Utah Aspire Plus
2020–2021 Technical Report

2021

TABLE OF CONTENTS

List of Tables

1. Introduction

1.1 Background

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

Prior to 2019, students were assessed on the core standards through the Utah Student Assessment of Growth and Excellence (SAGE) assessment program. The Utah Aspire Plus assessment program is an extension of the Utah SAGE, still intended to measure student performance in relation to the Utah Core Standards (https://www.uen.org/core/), but also intending to measure students' preparedness for meeting college readiness benchmarks. As such, the assessment content from Utah SAGE is used as one component of the Utah Aspire Plus assessments.

Additional content from ACT Aspire is used to provide predictions of performance on the ACT®. This content also aligns to the Utah Core Standards and is counted toward Utah Aspire Plus scores too. The ACT® is the primary college readiness assessment submitted to local universities in Utah. As such, the Utah Aspire Plus assessments incorporate test questions from the ACT Aspire assessments that are used not only to contribute to student overall scores but also to provide a predictive indicator of performance on the ACT®. Students receive predicted ACT® score ranges for each ACT® subtest (English, reading, mathematics, and science), as well as an overall predicted composite ACT® score range.

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

As stated, the first operational administration was conducted in the spring of 2019 at grades 9 and 10 for English, reading, mathematics, and science. Data from that administration were used to establish the initial Utah Aspire reporting scales and the setting of performance levels. Technical details of these features and activities are presented in the *2018-2019 Utah Aspire Plus Technical Report* (http://utah.pearsonaccessnext.com/resources/additional-services/2018-19%20UA+%20Tech%20Report_Web.pdf).

Note that spring 2020 was intended to be the second operational administration of the Utah Aspire Plus tests. In spring of 2020, Senate Bill 3005, which included a waiver of the Utah Aspire Plus assessment requirements, was passed during the Utah Legislature's 3rd Special

Session of 2020 and signed into law on April 22, 2020. As a result, the spring testing of Utah Aspire Plus was cancelled. As a result, spring 2021 marked the second administration of the Utah Aspire Plus assessments. However, it should be noted that a waiver was sought and granted by the U.S. Department of Education (Department) to waive the accountability, school identification, and related reporting requirements for the 2020-2021 school year (https://www.schools.utah.gov/file/829f7300-020d-456e-85ac-49e85ef0795a).

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

The tests were administered as an operational field test, meaning that items used to provide scores for students were identified after the administration. That identification activity was akin to the standard test construction process involving Pearson and USBE content experts and psychometricians working to identify the best forms based on match to blueprint and statistical indices. After these forms were determined they were then used to set performance standards in August of 2021. 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).

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 (http://utah.pearsonaccessnext.com/additional-services/). All blueprints were designed to measure knowledge and skills described in the Utah Core Standards (https://www.uen.org/core/). For science, the operational assessments were created to measure the new Science with Engineering Education Standards (SEEds). The standards were derived from several research-based sources such as A Framework for K–12 Science Education and the Next Generation Science Standards (NGSS).

The Utah Aspire Plus tests are composed of several different types of items to measure student performance. These include multiple choice, multiple select, evidence-based selected response, and technology enhanced (TE). Multiple-choice items present students with four or five

responses, of which there is one correct answer. Multiple-select items require students to select two or three correct choices from several presented choices. Evidence-based selected response items have two parts: Part A is designed as an *identification* component, where Part B is designed to elicit an *evidence*-based component. Further, these types can be designed as two multiple-choice items, or a combination of multiple-choice and technology-enhanced (TE) items. Technology-enhanced (TE) items require specialized interactions within the online presentation for capturing student responses (e.g., drag and drop).

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

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

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

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

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

2. Test Development

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

The Utah Aspire Plus assessments are aligned to the Utah Core Standards and designed to measure the breadth and depth of the Utah Core Standards across all levels of student performance, to provide awareness of individual achievement in relation to stated performance expectations, and to provide evidence of whether students are on track for college and career readiness. Utah Aspire Plus content follows a rigorous development process that meets and often exceeds industry standards for best practices in assessment. Every item, written by Utah teachers, goes through an extensive review designed to ensure adherence to high quality and the principles of universal design.

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

It should be noted that while both claims and sub claims are presented here for each subject, only the claims are reported on individual student reports (ISR). Sub claims currently only provide structure within the respective blueprints but are <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
- 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.

* 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.

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

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

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

- Gathering and Investigating –

  SEPs: Asking questions and defining problems; Obtaining, evaluating, and communicating information; Planning and carrying out investigations

  CCCs: Patterns; Cause and effect; Systems and system models; Energy and matter; Structure and function; Stability and change Use Science Process and Thinking Skills

- Developing Models –

  SEPs: Developing and using models

  CCCs: Patterns; Cause and effect; Scale, proportion and quantity; Systems and system models; Energy and matter; Stability and change

- Using Mathematical Thinking –

  SEPs: Analyzing and interpreting data; Using mathematics and computational thinking

  CCCs: Patterns; Cause and effect; Scale, proportion, and quantity; Systems and system models; Energy and matter; Stability and change

- Constructing Explanations –

  SEPs: Constructing explanations and designing solutions; Engaging in argument from evidence

  CCCs: Patterns; Cause and effect; Systems and system models; Energy and matter; Structure and function; Stability and change

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

2.2 Utah Aspire Plus Blueprints

The Utah Aspire Plus tests are administered in English, reading, mathematics, and science in grades 9 and 10 and are described in Section 1.3. 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 new science tests, The SEEds blueprints assume a design in which one of the three DCIs will be assessed by two clusters and the other two DCIs with a single cluster. Coverage of the respective DCIs will rotate across forms (either within a given year or across years) to ensure the standards are fully represented over time. For 2021 the intention was to have three and five forms at grades 9 and 10 respectively, where the ACT clusters (sets of items associated with common stimuli) served as common linkages across all forms.

The 2021 Utah Aspire Plus blueprints can be found at:
http://utah.pearsonaccessnext.com/additional-services/.

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 Mathematics, Reading, and English assessments were built from existing Utah SAGE banked content combined with items from ACT Aspire to allow for predictions of students' preparedness for meeting college readiness. All available content for creation of the 2020 Utah Aspire Plus tests was based on the existing item banks described in the *2018-2019 Utah Aspire Plus Technical Report*.

For creation of the 2021 tests for mathematics, reading and English, two important design elements are worth noting. The first is that sets of items administered in 2019 were selected to serve as linking or common items that would be used to equate the 2021 Utah Aspire Plus tests to the 2019 base scales within a common item non-equivalent groups equating design (Kolen and Brennan, 2014). For test development purposes, this meant selecting sets of items to ideally reflect a miniature version of the overall test (typically at least 20 percent) in content as well as statistical characteristics. The second element worth noting is that a different set of ACT Aspire content was used for this second-year forms development activity. This helped limit exposure of the Aspire content that might otherwise negatively impact ACT predication score activities. However, it also meant linking sets used for equating did not have any ACT content available to serve as common items for the 2021 test forms. Still, final linking sets that reflected at least 20 percent of the overall tests and of comparable content were able to be selected for the Utah Aspire Plus tests.

2.3.1 Operational Forms Development

The construction of test forms for the 2021 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. 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.

The general test development process for Utah Aspire Plus was initiated with the selection of items from ACT Aspire. Items were selected based on match to blueprint, as well as statistical indicators of item quality and fairness provided from the SAGE and ACT Aspire banks, respectively. ACT Aspire items were positioned within each form in the same locations as originally administered within ACT Aspire forms to help facilitate the derivation of the predictive scores on Utah Aspire Plus.

Once the ACT Aspire items were selected, Pearson psychometrics selected sets of items common to 2019 that would be used to equate the 2021 tests to the 2019 base scales. In addition to selecting items to be as similar as possible to the overall blueprints, but they were also targeted to the original base scale difficulties.

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 in supporting the intended claims without negative impact. Moving forward, newly developed content will fill gaps and address such limitations as the assessments mature.

The new SEEds science assessments for Utah Aspire Plus were derived from test forms administered as an operational field test. This was necessary given the dual requirement of having to report student science scores on the new standards in 2021 and the desire to evaluate item level statistical performance and select the best overall test forms for reporting scores. For the spring 2021 initial administration of the SEEds science assessments, item statistics were derived and a test construction activity was conducted following the same process used to create the other Utah Aspire Plus tests. This process followed immediately after the administration window closed and prior to the SEEds standard setting meetings held in August of 2021.

2.3.2 Statistical Guidelines

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

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

- **Target item difficulty range of between 0.30 and 0.85.** Based on *p*-values, where the percentage reflects the percentage of students correctly responding to the item. Items

awarding more than one point used the item mean divided by the maximum points possible to place on the *p*-value metric.

- **Target threshold for item discrimination of 0.20 and above.** Where item discrimination is defined by item-total score correlations.
- **Extreme differential item functioning (DIF) indices should be avoided**. A standard flagging convention indicates differences of magnitude and classifies the most extreme cases of DIF as "C," moderate DIF as "B," and minor to no DIF as "A." As such, items flagged "C" should be avoided and minimal use of items flagged "B" should be used and/or balanced within a form where possible.

More detailed description of the statistical indices reflecting item functioning for the Utah Aspire Plus tests appears later in this report, and distributional results by grade and subject test from the 2021 operational administration are presented in Appendix C. *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.

2.3.3 2021 Match to Test Blueprint

Table 1 through 8 present the match between the final 2021 operational forms of Utah Aspire Plus and the test blueprints. English, reading, math, and science final forms reasonably matched all targets by item type, depth of knowledge, and reporting category (within 3 percent).

Table 1. Utah Aspire Plus English Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form
Item Type				
Multiple Choice	24–31	48%	62%	58%
Technology Enhanced	20–26	40%	52%	42%
Depth of Knowledge				
Level 1	22–33	44%	66%	57%
Level 2	5–12	10%	24%	16%
Level 3	12–17	24%	34%	27%
Reporting Categories				
Production of Writing	9–14	18%	28%	20%
Knowledge of Language	4–10	8%	20%	9%
Conventions of Standard English	28–38	56%	76%	71%

Table 2. Utah Aspire Plus English Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form
Item Type				
Multiple Choice	24–31	48%	62%	56%
Technology Enhanced	20–26	40%	52%	44%
Depth of Knowledge				
Level 1	22–33	44%	66%	54%
Level 2	5–12	10%	24%	15%
Level 3	12–17	24%	34%	30%
Reporting Categories				
Production of Writing	9–14	18%	28%	24%
Knowledge of Language	4–10	8%	20%	13%
Conventions of Standard English	28–38	56%	76%	63%

Table 3. Utah Aspire Plus Reading Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form
Item Type				
Multiple Choice	22–29	62%	82%	69%
Technology Enhanced	2–5	6%	14%	17%
Evidence-Based Selected Response	4–6	10%	17%	14%
Depth of Knowledge				
Level 1	4–10	11%	28%	11%
Level 2	12–20	34%	57%	49%
Level 3	9–14	25%	40%	40%
Reporting Categories				
Key Ideas	9–18	26%	51%	51%
Craft and Structure	14–20	40%	57%	37%
Integration of Knowledge and Ideas	3–5	9%	14%	11%

Table 4. Utah Aspire Plus Reading Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form
Item Type				
Multiple Choice	22–29	62%	82%	83%
Technology Enhanced	2–5	6%	14%	5%
Evidence-Based Selected Response	4–6	10%	17%	11%
Depth of Knowledge				
Level 1	4–10	11%	28%	14%
Level 2	12–20	34%	57%	47%
Level 3	9–14	25%	40%	39%
Reporting Categories				
Key Ideas	9–18	26%	51%	50%
Craft and Structure	14–20	40%	57%	38%
Integration of Knowledge and Ideas	3–5	9%	14%	11%

Table 5. Utah Aspire Plus Mathematics Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form
Item Type				
Multiple Choice	30–33	75%	83%	75%
Technology Enhanced	7–10	18%	25%	25%
Depth of Knowledge				
Level 1	8–12	20%	30%	28%
Level 2	15–20	38%	50%	50%
Level 3	9–13	23%	33%	23%
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 6. Utah Aspire Plus Mathematics Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form
Item Type				
Multiple Choice	30–33	75%	83%	78%
Technology Enhanced	7–10	18%	25%	23%
Depth of Knowledge				
Level 1	8–12	20%	30%	30%
Level 2	15–20	38%	50%	48%
Level 3	9–13	23%	33%	23%
Reporting Categories				
Number and Quantity	2–4	5%	10%	10%
Algebra	9–11	23%	28%	25%
Functions	10–12	25%	30%	28%
Geometry	11–13	28%	33%	30%
Statistics and Probability	2–4	5%	10%	8%

Table 7. Utah Aspire Plus Science Grade 9 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form 1	2021 Form 2	2021 Form 3
Item Type						
Multiple Choice/Select	19-22	76%	88%	96%	88%	92%
Technology Enhanced	3-6	12%	24%	4%	12%	8%
DCI						
Life	13-15	52%	60%	56%	56%	58%
Earth and Space	5-7	20%	28%	20%	20%	23%
Physical	5-7	20%	28%	24%	24%	19%
Reporting Categories						
Gathering & Investigating	6-8	24%	32%	28%	28%	27%
Developing Models	3-6	12%	24%	20%	24%	19%
Using Mathematical Thinking	9-11	36%	44%	36%	36%	38%
Construct Explanations	3-6	12%	24%	16%	12%	15%

Table 8. Utah Aspire Plus Science Grade 10 Operational Test Blueprint Match

	Number of Items	Minimum %	Maximum %	2021 Form 1	2021 Form 2	2021 Form 3	2021 Form 4	2021 Form 5
Item Type								
Multiple Choice/Select	20-21	87%	91%	100%	96%	96%	100%	96%
Technology Enhanced	2-3	9%	13%	0%	4%	4%	0%	4%
DCI								
Life	6-8	26%	35%	44%	32%	32%	30%	29%
Earth and Space	5-6	22%	26%	13%	18%	18%	26%	29%
Physical	10-11	43%	48%	44%	50%	50%	44%	42%
Reporting Categories								
Gathering & Investigating	2-3	9%	13%	9%	9%	9%	9%	8%
Developing Models	2-5	9%	22%	4%	5%	5%	17%	17%
Using Mathematical Thinking	5-9	22%	39%	31%	32%	32%	26%	38%
Construct Explanations	9-11	39%	48%	56%	55%	55%	48%	38%

For additional information on the 2021 operational forms, Appendix A contains a breakdown reporting categories and standards by item type and depth of knowledge (DOK), with the exception of science that does not use DOK.

3. Operational Administration

3.1 Testing Window

The 2021 administration of the Utah Aspire Plus assessments was March 8 - May 14, 2021. Utah Aspire Plus can be administered on a subject-by-subject basis or as a complete battery with all tests administered in one sitting. Each subject test, however, must be administered in one sitting. In other words, once a subject test is started, it must be completed within that sitting.

3.2 Test Administration and Security Policies

Comprehensive details of the Utah Aspire Plus test administration are 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://www.utah.gov/pmn/files/704165.pdf.

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 2020–2021 administration, the testing times were: 45 minutes for English,75 minutes each for Reading and Mathematics, and 60 minutes for Science. It should be noted that students whose IEP, Section 504, or English Learner plan specified an accommodation for extended time were able to use extended time accommodations on Utah Aspire Plus as appropriate.

3.2.1 Online Administration and Monitoring

The Utah Aspire Plus tests are administered online via the Pearson 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/UT1130483_SummSp21TAN_WebTag.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 9 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. It is important to note that roughly 4,000 fewer valid test scores were observed in 2021 compared to the previous administration of Utah Aspire Plus in 2019. This is clearly the result of the impact of Covid-19.

Table 9. Spring 2021 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	42,964	39,286	42,045	38,573	43,214	39,417	42,635	39,067
Gender	Female	47.85	48.31	47.51	48.11	47.74	48.22	47.68	48.13
	Male	52.15	51.69	52.49	51.89	52.26	51.78	52.32	51.87
Ethnicity	Hispanic or Latino Ethnicity	16.88	16.35	16.86	16.27	17.17	16.55	17.13	16.62
	Asian	1.66	1.72	1.67	1.73	1.67	1.73	1.67	1.73
	Native Hawaiian or Other Pacific Islander	1.39	1.25	1.34	1.31	1.37	1.25	1.38	1.28
	Black or African American	1.24	1.22	1.24	1.22	1.24	1.23	1.25	1.23
	American Indian or Alaska Native	0.74	0.68	0.76	0.68	0.76	0.69	0.76	0.69
	White	75.32	75.95	75.36	76.00	75.03	75.72	75.04	75.64
	Other	2.77	2.83	2.76	2.78	2.76	2.81	2.77	2.80
Limited English Proficiency	No	94.69	95.79	94.62	95.71	94.57	95.79	94.50	95.68
	Yes	5.31	4.21	5.38	4.29	5.43	4.21	5.50	4.32
Economic Disadvantage	No	74.91	76.91	74.98	76.94	74.64	76.73	74.64	76.62
	Yes	25.09	23.09	25.02	23.06	25.36	23.27	25.36	23.38
Special Education	No	90.47	91.23	90.51	91.23	90.39	91.18	90.41	91.29
	Yes	9.53	8.77	9.49	8.77	9.61	8.82	9.59	8.71

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, 75 minutes for Reading, 75 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 2021 test administration, student total testing time was analyzed for each test. Overall, students completed the assessments within the recommended testing times. Table 10 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 43 minutes or less.

Additional information is presented in Appendix B, which provides a graphical display (box-and-whisker plot) of student testing time for each test. Box-and-whisker plots present the same information at each respective quartile, where the middle 50% of the given distribution is the box, and the whiskers represent the bottom 25% and top 25% of the distribution. Dots represent outliers and reflect very few overall cases. Most outliers still for regular testers are within the time allotment for the subject. For example, the outliers for grade 9 English for regular testers are all below the 75-minute time threshold. Based on these data and plots, the evidence suggests students in general had enough time to complete each respective test within the given allotments.

Table 10. Student Testing Time for Spring 2021 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	38859	1	69	31	8	31	37	38	39	42	43
		Time and a Half	3258	1	83	35	14	35	44	47	50	55	62
		Double Time	518	6	122	38	17	38	46	51	55	61	70
		Triple Time	150	5	133	34	23	34	40	44	48	58	94
	10	Regular Time	35745	1	56	29	8	29	35	36	38	40	42
		Time and a Half	2990	2	67	33	14	33	41	44	47	52	60
		Double Time	255	2	86	34	15	34	40	43	48	53	62
		Triple Time	160	3	106	32	20	32	40	43	47	57	69
Reading	9	Regular Time	39064	1	107	46	15	46	56	59	62	66	70
		Time and a Half	3309	2	112	44	22	44	58	62	67	73	84
		Double Time	511	3	149	50	26	50	63	68	74	83	100
		Triple Time	141	4	213	47	33	47	59	63	66	74	110
	10	Regular Time	35905	1	84	41	14	41	50	53	56	60	66
		Time and a Half	2937	2	112	43	22	43	55	59	65	71	81
		Double Time	247	2	127	44	26	44	59	66	70	81	95
		Triple Time	163	3	158	44	28	44	55	61	67	78	95
Math	9	Regular Time	38050	2	74	50	14	50	60	63	65	68	71
		Time and a Half	3168	2	111	48	21	48	62	65	69	75	86
		Double Time	509	2	148	50	26	50	64	69	76	84	99
		Triple Time	137	4	167	54	29	54	68	73	79	85	108
	10	Regular Time	35075	1	76	46	16	46	58	61	64	68	71
		Time and a Half	2923	2	111	42	22	42	54	59	64	72	85
		Double Time	252	3	138	46	24	46	57	62	71	81	87
		Triple Time	156	3	169	34	25	34	42	44	47	53	71

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	38543	1	70	36	12	36	45	47	49	52	56
		Time and a Half	3252	2	93	35	17	35	45	49	53	58	66
		Double Time	515	2	123	39	20	39	49	53	58	64	76
		Triple Time	145	5	175	36	26	36	45	47	53	58	72
	10	Regular Time	35605	1	70	30	12	30	37	39	42	46	51
		Time and a Half	2889	2	96	28	17	28	37	41	45	50	61
		Double Time	249	3	87	31	17	31	41	44	48	55	64
		Triple Time	159	2	155	23	20	23	30	34	38	46	52

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 C.

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 11 the number of focal and reference group members scoring in each possible item response is captured.

Table 11. 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 C. Examination of the distribution of items by difficulty across each test shows that items do vary in difficulty across each test, with most items between 0.30 and 0.75. There are items that did fall outside the guidelines outlined previously, 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 tests that fell outside the guidelines but were ultimately included on final forms as the best available. Overall, even where items fell outside the guidelines, they were still useful. This was particularly true for the new science assessments, where due to the operational field testing, some very difficult items and items with low discrimination were included on final operational forms to help hit blueprint targets.

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 D 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*).

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 E, each including a line indicating the scale score cut score for Proficient. Ideally, the lowest value of conditional standard error of measurement occurs at the location of Proficient.

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 12 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 12. 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 13 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 13. 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 14 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 14. 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 14, 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 14 is 0.548. Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Estimates of classification accuracy and consistency indices—including kappa coefficients—for overall performance level classification and at the Proficient cut point are provided in Appendix F.

6. IRT Calibration, Equating, and Scaling

6.1 Overview

Item response theory (IRT) was used to create the base scales for the Utah Aspire Plus assessments. For the 2021 administration, IRT was used to place mathematics, reading, and English assessments onto their respective base scales through a process of equating. The new SEEds science assessment base scales were established this year using the same IRT methodology used in 2019 and will serve as the base for with future assessment will be compared to. After calibration and equating, the item parameters are used to compute a student's score in the IRT metric and then transformed to the final Utah Aspire Plus scale score reporting metric.

