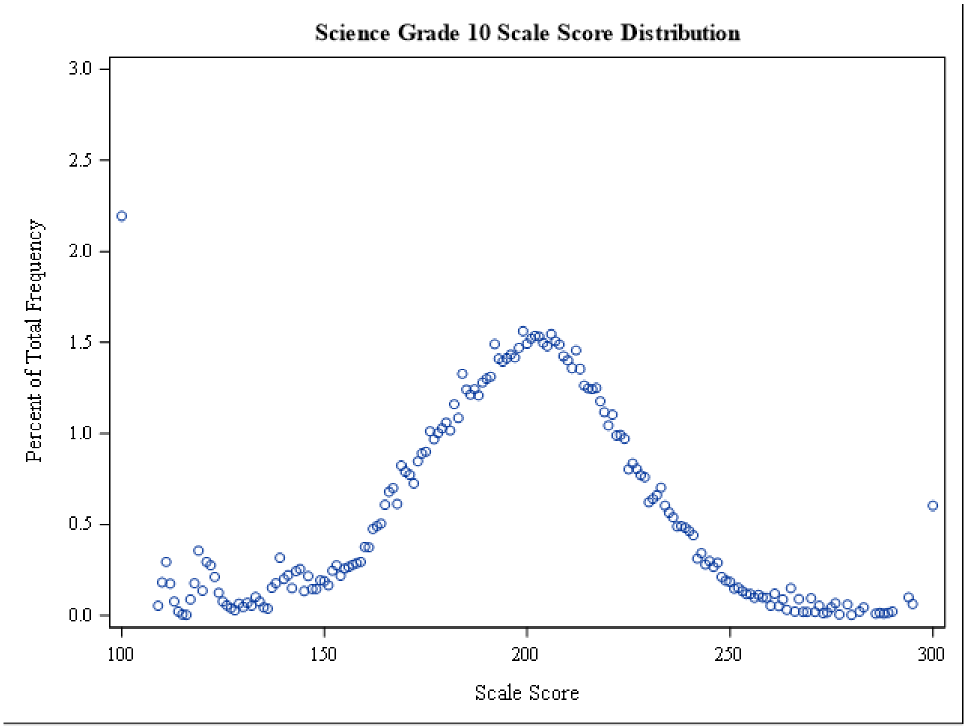

Figure I.8. Science Grade 10 Scale Score Distribution

Appendix J: Performance Level Distributions

Table J.1. English Grade 9 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	45,391	15.27	45.15	35.37	4.20
Sex	Female	21,705	11.31	44.20	39.16	5.34
	Male	23,629	18.86	46.06	31.92	3.16
	Unknown	57	36.84	35.09	26.32	1.75
Ethnicity	Hispanic or Latino Ethnicity	8,943	29.24	52.29	17.43	1.04
	Asian	764	15.18	43.19	35.34	6.28
	Native Hawaiian or Other Pacific Islander	584	28.42	57.19	13.36	1.03
	Black or African American	590	36.27	48.14	15.42	0.17
	American Indian or Alaska Native	422	34.60	52.13	12.56	0.71
	White	32,473	10.59	42.93	41.27	5.20
	Other	1,615	14.61	44.09	37.34	3.96
Limited English Proficiency	No	41,574	11.79	45.26	38.37	4.58
	Yes	3,817	53.24	44.04	2.72	0.00
Economic Disadvantage	No	33,168	11.14	43.05	40.69	5.12
	Yes	12,223	26.49	50.85	20.95	1.70
Special Education	No	40,876	11.41	45.38	38.59	4.62
	Yes	4,515	50.23	43.15	6.22	0.40

Table J.2. English Grade 10 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,431	10.17	42.87	42.42	4.53
Sex	Female	20,637	6.83	41.83	46.04	5.30
	Male	22,752	13.18	43.83	39.16	3.83
	Unknown	42	23.81	38.10	33.33	4.76
Ethnicity	Hispanic or Latino Ethnicity	8,476	18.23	56.35	24.15	1.27
	Asian	734	8.17	38.96	46.19	6.68
	Native Hawaiian or Other Pacific Islander	603	15.42	63.52	20.73	0.33
	Black or African American	584	25.51	55.48	18.49	0.51
	American Indian or Alaska Native	411	19.46	61.80	18.00	0.73
	White	31,077	7.56	38.41	48.45	5.59
	Other	1,546	9.18	42.76	43.73	4.33
Limited English Proficiency	No	39,926	8.06	41.30	45.71	4.92
	Yes	3,505	34.18	60.77	4.99	0.06
Economic Disadvantage	No	32,405	8.03	39.41	47.13	5.44
	Yes	11,026	16.46	53.07	28.61	1.87
Special Education	No	39,328	7.87	41.22	45.96	4.95
	Yes	4,103	32.22	58.76	8.53	0.49

Table J.3. Reading Grade 9 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	45,559	16.98	43.12	29.11	10.79
Sex	Female	21,844	12.92	43.84	30.98	12.26
	Male	23,661	20.70	42.47	27.39	9.44
	Unknown	54	31.48	37.04	24.07	7.41
Ethnicity	Hispanic or Latino Ethnicity	9,050	30.07	49.86	16.76	3.31
	Asian	770	15.45	41.95	30.13	12.47
	Native Hawaiian or Other Pacific Islander	588	29.25	55.61	12.93	2.21
	Black or African American	593	38.79	44.52	13.32	3.37
	American Indian or Alaska Native	423	29.55	55.32	13.71	1.42
	White	32,516	12.65	40.88	33.22	13.26
	Other	1,619	15.87	42.87	30.76	10.50
Limited English Proficiency	No	41,626	13.74	42.96	31.50	11.79
	Yes	3,933	51.26	44.83	3.76	0.15
Economic Disadvantage	No	33,239	12.98	41.23	32.86	12.92
	Yes	12,320	27.76	48.21	19.00	5.02
Special Education	No	41,026	13.54	42.99	31.62	11.84
	Yes	4,533	48.09	44.28	6.38	1.26

Table J.4. Reading Grade 10 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,594	25.19	33.28	33.91	7.63
Sex	Female	20,741	18.91	35.67	37.64	7.77
	Male	22,810	30.86	31.13	30.51	7.50
	Unknown	43	41.86	18.60	32.56	6.98
Ethnicity	Hispanic or Latino Ethnicity	8,544	41.71	35.38	20.17	2.74
	Asian	735	20.95	30.34	36.46	12.24
	Native Hawaiian or Other Pacific Islander	624	45.19	36.86	16.51	1.44
	Black or African American	590	48.98	35.42	13.56	2.03
	American Indian or Alaska Native	417	40.05	41.25	16.55	2.16
	White	31,129	19.76	32.59	38.48	9.17
	Other	1,555	23.99	32.41	36.08	7.52
Limited English Proficiency	No	40,039	21.29	33.88	36.53	8.29
	Yes	3,555	69.06	26.50	4.33	0.11
Economic Disadvantage	No	32,483	20.93	32.36	37.72	8.99
	Yes	11,111	37.65	35.96	22.75	3.65
Special Education	No	39,450	21.44	33.63	36.62	8.31
	Yes	4,144	60.86	29.95	8.04	1.16

