Utah Aspire Plus
2018–2019 Technical Report

2020

TABLE OF CONTENTS

LIST OF TABLES **5**

1. INTRODUCTION **6**

1.1 Background 6
1.2 Purpose of the Operational Tests 7
1.3 Composition of the Operational Tests 7
1.4 Intended Population of the Operational Tests 8
1.5 Overview of the Technical Report 8

2. TEST DEVELOPMENT **9**

2.1 Overview of the Utah Aspire Plus Assessments, Claims, and Blueprints 9
2.1.1 English Assessment Claims *9*
2.1.2 Reading Assessment Claims *10*
2.1.3 Mathematics Assessment Claims *10*
2.1.4 Science Assessment Claims *11*
2.2 Utah Aspire Plus Blueprint Creation 12
2.3 Test Development Activities 15
2.3.1 Item Development *15*
2.3.2 ACT Aspire Item Alignment *15*
2.3.3 Operational Forms Development *16*
2.3.4 Statistical Guidelines *16*
2.3.5 2019 Match to Test Blueprint *17*

3. OPERATIONAL ADMINISTRATION **22**

3.1 Testing Window 22
3.2 Test Administration and Security Policies 22
3.2.1 Online Administration and Monitoring *23*
3.3 Test Accommodations and Supports 23
3.4 Test Taking Irregularities and Security Breaches 25
3.4.1 Test Interruptions *25*
3.4.2 Scoring of Interrupted Tests *26*
3.4.3 Wrong Test Form/Accommodation *26*
3.4.4 Extended Time Accommodation Issues *26*
3.4.5 Test Invalidation *26*
3.5 Test Taker Characteristics 26
3.6 Testing Time 30

4. CLASSICAL ITEM ANALYSES **33**

4.1 Item Analyses 33
4.1.1 p-Value and Item Mean Scores *33*
4.1.2 Item-Test Score Correlations *33*
4.1.3 Differential Item Functioning *33*
4.2 Classical Item Summaries for Operational Administration 35

5. RELIABILITY **36**

5.1 Classical Definition of Reliability 36
5.2 Classical Test Theory Reliability Estimates 36
5.2.1 Cronbach's Alpha *36*
5.2.2 Standard Error of Measurement *37*
5.3 IRT Based Reliability 37
5.4 Reliability of Performance Level Categorization 38
5.4.1 Accuracy and Consistency *38*
5.4.2 Calculating Accuracy *38*
5.4.3 Calculating Consistency *40*

5.4.4 Calculating Kappa *40*

6. IRT CALIBRATION AND SCALING **41**

6.1 Overview 41
6.2 IRT Data Preparation 41
6.2.1 Student Inclusion/Exclusion Rules *41*
6.2.2 Quality Control of the IRT Data Matrix Files *41*
6.3 Description of the Calibration Process 42
6.3.1 IRT Models *42*
6.3.2 IRTPRO Calibration Procedures and Convergence Criteria *42*
6.3.3 Calibration Quality Control *43*
6.4 Model Fit Evaluation Criteria 43
6.5 Summary Statistics and Distributions from IRT Analyses 44
6.6 IRT Pattern Scoring 46
6.6.1 Quality Control of IRT Scoring *46*

7. SCORE REPORTING **47**

7.1 Appropriate Uses for Scores and Reports 47
7.2 Utah Aspire Plus Reporting Scale 47
7.3 Standard setting 48
7.4 ACT Predicted Score Ranges 48
7.5 2018–2019 Utah Aspire Plus Performance Results 49

8. QUALITY CONTROL **50**

8.1 Online Assessment Delivery 50
8.1.1 Item Validation *50*
8.1.2 Test Administration *50*
8.1.3 Operational Monitoring *51*
8.2 Production System Testing 51
8.2.1 Functional Testing *51*
8.2.2 Integration Testing *51*
8.2.3 Program Validation End-to-End Testing *52*
8.2.4 Load Testing *52*
8.2.5 Performance Monitoring *52*
8.2.6 Regression Testing *52*
8.2.7 User Acceptance Testing *53*
8.3 Reporting 53
8.4 Quality Control of Psychometric Processes 53

9. VALIDITY **54**

9.1 Evidence Based on Test Content 54
9.2 Evidence Based on Cognitive Process 55
9.3 Evidence Based on Internal Structure 56
9.3.1 Reliability *58*
9.4 Evidence Based on Different Student Populations 58
9.5 Summary 58

10. REFERENCES **60**

APPENDIX A: TEST BLUEPRINT EDUCATOR COMMITTEE **62**

APPENDIX B: ITEM ALIGNMENT EDUCATOR COMMITTEE **78**

APPENDIX C: TEST-LEVEL REPORTING CATEGORIES AND STANDARDS BY ITEM TYPE AND DOK **86**

APPENDIX D: STUDENT TESTING TIME **92**

APPENDIX E: ITEM STATISTICS SUMMARIES **95**

APPENDIX F: RELIABILITY AND STANDARD ERROR BY SUBGROUP 98
APPENDIX G: CONDITIONAL STANDARD ERROR OF SCALE SCORES 107
APPENDIX H: ACCURACY AND CONSISTENCY 110
APPENDIX I: PERFORMANCE LEVEL DESCRIPTOR EDUCATOR COMMITTEE 118
APPENDIX J: PREDICTING ACT TEST SCORES FROM THE UTAH ASPIRE PLUS HIGH SCHOOL ASSESSMENT 126
APPENDIX K: UTAH-TO-ACT CONCORDANCE TABLES 144
APPENDIX L: SCALE SCORE DESCRIPTIVE STATISTICS BY SUBGROUP 154
APPENDIX M: SCALE SCORE DISTRIBUTIONS FOR OVERALL TESTING POPULATION 166
APPENDIX N: PERFORMANCE LEVEL DISTRIBUTIONS 169
APPENDIX O: STANDARD PROCESSES AND QUALITY MANAGEMENT 181
APPENDIX P: PRINCIPAL COMPONENTS SCREE PLOTS 192
APPENDIX Q: SUBSCORE CORRELATIONS 195

List of Tables

Table 1. Utah Aspire Plus English (Grades 9 and 10) Test Design and Blueprint 13
Table 2. Utah Aspire Plus Reading (Grades 9 and 10) Test Design and Blueprint 13
Table 3. Utah Aspire Plus Mathematics (Grade 9) Test Design and Blueprint 13
Table 4. Utah Aspire Plus Mathematics (Grade 10) Test Design and Blueprint 14
Table 5. Utah Aspire Plus Science (Grades 9 and 10) Test Design and Blueprint 14
Table 6. Utah Aspire Plus English Grade 9 Operational Test Blueprint Match 18
Table 7. Utah Aspire Plus English Grade 10 Operational Test Blueprint Match 18
Table 8. Utah Aspire Plus Reading Grade 9 Operational Test Blueprint Match 19
Table 9. Utah Aspire Plus Reading Grade 10 Operational Test Blueprint Match 19
Table 10. Utah Aspire Plus Mathematics Grade 9 Operational Test Blueprint Match 20
Table 11. Utah Aspire Plus Mathematics Grade 10 Operational Test Blueprint Match 20
Table 12. Utah Aspire Plus Science Grade 9 Operational Test Blueprint Match 21
Table 13. Utah Aspire Plus Science Grade 10 Operational Test Blueprint Match 21
Table 14. Spring 2019 Participation Rates for Utah Aspire Plus 27
Table 15. Spring 2019 Accommodation Rates for Utah Aspire Plus 28
Table 16. Student Testing Time for Spring 2019 Utah Aspire Plus 31
Table 17. Item 2x2 Contingency Table for the *k*th Score Level 34
Table 18. Example Accuracy Classification Table 39
Table 19. Example Accuracy Classification Table for Proficient Cut Point 39
Table 20. Example Consistency Classification Table 40
Table 21. IRT Summary Parameter Estimates for Utah Aspire Plus Operational Items 45
Table 22. IRT Standard Errors of Parameter Estimates for Utah Aspire Plus Operational Items 45
Table 23. IRT Model Fit for Utah Aspire Plus Operational Items 45
Table 24. Model Fit Indices for Confirmatory Factor Analyses 57

1. Introduction

1.1 Background

The Utah Aspire Plus summative assessments were created out of Utah Statute 53E-4-304. 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.

Summative assessments for the first operational Utah Aspire Plus administration were created in 2018. The first operational administration was conducted in the spring of 2019 at grades 9 and 10 for English, reading, mathematics, and science. The majority of students (over 99%) took the tests on computer. Data from this inaugural administration were used to establish the initial Utah Aspire reporting scales and scores as described in this report. They were also used in the setting of performance levels. Technical details of these features and activities are presented in this report with the exception of use within Accountability (e.g., growth between 9th and 10th grade). Readers can find information on Utah annual accountability determinations at https://www.schools.utah.gov/assessment/resources.

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 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. All blueprints were designed to measure knowledge and skills described in the Utah Core Standards (https://www.uen.org/core/). For science, the Utah Aspire Plus blueprints are further explicated to measure 1) science content specific to biology, chemistry, Earth science, or physics; and 2) Intended Learning Outcomes (ILOs). The ILOs describe the goals for science skills and attitudes. They are defined for each grade and are an integral part of the standards that are used to guide science instruction (https://www.schools.utah.gov/File/8cf206d1-022d-42ec-b02d-3cbad59ecb79). Additionally, the tests are designed to focus on the underlying skills of science as defined in the ILOs (e.g., science process and thinking skills, etc.) and not require specific knowledge of the scientific discipline (meaning a chemistry student ought to possess the skills necessary to answer a biology question).

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 Utah Aspire Plus Science tests is on the Intended Learning Outcomes (ILOs), which describe the skills students should learn from science instruction. From ILOs, students should use science as a process of obtaining knowledge based upon observable evidence. As noted, these skills are applicable regardless of domain (i.e., Biology, Physics, etc.).

1.4 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).

1.5 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 creation of the inaugural Utah Aspire Plus testing system at grades 9 and 10. In addition to a general overview that provides an initial frame of reference around key attributes of the system, the report provides details around development of items and test forms, the administration of operational tests, and scoring and reporting. 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 and scaling, standard setting, 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 was designed according to a principled assessment design framework. 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 <u>not</u> 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

* 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.

- Conventions of Standard English

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

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 Math 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 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 Math tests and are grouped into the following categories:

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 Intended Learning Outcomes (ILOs). From ILOs, students are expected to use science as a process of obtaining knowledge based upon observable evidence. As noted, 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 ILOs 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 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 ILOs 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 ILOs 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 Science tests and are grouped into the following categories:

- ILO 1 – Use Science Process and Thinking Skills
- ILO 3 – Demonstrate Understanding of Science Concepts, Principles, and Systems
- ILO 4 – Communicate Effectively Using Science Language and Reasoning

- ILO 5/6 – Demonstrate Awareness of Social and Historical Aspects of Science/Demonstrate Understanding of the Nature of Science

2.2 Utah Aspire Plus Blueprint Creation

The Utah Aspire Plus tests are administered in English, reading, mathematics, and science in grades 9 and 10. Tests are composed of multiple-choice, multiple-select, evidence-based selected response, and technology-enhanced (TE) items. 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, whereas 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). Examples of Utah Aspire Plus technology-enhanced items are:

- Inline choice interaction: drop-down multiple choice
- Text-entry interaction: fill-in-the-blank item presentation
- Hot text interaction: selecting elements within a given image (e.g., reading passage)
- Plot/drawing interaction: plotting/drawing on a grid
- Matching interaction: matching response elements to the appropriate category

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 creation of the Utah Aspire Plus blueprints, USBE invited Utah educators to participate in workshops where they reviewed the state's standards (including content breakout categories), in addition to the Utah SAGE and ACT Aspire test blueprints. Panelists were chosen to reflect Utah's educator populations by subject according to characteristics such as grades and subjects taught, years of teaching, rural/suburban/urban district, experience with test development, regular/charter, special education experience, and English as a second language endorsement.

During review and discussion of these materials, educators provided recommendations for creation of blueprints that would support the intended claims and appropriately sample content that covered the respective standards. Specifically, they recommended content domain coverage, item type distribution, overall number of items and points for each test, and testing time.

Appendix A contains the agenda for each educator group convened to discuss the test blueprint and provides the general training used to introduce the educators to this process. At the conclusion of the test blueprint workshops, Pearson and USBE reviewed the recommendations and finalized the test blueprints for each Utah Aspire Plus test (see tables below).

Table 1. Utah Aspire Plus English (Grades 9 and 10) Test Design and Blueprint

	Number of Items	Minimum %	Maximum %
Item Type			
Multiple Choice	48–50	96%	100%
Technology Enhanced	0–2	0%	4%
Depth of Knowledge			
Level 1	22–24	44%	48%
Level 2	10–12	20%	24%
Level 3	15–17	30%	34%
Reporting Categories			
Production of Writing	12–14	24%	28%
Knowledge of Language	7–10	14%	20%
Conventions of Standard English	28–30	56%	60%

Table 2. Utah Aspire Plus Reading (Grades 9 and 10) Test Design and Blueprint

	Number of Items	Minimum %	Maximum %
Item Type			
Multiple Choice	22–29	62%	82%
Technology Enhanced	2–5	6%	14%
Evidence-Based Selected Response	4–6	10%	17%
Depth of Knowledge			
Level 1	4–10	11%	28%
Level 2	12–20	34%	57%
Level 3	9–14	25%	40%
Reporting Categories			
Key Ideas	9–18	26%	51%
Craft and Structure	14–20	40%	57%
Integration of Knowledge and Ideas	3–5	9%	14%

Table 3. Utah Aspire Plus Mathematics (Grade 9) Test Design and Blueprint

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

Table 4. Utah Aspire Plus Mathematics (Grade 10) Test Design and Blueprint

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

Table 5. Utah Aspire Plus Science (Grades 9 and 10) Test Design and Blueprint

	Number of Items	Minimum %	Maximum %
Item Type			
Multiple Choice	29–34	81%	94%
Technology Enhanced	2–3	6%	8%
Depth of Knowledge			
Level 1	3–9	8%	25%
Level 2	12–23	33%	64%
Level 3	8–13	22%	36%
Reporting Categories			
(ILO) 1: Use Science Process and Thinking Skills	15–23	42%	64%
(ILO) 3: Demonstrate Understanding of Science Concepts, Principles, and Systems	4–6	11%	17%
(ILO) 4: Communicate Effectively Using Science Language and Reasoning	7–10	19%	28%
(ILO) 5/6: Demonstrate Awareness of Social and Historical Aspects of Science/Demonstrate Understanding of the Nature of Science	3–4	8%	11%

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.

2.3.1 Item Development

As noted, item development for the Utah Aspire Plus assessments came from the SAGE and ACT Aspire testing programs, respectively. SAGE content targets of development were based on the same Utah Core Standards as intended for use on Utah Aspire Plus. ACT Aspire content was developed around the Common Core State Standards and the Next Generation Science Standards, which align with the Utah Aspire Plus standards. ACT Aspire content was also developed in alignment with the ACT College and Career Readiness Standards.

The process of developing items that make up the 2018–2019 Utah Aspire tests followed principled design procedures in line with industry standards for producing items, passages, and stimuli developed using the principles of universal design for assessing the Utah Core Standards. For both SAGE and ACT Aspire content, steps included the hiring and training of highly qualified item writers, extensive expert content review at all points of development (including separate committee review of content as well as bias and sensitivity review), field testing, and data review.

Specific details of the processes by which the Utah SAGE content was developed are described in the 2016–2017 Utah State Assessments annual technical report published by Cambium Assessment (formerly the American Institutes of Research®; interested parties should contact USBE for a copy of this report). Volume 2: Test Development describes in detail all activities pertaining to creation of SAGE item banks aligned to the Utah Core Standards.

Specific details of the process of developing ACT Aspire content are documented in their technical manual: https://www.act.org/content/dam/act/unsecured/documents/2019/aspire/Aspire-Summative-Technical-Manual.pdf. Detailed test development procedures are described in Chapter 2 of that manual.

2.3.2 ACT Aspire Item Alignment

ACT Aspire items form part of the Utah Aspire Plus tests. Here they serve dual purposes. To provide Utah students a measure of college readiness, ACT Aspire test items are included on the Utah tests to facilitate linking from Utah test scores to predicted ACT scores. They also count toward students' overall scores on the respective Utah Aspire Plus score scales. In order to ensure that specific items were aligned specifically with the Utah Core Standards and Intended Learning Outcomes (science), special meetings were conducted.

Experts from USBE, Pearson, and ACT initially matched items to their respective standards and ILOs. Then expert panels of Utah educators were convened to review the proposed item alignment designations for approval or suggest modification of a given alignment designation. Panelists were selected to represent the field of Utah educators. Appendix B contains the agenda

for educator groups convened to discuss the item alignment and the general training used to introduce the educators to the process. The result of the process was sufficient alignment of ACT Aspire items to the Utah Core Standards and ILOs to fulfill the Utah Aspire Plus blueprints.

2.3.3 Operational Forms Development

The construction of test forms for 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 and science ILOs. 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 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 (described later in this report).

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, overall match to linking set design, 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. Once the ACT Aspire items were selected, the remaining portion of the test forms were completed with Utah SAGE items.

This procedure was an iterative process whereby the first proposed form is 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 without new development of content, bank limitations meant an inability to strictly meet the new blueprint in all cases (see below). It also meant that 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 to contribute positively to supporting the intended claims. Moving forward, newly developed content will fill gaps and address such limitations as the assessments mature.

2.3.4 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 2019 operational administration are presented in Appendix E. *It should be noted that Appendix E 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.

2.3.5 2019 Match to Test Blueprint

Tables 6 through 13 present the match between the final 2019 operational forms of Utah Aspire Plus and the test blueprints. Reading, math, and science final forms matched almost all targets by item type, depth of knowledge, and reporting category (within 2% on one reporting category for grade 10 math). The limitations of the SAGE and ACT Aspire item banks for English resulted in the larger differences compared to the target blueprints. For example, items available from the SAGE banks for English forms were all technology-enhanced item types, which meant exceeding the blueprint by over 25%. This also resulted in the differences observed for depth of knowledge as well as Conventions of Standard English.

As noted, current item development planning is intended to address all shortcomings in meeting each blueprint moving forward. Field testing of this content should allow for fully matched operational forms to be available in spring 2022.

Table 6. Utah Aspire Plus English Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	48–50	96%	100%	69%
Technology Enhanced	0–2	0%	4%	31%
Depth of Knowledge				
Level 1	22–24	44%	48%	60%
Level 2	10–12	20%	24%	13%
Level 3	15–17	30%	34%	27%
Reporting Categories				
Production of Writing	12–14	24%	28%	20%
Knowledge of Language	7–10	14%	20%	9%
Conventions of Standard English	28–30	56%	60%	71%

Table 7. Utah Aspire Plus English Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	48–50	96%	100%	65%
Technology Enhanced	0–2	0%	4%	35%
Depth of Knowledge				
Level 1	22–24	44%	48%	58%
Level 2	10–12	20%	24%	15%
Level 3	15–17	30%	34%	27%
Reporting Categories				
Production of Writing	12–14	24%	28%	19%
Knowledge of Language	7–10	14%	20%	10%
Conventions of Standard English	28–30	56%	60%	71%

Table 8. Utah Aspire Plus Reading Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	22–29	62%	82%	74%
Technology Enhanced	2–5	6%	14%	11%
Evidence-Based Selected Response	4–6	10%	17%	14%
Depth of Knowledge				
Level 1	4–10	11%	28%	14%
Level 2	12–20	34%	57%	49%
Level 3	9–14	25%	40%	37%
Reporting Categories				
Key Ideas	9–18	26%	51%	51%
Craft and Structure	14–20	40%	57%	40%
Integration of Knowledge and Ideas	3–5	9%	14%	9%

Table 9. Utah Aspire Plus Reading Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	22–29	62%	82%	77%
Technology Enhanced	2–5	6%	14%	9%
Evidence-Based Selected Response	4–6	10%	17%	14%
Depth of Knowledge				
Level 1	4–10	11%	28%	14%
Level 2	12–20	34%	57%	46%
Level 3	9–14	25%	40%	40%
Reporting Categories				
Key Ideas	9–18	26%	51%	51%
Craft and Structure	14–20	40%	57%	40%
Integration of Knowledge and Ideas	3–5	9%	14%	9%

Table 10. Utah Aspire Plus Mathematics Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
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%	45%
Level 3	9–13	23%	33%	25%
Reporting Categories				
Algebra	9–11	23%	28%	28%
Functions	10–12	25%	30%	28%
Geometry	9–11	23%	28%	25%
Statistics and Probability	7–9	18%	23%	20%

Table 11. Utah Aspire Plus Mathematics Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	30–33	75%	83%	79%
Technology Enhanced	7–10	18%	25%	21%
Depth of Knowledge				
Level 1	8–12	20%	30%	28%
Level 2	15–20	38%	50%	51%
Level 3	9–13	23%	33%	21%
Reporting Categories				
Number and Quantity	2–4	5%	10%	10%
Algebra	9–11	23%	28%	28%
Functions	10–12	25%	30%	26%
Geometry	11–13	28%	33%	26%
Statistics and Probability	2–4	5%	10%	10%

Table 12. Utah Aspire Plus Science Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	29–34	81%	94%	92%
Technology Enhanced	2–3	6%	8%	8%
Depth of Knowledge				
Level 1	3–9	8%	25%	6%
Level 2	12–23	33%	64%	64%
Level 3	8–13	22%	36%	31%
Reporting Categories				
(ILO) 1: Use Science Process and Thinking Skills	15–23	42%	64%	58%
(ILO) 3: Demonstrate Understanding of Science Concepts, Principles, and Systems	4–6	11%	17%	11%
(ILO) 4: Communicate Effectively Using Science Language and Reasoning	7–10	19%	28%	22%
(ILO) 5/6: Demonstrate Awareness of Social and Historical Aspects of Science/Demonstrate Understanding of the Nature of Science	3–4	8%	11%	8%

Table 13. Utah Aspire Plus Science Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	Actual 2019
Item Type				
Multiple Choice	29–34	81%	94%	92%
Technology Enhanced	2–3	6%	8%	8%
Depth of Knowledge				
Level 1	3–9	8%	25%	17%
Level 2	12–23	33%	64%	64%
Level 3	8–13	22%	36%	19%
Reporting Categories				
(ILO) 1: Use Science Process and Thinking Skills	15–23	42%	64%	58%
(ILO) 3: Demonstrate Understanding of Science Concepts, Principles, and Systems	4–6	11%	17%	14%
(ILO) 4: Communicate Effectively Using Science Language and Reasoning	7–10	19%	28%	19%
(ILO) 5/6: Demonstrate Awareness of Social and Historical Aspects of Science/Demonstrate Understanding of the Nature of Science	3–4	8%	11%	8%

For additional information on the 2019 operational forms, Appendix C contains a breakdown reporting categories and standards by item type and depth of knowledge (DOK).

3. Operational Administration

3.1 Testing Window

The inaugural administration of the Utah Aspire Plus assessments was March 25–May 17, 2019. Utah Aspire Plus may 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 detailed in the Test Administration Manual (TAM, http://utah.pearsonaccessnext.com/training/) as well as via the Utah Aspire Plus Resource Center (http://utah.pearsonaccessnext.com/training/). 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 here: https://schools.utah.gov/file/47844e6b-59f1-4213-8701-1c1edf5b8423.

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–assessed courses. This ensures that each student had a standardized (similar and fair) testing experience. 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 2018–2019 administration, the testing times were: 45 minutes for English, 90 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 test management and delivery systems. PearsonAccess^{next} is the web application used by test staff (i.e., test coordinators, room supervisors) to manage online testing and start and monitor tests. TestNav is the test delivery engine used by examinees to take the tests. 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.

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.

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 and 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 the accessibility and accommodations manual insert at http://utah.pearsonaccessnext.com/resources/training/UtahAspirePlusAccessibilityAccommodationsManualInsertExternal_FORWEB.pdf.)

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

- 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
- 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.

- Extra time
- Personalized auditory notification of remaining time
- Breaks: stop the clock
- Breaks: extending over multiple days
- Human scribe
- Home administration
- Word-to-word dictionary
- Human reader
- Signed exact English (directions only)
- Sign language interpretation
- Cued speech
- Auditory notification of remaining time
- 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 preparing and starting sessions. In the event 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.). Where a test is invalidated, the student is not given another opportunity to take the test. Invalidating a test had to be completed by the district testing administrator.

3.5 Test Taker Characteristics

Table 14 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 15 provides support and accommodation rates for each test.

Table 14. Spring 2019 Participation Rates for Utah Aspire Plus

Students	Subgroup	English		Reading		Math		Science	
		Gr. 9	Gr.10	Gr. 9	Gr. 10	Gr. 9	Gr. 10	Gr. 9	Gr. 10
All	Students Scored	46,050	43,836	46,238	44,132	45,590	43,705	46,149	43,901
Gender	Female	49.1%	49.2%	49.1%	49.2%	49.1%	49.2%	49.2%	49.2%
	Male	50.9%	50.8%	50.9%	50.8%	50.9%	50.8%	50.8%	50.8%
Ethnicity	Hispanic or Latino Ethnicity	17.1%	17.2%	17.3%	17.4%	17.1%	17.3%	17.2%	17.3%
	Asian	1.8%	1.9%	1.8%	1.9%	1.8%	1.9%	1.8%	1.9%
	Native Hawaiian or Other Pacific Islander	1.6%	1.6%	1.6%	1.6%	1.6%	1.6%	1.6%	1.6%
	Black or African American	1.3%	1.3%	1.4%	1.3%	1.3%	1.3%	1.4%	1.3%
	American Indian or Alaska Native	1.1%	1.1%	1.1%	1.1%	1.1%	1.1%	1.2%	1.1%
	White	74.4%	74.5%	74.2%	74.2%	74.4%	74.3%	74.2%	74.3%
	Other	2.7%	2.5%	2.7%	2.5%	2.7%	2.4%	2.7%	2.5%
Limited English Proficiency	No	95%	95%	95%	94.9%	95%	95%	94.9%	95%
	Yes	5.0%	5.0%	5.0%	5.1%	5.0%	5.0%	5.1%	5.0%
Economic Disadvantage	No	68.8%	70.9%	68.6%	70.6%	68.8%	70.8%	68.7%	70.8%
	Yes	31.2%	29.1%	31.4%	29.4%	31.2%	29.2%	31.3%	29.2%
Special Education	No	90.1%	90.8%	90.1%	90.7%	90.2%	90.7%	90.1%	90.7%
	Yes	9.9%	9.2%	9.9%	9.3%	9.8%	9.3%	9.9%	9.3%

Table 15. Spring 2019 Accommodation Rates for Utah Aspire Plus

		English		Reading		Math		Science	
Accommodation	**Test Group**	**Gr. 9**	**Gr. 10**	**Gr. 9**	**Gr. 10**	**Gr. 9**	**Gr. 10**	**Gr. 9**	**Gr. 10**
All	Students Tested	46,057	43,846	46,241	44,137	45,592	43,712	46,152	43,911
Alternate Language	English	99.61%	99.61%	99.61%	99.56%	99.61%	99.59%	99.6%	99.6%
	Spanish	0.39%	0.39%	0.39%	0.44%	0.39%	0.41%	0.40%	0.40%
Text-to-Speech Translation	No	99.61%	99.61%	99.61%	99.56%	99.61%	99.59%	99.6%	99.6%
	Yes	0.39%	0.39%	0.39%	0.44%	0.39%	0.41%	0.40%	0.40%
Translated Test Navigation	English	99.61%	99.61%	99.61%	99.56%	99.61%	99.59%	99.6%	99.6%
	Spanish	0.39%	0.39%	0.39%	0.44%	0.39%	0.41%	0.40%	0.40%
Word-to-Word Dictionary	No	99.33%	99.36%	99.33%	99.32%	99.4%	99.38%	99.38%	99.39%
	Yes	0.67%	0.64%	0.67%	0.68%	0.60%	0.62%	0.62%	0.61%
Screen Reader	No	99.96%	99.87%	99.96%	99.87%	99.96%	99.87%	99.96%	99.86%
	Yes	0.04%	0.13%	0.04%	0.13%	0.04%	0.13%	0.04%	0.14%
Non-Screen Reader	No	99.95%	99.95%	99.95%	99.95%	99.95%	99.96%	99.95%	99.95%
	Yes	0.05%	0.05%	0.05%	0.05%	0.05%	0.04%	0.05%	0.05%
Audio and Orienting Description	No	99.99%	100%	99.98%	99.99%	99.98%	100%	99.99%	99.99%
	Yes	0.01%	0.00%	0.02%	0.01%	0.02%	0.00%	0.01%	0.01%
Magnification	No	100%	100%	100%	100%	100%	100%	100%	100%
Scribe (speech-to-text)	No	100%	100%	100%	100%	100%	100%	100%	100%
Other Assistive Technology	No	100%	100%	100%	100%	100%	100%	100%	100%
Braille	Non-Braille	100%	100%	100%	100%	100%	100%	100%	100%
Paper Accommodation	Regular Print	99.96%	99.96%	99.96%	99.97%	99.96%	99.97%	99.95%	99.96%
	Large Print	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Standard Print	0.04%	0.03%	0.04%	0.03%	0.04%	0.03%	0.04%	0.03%
Extra Time	Regular Time	90.62%	91.25%	90.59%	91.05%	90.75%	91.1%	90.65%	91.33%
	Time and a Half	7.48%	7.57%	7.5%	7.75%	7.31%	7.71%	7.44%	7.47%
	Double Time	1.31%	0.84%	1.33%	0.87%	1.34%	0.87%	1.34%	0.87%
	Triple Time	0.58%	0.33%	0.58%	0.33%	0.59%	0.33%	0.56%	0.33%
Stop clock - Supervised Breaks	No	97.74%	98.85%	97.72%	98.85%	97.71%	98.84%	97.73%	98.84%

Accommodation	Test Group	English		Reading		Math		Science	
		Gr. 9	Gr. 10	Gr. 9	Gr. 10	Gr. 9	Gr. 10	Gr. 9	Gr. 10
	Yes	2.26%	1.15%	2.28%	1.15%	2.29%	1.16%	2.27%	1.16%
Secure Multi-day Break	No	98.17%	98.5%	98.17%	98.57%	98.23%	98.63%	98.22%	98.7%
	Yes	1.83%	1.5%	1.83%	1.43%	1.77%	1.37%	1.78%	1.3%
Signed Exact English - Directions Only	No	100%	99.99%	100%	99.99%	100%	99.99%	100%	99.99%
	Yes	–	0.01%	–	0.01%	–	0.01%	–	0.01%
Sign Language Interpretation	No	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	100%	99.99%
	Yes	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0%	0.01%
Cued Speech	No	100%	99.99%	100%	99.99%	100%	99.99%	100%	99.99%
	Yes	–	0.01%	–	0.01%	–	0.01%	–	0.01%
Auditory Notification of Remaining Time	No	99.89%	99.98%	99.89%	99.98%	99.9%	99.97%	99.9%	99.97%
	Yes	0.11%	0.02%	0.11%	0.02%	0.1%	0.03%	0.1%	0.03%
Abacus	No	100%	100%	100%	100%	100%	100%	100%	100%
Human Reader	No	100%	100%	100%	100%	100%	100%	100%	100%
	Yes	0%	–	0%	–	0%	–	0%	–
Human Scribe	No	99.99%	100%	99.99%	100%	100%	100%	99.99%	100%
	Yes	0.01%	0%	0.01%	0%	0%	0%	0.01%	0%
Home Administration	No	100%	100%	100%	100%	100%	100%	100%	100%

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 earlier in this chapter, the operational testing times for the Utah Aspire Plus tests are: 45 minutes for English, 90 minutes for Reading, 90 minutes for Math, 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 2019 test administration, student total testing time was analyzed for each test. Overall, students completed the assessments within the recommended testing times. Table 16 provides 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 42 minutes or less.