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

- IRT Data Preparation
- Description of the Calibration Process
- Model Fit Evaluation Criteria
- Summary Statistics and Distributions from IRT Analyses
- Common-Item Non-Equivalent Groups Equating
- IRT Pattern 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, Equating, and Scaling 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

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

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 noted, for mathematics, reading, and English, the additional step of equating was also completed to place these 2021 scores onto the original base Utah Aspire Plus base scales respectively. Whereas for science, this was the first administration of Utah Aspire Plus SEEds assessments, and these calibrations resulted in the base scales to which future assessments will be related to.

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 was checked.

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

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

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.

6.3.4 Equating

A common item non-equivalent groups approach (Kolen and Brennan, 2014) was used for equating operational forms for the Utah Aspire Plus mathematics, reading, and English assessments. Each form from the 2021 administration contains a set of items that are the same as appeared on the respective 2019 forms. These common (anchor) sets of items for Utah Aspire Plus were selected to represent a given blueprint in terms of content and each were roughly 20 percent or more of a full form.

The Stocking and Lord (1983) test characteristic curve methodology was used to derive equating constants for each grade-subject test. Using the 2019 IRT item parameter estimates for each of the Utah Aspire Plus anchor sets and the respective item parameter estimates from the 2021 administration described in the previous section, were used to obtain transformation constants. This was conducted using the computer program STUIRT (Kim & Kolen, 2004). Procedurally this was carried out in conjunction with an anchor item stability check procedure (described next) that resulted in a final set of transformation constants that were then applied to all 2021 calibrated items to complete the respective equating. Once completed, items and scores were on the original base Utah Aspire Plus reporting scales.

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

1. Obtain the theoretically weighted estimated posterior theta distribution using 31 quadrature points (-5 to 5).
2. Compute the slope and intercept constants using Stocking and Lord in STUIRT with all anchor items in the linking set.
3. Place the original anchor item parameter estimates onto the baseline scale by applying the constants obtained in Step 2.

4. For each anchor item, calculate D^2 between the ICCs based on old (x) and new (y) parameters at each point in this theta distribution:

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

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

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

Scatterplots of the common items can be found in Appendix G. Overall, item functioning of common items from 2019 to 2021 can be described as typical and stable. No more than one item in any of the common item sets were removed from final linking solutions. Scatterplots and correlations of IRT difficulty and discrimination parameters were highly related, where the lowest correlation on any set was .96 and the remaining were roughly .99. Final Stocking and Lord scaling constants used for placing tests onto the 2019 Utah Aspire Plus base scales are presented in Table 15.

Table 15. 2021 Final Stocking and Lord Scaling Constants

Subject	Grade	Slope	Intercept
English	**9**	0.931	-0.058
	10	0.969	-0.107
Reading	**9**	0.961	-0.085
	10	0.907	0.076
Math	**9**	1.010	-0.190
	10	0.998	-0.156

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, two of the grade 9 items flagged for misfit were short text-entry items on Conventions of Standard English. The other flagged item was a Production of Writing multiple-choice item. In grade 10, all 3 flagged items were Conventions of Standard English Items. One was a short text-entry and two were drop-down select 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, the grade 9 item that was flagged was a multi-select item. The grade 10 items one flagged item involved students using the equation editor tool to construct the response to the item. The grade 9 Science items were primarily the two-part EBSR items that require a correct response to Part A to receive credit on Part B. The other two items included one multiple-choice item and one multi-select item. A total of 16 items were flagged for grade 10 science. Five of these items were statistically close to having adequate model fit. Of the items flagged, five were two-part EBSR items, 5 were technology-enhanced items and 6 were multiple-choice items.

6.5 Steps Taken to Ensure Stability of 2021 Equating

Given the extraordinary circumstances faced by schools and districts to carry out instruction in the face of a global pandemic, there were countless questions around how to help guarantee that assessments were not only carried out as closely as possible to how they were previously (in 2019), but also around how to prepare to maintain score scales in the face of substantive differences to instruction and participation. As a result, there were several steps that were taken to identify and overcome potential negative impacts to the equating work.

The first concern was the representativeness of the testing populations. Evaluation of this was carried out by comparing demographic level characteristics from testers to the overall population characteristics (sex, race/ethnicity, English language learner, economically disadvantaged status, students with disabilities). While overall numbers of testers were down comparatively, (where counts were lower overall by roughly 4,000 students per grade compared to 2019), the demographics of the population characteristics matched to within 2 percent across all demographic groups and were deemed reasonably comparable. Additional inspection of regional representativeness did not suggest concerns were warranted and the equating work moved forward. It should also be noted that one of the most notable attributes of relying on item response theory as a means of scale score systems is that the methodology is not dependent on the underlying population of students responding to test questions. This attribute further supported the decision to proceed with the equating process.

Secondly, there was some concern that if instruction were incomplete (for example if not all standards were taught), then it was possible that common items measuring those standards could be directly impacted and need to be removed from contributing to equating. Beyond the direct

impact to a given anchor item set, an additional concern was that of the anchor set potentially not reflecting the overall test characteristics from a content standpoint. Examination of anchor item functioning is described above and was found to be stable in all instances compared to functioning in 2019. Special emphasis was also placed on evaluation of any groupings of items that might indicate a negative opportunity to learn effect and warrant closer scrutiny. None were identified.

6.6 Summary Statistics and Distributions from IRT Analyses

Tables 16 and 17 present the summary statistics for the IRT (*a*-, and *b*-) parameter estimates, and standard errors (SE) of the parameter estimates, and model fit information for the spring 2021 operational items. The summary statistics shown include the total number of items, along with the mean, standard deviation (SD), minimum, and maximum. As mentioned previously, exceptions were made in science in finalizing operational form selection. As such, one item in both grade 9 and 10 with extreme difficulty parameters were included to better meet blueprint targets.

Table 16. 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	44	0.91	0.37	0.32	1.77	–0.13	1.05	–2.70	2.83
	Reading	35	0.85	0.44	0.20	1.86	0.21	1.32	–1.87	5.26
	Mathematics	40	0.96	0.33	0.33	1.71	0.21	0.72	–0.83	2.49
	Science	60*	0.71	0.47	0.04	2.25	1.39	2.39	–0.77	10.15
10	English	46	0.83	0.33	0.34	1.51	–0.13	1.09	–1.99	4.30
	Reading	35	1.11	0.46	0.21	2.09	–0.31	0.71	–1.49	1.75
	Mathematics	40	1.10	0.30	0.48	1.67	0.35	0.83	–0.99	2.69
	Science	54*	0.85	0.48	0.02	2.39	1.68	6.43	–1.35	41.78

**Item counts for science reflect total unique items across all operational forms*

Table 17. 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	44	0.04	0.02	0.01	0.12	0.03	0.02	0.01	0.08
	Reading	35	0.04	0.03	0.01	0.1	0.04	0.04	0.01	0.27
	Mathematics	40	0.04	0.02	0.02	0.08	0.03	0.02	0.01	0.09
	Science	60*	0.11	0.17	0.02	1.13	0.11	0.15	0.02	0.95
10	English	46	0.04	0.02	0.01	0.08	0.04	0.04	0.01	0.24
	Reading	35	0.05	0.02	0.01	0.11	0.03	0.04	0.01	0.23
	Mathematics	40	0.05	0.02	0.02	0.10	0.03	0.02	0.01	0.12
	Science	54*	0.09	0.05	0.03	0.30	0.06	0.08	0.01	0.5

**Item counts for science reflect total unique items across all operational forms*

6.7 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* 2021 Utah Aspire Plus operational items were used for estimating individual student scores for all regular forms.

6.7.1 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:

$$SS = Theta * 25 + 200$$

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 April 2021, Utah educators were convened to create and recommend the PLDs for Utah Aspire Plus new SEEds assessments. 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 G contains the training slides and agenda for educator groups convened. The approved final PLDs can be found online. Final scale score cuts for science, English, reading and mathematics are presented in Table 18.

In August 2021, 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. Appendix I contains the standard setting executive summary. 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.

Table 18. Utah Aspire Plus Scale Score Cuts by Grade and Subject

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

7.4 ACT Predicted Score Ranges

As noted, one of the goals of the Utah Aspire Plus assessments is to be predictive of college readiness at grades 9 and 10, and the means of this is in terms of providing prediction score ranges of performance on the ACT for the four subject tests (English, 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. Specific details from the original prediction score studies can be found in the 2018-2019 Utah Aspire Technical Report.

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

Appendix J provides the details of the second longitudinal study from spring 2021. This study included students who were in 9th grade in 2019 and took the ACT as 11th grade students in spring 2021. Within it are described steps taken in evaluating the ACT samples in relation to previous administrations and efforts to improve predications based on a weighting procedure.

Generally, these updated prediction score ranges are tighter than the original ranges derived indirectly through the ACT Aspire scales.

7.5 2020–2021 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. Scale score distributions for each Utah Aspire Plus assessment are provided in Appendix M, for the overall testing population. Appendix N contains the performance level distributions of each Utah Aspire Plus. The tables contain the percentages of students being classified into each respective performance level.

While results can be compared directly to 2019 performance within same subject and grade, extra cautions should be taken with respect to interpretations beyond high-level due to impacts from the pandemic. These opportunity-to-learn (OTL) impacts are multi-faceted and differential across the state. Self-reported OTL data were collected from students taking the Utah Aspire Plus tests this year by USBE to help gain insight into how the pandemic impacted student learning experiences. A link to the resulting data can be found at: https://public.tableau.com/app/profile/data.and.statistics/viz/OpportunityToLearnPreviewDemographicPublic2021-07-28/OpportunitytoLearnDashboard . Users can explore summarized responses to the respective questions overall and by LEA. However, it should be noted that not all students or districts responded to the survey.

In addition to the fact that fewer students tested compared to 2019 and the fact that there could be effects tied to having a waiver for accountability are all reasons to be cautious and avoid drawing conclusions when interpreting or comparing scores at other aggregations (e.g., school, LEA, subgroup, etc.), as differences could be magnified. And while generally speaking performance was lower this year compared to 2019, it's possible that results could be more marked if all eligible testers had participated and in ways that may not be obvious. For example, if it were shown that missing testers reflected a part of the overall population who would tend to score lower, it would mean performance results would appear better in turn (and vice versa).

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. Details of Pearson's quality and control procedures for all psychometric tasks conducted, to include test construction, calibration, equating, scaling, field test analysis, data review, item bank creation and management, standard setting, and technical reporting, can be found in the Utah Aspire Plus 2018-2019 technical report.

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 A.

Also described in Chapters 2, Utah educators created and recommended performance level descriptors for the Utah Aspire Plus tests, which provide a description of typical end-of-grade performance expectations for each level of achievement in relation to the Utah Core Standards. The PLDs are descriptions of the knowledge and skills demonstrated by students in each performance category. Higher scores translate to a greater level of knowledge and skills demonstrated. There is a link between the PLDs and the knowledge and skills required to meet proficiency according to the standards.

PLDs are used to relate performance on Utah Aspire Plus tests to the Utah Core Standards through the process of standard setting. As described, content experts and stakeholders participated in standard setting in August 2019 for mathematics, reading and English. In August 2021, similar meetings were conducted in support of the new Utah Aspire Plus SEEds science tests. 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 SAGE 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 it is noted that in the 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 O.

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 19.

Table 19. Model Fit Indices for Confirmatory Factor Analyses

Subject	Grade	Form	SRMR	RMSEA	RMSEA 90% Lower CL	RMSEA 90% Upper CL
English	9	1	0.0291	0.0332	0.0329	0.0335
	10	1	0.0242	0.0276	0.0274	0.0279
Reading	9	1	0.0198	0.0223	0.0219	0.0226
	10	1	0.0292	0.0304	0.0289	0.0319
Mathematics	9	1	0.0265	0.0268	0.0254	0.0283
	10	1	0.0306	0.0270	0.0254	0.0285
Science	9	1	0.0212	0.0250	0.0243	0.0257
		2	0.0243	0.0275	0.0268	0.0282
		3	0.0229	0.0268	0.0261	0.0275
	10	1	0.0272	0.0319	0.0309	0.0330
		2	0.0216	0.0216	0.0237	0.0259
		3	0.0219	0.0252	0.2694	0.0263
		4	0.0214	0.0242	0.0231	0.0253
		5	0.0225	0.0255	0.0244	0.0266

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 P.

Lastly, given the administration of the new Utah Aspire Plus SEEds assessments based on multidimensional standards and structured with item sets, there was a need to evaluate the potential dependency of items. That is, within the context of deriving score scales for the Utah Aspire Plus science assessments by applying a unidimensional IRT model, there is an assumption of local independence of items. In effect this means that the only thing that should make a difference to student performance on different items is their ability (specified by the model). That is, there is no dependency on contributing information from different items that influences performance above and beyond overall ability. To the extent that such dependencies do occur, this can have deleterious effects on the measurement characteristics of the assessment (such as standard errors) and of score interpretations. As such it is important to be able to demonstrate the level of local dependency (LD) that exists on a given assessment and determine the extent to which this may need to be managed explicitly.

To evaluate LD among the Utah Aspire Plus science test questions, Yen's Q_3 statistic (Yen, 1984, 1993) was applied to data from each respective core form. The statistic reflects the correlation between performance on two contrasted items after accounting for performance on the overall assessment (residual correlations). In theory, values of Q_3 for any pair of items should be generally close to zero (indicating they are uncorrelated/functioning independently).

For Q_3, a critical value of 0.20 is often used as a threshold to define meaningful LD (Chen & Thissen, 1997). For these analyses, a critical threshold based on 0.20 above the average Q_3 correlation was used to flag item contrasts as indicative of exhibiting LD. After examination of all item-by-item contrasts on each respective Utah Aspire Plus core form across grades 9 and 10, only one was flagged. Average Q_3 values in all instances were roughly -.01. Lastly, no set-based comparisons suggested there was any explicit dependency based on test structure.

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.

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.

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

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

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

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

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

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

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

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

Jöresky, K., & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Scientific Software International Inc.

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

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

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

Loehlin, J. C. (1987). *Latent Variable Models*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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

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

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

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

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

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

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

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

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

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

English

Grade	Reporting Category: Standard	Multiple Choice			Technology Enhanced		
		DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
9	Conventions of Standard English: L.9-10.1	6	4	0	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	0	0	0	1	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	1	0	0	0	0	0
	Knowledge of Language: L.9-10.3	1	2	3	0	0	0
	Production of Writing: W.9-10.4	5	0	0	0	0	0
	Production of Writing: W.9-10.5	3	0	0	0	0	0
	Total	**44**					
10	Conventions of Standard English: L.9-10.1	4	4	0	0	0	0
	Conventions of Standard English: L.9-10.1a	0	0	0	2	0	0
	Conventions of Standard English: L.9-10.1b	0	0	0	3	0	0
	Conventions of Standard English: L.9-10.2	0	0	0	2	0	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
	Knowledge of Language: L.9-10.3	6	0	0	0	0	0
	Production of Writing: W.9-10.4	3	8	0	0	0	0
	Total	**46**					

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	Craft and Structure: L.9-10.4a	1	0	0	0	0	0	0	0	0
	Craft and Structure: L.9-10.5a	1	0	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.4	0	0	0	2	0	0	0	0	1
	Craft and Structure: RI.9-10.5	1	0	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.6	1	0	0	0	0	0	0	0	1
	Craft and Structure: RL.9-10.4	2	0	0	0	0	0	0	0	0
	Craft and Structure: RL.9-10.5	1	0	0	0	0	0	0	0	0
	Craft and Structure: RL.9-10.6	1	0	0	0	0	0	0	0	0
	Integration of Knowledge and Ideas: RI.9-10.8	0	0	0	1	0	0	0	0	0
	Integration of Knowledge and Ideas: RL.9-10.9	0	0	0	0	0	0	0	0	1
	Key Ideas: RI.9-10.1	0	0	0	1	0	0	0	0	0
	Key Ideas: RI.9-10.2	0	0	0	1	0	0	0	0	1
	Key Ideas: RL.9-10.1	1	1	1	0	0	0	0	0	1
	Key Ideas: RL.9-10.2	1	1	0	0	0	0	0	0	0
	Key Ideas: RL.9-10.3	1	0	0	0	0	0	0	0	0
	Total	**35**								
10	Craft and Structure: L.9-10.4a	1	0	0	0	0	0	0	0	0
	Craft and Structure: L.9.10.6	1	0	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.4	1	0	0	0	0	0	0	0	0
	Craft and Structure: RI.9-10.5	0	0	0	1	0	0	0	0	0
	Craft and Structure: RI.9-10.6	1	1	0	0	0	0	0	0	1
	Craft and Structure: RL.9-10.4	1	1	0	0	0	0	0	0	1
	Craft and Structure: RL.9-10.5	2	0	0	0	0	0	0	0	0
	Craft and Structure: RL.9-10.6	1	0	0	0	0	0	0	0	0
	Integration of Knowledge and Ideas: CCRA.R.5	1	0	0	0	0	0	0	0	0

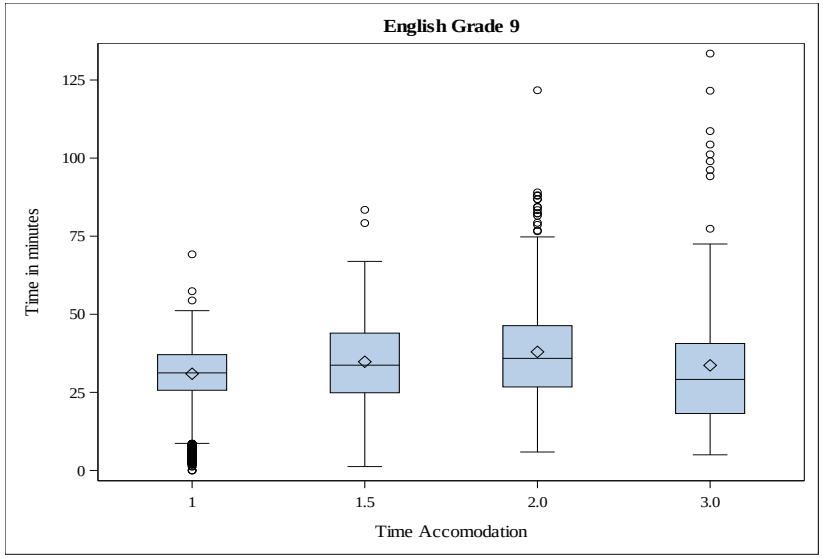
Grade	Reporting Category: Standard	Multiple Choice			Technology Enhanced			Evidence-Based Selected Response		
		DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
	Integration of Knowledge and Ideas: RI.9-10.8	1	1	0	0	0	0	0	0	0
	Integration of Knowledge and Ideas: RI.9-10.9	1	0	0	0	0	0	0	0	0
	Key Ideas: RI.9-10.1	3	2	0	0	0	0	0	0	0
	Key Ideas: RI.9-10.2	0	0	0	1	0	0	0	0	1
	Key Ideas: RI.9-10.3	1	0	0	0	0	0	0	0	1
	Total	**36**								

Mathematics

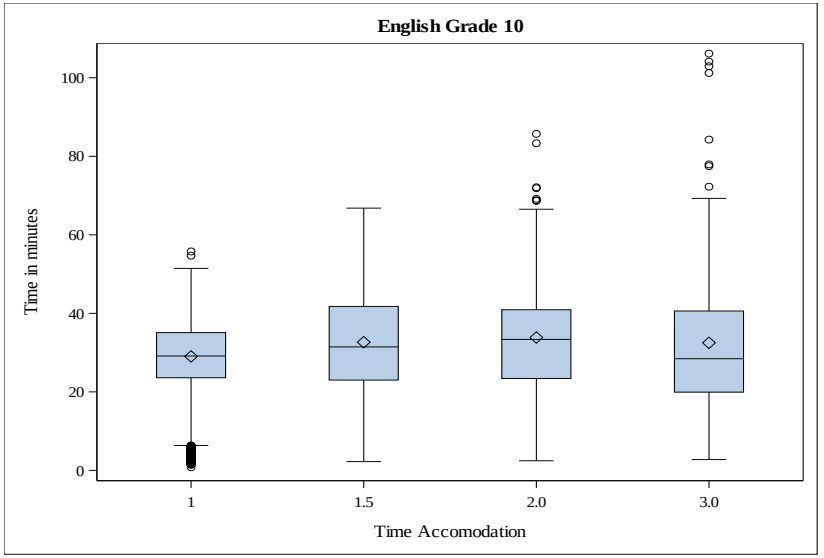
Grade	Reporting Category: Standard	Multiple Choice			Technology Enhanced		
		DOK 1	DOK 2	DOK 3	DOK 1	DOK 2	DOK 3
9	Algebra: MI.A.CED.1	1	0	0	0	0	0
	Algebra: MI.A.CED.2	0	0	0	1	0	0
	Algebra: MI.A.CED.4	1	0	0	0	0	0
	Algebra: MI.A.REI.1	1	0	0	0	0	0
	Algebra: MI.A.REI.10	1	0	0	0	0	0
	Algebra: MI.A.REI.3	0	0	0	1	0	0
	Algebra: MI.A.REI.6	0	0	0	1	0	0
	Algebra: MI.A.SSE.1b	1	0	0	0	0	0
	Functions: MI.F.BF.1a	1	0	0	0	0	0
	Functions: MI.F.BF.1b	0	0	0	1	0	0
	Functions: MI.F.IF.2	1	0	0	0	0	0
	Functions: MI.F.IF.4	1	0	0	0	0	0
	Functions: MI.F.IF.6	1	0	0	0	0	0
	Functions: MI.F.IF.7a	1	1	0	0	0	0
	Functions: MI.F.LE.1b	0	0	0	1	0	0
	Functions: MI.F.LE.1c	1	0	0	0	0	0
	Functions: MI.F.LE.2	1	0	0	0	0	0
	Functions: MI.F.LE.5	1	0	0	0	0	0
	Geometry: MI.G.CO.12	0	0	0	1	0	0
	Geometry: MI.G.CO.2	0	0	0	1	0	0
	Geometry: MI.G.CO.3	1	0	0	0	0	0
	Geometry: MI.G.CO.5	0	0	0	1	0	0
	Geometry: MI.G.CO.7	1	1	0	0	0	0
	Geometry: MI.G.GPE.4	0	0	0	1	0	0
	Geometry: MI.G.GPE.5	1	0	0	0	0	0
	Geometry: MI.G.GPE.7	1	0	0	0	0	0
	Statistics and Probability: MI.S.ID.1	0	0	0	1	0	0
	Statistics and Probability: MI.S.ID.2	2	0	0	0	0	0
	Statistics and Probability: MI.S.ID.6	1	0	0	0	0	0
	Statistics and Probability: MI.S.ID.6a	1	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	1	0	0	0	0	0
	Total	**40**					
10	Algebra: MII.A.APR.1	1	0	0	0	0	0