Table J.5. Mathematics Grade 9 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,674	22.57	41.65	28.73	7.05
Sex	Female	20,694	21.30	45.97	28.07	4.66
	Male	22,934	23.64	37.77	29.36	9.23
	Unknown	46	58.70	30.43	10.87	0.00
Ethnicity	Hispanic or Latino Ethnicity	8,477	43.40	43.02	12.07	1.51
	Asian	725	18.62	38.48	31.59	11.31
	Native Hawaiian or Other Pacific Islander	560	42.50	46.25	10.00	1.25
	Black or African American	555	52.07	38.02	8.83	1.08
	American Indian or Alaska Native	393	45.04	46.56	7.63	0.76
	White	31,417	15.85	41.29	34.15	8.71
	Other	1,547	23.08	41.31	27.86	7.76
Limited English Proficiency	No	39,958	18.55	42.55	31.19	7.70
	Yes	3,716	65.74	31.94	2.21	0.11
Economic Disadvantage	No	31,974	16.66	41.12	33.52	8.69
	Yes	11,700	38.71	43.10	15.62	2.57
Special Education	No	39,233	18.02	42.72	31.50	7.75
	Yes	4,441	62.71	32.20	4.21	0.88

Table J.6. Mathematics Grade 10 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	42,840	34.51	37.83	22.57	5.09
Sex	Female	20,294	34.19	40.49	21.70	3.61
	Male	22,507	34.75	35.46	23.36	6.43
	Unknown	39	53.85	25.64	17.95	2.56
Ethnicity	Hispanic or Latino Ethnicity	8,332	56.79	32.78	9.06	1.37
	Asian	720	27.64	32.08	27.22	13.06
	Native Hawaiian or Other Pacific Islander	610	59.34	32.30	8.03	0.33
	Black or African American	577	68.46	23.40	7.28	0.87
	American Indian or Alaska Native	398	58.29	32.66	8.04	1.01
	White	30,685	27.19	39.78	26.93	6.10
	Other	1,518	34.26	38.01	21.67	6.06
Limited English Proficiency	No	39,349	30.55	39.54	24.38	5.53
	Yes	3,491	79.03	18.59	2.18	0.20
Economic Disadvantage	No	31,974	28.57	39.21	25.97	6.25
	Yes	10,866	51.96	33.79	12.54	1.70
Special Education	No	38,741	30.03	39.77	24.62	5.58
	Yes	4,099	76.82	19.49	3.15	0.54

Table J.7. Science Grade 9 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	45,542	30.36	27.41	27.66	14.57
Sex	Female	21,825	29.56	29.90	28.22	12.32
	Male	23,659	31.05	25.12	27.16	16.67
	Unknown	58	48.28	22.41	25.86	3.45
Ethnicity	Hispanic or Latino Ethnicity	9,053	49.98	28.48	16.45	5.09
	Asian	772	27.33	23.96	29.15	19.56
	Native Hawaiian or Other Pacific Islander	598	54.18	28.43	14.55	2.84
	Black or African American	597	55.78	27.14	14.41	2.68
	American Indian or Alaska Native	422	53.08	29.62	13.74	3.55
	White	32,478	23.75	27.17	31.41	17.67
	Other	1,622	30.58	26.94	27.93	14.55
Limited English Proficiency	No	41,611	26.45	27.89	29.77	15.89
	Yes	3,931	71.71	22.26	5.39	0.64
Economic Disadvantage	No	33,225	24.97	26.85	30.77	17.42
	Yes	12,317	44.91	28.92	19.29	6.88
Special Education	No	41,014	26.33	27.89	29.88	15.90
	Yes	4,528	66.85	23.06	7.58	2.52

Table J.8. Science Grade 10 Performance Level Distribution

	Test Group	N	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,491	33.06	32.95	27.73	6.26
Sex	Female	20,686	31.90	35.28	28.18	4.65
	Male	22,764	34.09	30.85	27.34	7.72
	Unknown	41	48.78	21.95	24.39	4.88
Ethnicity	Hispanic or Latino Ethnicity	8,542	51.02	32.67	14.52	1.79
	Asian	736	27.04	28.67	33.29	11.01
	Native Hawaiian or Other Pacific Islander	614	53.91	35.67	9.77	0.65
	Black or African American	596	62.25	28.02	7.89	1.85
	American Indian or Alaska Native	413	51.33	34.14	13.08	1.45
	White	31,050	27.05	33.18	32.17	7.60
	Other	1,540	32.99	32.47	27.66	6.88
Limited English Proficiency	No	39,921	29.66	33.62	29.91	6.81
	Yes	3,570	71.06	25.43	3.39	0.11
Economic Disadvantage	No	32,402	28.56	32.90	31.15	7.39
	Yes	11,089	46.22	33.10	17.76	2.93
Special Education	No	39,376	29.74	33.44	30.01	6.81
	Yes	4,115	64.81	28.24	5.98	0.97