Additional information is presented in Appendix D, 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 (for example, one outlier is shown for grade 9 English regular testers above the distribution, of roughly 40,000 testers). 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 16. Student Testing Time for Spring 2019 Utah Aspire Plus

Subject	Grade	Group	*N*-count	Testing Time (minutes)									
				Descriptive Statistics				Percentiles					
				Minimum	Maximum	Mean	St. Dev.	P50	P75	P80	P85	P90	P95
English	9	Regular Time	40,387	1	74	30	8	30	36	37	39	40	42
		Time and a Half	3,242	2	100	32	13	31	40	43	45	50	56
		Double Time	562	2	91	34	16	32	42	44	47	54	63
		Triple Time	240	4	117	38	18	36	48	51	55	59	69
	10	Regular Time	38,543	1	63	30	8	30	36	37	39	41	42
		Time and a Half	3,142	2	74	34	14	33	43	45	49	53	59
		Double Time	347	1	157	35	16	34	42	44	47	54	65
		Triple Time	133	2	110	37	19	35	45	48	53	59	74
Reading	9	Regular Time	40,365	1	114	47	15	47	57	59	63	67	73
		Time and a Half	3,250	1	134	48	23	47	62	66	71	78	91
		Double Time	573	3	172	50	27	47	64	70	76	86	100
		Triple Time	243	4	186	50	31	43	64	73	78	88	108
	10	Regular Time	38,459	1	94	43	15	43	53	56	59	63	69
		Time and a Half	3,214	2	134	47	22	46	60	64	69	74	85
		Double Time	358	2	276	41	25	38	51	55	62	68	78
		Triple Time	134	2	155	46	31	39	61	68	73	85	109
Math	9	Regular Time	40,411	1	154	57	16	58	69	72	75	79	83
		Time and a Half	3,194	2	158	52	25	50	67	71	77	83	95
		Double Time	581	2	177	56	28	51	71	76	83	91	106
		Triple Time	245	4	166	56	30	51	75	79	84	92	113
	10	Regular Time	38,462	1	97	50	17	50	62	65	68	72	78
		Time and a Half	3,207	2	133	48	23	46	61	66	71	77	87
		Double Time	361	2	168	49	26	48	64	67	75	82	90
		Triple Time	135	3	263	49	34	46	60	65.5	70	81	97

Subject	Grade	Group	*N*-count	Testing Time (minutes)									
				Descriptive Statistics				Percentiles					
				Minimum	Maximum	Mean	St. Dev.	P50	P75	P80	P85	P90	P95
Science	9	Regular Time	40,382	1	59	33	10	33	40	42	43	46	50
		Time and a Half	3,224	1	108	33	16	31	43	45	49	53	61
		Double Time	583	3	118	37	18	35	47	51	55	60	70
		Triple Time	242	3	132	38	19	36	47	51	55	62	73
	10	Regular Time	38,561	1	83	32	11	32	39	41	43	46	50
		Time and a Half	3,111	1	89	32	17	30	42	46	49	54	62
		Double Time	360	2	114	32	19	31	42	45	49	56	67
		Triple Time	133	1	136	35	24	30	45	49	53	60	77

4. Classical Item Analyses

4.1 Item Analyses

In the Test Development chapter, 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 initial forms were created, there were instances in which not all statistical targets were fully met. This section describes in more detail those classical item statistics. Additionally, after the Utah Aspire Plus 2018–2019 operational administration, classical item statistics were also calculated and results are presented in Appendix E.

4.1.1 *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.

4.1.2 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 Appendix E, 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.

4.1.3 Differential Item Functioning

Differential item functioning (DIF) exists when an item functions differentially across identifiable subgroups (e.g., gender 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 17, the number of focal and reference group members scoring in each possible item response is captured.

Table 17. Item 2x2 Contingency Table for the *k*th 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.

4.2 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 E. 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, which 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 initial operational 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. For example, no item had an item-total correlation below 0.15 (below threshold, but still functional). And overall, the distributions of all of the statistics were within acceptable ranges for large-scale summative assessments.

5. 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.

5.1 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.

5.2 Classical Test Theory Reliability Estimates

5.2.1 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 F for the overall testing population as well as by gender, 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*).

Test-level reliabilities for all of the Utah Aspire Plus tests were reasonable in relation to what is generally found from high-stakes large-scale state summative tests, with most at roughly 0.90 and the lowest overall at 0.88 (Grade 9 Reading). And while scores are not reported to students for subscores, it is worth noting that many were very low (as low as 0.20, with several below 0.50). In all instances, these appear to be related directly to the low numbers of items making up the given score. Hence, to the extent that there is a desire to report subscore information to individual students moving forward, ensuring more items contribute to each would be encouraged.

5.2.2 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 F and are presented on the Utah Aspire Plus scale score metric ($s_x = 25$).

5.3 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 the 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 his or her 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 G, 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.

Conditional standard errors for the Utah Aspire Plus tests were all roughly 8 scale score points in the region of the Proficiency cuts. Examination of the curves showed these points were generally at the lowest conditional errors, respectively (highest measurement precision). And as in the nature of conditional errors along a score scale, the greatest error was at the lowest and highest ends of the scale and consistently low across the range 150 to 250 on each respective test. It should be noted that conditional errors in this area are also consistent with the SEMs as well and would be considered in line with what is generally observed on similar high-stakes state summative tests.

5.4 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 the section on the SEM, a student's true score is most likely to fall into a standard error band around his or her 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.

5.4.1 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—"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.

5.4.2 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 18 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 18. Example Accuracy Classification Table

True Score	Observed Score				Total
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
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 19 indicated 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 19. Example Accuracy Classification Table for Proficient Cut Point

True Score	Observed Score				Total
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
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

5.4.3 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 20 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 20. Example Consistency Classification Table

First Form	Second Form				Total
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
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

5.4.4 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 20, 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 20 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 H. For all Utah Aspire Plus tests, classifications were all moderate to substantial.

6. IRT Calibration and Scaling

6.1 Overview

The purpose of the item response theory (IRT) calibration and scaling was to place all operational items, Utah legacy and ACT Aspire items, for each test onto a common scale. This process was used to establish the base Utah Aspire Plus base scales to which future administrations can be compared. Once items are calibrated, the item parameters are used to compute a student's score in the IRT metric. In this section of the technical report, the following topics related to IRT calibration and scoring are discussed:

- IRT Data Preparation
- Description of the Calibration Process
- Model Fit Evaluation Criteria
- Summary Statistics and Distributions from IRT Analyses
- IRT Scoring

6.2 IRT Data Preparation

6.2.1 Student Inclusion/Exclusion Rules

The data preparation for the IRT calibration process began with all Utah students that were administered the "base" forms (i.e., online, English-language forms). Special handling for students taking the accommodation forms is discussed in a later section.

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 all subject tests within a grade.

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

6.2.2 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 (not applicable to Spring 2019), student records would first be sorted by form, then by student identification number. The array of item responses were presented in the order as administered in the test form, including items that are presented in field test slots (placeholders for Spring 2019).

The IRT data matrices were created independently by two Pearson psychometric staff. The matrices were checked for accuracy by comparing numbering of students (counts) and the item response arrays. Any discrepancy found was resolved. Final calibration data files matched perfectly.

6.3 Description of the Calibration Process

6.3.1 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).

6.3.2 IRTPRO Calibration Procedures and Convergence Criteria

The primary goal of IRT calibration is to place the operational items from a given test onto a common scale. As this was the first administration of Utah Aspire Plus assessments, these calibrations result in the base scales to which future assessments will be related to. The following is a description of the steps to calibrate the operational item response data. 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 the previous section. 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 is 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:

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

6.3.3 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.

6.4 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 (χ^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. MAP (maximum *a posteriori*) estimates from IRTPRO were used as student ability estimates. 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 degrees of freedom. The 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 noted, item fit plots were produced and reviewed in addition to Q_1. Upon inspection, item plots were reasonable and did not suggest model selection was a concern in any instance. Very few items were flagged during the Q_1 analyses, which was consistent with the item plots. Of those items that were flagged, for English, all grade 9 items flagged for misfit were short text-entry items on Conventions of Standard English. In grade 10, the flagged items were a mixture of short text-entry and multiple-choice items. For Reading, most of the items flagged for misfit were evidence-based selected response (EBSR) items. These are two-part items where students' response to Part B should depend on the response to Part A. Students can only get credit for Part B if Part A is correct. There were four such items flagged in grade 9 and three flagged items in grade 10. The remaining flagged items were a mixture of technology-enhanced type and multiple-choice items. For Mathematics, one flagged item in each grade involved students using the equation editor tool to construct the response to the item. Both items, though, were statistically close to adequate model fit. Two other items in grade 9 involved students' interpretation of a graph in order to select the correct response to the items.

It should be noted that evaluation of fit was part of the operational calibration process and not yet part of the test construction process. As the Utah Aspire Plus program matures, model fit information will be part of item selection and should lead to further improvement of the measurement characteristics of the assessments.

6.5 Summary Statistics and Distributions from IRT Analyses

Tables 21 through 23 present the summary statistics for the IRT (*a*-, and *b*-) parameter estimates, standard errors (SE) of the parameter estimates, and model fit information for the spring 2019 operational items. The summary statistics shown include the total number of items, along with the mean, standard deviation (SD), minimum, and maximum.

Table 21. 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	45	0.83	0.33	0.27	1.83	–0.19	1.02	–1.93	3.01
	Reading	35	0.84	0.39	0.18	1.56	0.17	1.31	–1.13	6.02
	Mathematics	40	1.05	0.32	0.54	1.82	0.17	0.93	–1.87	1.80
	Science	36	0.96	0.34	0.36	1.65	–0.17	0.59	–1.49	1.07
10	English	48	0.91	0.37	0.31	1.63	–0.13	1.17	–1.88	4.72
	Reading	35	0.97	0.38	0.22	2.07	–0.35	0.88	–1.98	1.28
	Mathematics	39	1.12	0.29	0.39	1.74	0.35	0.80	–1.26	1.55
	Science	36	1.05	0.65	-2.20	1.94	–0.06	0.81	–2.36	1.82

Table 22. IRT Standard Errors of Parameter Estimates for Utah Aspire Plus Operational Items

			SE of *a* Estimates				SE of *b* Estimates			
Grade	Subject	No. of Items	Mean	SD	Min	Max	Mean	SD	Min	Max
9	English	45	0.03	0.02	0.01	0.10	0.04	0.07	0.01	0.48
	Reading	35	0.03	0.02	0.01	0.08	0.04	0.05	0.01	0.29
	Mathematics	40	0.04	0.02	0.02	0.10	0.03	0.02	0.01	0.08
	Science	36	0.04	0.02	0.02	0.09	0.04	0.03	0.01	0.14
10	English	48	0.04	0.02	0.01	0.08	0.03	0.03	0.01	0.16
	Reading	35	0.04	0.02	0.01	0.07	0.03	0.03	0.01	0.15
	Mathematics	39	0.05	0.01	0.01	0.07	0.03	0.03	0.01	0.20
	Science	36	0.06	0.05	0.01	0.35	0.03	0.04	0.01	0.19

Table 23. IRT Model Fit for Utah Aspire Plus Operational Items

			Q_1			
Grade	Subject	No. of Items	Mean	SD	Min	Max
9	English	45	206.8	256.7	17.5	1532.5
	Reading	35	392.8	519.6	63.2	2009.7
	Mathematics	40	186.9	177.6	39.5	954.8
	Science	36	135.6	82.0	37.1	427.8
10	English	48	190.1	250.4	35.8	1321.4
	Reading	35	356.0	495.3	41.9	2470.7
	Mathematics	39	139.0	134.3	39.1	762.6
	Science	36	190.1	418.9	34.7	2601.8

6.6 IRT Pattern Scoring

Item parameters derived from the 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 to the item characteristics across all items.

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: student response file and 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 derived for *all* 2019 Utah Aspire Plus operational items were used for estimating individual student scores for all regular forms. It should be noted that several of the accommodated forms had items that differed substantively from the source items such that they were not included in providing individual student scores. This was due to the inability to freely estimate those items due to very low sample sizes (where for all other items, IRT parameters from the general forms were used for scoring of the accommodated forms). In those instances, roughly ninety percent or more of the operational items were used to produce student scores and determine performance level indices (a few instances where 4 items were impacted, mostly 1 or 2 items total). In all cases the blueprint coverage was reasonable and overall scores reflected the asserted claims. Moving forward, all future Utah Aspire Plus items that need modification for the specific accommodation will be created as "sisters" and based entirely on the originating source items as described previously in the section on test accommodations and supports.

6.6.1 Quality Control of IRT Scoring

IRT pattern scoring is replicated independently through two parties internally. This scoring was conducted at the overall test level as well as by reporting categories. Any differences are resolved and rerun until both parties' results are identical and deemed correct based on careful examination of output.

7. Score Reporting

7.1 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 incorporate 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 online. 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*.

7.2 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:

$$SS = Theta * 25 + 200 \cdot$$

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 Reading and English 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.

7.3 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.

In November 2018, Utah educators were convened to create and recommend the PLDs for Utah Aspire Plus. This process began with a review of the Utah SAGE PLDs in light of the context of college readiness within the Utah Aspire Plus framework. Appendix I contains the agenda for educator groups convened for this process and the general training used to introduce the educators to this process. The approved PLDs can be found online.

In August 2019, 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. A separate report of the standard-setting process includes a demographic summary of the educators that participated in that process, a detailed description of the standard-setting process, and the outcomes.

7.4 ACT Predicted Score Ranges

As noted throughout, 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, math, reading, 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. To provide the predicted performance on the ACT

tests in the first two administration years, a linking study was performed between scale scores of the Utah Aspire Plus assessments and established ACT score predictions for ACT Aspire® tests. The link was facilitated through common items between Utah Aspire Plus and ACT Aspire® test forms. The result of this linking study is a set of predicted ACT score ranges across the Utah Aspire Plus score scale (100–300) for each Utah Aspire Plus assessment. The predicted ACT score ranges will be updated when longitudinal data become available to directly link the Utah Aspire Plus scores of grade 9 and 10 to ACT scores at grade 11.

A separate report (see Appendix J) provides the details of the Utah-to-ACT linking study, including the rationale for establishing an appropriate width of the prediction ranges. Predicted ACT score ranges for Utah Aspire Plus scale scores were provided for each subject test and the Composite. For each score, student's predicted ACT score range contains the most likely ACT scores that the student would obtain when taking the ACT test during the 11th grade. Appendix K provides predicted ACT score ranges for Utah Aspire Plus scale scores for each test and the Composite score.

7.5 2018–2019 Utah Aspire Plus Performance Results

Descriptive statistics of the scale scores for each Utah Aspire Plus assessment are in Appendix L. The descriptive statistics are provided for the overall testing population, as well as by subgroups—gender, 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. With respect to skewness, values between –0.50 and 0.50 are generally considered approximately symmetric, whereas greater values indicate moderate or more skew.

Scale score distributions for each Utah Aspire Plus assessment are provided in Appendix M, for the overall testing population. Note that scores at respective ends of the plots are given the LOSS and HOSS, respectively, and appear as spikes on each graph. These reflect scores that fall below and above the bounds of the scale score ranges (100 to 300) and are each given the lowest or highest scores for convention.

Appendix N contains the performance level distributions of each Utah Aspire Plus using the approved cut points from the standard-setting process. The tables contain the percentages of students being classified into each respective performance level.

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

PearsonAccess is Pearson's next-generation system for managing student data, paper, and online test administration, scoring, and reporting high-stakes assessments. This system 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 solution. TestNav delivers online tests. 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.

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 his or her log-in and password on the testing workstation to gain access to the test. To further secure the testing environment, a blacklist capability sends

notifications when unapproved applications are running when the test is started. Once all blacklisted 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 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. Appendix O details 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.

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, 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 and the Intended Learning Outcomes for science 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 of this report, 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 (http://utah.pearsonaccessnext.com/additional-services/).

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, as measuring intended targets of measurement, appropriately aligned to a DOK level, and that vocabulary is appropriate for the given level, the content is accurate and straightforward, supporting graphics or stimuli are necessary to answer the question, and items are clear and concise. 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 gender, 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, to include 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 C.

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 in Chapter 2, content experts and stakeholders participated in standard setting in August 2019. These committee 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 7.

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, N. L., 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 2016–2017 technical report. In Section 2.3.4, 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 where it's noted that the content reflects expectations that students need to think, reason, and analyze at high levels of cognitive complexity in order to be college- and career-ready and that items and tasks require sampling different levels of cognitive complexity with most targeted at upper levels. Their 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://www.act.org/content/dam/act/unsecured/documents/2019/aspire/Aspire-Summative-Technical-Manual.pdf. Here they point to piloting of ACT Aspire CR items using think-aloud tasks, surveys, and interviews as providing evidence of process to intended targets.

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 P. Here, the first eigenvalue is substantially larger than the second in all instances and indicative of essential unidimensionality. This type of result in a scree plot is evidence the Utah Aspire Plus tests are measuring a single dimension and suggests the application of the IRT models is appropriate.

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 24. These estimates provide additional evidence of a one-factor construct for the Utah Aspire Plus tests.

Table 24. Model Fit Indices for Confirmatory Factor Analyses

Subject	Grade	SRMR	RMSEA	RMSEA 90% Lower CL	RMSEA 90% Upper CL
English	9	0.0308	0.0346	0.0344	0.0349
	10	0.0315	0.0344	0.0341	0.0346
Reading	9	0.0154	0.0186	0.0183	0.0189
	10	0.0315	0.0352	0.0348	0.0355
Mathematics	9	0.0298	0.0332	0.0329	0.0335
	10	0.0289	0.0316	0.0313	0.0319
Science	9	0.0180	0.0230	0.0227	0.0233
	10	0.0240	0.0286	0.0283	0.0289

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 Chapter 6 with Q_1 indices for all Utah Aspire Plus operational items. 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 Q. These correlations show that the subcomponents of the overall test are moderately to highly related to one another but more strongly related to the total test score.

9.3.1 Reliability

Additionally, the reliability analyses presented in Chapter 5 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. As noted, reliabilities for the overall test level scores were roughly 0.90 in all instances (where the lowest was 0.88). These reliability estimates further indicate that the items that form each Utah Aspire Plus test are measuring the same construct and provide further evidence of unidimensionality.

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 find 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 gender, 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. Chapter 4 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 initial creation and 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.

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.

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.

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

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.

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 Blueprint Educator Committee

Blueprint Review Meetings

English and Reading – June 7–8

Day 1

1. Welcome and overview, general purpose for the meeting, housekeeping (USBE and Pearson)
2. Math and Science break into content area rooms; English and Reading remain in room
3. Purpose and goals of the meeting (English/Reading combined)
4. Review of the agenda

    Proposed Agenda

    Overview of new High School Assessments—College and Career Readiness

    What is a blueprint and what is its purpose? (Example of a generic blueprint)

    Overview of inputs used to develop the new blueprint (the outcome)

    i. What is the desired outcome? A customized hybrid of ACT Aspire and UT SAGE
        1. Predictive ability of ACT Aspire for performance on the ACT. Discuss some high-level information as an overview and purpose of the new assessment.
        2. Measure of progress for students in Grade 9 and Grade 10. This is a shift from a traditional summative end-of-course assessment.
    ii. ACT Aspire Inputs
        1. The ACT Aspire Reading and English blueprints (for reporting categories, number of items, item types, and DOK distribution)
    iii. Utah SAGE Inputs
        1. The SAGE Grades 9 and 10 blueprints (for reporting categories, percentage of items per category, and DOK distribution)
    iv. Utah Core Standards
    v. Shift from one assessment to two: reading and English
    vi. Your Role for Break-Out Rooms:
        1. Respond to the following characteristics:
            a. timing
            b. number of items per category
            c. item types
            d. number of points
            e. standards within a reporting category
            f. reading load (word count, Lexile) vs. timing

Day 2 (English and Reading content areas break into separate rooms)

5. Begin Blueprint Review
   Process for developing blueprint –

   i. Thinking through intent of the assessment
      1. Predictive ability of ACT Aspire for performance on the ACT. Discuss some high-level information as an overview and purpose of the new assessment.
      2. Measure of progress for students in Grade 9 and Grade 10. This is a shift from a traditional summative end-of-course assessment.
      3. ACTIVITY: Utah Core Standards—which standards are eligible for assessment?
      4. ACTIVITY: How will the assessment approach each eligible standard? (Consider passage types, how many items, the item types, and DOK level that may be necessary to provide the most comprehensive information for students, teachers, and schools. Does the blueprint capture this?)
      5. ACTIVITY: The classroom perspective: What does this look like in the classroom? How would this change affect what is done in the classroom?

   ii. Understanding the impact of the removal of constructed response items.

Blueprint Review Meetings

Math – June 7–8

Day 1

1. Welcome and overview, general purpose for the meeting, housekeeping (USBE and Pearson)
2. Break into content area rooms
3. Introductions
4. Purpose and goals of the meeting
5. Overview of new High School Assessments—College and Career Readiness
    a. What is a blueprint and what is its purpose? (Example of a generic blueprint)
    b. Overview of inputs used to develop the new blueprint (the outcome)
        i. What is the desired outcome? A customized hybrid of ACT Aspire and UT SAGE
            1. Predictive ability of ACT Aspire for performance on the ACT. Discuss some high-level information as an overview and purpose of the new assessment.
            2. Measure of progress for students in Math I and Math II
        ii. Item Types
            1. ACT Aspire Items – Exemplar Items
            2. Utah SAGE – Practice Test
        iii. DOK overview
        iv. ACT Aspire Inputs
            1. The ACT Aspire Math blueprints (for reporting categories, number of items, item types, and DOK distribution)
        v. Utah SAGE Inputs
            1. The SAGE blueprints (for reporting categories, percentage of items per category, and DOK distribution)
        vi. Utah Core Standards
6. Begin Blueprint Review
   ACTIVITY: Breakout into small groups and review the following characteristics:

    a. Item types
    b. DOK
    c. Reporting categories
    d. Strands within a reporting category
    e. Number of items/points
    f. Number and types of items vs. timing

Day 2

Morning – Math I

Afternoon – Math II

1. Begin process of detailed blueprint review
   ACTIVITY: Utah Core Standards—which standards are eligible for assessment?

   ACTIVITY: Breakout into small groups and review each Math strand to determine point distribution within the strand and across each standard.

Utah Science Blueprint Meeting Agenda

Day 1

1. Test Blueprint 101
    a. Understanding an assessment blueprint
    b. Using the assessment blueprint
        i. Understanding stakeholder requirements
            1. The content perspective
            2. The psychometric perspective
            3. The basis for test production
            4. **Discussion topic:** Prioritizing stakeholder perspectives
        ii. Understanding how a blueprint is created for a new assessment program
        iii. Assessing the standards
        iv. DOK and cognitive complexity
        v. Timing considerations
        vi. **Discussion topic:** Balancing measuring student performance against time limits
2. The Hybrid test
    a. What it will measure
        i. Predictive ability of ACT Aspire
        ii. Grade level progress
    b. What inputs were used in creating the blueprint
        i. Review of the ACT Aspire blueprint
        ii. Review of the Utah Sage blueprint
    c. **Discussion topic:** The challenge of creating the hybrid blueprint

Day 2

1. Use of Intended Learning Outcomes (ILOs) in contrast to specific content standards
2. How ILOs match up with ACT Aspire Reporting categories
    a. Optimizing similarities
    b. **Discussion topic:** The benefits of using the ILOs to measure student science process
3. Cognitive complexity
    a. Considering a method based on levels of science engagement
4. Grade Level Content
    a. The challenge of assessing more than one course at each grade level
        i. Grade 9 blending Earth Science and Biology
        ii. Grade 10 blending Biology, Chemistry, and Physics
        iii. **Discussion topic:** Finding overlap in course content
5. Likely layout of the form
    a. Creating an integrated and seamless form with distinctly different components
6. Discussion about what can and can't be measured with this assessment
    a. Going from a computer-adaptive test (CAT) to a linear assessment
    b. Transitioning from teacher effectiveness to student progress
    c. **Discussion topic:** Views on the draft Utah hybrid blueprint for science

Utah Aspire Plus Assessments for Grades 9-10

Educator Blueprint Review

June 7-8, 2018

Pearson

Agenda Day 1

1. Welcome and housekeeping
2. Purpose and goals of meeting
3. History of Utah Assessments
4. Break into content area groups
5. Overview of Utah Aspire Plus Assessments
6. General background of blueprints
7. Process for development of new blueprints
8. Start detailed discussion of new blueprints

Pearson

Agenda Day 2

1. Breakfast
2. Break into content groups
3. Review of Day 1
4. Continued detailed review and discussion of new blueprints
5. Recommendations
6. Closing

History of Utah Assessments

Pearson

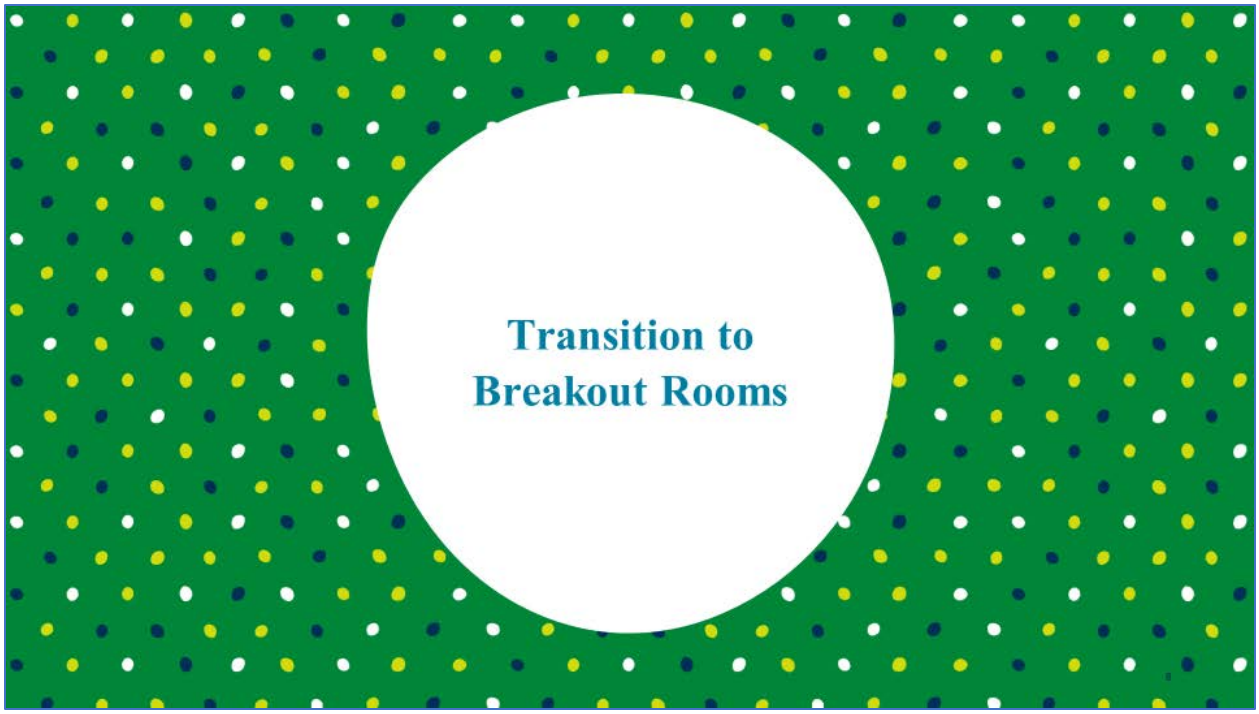
SAGE Assessment

New Standards
Blueprint CAT
Performance Level Descriptors
Item Development
Standard Setting
Legislation

- Predictive of College and Career Assessment
- Growth 9th to 10th

Timed
Hybrid- link to 3rd-8th and to ACT

Pearson


Transition to
Breakout Rooms

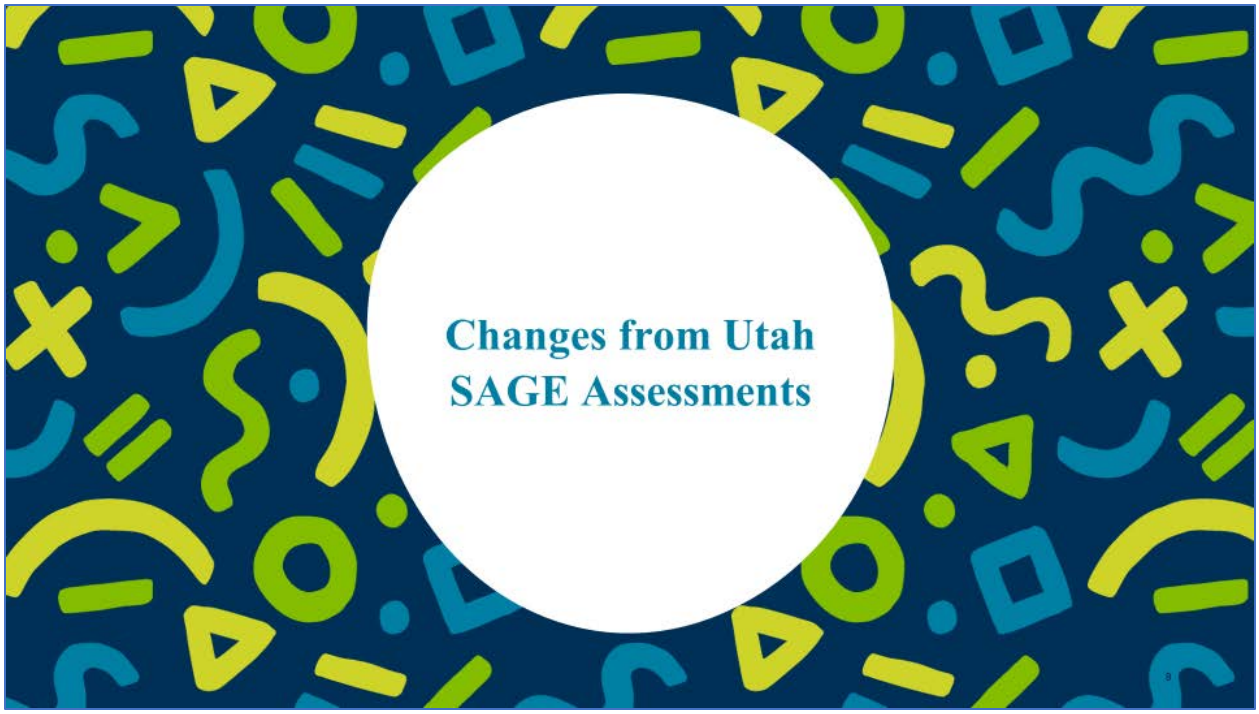
New Utah Assessments for Grades 9-10: Utah Aspire Plus

Pearson

Utah Aspire Plus 9-10 Assessments

- Utah-based test created by Utah educators with Utah-aligned ACT Aspire ® content embedded
- 100% aligned to the Utah Core Standards
- Measures growth and readiness
- Strong prediction of performance on The ACT® Test and projected ACT National Career Readiness Certificate (ACT NCRC®) achievement level
- Same platforms as The ACT Test
  - Seamless experience for students from grades 9-11
  - Simplifies training and technology support for high schools

Pearson



Changes from Utah SAGE to Utah Aspire Plus 9-10 Assessments

- Linear rather than CAT Assessment
- English and Reading are two separate assessments
- 50% of the content will come from ACT Aspire
- Going from EOC to Grade Level Assessments

Utah Blueprint Meeting – Final Summary

English

Overall, educators accepted this draft blueprint with few modifications. Since it is essentially the Aspire blueprint, the overall number of items and timing felt appropriate. There were a few modifications suggested.