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

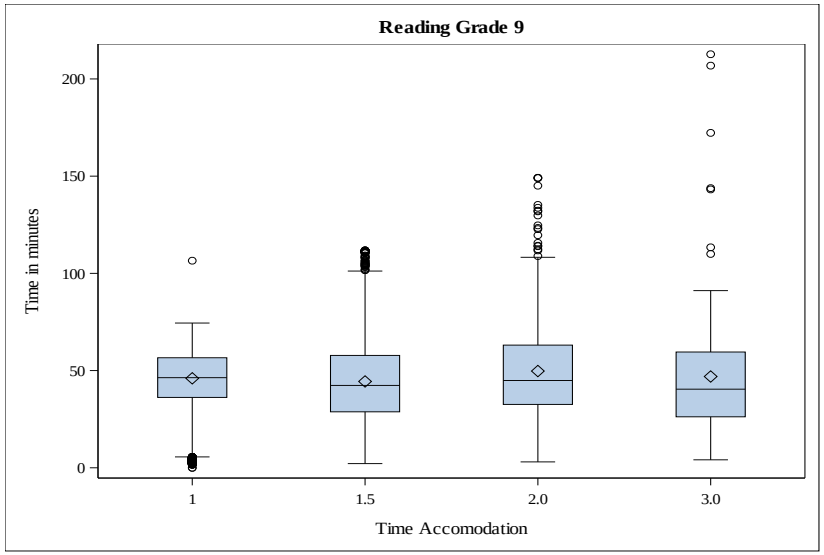
Appendix B: Student Testing Time



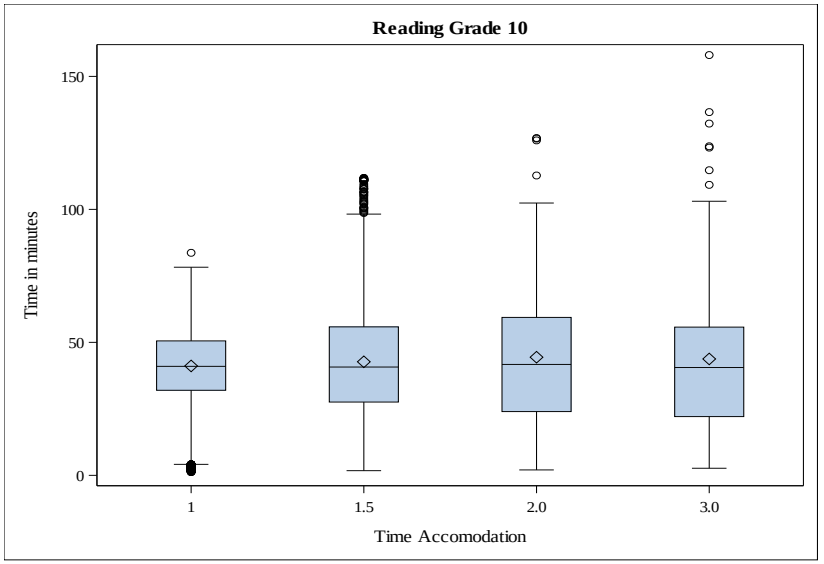
B-1. English Grade 9 Student Testing Time



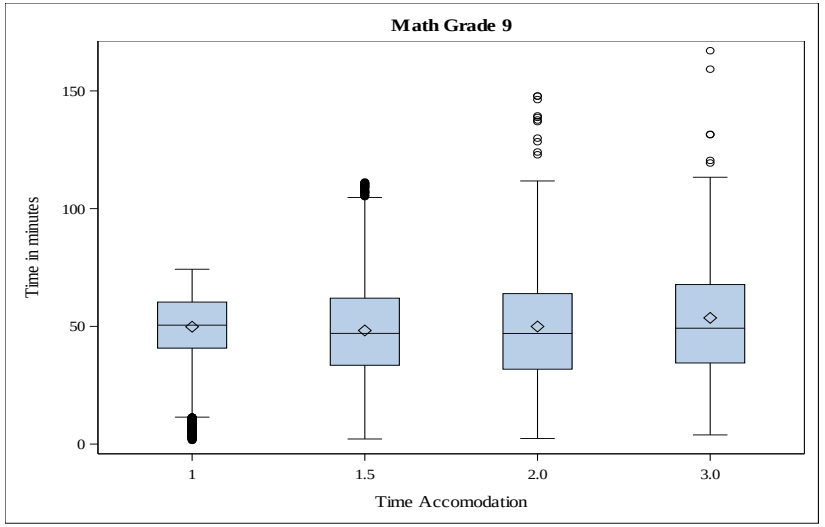
B-2. English Grade 10 Student Testing Time



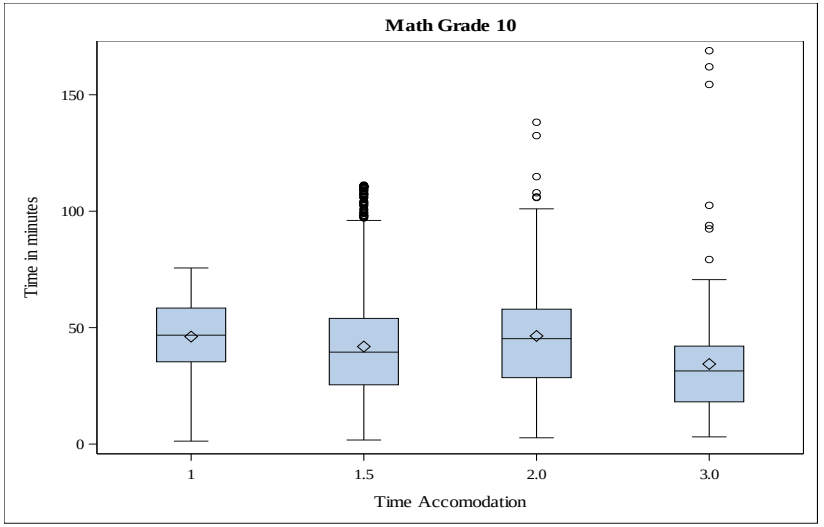
B-3. Reading Grade 9 Student Testing Time



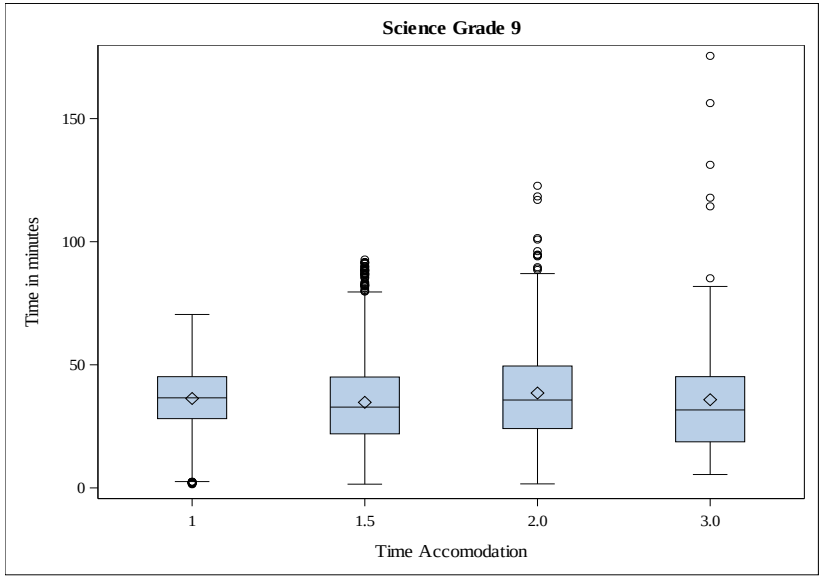
B-4. Reading Grade 10 Student Testing Time



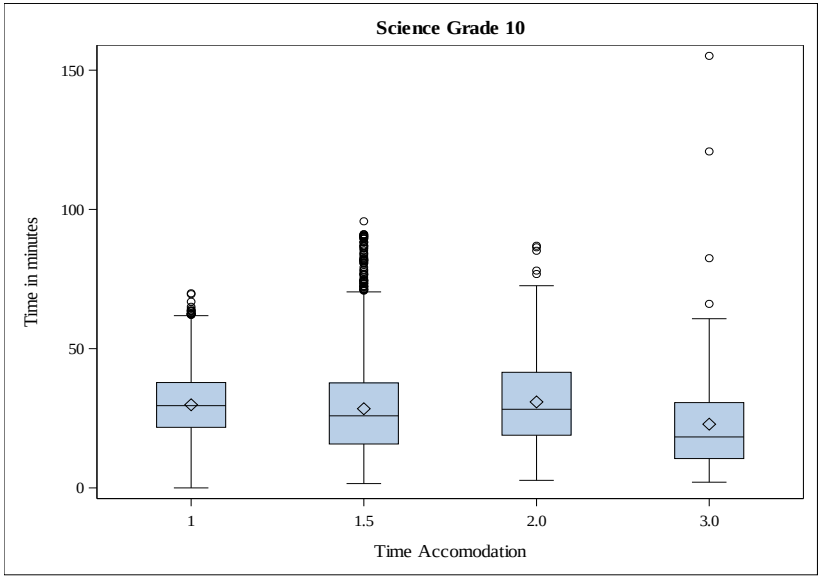
B-5. Math Grade 9 Student Testing Time



B-6. Math Grade 10 Student Testing Time



B-7. Science Grade 9 Student Testing Time



B-8. Science Grade 10 Student Testing Time

Appendix C: 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	1	8	12	9	–	0.62	30
	10	–	9	17	4	–	0.61	30
Reading	9	2	8	14	5	–	0.58	29
	10	1	9	12	10	–	0.65	32
Mathematics	9	6	15	19	–	–	0.49	40
	10	10	10	19	1	–	0.48	40
Science	9	5	13	14	–	–	0.46	32
	10	–	17	11	–	–	0.50	28

Two-Point Items

Subject	Grade	N	Mean	Min	Max
English	9	4	2.18	1.68	3.07
	10	4	2.38	1.59	3.42
Reading	9	6	0.61	0.43	0.77
	10	3	1.20	0.68	1.66
Mathematics	9	–	–	–	–
	10	–	–	–	–
Science	9	11	0.72	0.09	1.2
	10	14	0.84	0.43	1.66

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	14	–	–	–	0.39	39
	10	–	16	–	–	–	0.41	44
Reading	9	3	15	–	–	–	0.4	29
	10	1	26	2	–	–	0.48	31
Mathematics	9	1	25	1	–	–	0.45	40
	10	–	24	–	–	–	0.42	39
Science	9	3	9	–	–	–	0.31	36
	10	1	12	–	–	–	0.38	36

Two-Point Items

Subject	Grade	*N*	Median r	Min r	Max r
English	9	4	0.67	0.65	0.71
	10	4	0.59	0.58	0.7
Reading	9	6	0.43	0.25	0.51
	10	3	0.62	0.48	0.81
Mathematics	9	–	–	–	–
	10	–	–	–	–
Science	9	11	0.37	0.03	0.61
	10	14	0.35	0.02	0.74

Differential Item Functioning

Subject	Grade	Subgroups	DIF Categories				
			Negligible DIF	Moderate DIF		Substantial DIF	
				Focal	Reference	Focal	Reference
English	9	Male-Female	34	–	–	–	–
		White-Hispanic	34	–	–	–	–
	10	Male-Female	33	–	1	–	–
		White-Hispanic	33	–	1	–	–
Reading	9	Male-Female	35	–	–	–	–
		White-Hispanic	35	–	–	–	–
	10	Male-Female	31	2	1	1	–
		White-Hispanic	32	–	2	–	1
Mathematics	9	Male-Female	37	–	1	1	1
		White-Hispanic	40	–	–	–	–
	10	Male-Female	38	–	2	–	–
		White-Hispanic	40	–	–	–	–
Science	9	Male-Female	40	1	2	–	–
		White-Hispanic	43	–	–	–	–
	10	Male-Female	39	–	–	–	–
		White-Hispanic	42	–	3	–	–

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

Appendix D: Reliability and Standard Error by Subgroup

D-1. English Grade 9 Test Reliability

	Test Group	*N*	Alpha	SEM	Conventions of Standard English	Knowledge of Language	Production of Writing
All	Students Tested	42,964	0.88	8.77	0.83	0.59	0.70
Gender	Female	20,555	0.88	8.54	0.82	0.57	0.66
	Male	22,405	0.89	8.88	0.83	0.60	0.72
Ethnicity	Hispanic or Latino Ethnicity	7,251	0.87	8.82	0.79	0.54	0.68
	Asian	715	0.89	9.10	0.84	0.60	0.68
	Native Hawaiian or Other Pacific Islander	596	0.85	8.60	0.77	0.48	0.66
	Black or African American	534	0.87	9.37	0.80	0.55	0.68
	American Indian or Alaska Native	318	0.87	8.13	0.79	0.55	0.66
	White	32,361	0.88	8.69	0.82	0.57	0.68
	Other	1,189	0.87	8.83	0.82	0.56	0.67
Limited English Proficiency	No	40,684	0.88	8.68	0.82	0.57	0.68
	Yes	2,280	0.79	9.80	0.69	0.35	0.58
Economic Disadvantage	No	32,186	0.88	8.68	0.82	0.57	0.68
	Yes	10,778	0.88	8.88	0.82	0.58	0.70
Special Education	No	38,870	0.87	8.61	0.82	0.57	0.67
	Yes	4,094	0.83	9.44	0.75	0.43	0.62

D-2. English Grade 10 Test Reliability

	Test Group	*N*	Alpha	SEM	Conventions of Standard English	Knowledge of Language	Production of Writing
All	Students Tested	39,286	0.89	8.81	0.83	0.53	0.72
Gender	Female	18,975	0.88	8.66	0.82	0.50	0.72
	Male	20,305	0.89	8.88	0.84	0.56	0.72
Ethnicity	Hispanic or Latino Ethnicity	6,425	0.87	8.81	0.81	0.48	0.63
	Asian	677	0.90	9.19	0.85	0.53	0.76
	Native Hawaiian or Other Pacific Islander	491	0.83	8.47	0.79	0.41	0.55
	Black or African American	478	0.87	9.33	0.82	0.50	0.60
	American Indian or Alaska Native	268	0.85	8.48	0.79	0.40	0.63
	White	29,837	0.89	8.77	0.83	0.53	0.72
	Other	1,110	0.88	8.78	0.82	0.53	0.70
Limited English Proficiency	No	37,632	0.89	8.75	0.82	0.52	0.72
	Yes	1,654	0.75	10.11	0.70	0.37	0.27
Economic Disadvantage	No	30,214	0.89	8.78	0.83	0.53	0.72
	Yes	9,072	0.88	8.87	0.82	0.51	0.67
Special Education	No	35,842	0.88	8.72	0.82	0.51	0.72
	Yes	3,444	0.82	9.50	0.76	0.47	0.49

D-3. Reading Grade 9 Test Reliability

	Test Group	*N*	Alpha	SEM	Craft and Structure	Integration of Knowledge and Ideas	Key Ideas
All	Students Tested	43,214	0.86	10.30	0.67	0.21	0.80
Gender	Female	20,627	0.85	10.08	0.65	0.22	0.79
	Male	22,583	0.86	10.46	0.68	0.20	0.80
Ethnicity	Hispanic or Latino Ethnicity	7,418	0.84	10.93	0.65	0.17	0.76
	Asian	723	0.86	10.30	0.67	0.25	0.80
	Native Hawaiian or Other Pacific Islander	591	0.82	10.99	0.60	0.16	0.76
	Black or African American	537	0.84	11.28	0.66	0.13	0.77
	American Indian or Alaska Native	328	0.81	10.31	0.61	0.05	0.73
	White	32,424	0.85	10.13	0.65	0.21	0.79
	Other	1,193	0.86	10.10	0.68	0.25	0.79
Limited English Proficiency	No	40,868	0.85	10.16	0.66	0.21	0.79
	Yes	2,346	0.68	13.39	0.44	0.05	0.56
Economic Disadvantage	No	32,255	0.85	10.12	0.65	0.21	0.79
	Yes	10,959	0.85	10.76	0.67	0.17	0.78
Special Education	No	39,060	0.85	10.09	0.65	0.20	0.79
	Yes	4,154	0.76	12.21	0.54	0.10	0.66

D-4. Reading Grade 10 Test Reliability

	Test Group	*N*	Alpha	SEM	Craft and Structure	Integration of Knowledge and Ideas	Key Ideas
All	Students Tested	39,417	0.91	7.79	0.79	0.51	0.84
Gender	Female	19,003	0.90	7.72	0.77	0.48	0.82
	Male	20,408	0.92	7.85	0.81	0.53	0.85
Ethnicity	Hispanic or Latino Ethnicity	6,525	0.90	7.55	0.78	0.49	0.82
	Asian	683	0.91	7.85	0.79	0.50	0.83
	Native Hawaiian or Other Pacific Islander	493	0.89	8.01	0.75	0.46	0.81
	Black or African American	486	0.91	7.52	0.81	0.54	0.83
	American Indian or Alaska Native	273	0.90	7.59	0.78	0.53	0.81
	White	29,848	0.91	7.85	0.78	0.50	0.83
	Other	1,109	0.90	7.72	0.78	0.51	0.83
Limited English Proficiency	No	37,757	0.91	7.80	0.78	0.50	0.83
	Yes	1,660	0.85	8.25	0.66	0.36	0.74
Economic Disadvantage	No	30,246	0.91	7.84	0.78	0.50	0.83
	Yes	9,171	0.91	7.68	0.79	0.51	0.83
Special Education	No	35,941	0.90	7.81	0.77	0.49	0.83
	Yes	3,476	0.88	8.12	0.75	0.40	0.79

D-5. Math Grade 9 Test Reliability

	Test Group	N	Alpha	SEM	Algebra	Functions	Geometry	Number and Quantity
All	Students Tested	41,973	0.91	9.36	0.77	0.77	0.70	0.61
Gender	Female	19,946	0.90	9.24	0.75	0.76	0.67	0.59
	Male	22,024	0.92	9.42	0.79	0.78	0.72	0.64
Ethnicity	Hispanic or Latino Ethnicity	7,063	0.88	11.04	0.70	0.73	0.59	0.53
	Asian	703	0.92	9.08	0.80	0.77	0.72	0.60
	Native Hawaiian or Other Pacific Islander	565	0.87	11.66	0.66	0.72	0.59	0.47
	Black or African American	519	0.84	12.46	0.63	0.67	0.54	0.46
	American Indian or Alaska Native	320	0.86	10.89	0.68	0.71	0.58	0.48
	White	31,642	0.91	8.97	0.77	0.76	0.69	0.60
	Other	1,161	0.91	9.38	0.77	0.77	0.70	0.60
Limited English Proficiency	No	39,735	0.91	9.16	0.77	0.77	0.69	0.61
	Yes	2,238	0.76	14.61	0.49	0.55	0.39	0.33
Economic Disadvantage	No	31,497	0.91	9.01	0.77	0.76	0.70	0.61
	Yes	10,476	0.90	10.47	0.74	0.75	0.65	0.57
Special Education	No	38,032	0.90	8.95	0.76	0.76	0.69	0.60
	Yes	3,941	0.82	13.55	0.59	0.61	0.51	0.41

D-6. Math Grade 10 Test Reliability

	Test Group	N	Alpha	SEM	Algebra	Functions	Geometry	Number and Quantity	Statistics and Probability
All	Students Tested	38,573	0.90	10.55	0.70	0.67	0.80	0.47	0.37
Gender	Female	18,553	0.89	10.33	0.67	0.61	0.79	0.43	0.34
	Male	20,014	0.91	10.69	0.73	0.70	0.82	0.50	0.41
Ethnicity	Hispanic or Latino Ethnicity	6,275	0.85	13.55	0.59	0.52	0.74	0.34	0.22
	Asian	668	0.93	9.39	0.77	0.74	0.83	0.55	0.50
	Native Hawaiian or Other Pacific Islander	505	0.82	13.80	0.60	0.50	0.68	0.33	0.08
	Black or African American	470	0.82	14.81	0.54	0.50	0.69	0.38	0.16
	American Indian or Alaska Native	264	0.87	13.09	0.63	0.52	0.77	0.33	0.31
	White	29,317	0.90	9.99	0.70	0.66	0.80	0.46	0.37
	Other	1,074	0.90	10.71	0.69	0.66	0.80	0.45	0.35
Limited English Proficiency	No	36,917	0.90	10.29	0.70	0.66	0.80	0.46	0.37
	Yes	1,656	0.74	18.56	0.44	0.36	0.56	0.28	0.06
Economic Disadvantage	No	29,678	0.90	10.03	0.70	0.67	0.80	0.47	0.38
	Yes	8,895	0.88	12.52	0.64	0.58	0.77	0.37	0.30
Special Education	No	35,192	0.90	9.98	0.69	0.66	0.80	0.46	0.37
	Yes	3,381	0.77	17.48	0.47	0.39	0.61	0.22	0.19