Appendix K: Principal Components Scree Plot

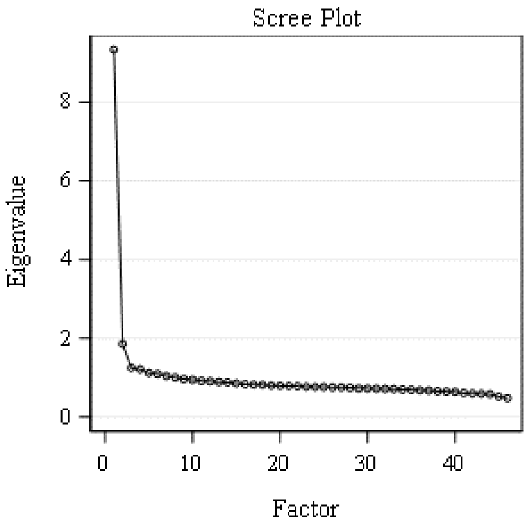

Figure K.1. English Grade 9 Principal Components Scree Plot

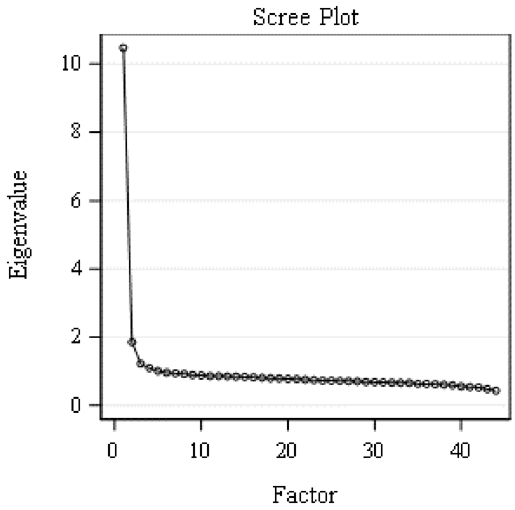

Figure K.2. English Grade 10 Principal Components Scree Plot

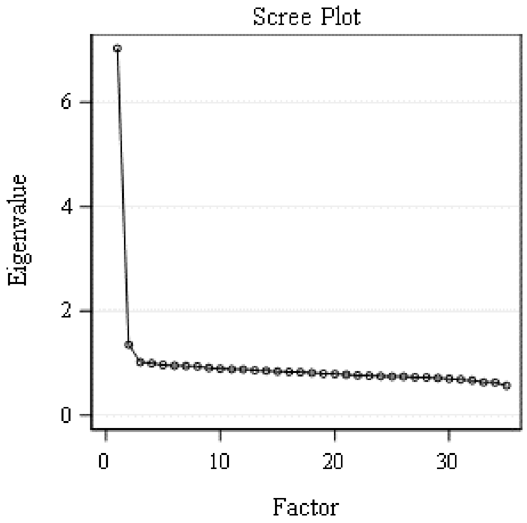

Figure K.3. Reading Grade 9 Principal Components Scree Plot

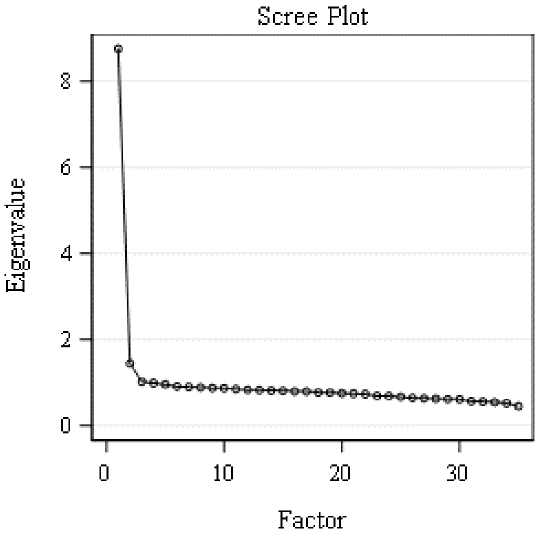

Figure K.4. Reading Grade 10 Principal Components Scree Plot

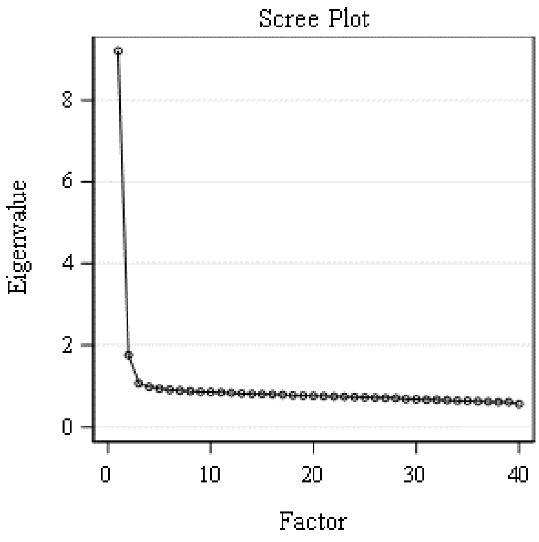

Figure K.5. Mathematics Grade 9 Principal Components Scree Plot

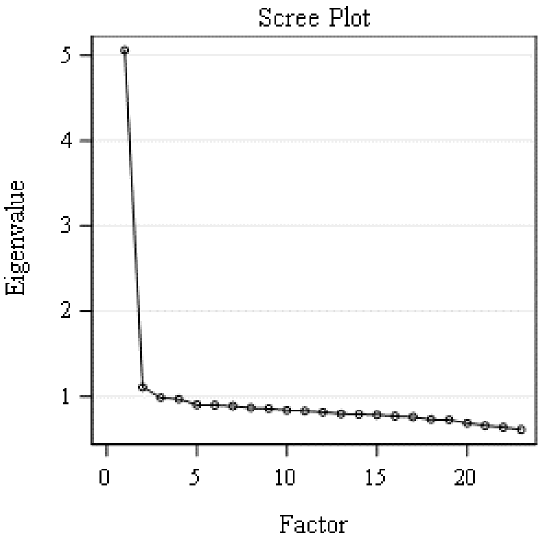

Figure K.7. Science Grade 9 Principal Components Scree Plot

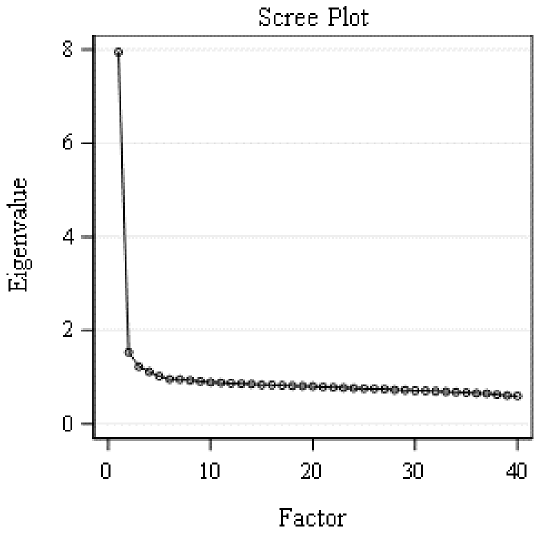

Figure K.6. Mathematics Grade 10 Principal Components Scree Plot

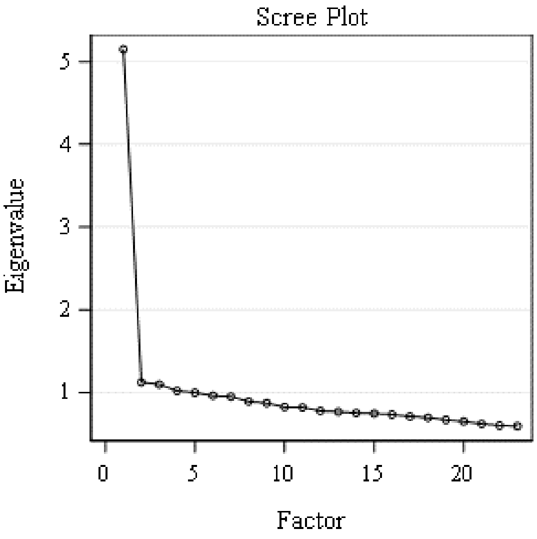

Figure K.8. Science Grade 10 Principal Components Scree Plot

Appendix L: Subscore Correlations

Table L.1. English Correlations of Total Score and Subscores

Grade	Subdomain	English Total	Production of Writing	Knowledge of Language	Conventions of Standard English
9	Total	1.00			
	Production of Writing	0.87	1.00		
	Knowledge of Language	0.76	0.64	1.00	
	Conventions of Standard English	0.94	0.77	0.68	1.00
10	Total	1.00			
	Production of Writing	0.84	1.00		
	Knowledge of Language	0.77	0.66	1.00	
	Conventions of Standard English	0.95	0.77	0.72	1.00