Reporting Categories—Overall these were accepted as they were represented, although educators felt the Knowledge of Language category should be emphasized slightly more. An adjustment was made to this category to include the potential for more items.

Grade 9 and Grade 10—Comments about the potential difference between the Grade 9 and Grade 10 assessments: Grade 9 is heavy on parallel structure, semicolon, and colon; Grade 10 is heavier on punctuating complex and compound sentences.

Language Progressive Skills, by Grade chart represented in the Core Standards—Educators agreed that we could "reach down" to include the standards represented at lower grades, but obviously in items written at a grade-appropriate level (e.g., L.3.1f Ensure subject-verb and pronoun-antecedent agreement. Although a Grade 3 standard, this skill is still very relevant to Grades 9 and 10 and can be tested with appropriately complex sentences).

OTHER AREAS OF DISCUSSION TO NOTE

Reporting—Educators were very interested in what the reports would look like. The look and feel of the ACT reports and information included in the ACT reporting was mentioned as being informative and helpful.

Accommodations—Educators were also quite interested in what accommodations would be available for the new assessment.

Functionality—Educators wondered about the functionality of the new assessment and whether students would be able to return to items and/or passage sets as they work through the test. In a timed situation, educators felt like this would be important for students to be able to toggle back and forth in order to attack what they felt most comfortable with first, perhaps, and then return to the more challenging passages/items.

Growth and Predictability—There was quite a bit of discussion about the ability to continue to see growth. How will this be messaged to the field? There was general agreement that the predictive nature of this new test would be appealing to the field.

Reading

Item types and number of items—Generally acceptable, although a good portion of the committee wanted to see more EBSRs since EBSRs may offer an opportunity to get to the higher DOK items. The number of EBSRs on the blueprint was adjusted slightly higher. For the same reason, the number of TEIs was adjusted slightly higher.

Depth of Knowledge spread—Slight modifications were made to these spreads. Educators felt that while we shouldn't go below 4 items for DOK 1, that category shouldn't represent as large of a percentage of the test. The overall percentage was reduced slightly and the other two categories were increased as a result.

Reporting categories—Educators were comfortable with the reporting categories since they align to the Utah Core Standards. There was general agreement that the Integration of Knowledge and Ideas (IKI) category was important and should be stressed more. The limitations of the items coming from Aspire in this category make it a bit challenging to increase as much as the committee might like, but additional items were added. It should be noted in test construction specifications that, when possible, passage sets with viable items in the IKI category should be prioritized.

Standards alignment—Overall, educators agreed that the new assessment should align with and assess the Grade 9/10 Utah Core Standards. One area of deviation was RI.9-10 (*Analyze seminal U.S. documents of historical and literary significance, including how they address related themes and concepts*). After much discussion, educators recommended that it was indeed important for students to analyze related historical documents, but that they don't necessarily need to be seminal to be appropriate for the assessment. (How a document is considered "seminal" and who makes that decision was a sticking point.) Educators agreed: Students need to be able to read and analyze historical documents.

Information vs. Literary balance—Educators felt this balance was appropriate.

Word count and timing—There was quite a bit of discussion about overall reading load and timing of the test. Ultimately, the committee agreed to move forward with a 90-minute testing session as a recommendation. The word count was considered and accepted when considering the timing, although as Pearson assessment specialists are able to access the SAGE bank more fully and see specific passages, it could be possible to adjust the overall load. The spread could be slightly lower depending on the typical length of existing passages.

Make-up of a form—The educators discussed this quite a bit. Seven passages (which includes one field test set) seemed like quite a lot to most of the committee. However, after discussing the content expected from Aspire, what the SAGE content will need to support, and considering what students were used to from the SAGE assessment, the committee agreed that the form was appropriate. The overall number of operational items was acceptable with the additional field test items.

Math

The Math Educator committee reviewed the high-level blueprints for Math I and Math II and came to the following conclusions:

- The information presented and proposed for the percent breakdown of technology-enhanced items and multiple-choice/multiple-select items was appropriate at both grade levels.
- The information presented and proposed for the percent breakdown between the different levels of depth of knowledge was appropriate at both grade levels.
- Reporting categories:

  Math I
  - It was determined that since there were only 3 standards that fall under the strand "Number and Quantity" and all 3 standards are better assessed in the classroom and weaved into other standards, the 3 standards should not be formally assessed. This strand was removed.
  - The items that were previously in the Number and Quantity strand were moved to Functions, which was deemed to be a strong focus within the Math I standards.
  - The information presented and proposed for the percent breakdown of the remaining strands stayed the same.
  - There are 4 proposed reporting categories: Functions, Algebra, Geometry, and Statistics and Probability.

  Math II
  - Discussion focused around decreasing the percent of items assessed in Statistics and Probability to 2–4 items (5%–10%) and increasing focus in Functions and Geometry.
  - Two of the strands (Number and Quantity and Statistics and Probability) have too few items to qualify them to be a stand-alone reporting category.
  - Number and Quantity, Algebra, and Statistics and Probability will collapse into one reporting category.
  - The committee requested that the breakdown of number of items and the percent min/max for each strand be shared publicly since the collapsing of strands into one reporting category fails to make this information accessible and clear.

Science

The committee agreed that aligning the new blueprint to science process was the correct course for the hybrid assessment. The discussion regarding how well the ACT Aspire reporting categories aligned to the Intended Learning Objectives (ILOs) in the Utah Science Standards resulted in a consensus that while they were not exactly the same, they were quite comparable. Thus, the ILOs and Aspire reporting categories could be used as the basis for the alignment to process on the Utah Aspire Plus Science assessment. The concern about which content to assess in each grade was addressed by agreeing that all high school science content would be represented in each grade, but that the focus of the assessment would be on science process. However, there may need to be some glossing for specific terms to provide a leveling effect for all students.

Appendix B: Item Alignment Educator Committee

Utah Aspire Plus

Alignment Review Meeting Agenda
July 10, 2018

8:00 a.m. – 8:30 a.m. (MST)	Breakfast/Registration
8:30 a.m. – 9:00 a.m.	Orientation a. Welcome and Introductions b. Purpose and Goals of Meeting c. Overview of Utah Aspire Plus Assessments d. Housekeeping
9:00 a.m. – 10:15 a.m.	Review of Item Alignment ***Breakout Rooms*** English/Reading 1 English/Reading 2 Math Science a. Overview of Process and Procedures b. Begin Alignment Review
10:15 a.m. – 10:30 a.m.	Break
10:30 a.m. – 12:00 p.m.	Alignment Reviews ***Breakout Rooms***
12:00 p.m. – 12:45 p.m.	Lunch
12:45 p.m. – 2:00 p.m.	Alignment Reviews ***Breakout Rooms***
2:00 p.m. – 2:15 p.m.	Break
2:30 – 4:15 p.m.	Alignment Reviews ***Breakout Rooms***
4:15 – 4:30 p.m.	Wrap up and Adjourn

Utah Aspire Plus Assessments for Grades 9–10

Educator Review of Alignment

July 10, 2018

Pearson

Agenda

1. Welcome and housekeeping
2. History of Utah Assessments
3. Overview of Utah Aspire Plus Assessments
4. General background and process of alignment
5. Security and confidentiality
6. Transition to breakout rooms
7. Review of item alignment coding
8. Closing

Pearson

History of Utah Assessments

Pearson

SAGE Assessment

- New standards
- Blueprint CAT
- Performance-level descriptors
- Item development
- Standard-setting
- Legislation
  - Predictive of college and career assessment
  - Growth from 9th to 10th
- Timed
- Hybrid—link to 3rd–8th and to ACT

Pearson

New Utah Assessments for Grades 9–10: Utah Aspire Plus

Pearson

Utah Aspire Plus 9–10 Assessments

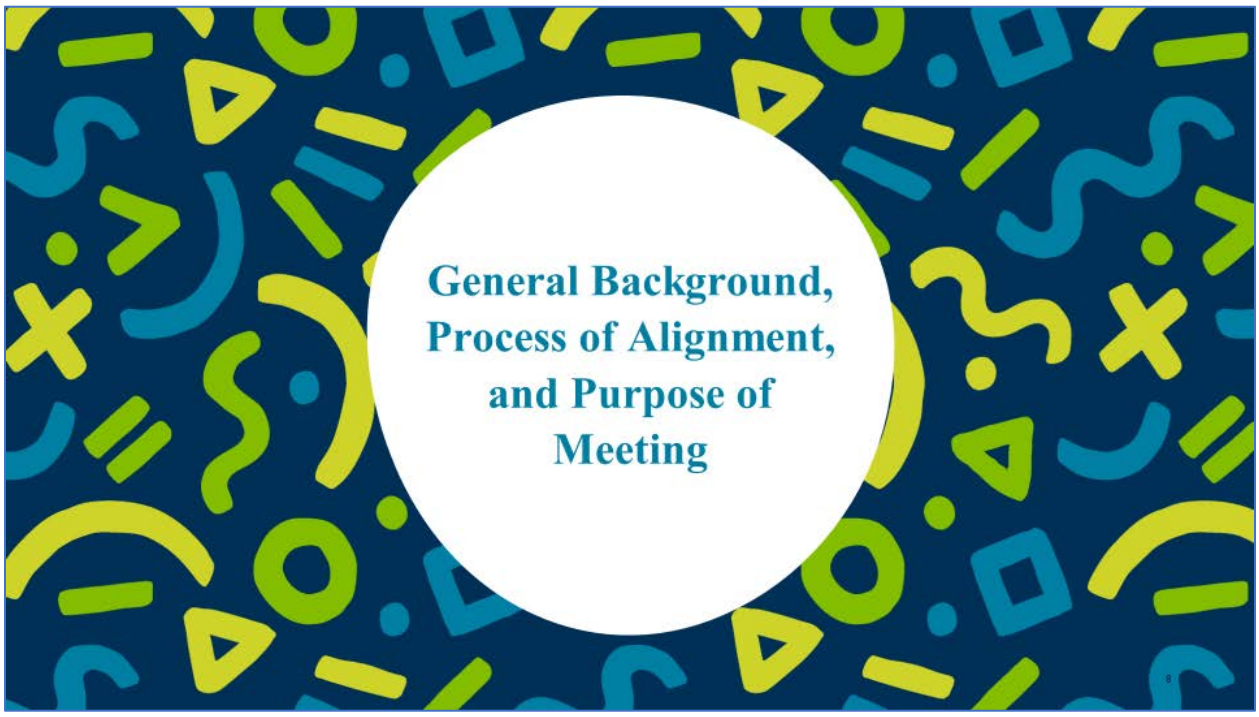
- Utah-based test created by Utah educators, with Utah-aligned ACT Aspire ® content embedded
- 100% aligned to the Utah Core Standards
- Measures growth and readiness
- Strong prediction of performance on the ACT® Test and projected ACT National Career Readiness Certificate (ACT NCRC®) achievement level
- Same platforms as the ACT Test
  - Seamless experience for students grades 9–11
  - Simplifies training and technology support for high schools

Pearson

Changes from Utah SAGE to Utah Aspire Plus 9–10 Assessments

- Linear rather than CAT assessment
- English and Reading are two separate assessments
- Approximately 50% ACT Aspire content
- Going from EOC to grade-level assessments

Pearson

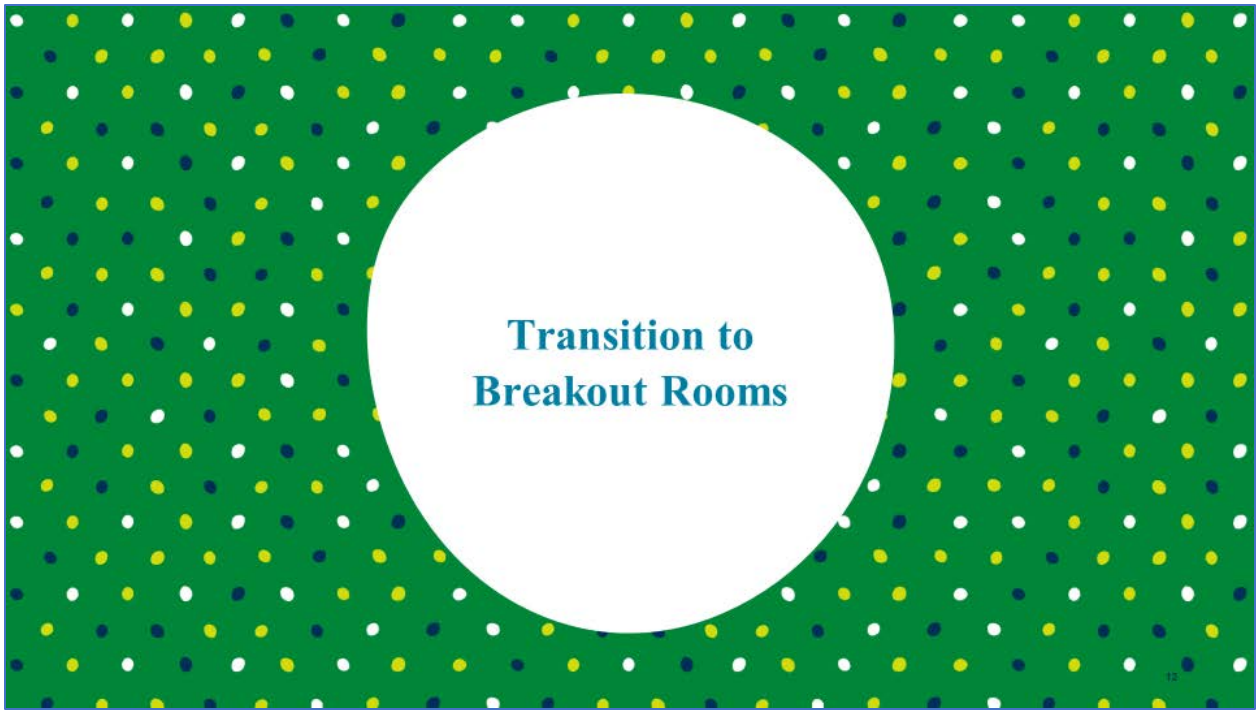


General Background and Process of Alignment

- The Pearson team reviewed and analyzed each ACT Aspire item and its corresponding ACT Aspire coding information in order to determine the alignment with a Utah Core Standard. The Utah coding was noted on spreadsheets.
- ACT reviewed each alignment and provided feedback.
- Today, each committee will review the Utah Core Standard alignment for each item. As a committee, we will determine whether the item alignment should be accepted or whether a different Utah Core Standard alignment should be recommended.

Security and Confidentiality

- Strict security procedures protect the content discussed.
  - Nondisclosure/confidentiality forms
- Technology considerations in committee meetings
  - Use of laptops
  - Use of personal electronics
- The process is not confidential; procedures and your training can be shared.
- The content of the items is confidential and secure. You have signed a confidentiality statement saying you will not share any information about the content of the items, passages, or stimuli.


Transition to
Breakout Rooms

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

English

Grade	Reporting Category: Standard	Multiple Choice			Technology Enhanced		
		DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
9	Production of Writing: W.9-10.4	1	0	8	0	0	0
	Knowledge of Language: L.9-10.3	1	1	2	0	0	0
	Conventions of Standard English: L.9-10.1	5	5	1	0	0	0
	Conventions of Standard English: L.9-10.1a	0	0	0	6	0	0
	Conventions of Standard English: L.9-10.1b	0	0	0	2	0	0
	Conventions of Standard English: L.9-10.2	6	0	0	0	0	0
	Conventions of Standard English: L.9-10.2a	0	0	0	5	0	0
	Conventions of Standard English: L.9-10.2c	0	0	0	1	0	0
	Conventions of Standard English: L.9-10.6	0	0	1	0	0	0
	Total	**45**					
10	Production of Writing: W.9-10.4	0	0	8	0	0	0
	Production of Writing: W.9-10.5	0	1	0	0	0	0
	Knowledge of Language: L.9-10.3	0	0	5	0	0	0
	Conventions of Standard English: L.9-10.1	8	5	0	0	0	0
	Conventions of Standard English: L.9-10.1a	0	0	0	3	0	0
	Conventions of Standard English: L.9-10.1b	0	0	0	2	0	0
	Conventions of Standard English: L.9-10.2	4	0	0	0	1	0
	Conventions of Standard English: L.9-10.2a	0	0	0	6	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	4	0	0
	Total	**48**					

Reading

Grade	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
9	Key Ideas: RI.9-10.1	3	3	0	0	1	0	0	0	0
	Key Ideas: RI.9-10.2	0	1	1	0	0	1	0	0	1
	Key Ideas: RI.9-10.3	0	2	0	0	0	0	0	0	0
	Key Ideas: RL.9-10.1	1	1	0	0	0	0	0	0	0
	Key Ideas: RL.9-10.2	0	1	0	0	0	0	0	0	0
	Key Ideas: RL.9-10.3	0	2	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.4	0	1	0	0	1	0	0	0	1
	Craft and Structure: RI.9-10.5	0	0	1	0	0	0	0	0	0
	Craft and Structure: RI.9-10.6	0	0	1	0	0	0	0	0	1
	Craft and Structure: RL.9-10.4	0	0	2	0	0	0	0	0	0
	Craft and Structure: RL.9-10.5	0	1	1	0	0	0	0	0	0
	Craft and Structure: RL.9-10.6	0	1	1	0	0	0	0	0	1
	Craft and Structure: L.9-10.5	0	0	0	0	0	0	0	0	1
	Integration of Knowledge and Ideas: RI.9-10.8	1	1	0	0	1	0	0	0	0
	Total	**35**								
10	Key Ideas: RI.9-10.1	2	1	1	0	1	0	0	0	0
	Key Ideas: RI.9-10.2	0	0	1	0	0	0	0	0	1
	Key Ideas: RI.9-10.3	0	1	0	0	2	0	0	0	0
	Key Ideas: RL.9-10.1	1	1	1	0	0	0	0	0	1
	Key Ideas: RL.9-10.2	0	1	1	0	0	0	0	0	0
	Key Ideas: RL.9-10.3	0	2	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.4	0	2	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.5	0	1	1	0	0	0	0	0	1
	Craft and Structure: RI.9-10.6	0	1	2	0	0	0	0	0	0
	Craft and Structure: RL.9-10.4	0	1	1	0	0	0	0	0	1

Grade	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
	Craft and Structure: RL.9-10.5	0	0	1	0	0	0	0	0	0
	Craft and Structure: RL.9-10.6	0	0	0	0	0	0	0	0	0
	Craft and Structure: L.9-10.4a	1	1	0	0	0	0	0	0	0
	Integration of Knowledge and Ideas: RI.9-10.7	0	0	0	0	0	0	0	0	1
	Integration of Knowledge and Ideas: RI.9-10.8	1	1	0	0	0	0	0	0	0
	Total	**35**								

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.2	0	0	1	0	0	1
Algebra: MI.A.CED.3	0	0	0	0	2	0
Algebra: MI.A.REI.12	0	1	0	0	0	0
Algebra: MI.A.REI.3	0	1	0	0	1	0
Algebra: MI.A.REI.3b	0	1	0	0	0	0
Algebra: MI.A.REI.6	0	0	0	0	0	2
Algebra: MI.A.SSE.1b	0	0	1	0	0	0
Functions: MI.F.IF.4	0	0	0	0	0	1
Functions: MI.F.IF.7a	0	1	0	0	0	0
Functions: MI.F.BF.2	0	1	0	0	0	0
Functions: MI.F.IF.1	1	0	0	0	0	0
Functions: MI.F.IF.2	1	0	0	0	0	0
Functions: MI.F.IF.6	1	0	0	0	0	0
Functions: MI.F.IF.7e	0	1	0	0	0	0
Functions: MI.F.LE.1b	0	0	0	0	0	1
Functions: MI.F.LE.2	1	1	0	0	0	0
Functions: MI.F.LE.5	0	1	0	0	0	0
Geometry: MI.G.CO.3	0	0	1	0	0	1
Geometry: MI.G.CO.5	1	1	0	0	0	0
Geometry: MI.G.CO.7	2	0	0	0	0	0
Geometry: MI.G.GPE.4	0	0	0	0	1	0
Geometry: MI.G.GPE.5	1	1	0	0	0	0
Geometry: MI.G.GPE.7	0	0	1	0	0	0
Statistics and Probability: MI.S.ID.1	1	1	0	0	0	0
Statistics and Probability: MI.S.ID.2	0	1	0	0	0	0
Statistics and Probability: MI.S.ID.6a	0	1	0	0	0	0
Statistics and Probability: MI.S.ID.6c	2	0	0	0	0	0
Statistics and Probability: MI.S.ID.7	1	0	0	0	0	0
Statistics and Probability: MI.S.ID.8	0	1	0	0	0	0
Total	**40**					

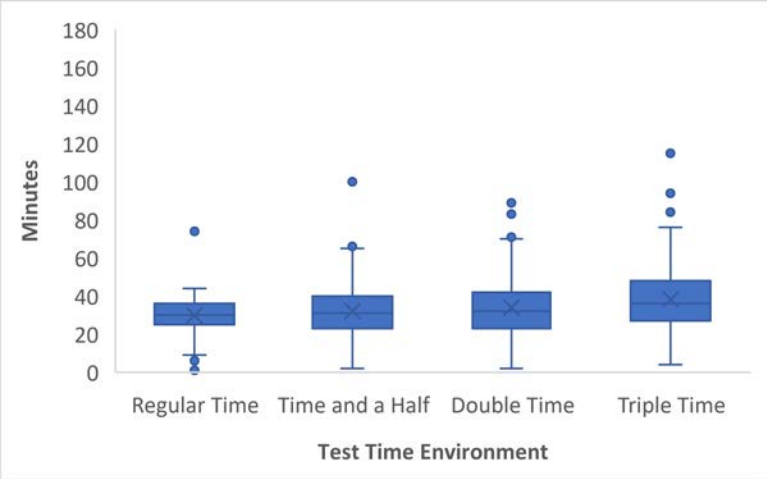
Mathematics – Grade 10

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.1	0	0	0	0	0	1
Number and Quantity: MII.N.RN.2	0	2	0	0	0	0
Number and Quantity: MII.N.RN.3	0	0	0	0	1	0
Algebra: MII.A.APR.1	1	1	0	0	0	0
Algebra: MII.A.CED.2	0	1	0	0	0	0
Algebra: MII.A.CED.4	1	0	0	0	0	0
Algebra: MII.A.REI.4a	0	1	0	0	0	0
Algebra: MII.A.REI.4b	0	1	0	0	0	0
Algebra: MII.A.REI.7	1	0	0	0	0	0
Algebra: MII.A.SSE.1a	0	0	0	0	0	1
Algebra: MII.A.SSE.2	0	1	0	0	0	0
Algebra: MII.A.SSE.3a	0	1	0	0	0	0
Algebra: MII.A.SSE.3b	0	1	0	0	0	0
Functions: MII.F.BF.1a	0	0	0	0	0	1
Functions: MII.F.BF.1b	0	1	0	0	0	0
Functions: MII.F.BF.3	1	1	0	0	0	0
Functions: MII.F.IF.4	1	1	0	0	0	0
Functions: MII.F.IF.5	0	1	0	0	0	0
Functions: MII.F.IF.7a	1	0	0	0	0	0
Functions: MII.F.IF.7b	1	0	0	0	0	0
Functions: MII.F.IF.8b	0	0	1	0	0	0
Geometry: MII.G.C.2	1	0	0	0	0	0
Geometry: MII.G.C.4	1	0	0	0	0	0
Geometry: MII.G.CO.10	0	2	0	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.1b	0	0	0	0	0	1
Geometry: MII.G.SRT.4	0	0	0	0	0	1
Geometry: MII.G.SRT.5	0	1	0	0	0	0
Geometry: MII.G.SRT.8	0	0	0	0	0	1
Statistics and Probability: MII.S.CP.1	1	0	0	0	0	0
Statistics and Probability: MII.S.CP.5	0	1	0	0	0	0
Statistics and Probability: MII.S.CP.6	0	0	0	0	0	1
Statistics and Probability: MII.S.ID.5	0	1	0	0	0	0
Total	**39**					

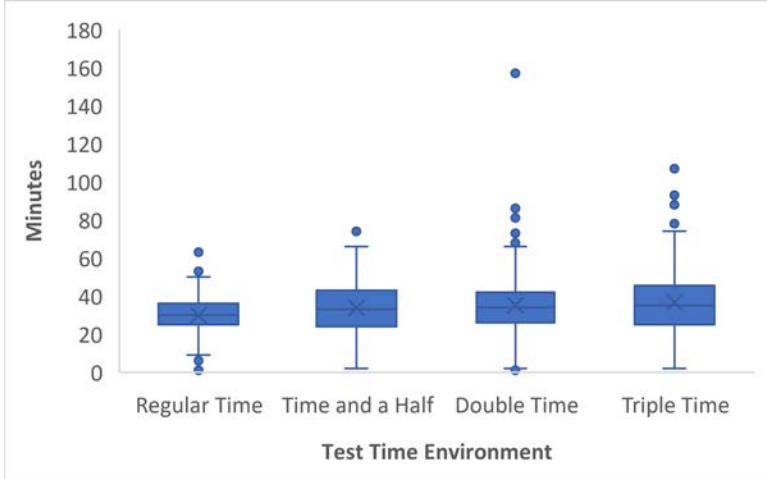
Science

		Multiple Choice			Technology Enhanced		
Grade	**Reporting Category**	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
9	Intended Learning Outcome 1	2	10	7	0	1	1
	Intended Learning Outcome 3	0	3	0	0	1	0
	Intended Learning Outcome 4	0	5	3	0	0	0
	Intended Learning Outcome 5/6	0	3	0	0	0	0
	Total	**36**					
10	Intended Learning Outcome 1	4	12	3	0	2	0
	Intended Learning Outcome 3	1	2	2	0	0	0
	Intended Learning Outcome 4	0	6	1	0	0	0
	Intended Learning Outcome 5/6	1	0	1	0	1	0
	Total	**36**					

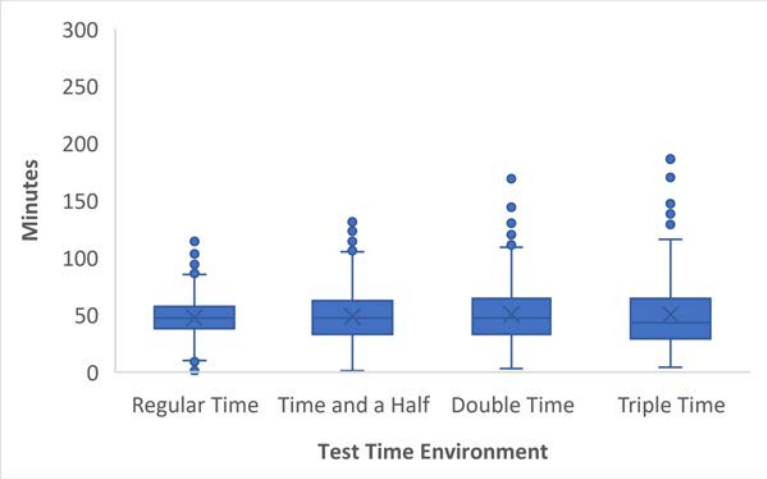
Appendix D: Student Testing Time



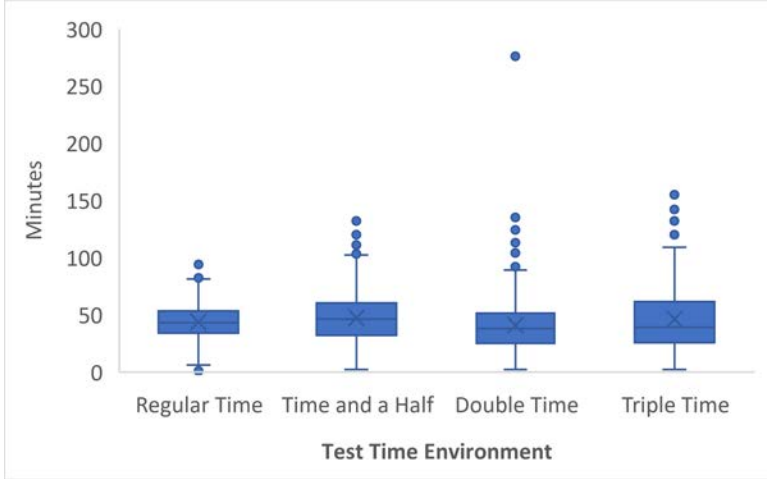
D-1. English Grade 9 Student Testing Time