D-7. Science Grade 9 Form 1 Test Reliability

	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	14,062	0.79	14.80	0.39	0.45	0.73	0.41
Gender	Female	6,695	0.77	14.79	0.33	0.41	0.72	0.33
	Male	7,366	0.81	14.81	0.43	0.48	0.73	0.45
Ethnicity	Hispanic or Latino Ethnicity	2,381	0.71	16.95	0.31	0.35	0.64	0.22
	Asian	256	0.80	14.77	0.31	0.52	0.72	0.51
	Native Hawaiian or Other Pacific Islander	201	0.68	17.74	0.33	0.49	0.49	0.31
	Black or African American	175	0.64	19.16	0.20	0.25	0.55	0.14
	American Indian or Alaska Native	112	0.69	17.62	0.35	0.39	0.61	0.17
	White	10,553	0.79	14.32	0.38	0.44	0.72	0.41
	Other	384	0.78	14.55	0.43	0.45	0.74	0.40
Limited English Proficiency	No	13,325	0.79	14.61	0.38	0.44	0.72	0.41
	Yes	737	0.49	20.53	0.16	0.30	0.35	0.14
Economic Disadvantage	No	10,571	0.79	14.40	0.37	0.45	0.73	0.42
	Yes	3,491	0.76	16.09	0.39	0.40	0.67	0.31
Special Education	No	12,741	0.79	14.46	0.38	0.44	0.72	0.41
	Yes	1,321	0.66	18.21	0.32	0.31	0.52	0.21

D-7. Science Grade 9 Form 2 Test Reliability

	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	14,103	0.78	15.08	0.14	0.34	0.72	0.55
Gender	Female	6,817	0.75	15.20	0.10	0.30	0.71	0.46
	Male	7,285	0.81	14.97	0.18	0.38	0.73	0.61
Ethnicity	Hispanic or Latino Ethnicity	2,361	0.70	17.21	0.12	0.25	0.63	0.40
	Asian	219	0.79	14.52	0.21	0.40	0.73	0.61
	Native Hawaiian or Other Pacific Islander	194	0.57	17.67	0.11	0.21	0.55	0.16
	Black or African American	196	0.60	19.87	0.05	0.27	0.52	0.28
	American Indian or Alaska Native	100	0.69	17.26	-0.06	0.19	0.60	0.50
	White	10,646	0.78	14.62	0.14	0.34	0.72	0.56
	Other	387	0.78	14.74	0.09	0.31	0.71	0.52
Limited English Proficiency	No	13,385	0.78	14.87	0.14	0.33	0.72	0.55
	Yes	718	0.49	20.41	0.07	0.22	0.33	0.25
Economic Disadvantage	No	10,493	0.78	14.66	0.14	0.35	0.72	0.56
	Yes	3,610	0.75	16.32	0.13	0.29	0.67	0.48
Special Education	No	12,765	0.78	14.83	0.14	0.33	0.72	0.55
	Yes	1,338	0.66	17.99	0.13	0.28	0.51	0.42

D-7. Science Grade 9 Form 3 Test Reliability

	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	14,290	0.83	12.82	0.42	0.47	0.72	0.59
Gender	Female	6,752	0.81	12.69	0.40	0.44	0.70	0.51
	Male	7,536	0.85	12.87	0.44	0.49	0.74	0.63
Ethnicity	Hispanic or Latino Ethnicity	2,428	0.77	14.31	0.37	0.37	0.65	0.44
	Asian	237	0.84	12.52	0.42	0.38	0.75	0.67
	Native Hawaiian or Other Pacific Islander	195	0.75	14.25	0.26	0.35	0.58	0.44
	Black or African American	157	0.79	15.88	0.46	0.37	0.61	0.48
	American Indian or Alaska Native	113	0.74	14.96	0.21	0.42	0.52	0.39
	White	10,752	0.83	12.46	0.41	0.46	0.72	0.59
	Other	408	0.82	12.90	0.47	0.52	0.70	0.54
Limited English Proficiency	No	13,531	0.83	12.66	0.41	0.46	0.72	0.59
	Yes	759	0.54	17.97	0.18	0.26	0.36	0.15
Economic Disadvantage	No	10,683	0.83	12.46	0.41	0.46	0.72	0.59
	Yes	3,607	0.80	13.95	0.39	0.42	0.69	0.51
Special Education	No	12,922	0.83	12.46	0.41	0.46	0.72	0.59
	Yes	1,368	0.64	17.21	0.24	0.26	0.48	0.27

D-7. Science Grade 10 Form 1 Test Reliability

	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models[a]	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	7,882	0.78	17.45	0.67	—	0.51	0.54
Gender	Female	3,780	0.74	17.84	0.63	—	0.47	0.48
	Male	4,102	0.81	17.16	0.71	—	0.55	0.58
Ethnicity	Hispanic or Latino Ethnicity	1,285	0.68	19.60	0.56	—	0.44	0.38
	Asian	140	0.82	17.76	0.73	—	0.47	0.58
	Native Hawaiian or Other Pacific Islander	100	0.70	21.04	0.61	—	0.32	0.46
	Black or African American	97	0.60	21.73	0.48	—	0.45	0.24
	American Indian or Alaska Native	57	0.69	22.74	0.56	—	0.33	0.42
	White	5,977	0.78	16.98	0.68	—	0.52	0.55
	Other	226	0.79	19.49	0.67	—	0.51	0.57
Limited English Proficiency	No	7,582	0.78	17.35	0.68	—	0.51	0.54
	Yes	300	0.46	24.12	0.22	—	0.13	0.17
Economic Disadvantage	No	6,031	0.79	16.98	0.68	—	0.51	0.55
	Yes	1,851	0.72	19.40	0.60	—	0.49	0.45
Special Education	No	7,201	0.78	17.07	0.67	—	0.51	0.54
	Yes	681	0.57	23.53	0.44	—	0.31	0.22

[a] Too few items to calculate Cronbach's alpha

D-7. Science Grade 10 Form 2 Test Reliability

	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models[b]	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	7,761	0.78	16.00	0.68	—	0.49	0.54
Gender	Female	3,686	0.75	16.26	0.65	—	0.47	0.49
	Male	4,073	0.80	15.80	0.71	—	0.50	0.57
Ethnicity	Hispanic or Latino Ethnicity	1,247	0.64	19.74	0.49	—	0.41	0.33
	Asian	126	0.84	15.80	0.76	—	0.57	0.66
	Native Hawaiian or Other Pacific Islander	100	0.59	22.02	0.45	—	0.50	0.28
	Black or African American	97	0.61	18.65	0.45	—	0.11	0.26
	American Indian or Alaska Native	58	0.65	17.31	0.46	—	0.50	0.39
	White	5,909	0.79	15.36	0.69	—	0.49	0.56
	Other	224	0.74	15.95	0.64	—	0.31	0.51
Limited English Proficiency	No	7,469	0.78	15.86	0.68	—	0.48	0.54
	Yes	292	0.34	23.57	0.11	—	0.31	0.12
Economic Disadvantage	No	5,996	0.78	15.63	0.69	—	0.49	0.55
	Yes	1,765	0.73	17.59	0.61	—	0.43	0.45
Special Education	No	7,113	0.78	15.74	0.69	—	0.48	0.54
	Yes	648	0.51	21.07	0.37	—	0.20	0.31

[b] Too few items to calculate Cronbach's alpha

D-7. Science Grade 10 Form 3 Test Reliability

	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models[c]	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	7,643	0.78	15.78	0.67	—	0.50	0.53
Gender	Female	3,667	0.74	16.09	0.63	—	0.48	0.48
	Male	3,976	0.80	15.56	0.71	—	0.53	0.56
Ethnicity	Hispanic or Latino Ethnicity	1,270	0.68	17.96	0.54	—	0.40	0.42
	Asian	134	0.82	14.05	0.75	—	0.60	0.54
	Native Hawaiian or Other Pacific Islander	95	0.45	21.56	0.35	—	0.51	0.28
	Black or African American	89	0.69	17.38	0.60	—	0.40	0.44
	American Indian or Alaska Native	45	0.64	23.08	0.47	—	0.46	0.48
	White	5,795	0.78	15.34	0.68	—	0.51	0.54
	Other	215	0.78	16.88	0.67	—	0.53	0.53
Limited English Proficiency	No	7,320	0.78	15.60	0.68	—	0.50	0.53
	Yes	323	0.29	24.65	0.14	—	0.28	0.23
Economic Disadvantage	No	5,887	0.78	15.38	0.68	—	0.51	0.54
	Yes	1,756	0.72	17.37	0.60	—	0.44	0.46
Special Education	No	6,963	0.78	15.50	0.67	—	0.50	0.53
	Yes	680	0.59	19.91	0.48	—	0.39	0.35

[c] Too few items to calculate Cronbach's alpha

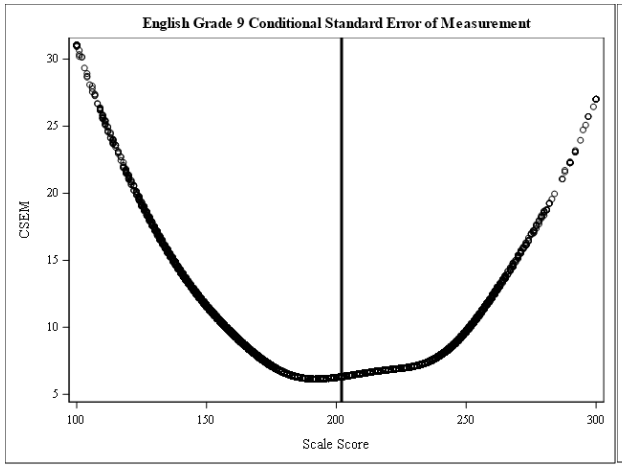
D-7. Science Grade 10 Form 4 Test Reliability

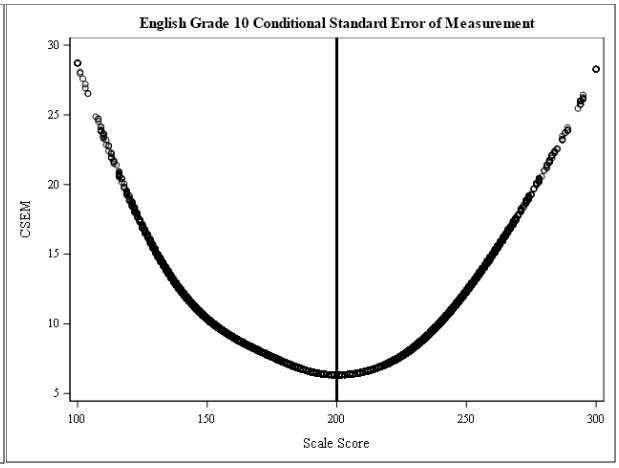
	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	7,810	0.82	13.65	0.73	0.35	0.49	0.57
Gender	Female	3,845	0.80	13.52	0.71	0.28	0.47	0.53
	Male	3,963	0.84	13.70	0.75	0.41	0.51	0.60
Ethnicity	Hispanic or Latino Ethnicity	1,288	0.75	15.48	0.64	0.17	0.42	0.44
	Asian	138	0.86	13.50	0.81	0.33	0.53	0.62
	Native Hawaiian or Other Pacific Islander	108	0.72	15.83	0.55	0.20	0.36	0.55
	Black or African American	106	0.72	15.18	0.67	0.12	0.35	0.47
	American Indian or Alaska Native	48	0.71	17.49	0.55	-0.01	0.24	0.35
	White	5,889	0.82	13.30	0.73	0.37	0.50	0.57
	Other	233	0.81	14.30	0.73	0.19	0.50	0.58
Limited English Proficiency	No	7,468	0.82	13.53	0.73	0.35	0.49	0.57
	Yes	342	0.35	21.40	0.28	0.03	0.20	0.20
Economic Disadvantage	No	5,981	0.82	13.36	0.73	0.36	0.49	0.57
	Yes	1,829	0.79	14.84	0.69	0.26	0.45	0.50
Special Education	No	7,122	0.82	13.30	0.73	0.35	0.49	0.57
	Yes	7,810	0.82	13.65	0.73	0.35	0.49	0.57

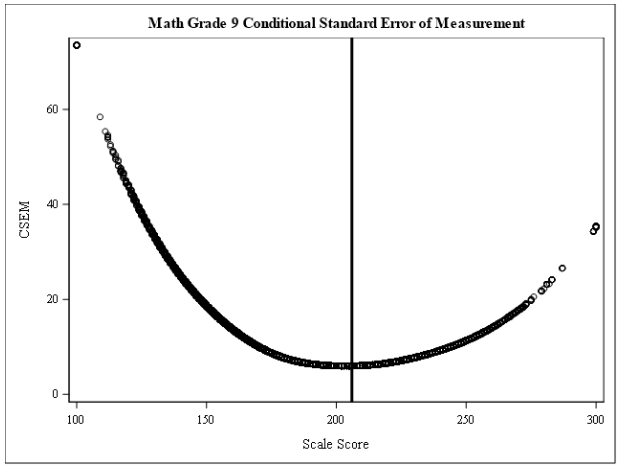
D-7. Science Grade 10 Form 5 Test Reliability

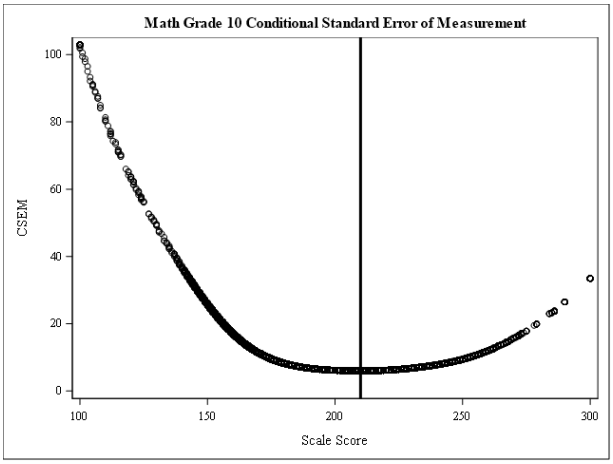
	Test Group	N	Alpha	SEM	Construct Explanations	Developing Models	Gathering and Investigating	Using Mathematical Thinking
All	Students Tested	7,806	0.84	11.58	0.66	0.47	0.51	0.68
Gender	Female	3,751	0.81	11.35	0.62	0.41	0.49	0.64
	Male	4,053	0.86	11.72	0.70	0.52	0.52	0.71
Ethnicity	Hispanic or Latino Ethnicity	1,278	0.77	12.64	0.55	0.36	0.40	0.59
	Asian	137	0.86	11.78	0.72	0.51	0.60	0.72
	Native Hawaiian or Other Pacific Islander	99	0.70	12.67	0.46	0.11	0.29	0.53
	Black or African American	91	0.75	15.34	0.63	0.27	0.47	0.51
	American Indian or Alaska Native	59	0.80	12.14	0.66	0.44	0.60	0.51
	White	5,946	0.84	11.34	0.66	0.47	0.51	0.68
	Other	196	0.86	11.97	0.67	0.56	0.63	0.73
Limited English Proficiency	No	7,505	0.84	11.47	0.66	0.47	0.51	0.68
	Yes	301	0.53	16.04	0.10	0.24	0.19	0.44
Economic Disadvantage	No	5,972	0.84	11.32	0.67	0.48	0.51	0.68
	Yes	1,834	0.80	12.60	0.59	0.35	0.47	0.64
Special Education	No	7,140	0.84	11.32	0.66	0.47	0.50	0.68
	Yes	666	0.66	14.96	0.42	0.26	0.30	0.44

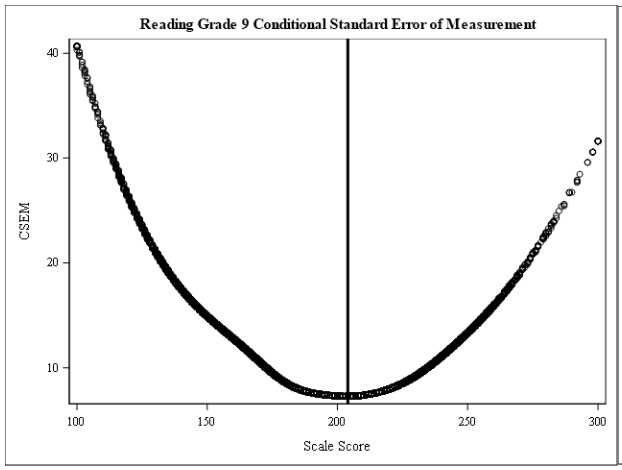
Appendix E: Conditional Standard Error of Scale Scores

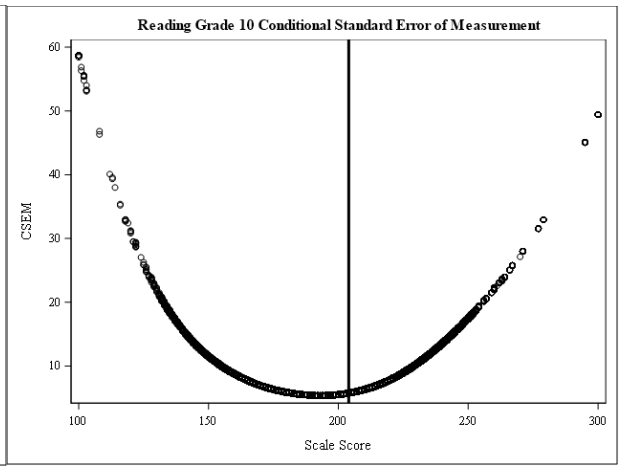


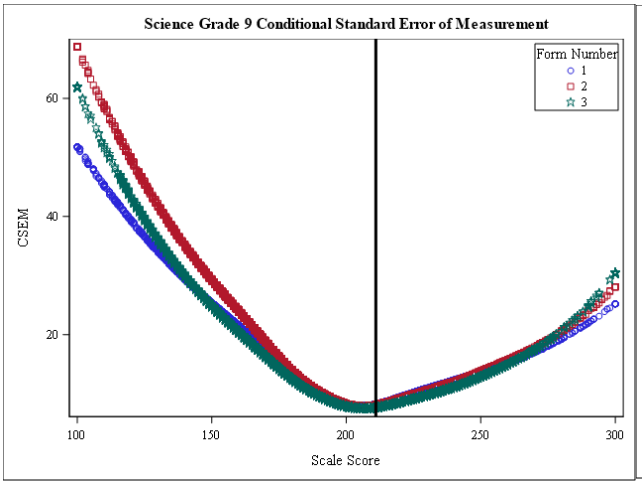


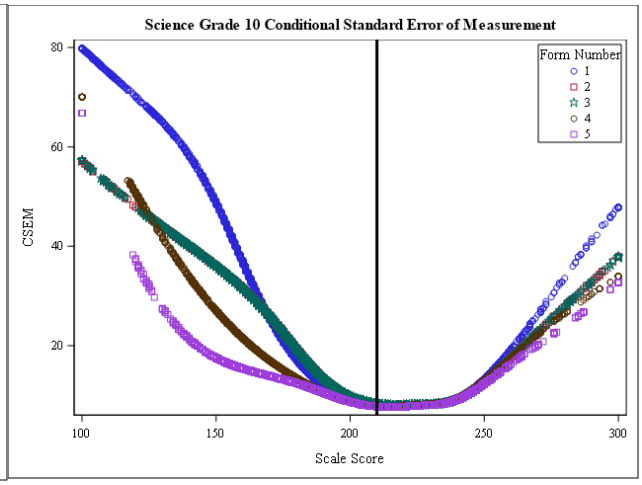









Science Grade 9 Conditional Standard Error of Measurement
Form Number
1
2
3
CSEM
60
40
20
100
150
200
250
300
Scale Score


Science Grade 10 Conditional Standard Error of Measurement
Form Number
1
2
3
4
5
CSEM
80
60
40
20
100
150
200
250
300
Scale Score

Appendix F: Accuracy and Consistency

F-1. Accuracy Classification for English Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.178	0.030	0.002	0.000	79.10
Approaching Proficient	0.048	0.105	0.042	0.002	
Proficient	0.004	0.053	0.139	0.049	
Highly Proficient	0.000	0.002	0.050	0.300	

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

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.178	0.030	0.002	0.000	89.66
Approaching Proficient	0.048	0.105	0.042	0.002	
Proficient	0.004	0.053	0.139	0.049	
Highly Proficient	0.000	0.002	0.050	0.300	

F-3. Consistency Classification for English Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.172	0.047	0.010	0.001	63.70	0.506
Approaching Proficient	0.047	0.080	0.053	0.009		
Proficient	0.010	0.053	0.107	0.062		
Highly Proficient	0.001	0.009	0.062	0.279		

F-4. Accuracy Classification for English Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.161	0.022	0.001	0.000	73.14
Approaching Proficient	0.049	0.113	0.030	0.000	
Proficient	0.003	0.063	0.170	0.030	
Highly Proficient	0.000	0.001	0.70	0.285	

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

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.161	0.022	0.001	0.000	90.13
Approaching Proficient	0.049	0.113	0.030	0.000	
Proficient	0.003	0.063	0.170	0.030	
Highly Proficient	0.000	0.001	0.70	0.285	

F-6. Consistency Classification for English Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.161	0.045	0.007	0.000	65.23	0.532
Approaching Proficient	0.045	0.095	0.056	0.004		
Proficient	0.007	0.056	0.147	0.061		
Highly Proficient	0.000	0.004	0.061	0.250		

F-7. Accuracy Classification for Math Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.203	0.018	0.000	0.000	73.20
Approaching Proficient	0.056	0.104	0.020	0.001	
Proficient	0.004	0.065	0.079	0.017	
Highly Proficient	0.000	0.010	0.079	0.344	

F-8. Accuracy Classification at Proficient Cut Point for Math Grade 9

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.203	0.018	0.000	0.000	90.26
Approaching Proficient	0.056	0.104	0.020	0.001	
Proficient	0.004	0.065	0.079	0.017	
Highly Proficient	0.000	0.010	0.079	0.344	