Table L.2. Reading Correlations of Total Score and Subscores

Grade	Subdomain	Reading Total	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
9	Total	1.00			
	Key Ideas	0.92	1.00		
	Craft and Structure	0.89	0.76	1.00	
	Integration of Knowledge and Ideas	0.62	0.53	0.50	1.00
10	Total	1.00			
	Key Ideas	0.93	1.00		
	Craft and Structure	0.90	0.80	1.00	
	Integration of Knowledge and Ideas	0.72	0.62	0.61	1.00

Table L.3. Mathematics Correlations of Total Score and Subscores

Grade	Subdomain	Math Total	Number and Quantity	Algebra	Functions	Geometry	Statistics and Probability
9	Total	1.00	–				
	Algebra	0.81	–	1.00			
	Functions	0.79	–	0.74	1.00		
	Geometry	0.85	–	0.71	0.68	1.00	
	Statistics and Probability	0.82	–	0.70	0.68	0.70	1.00
10	Total	1.00					
	Number and Quantity	0.66	1.00				
	Algebra	0.68	0.53	1.00			
	Functions	0.71	0.58	0.67	1.00		
	Geometry	0.80	0.59	0.60	0.66	1.00	
	Statistics and Probability	0.69	0.46	0.47	0.50	0.56	1.00

Table L.4. Science Correlations of Total Score and Subscores

Grade	Subdomain	Science Total	Gathering & Investigating	Developing Models	Using Mathematical Thinking	Construct Explanation
9	Total	1.00				
	Gathering & Investigating	0.68	1.00			
	Developing Models	0.71	0.47	1.00		
	Using Mathematical Thinking	0.84	0.53	0.51	1.00	
	Construct Explanation	0.81	0.49	0.49	0.60	1.00
10	Total	1.00				
	Gathering & Investigating	0.75	1.00			
	Developing Models	0.75	0.51	1.00		
	Using Mathematical Thinking	0.64	0.49	0.48	1.00	
	Construct Explanation	0.68	0.58	0.50	0.52	1.00