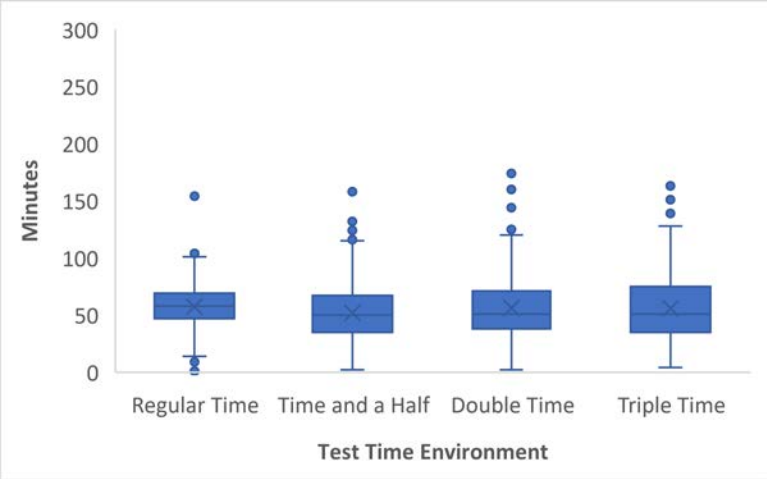
D-2. English Grade 10 Student Testing Time



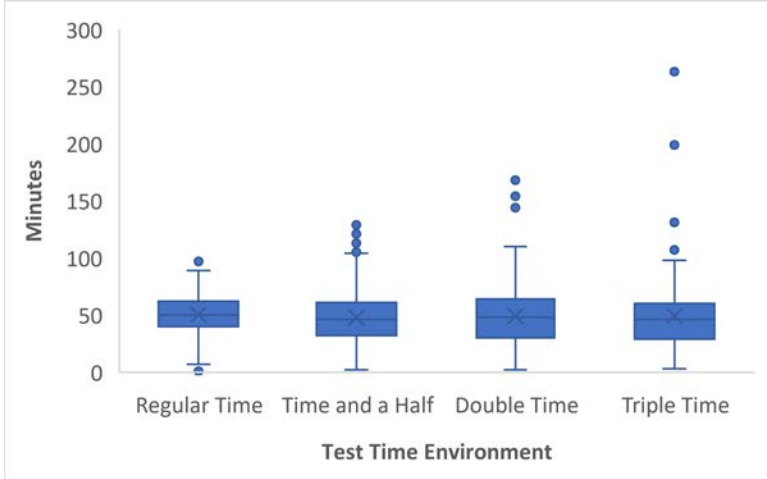
D-3. Reading Grade 9 Student Testing Time



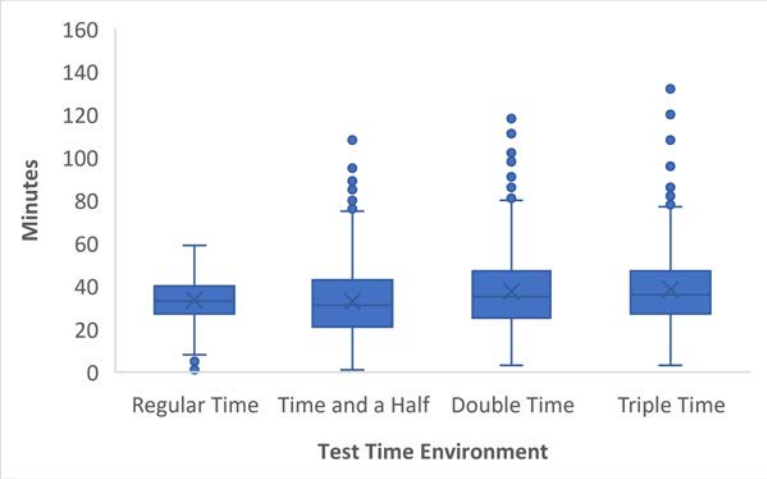
D-4. Reading Grade 10 Student Testing Time



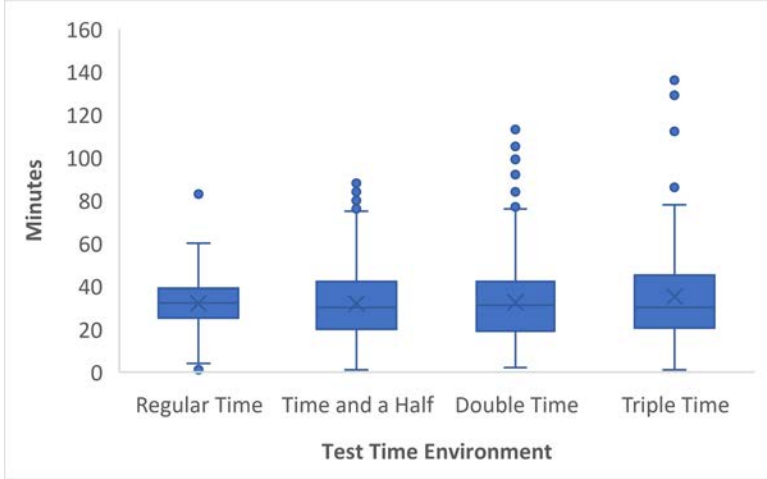
D-5. Mathematics Grade 9 Student Testing Time



D-6. Mathematics Grade 10 Student Testing Time



D-7. Science Grade 9 Student Testing Time



D-8. Science Grade 10 Student Testing Time

Appendix E: Item Statistics Summaries

Item Mean

One-Point Items

Subject	Grade	$p<0.30$	$0.30\leq p<0.55$	$0.55\leq p<0.75$	$0.75\leq p<0.95$	$p\geq 0.95$	Mean p	N
English	9	2	14	12	11	–	0.61	39
	10	4	13	18	9	–	0.59	44
Reading	9	1	5	18	5	–	0.62	29
	10	2	6	14	9	–	0.64	31
Mathematics	9	7	15	11	7	–	0.53	40
	10	8	15	13	3	–	0.50	39
Science	9	--	7	25	4	–	0.63	36
	10	1	18	13	4	–	0.57	36

Two-Point Items

Subject	Grade	N	Mean	Min	Max
English	9	6	0.48	0.25	0.83
	10	4	0.58	0.38	0.80
Reading	9	6	0.35	0.25	0.48
	10	4	0.61	0.52	0.80
Mathematics	9	N/A	–	–	–
	10	N/A	–	–	–
Science	9	N/A	–	–	–
	10	N/A	–	–	–

Item-Total Correlation

One-Point Items

Subject	Grade	$r<0.20$	$0.20\leq r<0.40$	$0.40\leq r<0.60$	$0.60\leq r<0.80$	$r\geq 0.80$	Median Pt.Bis	*N*
English	9	1	16	22	–	–	0.46	39
	10	1	12	31	–	–	0.45	44
Reading	9	1	2	25	1	–	0.49	29
	10	–	8	22	1	–	0.47	31
Mathematics	9	1	9	28	2	–	0.46	40
	10	–	6	31	2	–	0.47	39
Science	9	–	6	26	4	–	0.47	36
	10	1	9	17	9	–	0.51	36

Two-Point Items

Subject	Grade	*N*	Median r	Min r	Max r
English	9	6	0.54	0.38	0.67
	10	4	0.54	0.45	0.64
Reading	9	6	0.37	0.31	0.46
	10	4	0.63	0.54	0.70
Mathematics	9	N/A	–	–	–
	10	N/A	–	–	–
Science	9	N/A	–	–	–
	10	N/A	–	–	–

Differential Item Functioning

Subject	Grade	Subgroups	DIF Categories				
			Negligible DIF	Moderate DIF		Substantial DIF	
				Focal	Reference	Focal	Reference
English	9	Male-Female	45	–	–	–	–
		White-Hispanic	44	–	1	–	–
	10	Male-Female	47	–	1	–	–
		White-Hispanic	47	–	1	–	–
Reading	9	Male-Female	34	1	–	–	–
		White-Hispanic	35	–	–	–	–
	10	Male-Female	31	1	1	2	–
		White-Hispanic	34	–	1	–	–
Mathematics	9	Male-Female	38	–	1	1	–
		White-Hispanic	40	–	–	–	–
	10	Male-Female	38	–	–	–	1
		White-Hispanic	38	–	1	–	–
Science	9	Male-Female	35	–	1	–	–
		White-Hispanic	36	–	–	–	–
	10	Male-Female	34	–	2	–	–
		White-Hispanic	36	–	–	–	–

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

Appendix F: Reliability and Standard Error by Subgroup

F-1. English Grade 9 Test Reliability

	Test Group	*N*	Alpha	SEM	Conventions of Standard English	Knowledge of Language	Production of Writing
All	Students Tested	46,050	0.90	8.49	0.88	0.35	0.67
Gender	Female	22,626	0.90	8.37	0.87	0.30	0.64
	Male	23,422	0.91	8.58	0.88	0.39	0.67
Ethnicity	Hispanic or Latino Ethnicity	7,865	0.89	8.71	0.85	0.31	0.62
	Asian	815	0.92	8.70	0.90	0.27	0.69
	Native Hawaiian or Other Pacific Islander	716	0.88	8.41	0.85	0.30	0.62
	Black or African American	616	0.90	9.21	0.88	0.46	0.62
	American Indian or Alaska Native	518	0.87	8.65	0.83	0.29	0.60
	White	34,277	0.90	8.42	0.87	0.33	0.65
	Other	1,235	0.90	8.15	0.88	0.31	0.66
Limited English Proficiency	No	43,754	0.90	8.42	0.87	0.34	0.65
	Yes	2,296	0.81	9.94	0.76	0.21	0.42
Economic Disadvantage	No	31,683	0.90	8.40	0.87	0.32	0.65
	Yes	14,367	0.90	8.68	0.87	0.35	0.65
Special Education	No	41,505	0.89	8.42	0.87	0.31	0.64
	Yes	4,545	0.84	9.28	0.80	0.30	0.51

F-2. English Grade 10 Test Reliability

	Test Group	*N*	Alpha	SEM	Conventions of Standard English	Knowledge of Language	Production of Writing
All	Students Tested	43,836	0.92	7.89	0.89	0.48	0.70
Gender	Female	21,565	0.91	7.84	0.88	0.43	0.68
	Male	22,270	0.92	7.92	0.89	0.50	0.71
Ethnicity	Hispanic or Latino Ethnicity	7,518	0.90	7.91	0.86	0.44	0.65
	Asian	822	0.92	7.97	0.89	0.48	0.72
	Native Hawaiian or Other Pacific Islander	694	0.88	7.87	0.85	0.41	0.60
	Black or African American	582	0.90	8.09	0.87	0.52	0.65
	American Indian or Alaska Native	483	0.88	7.73	0.84	0.37	0.64
	White	32,653	0.91	7.86	0.89	0.46	0.69
	Other	1,078	0.92	7.98	0.89	0.51	0.70
Limited English Proficiency	No	41,663	0.91	7.87	0.88	0.46	0.69
	Yes	2,173	0.82	8.61	0.77	0.33	0.46
Economic Disadvantage	No	31,083	0.91	7.87	0.88	0.46	0.69
	Yes	12,753	0.91	7.94	0.88	0.46	0.67
Special Education	No	39,798	0.91	7.88	0.88	0.45	0.69
	Yes	4,038	0.86	8.30	0.81	0.37	0.57

F-3. Reading Grade 9 Test Reliability

	Test Group	*N*	Alpha	SEM	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
All	Students Tested	46,238	0.88	10.1	0.82	0.71	0.20
Gender	Female	22,724	0.87	9.96	0.80	0.69	0.20
	Male	23,513	0.89	10.2	0.83	0.73	0.20
Ethnicity	Hispanic or Latino Ethnicity	7,985	0.87	10.2	0.80	0.70	0.23
	Asian	815	0.89	10.2	0.84	0.72	0.16
	Native Hawaiian or Other Pacific Islander	724	0.86	10.2	0.79	0.71	0.22
	Black or African American	632	0.88	11.1	0.81	0.74	0.21
	American Indian or Alaska Native	529	0.85	10.3	0.77	0.67	0.15
	White	34,310	0.87	10.1	0.81	0.69	0.18
	Other	1,235	0.88	10.0	0.82	0.71	0.19
Limited English Proficiency	No	43,906	0.87	10.0	0.81	0.70	0.19
	Yes	2,332	0.77	12.1	0.65	0.56	0.13
Economic Disadvantage	No	31,723	0.87	10.1	0.81	0.68	0.18
	Yes	14,515	0.88	10.3	0.81	0.72	0.22
Special Education	No	41,675	0.87	10.0	0.81	0.69	0.18
	Yes	4,563	0.83	11.3	0.73	0.64	0.19

F-4. Reading Grade 10 Test Reliability

	Test Group	*N*	Alpha	SEM	Key Ideas	Craft and Structure	Integration of Knowledge and Ideas
All	Students Tested	44,132	0.90	8.87	0.84	0.73	0.44
Gender	Female	21,711	0.88	8.62	0.82	0.70	0.42
	Male	22,420	0.91	9.06	0.86	0.76	0.45
Ethnicity	Hispanic or Latino Ethnicity	7,674	0.88	8.84	0.83	0.68	0.33
	Asian	827	0.90	8.86	0.85	0.74	0.46
	Native Hawaiian or Other Pacific Islander	719	0.88	8.66	0.82	0.68	0.34
	Black or African American	593	0.89	8.94	0.83	0.70	0.32
	American Indian or Alaska Native	492	0.87	8.78	0.82	0.64	0.33
	White	32,739	0.89	8.88	0.84	0.72	0.43
	Other	1,082	0.90	8.94	0.85	0.75	0.47
Limited English Proficiency	No	41,899	0.89	8.86	0.84	0.73	0.43
	Yes	2,233	0.80	9.36	0.75	0.52	0.15
Economic Disadvantage	No	31,175	0.89	8.91	0.83	0.72	0.43
	Yes	12,957	0.89	8.80	0.84	0.71	0.39
Special Education	No	40,044	0.89	8.87	0.83	0.72	0.43
	Yes	4,088	0.85	9.14	0.79	0.61	0.26

F-5. Mathematics Grade 9 Test Reliability

	Test Group	*N*	Alpha	SEM	Algebra	Statistics and Probability	Functions	Geometry
All	Students Tested	45,590	0.91	8.59	0.80	0.69	0.69	0.68
Gender	Female	22,386	0.90	8.31	0.78	0.66	0.66	0.64
	Male	23,203	0.92	8.77	0.83	0.71	0.72	0.71
Ethnicity	Hispanic or Latino Ethnicity	7,801	0.88	9.67	0.76	0.66	0.59	0.57
	Asian	812	0.92	8.34	0.82	0.68	0.75	0.71
	Native Hawaiian or Other Pacific Islander	710	0.87	9.83	0.73	0.66	0.58	0.56
	Black or African American	612	0.87	10.9	0.74	0.69	0.57	0.51
	American Indian or Alaska Native	513	0.86	10.1	0.72	0.66	0.55	0.56
	White	33,914	0.90	8.29	0.80	0.66	0.69	0.67
	Other	1,220	0.91	8.77	0.80	0.68	0.70	0.68
Limited English Proficiency	No	43,311	0.91	8.45	0.80	0.67	0.69	0.67
	Yes	2,279	0.79	11.8	0.63	0.58	0.40	0.37
Economic Disadvantage	No	31,366	0.90	8.23	0.80	0.65	0.69	0.67
	Yes	14,224	0.89	9.38	0.78	0.69	0.63	0.63
Special Education	No	41,116	0.90	8.29	0.79	0.64	0.68	0.67
	Yes	4,474	0.83	11.2	0.66	0.61	0.47	0.50

F-6. Mathematics Grade 10 Test Reliability

	Test Group	*N*	Alpha	SEM	Algebra	Statistics and Probability	Functions	Geometry	Number and Quantity
All	Students Tested	43,705	0.90	9.46	0.76	0.41	0.66	0.73	0.52
Gender	Female	21,504	0.89	9.14	0.74	0.41	0.63	0.71	0.48
	Male	22,200	0.91	9.67	0.78	0.42	0.69	0.75	0.56
Ethnicity	Hispanic or Latino Ethnicity	7,542	0.85	11.5	0.67	0.33	0.53	0.63	0.45
	Asian	824	0.92	9.01	0.81	0.48	0.72	0.77	0.56
	Native Hawaiian or Other Pacific Islander	710	0.84	11.1	0.66	0.39	0.51	0.58	0.43
	Black or African American	581	0.85	13.0	0.67	0.28	0.49	0.58	0.46
	American Indian or Alaska Native	490	0.84	11.7	0.65	0.22	0.53	0.62	0.41
	White	32,484	0.90	9.00	0.76	0.40	0.66	0.73	0.50
	Other	1,068	0.90	9.64	0.77	0.40	0.64	0.74	0.53
Limited English Proficiency	No	41,501	0.90	9.23	0.76	0.40	0.66	0.73	0.51
	Yes	2,204	0.72	15.7	0.53	0.18	0.30	0.38	0.32
Economic Disadvantage	No	30,936	0.90	8.93	0.76	0.40	0.67	0.73	0.51
	Yes	12,769	0.88	10.9	0.71	0.37	0.57	0.68	0.47
Special Education	No	39,653	0.90	9.01	0.75	0.40	0.66	0.73	0.50
	Yes	4,052	0.75	15.1	0.53	0.23	0.33	0.43	0.36

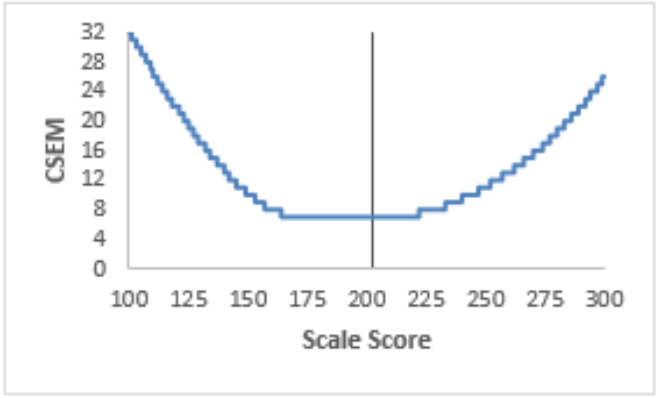
F-7. Science Grade 9 Test Reliability

	Test Group	*N*	Alpha	SEM	ILO 1	ILO 3	ILO 4	ILO 5/6
All	Students Tested	46,149	0.90	9.30	0.86	0.45	0.65	0.33
Gender	Female	22,683	0.89	9.14	0.85	0.40	0.61	0.30
	Male	23,465	0.91	9.39	0.87	0.49	0.68	0.35
Ethnicity	Hispanic or Latino Ethnicity	7,949	0.88	9.88	0.83	0.39	0.60	0.31
	Asian	820	0.91	8.90	0.88	0.51	0.70	0.34
	Native Hawaiian or Other Pacific Islander	721	0.87	9.89	0.81	0.41	0.57	0.30
	Black or African American	630	0.87	10.6	0.80	0.40	0.57	0.37
	American Indian or Alaska Native	532	0.84	10.0	0.78	0.29	0.51	0.32
	White	34,250	0.89	9.15	0.85	0.43	0.63	0.30
	Other	1,239	0.90	9.40	0.86	0.43	0.67	0.36
Limited English Proficiency	No	43,816	0.90	9.21	0.86	0.43	0.64	0.31
	Yes	2,333	0.77	12.1	0.68	0.24	0.40	0.26
Economic Disadvantage	No	31,713	0.89	9.10	0.85	0.42	0.63	0.29
	Yes	14,436	0.89	9.74	0.84	0.43	0.62	0.34
Special Education	No	41,601	0.89	9.14	0.85	0.42	0.63	0.29
	Yes	4,548	0.84	11.1	0.77	0.36	0.49	0.32

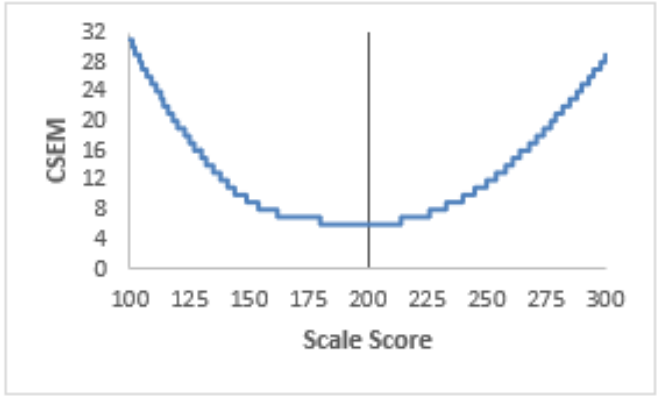
F-8. Science Grade 10 Test Reliability

	Test Group	*N*	Alpha	SEM	ILO 1	ILO 3	ILO 4	ILO 5/6
All	Students Tested	43,901	0.91	8.86	0.84	0.51	0.76	0.48
Gender	Female	21,581	0.90	8.67	0.83	0.45	0.74	0.45
	Male	22,319	0.92	8.99	0.85	0.56	0.79	0.51
Ethnicity	Hispanic or Latino Ethnicity	7,582	0.87	9.76	0.77	0.40	0.70	0.38
	Asian	828	0.92	8.75	0.87	0.55	0.78	0.50
	Native Hawaiian or Other Pacific Islander	713	0.86	9.79	0.75	0.32	0.69	0.41
	Black or African American	591	0.86	10.2	0.74	0.43	0.68	0.33
	American Indian or Alaska Native	485	0.87	9.54	0.76	0.41	0.71	0.40
	White	32,617	0.91	8.67	0.83	0.50	0.76	0.48
	Other	1,079	0.91	9.16	0.83	0.49	0.76	0.52
Limited English Proficiency	No	41,687	0.91	8.78	0.83	0.50	0.76	0.48
	Yes	2,214	0.70	12.7	0.52	0.19	0.50	0.20
Economic Disadvantage	No	31,079	0.91	8.69	0.84	0.50	0.76	0.48
	Yes	12,822	0.89	9.37	0.80	0.46	0.74	0.42
Special Education	No	39,834	0.91	8.71	0.83	0.50	0.76	0.47
	Yes	4,067	0.81	11.2	0.67	0.34	0.58	0.31

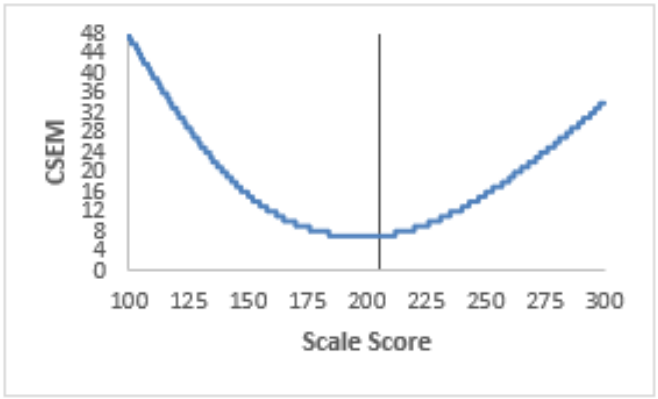
Appendix G: Conditional Standard Error of Scale Scores



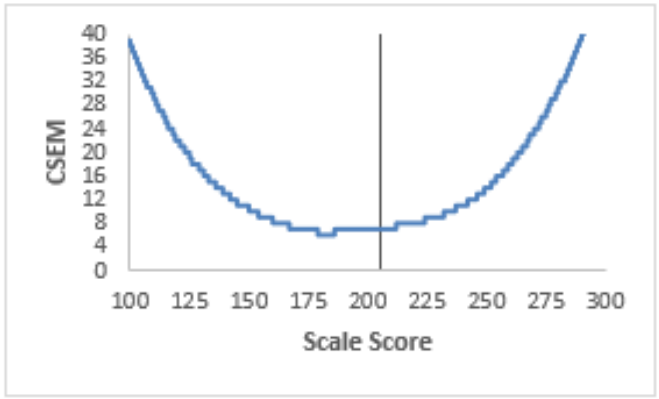
G-1. English Grade 9 Conditional Standard Error of Measurement



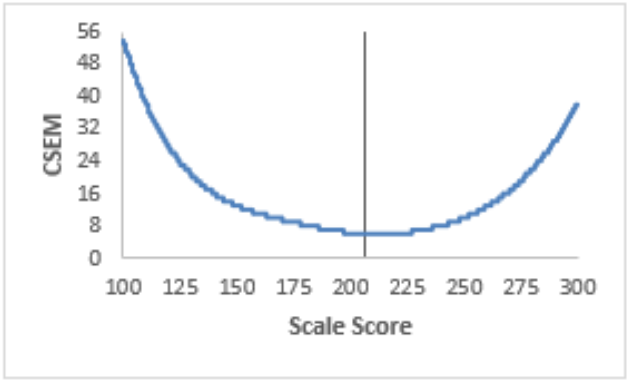
G-2. English Grade 10 Conditional Standard Error of Measurement



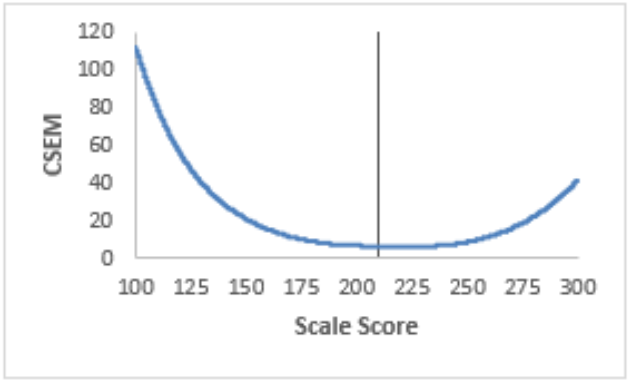
G-3. Reading Grade 9 Conditional Standard Error of Measurement



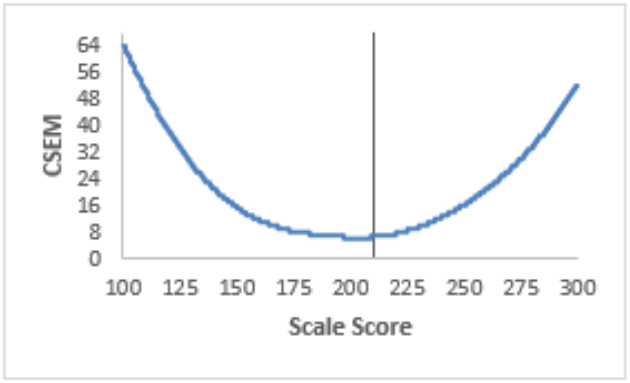
G-4. Reading Grade 10 Conditional Standard Error of Measurement



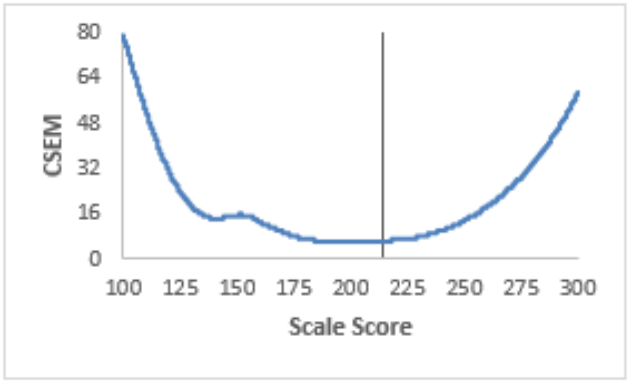
G-5. Mathematics Grade 9 Conditional Standard Error of Measurement



G-6. Mathematics Grade 10 Conditional Standard Error of Measurement



G-7. Science Grade 9 Conditional Standard Error of Measurement



G-8. Science Grade 10 Conditional Standard Error of Measurement

Appendix H: Accuracy and Consistency

H-1. Accuracy Classification for English Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.07472	0.01791	0.00000	0.00001	82.23
Approaching Proficient	0.02366	0.35035	0.04851	0.00000	
Proficient	0.00000	0.05253	0.36763	0.02498	
Highly Proficient	0.00000	0.00000	0.01009	0.02962	

H-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	0.07472	0.01791	0.00000	0.00001	89.90
Approaching Proficient	0.02366	0.35035	0.04851	0.00000	
Proficient	0.00000	0.05253	0.36763	0.02498	
Highly Proficient	0.00000	0.00000	0.01009	0.02962	

H-3. Consistency Classification for English Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.07136	0.03356	0.00006	0.00000	74.82	0.60
Approaching Proficient	0.02697	0.31439	0.06877	0.00003		
Proficient	0.00005	0.07281	0.33335	0.02547		
Highly Proficient	0.00000	0.00003	0.02405	0.02911		

H-4. Accuracy Classification for English Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.05825	0.01235	0.00000	0.00001	85.41
Approaching Proficient	0.02033	0.35377	0.04602	0.00000	
Proficient	0.00000	0.04395	0.41161	0.01516	
Highly Proficient	0.00000	0.00000	0.00807	0.03049	

H-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	0.05825	0.01235	0.00000	0.00001	91.00
Approaching Proficient	0.02033	0.35377	0.04602	0.00000	
Proficient	0.00000	0.04395	0.41161	0.01516	
Highly Proficient	0.00000	0.00000	0.00807	0.03049	

H-6. Consistency Classification for English Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.05569	0.02390	0.00001	0.00000	79.32	0.66
Approaching Proficient	0.02287	0.32458	0.06481	0.00000		
Proficient	0.00001	0.06159	0.38350	0.01626		
Highly Proficient	0.00000	0.00000	0.01737	0.02940		

H-7. Accuracy Classification for Reading Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.09355	0.02179	0.00000	0.00000	75.99
Approaching Proficient	0.02712	0.33231	0.05399	0.00017	
Proficient	0.00000	0.06000	0.25658	0.04387	
Highly Proficient	0.00000	0.00013	0.03301	0.07747	

H-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	0.09355	0.02179	0.00000	0.00000	88.57
Approaching Proficient	0.02712	0.33231	0.05399	0.00017	
Proficient	0.00000	0.06000	0.25658	0.04387	
Highly Proficient	0.00000	0.00013	0.03301	0.07747	

H-9. Consistency Classification for Reading Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.08937	0.03990	0.00023	0.00000	66.37	0.51
Approaching Proficient	0.03112	0.29286	0.07598	0.00205		
Proficient	0.00018	0.07892	0.20990	0.04791		
Highly Proficient	0.00000	0.00255	0.05748	0.07156		

H-10. Accuracy Classification for Reading Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.14333	0.02814	0.00002	0.00001	78.44
Approaching Proficient	0.03674	0.27881	0.05646	0.00001	
Proficient	0.00002	0.04917	0.30140	0.02506	
Highly Proficient	0.00000	0.00000	0.02002	0.06081	

H-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	0.14333	0.02814	0.00002	0.00001	89.43
Approaching Proficient	0.03674	0.27881	0.05646	0.00001	
Proficient	0.00002	0.04917	0.30140	0.02506	
Highly Proficient	0.00000	0.00000	0.02002	0.06081	

H-12. Consistency Classification for Reading Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.13621	0.04791	0.00079	0.00000	69.47	0.56
Approaching Proficient	0.04324	0.23847	0.07645	0.00029		
Proficient	0.00064	0.06937	0.26206	0.02769		
Highly Proficient	0.00000	0.00036	0.03859	0.05791		

H-13. Accuracy Classification for Mathematics Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.12935	0.02462	0.00000	0.00000	78.25
Approaching Proficient	0.02493	0.33443	0.04309	0.00003	
Proficient	0.00000	0.05642	0.26719	0.04356	
Highly Proficient	0.00000	0.00002	0.02480	0.05156	