F-9. Consistency Classification for Math Grade 9

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.211	0.046	0.006	0.000	68.03	0.561
Approaching Proficient	0.046	0.095	0.047	0.010		
Proficient	0.006	0.047	0.074	0.051		
Highly Proficient	0.000	0.010	0.051	0.300		

F-10. Accuracy Classification for Math Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.270	0.022	0.001	0.000	73.83
Approaching Proficient	0.056	0.083	0.022	0.001	
Proficient	0.006	0.057	0.074	0.020	
Highly Proficient	0.000	0.009	0.067	0.311	

F-11. Accuracy Classification at Proficient Cut Point for Math Grade 10

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.270	0.022	0.001	0.000	90.29
Approaching Proficient	0.056	0.083	0.022	0.001	
Proficient	0.006	0.057	0.074	0.020	
Highly Proficient	0.000	0.009	0.067	0.311	

F-12. Consistency Classification for Math Grade 10

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.275	0.047	0.009	0.001	68.47	0.564
Approaching Proficient	0.047	0.071	0.042	0.011		
Proficient	0.009	0.042	0.065	0.048		
Highly Proficient	0.001	0.011	0.048	0.273		

F-13. Accuracy Classification for Science Grade 9 Form 1

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.248	0.035	0.013	0.002	65.43
Approaching Proficient	0.056	0.044	0.033	0.014	
Proficient	0.024	0.041	0.057	0.052	
Highly Proficient	0.004	0.016	0.056	0.305	

F-14. Accuracy Classification at Proficient Cut Point for Science Grade 9 Form 1

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.248	0.035	0.013	0.002	85.37
Approaching Proficient	0.056	0.044	0.033	0.014	
Proficient	0.024	0.041	0.057	0.052	
Highly Proficient	0.004	0.016	0.056	0.305	

F-15. Consistency Classification for Science Grade 9 Form 1

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.240	0.048	0.030	0.015	58.45	0.412
Approaching Proficient	0.048	0.031	0.030	0.026		
Proficient	0.030	0.030	0.040	0.058		
Highly Proficient	0.015	0.026	0.058	0.274		

F-16. Accuracy Classification for Science Grade 9 Form 2

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.243	0.034	0.013	0.003	66.64
Approaching Proficient	0.053	0.042	0.032	0.014	
Proficient	0.022	0.038	0.055	0.053	
Highly Proficient	0.004	0.015	0.052	0.327	

F-17. Accuracy Classification at Proficient Cut Point for Science Grade 9 Form 2

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.243	0.034	0.013	0.003	85.84
Approaching Proficient	0.053	0.042	0.032	0.014	
Proficient	0.022	0.038	0.055	0.053	
Highly Proficient	0.004	0.015	0.052	0.327	

F-18. Consistency Classification for Science Grade 9 Form 2

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.232	0.046	0.029	0.015	59.84	0.426
Approaching Proficient	0.046	0.029	0.029	0.026		
Proficient	0.029	0.029	0.038	0.057		
Highly Proficient	0.015	0.023	0.057	0.299		

F-19. Accuracy Classification for Science Grade 9 Form 3

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.266	0.031	0.009	0.001	68.52
Approaching Proficient	0.054	0.047	0.030	0.008	
Proficient	0.019	0.044	0.061	0.041	
Highly Proficient	0.002	0.015	0.059	0.311	

F-20. Accuracy Classification at Proficient Cut Point for Science Grade 9 Form 3

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.266	0.031	0.009	0.001	87.09
Approaching Proficient	0.054	0.047	0.030	0.008	
Proficient	0.019	0.044	0.061	0.041	
Highly Proficient	0.002	0.015	0.059	0.311	

F-21. Consistency Classification for Science Grade 9 Form 3

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.267	0.047	0.025	0.009	61.81	0.461
Approaching Proficient	0.047	0.036	0.033	0.022		
Proficient	0.025	0.033	0.046	0.056		
Highly Proficient	0.009	0.022	0.056	0.275		

F-22. Accuracy Classification for Science Grade 10 Form 1

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.257	0.037	0.017	0.003	64.26
Approaching Proficient	0.056	0.043	0.041	0.014	
Proficient	0.023	0.040	0.073	0.063	
Highly Proficient	0.003	0.011	0.051	0.269	

F-23. Accuracy Classification at Proficient Cut Point for Science Grade 10 Form 1

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.257	0.037	0.017	0.003	84.89
Approaching Proficient	0.056	0.043	0.041	0.014	
Proficient	0.023	0.040	0.073	0.063	
Highly Proficient	0.003	0.011	0.051	0.269	

F-24. Consistency Classification for Science Grade 10 Form 1

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.243	0.047	0.035	0.014	57.16	0.399
Approaching Proficient	0.047	0.028	0.032	0.023		
Proficient	0.035	0.032	0.052	0.062		
Highly Proficient	0.014	0.023	0.062	0.249		

F-25. Accuracy Classification for Science Grade 10 Form 2

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.238	0.0322	0.013	0.001	65.31
Approaching Proficient	0.055	0.041	0.035	0.011	
Proficient	0.025	0.041	0.072	0.053	
Highly Proficient	0.003	0.013	0.063	0.302	

F-26. Accuracy Classification at Proficient Cut Point for Science Grade 10 Form 2

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.238	0.0322	0.013	0.001	85.65
Approaching Proficient	0.055	0.041	0.035	0.011	
Proficient	0.025	0.041	0.072	0.053	
Highly Proficient	0.003	0.013	0.063	0.302	

F-27. Consistency Classification for Science Grade 10 Form 2

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.231	0.045	0.032	0.013	58.13	0.411
Approaching Proficient	0.045	0.028	0.032	0.022		
Proficient	0.032	0.032	0.054	0.065		
Highly Proficient	0.013	0.022	0.065	0.268		

F-28. Accuracy Classification for Science Grade 10 Form 3

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.235	0.034	0.014	0.002	63.97
Approaching Proficient	0.058	0.043	0.038	0.012	
Proficient	0.026	0.043	0.075	0.058	
Highly Proficient	0.003	0.013	0.286	0.363	

F-29. Accuracy Classification at Proficient Cut Point for Science Grade 10 Form 3

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.235	0.034	0.014	0.002	84.95
Approaching Proficient	0.058	0.043	0.038	0.012	
Proficient	0.026	0.043	0.075	0.058	
Highly Proficient	0.003	0.013	0.286	0.363	

F-30. Consistency Classification for Science Grade 10 Form 3

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.228	0.047	0.034	0.014	56.66	0.394
Approaching Proficient	0.047	0.029	0.033	0.023		
Proficient	0.034	0.033	0.055	0.066		
Highly Proficient	0.014	0.023	0.066	0.255		

F-31. Accuracy Classification for Science Grade 10 Form 4

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.234	0.030	0.008	0.000	66.82
Approaching Proficient	0.058	0.053	0.037	0.005	
Proficient	0.020	0.051	0.093	0.045	
Highly Proficient	0.001	0.010	0.067	0.288	

F-32. Accuracy Classification at Proficient Cut Point for Science Grade 10 Form 4

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.234	0.030	0.008	0.000	86.79
Approaching Proficient	0.058	0.053	0.037	0.005	
Proficient	0.020	0.051	0.093	0.045	
Highly Proficient	0.001	0.010	0.067	0.288	

F-33. Consistency Classification for Science Grade 10 Form 4

First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.232	0.048	0.027	0.006	59.49	0.441
Approaching Proficient	0.048	0.038	0.041	0.017		
Proficient	0.027	0.041	0.073	0.064		
Highly Proficient	0.006	0.017	0.064	0.252		

F-34. Accuracy Classification for Science Grade 10 Form 5

True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.248	0.028	0.005	0.000	66.74
Approaching Proficient	0.064	0.060	0.033	0.003	
Proficient	0.020	0.058	0.102	0.034	
Highly Proficient	0.001	0.010	0.077	0.258	

F-35. Accuracy Classification at Proficient Cut Point for Science Grade 10 Form 5

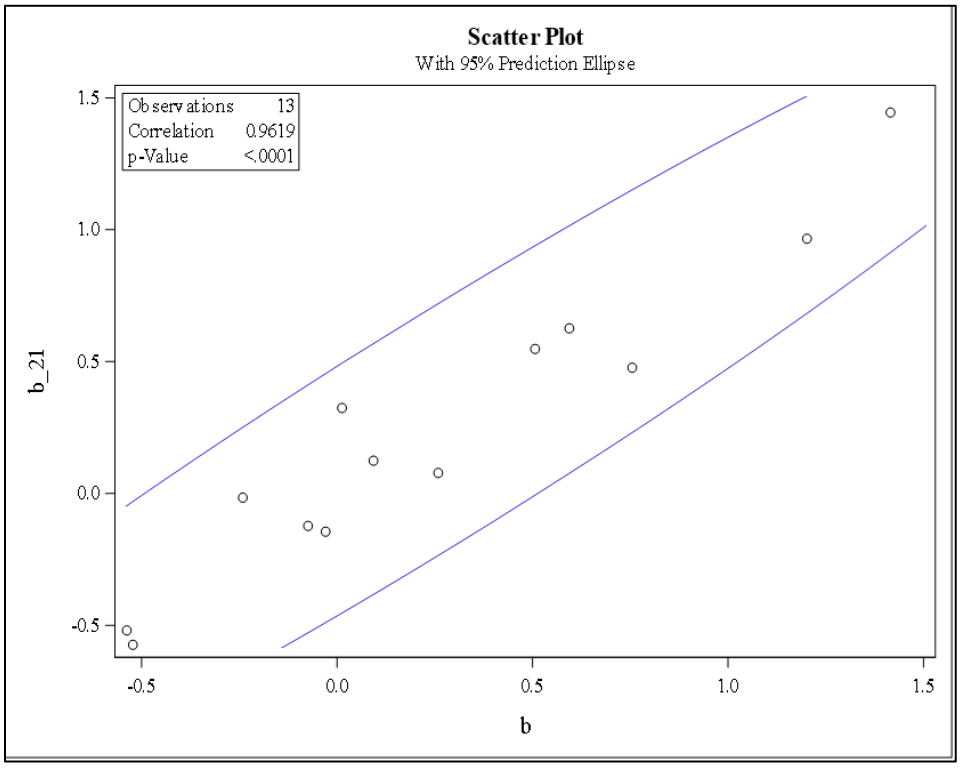
True Score	Observed Score				Accuracy %
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient	
Below Proficient	0.248	0.028	0.005	0.000	86.96
Approaching Proficient	0.064	0.060	0.033	0.003	
Proficient	0.020	0.058	0.102	0.034	
Highly Proficient	0.001	0.010	0.077	0.258	

F-36. Consistency Classification for Science Grade 10 Form 5

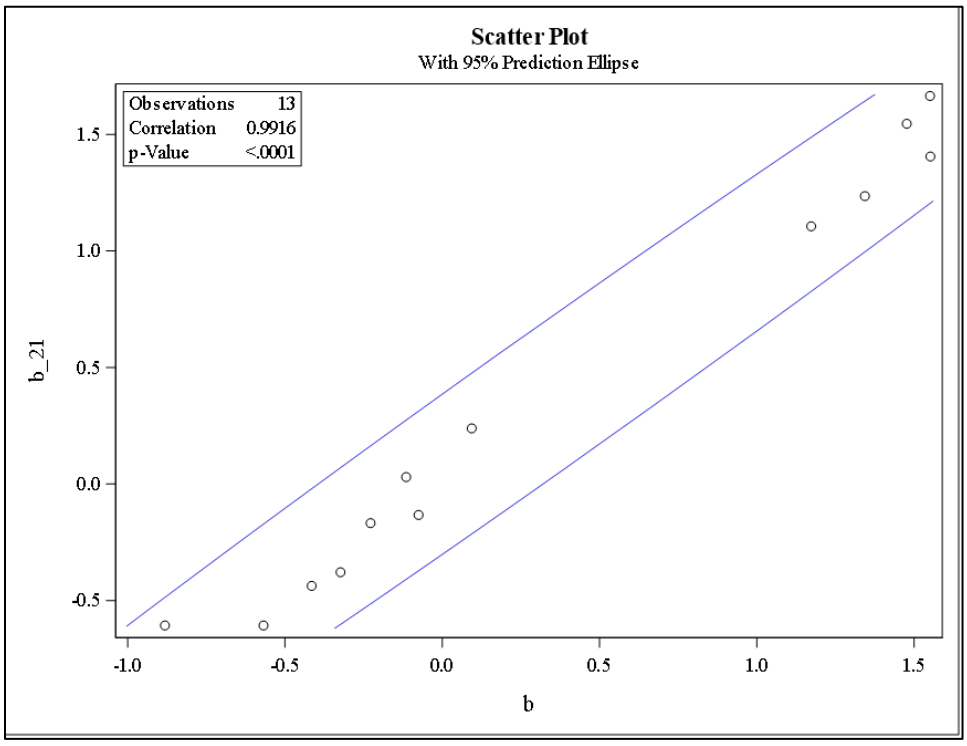
First Form	Alternate Form				Consistency %	Kappa
	Below Proficient	Approaching Proficient	Proficient	Highly Proficient		
Below Proficient	0.253	0.051	0.025	0.004	59.97	0.453
Approaching Proficient	0.051	0.046	0.045	0.014		
Proficient	0.025	0.045	0.085	0.062		
Highly Proficient	0.004	0.014	0.062	0.216		

Appendix G: Common Item Scatterplots for 2021 Anchor Items

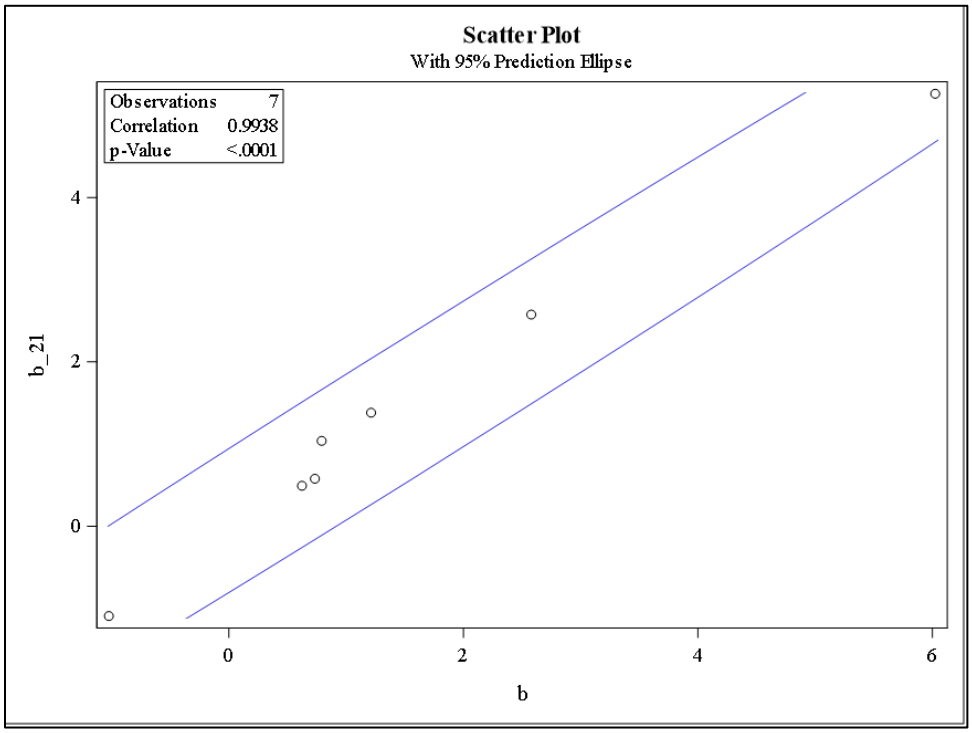
G-1. Scatterplot of Anchor Items for Math Grade 9



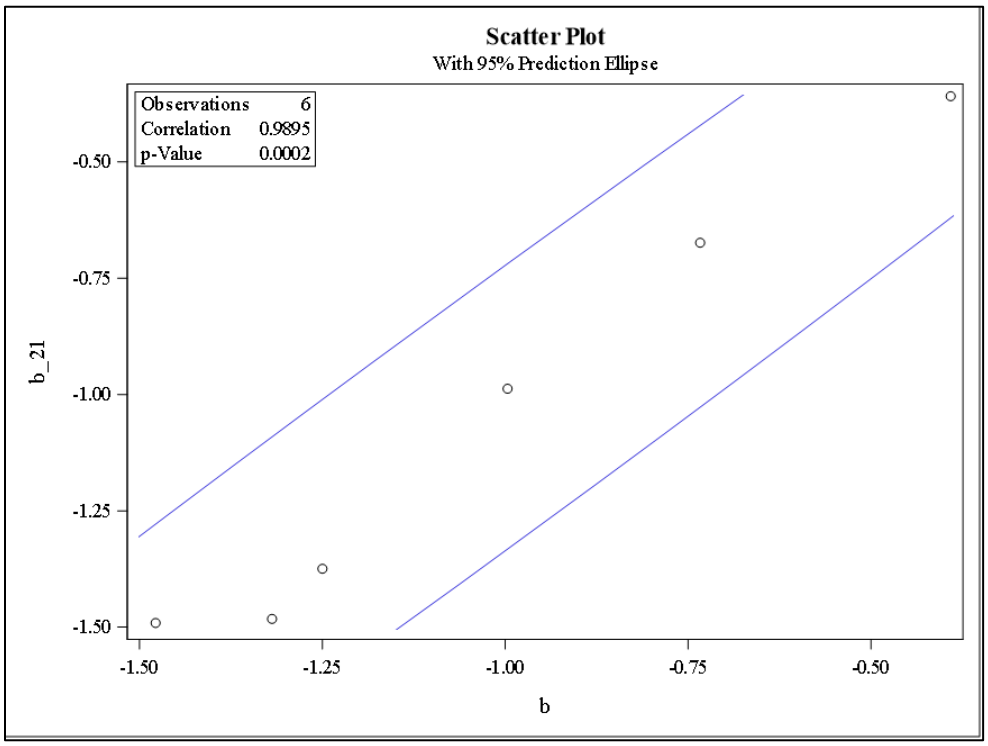
G-2. Scatterplot of Anchor Items for Math Grade 10



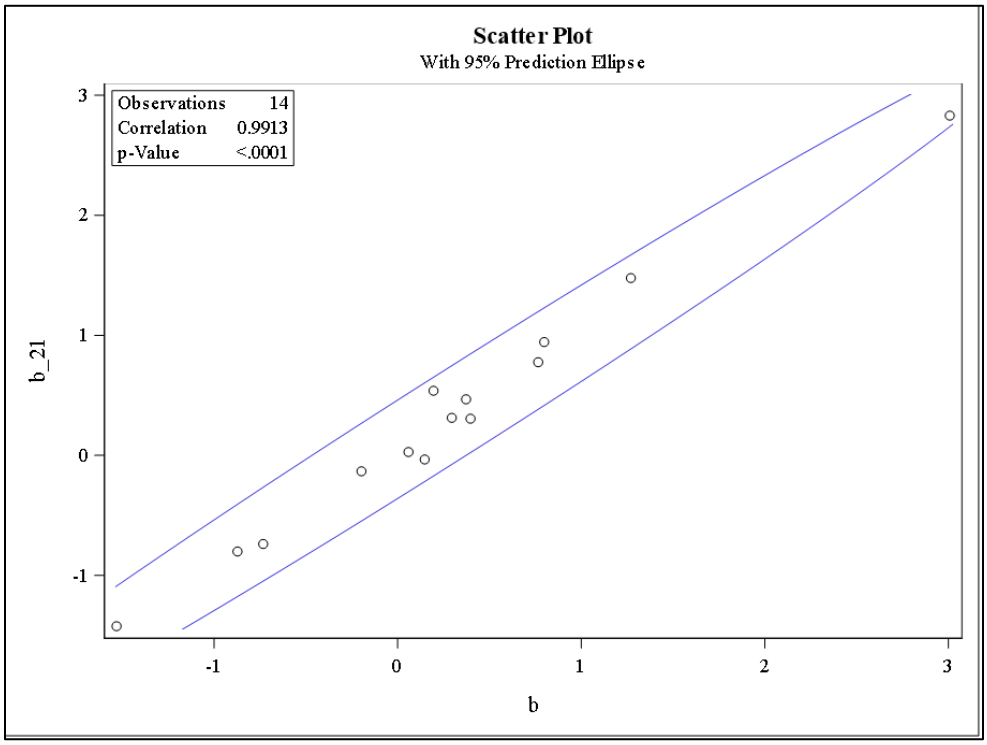
G-3. Scatterplot of Anchor Items for Reading Grade 9



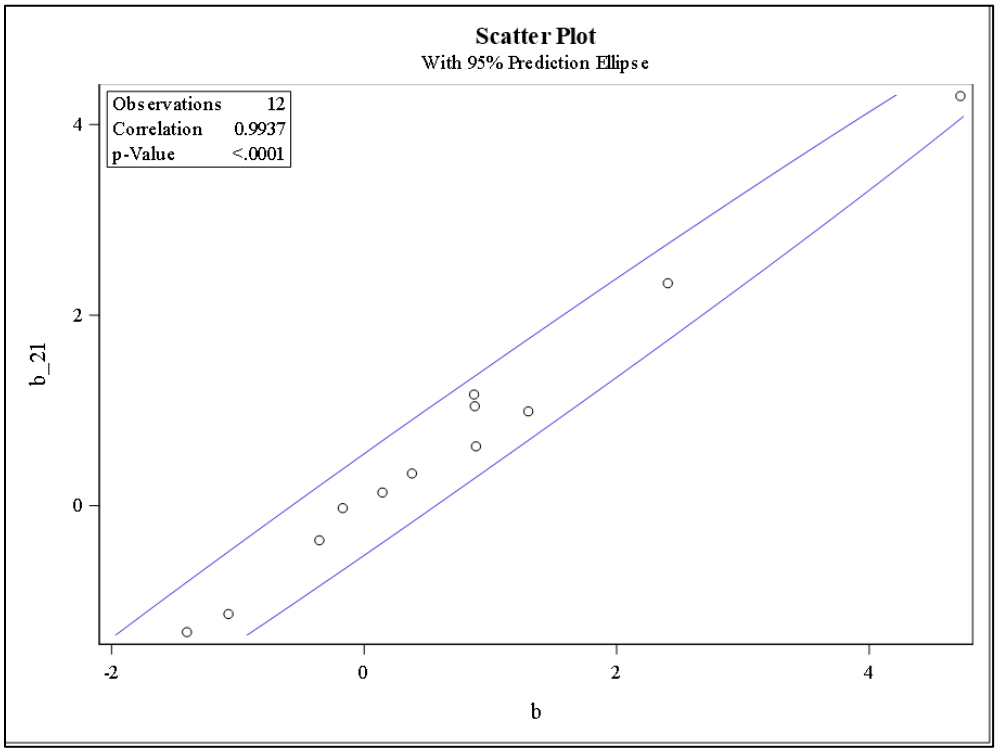
G-4. Scatterplot of Anchor Items for Reading Grade 10



G-5. Scatterplot of Anchor Items for English Grade 9



G-6. Scatterplot of Anchor Items for English Grade 10



Appendix H: SEEds Performance Level Descriptor Educator Committee Training

Utah Aspire Plus Assessments for Science Grades 9–10

Educator Review of Performance Level Descriptors

April 6, 2021

1

What to Expect?