Appendix M: Item Drift

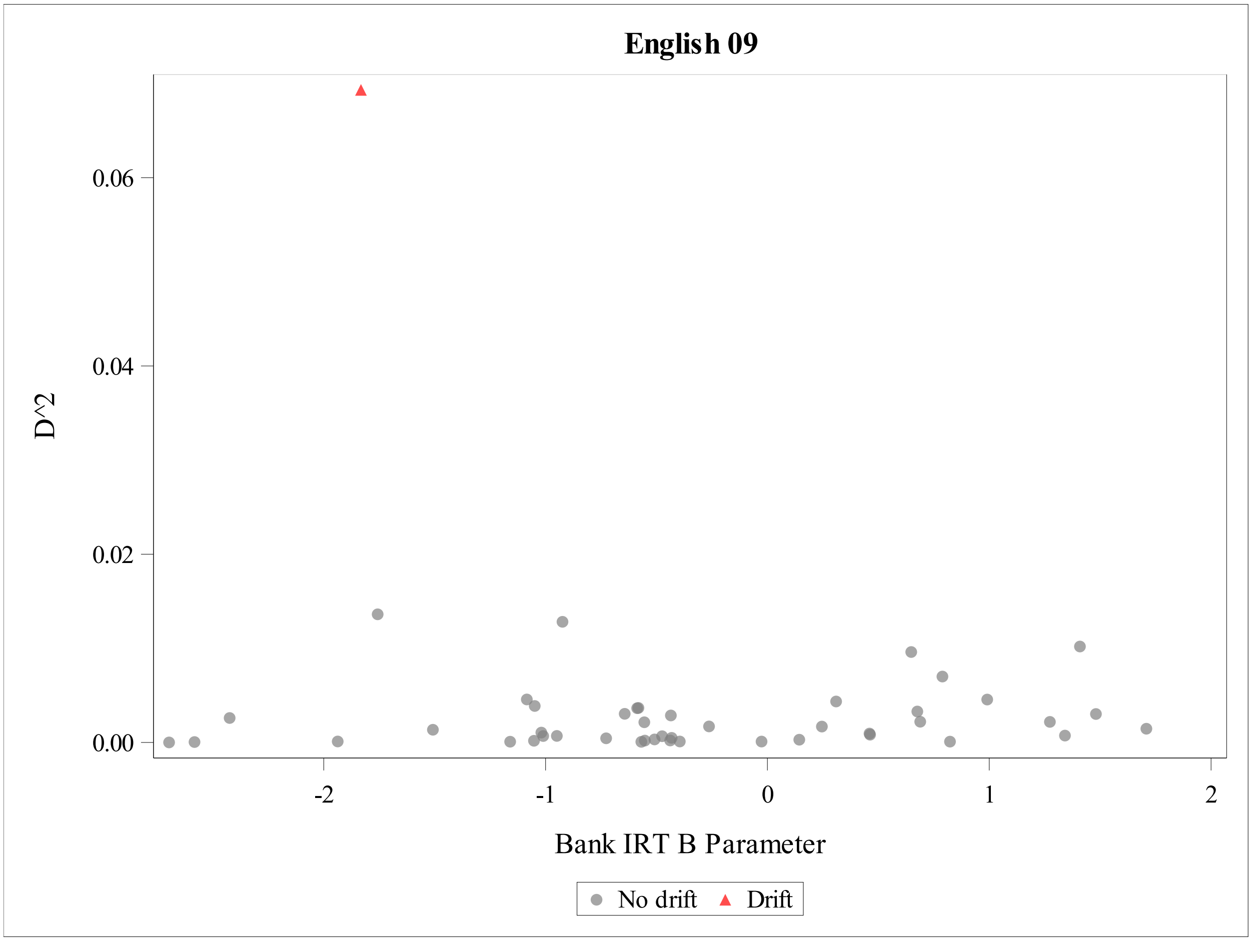

Figure M.1. English Grade 9 Item Drift

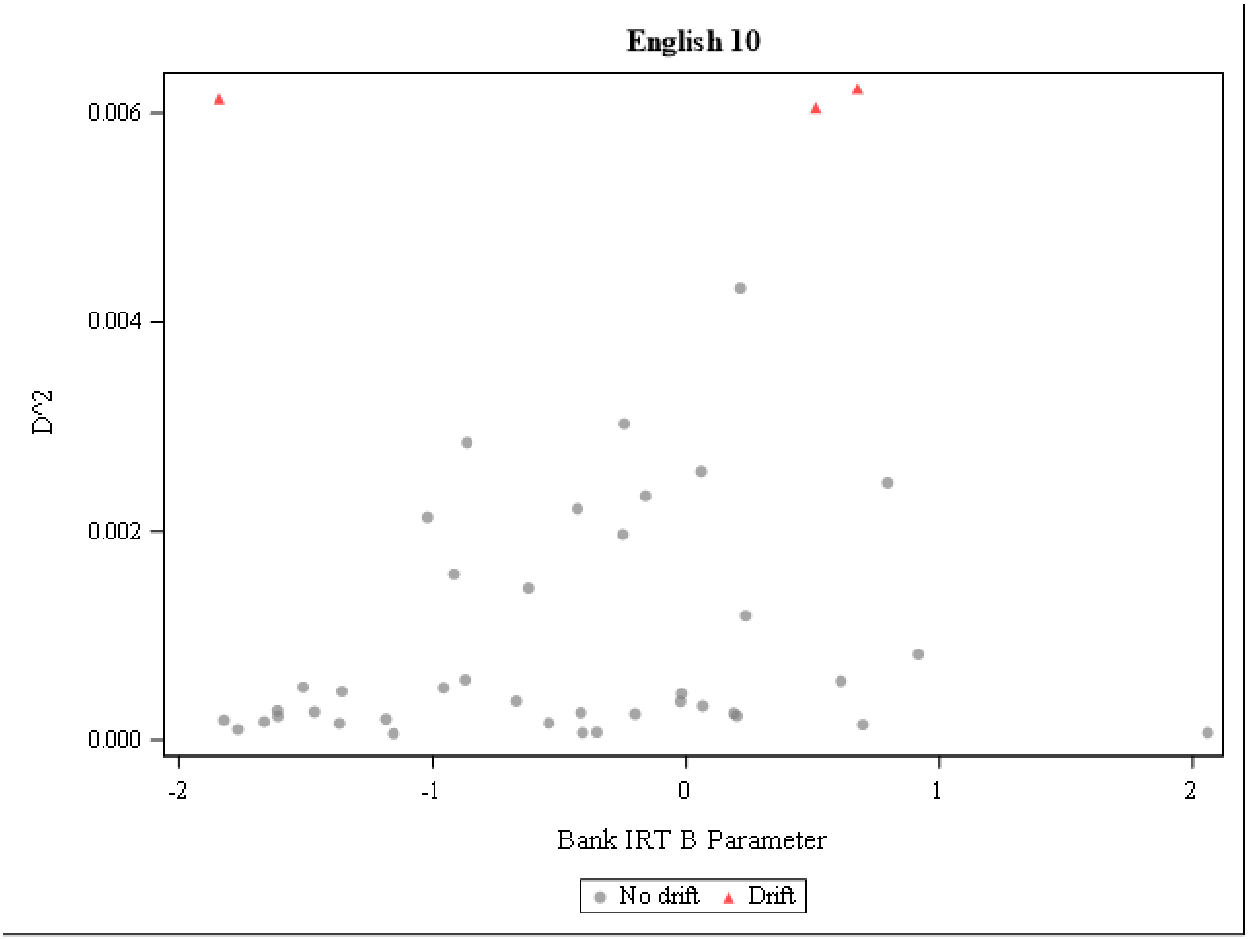

Figure M.2. English Grade 10 Item Drift