H-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	0.12935	0.02462	0.00000	0.00000	90.04
Approaching Proficient	0.02493	0.33443	0.04309	0.00003	
Proficient	0.00000	0.05642	0.26719	0.04356	
Highly Proficient	0.00000	0.00002	0.02480	0.05156	

H-15. Consistency Classification for Mathematics Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.12402	0.04170	0.00012	0.00000	69.59	0.56
Approaching Proficient	0.03015	0.29752	0.06230	0.00080		
Proficient	0.00011	0.07529	0.22621	0.04621		
Highly Proficient	0.00000	0.00098	0.04645	0.04813		

H-16. Accuracy Classification for Mathematics Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.20974	0.03362	0.00003	0.00000	73.61
Approaching Proficient	0.03117	0.27897	0.03653	0.00027	
Proficient	0.00008	0.08355	0.24737	0.07868	
Highly Proficient	0.00000	0.00000	0.00000	0.00000	

H-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	0.20974	0.03362	0.00003	0.00000	87.96
Approaching Proficient	0.03117	0.27897	0.03653	0.00027	
Proficient	0.00008	0.08355	0.24737	0.07868	
Highly Proficient	0.00000	0.00000	0.00000	0.00000	

H-18. Consistency Classification for Mathematics Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.20229	0.05443	0.00084	0.00000	65.18	0.50
Approaching Proficient	0.03770	0.23895	0.05787	0.00448		
Proficient	0.00100	0.09893	0.19549	0.05937		
Highly Proficient	0.00000	0.00383	0.02972	0.01510		

H-19. Accuracy Classification for Science Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.08418	0.02002	0.00000	0.00000	81.05
Approaching Proficient	0.02422	0.46127	0.05088	0.00005	
Proficient	0.00000	0.05075	0.22536	0.02595	
Highly Proficient	0.00000	0.00002	0.01761	0.03967	

H-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	0.08418	0.02002	0.00000	0.00000	89.83
Approaching Proficient	0.02422	0.46127	0.05088	0.00005	
Proficient	0.00000	0.05075	0.22536	0.02595	
Highly Proficient	0.00000	0.00002	0.01761	0.03967	

H-21. Consistency Classification for Science Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.08039	0.03672	0.00001	0.00000	73.00	0.57
Approaching Proficient	0.02800	0.42156	0.06813	0.00073		
Proficient	0.00001	0.07278	0.19029	0.02718		
Highly Proficient	0.00000	0.00102	0.03543	0.03776		

H-22. Accuracy Classification for Science Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.09850	0.02376	0.00000	0.00000	81.86
Approaching Proficient	0.02573	0.47949	0.05000	0.00005	
Proficient	0.00000	0.04505	0.19742	0.02021	
Highly Proficient	0.00000	0.00002	0.01664	0.04314	

H-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	0.09850	0.02376	0.00000	0.00000	90.49
Approaching Proficient	0.02573	0.47949	0.05000	0.00005	
Proficient	0.00000	0.04505	0.19742	0.02021	
Highly Proficient	0.00000	0.00002	0.01664	0.04314	

H-24. Consistency Classification for Science Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.09390	0.04205	0.00001	0.00000	74.03	0.58
Approaching Proficient	0.03032	0.43835	0.06489	0.00064		
Proficient	0.00001	0.06693	0.16705	0.02172		
Highly Proficient	0.00000	0.00098	0.03211	0.04104		

Appendix I: Performance Level Descriptor Educator Committee

Utah Aspire Plus
Performance Level Descriptors (PLDs) Review Meeting Agenda
November 14, 2018

8:00 a.m. – 8:30 a.m. (MST)	Breakfast/Registration
8:30 a.m. – 9:00 a.m.	Orientation a. Welcome and Introductions b. Purpose and Goals of Meeting c. Overview of Utah Aspire Plus Assessments d. Housekeeping
9:00 a.m. – 10:15 a.m.	Review of PLDs Breakout Rooms a. English b. Reading c. Math d. Science Overview of Process and Procedures Begin PLDs Review
10:15 a.m. – 10:30 a.m.	Break
10:30 a.m. – 12:00 p.m.	PLDs Review Breakout Rooms
12:00 p.m. – 12:45 p.m.	Lunch
12:45 p.m. – 2:00 p.m.	PLDs Review Breakout Rooms
2:00 p.m. – 2:15 p.m.	Break
2:30 – 4:15 p.m.	PLDs Review Breakout Rooms
4:15 – 4:30 p.m.	Wrap up and Adjourn

Utah Aspire Plus Assessments for Grades 9–10

Educator Review of Performance Level Descriptors

November 14, 2018

Pearson

Agenda

1. Welcome
2. History of Utah Assessments
3. Overview of Utah Aspire Plus Assessments
4. General overview of performance level descriptors (PLDs)
5. Policy and Range PLDs
6. PLDs Process, Standard Setting, and Recap
7. Security and confidentiality
8. Housekeeping and transition to breakout rooms
9. Review and evaluate PLDs; approve verbiage or recommend edits
10. Closing

Pearson

History of Utah Assessments

SAGE Assessment

- New standards
- Blueprint CAT
- Performance-level descriptors
- Item development
- Standard-setting
- Legislation
  - Predictive of college and career assessment
  - Growth from 9th to 10th
- Timed
- Hybrid—link to 3rd–8th and to ACT

Pearson

Utah Aspire Plus 9–10 Assessments

- Utah-based test created by Utah educators, with Utah-aligned ACT Aspire ® content embedded
- 100% aligned to the Utah Core Standards
- Measures growth and readiness
- Strong prediction of performance on the ACT® Test
- Same platforms as the ACT Test
  - Seamless experience for students grades 9–11
  - Simplifies training and technology support for high schools

Pearson

Changes from Utah SAGE to Utah Aspire Plus 9–10 Assessments

- Linear rather than CAT assessment
- English and Reading are two separate assessments
- Approximately 50% ACT Aspire content
- Going from EOC to grade-level assessments

Performance Level Descriptors (PLDs)- General Overview

- Differentiate content mastery into a number of categories (lowest to highest mastery)
- Number and names of levels vary across testing programs
- Utah Aspire Plus includes 4 proficiency levels:
  - Below Proficient
  - Approaching Proficient
  - Proficient
  - Highly Proficient

Performance Level Descriptors (PLDs)- Policy vs. Range

- Policy Descriptions
  - Broad statements that define level of rigor and/or policy implications for each performance level
  - May or may not be specific to content area
- Range Descriptions
  - Subject and grade-specific knowledge and skills for the full range of each performance level
  - Used during the process of standard setting
  - Posted to provide instructional guidance and allow stakeholders to attach contextual meaning to students' test scores.

Range PLDs

- communicate how student knowledge and skill are observed in student work
- include characteristics at each level to distinguish one from the other
- use observable verbs
- include what students CAN do, not what they CAN'T do
- not include adverbs of frequency (usually, sometimes, generally) to discriminate between PLDs since the assessment is a one-time event
- range across the full level of performance for that level

Performance Level Descriptors (PLDs)- Process

- The Pearson team reviewed SAGE PLDs for Grades 9 and 10. PLDs for standards included on the Utah Aspire Plus blueprint were retained. The Pearson team drafted new PLDs when necessary.
- USBE reviewed and edited the PLDs draft.
- Each committee will review the PLDs draft and recommend use for the Utah Aspire Plus assessments beginning in spring 2019.
- The committee will determine if there is ACT Aspire® PLDs verbiage that should be incorporated into the Policy PLDs.

Performance Level Descriptors (PLDs) & Standard Setting

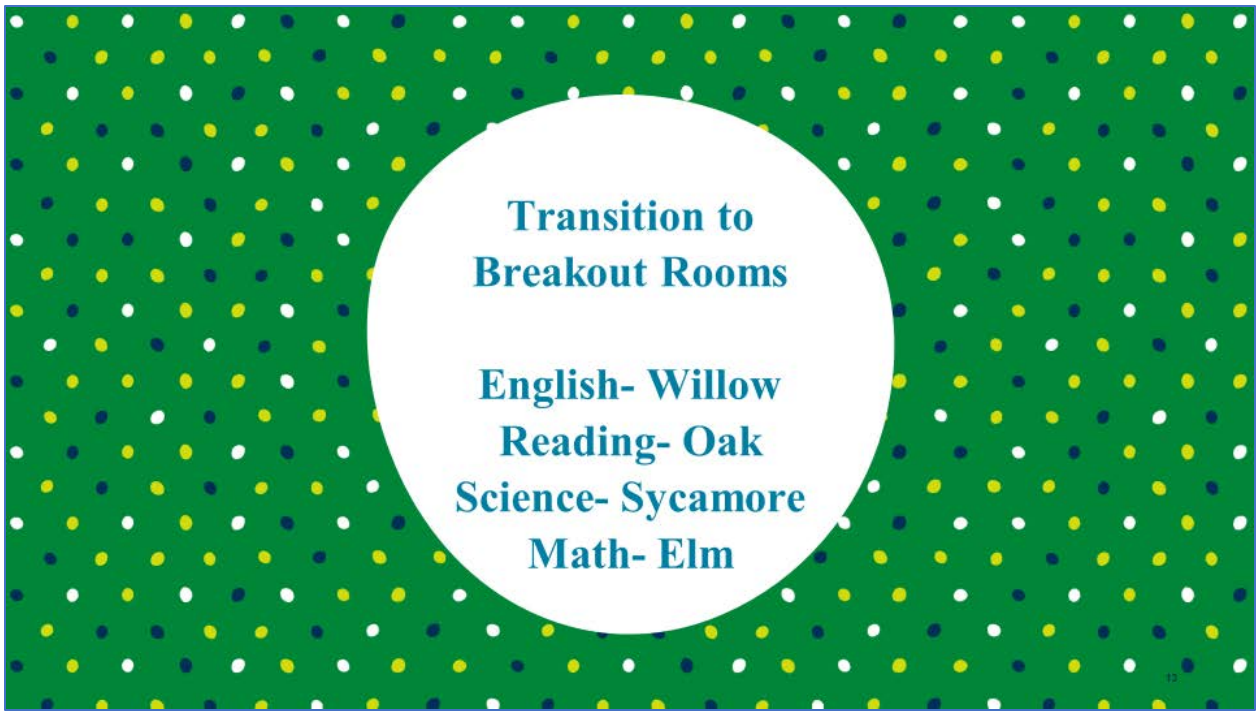
- The range PLDs will be used during the process of standard setting. PLDs are used to establish cut scores that differentiate test performance into the specified performance levels. For Utah Aspire Plus, there are 4 cut scores.
- Cut scores define transitional points between adjacent performance levels.
- Threshold Descriptions
  - Minimal knowledge and skills for entry into a performance level
  - Only used during standard setting, not used for test score reporting
  - Are not the emphasis of today's workshop. However, borderline knowledge and skills (i.e., skills that cannot be directly classified into a performance level) may be brought up in discussions during review of range PLDs.

Performance Level Descriptors (PLDs) Recap

- Policy Definitions
  - Broad statements that define level of rigor and/or policy implication for each performance level.
- Range PLDs
  - Detailed descriptions of subject and grade-specific knowledge and abilities for the full range of each performance level. **(primary goal for today)**
- Threshold PLDs
  - Describe the student who just barely makes it into a performance level (i.e., the borderline student) **(this is a secondary goal)**

Security and Confidentiality

- Strict security procedures protect the content discussed.
  - Nondisclosure/confidentiality forms
- Technology considerations in committee meetings
  - Use of laptops
  - Use of personal electronics
- The process is not confidential; procedures and your training can be shared.
- Items are confidential and secure. You have signed a confidentiality statement saying you will not share any information about the content of the items, passages, or stimuli.


Transition to
Breakout Rooms
English- Willow
Reading- Oak
Science- Sycamore
Math- Elm

Appendix J: Predicting ACT Test Scores from the Utah Aspire Plus High School Assessment

Abstract

The new Utah Aspire Plus high school assessment comprises items from the ACT Aspire Early High School tests and items from the Utah item bank. One of the primary requirements for the new assessment is to provide test users with predictive information about a student's expected performance on the ACT. This document presents the study for predicting the ACT test scores from the new Utah high school assessment. In particular, an indirect linkage was used in the absence of longitudinal data.

Keywords: Concordance, Predicted ACT scores

[Please note that the table and figures referred to in this report are shown at the end of Appendix J.]

Predicting ACT Test Scores from the Utah Aspire Plus High School Assessment

The new Utah Aspire Plus high school assessment—testing Utah's grade 9 and 10 students in English, mathematics, reading, and science—comprises items from the ACT Aspire Early High School tests and items from the Utah item bank. Only multiple-choice and technology-enhanced items from Aspire that are aligned to the Utah standards are included in the new assessment. The constructed-response items in Aspire mathematics, reading, and science tests are not used. One of the primary requirements for the new assessment is to provide test users with predictive information about a student's expected performance on the ACT. This document presents the study for predicting ACT test scores from the Utah Aspire Plus high school assessment used in the first test administration year.

Methodology

Because the Utah Aspire Plus high school assessment is new and no student has the opportunity to take both the Utah assessment and the ACT test, prediction on a student's expected performance on the ACT test cannot be established directly. When matched longitudinal data become available for Utah students taking both the new assessment and the ACT, the prediction can be obtained using the same methodology as used for predicting students' performance on the ACT based on their Aspire scores (ACT, 2019).

With the absence of longitudinal data, however, the predication was estimated through an indirect way utilizing the existing Aspire to the ACT prediction, as approved by the Utah Technical Advisory Committee. Because the Utah assessment and the Aspire assessment share common test items, a concordance can be conducted mapping Utah scale scores to Aspire scale scores, using the common items. The mapped Aspire scale scores can be used to look up the predicted ACT scale score ranges from the existing table (ACT, 2019). Figure 1 illustrates the

two-step process. Following this process, a predicted ACT score range can be established for each Utah scale score point.

The Aspire to the ACT Prediction

Both the ACT and Aspire assessments measure students' achievement in the same four subject areas of English, mathematics, reading, and science. For Aspire examinees in grades 9 and 10, a predicted ACT score range is provided for each corresponding Aspire subject they have taken; if a student takes all four subjects, a predicted ACT Composite score range is also reported. The predicted score range indicates students' expected performance on the ACT when they are in grade 11. For 9th graders, the prediction assumes that the ACT test will be taken 22 to 26 months after the Aspire test. For 10th graders, the prediction assumes that the ACT test will be taken 10 to 14 months after the Aspire test.

The longitudinal sample used for developing ACT test score predictions for grade 9 and 10 students is formed by matching Aspire student records to the ACT test records. Predicted ACT score ranges for grade 10 are based on students tested in consecutive years with Aspire and the ACT, while predicted ACT score ranges for grade 9 are based on students who tested two years apart with Aspire and the ACT. Quantile regression models are used to estimate the percentiles of the ACT test score distribution, conditional on Aspire score. The predicted score range endpoints are defined as the scores closest to the 25th and 75th percentiles of the ACT score distribution, conditional on ACT Aspire score. The prediction analysis is conducted every year after the spring administration and the updated prediction is applied in reporting for the following academic year. For reporting during the 2017–2018 academic year, the longitudinal sample for grade 10 included over 200,000 students.

Concordance between Utah and Aspire Scale Scores

Concordance is typically used to link two tests measuring similar constructs and intending for similar populations (Dorans, 2004; Holland & Dorans, 2006; Kolen & Brennan, 2014), as in the case of the new Utah assessment and the Aspire assessment. With the help of common items between the two tests, a chained equipercentile concordance can be performed to link Utah scale scores and Aspire scale scores. The procedure of the chained concordance is described in the following subsection, separately for the four subject tests and Composite score since they follow different procedures.

Concordance of four subject tests.

Common items between Utah and Aspire.

The Aspire items selected for the Utah assessment were all aligned to the Utah standards and met statistical requirements documented in the Utah Aspire Plus High School Assessment test specifications. To include a sufficient number of common items for a solid linking, a target was set so that the selected common items accounted for at least 50% of the score points on the intact Aspire form. This was accomplished for all four subjects except math. To maximize the proportion of common items for math, five additional Aspire items were placed in the pretest slots. These items were used only for the purpose of linking and did not count toward students' scores on the Utah assessment. Although the 50% target was still not reached for math, it did increase the common item percentage to at least 25% on the Aspire test and 33% on the Utah test.

Table 1 shows the number of Aspire (common) items for each subject. After the administration of the Utah assessment in 2019, *p*-value and point-biserial statistics were computed for the common items based on the Utah sample. The results were compared with the

statistics for the same items based on the ACT Aspire national sample from the spring 2018 administration. After examining the bivariate plots of *p*-value and point-biserial between the Utah sample and Aspire sample, as well as the absolute difference in *p*-value (Kolen and Brennan, 2014), one common item on the grade 9 reading test appeared to perform differently than other common items and was excluded from the concordance analysis. Figure 2 shows the bivariate plot of the *p*-value statistic for this item.

Chained equipercentile concordance.

For the Utah sample, a total raw score on the common item set—that is, sum scores of individual items—was computed for each student. With distributions of the common item score and the Utah scale score, an equipercentile concordance was carried out, resulting in a concordance table mapping the Utah scale score to the common item score. According to this table, students should be ranked in the same order with Utah scale scores and with the corresponding concordant common item scores.

Similarly, a concordance table mapping the common item score to the ACT Aspire scale score was obtained using the Aspire sample. Putting two concordance tables together yielded a concordance table mapping the Utah scale score to the Aspire scale score, referred to as the Utah-to-Aspire concordance table. For example, suppose the Utah scale score of 200 was mapped to the common item score of 21 and the same common item score was mapped to the Aspire scale score of 420. The Utah-to-Aspire concordance table would have the Utah scale score of 200 mapped to the Aspire scale score of 420. When a Utah scale score was mapped to a non-integer common item score, as in most cases, linear interpolation was used to find the corresponding Aspire scale score. This step was needed because the concordance table mapping the common item score to

the Aspire scale score only contains integer scores. Putting the Utah-to-Aspire concordance table and the Aspire to the ACT prediction table together yielded a Utah-to-ACT table, predicting ACT scale scores from Utah scale scores. A subsequent subsection presents the details.

Concordance of the Composite score.

With the Utah-to-Aspire concordance tables available for the four subjects, each Utah student was assigned an Aspire scale score for each of the four subjects. The Aspire Composite score was then computed by averaging the four Aspire subject scale scores rounded to an integer. Now that each Utah student had an observed Utah Composite score and an Aspire Composite score, concordance was conducted using the single group design, leading to a Utah-to-Aspire concordance table for the Composite score.

Predicting ACT scale scores from Utah scale scores.

For each Utah scale score, a concordant Aspire scale score can be found through the Utah-to-Aspire concordance table. A corresponding predicted ACT scale score range can be found by looking up the concordant Aspire scale score in the Aspire to the ACT prediction table.

Because the Utah-to-ACT prediction table was estimated through an indirect linkage— with the help of the ACT Aspire common items— the predicted scale score range was widened by adding one standard error of measurement, taking into account the additional linking error resulted from indirect linking. This adjustment was researched in a separate study and endorsed by the Utah Technical Advisory Committee.

As mentioned in the previous section, the current Aspire to the ACT prediction range covers the 25th to 75th percentile band conditional on Aspire scale score. For the Utah

prediction, the existing 25th–75th percentile prediction band was adjusted by taking into account the standard error of measurement (SEM) on the Aspire scale scores. Specifically, the updated lower boundary of the predicted range is mapped from the value of the original Aspire scale score minus one SEM and the updated upper boundary is mapped from the value of the original Aspire scale score plus one SEM. For example, if the original 25th–75th predicted ACT score range for the Aspire scale score of 421 is 15 to 18 and the SEM for the Aspire scale score is one, the updated predicted ACT score range for the Aspire score of 421 would be 14 (25th percentile mapped from the Aspire score of 420: 421 minus 1) to 19 (75th percentile mapped from the Aspire score of 422: 421 plus 1).

Results

Score Distributions

Only on-grade students who took the standard test forms were included in the concordance analysis. Figures 3–6 show the score distributions of the Utah scale scores, the Aspire scale scores, and common item scores on both assessments. The sample size for the Utah assessment ranged from 43,456 to 46,019; and the sample size for the Aspire assessment ranged from 12,172 to 22,573. Note that the Utah scale score distributions show some jumps for the lowest and highest scale scores. Because they were valid scale scores reported to students, concordance was carried out with these scores included.

Utah to the ACT Prediction

Figures 7–11 show the predicted ACT score ranges conditional on Utah scale scores. The ranges for the top Utah scale score were somehow much smaller than the ranges for adjacent scores, especially for math, reading, and science. Several reasons could contribute to this phenomenon: (1) the upper bound of the range had reached the highest ACT scale score, 36, and there was no room to increase; (2) there were much more students who scored at the highest

possible Utah scale score than at adjacent scores, which inevitably led to a jump of the concordant scores.

Summary

In this document, the analysis was described for developing the ACT scale score predictions based on scale scores from the Utah Aspire Plus high school assessment. An indirect linkage was used in the absence of longitudinal data. A direct linkage should be used when the longitudinal data become available, after two or three years of administration of the new assessment.

Cautions should be taken when using results from this study. The predictions were based on a chained process with the concordance between the Utah assessment scale and Aspire scale chained to the existing Aspire to the ACT predictions developed based on a sample not specific to Utah. Also, the common items used in creating the concordance were likely different in terms of font sizes and positions they appeared on the two tests. The common item set may not be treated as a miniature of the intact ACT Aspire test forms due to potential differences in blueprints and test specifications between the two tests. These issues may affect the quality of the linkage and undermine the common item design.

To mitigate these potential concerns, a larger proportion of common items were used than typically required for common-item non-equivalent groups design among parallel forms. However, the prediction analysis based on Utah-specific longitudinal data in Year 2 (for grade 10) and Year 3 (for grade 9) will likely result in more accurate predictions.

References

ACT. (2019). *ACT Aspire technical manual*. Iowa City: ACT, Inc.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227–246. https://doi.org/10.1177/0146621604265031

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187–220). Praeger, CT: Westport.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking. Methods and Practices* (3rd ed.). New York: Springer.

Tables

Table 1

Number of Common Items between Utah Assessment and Aspire Assessment

	Subject	Operational Item Points		Common Aspire Items		
		Utah	Aspire	Points	% in Utah	% in Aspire
Grade 9						
	English	51	50	31	61%	62%
	Math	40	51	13	33%	25%
	Reading	35	30	22	63%	73%
	Science	36	40	19	53%	48%
Grade 10						
	English	52	50	32	62%	64%
	Math	40	51	14	35%	27%
	Reading	35	30	20	57%	67%
	Science	36	40	20	56%	50%

Note: Five Aspire items were placed in pretest slots for math in both grades.

Figures

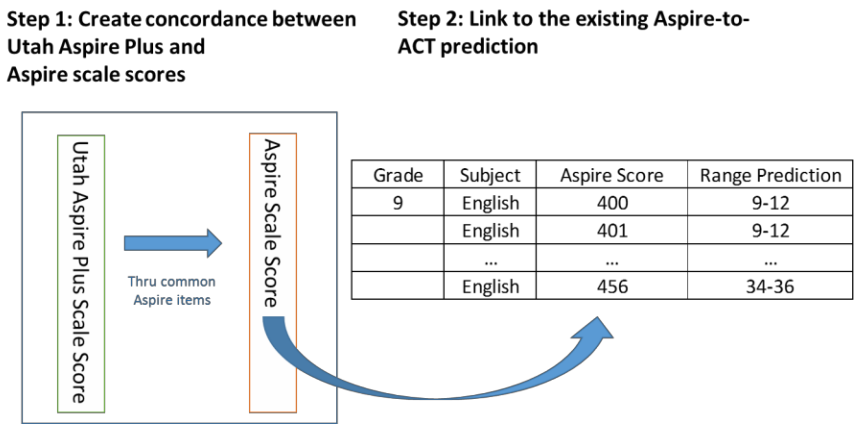


Figure 1. Two-step linking to predict ACT scores from Utah scores.

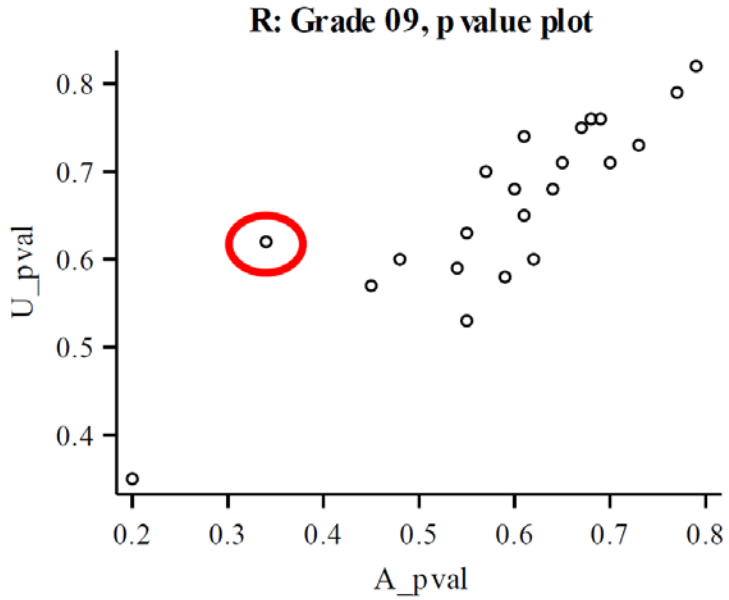


Figure 2. Bivariate plot of *p*-value for an excluded common item. A_pval denotes *p*-value based on the Aspire sample, and U_pval on the Utah sample.

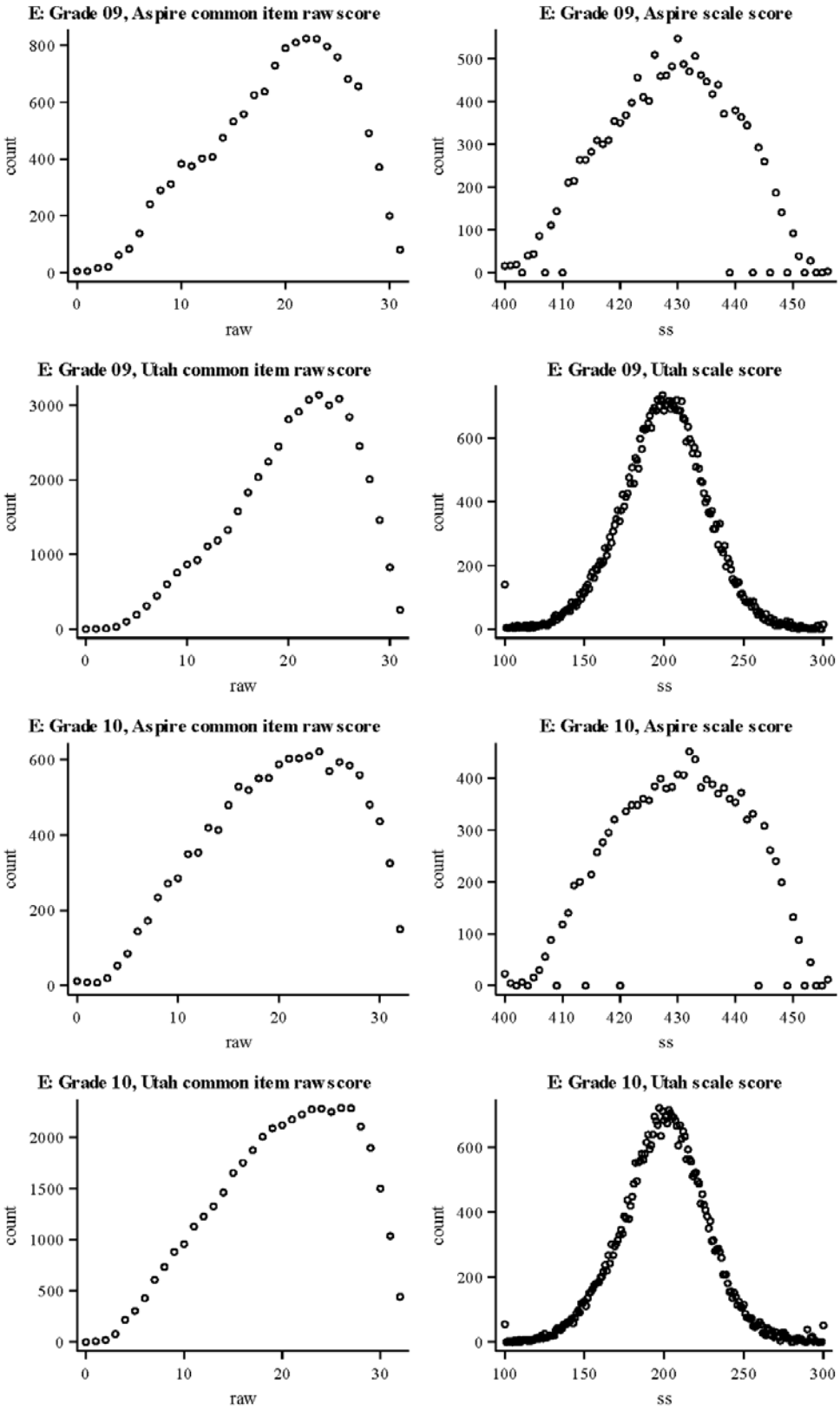


Figure 3. Score distributions for the English test.