1. Welcome
2. Housekeeping
3. Schedule
4. Training
5. Review and evaluate PLDs
6. Approve verbiage or recommend edits
7. Closing

2

Introductions

Pearson Development Team
Christopher Altermatt, Assessment Specialist- Science

Steven "Brice" Howell, Content Specialist- Science
Katie Mohasci, Senior Test Development Manager

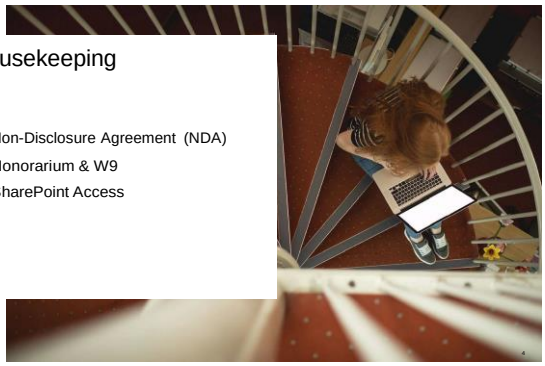
USBE
Scott Roskelley, Science Assessment Specialist
Cydnee Carter, Assessment Development Coordinator

Educator Introductions
Introductions of Educators will occur following our break

3

Housekeeping

1. Non-Disclosure Agreement (NDA)
2. Honorarium & W9
3. SharePoint Access



4

Today's Schedule

8:30 a.m. –9:15 a.m.	Welcome and Introductions Housekeeping Schedule Overview of Utah Aspire Plus Assessments General Training
9:15 a.m. 10:15 a.m.	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
12:00 p.m. –12:45 p.m.	Lunch Break
12:45 p.m. –2:00 p.m.	PLDs Review
2:00 p.m. –2:15 p.m.	Break
2:30 p.m. –4:15 p.m.	PLDs Review
4:15 p.m. –4:30 p.m.	Wrap up and Adjourn

5

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 ~~-~~1
- Simplifies training and technology support for high schools

6

Appendix I: Utah Aspire Plus 2121 Science Standard Setting Executive Summary

Utah Aspire Plus Science
Summer 2021 Standard Setting Meeting
Executive Summary

August 2021

This report summarizes the process and results of setting performance levels for the Utah Aspire Plus Science assessment for Grades 9 and 10. The Utah State Board of Education (USBE) and Pearson (assessment contractor) recommend the performance levels shown in Table 3 of this report.

Utah Aspire Plus Standard Setting Process and Results

Performance levels are used to classify and describe student performance on an assessment. In order to classify student performance into the different performance levels, the following components are generally required: 1) Policy Level Performance Level Descriptors, 2) Range Performance Level Descriptors (PLDs), and 3) cut scores. Policy level performance level descriptors provide descriptions of what students at each performance level know and what they are able to do. Range PLDs illustrate the performance levels in terms that are specific to a grade and subject. Cut scores represent the lowest boundary of each performance level on the scale.

The process of recommending performance standards for the Utah Aspire Plus science assessments was in line with national best practice for standard setting. Results and details of the process are presented in the following sections.

Policy Definitions

Policy Level Performance Level Descriptors for the Utah Aspire Plus assessments are shown in Table 1. The titles and descriptions of the performance levels were defined to be part of a cohesive assessment system.

Table 1. Policy level descriptors for Utah Aspire Plus Science

Below Proficient	Approaching Proficient	Proficient	Highly Proficient
The Level 1 students are below proficient in achieving or applying the science attitudes and knowledge/ skills as specified in the Utah Core Standards. The students generally perform significantly below the standard for their grade level, are able to engage with higher-order thinking skills for all science contexts with extensive support.	The Level 2 students are approaching proficient in achieving or applying the science attitudes and knowledge/ skills as specified in the Utah Core Standards. The students generally perform slightly below the standard for their grade level, are likely able to engage in higher-order thinking skills for all science contexts with support.	The Level 3 students are proficient in achieving or applying the science attitudes and knowledge/skills as specified in the Utah Core Standards. The students generally perform at the standard for their grade level, are able to engage in higher order thinking-skills for all science contexts with independence and minimal support. This level of science performance also likely indicates students are on track to be sufficiently prepared for college or career.	The Level 4 students are highly proficient in achieving or applying the science attitudes and knowledge/skills as specified in the Utah Core Standards. The students generally perform above the standard for their grade level, are able to engage in higher-order thinking skills involving all science contexts independently. This level of science performance also likely indicates students are on track to be well-prepared for college or career.

Performance Level Descriptors (PLDs)

The Utah State Board of Education (USBE) and Pearson (assessment contractor) drafted the Utah Aspire Plus Science Performance Level Descriptors (PLDs) for science grades 9 and 10 in March 2021. The new PLDs were written to correspond to the newly adopted Science standards. In April 2021, Utah educators reviewed the PLDs and recommended adjustments to adhere to the goals of the Utah Aspire Plus assessments. Pearson and USBE reconciled the recommendations of the Utah educators in finalizing the PLDs for Utah Aspire Plus.

General Method

From August 9 to August 12, 2021, after the first year of operational administration, a virtual standard setting workshop was conducted to provide cut score recommendations for the Utah Aspire Plus science assessments for grades 9 and 10. The participants, including teachers and non-teacher educators, were selected for the standard setting committee to provide content and grade-level expertise during the workshop and be representative of the state teaching population, including geographic region, gender, ethnicity, educational experience, community size, and community socioeconomic status.

The Extended Modified (Yes/No) Angoff standard setting method was used at the standard setting meeting (Davis & Moyer, 2015; Plake, Ferdous, Impara, & Buckendahl, 2005). This is a content- and item-based method that leads participants through a standardized process in which they consider expectations of student performance, as defined by the borderline performance level descriptions, and the individual items administered to students to recommend cut scores for each performance level. The standardized process was used by the science committee, which resulted in cut score recommendations.

The process started with participants experiencing an assessment through an online testing environment similar to the one used to administer the items to students. Participants then spent time drafting borderline descriptions that identify the knowledge and skills needed to 'just barely' be classified into a performance level. Based on their experience with the test items and a review of the borderline performance level descriptions, participants reviewed each item on the test and answered the following question for each performance level:

"How many points would a student performing at the borderline of the [specific] performance level likely earn if they answered the question?"

The cut score recommendation for each individual participant was the expected scale score a student performing at the borderline of the respective performance level would likely earn, calculated using their pattern of responses for the items based on the judgment question. For the purposes of the standard setting, "likely" was defined as 2 out of 3 students with performance at the borderline of the performance level. Each recommended cut score from the standard setting committee is the median of the recommendations from the individual participants in the committee. Each committee completed three rounds of judgments, with feedback data and panelist discussions between each round. The standard errors of judgment (SEJ) for the recommended scale score cuts based on Round 3 of the committee were calculated. The recommended Round 3 cut scores plus and minus 1 to 3 SEJs is shown in Table 3 of Appendix A.

Round 3 Committee Cut Scores

The Round 3 recommended cut scores by the teacher committee for science and STEM composite are shown in Table 2. This table shows the scale score ranges corresponding to each performance level. The reporting scale ranges from a lowest obtainable scale score (LOSS) of 100 to a highest obtainable scale score of (300). The cut scores for the performance levels are the lowest score within each range. The cut scores for the STEM composite were calculated by averaging the math cut scores established in 2019 and the round 3 science cut scores by performance level.

Table 2. Round 3 Cut Score Ranges for Utah Aspire Plus Performance Levels

		Performance Levels			
Subject	Grade	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
Science	9	100-186	187-210	211-236	237-300
	10	100-172	173-216	217-251	252-300
STEM	9	100-179	180-208	209-234	235-300
	10	100-176	177-213	214-243	244-300

Reasonableness Review

A reasonableness review was conducted by USBE and Pearson following the committee meeting. The review considered the final cut score recommendations of the committee, the standard errors of judgment, the committee evaluations regarding cut score confidence, and the ACT concordance data. Taking into consideration these elements, USBE took the cut scores shown in Table 3 to a cross-grade articulation meeting held on August 13, 2021. Cut scores that were modified from Round 3 are highlighted.

Table 3. Cut Score Ranges for Utah Aspire Plus Science Performance Levels After Reasonableness Review

Subject	Grade	Performance Levels			
		Below Proficient	Approaching Proficient	Proficient	Highly Proficient
Science	9	100-186	187-210	211-236	237-300
	10	100-186	187-209	210-239	240-300
STEM	9	100-179	180-208	209-234	235-300
	10	100-183	184-209	210-237	238-300

The cut scores for grade 9 were maintained from Round 3 as the committee indicated high levels of confidence in the cut scores. After Round 3, the committee indicated the percent of students classified into each performance level at grade 10 should more closely mirror the results from grade 9. The grade 10 scores for Advanced, Proficient and Approaching Proficient were adjusted accordingly and within +/- 2 SEJ of the Round 3 cut scores.

The upper and lower bounds of the predicted ACT concordance scores based on the final scale score cuts for Proficient from the reasonableness review are shown in Table 4. The ACT college and career benchmarks are also included in the table.

Table 4. Predicted ACT Concordance Scores for the Proficient Scale Score Cut

Grade	Proficient Scale Score Cut	Predicted ACT Score Lower Bound	Predicted ACT Score Upper Bound	ACT College and Career Benchmark
9	211	19	23	23
10	210	18	23	23

Results for Utah Aspire Plus Assessments

Table 5 shows the percentage of students who took the Utah Aspire Plus Science assessments during the Spring 2021 administration that would be classified into each performance level based on the recommended cut scores after the Reasonableness Review. The percentage of students in a performance level is not directly comparable across grades and subjects. The population of students tested is different for each assessment. Performance levels from different tests are not comparable because the cut scores for these tests are criterion-referenced—they are based on skill/content-specific expectations of what students should know and be able to do.

Table 5. Percentage of Students in Each Performance Level

Subject	Grade	Performance Levels			
		Below Proficient	Approaching Proficient	Proficient	Highly Proficient
Science	9	30.7%	33.3%	27.1%	8.9%
	10	31.2%	31.7%	30.1%	7.0%
STEM	9	25.5%	39.0%	28.4%	7.1%
	10	30.4%	36.2%	27.6%	5.9%

Cross-grade Articulation

Following the Reasonableness Review, USBE conducted a cross-grade articulation meeting including the review of cut scores from grade 4 through 10. The results of that meeting are shared in a separate document.

References

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

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

Appendix A

The standard errors of judgment (SEJ) for the recommended scale score cuts based on Round 3 of the committee were calculated. The recommended Round 3 cut plus and minus 1 to 3 SEJs is shown in Table 6.

Table 6. Standard Errors of Judgment based on Round 3 Recommended Cut Scores

Subject	Grade	Performance Level	-3 SEJ	-2 SEJ	-1 SEJ	Cut Score	+1 SEJ	+2 SEJ	+3 SEJ
Science	9	Approaching Proficient	152	164	175	187	199	210	222
		Proficient	201	204	207	211	214	217	220
		Highly Proficient	223	228	232	237	241	245	250
	10	Approaching Proficient	131	145	159	173	187	201	215
		Proficient	206	210	213	217	221	224	228
		Highly Proficient	240	244	248	252	256	260	264

Appendix J: Updating ACT Score Predictions for Utah Grade 9 Aspire Plus

May 2021

Jeff Allen, ACT

We document the data and procedures used to generate updated ACT score predictions for the Utah Aspire Plus 9th grade assessments. Included in this documentation are:

- A description of the methodological approach
- Descriptions of the samples used to generate the predictions
- Description of weighting procedure to ensure samples are representative of 9th grade population
- Description of updated predicted ACT score ranges
    - Comparison to previously derived predicted ACT score ranges
    - Accuracy statistics

The updated ACT score predictions can be used for reporting Utah Aspire Plus results for spring 2021 9th and 10th graders. However, because of the changes to the Utah Aspire Plus science test and its score scale, the concordance of the old and updated Utah Aspire Plus science scores must first be applied to report the predicted ACT science score ranges.

General description of methodological approach

The following steps were taken:

1) Match the spring 2019 grade 9 Utah Aspire Plus records to the spring 2021 grade 11 ACT test records. Student state ID was used to match the records.

2) Compare the matched sample to the spring 2019 grade 9 data to assess how representative the matched sample is to the target population (the spring grade 9 data is used as the target population). The matched sample is different than the target population because some students were lost to follow-up (e.g., moved out of state, were absent on test day, or did not take the ACT for some other reason).

3) Weight the matched sample to be representative of the target population with respect to 9th grade test scores, gender, limited English proficient status, special education status,

and race/ethnicity. Propensity scores, based on logistic regression models, are used to derive the weights.

4) Using the weighted data, use quantile regression to estimate the percentiles of ACT scores, conditional on Utah Aspire Plus scores. The SGP R package is used to obtain the quantile of each possible ACT score for each possible Utah Aspire Plus score. Quantile regression using the SGP package is preferred over linear regression because it does not impose assumptions of linearity or homoskedasticity.

5) Using the SGP model results, find the 25th and 75th percentiles of the ACT score distribution for each possible 9th grade test score. The 25th percentile serves as the lower bound of the predicted ACT score range and the 75th percentile serves as the upper bound of the predicted ACT score range. If ACT scores were reported on a continuous scale, this would produce predicted score ranges with 50% coverage. Because ACT scores are reported to the nearest integer, this produces predicted score ranges with approximately 60% coverage.

6) Adjust the predicted ACT score ranges to ensue that they do not decrease with Utah Aspire Plus score. This step is necessary because the SGP procedure may result in conditional percentiles that are not monotonically increasing (this only tends to happen for areas of the score distribution where the data are very sparse).

7) Assess the accuracy of the predicted ACT score ranges and compare the ranges to those that were estimated in 2019. Note that steps 2 through 6 are completed for each subject for which predicted ACT scores are estimated (English, math, reading, science, and composite).

Samples used to generate the predictions

Table 1 shows the number of students with reported scale scores, by subject and assessment. Table 1 also provides the number of students with scale scores on both tests (Matched Sample).

Table 1: Number of students tested and matched, by subject

Subject	Grade 9 Aspire Plus	Grade 11 ACT	Matched Sample
Composite	44,429	41,357	36,273
English	46,050	41,497	37,348
Mathematics	45,590	41,460	37,039
Reading	46,238	41,400	37,411
Science	46,149	41,369	37,338

Not all students who tested in 9th grade had a matching 11th grade ACT record. This is expected for students who migrated out of Utah after the 9th grade test, or who did not take the ACT test for any other reason. Similarly, not all ACT-tested students had matching 9th grade records. This is expected for students who migrated into Utah after the 9th grade test, or who did not take the 9th grade test for any other reason.

Because the matched samples are large and representative of the target population (described later), we do not expect that this missing data would have much impact on the updated predictions.

Because the predictions are reported to 9th grade students, we used the 9th grade data as the target population. In Table 2, we compare the matched sample to the target population on 9th grade test score quintile, gender, limited English proficient status, special education status, and race/ethnicity. Table 2 only reflects the composite score analysis, but the comparison is similar across the other subject areas.

Table 2: Comparing the matched sample to the target population (composite score)

Variable	Matched Sample	Target Population	Matched Sample, Weighted
Grade 9 Aspire Plus score, quintile			
1st	15.3%	20.0%	19.9%
2nd	19.5%	20.3%	20.3%
3rd	20.2%	19.4%	19.3%
4th	22.7%	20.7%	20.8%
5th	22.3%	19.6%	19.7%
Female	49.9%	49.2%	49.2%
Limited English proficient	3.9%	4.8%	4.9%
Special education	7.6%	9.5%	9.6%
Race/ethnicity			
African American	1.2%	1.3%	1.3%
Asian	1.8%	1.8%	1.8%
Hispanic	14.8%	16.8%	16.8%
Two or more races	2.6%	2.7%	2.7%
Other	2.2%	2.6%	2.6%
White	77.4%	74.8%	74.8%

Table 2 shows that students in the matched sample tend to have higher 9th grade test scores, are slightly less likely to have limited English proficient status, are slightly less likely to have special education status, and are slightly more likely to be White. Table 2 also shows the comparison after weighting the matched sample to be more like the target population on these characteristics.

After weighting, the matched sample is nearly identical to the target population. Later, we describe the method used to weight the matched sample.

Table 3 shows summary statistics for the 9th and 11th grade test scores for the matched samples and weighed matched samples. In addition to test score means and standard deviations, the correlations are also presented. The correlations range from 0.69 for science to 0.85 for composite.

Because the weighting procedure assigns larger weights to lower-achieving students, the mean test scores for the weighted matched sample are lower than those for the matched sample. Weighting also slightly increases the standard deviation of the 9th grade test scores but has very little impact on the correlations.

Table 3: Matched sample summary statistics

Sample	Subject	Grade 9 Aspire Plus		Grade 11 ACT		*r*
		Mean	SD	Mean	SD	
Matched Sample	Composite	203.15	24.10	19.79	5.14	0.85
	English	202.62	26.13	18.59	6.08	0.78
	Mathematics	202.83	26.40	19.44	5.07	0.77
	Reading	202.48	27.77	20.26	6.35	0.72
	Science	202.85	28.13	20.06	5.25	0.69
Matched Sample, Weighted	Composite	200.17	25.02	19.27	5.15	0.85
	English	199.74	26.83	18.04	6.09	0.78
	Mathematics	199.27	27.61	18.94	5.01	0.77
	Reading	199.41	28.74	19.73	6.35	0.72
	Science	199.57	29.15	19.61	5.26	0.69

Note: SD = standard deviation, *r* = Pearson correlation

Weighting procedure

For weighting, we used the inverse probability of treatment weight (IPTW) based on propensity scores.[4] This involves the following steps:

1) Fit a logistic regression model for the target population where the dependent variable is whether the student is included in the matched sample, and the independent variables are the demographic and achievement variables listed in Table 2.
2) Use the predicted probability from the logistic regression model as the propensity score (*ps*).
3) For students in the matched sample, assign weights as weight = 1/*ps*.

The weighted matched sample is a synthetic sample in which the distribution of covariates is independent of inclusion in the matched sample. The logistic regression model estimates used to

[4] Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

generate the propensity scores for the composite analysis are presented in Table 4. The results are similar for the other subject areas.

The logistic regression model shows that the following variables are associated with a lower probability of being in the matched sample: Inclusion in the 1st, 2nd, 3rd, or 4th quintile of 9th grade test scores; special education status; and membership in all racial/ethnic minority groups. The weighting procedure up-weights students in these groups to make the matched sample more like the target population.

Table 4: Logistic regression propensity score model estimates, composite score

Variable	Beta	SE	p-value
Intercept	2.587	0.044	<.001
Grade 9 Aspire Plus score, quintile*			
1st	-1.872	0.050	<.001
2nd	-1.168	0.049	<.001
3rd	-0.764	0.051	<.001
4th	-0.383	0.054	<.001
Female	0.002	0.026	0.952
Limited English proficient	0.121	0.054	0.025
Special education	-0.239	0.039	<.001
Race/ethnicity*			
African American	-0.184	0.098	0.060
Asian	-0.191	0.099	0.052
Hispanic	-0.325	0.034	<.001
Two or more races	-0.252	0.077	0.001
Other	-0.486	0.068	<.001

*Reference groups are 5th quintile and White.

Updated predicted ACT score ranges

Table 5 shows statistics related to the accuracy of the ACT score predictions and compares the accuracy of the updated predictions to those that were estimated in 2019. The statistics include:

- Mean width: the average width of the predicted ACT score range
- %Within: the percentage of students in the matched sample whose ACT score was within the predicted ACT score range
- %Below: the percentage of students in the matched sample whose ACT score was below the predicted ACT score range (% over-predicted)
- %Above: the percentage of students in the matched sample whose ACT score was above the predicted ACT score range (% under-predicted)

Table 5: Prediction accuracy

Subject	Previously Derived Prediction				Updated Predictions			
	Mean width	%Within	%Below	%Above	Mean width	%Within	%Below	%Above
Composite	5.7	75.9%	20.5%	3.6%	3.2	63.6%	17.7%	18.7%
English	8.2	74.9%	20.7%	4.3%	4.8	59.5%	19.5%	21.0%
Math	6.9	78.5%	17.9%	3.6%	3.6	63.6%	17.6%	18.8%
Reading	9.7	77.8%	14.7%	7.5%	5.7	58.1%	20.1%	21.7%
Science	7.3	73.1%	18.3%	8.6%	4.7	59.3%	20.1%	20.6%

Table 5 shows that:

- The updated predicted score ranges are much tighter than the previous score ranges, as shown by the decrease in the mean width of the prediction intervals.
- The updated predicted score ranges include 58-64% of actual ACT scores. Recall that the 25th and 75th conditional percentiles were used, resulting in a typical coverage percentage of around 60%. The predicted ACT score ranges could be widened to increase the percentage of students scoring within their predicted range.
- The updated predictions result in mostly symmetric prediction error percentages (e.g., similar percent over- and under predicted). Because the weighted matched sample has lower mean achievement than the matched sample, the updated predictions are slightly more likely to under-predict for the matched sample. (But should be more symmetric for the weighted matched sample and target population).
- While the previous predictions included a larger percentage of actual ACT scores, the predicted ranges were wider and resulted in asymmetric prediction errors, with overprediction more likely than underprediction for all subjects and for the composite score.