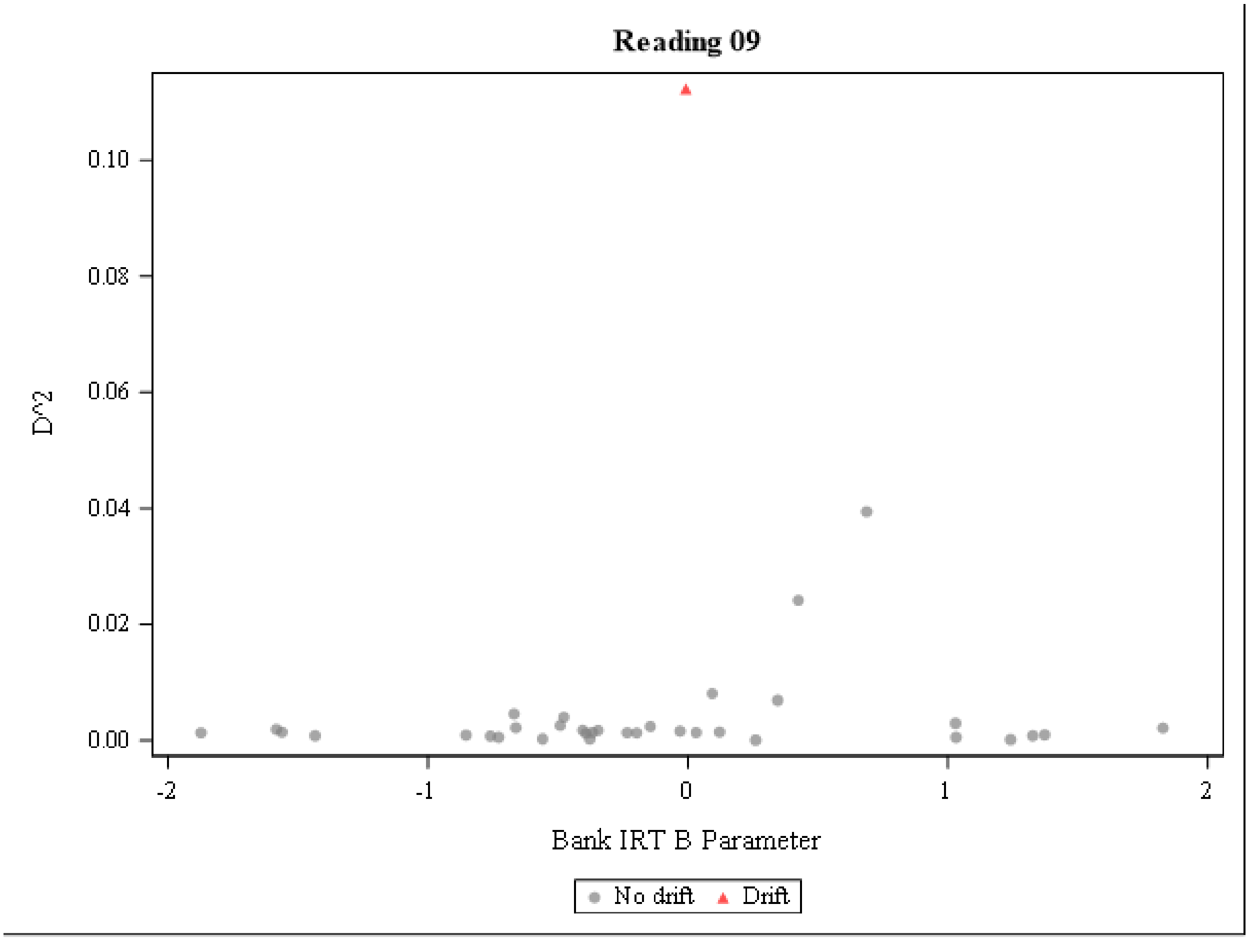

Figure M.3. Reading Grade 9 Item Drift

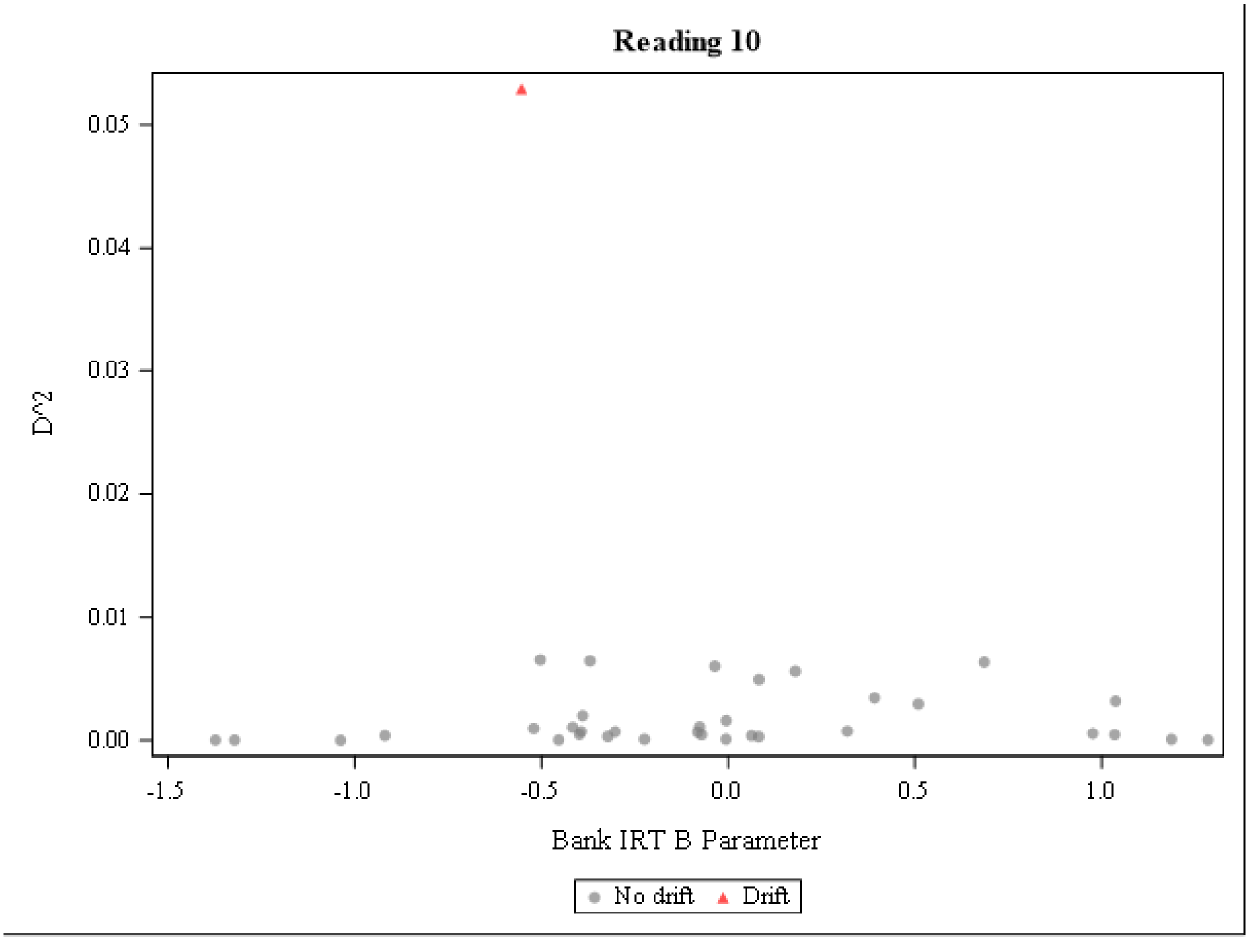

Figure M.4. Reading Grade 10 Item Drift