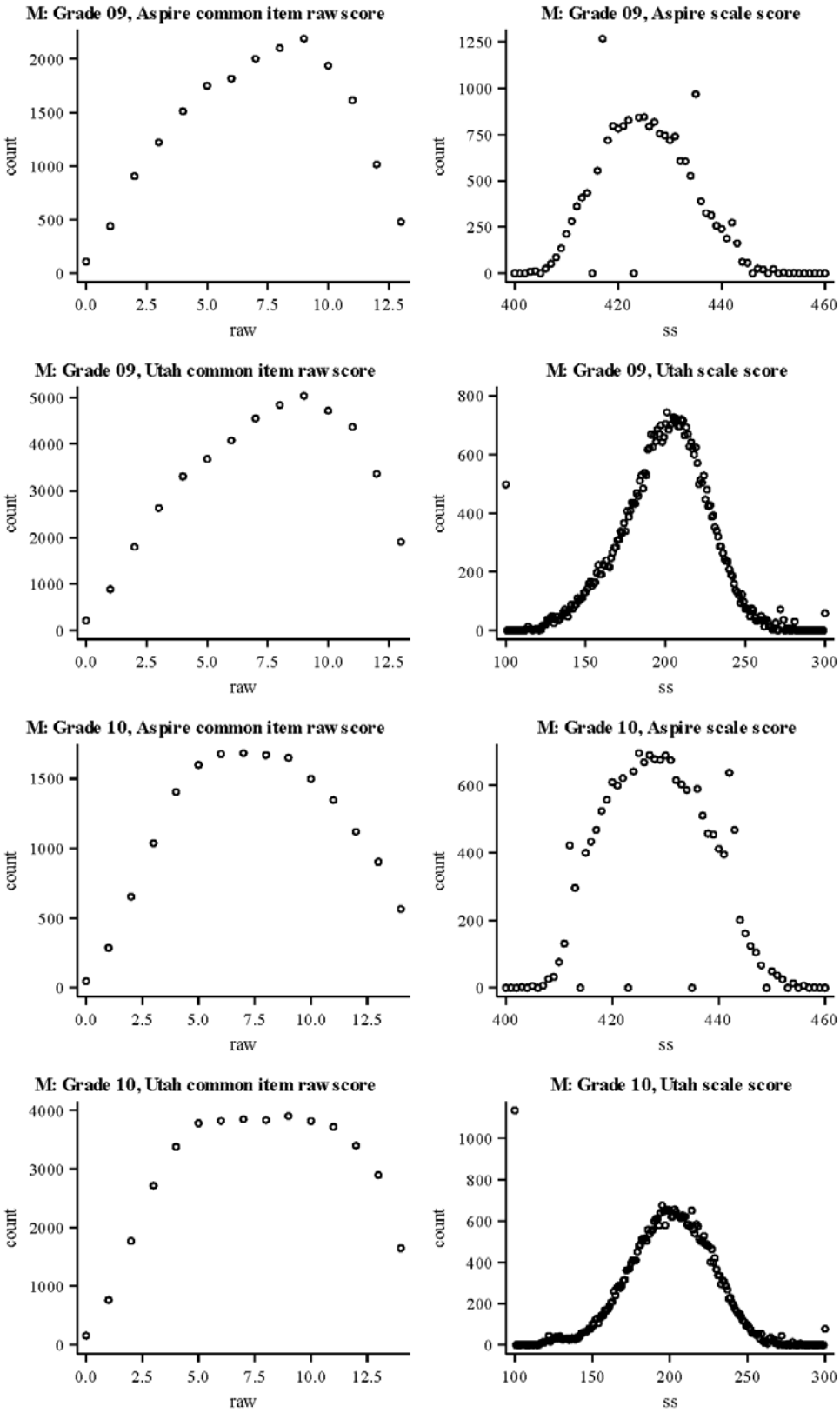


Figure 4. Score distributions for the math test.

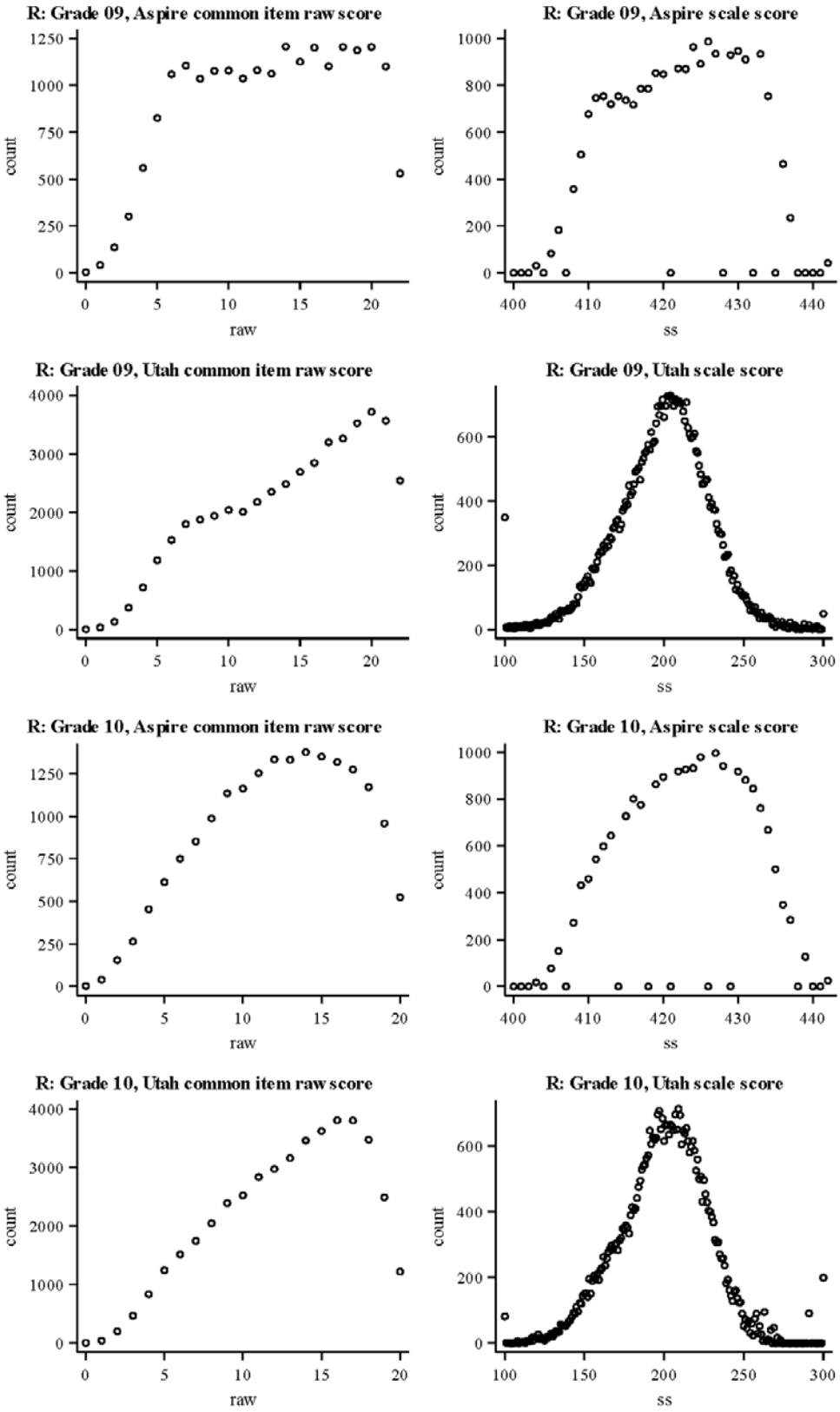


Figure 5. Score distributions for the reading test.

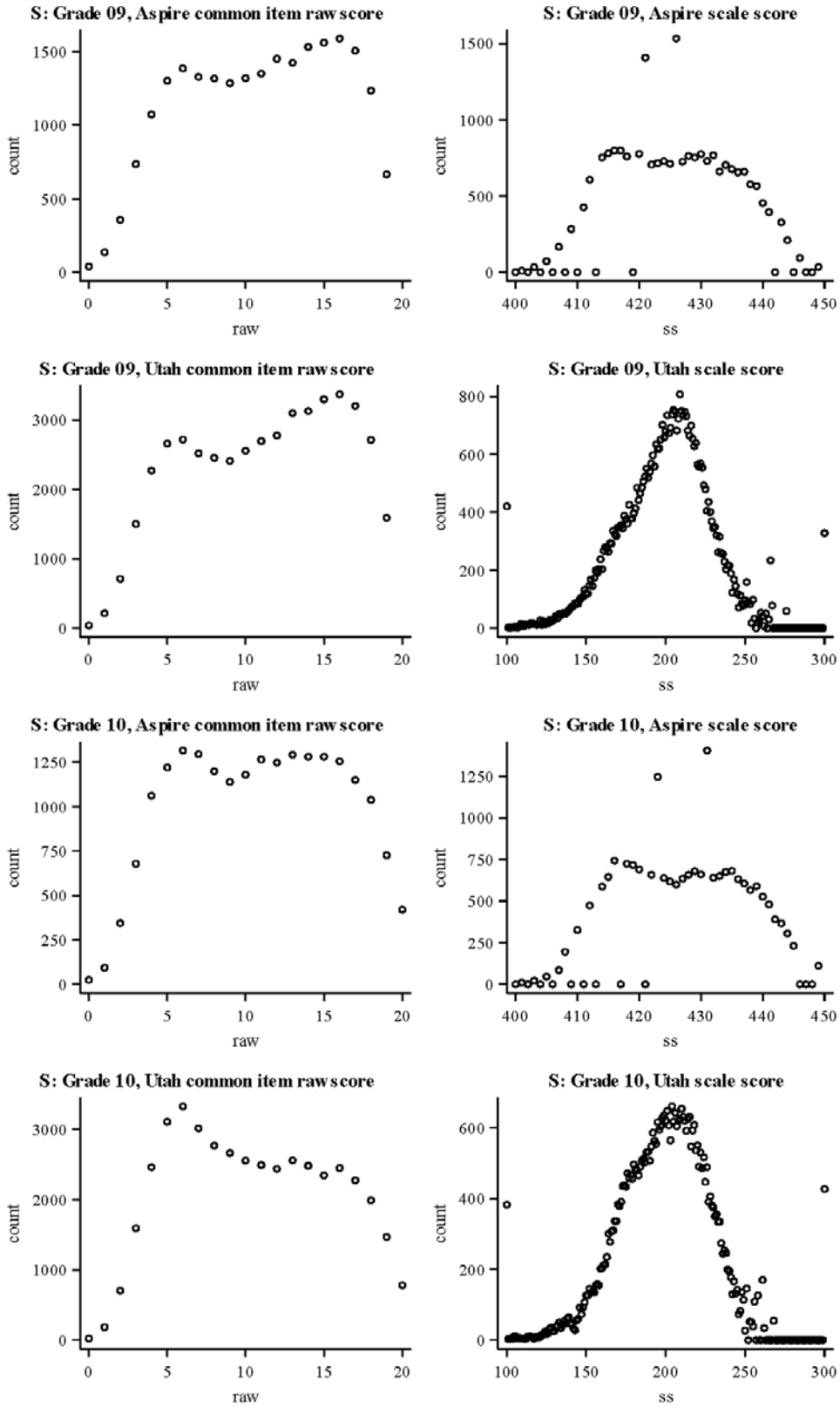


Figure 6. Score distributions for the science test.

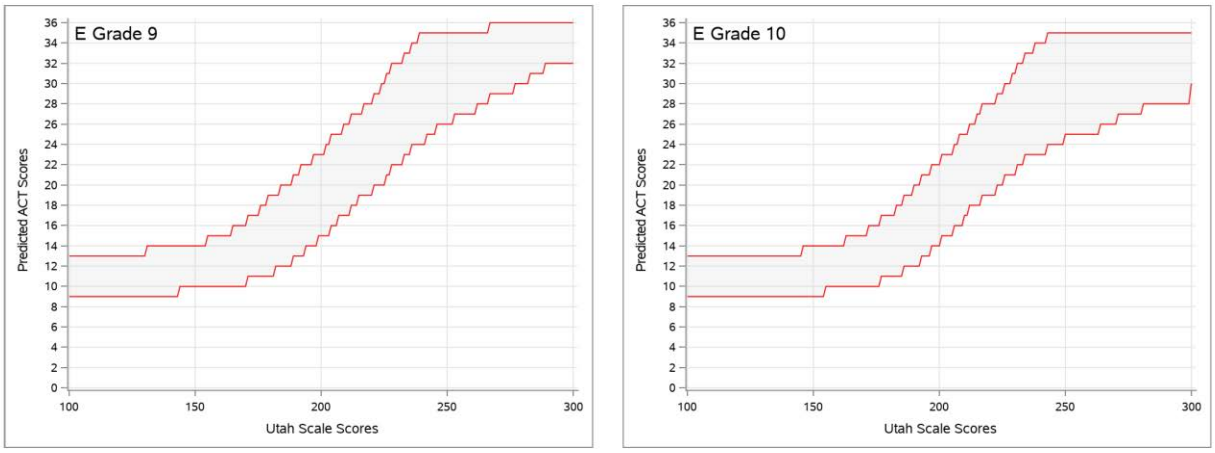


Figure 7. Utah to ACT prediction for the English test.

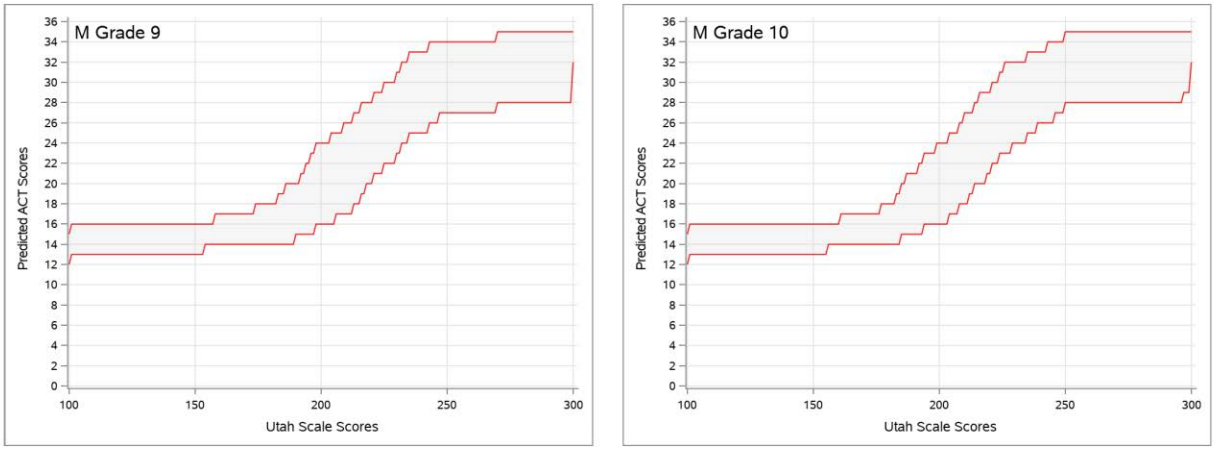


Figure 8. Utah to ACT prediction for the math test.

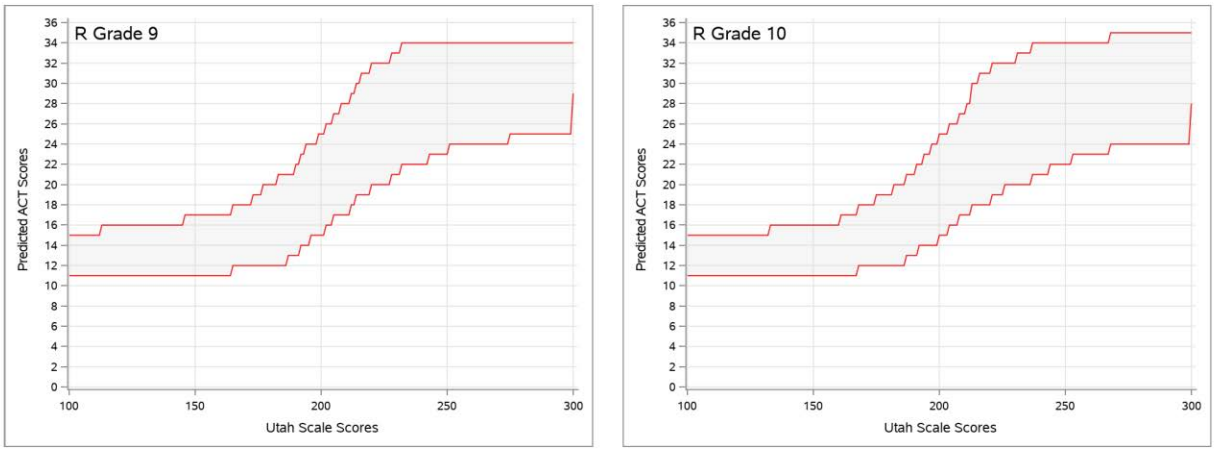


Figure 9. Utah to ACT prediction for the reading test.

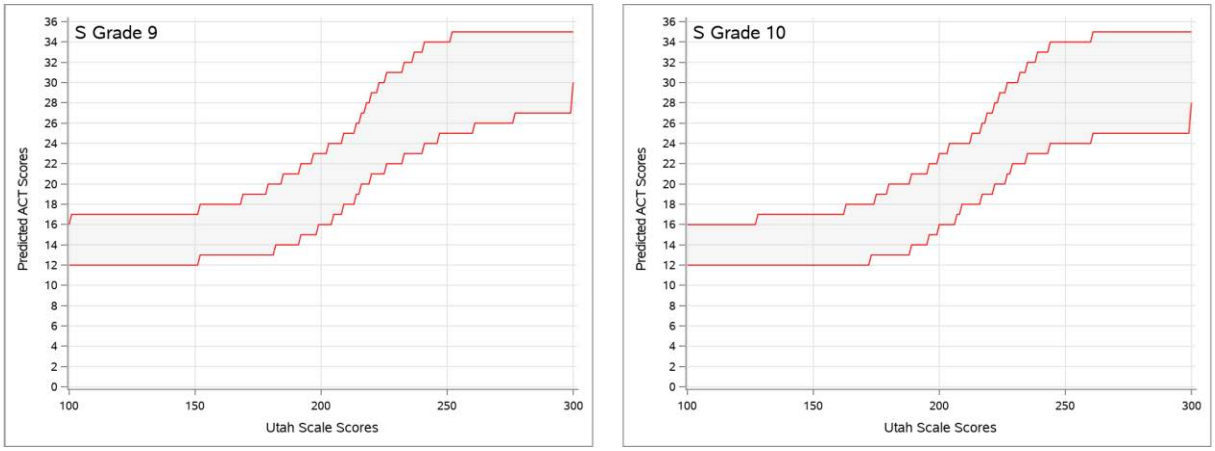


Figure 10. Utah to ACT prediction for the science test.

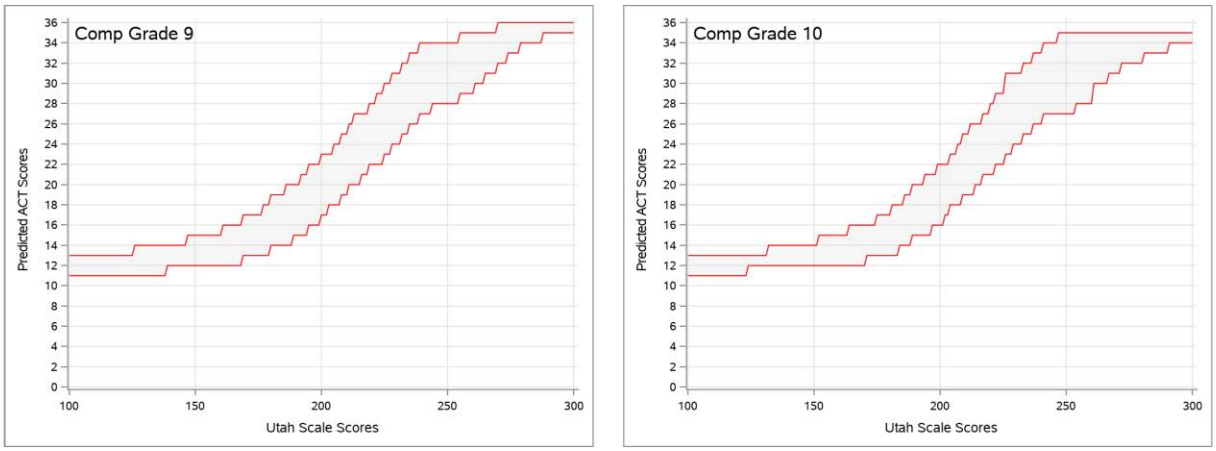


Figure 11. Utah to ACT prediction for the composite score.

Appendix K: Utah-to-ACT Concordance Tables

K-1. English Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-130	9-13	5
131-143	9-14	6
144-154	10-14	5
155-164	10-15	6
165-170	10-16	7
171-175	11-17	7
176-178	11-18	8
179-181	11-19	9
182-183	12-19	8
184-188	12-20	9
189-191	13-21	9
192-193	13-22	10
194-196	14-22	9
197-198	14-23	10
199-201	15-23	9
202-203	15-24	10
204-206	16-25	10
207-208	17-25	9
209-211	17-26	10
212-214	18-27	10
215-216	19-27	9
217-220	19-28	10
221-223	20-29	10
224-225	20-30	11
226-227	21-31	11
228-232	22-32	11
233-235	23-33	11
236-238	24-34	11
239-241	24-35	12
242-245	25-35	11
246-252	26-35	10
253-261	27-35	9
262-266	28-35	8
267-276	29-36	8
277-282	30-36	7
283-288	31-36	6
289-300	32-36	5

K-2. English Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-145	9-13	5
146-154	9-14	6
155-162	10-14	5
163-171	10-15	6
172-176	10-16	7
177-182	11-17	7
183-185	11-18	8
186-189	12-19	8
190-192	12-20	9
193-196	13-21	9
197-200	14-22	9
201-205	15-23	9
206-207	16-24	9
208-209	16-25	10
210-211	17-25	9
212-214	18-26	9
215-216	18-27	10
217-222	19-28	10
223-225	20-29	10
226-228	21-30	10
229-230	21-31	11
231-233	22-32	11
234-237	23-33	11
238-242	23-34	12
243-249	24-35	12
250-263	25-35	11
264-270	26-35	10
271-280	27-35	9
281-299	28-35	8
300	30-35	6

K-3. Reading Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-112	11-15	5
113-145	11-16	6
146-164	11-17	7
165-172	12-18	7
173-176	12-19	8
177-182	12-20	9
183-186	12-21	10
187-189	13-21	9
190-191	13-22	10
192-193	14-23	10
194-195	14-24	11
196-198	15-24	10
199-201	15-25	11
202-204	16-26	11
205-207	17-27	11
208-211	17-28	12
212-213	18-29	12
214-215	19-30	12
216-219	19-31	13
220-227	20-32	13
228-231	21-33	13
232-242	22-34	13
243-250	23-34	12
251-274	24-34	11
275-299	25-34	10
300	29-34	6

K-4. Reading Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-132	11-15	5
133-160	11-16	6
161-167	11-17	7
168-174	12-18	7
175-181	12-19	8
182-186	12-20	9
187-190	13-21	9
191	13-22	10
192-193	14-22	9
194-196	14-23	10
197-199	14-24	11
200-203	15-25	11
204-207	16-26	11
208-210	17-27	11
211-212	17-28	12
213-215	18-30	13
216-220	18-31	14
221-225	19-32	14
226-230	20-32	13
231-236	20-33	14
237-243	21-34	14
244-252	22-34	13
253-267	23-34	12
268-299	24-35	12
300	28-35	8

K-5. Mathematics Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100	12-15	4
101-153	13-16	4
154-157	14-16	3
158-173	14-17	4
174-182	14-18	5
183-185	14-19	6
186-189	14-20	7
190-191	15-20	6
192-193	15-21	7
194-195	15-22	8
196-197	15-23	9
198-203	16-24	9
204-205	16-25	10
206-208	17-25	9
209-212	17-26	10
213-215	18-27	10
216-217	19-28	10
218-220	20-28	9
221-224	21-29	9
225-229	22-30	9
230-231	23-31	9
232-234	24-32	9
235-242	25-33	9
243-246	26-34	9
247-269	27-34	8
270-299	28-35	8
300	32-35	4

K-6. Mathematics Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100	12-15	4
101-155	13-16	4
156-160	14-16	3
161-176	14-17	4
177-182	14-18	5
183-184	14-19	6
185-186	15-20	6
187-191	15-21	7
192-193	15-22	8
194-198	16-23	8
199-203	16-24	9
204-207	17-25	9
208-209	18-26	9
210-211	18-27	10
212-213	19-27	9
214-215	20-28	9
216-218	20-29	10
219-220	21-29	9
221-223	22-30	9
224-225	23-31	9
226-228	23-32	10
229-234	24-32	9
235-238	25-33	9
239-242	26-33	8
243-245	26-34	9
246-249	27-34	8
250-296	28-35	8
297-299	29-35	7
300	32-35	4

K-7. Science Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100	12-16	5
101-151	12-17	6
152-168	13-18	6
169-178	13-19	7
179-181	13-20	8
182-184	14-20	7
185-191	14-21	8
192-196	15-22	8
197-198	15-23	9
199-202	16-23	8
203-204	16-24	9
205-208	17-24	8
209-213	18-25	8
214-215	19-26	8
216-217	20-27	8
218-219	20-28	9
220-222	21-29	9
223-225	21-30	10
226-232	22-31	10
233-236	23-32	10
237-240	23-33	11
241-246	24-34	11
247-251	25-34	10
252-260	25-35	11
261-276	26-35	10
277-299	27-35	9
300	30-35	6

K-8. Science Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-127	12-16	5
128-162	12-17	6
163-172	12-18	7
173-174	13-18	6
175-179	13-19	7
180-188	13-20	8
189-195	14-21	8
196-199	15-22	8
200-203	16-23	8
204-206	16-24	9
207-208	17-24	8
209-212	18-24	7
213-216	18-25	8
217-218	19-26	8
219-221	19-27	9
222-223	20-28	9
224-226	20-29	10
227-228	21-30	10
229-231	22-30	9
232-234	22-31	10
235-238	23-32	10
239-243	23-33	11
244-260	24-34	11
261-299	25-35	11
300	28-35	8

K-9. Grade 9 Predicted ACT Composite Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-125	11-13	3
126-138	11-14	4
139-146	12-14	3
147-160	12-15	4
161-168	12-16	5
169-176	13-17	5
177-179	13-18	6
180-185	14-19	6
186-188	14-20	7
189-191	15-20	6
192-194	15-21	7
195-199	16-22	7
200-202	17-23	7
203-204	18-23	6
205-207	18-24	7
208-210	19-25	7
211-212	20-26	7
213-215	20-27	8
216-218	21-27	7
219-221	22-28	7
222-224	22-29	8
225-227	23-30	8
228-231	24-31	8
232-234	25-32	8
235-238	26-33	8
239-243	27-34	8
244-254	28-34	7
255-260	29-35	7
261-264	30-35	6
265-269	31-35	5
270-273	32-36	5
274-278	33-36	4
279-287	34-36	3
288-300	35-36	2

K-10. Grade 10 Predicted ACT Composite Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-123	11-13	3
124-131	12-13	2
132-151	12-14	3
152-163	12-15	4
164-170	12-16	5
171-174	13-16	4
175-180	13-17	5
181-183	13-18	6
184-185	14-18	5
186-188	14-19	6
189-193	15-20	6
194-196	15-21	7
197-198	16-21	6
199-201	16-22	7
202-203	17-22	6
204-206	18-23	6
207-208	18-24	7
209-211	19-25	7
212-213	19-26	8
214-216	20-26	7
217-219	21-27	7
220-221	21-28	8
222-225	22-29	8
226-228	23-31	9
229-232	24-31	8
233-236	25-32	8
237-240	26-33	8
241-246	27-34	8
247-253	27-35	9
254-260	28-35	8
261-266	30-35	6
267-271	31-35	5
272-280	32-35	4
281-290	33-35	3
291-300	34-35	2

Appendix L: Scale Score Descriptive Statistics by Subgroup

L-1. English Grade 9 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	46,050	199.45	27.4	182	200	217	-0.21
Gender	Female	22,626	204.40	25.9	188	205	221	-0.11
	Male	23,422	194.66	27.9	177	196	213	-0.22
Ethnicity	Hispanic or Latino Ethnicity	7,865	184.57	25.7	169	185	201	-0.23
	Asian	815	203.63	30.1	185	205	223	0.04
	Native Hawaiian or Other Pacific Islander	716	187.66	24.5	172	189	204	-0.01
	Black or African American	616	179.06	29.7	160	181	199	-0.24
	American Indian or Alaska Native	518	182.20	23.7	168	183	197	-0.30
	White	34,277	203.57	26.3	188	204	220	-0.21
	Other	1,235	201.60	26.2	185	203	220	-0.27
Limited English Proficiency	No	43,754	201.31	26.3	185	202	218	-0.16
	Yes	2,296	164.03	22.6	151	165	179	-0.39
Economic Disadvantage	No	31,683	204.43	25.9	189	205	221	-0.18
	Yes	14,367	188.47	27.4	171	189	206	-0.14
Special Education	No	41,505	202.73	25.8	187	203	219	-0.17
	Yes	4,545	169.48	23.5	155	169	183	0.00

L-2. English Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,836	199.66	27.3	183	200	217	-0.04
Gender	Female	21,565	204.48	25.8	188	204	221	0.12
	Male	22,270	194.99	27.9	177	196	213	-0.09
Ethnicity	Hispanic or Latino Ethnicity	7,518	185.92	24.5	170	187	201	-0.03
	Asian	822	203.58	28.2	186	203	222	0.10
	Native Hawaiian or Other Pacific Islander	694	187.14	22.6	174	189	201	-0.34
	Black or African American	582	178.93	26.1	160	181	197	0.06
	American Indian or Alaska Native	483	183.14	22.3	168	183	199	-0.08
	White	32,653	203.57	26.7	187	204	220	-0.08
	Other	1,078	200.99	27.6	183	201	219	-0.07
LEP	No	41,663	201.32	26.6	185	202	218	-0.03
	Yes	2,173	167.85	20.4	155	169	181	-0.13
Economic Disadvantage	No	31,083	203.88	26.6	188	204	221	-0.06
	Yes	12,753	189.39	26.3	172	190	206	0.03
Special Education	No	39,798	202.50	26.2	187	203	219	-0.03
	Yes	4,038	171.62	22.1	158	172	185	0.17

L-3. Reading Grade 9 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	46,238	199.16	29.0	182	201	218	-0.37
Gender	Female	22,724	202.28	27.2	186	204	219	-0.30
	Male	23,513	196.15	30.4	177	199	217	-0.36
Ethnicity	Hispanic or Latino Ethnicity	7,985	185.34	28.0	168	187	204	-0.37
	Asian	815	203.54	30.4	185	206	223	-0.28
	Native Hawaiian or Other Pacific Islander	724	184.28	27.8	167	186	204	-0.39
	Black or African American	632	178.97	32.2	161	180	200	-0.23
	American Indian or Alaska Native	529	182.36	26.3	166	185	199	-0.40
	White	34,310	203.18	27.9	187	205	221	-0.38
	Other	1,235	200.49	28.3	184	202	218	-0.41
Limited English Proficiency	No	43,906	200.97	28.1	184	203	219	-0.35
	Yes	2,332	165.08	25.1	152	167	181	-0.50
Economic Disadvantage	No	31,723	204.14	27.5	188	206	222	-0.36
	Yes	14,515	188.28	29.3	170	190	208	-0.31
Special Education	No	41,675	202.38	27.4	186	204	220	-0.33
	Yes	4,563	169.77	27.1	155	170	187	-0.24