Figure 1 below shows the updated predicted ACT score ranges (dotted green lines) as compared to those estimated in 2019 (solid black line) for the composite scores. The figure also shows a histogram of 9th grade composite scores to show where the differences are most consequential. Relative to the original predictions from 2019, the updated predictions have smaller ranges (tighter prediction intervals), similar values for the lower bounds of the predicted score range, and lower values for the upper bounds of the predicted score range.

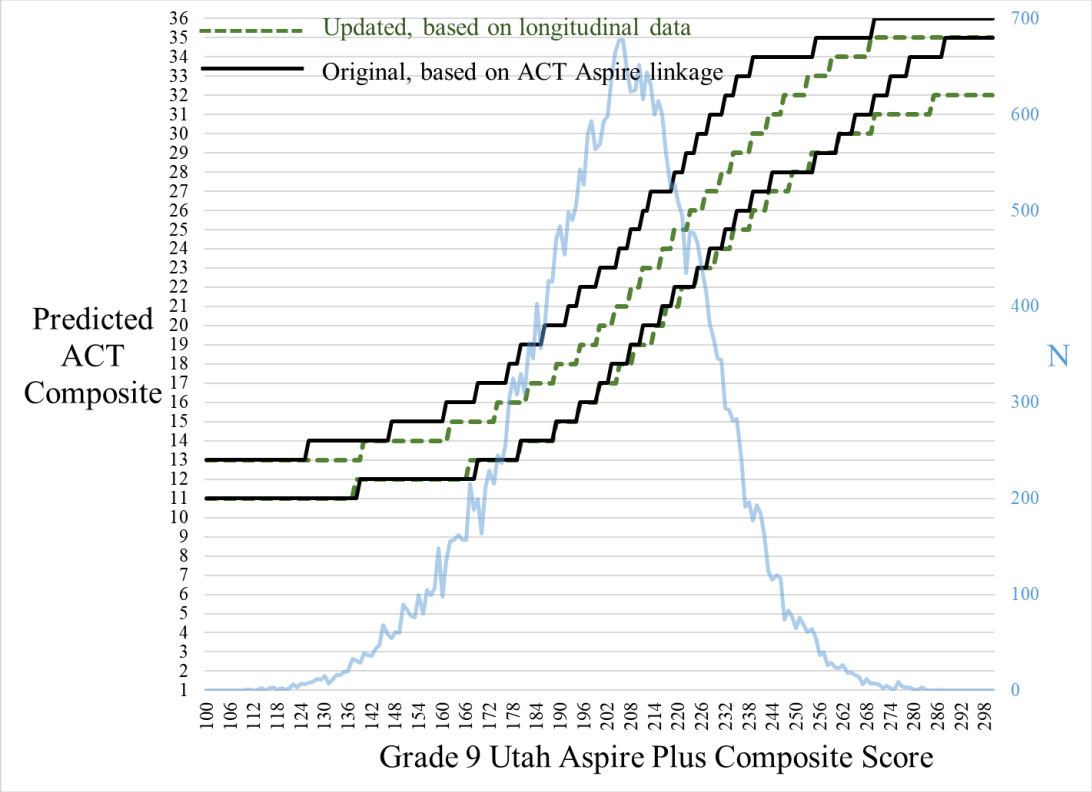


Figure 1: Predicted ACT Composite Scores

Figure 2 below shows the same information for the math scores. The math predictions changed considerably. Similar to the composite score predictions, the updated predictions have smaller ranges (tighter prediction intervals) relative to the original predictions, similar values for the lower bounds of the predicted score range, and lower values for the upper bounds of the predicted score range. Figure 2 also shows that many students (N=235) in the matched sample had the lowest possible 9th grade math score (100). It's possible that these students had a large influence on the SGP model results and updated predicted ACT score ranges. The mean ACT math score for these students was 14.3, with standard deviation 2.6. The updated predicted score range for students with a 9th grade math score of 100 (13-15) seems reasonable.

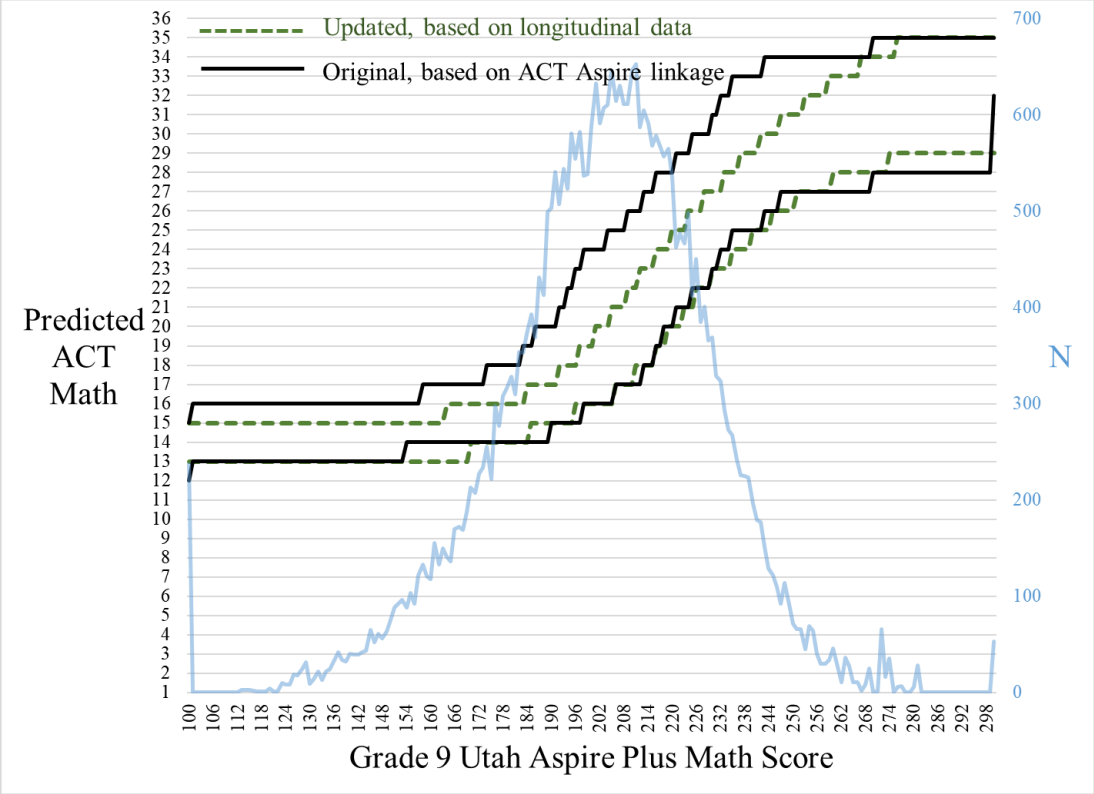


Figure 2: Predicted ACT Math Scores

Figures 1 and 2 show the results for composite and math. Comparisons of the updated and previous predicted ACT score ranges are available for the other subjects in the enclosed spreadsheet.

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
100-154	9-13	4
155-165	10-13	3
166-173	10-14	4
174	11-14	3
175-180	11-15	4
181-182	11-16	5
183-185	12-16	4
186-188	12-17	5
189-190	13-17	4
191-193	13-18	5
194	14-18	4
195-198	14-19	5
199-202	15-20	5
203-207	16-21	5
208-211	17-22	5
212-215	18-23	5
216-219	19-24	5
220	20-24	4
221-224	20-25	5
225-228	21-26	5
229-232	22-27	5
233	22-28	6
234-235	23-28	5
236-239	23-29	6
240	24-29	5
241-244	24-30	6
245-248	25-31	6
249-251	25-32	7
252-253	26-32	6
254-258	26-33	7
259	27-33	6
260-265	27-34	7
266	27-35	8
267-274	28-35	7
275-277	29-35	6
278-284	29-36	7
285-295	30-36	6
296-300	31-36	5

K-2. English Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-161	10-13	3
162-169	10-14	4
170-173	11-14	3
174-178	11-15	4
179-181	11-16	5
182-183	12-16	4
184-187	12-17	5
188	12-18	6
189-191	13-18	5
192-193	13-19	6
194-195	14-19	5
196	14-20	6
197-200	15-20	5
201-203	16-21	5
204	17-21	4
205-206	17-22	5
207-208	18-22	4
209	18-23	5
210-212	19-23	4
213-213	19-24	5
214-215	20-24	4
216-217	20-25	5
218-219	21-25	4
220-222	21-26	5
223-226	22-27	5
227-229	23-28	5
230-231	23-29	6
232-233	24-29	5
234-236	24-30	6
237	25-30	5
238-241	25-31	6
242-246	26-32	6
247-252	27-33	6
253-258	28-34	6
259	29-34	5
260-265	29-35	6
266-273	30-35	5
274-282	31-36	5
283-291	32-36	4
292-300	33-36	3

K-3. Reading Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-145	11-14	3
146-161	11-15	4
162-165	12-15	3
166-174	12-16	4
175-179	12-17	5
180-185	13-18	5
186-188	13-19	6
189	14-19	5
190-194	14-20	6
195-198	15-21	6
199-200	16-21	5
201-203	16-22	6
204	17-22	5
205-206	17-23	6
207-208	18-23	5
209-210	18-24	6
211-212	19-24	5
213-214	19-25	6
215-215	20-25	5
216-218	20-26	6
219-221	21-27	6
222	21-28	7
224-224	22-28	6
225-228	22-29	7
229-232	23-30	7
233	23-31	8
234-238	24-31	7
239-240	24-32	8
241-245	25-32	7
246-249	25-33	8
250-257	26-33	7
258-259	26-34	8
260-273	27-34	7
274-288	28-34	6
289-299	29-34	5
300	30-35	5

K-4. Reading Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-140	11-15	4
141-152	11-16	5
153-161	12-16	4
162-173	12-17	5
174-174	12-18	6
175-181	13-18	5
182-184	13-19	6
185-187	14-19	5
188-190	14-20	6
191-193	15-20	5
194-195	15-21	6
196-199	16-21	5
200	16-22	6
201-203	17-22	5
204-206	18-23	5
207	18-24	6
208-210	19-24	5
211-212	19-25	6
213	20-25	5
214-215	20-26	6
216	20-27	7
217-218	21-27	6
219-221	21-28	7
222-224	22-29	7
225-225	22-30	8
226-227	23-30	7
228-229	23-31	8
230-231	24-31	7
232-234	24-32	8
235-236	25-32	7
237-238	25-33	8
239-242	26-33	7
243	26-34	8
244-248	27-34	7
249-252	28-34	6
253	28-35	7
254-258	29-35	6
259-265	30-35	5
266-273	31-35	4
274-299	32-35	3
300	32-36	4

K-5. Math Grade 09 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-163	13-15	2
164-169	13-16	3
170-183	14-16	2
184	14-17	3
185-191	15-17	2
192-195	15-18	3
196	16-18	2
197-200	16-19	3
201-204	16-20	4
205	16-21	5
206-208	17-21	4
209-210	17-22	5
211	18-22	4
212-215	18-23	5
216-218	19-24	5
219	20-24	4
220-222	20-25	5
223-223	21-25	4
224-225	21-26	5
226-227	22-26	4
228-229	22-27	5
230-232	23-27	4
233-234	23-28	5
235-236	24-28	4
237-239	24-29	5
240-241	25-29	4
242-244	25-30	5
245-246	26-30	4
247-250	26-31	5
251-252	27-31	4
253-258	27-32	5
259	27-33	6
260-266	28-33	5
267-273	28-34	6
274-275	29-34	5
276-300	29-35	6

K-6. Math Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-156	13-15	2
157-162	13-16	3
163-181	14-16	2
182-189	15-17	2
190-192	15-18	3
193-195	16-18	2
196-198	16-19	3
199-202	16-20	4
203-205	17-21	4
206-208	17-22	5
209-212	18-23	5
213-215	19-24	5
216	20-24	4
217-218	20-25	5
219-220	21-25	4
221-221	21-26	5
222-224	22-26	4
225	22-27	5
226-229	23-27	4
230-233	24-28	4
234-237	25-29	4
238	26-29	3
239-242	26-30	4
243-247	27-31	4
248	27-32	5
249-253	28-32	4
254-256	28-33	5
257-260	29-33	4
261-265	29-34	5
266-267	30-34	4
268	30-35	5
284-288	30-36	6
289-300	31-36	5

K-7. Science Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-159	12-16	4
160-168	12-17	5
169-177	13-17	4
178-183	13-18	5
184-185	14-18	4
186-191	14-19	5
192	15-19	4
193-197	15-20	5
198-201	16-21	5
202	17-21	4
203-206	17-22	5
207-208	18-22	4
209	18-23	5
210-212	19-23	4
213-214	19-24	5
215-216	20-24	4
217-218	20-25	5
219-221	21-25	4
222	21-26	5
223-224	22-26	4
225-228	22-27	5
229-230	23-27	4
231-234	23-28	5
235-235	23-29	6
236-238	24-29	5
239-243	24-30	6
244-247	24-31	7
248	25-31	6
249-255	25-32	7
256-265	25-33	8
266-286	25-34	9
287-191	26-34	8
192-237	16-20	4
238-245	24-28	4
246-254	25-30	5
255-260	26-32	6
261-267	26-33	7
268-277	27-34	7
278-297	27-35	8
298-300	28-35	7

K-8. Science Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-160	12-17	5
161-171	13-17	4
172-176	13-18	5
177-182	14-18	4
183-185	14-19	5
186-189	15-19	4
190-191	15-20	5
192-196	16-20	4
197-197	16-21	5
198-203	17-21	4
204-204	17-22	5
205-209	18-22	4
210	18-23	5
211-214	19-23	4
215	19-24	5
216-219	20-24	4
220-221	20-25	5
222-225	21-25	4
226	21-26	5
227-230	22-26	4
231	22-27	5
232-235	23-27	4
236-237	23-28	5
238-239	24-28	4
240-244	24-29	5
245	24-30	6
246-249	25-30	5
250-254	25-31	6
255-260	26-32	6
261-267	26-33	7
268-277	27-34	7
278-297	27-35	8
298-300	28-35	7

K-9. Composite Grade 9 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-137	11-13	2
138-139	12-13	1
140-161	12-14	2
162-166	12-15	3
167-173	13-15	2
174-179	13-16	3
180-181	14-16	2
182-188	14-17	3
189-194	15-18	3
195-199	16-19	3
200-203	17-20	3
204-204	17-21	4
205-207	18-21	3
208-208	18-22	4
209-210	19-22	3
211-213	19-23	4
214-215	20-23	3
216-216	20-24	4
217-218	21-24	3
219-220	21-25	4
221-222	22-25	3
223-224	22-26	4
225-226	23-26	3
227-229	23-27	4
230	24-27	3
231-233	24-28	4
234-238	25-29	4
239-242	26-30	4
243-246	27-31	4
247-248	27-32	5
249-252	28-32	4
253	28-33	5
254-258	29-33	4
259-260	29-34	5
261-268	30-34	4
269-269	30-35	5
270-284	31-35	4
285-300	32-35	3

K-10. Composite Grade 10 Predicted ACT Score Ranges

Utah Aspire Plus Scale Score	Predicted ACT Score Range	Predicted Score Width
100-112	11-14	3
113-150	12-14	2
151-162	12-15	3
163-169	13-15	2
170-177	13-16	3
178	14-16	2
179-185	14-17	3
186-190	15-18	3
191	15-19	4
192-196	16-19	3
197-200	17-20	3
201-201	17-21	4
202-205	18-21	3
206-208	19-22	3
209	19-23	4
210-212	20-23	3
213-213	20-24	4
214-215	21-24	3
216	21-25	4
217-218	22-25	3
219-220	22-26	4
221-222	23-26	3
223-224	23-27	4
225	24-27	3
226-228	24-28	4
229	25-28	3
230-232	25-29	4
233	26-29	3
234-236	26-30	4
237-238	27-30	3
239-241	27-31	4
242-243	28-31	3
244-246	28-32	4
247-248	29-32	3
249-252	29-33	4
253-256	30-33	3
257-259	30-34	4
260-268	31-34	3
269-270	32-34	2
271-300	32-35	3

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	42,964	198	25.75	182	199	214	-0.23
Gender	Female	20,555	202	24.28	187	202	217	-0.11
	Male	22,405	194	26.48	178	196	211	-0.27
Ethnicity	Hispanic or Latino Ethnicity	7,251	184	24.20	169	185	200	-0.23
	Asian	715	202	27.21	186	203	218	-0.07
	Native Hawaiian or Other Pacific Islander	596	187	21.93	173	187	200	-0.11
	Black or African American	534	179	25.96	163	181	196	-0.26
	American Indian or Alaska Native	318	182	22.20	168	183	196	0.10
	White	32,361	201	24.92	187	202	217	-0.26
	Other	1,189	199	24.92	185	200	215	-0.40
Limited English Proficiency	No	40,684	199	24.89	185	200	215	-0.21
	Yes	2,280	167	21.45	155	169	181	-0.45
Economic Disadvantage	No	32,186	201	24.80	187	202	217	-0.22
	Yes	10,778	187	25.61	171	188	204	-0.19
Special Education	No	38,870	201	24.19	186	201	216	-0.18
	Yes	4,094	170	23.04	156	170	184	-0.01

L-2. English Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	39,286	197	26.62	180	197	214	-0.01
Gender	Female	18,975	201	25.27	185	200	216	0.13
	Male	20,305	193	27.26	175	193	211	-0.05
Ethnicity	Hispanic or Latino Ethnicity	6,425	184	23.97	168	184	199	-0.03
	Asian	677	204	29.46	184	206	223	0.11
	Native Hawaiian or Other Pacific Islander	491	185	20.81	172	187	199	-0.19
	Black or African American	478	180	25.68	164	181.5	196	-0.12
	American Indian or Alaska Native	268	181	22.02	166	181	196	-0.08
	White	29,837	200	26.16	184	200	217	-0.04
	Other	1,110	197	25.87	181	196	215	-0.11
Limited English Proficiency	No	37,632	198	25.98	182	198	215	0.01
	Yes	1,654	165	20.10	152	165	178	-0.02
Economic Disadvantage	No	30,214	200	26.22	184	200	216	-0.03
	Yes	9,072	187	25.36	170	187	203	0.05
Special Education	No	35,842	199	25.56	184	199	215	0.02
	Yes	3,444	170	22.35	157	170	183	0.22

L-3. Math Grade 9 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	42,045	192	31.30	176	196	213	-0.77
Gender	Female	19,975	193	29.44	178	196	212	-0.87
	Male	22,067	192	32.89	174	195	214	-0.69
Ethnicity	Hispanic or Latino Ethnicity	7,088	174	31.58	158	176	194.5	-0.61
	Asian	704	199	31.66	184	202	218	-0.66
	Native Hawaiian or Other Pacific Islander	565	174	31.93	159	179	195	-0.77
	Black or African American	523	167	31.43	152	172	189	-0.62
	American Indian or Alaska Native	320	176	29.29	160	178	195	-0.78
	White	31,684	197	29.20	183	200	216	-0.84
	Other	1,161	193	30.67	178	196	213	-0.80
Limited English Proficiency	No	39,782	194	30.14	179	198	214	-0.79
	Yes	2,263	157	29.79	142	162	177	-0.46
Economic Disadvantage	No	31,524	197	29.53	182	200	216	-0.84
	Yes	10,521	179	32.35	162	181	200	-0.58
Special Education	No	38,053	196	28.99	181	199	215	-0.77
	Yes	3,992	158	31.93	142	163	179	-0.35

L-4. Math Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	38,573	192	34.02	177	196	213	-0.97
Gender	Female	18,553	192	31.30	179	196	212	-1.11
	Male	20,014	192	36.37	176	196	215	-0.87
Ethnicity	Hispanic or Latino Ethnicity	6,275	173	35.45	162	179	196	-0.84
	Asian	668	204	34.82	185	205	225	-0.59
	Native Hawaiian or Other Pacific Islander	505	175	32.91	168	182	193	-1.08
	Black or African American	470	168	34.90	156	175	190	-0.79
	American Indian or Alaska Native	264	175	36.39	161	181	198.5	-0.78
	White	29,317	196	31.87	182	200	216	-1.04
	Other	1,074	191	34.27	178	196	211	-1.02
Limited English Proficiency	No	36,917	194	32.90	179	197	214	-0.99
	Yes	1,656	154	36.09	100	165	179	-0.44
Economic Disadvantage	No	29,678	196	32.19	182	200	216	-1.02
	Yes	8,895	177	35.95	165	183	200	-0.83
Special Education	No	35,192	195	31.49	181	198	215	-1.00
	Yes	3,381	155	36.82	100	165	179	-0.35