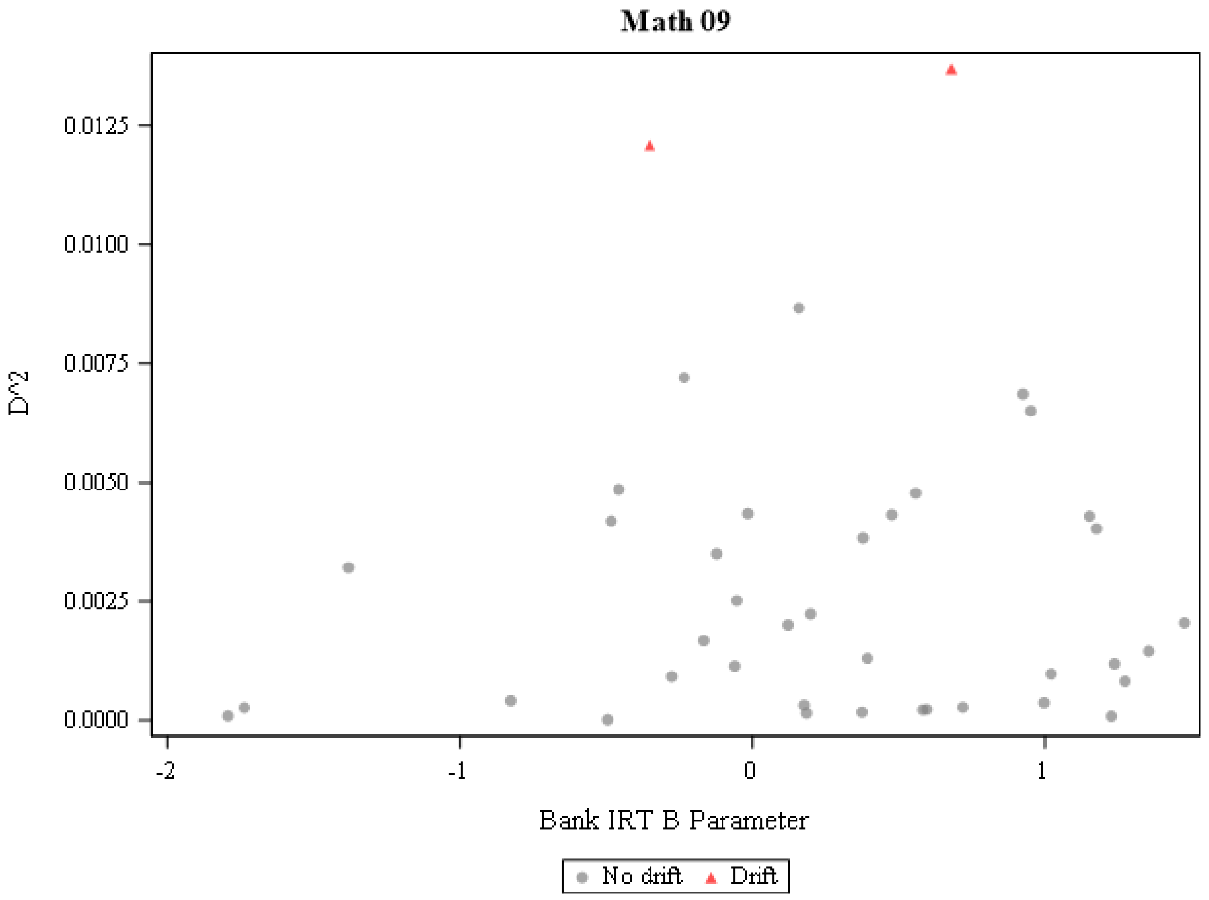

Figure M.5. Mathematics Grade 9 Item Drift

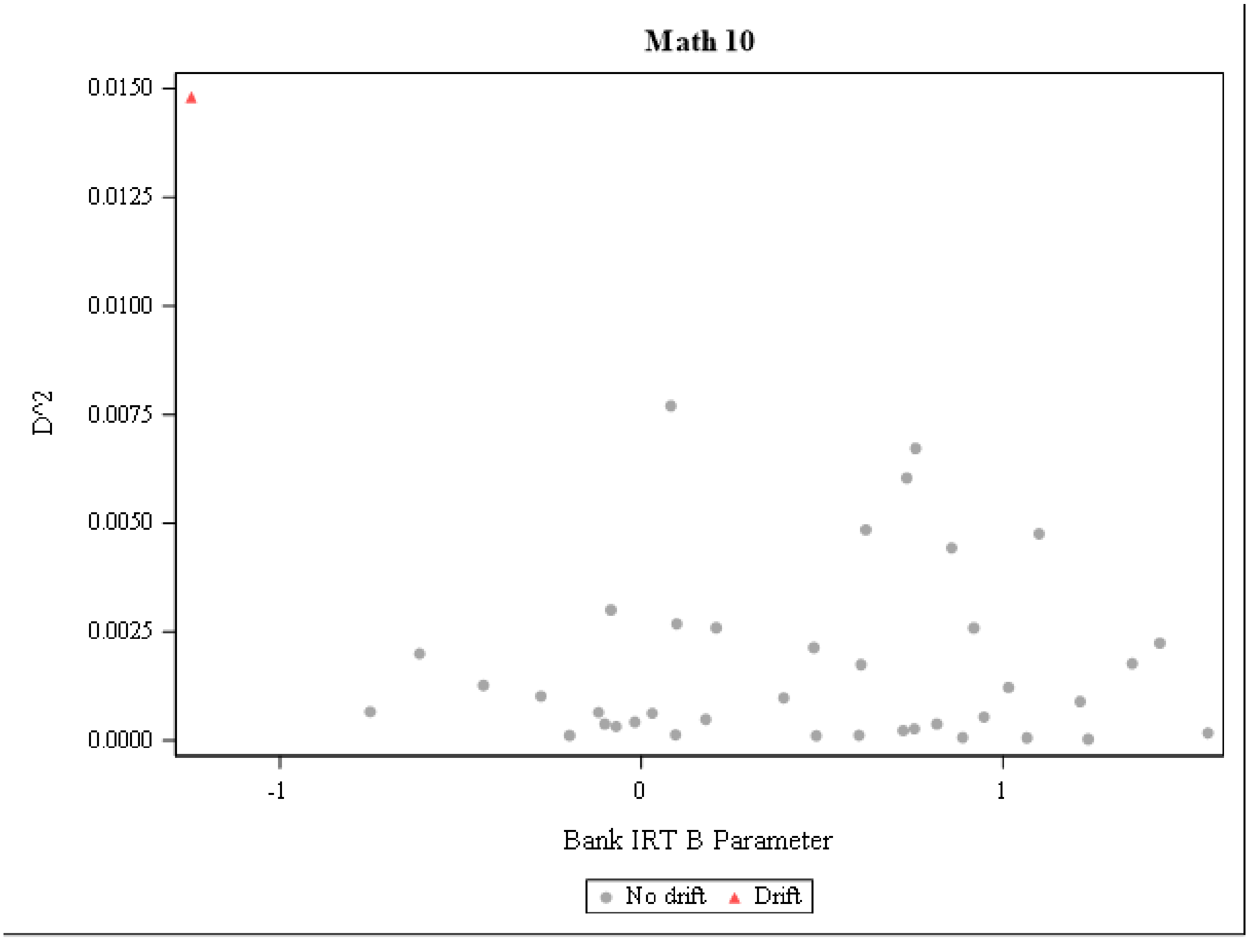

Figure M.6. Mathematics Grade 10 Item Drift