L-4. Reading Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	44,132	199.72	27.9	183	201	218	-0.07
Gender	Female	21,711	202.84	25.3	188	204	219	-0.02
	Male	22,420	196.69	29.9	176	198	217	-0.02
Ethnicity	Hispanic or Latino Ethnicity	7,674	186.59	25.8	169	188	204	-0.03
	Asian	827	202.89	28.2	185	205	221	-0.00
	Native Hawaiian or Other Pacific Islander	719	184.82	24.8	168	186	203	-0.30
	Black or African American	593	181.18	26.4	162	181	199	0.36
	American Indian or Alaska Native	492	184.75	24.4	169	187	201	-0.26
	White	32,739	203.56	27.2	188	205	221	-0.11
	Other	1,082	201.15	28.6	184	202	220	-0.08
Limited English Proficiency	No	41,899	201.34	27.3	185	203	219	-0.08
	Yes	2,233	169.30	21.1	155	170	184	-0.08
Economic Disadvantage	No	31,175	203.83	27.3	188	205	221	-0.09
	Yes	12,957	189.82	26.9	172	191	208	-0.03
Special Education	No	40,044	202.34	27.0	186	204	220	-0.08
	Yes	4,088	174.07	23.3	158	174	189	0.12

L-5. Mathematics Grade 9 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	45,590	198.77	28.4	182	201	218	-0.50
Gender	Female	22,386	199.61	25.7	185	202	216	-0.53
	Male	23,203	197.96	30.7	180	200	219	-0.45
Ethnicity	Hispanic or Latino Ethnicity	7,801	183.02	27.5	167	185	201	-0.46
	Asian	812	206.26	29.6	188	207	226	-0.31
	Native Hawaiian or Other Pacific Islander	710	184.40	27.1	169	187	203	-0.69
	Black or African American	612	174.97	29.9	156	177	196	-0.43
	American Indian or Alaska Native	513	181.62	27.5	166	185	200	-0.54
	White	33,914	203.18	26.8	188	205	221	-0.53
	Other	1,220	199.47	28.9	184	202	217	-0.60
Limited English Proficiency	No	43,311	200.47	27.5	185	202	218	-0.50
	Yes	2,279	166.35	25.9	152	168	183	-0.41
Economic Disadvantage	No	31,366	204.38	26.4	190	206	221	-0.50
	Yes	14,224	186.38	28.6	169	189	206	-0.44
Special Education	No	41,116	202.24	26.3	187	204	219	-0.47
	Yes	4,474	166.85	27.0	151	167	184	-0.14

L-6. Mathematics Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,705	197.63	30.5	181	200	218	-0.76
Gender	Female	21,504	198.47	27.9	184	201	217	-0.85
	Male	22,200	196.82	32.7	179	199	219	-0.67
Ethnicity	Hispanic or Latino Ethnicity	7,542	181.49	30.0	167	184	201	-0.74
	Asian	824	205.32	32.6	188	208	227	-0.76
	Native Hawaiian or Other Pacific Islander	710	184.42	27.9	172	189	202	-0.96
	Black or African American	581	173.96	33.0	160	177	196	-0.74
	American Indian or Alaska Native	490	181.93	29.2	169	185	201	-0.93
	White	32,484	202.12	28.8	187	204	221	-0.80
	Other	1,068	198.34	30.8	183	202	218	-0.84
Limited English Proficiency	No	41,501	199.40	29.5	183	201	219	-0.77
	Yes	2,204	164.46	29.8	153	169	183	-0.69
Economic Disadvantage	No	30,936	202.73	28.7	187	205	222	-0.80
	Yes	12,769	185.28	31.0	170	188	206	-0.72
Special Education	No	39,653	200.94	28.5	186	203	220	-0.76
	Yes	4,052	165.33	30.3	153	170	184	-0.62

L-7. Science Grade 9 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	46,149	199.25	29.5	182	202	218	-0.25
Gender	Female	22,683	200.19	27.3	185	202	217	-0.33
	Male	23,465	198.34	31.5	178	201	219	-0.18
Ethnicity	Hispanic or Latino Ethnicity	7,949	183.68	28.3	166	185	203	-0.34
	Asian	820	205.01	29.9	186	207	225	-0.11
	Native Hawaiian or Other Pacific Islander	721	181.79	27.4	165	182	200	-0.40
	Black or African American	630	177.92	29.2	161	179	197	-0.21
	American Indian or Alaska Native	532	182.17	25.5	167	183	199	-0.36
	White	34,250	203.72	28.2	188	206	221	-0.23
	Other	1,239	200.35	30.3	183	204	218	-0.46
Limited English Proficiency	No	43,816	201.03	28.6	184	203	219	-0.23
	Yes	2,333	165.81	25.3	153	167	182	-0.41
Economic Disadvantage	No	31,713	204.56	27.9	189	206	221	-0.17
	Yes	14,436	187.58	29.5	169	189	208	-0.32
Special Education	No	41,601	202.34	28.0	186	204	219	-0.20
	Yes	4,548	170.93	28.0	156	171	188	-0.16

L-8. Science Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,901	199.09	29.4	180	200	218	-0.14
Gender	Female	21,581	199.44	27.3	183	201	217	-0.27
	Male	22,319	198.75	31.3	178	200	220	-0.04
Ethnicity	Hispanic or Latino Ethnicity	7,582	183.33	27.2	168	184	200	-0.27
	Asian	828	205.11	31.2	185	205	224	0.22
	Native Hawaiian or Other Pacific Islander	713	181.26	25.8	167	182	199	-0.63
	Black or African American	591	178.77	27.1	164	178	194	0.18
	American Indian or Alaska Native	485	182.48	26.2	167	184	200	-0.38
	White	32,617	203.56	28.3	186	205	221	-0.13
	Other	1,079	200.69	30.2	183	202	219	-0.19
Limited English Proficiency	No	41,687	200.78	28.8	183	202	219	-0.13
	Yes	2,214	167.27	23.2	157	170	181	-0.67
Economic Disadvantage	No	31,079	203.76	28.6	186	205	222	-0.12
	Yes	12,822	187.78	28.4	171	188	207	-0.20
Special Education	No	39,834	201.82	28.4	184	203	220	-0.12
	Yes	4,067	172.36	25.7	160	173	186	-0.16

L-9. English Language Arts (ELA) Grade 9 Composite Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	47,247	199.74	26.5	183	201	218	-0.26
Gender	Female	23,185	203.74	24.9	188	205	220	-0.23
	Male	24,060	195.88	27.3	178	198	215	-0.23
Ethnicity	Hispanic or Latino Ethnicity	8,196	185.41	24.9	169	187	203	-0.15
	Asian	834	204.06	28.5	185	206	222	-0.11
	Native Hawaiian or Other Pacific Islander	748	186.29	24.4	169	187	203	-0.07
	Black or African American	646	179.60	29.0	159	182	199	-0.08
	American Indian or Alaska Native	549	182.75	23.3	168	183	198	-0.23
	White	34,999	203.75	25.3	188	205	221	-0.30
	Other	1,267	201.43	25.4	185	203	220	-0.30
Limited English Proficiency	No	44,835	201.55	25.5	186	203	219	-0.24
	Yes	2,412	164.93	20.9	152	165	179	-0.06
Economic Disadvantage	No	32,333	204.65	24.9	189	206	221	-0.29
	Yes	14,914	188.86	26.5	171	190	207	-0.10
Special Education	No	42,466	202.96	24.8	188	204	219	-0.24
	Yes	4,781	169.88	22.9	155	169	185	0.19

L-10. English Language Arts (ELA) Grade 10 Composite Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	45,442	200.18	25.9	183	201	218	-0.07
Gender	Female	22,311	204.15	24.0	189	204	220	0.02
	Male	23,130	196.34	27.2	178	197	215	-0.04
Ethnicity	Hispanic or Latino Ethnicity	7,927	186.75	23.5	171	187	202	0.06
	Asian	850	203.72	26.7	186	204	222	0.05
	Native Hawaiian or Other Pacific Islander	747	186.73	21.4	172	188	201	-0.11
	Black or African American	618	180.76	24.6	162	180	197	0.33
	American Indian or Alaska Native	508	184.34	21.8	170	184	200	-0.06
	White	33,661	203.99	25.2	188	205	221	-0.14
	Other	1,123	201.54	26.4	184	203	219	-0.04
Limited English Proficiency	No	43,115	201.80	25.2	186	203	218	-0.07
	Yes	2,327	168.97	18.5	156	169	181	0.16
Economic Disadvantage	No	32,006	204.28	25.2	188	205	221	-0.12
	Yes	13,436	190.14	25.0	173	190	207	0.06
Special Education	No	41,126	202.89	24.8	187	204	219	-0.08
	Yes	4,316	173.28	20.8	159	173	186	0.34

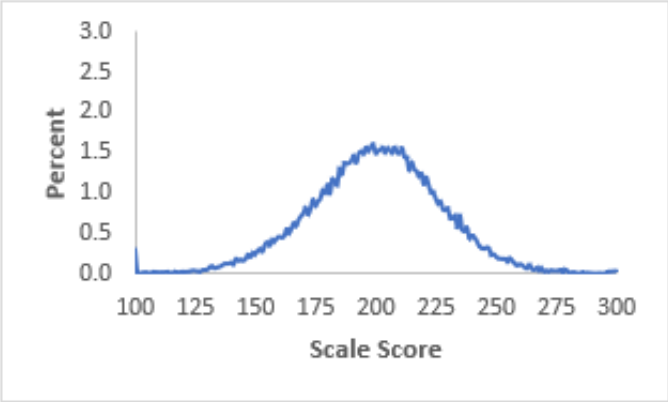
L-11. Science, Technology, Engineering and Mathematics (STEM) Grade 9 Composite Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	47,247	199.47	27.1	183	202	217	-0.30
Gender	Female	23,185	200.32	24.8	186	202	217	-0.36
	Male	24,060	198.65	29.1	180	201	218	-0.23
Ethnicity	Hispanic or Latino Ethnicity	8,196	183.81	25.8	166	185	202	-0.17
	Asian	834	206.10	27.8	189	207	226	-0.18
	Native Hawaiian or Other Pacific Islander	748	183.56	24.8	167	185	201	-0.23
	Black or African American	646	176.82	27.3	159	177	196	-0.10
	American Indian or Alaska Native	549	182.43	24.3	168	183	199	-0.15
	White	34,999	203.86	25.6	189	206	220	-0.33
	Other	1,267	200.43	27.3	185	203	217	-0.41
Limited English Proficiency	No	44,835	201.19	26.2	186	203	218	-0.29
	Yes	2,412	166.41	22.5	152	166	181	0.02
Economic Disadvantage	No	32,333	204.88	25.3	190	206	221	-0.29
	Yes	14,914	187.44	27.1	169	189	206	-0.19
Special Education	No	42,466	202.72	25.3	188	204	219	-0.27
	Yes	4,781	169.14	24.6	153	168	185	0.23

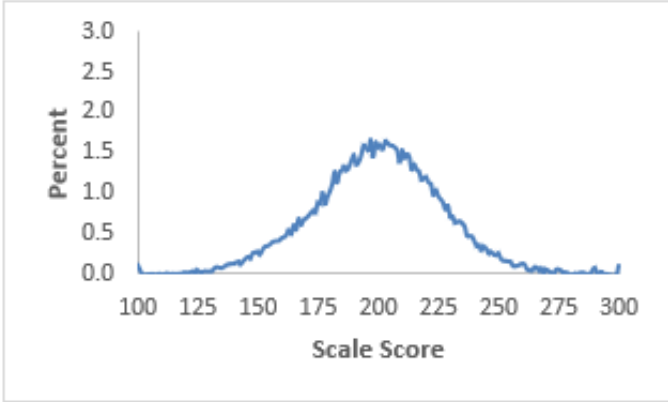
L-12. Science, Technology, Engineering and Mathematics (STEM) Grade 10 Composite Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	45,442	198.85	27.8	181	200	218	-0.26
Gender	Female	22,311	199.41	25.6	184	201	217	-0.36
	Male	23,130	198.31	29.8	179	200	219	-0.17
Ethnicity	Hispanic or Latino Ethnicity	7,927	182.86	25.9	167	183	200	-0.12
	Asian	850	205.58	29.9	186	207	225	-0.16
	Native Hawaiian or Other Pacific Islander	747	183.30	24.0	169	184	200	-0.35
	Black or African American	618	176.77	26.9	160	177	194	0.02
	American Indian or Alaska Native	508	182.84	24.6	168	185	199	-0.28
	White	33,661	203.29	26.5	187	205	221	-0.32
	Other	1,123	200.02	28.4	183	202	219	-0.27
Limited English Proficiency	No	43,115	200.56	27.0	184	202	219	-0.26
	Yes	2,327	166.27	22.3	153	168	180	-0.09
Economic Disadvantage	No	32,006	203.67	26.6	188	205	221	-0.30
	Yes	13,436	187.05	27.1	170	188	205	-0.15
Special Education	No	41,126	201.84	26.3	186	203	219	-0.25
	Yes	4,316	169.16	24.1	155	170	183	0.11

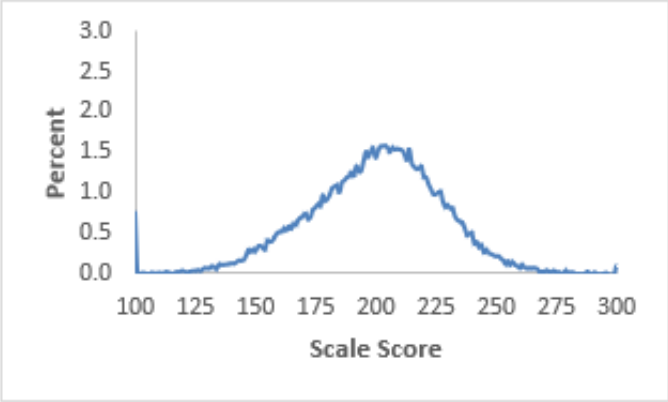
Appendix M: Scale Score Distributions for Overall Testing Population



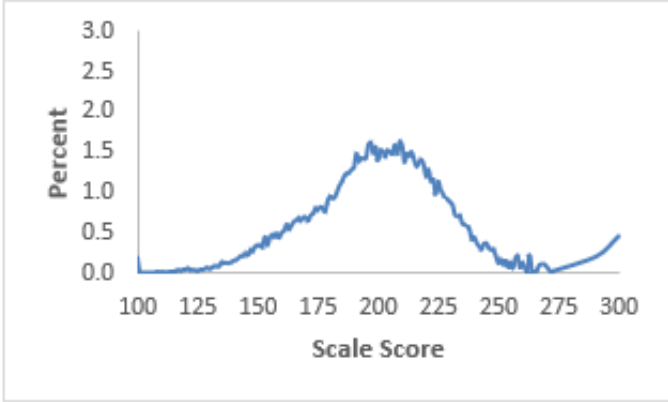
M-1. English Grade 9 Scale Score Distribution



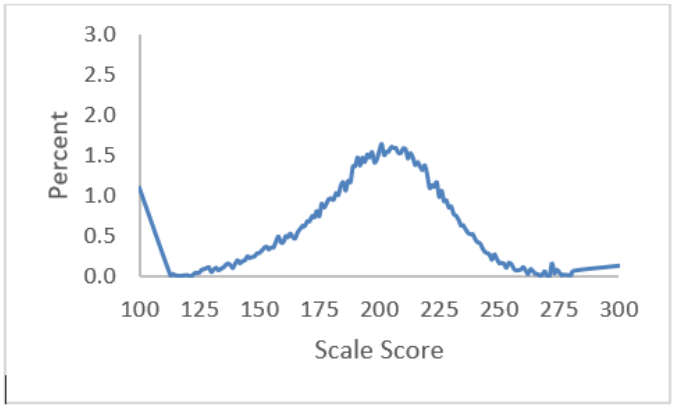
M-2. English Grade 10 Scale Score Distribution



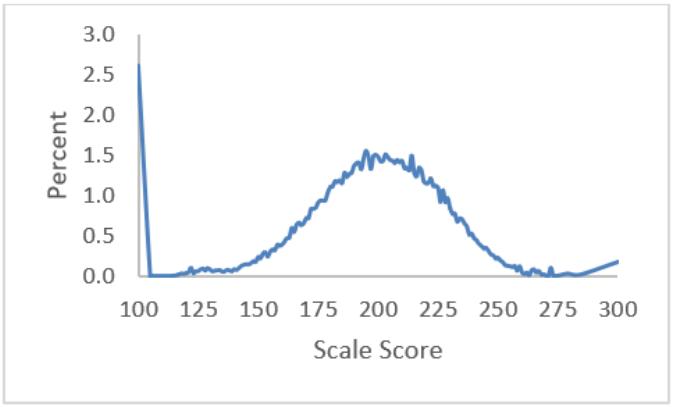
M-3. Reading Grade 9 Scale Score Distribution



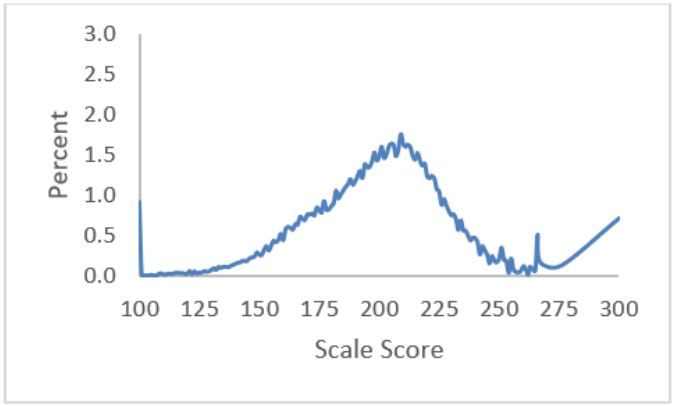
M-4. Reading Grade 10 Scale Score Distribution



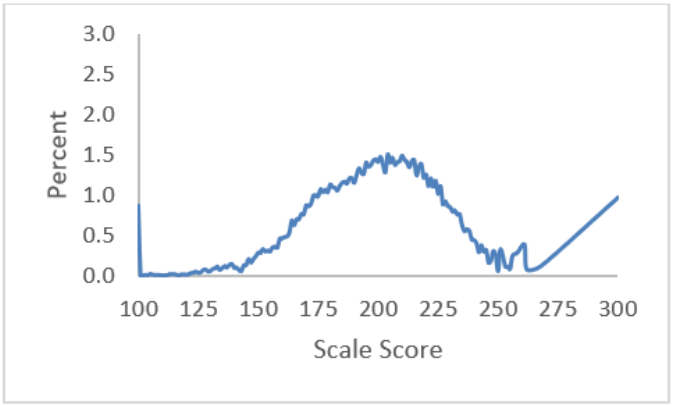
M-5. Mathematics Grade 9 Scale Score Distribution



M-6. Mathematics Grade 10 Scale Score Distribution



M-7. Science Grade 9 Scale Score Distribution



M-8. Science Grade 10 Scale Score Distribution

Appendix N: Performance Level Distributions

N-1. English Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	46,050	9.8%	42.1%	42.6%	5.5%
Gender	Female	22,626	6.1%	38.7%	48.1%	7.1%
	Male	23,422	13.5%	45.3%	37.3%	3.9%
Ethnicity	Hispanic or Latino Ethnicity	7,865	19.8%	55.6%	23.3%	1.2%
	Asian	815	9.1%	37.7%	43.6%	9.7%
	Native Hawaiian or Other Pacific Islander	716	17.2%	55.6%	26.3%	1.0%
	Black or African American	616	29.2%	50.2%	19.2%	1.5%
	American Indian or Alaska Native	518	21.0%	58.7%	19.9%	0.4%
	White	34,277	6.9%	38.5%	48.0%	6.5%
	Other	1,235	8.4%	39.9%	45.6%	6.1%
Limited English Proficiency	No	43,754	7.8%	41.8%	44.7%	5.7%
	Yes	2,296	48.2%	48.1%	3.7%	0.0%
Economic Disadvantage	No	31,683	6.2%	38.0%	49.0%	6.8%
	Yes	14,367	17.9%	51.1%	28.5%	2.5%
Special Education	No	41,505	6.5%	41.0%	46.5%	6.0%
	Yes	4,545	40.4%	51.5%	7.5%	0.5%

N-2. English Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,836	7.9%	41.0%	46.6%	4.6%
Gender	Female	21,565	4.2%	38.0%	51.9%	5.9%
	Male	22,270	11.4%	43.9%	41.4%	3.3%
Ethnicity	Hispanic or Latino Ethnicity	7,518	14.4%	58.2%	26.4%	1.0%
	Asian	822	4.9%	40.5%	47.1%	7.5%
	Native Hawaiian or Other Pacific Islander	694	11.5%	60.5%	27.1%	0.9%
	Black or African American	582	26.3%	53.6%	19.4%	0.7%
	American Indian or Alaska Native	483	14.9%	61.3%	23.6%	0.2%
	White	32,653	5.9%	36.2%	52.4%	5.5%
	Other	1,078	7.3%	40.1%	47.3%	5.3%
Limited English Proficiency	No	41,663	6.5%	39.9%	48.8%	4.8%
	Yes	2,173	33.7%	61.8%	4.5%	0.0%
Economic Disadvantage	No	31,083	5.6%	36.2%	52.6%	5.6%
	Yes	12,753	13.4%	52.8%	31.9%	2.0%
Special Education	No	39,798	5.7%	39.0%	50.4%	5.0%
	Yes	4,038	29.6%	61.2%	8.9%	0.3%

N-3. Reading Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	46,238	12.1%	41.4%	34.4%	12.2%
Gender	Female	22,724	8.4%	41.6%	36.7%	13.3%
	Male	23,513	15.6%	41.3%	32.0%	11.1%
Ethnicity	Hispanic or Latino Ethnicity	7,985	22.6%	51.4%	22.0%	4.0%
	Asian	815	10.7%	35.8%	38.8%	14.7%
	Native Hawaiian or Other Pacific Islander	724	23.5%	50.8%	22.4%	3.3%
	Black or African American	632	31.3%	49.2%	14.9%	4.6%
	American Indian or Alaska Native	529	24.2%	56.5%	17.0%	2.3%
	White	34,310	8.9%	38.7%	38.0%	14.4%
	Other	1,235	10.9%	40.8%	35.1%	13.3%
Limited English Proficiency	No	43,906	10.2%	41.1%	36.0%	12.8%
	Yes	2,332	47.3%	48.3%	4.0%	0.4%
Economic Disadvantage	No	31,723	8.2%	38.1%	38.8%	15.0%
	Yes	14,515	20.6%	48.7%	24.7%	6.0%
Special Education	No	41,675	8.8%	40.6%	37.2%	13.4%
	Yes	4,563	41.9%	48.5%	8.5%	1.1%

N-4. Reading Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	44,132	18.0%	35.6%	37.8%	8.6%
Gender	Female	21,711	12.4%	37.6%	41.5%	8.5%
	Male	22,420	23.5%	33.7%	34.2%	8.6%
Ethnicity	Hispanic or Latino Ethnicity	7,674	30.9%	44.0%	22.5%	2.6%
	Asian	827	16.0%	31.9%	41.2%	10.9%
	Native Hawaiian or Other Pacific Islander	719	32.7%	43.8%	22.0%	1.5%
	Black or African American	593	41.1%	38.3%	18.5%	2.0%
	American Indian or Alaska Native	492	31.9%	46.1%	20.7%	1.2%
	White	32,739	14.1%	33.4%	42.2%	10.3%
	Other	1,082	17.4%	34.5%	38.1%	10.1%
Limited English Proficiency	No	41,899	15.8%	35.6%	39.6%	9.0%
	Yes	2,233	59.2%	36.1%	4.5%	0.2%
Economic Disadvantage	No	31,175	13.8%	33.5%	42.3%	10.5%
	Yes	12,957	28.2%	40.8%	27.1%	4.0%
Special Education	No	40,044	14.6%	35.3%	40.7%	9.4%
	Yes	4,088	51.7%	38.5%	9.1%	0.7%

N-5. Mathematics Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	45,590	15.4%	41.5%	33.5%	9.5%
Gender	Female	22,386	12.8%	44.0%	35.5%	7.7%
	Male	23,203	18.0%	39.2%	31.6%	11.2%
Ethnicity	Hispanic or Latino Ethnicity	7,801	30.7%	49.4%	17.5%	2.4%
	Asian	812	11.7%	35.0%	35.6%	17.7%
	Native Hawaiian or Other Pacific Islander	710	27.9%	51.1%	18.9%	2.1%
	Black or African American	612	42.8%	41.8%	14.7%	0.7%
	American Indian or Alaska Native	513	32.2%	50.1%	15.4%	2.3%
	White	33,914	11.0%	39.6%	38.1%	11.4%
	Other	1,220	14.4%	41.5%	33.9%	10.2%
Limited English Proficiency	No	43,311	13.3%	41.7%	35.0%	10.0%
	Yes	2,279	55.7%	39.3%	4.5%	0.5%
Economic Disadvantage	No	31,366	10.0%	39.0%	38.9%	12.1%
	Yes	14,224	27.5%	47.1%	21.6%	3.8%
Special Education	No	41,116	10.9%	42.0%	36.6%	10.4%
	Yes	4,474	56.7%	37.1%	5.0%	1.1%

N-6. Mathematics Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,705	24.1%	39.6%	28.4%	7.9%
Gender	Female	21,504	21.5%	42.5%	29.7%	6.3%
	Male	22,200	26.6%	36.8%	27.2%	9.5%
Ethnicity	Hispanic or Latino Ethnicity	7,542	44.6%	40.0%	13.5%	1.9%
	Asian	824	19.3%	32.2%	33.5%	15.0%
	Native Hawaiian or Other Pacific Islander	710	37.9%	47.3%	13.2%	1.5%
	Black or African American	581	54.2%	35.1%	9.8%	0.9%
	American Indian or Alaska Native	490	41.2%	44.9%	12.9%	1.0%
	White	32,484	18.4%	39.6%	32.6%	9.4%
	Other	1,068	23.0%	39.3%	29.3%	8.3%
Limited English Proficiency	No	41,501	21.6%	40.4%	29.7%	8.3%
	Yes	2,204	71.4%	25.1%	3.2%	0.3%
Economic Disadvantage	No	30,936	17.9%	39.2%	33.0%	9.9%
	Yes	12,769	39.1%	40.7%	17.2%	3.0%
Special Education	No	39,653	19.5%	41.0%	30.9%	8.6%
	Yes	4,052	69.3%	26.4%	3.7%	0.5%

N-7. Science Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	46,149	10.8%	53.2%	29.4%	6.6%
Gender	Female	22,683	8.7%	55.7%	30.0%	5.6%
	Male	23,465	13.0%	50.8%	28.7%	7.5%
Ethnicity	Hispanic or Latino Ethnicity	7,949	21.9%	62.1%	14.3%	1.7%
	Asian	820	8.3%	46.3%	34.5%	10.9%
	Native Hawaiian or Other Pacific Islander	721	23.6%	61.9%	13.0%	1.5%
	Black or African American	630	27.9%	61.6%	8.7%	1.7%
	American Indian or Alaska Native	532	20.5%	66.7%	11.8%	0.9%
	White	34,250	7.6%	50.8%	33.7%	7.9%
	Other	1,239	9.9%	52.2%	31.3%	6.5%
Limited English Proficiency	No	43,816	9.1%	53.2%	30.8%	6.9%
	Yes	2,333	44.1%	53.2%	2.5%	0.2%
Economic Disadvantage	No	31,713	7.0%	50.4%	34.3%	8.3%
	Yes	14,436	19.3%	59.4%	18.5%	2.9%
Special Education	No	41,601	8.0%	52.9%	31.9%	7.2%
	Yes	4,548	37.1%	56.0%	6.0%	0.9%

N-8. Science Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,901	12.4%	54.8%	26.4%	6.3%
Gender	Female	21,581	10.7%	57.6%	26.8%	4.9%
	Male	22,319	14.1%	52.1%	26.0%	7.8%
Ethnicity	Hispanic or Latino Ethnicity	7,582	24.3%	62.7%	11.2%	1.8%
	Asian	828	10.1%	49.4%	27.7%	12.8%
	Native Hawaiian or Other Pacific Islander	713	25.1%	64.9%	9.7%	0.3%
	Black or African American	591	31.1%	59.6%	8.3%	1.0%
	American Indian or Alaska Native	485	26.4%	60.8%	12.2%	0.6%
	White	32,617	8.9%	52.7%	30.8%	7.5%
	Other	1,079	11.2%	54.5%	27.2%	7.1%
Limited English Proficiency	No	41,687	10.7%	54.9%	27.7%	6.7%
	Yes	2,214	45.3%	52.7%	1.8%	0.2%
Economic Disadvantage	No	31,079	9.1%	52.4%	30.7%	7.9%
	Yes	12,822	20.6%	60.8%	16.1%	2.6%
Special Education	No	39,834	9.8%	54.6%	28.6%	6.9%
	Yes	4,067	38.1%	56.6%	4.6%	0.7%

N-9. English Language Arts (ELA) Grade 9 Composite Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	45,569	10.7%	41.0%	41.2%	7.1%
Gender	Female	22,408	6.8%	39.5%	45.2%	8.6%
	Male	23,160	14.4%	42.6%	37.4%	5.6%
Ethnicity	Hispanic or Latino Ethnicity	7,753	21.6%	53.3%	23.6%	1.5%
	Asian	806	9.2%	37.3%	41.4%	12.0%
	Native Hawaiian or Other Pacific Islander	701	20.3%	53.5%	25.0%	1.3%
	Black or African American	608	30.8%	48.8%	17.8%	2.6%
	American Indian or Alaska Native	507	20.9%	58.6%	19.7%	0.8%
	White	33,968	7.6%	37.7%	46.2%	8.5%
	Other	1,218	8.9%	40.1%	43.8%	7.2%
Limited English Proficiency	No	43,317	8.6%	40.8%	43.2%	7.4%
	Yes	2,252	50.4%	46.2%	3.3%	0.1%
Economic Disadvantage	No	31,416	6.7%	37.3%	47.2%	8.7%
	Yes	14,153	19.4%	49.3%	28.0%	3.3%
Special Education	No	41,135	7.1%	40.3%	44.9%	7.8%
	Yes	4,434	43.8%	48.2%	7.4%	0.5%

N-10. English Language Arts (ELA) Grade 10 Composite Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,252	11.1%	39.4%	44.0%	5.5%
Gender	Female	21,285	6.5%	38.4%	48.8%	6.3%
	Male	21,966	15.5%	40.4%	39.3%	4.8%
Ethnicity	Hispanic or Latino Ethnicity	7,393	20.5%	53.3%	24.8%	1.3%
	Asian	816	8.6%	38.1%	44.9%	8.5%
	Native Hawaiian or Other Pacific Islander	680	20.3%	55.3%	23.8%	0.6%
	Black or African American	569	32.0%	48.7%	18.3%	1.1%
	American Indian or Alaska Native	475	22.9%	55.6%	21.1%	0.4%
	White	32,256	8.3%	35.5%	49.5%	6.6%
	Other	1,057	10.0%	38.5%	44.8%	6.6%
Limited English Proficiency	No	41,118	9.3%	38.8%	46%	5.8%
	Yes	2,134	45.7%	50.2%	4.0%	0.0%
Economic Disadvantage	No	30,716	7.9%	35.8%	49.6%	6.8%
	Yes	12,536	19.0%	48.3%	30.3%	2.3%
Special Education	No	39,301	8.2%	38.2%	47.6%	6.0%
	Yes	3,951	40.1%	51.1%	8.3%	0.5%

N-11. Science, Technology, Engineering, and Math (STEM) Grade 9 Composite Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	45,137	12.4%	48.7%	31.5%	7.4%
Gender	Female	22,196	9.8%	51.4%	32.8%	5.9%
	Male	22,940	14.9%	46.0%	30.3%	8.8%
Ethnicity	Hispanic or Latino Ethnicity	7,683	26.2%	57.0%	15.2%	1.7%
	Asian	807	8.7%	43.1%	34.8%	13.4%
	Native Hawaiian or Other Pacific Islander	698	25.5%	58.5%	14.8%	1.3%
	Black or African American	607	35.4%	52.2%	11.5%	0.8%
	American Indian or Alaska Native	503	24.7%	62.2%	11.9%	1.2%
	White	33,619	8.5%	46.5%	36.2%	8.9%
	Other	1,212	11.4%	48.0%	33.0%	7.6%
Limited English Proficiency	No	42,902	10.4%	48.9%	33.0%	7.7%
	Yes	2,235	52.2%	44.7%	2.8%	0.2%
Economic Disadvantage	No	31,141	7.6%	46.0%	36.9%	9.4%
	Yes	13,996	23.1%	54.5%	19.6%	2.8%
Special Education	No	40,772	8.5%	49.0%	34.4%	8.1%
	Yes	4,365	49.0%	45.5%	4.7%	0.8%

N-12. Science, Technology, Engineering, and Math (STEM) Grade 10 Composite Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,017	18.0%	48.3%	27.3%	6.3%
Gender	Female	21,164	15.4%	51.6%	28.1%	4.8%
	Male	21,852	20.5%	45.1%	26.5%	7.8%
Ethnicity	Hispanic or Latino Ethnicity	7,366	35.7%	51.7%	11.1%	1.5%
	Asian	818	15.0%	41.2%	30.7%	13.1%
	Native Hawaiian or Other Pacific Islander	692	34.0%	54.2%	11.4%	0.4%
	Black or African American	570	45.8%	44.4%	9.1%	0.7%
	American Indian or Alaska Native	477	34.0%	55.1%	10.5%	0.4%
	White	32,039	13.0%	47.6%	31.8%	7.6%
	Other	1,049	16.6%	47.4%	28.5%	7.5%
Limited English Proficiency	No	40,877	15.5%	49.1%	28.6%	6.7%
	Yes	2,140	65.0%	32.9%	1.9%	0.1%
Economic Disadvantage	No	30,540	12.8%	47.1%	32.1%	8.0%
	Yes	12,477	30.8%	51.3%	15.6%	2.3%
Special Education	No	39,080	13.9%	49.5%	29.7%	6.9%
	Yes	3,937	59.3%	36.6%	3.6%	0.5%

Appendix O: Standard Processes and Quality Management

Psychometrics Services, Pearson

Overview

Pearson Psychometrics Services conducts a variety of technical analyses and activities in supporting our programs and customers. In addition to test construction, Psychometric Services typically engaged in equating, scaling, field test analysis, data review, item bank creation and management, standard setting, and technical reporting. For all psychometric tasks, quality management is central to ensuring on-time and error-free results. The following are examples of some of the standard quality steps used by Psychometric Services to this end.

- Detailed technical specifications are created, reviewed, and followed for all psychometric activities.
- Standard statistical key check and adjudication analyses are conducted on operational data to ensure multiple-choice items have the correct scoring key applied and that technology-enhanced items are scored correctly.
- Standard quality checklists are used to document end-to-end data management, equating and test construction activities and ensure standard processes are conducted and details notated.
- For operational equating, scaling, and score file production, all plans are initially reviewed by senior management. All related analyses are independently reproduced either internally or externally to validate results. Senior staff review that all related work has been conducted as intended, all procedures followed, and results appear reasonable.
- For classical analysis or other analysis that do not directly contribute to student scores reasonableness checks are conducted.
- Standard-setting plans are developed according to industry standards. All technical analyses produced during the meetings are independently verified.
- For technical reports, all technical analyses are reviewed for reasonableness.

Psychometric Services also follows a continuous improvement model in supporting our cyclical project work. This includes utilizing standardized solutions instead of developing custom solutions for each project where possible. Also, with respect to data management, Psychometric Services follows strict adherence to responsible management of all personally identifiable information (PII).

The purpose of this document is to provide a high-level description of the processes and quality management steps of Pearson's Psychometric Services group. It includes the main tasks noted above, in addition to others such as data handling and general quality management. This document is provided for informational purposes only. It is not intended to serve as a contract document, as an amendment to any contracts, or to replace comprehensive process documentations for the various steps outlined in this document.

1. Data Inspection and Management

Most standard psychometric activities rely on data files containing student responses to test questions. All data files are specified and approved by Psychometric Services (among other groups) ahead of time in the form of file layouts. Data files are processed by the Assessment Technology Engineering (ATE) group and posted to Psychometric Services for use. This section describes Psychometric Services data file inspection and management.

1.1 Standard Data Inspection Checks

Upon receipt of any data file, it is imperative that the file be evaluated for completeness and correctness. Data Inspection Checks are meant to be applied to all data files we use to do the majority of our psychometric work: Student Data Files (SDF) and Item Response Files (IRF), and to be integrated with the Universal Work Process Specifications Document and the End to End Checklist. This process is completed independently by two different data analysts and the results verified.

These checks include:

- Data Read-in and Layout/Length checks
- General FREQs and Null Value Identification
- SDF and IRF Merge and Score Point validations
- Complex FREQs and Raw Score Recalculation
- IDM and Score Point Item Mean Compare

1.2 SAS Grid – Data Flow and Network File Archiving

Prior to delivery to Psychometric Services, data files are validated by the Customer Data Quality (CDQ) group, a group within ATE. Once verified, all original data files are passed from ATE to Psychometric Services on the SAS Grid. No original data files are moved and stored in Psychometric Services folders. They are held in SECURE_DATA folders and removed after 90 days. Only processed SAS datasets are stored within the appropriate folders on Grid. No student PII information (such as student name) can be contained in the files Psychometric Services processes, analyzes, and stores.

2. Statistical Key Check and Adjudication

Two quality checks are performed on operational and field test multiple-choice and technology-enhanced item types. Often these are conducted prior to the end of scoring in the event that issues are discovered to help ensure they can be addressed in a timely manner such that schedules are maintained.

Compare *p*-value, Point Biserial, and score point distribution to ABBI export

Prior to running keycheck and adjudication, an ABBI statistical report is pulled that contains the most recent *p*-value and pt-biserial. This is compared to the statistics generated for the current administration and flagging is implemented based on a compare of the values, according to project criteria. Additionally, score point distributions are generated and similarly flagged for proportions that are below a provided threshold.

2.1 TRIAN Standard Keycheck Process

The TRIAN analysis is used to evaluate multiple-choice items and identify items whose statistics fall outside typical thresholds such that they may be indicative of an incorrect answer key being applied. These analyses and reports are produced using standardized code (see Section 4). Flags are generated with respect to out of range *p*-values, point biserial correlations, score distributions, and across form differences.

These are run both at overall test (across all forms) and on a form by form basis. Flagged items are sent to the content group who inspect each item in its live version (e.g. reviewing the specific paper form a given item appears on, for example). While the flagged item may be indicative of an incorrect key being applied, it may also indicate administration anomalies such as misprints. This process is completed independently using the standardized code by two different data analysts and the results verified.

2.2 Adjudication Process

For technology-enhanced items, student responses are captured and evaluated by running standardized code (see Section 4) that produces frequency distributions of each. The lists are sorted by frequency, high to low, and indicate what score was assigned for each response. Responses that have been previously adjudicated for an item are not included on subsequent adjudication reports, as long as both the response and score match. These are sent to the content group to evaluate against the existing scoring logic.

3. Calibration, Equating, Scaling

The Pearson Scaling and Equating Process is used by Psychometric Services Research Scientists to scale and equate test scores. The process is applicable to all measurement models and all scaling and equating situations, however, several process steps assume the use of Item Response Theory methods.

Scaling/equating activities require appropriate research designs, sophisticated data analysis and finesse. Assigned research scientists must design the equating study to meet program and professional requirements, specify data collection activities based on design (i.e., sampling, forms design, timing of data collection), maintain awareness of test administration activities, evaluate data quality, compare “black box” computational routines to empirical evidence from other sources and a priori expectations, and make adjustments to computational output based on best professional judgment. Equating results are estimations of “true” equating relationships. Because of this, equating/scaling activities that lead to the production of scores used for high-stakes decisions must be verified by a psychometric staff member implementing the same procedures yet operating under independent conditions. This verification activity is not a replacement for quality control, and quality control is not a replacement for verification.

Since this activity directly results in the production of scores, independent internal verification of the analyses is completed by two psychometricians. This includes operational testing or, in the case of field testing where preequating is directly or indirectly used in score reporting. Results of the equating are verified at various times throughout the process, including:

- After initial calibration of items, verification of independent item parameters
- After stability check, verification of items included in equating process and final Stocking-Lord equating constants
- After item parameter tables for ATE for scoring, verification of final item parameters and exact format of files

For a given project, the Lead Research Scientist designs a system for processing student data and producing scaling/equating results. The system must include as a minimum:

- Systematic evaluation of the accuracy of data prior to beginning scaling/equating procedures (e.g., frequency distributions of all required variables, check the sum scores, etc.)
- Statistical key check
- Adjudications
- Indicates the points within the processing stream where assigned Research Scientists will analyze results
- Tested and verifiable computations performed by the software tools utilized
- Scaling/equating output in an agreed-upon format that can be readily consumable by technology

Participants, including assigned Research Scientists and Stat analysts review the system design.

3.1 Equating and Scaling Specifications

The Lead Research Scientist creates/updates specifications that document the process. The specifications must include details referring to every step of the process, from initial planning, to archiving of files and documentation.

All Psychometric Services staff assigned to the project should review the specifications prior to beginning the execution phase. The specifications should be reviewed for:

- Accuracy
- Clarity
- Completeness
- Attending to all requirements
- Compliance to this process, or documentation as to otherwise

3.2 Equating Dry Run (Practice) Prior to Live Administration

An assigned Research Scientist (or Research Associate) creates realistic mock data using the student data file layout. Alternatively, previous year's data can be used if it meets current year requirements.

Mock data should contain:

- Scored responses
- Total score
- Demographics: ethnicity, gender (others depending on project)
- Variables needed to test exclusion criteria (project specific)

The Lead Research Scientist develops a testing plan to thoroughly evaluate the system. Assigned Research Scientists and Stat analysts carry out the testing using the mock data.

The Lead Research Scientist schedules and monitors the completion of a practice scaling/equating session that includes all participants and every step of the process that the mock data will support, and begins at data receipt. This activity is scheduled early enough that corrections can be made to the process, but not so early that participants have to refresh themselves on the system when the operational work begins.

3.3 Equating Checklists

The Lead Research Scientist creates a customized equating checklist(s) using the Pearson standard equating checklist template to ensure that each step of the process is followed and that observations specific to each equated test are documented. Additionally, any additional customization pertaining to their specific program is addressed. The Prime fills out the checklist(s) and archives them with the scaling/equating results. Additionally, where a software tool generates a log file, the Prime archives this file in the same location as the scaling/equating results.

3.4 Senior Level External Review and Sign Off

Internally, a senior level external review of all equating is part of standard process. For this review, all specifications, documentation, inputs, and outputs from the process are made available to the reviewer. Once the reviewer has had a chance to evaluate every aspect of a given equating effort, a meeting is held to answer any outstanding questions, discuss findings, and ideally to affirm the reasonableness of the work and the final results. Once final approval is conferred, written approval is sent to the Lead Research Scientist as well as a completed reviewer checklist. The Lead Research Scientist archives the written agreement and approval.

3.5 Scoring File Upload to ePRS and Confirmation

All scoring files applied through ATE must be loaded to ePRS. Once equating is officially signed off, scoring files can be loaded to ePRS following the scoring table/score file loading process. Once loaded, all scoring files are extracted from ePRS and then compared against the original source files. Only after confirming all files match 100% from source to ePRS extract are they considered ready for scoring.

3.6 Verification of ISE Scoring (When Pattern Scoring is Used)

After ATE completes scoring of the student responses with the score tables and item parameter files provided, a verification of the assignment of scale scores to students is completed by psychometrics. A research scientist independently uses ISE to calculate a scale score and performance level to each student. This verifies that the calculation of scale scores by ATE is correct and that the correct item parameter tables are being used. Only after confirming all scale scores match 100% to ATE are they verified.

4. Item Analysis and Item Banking

Item level statistical analysis typically occurs in support of several tasks, such as part of the statistical key check (described earlier), as part of the calibration, equating, and scaling activities, and as part of field test analysis and test construction set up. The processes described in this section address use of standardized data and item analysis code for these purposes. Steps related to set up of the ABBI system for banking of item level statistics is also described.

4.1 Item Analysis using Standardized Data and Code

The two main purposes of standardized data formatting and item analysis code is to reduce the need to re-create programming of common psychometric tasks across projects. Doing so helps reduce errors as well as increase efficiency of processing.

Standard Data File Creation

The initial set up for running item analyses using the standardized code involves creation of the standard input data files. This is an activity that is independently replicated by two stat analysts and based on raw data files received from ATE. Variables within the originating data file are mapped to the standardized layout and the standardized data files are created in conjunction with the respective Pearson Test Maps. Once produced, data validation code is run to evaluate values in the standardized data file to confirm they are formatted appropriately and to check for any missing variables or data elements. Once validated, data are compared across replicators to ensure 100% match.

Standard Item Analysis

The code is used with the standard input data in deriving item level statistics used for evaluating item characteristics and quality within the context of a given assessment. Statistics are loaded into an item bank for use in data review, test construction, and/or for historical record. In addition to classical item statistics such as *p*-value and point biserial, option and score point distributions are available as well as differential item functioning. Also, the standardized code is used for producing statistical key check and adjudication reports (described previously).

4.2 Item Banking in ABBI

School Assessment utilizes the Assessment Banking and Building for Interoperability (ABBI) system as a single unified interface for authoring test content, banking content elements and metadata (including item level statistics), building test forms, and publishing tests. ABBI supports a full spectrum of item types, from simple, multiple-choice items to complex technology-enhanced items. The ABBI system was built to support common interoperability standards for content encoding and accessibility (QTI, APIP, and WCAG) promote content portability across a wide range of platforms and provides robust quality management support for publishing. One of the primary tasks of Psychometric Services is to upload item level statistics into ABBI.

Stats Registry and Upload Specifications

Each grade and subject test needs to be set up with a registry that aligns statistics to be uploaded to standard variable names that exist in ABBI. This alignment is established by mapping variables as named in an upload file (e.g. as output from the standard item analysis code described previously) to the given ABBI variable name. Once created, the registry is checked to confirm display names are populated correctly and the originating dataset file is saved for comparison to registry to ensure all fields get populated.

All source files used to create the ABBI Upload file need to be documented in the ABBI Stats Upload Specs Document. This can include, but is not limited to, testmaps, classical statistics, DIF, Common Response Analysis, IRT parameters, and item fit files. Paths of the locations of the files are to be included in the Specs document. Additionally, if certain data columns are to be constructed by the Stat Analyst (e.g., flags for low *p*-value), the rules for those columns would be captured as well. The ABBI Specs document is also an important resource when migrating an existing item bank or combining files from different sources (e.g., an external vendor's files for older stats with ITTB extracts for newer stats) to load to ABBI for the first time. The ABBI Specs document reflects all rules needed for combining these files.

Uploading Statistics to ABBI and QC

Based on the registry and specifications, an ABBI upload file is created. Once created, a summary quality control report is produced that summarizes characteristics of the file's contents. Summaries are inspected for reasonability and/or to identify anomalies or missing components. Once confirmed as accurate, the file is uploaded to ABBI and a reasonability inspection is conducted within ABBI. A final quality control step is taken by exporting a file from ABBI with all the statistics. This is compared to the upload file itself.

5. Test Construction

The Pearson test construction process is used by Pearson research scientists, stat analysts, and content experts to support the assignment of test items to operational or field test forms.

Test construction is a complex, interactive task that requires both content and psychometric expertise. Test items often serve a variety of functions beyond establishing estimates of student ability, such as supporting scaling and equating activities. Research scientists must, therefore, evaluate the statistical quality of individual test items and test forms to ensure all defined psychometric requirements will be met. Content experts review individual items and the test form as a whole for appropriateness, clarity, and overall flow. They select replacement items and passages as needed to improve content integrity and support the intended purpose of the assessment, and consult with the customer on all content-related assessment issues. In addition, they document and track customer-requested changes to selected items and forms throughout the test creation process. After test form approval, these changes are provided to the forms department for use in authoring.

If the unique requirements of an assessment program require modification to one or more of the steps defined in this process, the modifications should be documented in the project Test Creation Specifications and approved by the content and psychometric functional managers supervising the project. The outputs of this process lead directly into the Test Map creation.

5.1 Test Construction Standard Specifications Template

The lead research scientist and test development manager collaborate in the development of the Test Creation Specifications (TCS) using the TCS Template as a guide. This template provides a suggested format for the specification document and summarizes best practices in the selection and sequencing of operational items.

The lead research scientist is responsible for the components of the TCS related to test design and the program in general. The test development manager creates the portions of the TCS related to content considerations for item selection (operational and field test) and the guidelines for selecting and sequencing field test items.

The TCS should provide details related to each of the steps discussed in this document, as well as the following:

- Purpose of the assessment
- General description of the assessment
- Administration considerations
- Measurement theory or mathematical model employed
- Tools available to support development (e.g., Tracker/Builder)
- The operational test creation and review process
- The field test creation and review process
- Content and statistical considerations in the test creation process
- Samples of supporting materials (checklists, approval forms, etc.)

After internal review, the TCS-approved specifications document is published to the project's central location. A consistent version control method should be followed for development and review of the TCS to ensure the most current version is in use. For a given test administration cycle, any deviations from the procedures/practices outlined in the TCS must be documented with rationale in the test construction checklist.

5.2 Test Construction Checklist

The standard Psychometric Services checklist is completed for each grade and subject test form created. All responsible parties complete the checklists according to the work they perform and enter comments regarding the particular characteristics of a given build as warranted. The checklist contains the minimal standard elements important to fulfill for a given test construction activity. It also provides a place to document specific technical and statistical characteristics of the tests. Completed checklists should be archived within the appropriate folder on the network drive.

5.3 Test Map QC

After test construction activities have concluded and test maps finalized and loaded, the research associate conducts several quality control checks to help ensure any anomalies are identified and errors corrected. Working from extracts of the test maps from the system, the following general checks are conducted:

- Frequencies – Reasonability checking of expected frequencies, for example:

    - Question No. by Section crosstab
    - Publishing Form Number check (paper only)
    - Form ID by Section crosstab
    - Section Name by Section Segment crosstab (online only)
    - Calculator by Section crosstab (paper only)
    - Item Type by Max Score Points crosstab
    - Publish Format by Scoring Destination crosstab
    - Item Type by Scoring Destination crosstab
- Comparisons – Compares one dataset to another
    - Identifies changes in a new extract compared to a previous version
    - Compares datasets from different administrations or accommodations
- Metadata Check – Used to compare metadata (fields) across forms for items
- Ineligible Check – Used to compare to a bank and identify if there are items on a test that may be marked ineligible in the bank
- Passage Check – Various checks to inspect the accuracy of passage associations and metadata
- Item Enemy Check – Verifies items listed as enemies don't appear on the same form

6. General Quality Management and Oversight

In addition to task-specific processes and quality management steps, there are other important activities that help ensure ongoing success in supporting our various projects.

6.1 Project Documentation

In addition to the process documents, specifications, and checklists outlined in previous sections, it is just as important for all team members to have access to the documentation that supports the larger project. These include such things as the following:

- Proposal and Contract Documents
- Customer Requirements
- Project Schedules and Critical Milestones
- Customer and Team Meeting Minutes
- Scoring and Aggregation Specifications
- Data File Layouts
- List of Functional Team Leads

6.2 Monthly Psychometric Metrics Collection

Project leads are responsible for providing weekly updates to ongoing project activities and accomplishments for visibility to leadership and to share any concerns or risks. For additional oversight, monthly metrics are collected to provide a comparative snapshot across projects of preparedness and customer support. Metrics include preparedness toward equating, the number of TAC meetings supported, the number of customer facing meetings supported, the number of formal documents created and delivered, and the number of additional analyses conducted that fall outside of standard delivery.

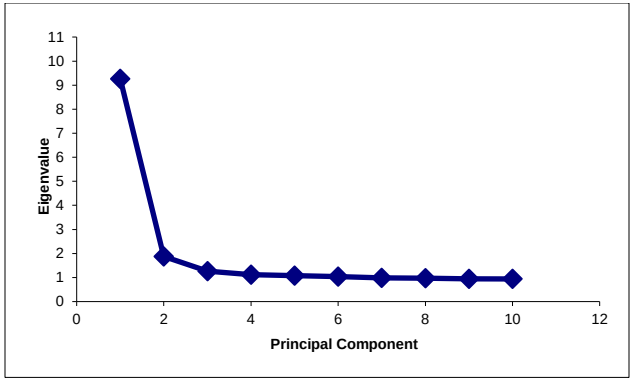
6.3 Service Level Agreement Metrics Tracking

Given the critical importance of the handoffs between ATE and Psychometric Services, a partnership was formed between the functional groups that specifies several task-specific expectations that have been agreed upon and captured in a Service Level Agreement (SLA). The agreement was established to ensure clean hand-offs across groups and allow for documentation and oversight moving forward. In addition to general expectations, the SLA applies to scoring files, keychecks and adjudication, and test construction. Metrics will be collected and discussed quarterly by the review committee.

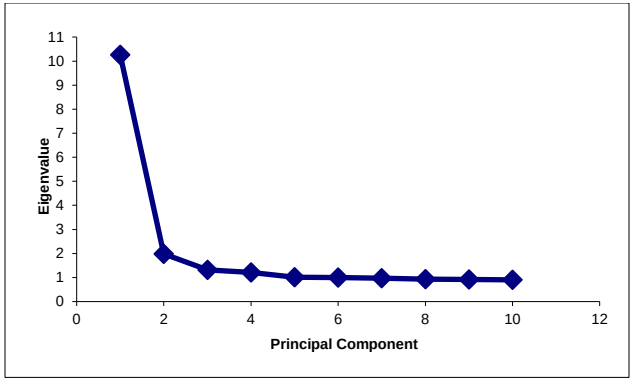
6.4 Risk and Issues Management

All formal process documents, flow charts, and checklists are maintained in a single repository within Compliant Pro. All Psychometric Services staff should be familiar with these and refer to them regularly when engaging in the setup and delivery of psychometric tasks. Risk and issues management, mitigation, and resolution is a continuous effort and requires the dedication and awareness of all. Noteworthy risks and/or identified issues are documented and managed through ServiceNow.

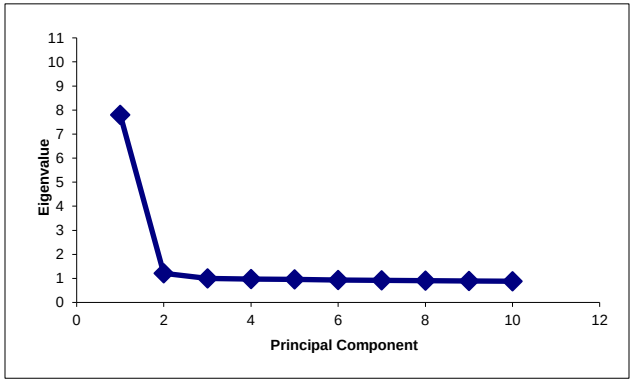
Appendix P: Principal Components Scree Plots



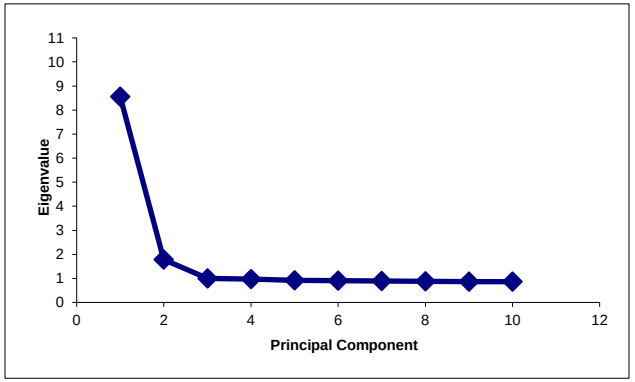
P-1. English Grade 9 Principal Components Scree Plot



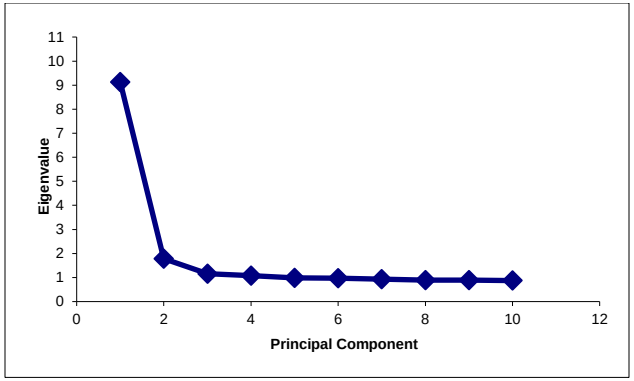
P-2. English Grade 10 Principal Components Scree Plot



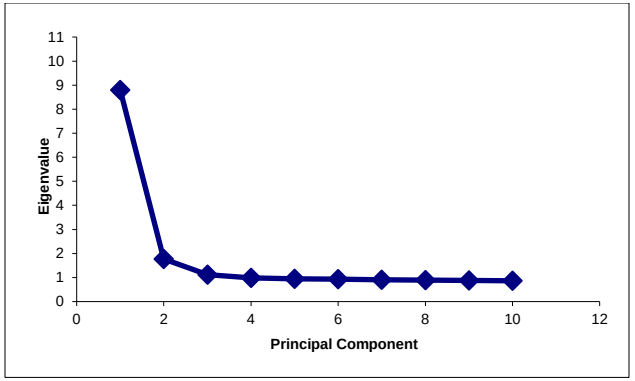
P-3. Reading Grade 9 Principal Components Scree Plot



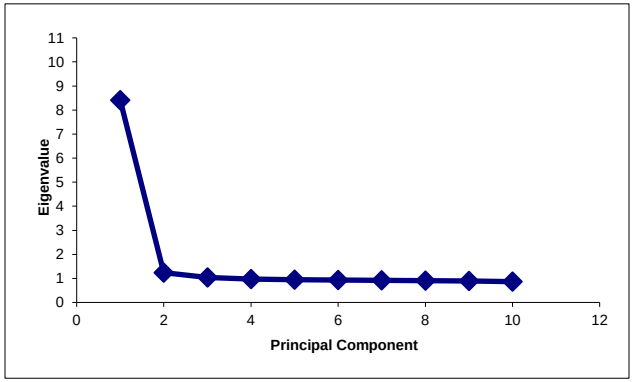
P-4. Reading Grade 10 Principal Components Scree Plot



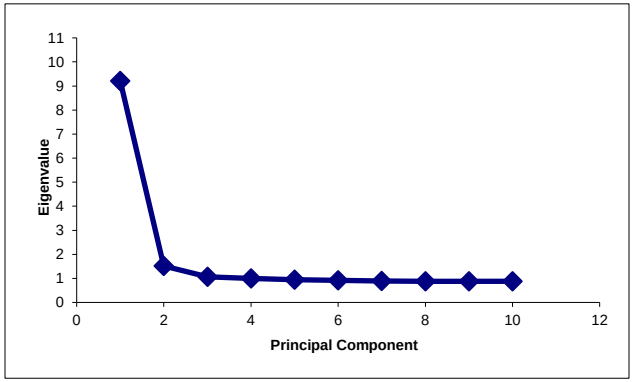
P-5. Mathematics Grade 9 Principal Components Scree Plot



P-6. Mathematics Grade 10 Principal Components Scree Plot



P-7. Science Grade 9 Principal Components Scree Plot



P-8. Science Grade 10 Principal Components Scree Plot

Appendix Q: Subscore Correlations

Q-1. English Correlations of Total Score and Subscores

Grade	Subdomain	English Total	Conventions of Standard English	Knowledge of Language	Production of Writing
9	Total	1.00	–	–	–
	Conventions of Standard English	0.98	1.00	–	–
	Knowledge of Language	0.63	0.54	1.00	–
	Production of Writing	0.81	0.71	0.49	1.00
10	Total	1.00	–	–	–
	Conventions of Standard English	0.98	1.00	–	–
	Knowledge of Language	0.73	0.65	1.00	--
	Production of Writing	0.84	0.74	0.59	1.00

Q-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	0.95	1.00	–	–
	Craft and Structure	0.92	0.77	1.00	–
	Integration of Knowledge and Ideas	0.54	0.46	0.41	1.00
10	Total	1.00	–	–	–
	Key Ideas	0.96	1.00	–	–
	Craft and Structure	0.92	0.80	1.00	–
	Integration of Knowledge and Ideas	0.71	0.60	0.58	1.00

Q-3. Mathematics Correlations of Total Score and Subscores

Grade	Subdomain	Mathematics Total	Algebra	Statistics and Probability	Functions	Geometry	Number and Quantity
9	Total	1.00	–	–	–	–	N/A
	Algebra	0.92	1.00	–	–	–	N/A
	Statistics and Probability	0.81	0.68	1.00	–	–	N/A
	Functions	0.88	0.75	0.63	1.00	–	N/A
	Geometry	0.86	0.71	0.61	0.67	1.00	N/A
10	Total	1.00	–	–	–	–	–
	Algebra	0.91	1.00	–	–	–	–
	Statistics and Probability	0.66	0.53	1.00	–	–	–
	Functions	0.86	0.71	0.48	1.00	–	–
	Geometry	0.89	0.73	0.54	0.68	1.00	–
	Number and Quantity	0.76	0.64	0.47	0.57	0.62	1.00

Q-4. Science Correlations of Total Score and Subscores

Grade	Subdomain	Science Total	ILO 1	ILO 3	ILO 4	ILO 5/6
9	Total	1.00	–	–	–	–
	ILO 1	0.97	1.00	–	–	–
	ILO 3	0.72	0.62	1.00	–	–
	ILO 4	0.86	0.75	0.55	1.00	--
	ILO 5/6	0.56	0.46	0.36	0.42	1.00
10	Total	1.00	–	–	–	–
	ILO 1	0.97	1.00	–	–	–
	ILO 3	0.79	0.69	1.00	–	–
	ILO 4	0.88	0.78	0.64	1.00	–
	ILO 5/6	0.72	0.63	0.52	0.60	1.00

ILO: "Intended Learning Outcome"