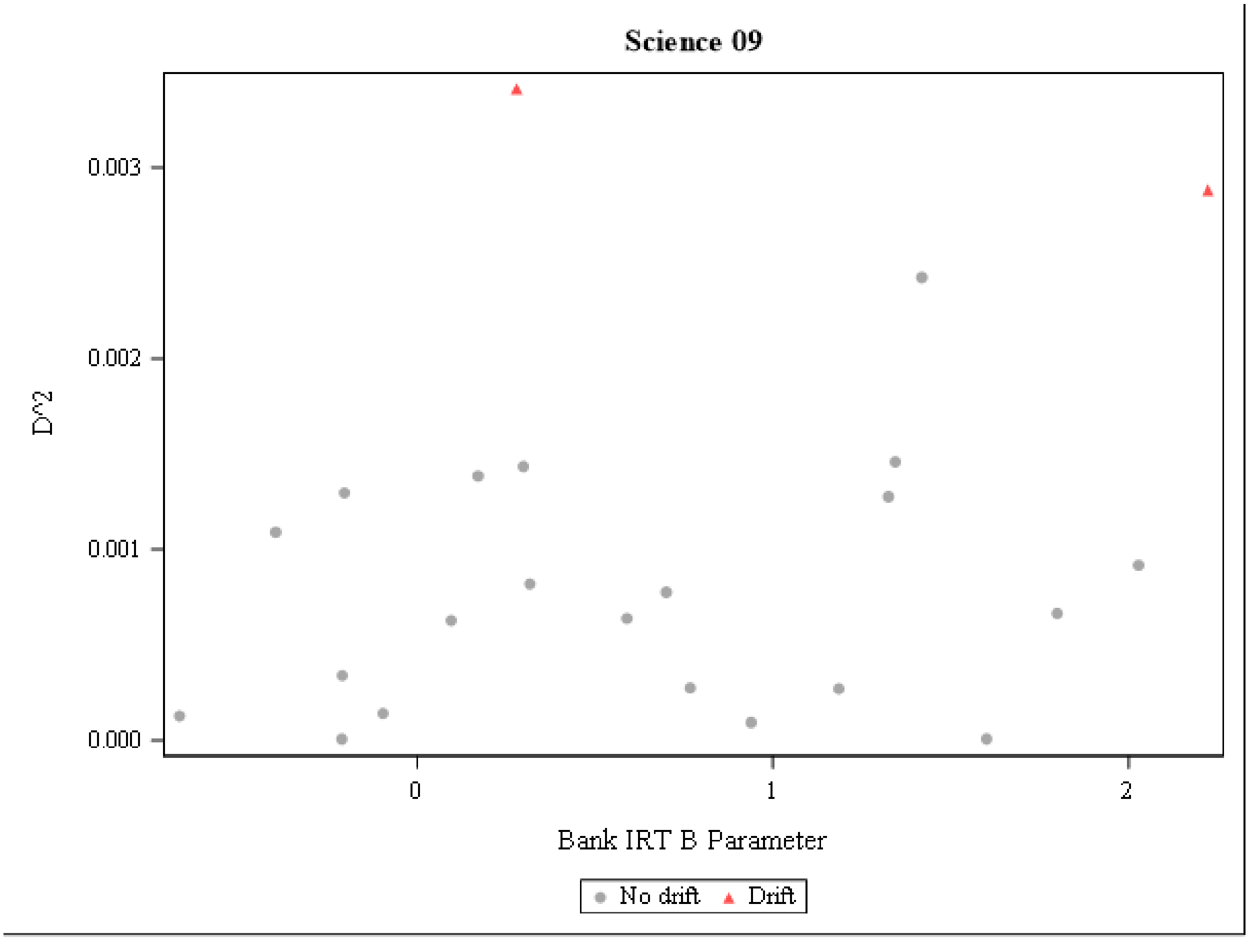

Figure M.7. Science Grade 9 Item Drift

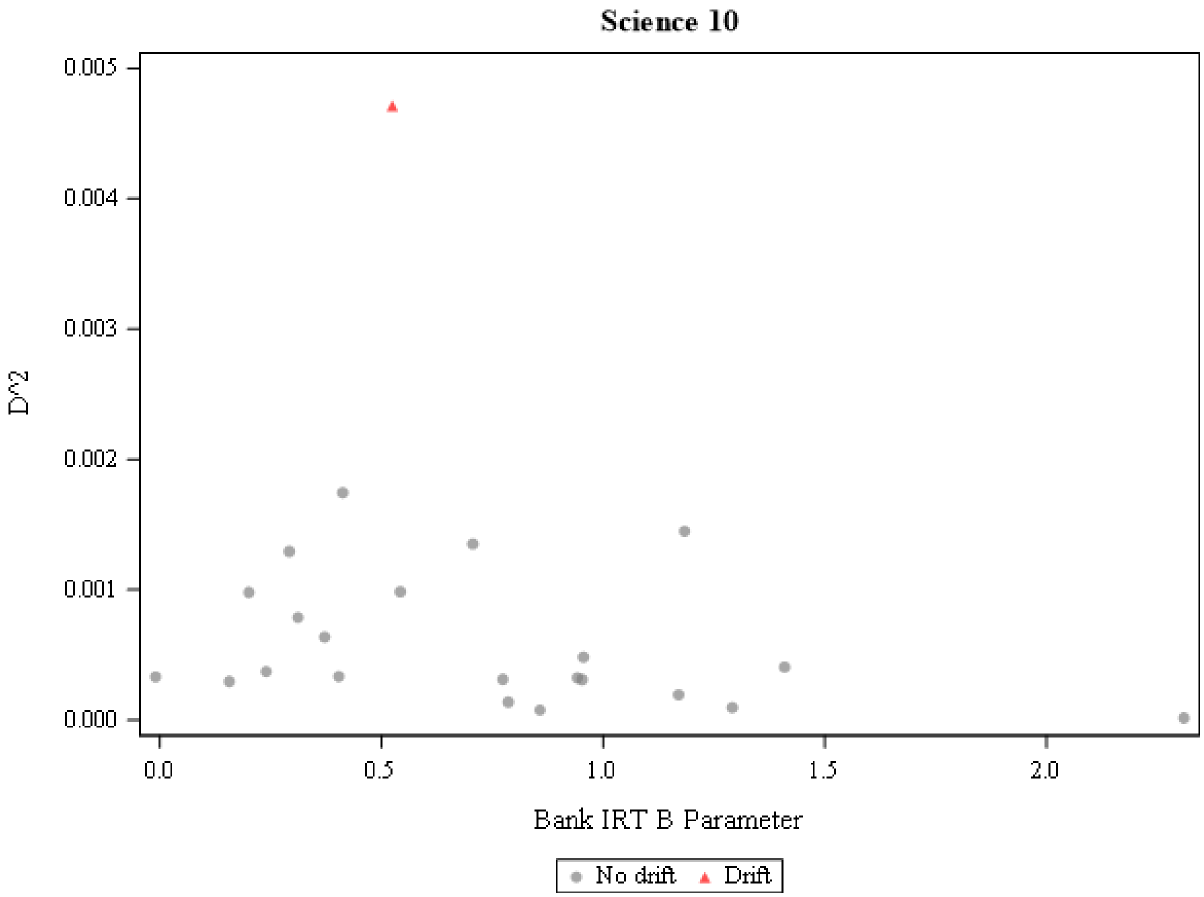

Figure M.8. Science Grade 10 Item Drift