L-5. Reading Grade 9 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	43,214	197	27.48	180	198	215	-0.29
Gender	Female	20,627	200	26.29	184	201	218	-0.23
	Male	22,583	193	28.11	176	195	213	-0.31
Ethnicity	Hispanic or Latino Ethnicity	7,418	183	26.92	166	184	201	-0.19
	Asian	723	203	27.77	185	206	222	-0.44
	Native Hawaiian or Other Pacific Islander	591	183	25.99	168	183	200	-0.25
	Black or African American	537	181	28.01	163	181	199	-0.08
	American Indian or Alaska Native	328	183	23.60	168	183	200	-0.17
	White	32,424	200	26.44	184	202	218	-0.32
	Other	1,193	198	26.96	180	200	216	-0.23
Limited English Proficiency	No	40,868	199	26.61	182	200	216	-0.28
	Yes	2,346	166	23.50	152	168	181	-0.40
Economic Disadvantage	No	32,255	201	26.40	184	202	218	-0.31
	Yes	10,959	186	27.67	168	186	205	-0.18
Special Education	No	39,060	200	26.05	184	201	217	-0.28
	Yes	4,154	169	24.85	155	169	184	-0.09

L-6. Reading Grade 10 Scale Score Descriptive Statistics

	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	39,417	202	26.04	187	203	218	-0.24
Gender	Female	19,003	205	24.24	191	206	219	-0.04
	Male	20,408	199	27.26	182	201	217	-0.31
Ethnicity	Hispanic or Latino Ethnicity	6,525	189	24.09	174	191	206	-0.28
	Asian	683	206	26.00	190	208	222	-0.26
	Native Hawaiian or Other Pacific Islander	493	186	24.57	172	189	203	-0.68
	Black or African American	486	187	25.64	169	188	204	-0.05
	American Indian or Alaska Native	273	188	23.67	174	190	204	-0.47
	White	29,848	205	25.51	190	206	220	-0.25
	Other	1,109	201	24.91	188	203	216	-0.64
Limited English Proficiency	No	37,757	203	25.45	188	204	219	-0.24
	Yes	1,660	172	21.03	160	172	186	-0.48
Economic Disadvantage	No	30,246	204	25.58	190	206	220	-0.25
	Yes	9,171	192	25.27	176	193	209	-0.24
Special Education	No	35,941	204	24.98	190	205	219	-0.22
	Yes	3,476	177	23.64	163	176.5	191	-0.21

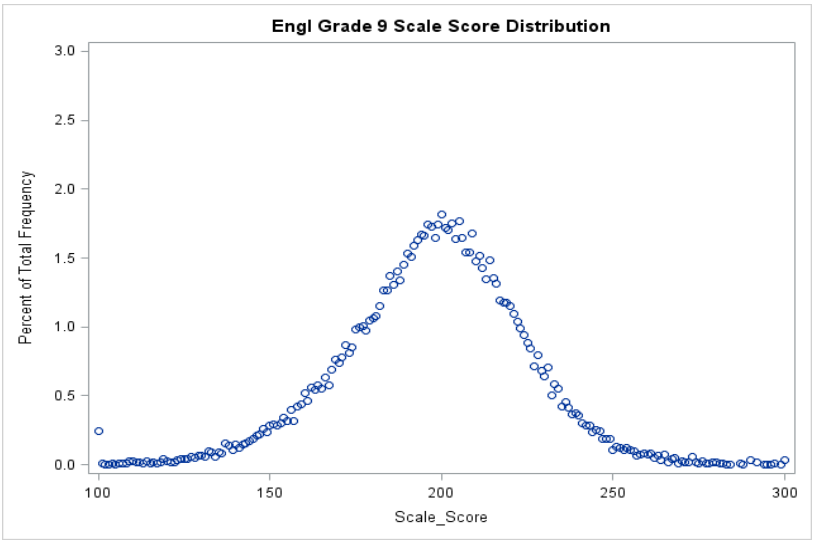
L-7. Science Grade 9 Scale Score Descriptive Statistics

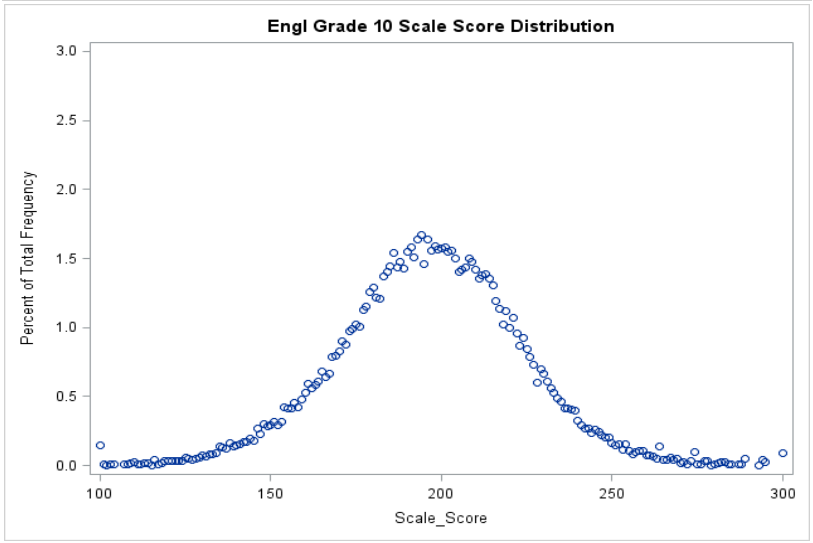
	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	42,635	197	32.06	180	200	218	-0.50
Gender	Female	20,328	197	29.93	182	201	217	-0.71
	Male	22,303	197	33.89	178	200	219	-0.36
Ethnicity	Hispanic or Latino Ethnicity	7,304	181	30.88	165	184	202	-0.47
	Asian	713	203	31.92	185	206	222	-0.33
	Native Hawaiian or Other Pacific Islander	590	180	29.20	165	184	200	-0.60
	Black or African American	532	175	32.61	156	179	197.5	-0.45
	American Indian or Alaska Native	325	181	30.58	165	185	202	-0.59
	White	31,992	202	30.92	186	204	221	-0.54
	Other	1,179	198	30.96	182	201	217	-0.49
Limited English Proficiency	No	40,291	199	31.41	182	202	219	-0.52
	Yes	2,344	167	27.72	153	171	186	-0.62
Economic Disadvantage	No	31,824	201	31.04	185	204	221	-0.54
	Yes	10,811	185	32.05	167	188	206	-0.40
Special Education	No	38,548	200	30.98	184	203	220	-0.53
	Yes	4,087	171	30.21	155	174	190	-0.29

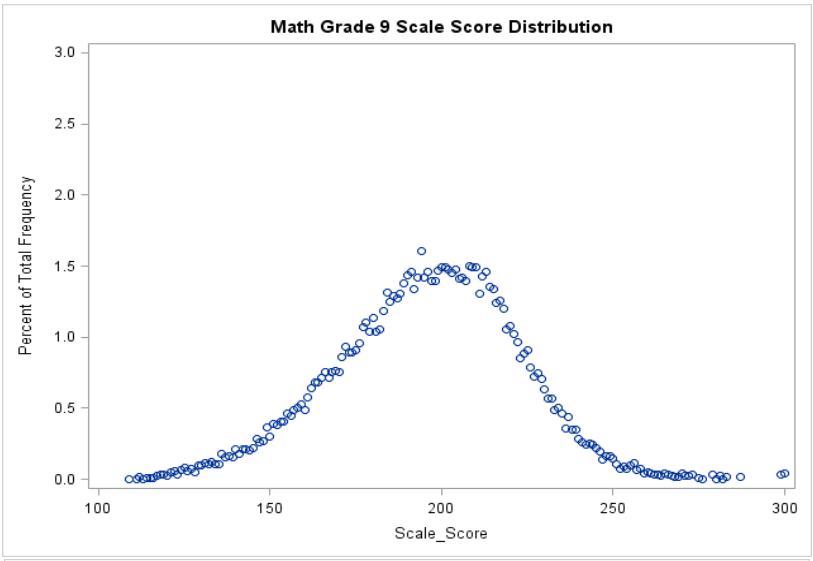
L-8. Science Grade 10 Scale Score Descriptive Statistics

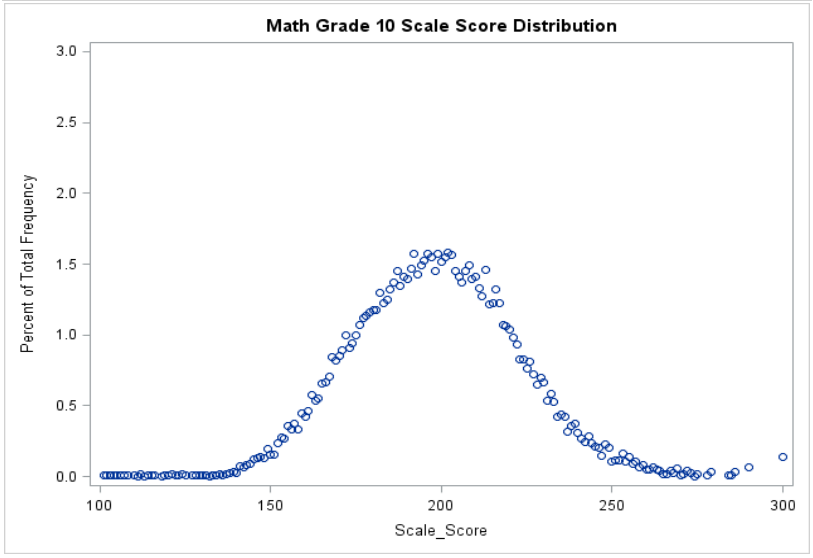
	Test Group	*N*	Mean	SD	P25	Median	P75	Skew
All	Students Scored	39,067	197	33.45	180	200	218	-0.51
Gender	Female	18,800	197	31.44	181	200	217	-0.72
	Male	20,261	197	35.21	178	200	220	-0.37
Ethnicity	Hispanic or Latino Ethnicity	6,494	182	31.49	166	186	203	-0.56
	Asian	676	204	36.53	184	207	227	-0.31
	Native Hawaiian or Other Pacific Islander	502	179	31.46	165	184	199	-0.73
	Black or African American	480	178	31.21	162	182	198.5	-0.42
	American Indian or Alaska Native	269	181	33.86	164	185	202	-0.36
	White	29,551	201	32.64	185	204	221	-0.56
	Other	1,095	196	35.29	179	200	219	-0.56
Limited English Proficiency	No	37,381	198	33.10	181	201	219	-0.54
	Yes	1,686	169	28.21	155	173	188	-0.68
Economic Disadvantage	No	29,934	200	32.89	184	203	221	-0.54
	Yes	9,133	186	32.92	169	190	207	-0.50
Special Education	No	35,666	199	32.62	183	203	220	-0.55
	Yes	3,401	171	30.88	156	175	191	-0.43

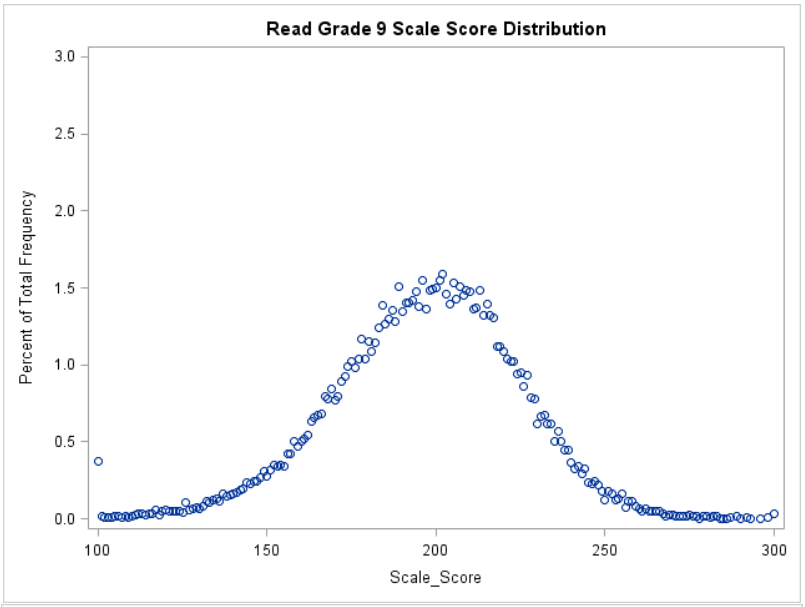
Appendix M: Scale Score Distributions for Overall Testing Population

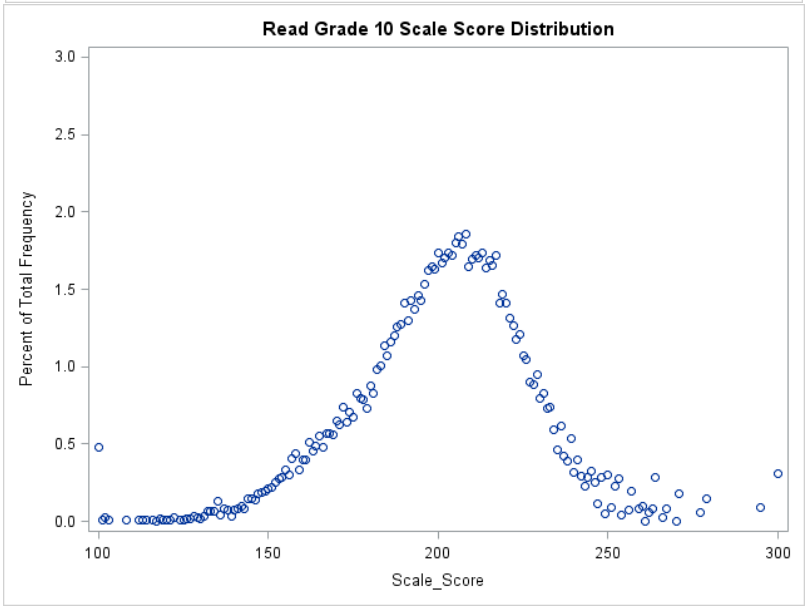


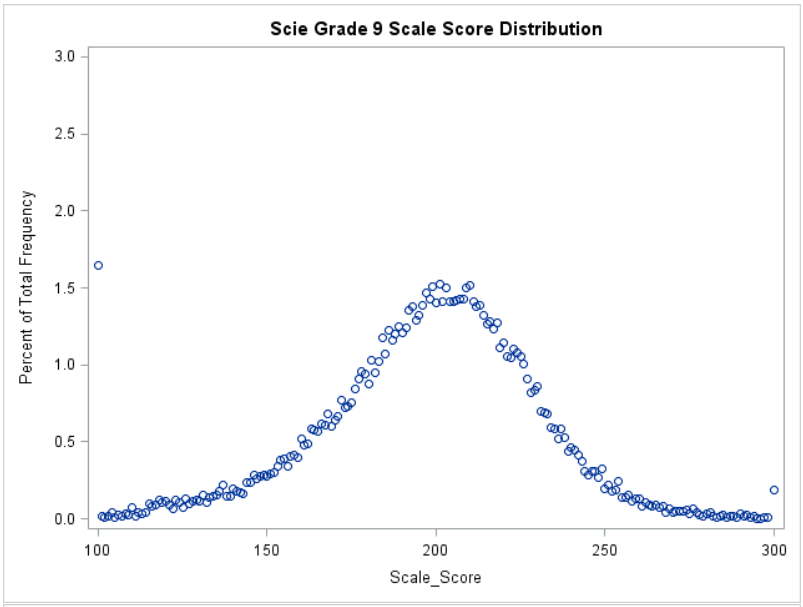


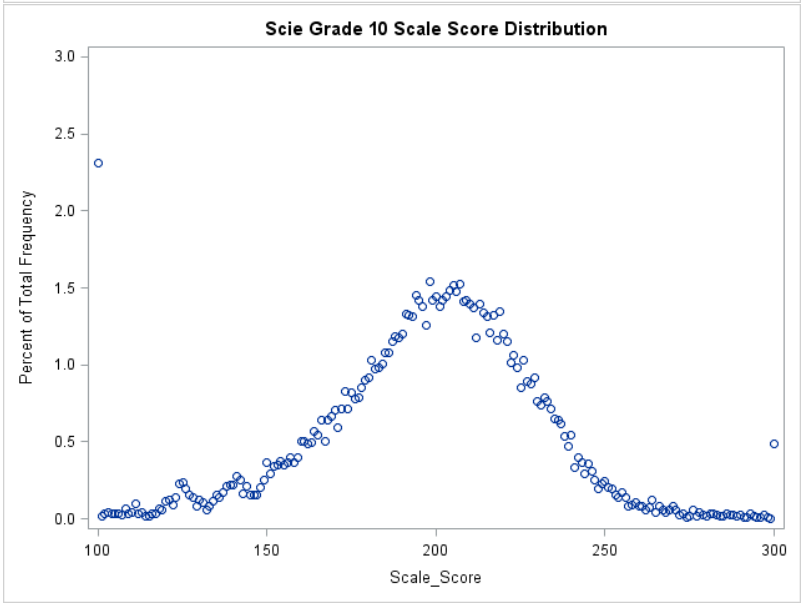

Math Grade 9 Scale Score Distribution
Percent of Total Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100
150
200
250
300
Scale_Score


Math Grade 10 Scale Score Distribution
Percent of Total Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100
150
200
250
300
Scale_Score


Read Grade 9 Scale Score Distribution
Percent of Total Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100
150
200
250
300
Scale_Score


Read Grade 10 Scale Score Distribution
Percent of Total Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100
150
200
250
300
Scale_Score


Scie Grade 9 Scale Score Distribution
Percent of Total Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100
150
200
250
300
Scale_Score


Scie Grade 10 Scale Score Distribution
Percent of Total Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100
150
200
250
300
Scale_Score

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	42,964	9.77	44.94	41.55	3.74
Gender	Female	20,555	6.09	42.79	46.34	4.77
	Male	22,405	13.14	46.91	37.15	2.80
Ethnicity	Hispanic or Latino Ethnicity	7,251	19.62	57.52	21.97	0.88
	Asian	715	6.57	40.84	45.73	6.85
	Native Hawaiian or Other Pacific Islander	596	14.60	62.08	22.32	1.01
	Black or African American	534	28.28	52.43	18.54	0.75
	American Indian or Alaska Native	318	21.70	58.18	19.50	0.63
	White	32,361	7.22	41.61	46.73	4.44
	Other	1,189	7.06	45.50	43.48	3.95
Limited English Proficiency	No	40,684	7.99	44.42	43.65	3.95
	Yes	2,280	41.49	54.21	4.21	0.09
Economic Disadvantage	No	32,186	6.99	42.07	46.44	4.50
	Yes	10,778	18.06	53.49	26.96	1.48
Special Education	No	38,870	6.54	44.19	45.16	4.11
	Yes	4,094	40.40	52.03	7.28	0.29

N-2. English Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	39,286	8.34	45.62	42.46	3.58
Gender	Female	18,975	5.01	43.43	47.05	4.51
	Male	20,305	11.45	47.68	38.16	2.71
Ethnicity	Hispanic or Latino Ethnicity	6,425	15.98	59.21	24.09	0.72
	Asian	677	5.91	36.93	49.19	7.98
	Native Hawaiian or Other Pacific Islander	491	13.03	63.14	23.63	0.20
	Black or African American	478	20.92	59.21	19.25	0.63
	American Indian or Alaska Native	268	16.42	64.55	19.03	
	White	29,837	6.40	42.15	47.21	4.25
	Other	1,110	8.11	47.57	41.08	3.24
Limited English Proficiency	No	37,632	6.95	45.15	44.16	3.73
	Yes	1,654	39.84	56.35	3.69	0.12
Economic Disadvantage	No	30,214	6.58	42.36	46.82	4.24
	Yes	9,072	14.19	56.47	27.95	1.39
Special Education	No	35,842	6.16	44.10	45.85	3.88
	Yes	3,444	30.98	61.44	7.14	0.44

N-3. Math Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	42,045	20.74	43.22	29.39	6.65
Gender	Female	19,975	18.94	45.86	29.91	5.29
	Male	22,067	22.37	40.83	28.91	7.89
Ethnicity	Hispanic or Latino Ethnicity	7,088	42.41	43.51	12.70	1.38
	Asian	704	15.20	41.76	30.97	12.07
	Native Hawaiian or Other Pacific Islander	565	39.65	47.08	12.21	1.06
	Black or African American	523	49.33	42.26	7.84	0.57
	American Indian or Alaska Native	320	39.06	46.88	13.44	0.63
	White	31,684	15.08	43.04	33.90	7.97
	Other	1,161	19.12	44.53	29.63	6.72
Limited English Proficiency	No	39,782	18.09	43.98	30.90	7.02
	Yes	2,263	67.34	29.70	2.83	0.13
Economic Disadvantage	No	31,524	15.57	42.82	33.70	7.91
	Yes	10,521	36.25	44.41	16.46	2.88
Special Education	No	38,053	16.18	44.51	32.03	7.28
	Yes	3,992	64.23	30.91	4.23	0.63

N-4. Math Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	38,573	28.58	41.65	24.02	5.76
Gender	Female	18,553	27.33	44.64	23.62	4.40
	Male	20,014	29.73	38.87	24.39	7.02
Ethnicity	Hispanic or Latino Ethnicity	6,275	52.05	37.23	9.59	1.13
	Asian	668	20.96	36.23	28.14	14.67
	Native Hawaiian or Other Pacific Islander	505	46.53	44.16	8.51	0.79
	Black or African American	470	59.57	33.62	6.38	0.43
	American Indian or Alaska Native	264	47.35	40.53	10.23	1.89
	White	29,317	22.75	42.76	27.72	6.77
	Other	1,074	28.68	43.11	22.91	5.31
Limited English Proficiency	No	36,917	26.38	42.62	24.99	6.02
	Yes	1,656	77.66	19.87	2.42	0.06
Economic Disadvantage	No	29,678	23.18	42.77	27.18	6.88
	Yes	8,895	46.59	37.90	13.48	2.03
Special Education	No	35,192	23.93	43.76	26.04	6.26
	Yes	3,381	76.93	19.61	2.93	0.53

N-5. Reading Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	43,214	12.19	45.85	32.14	9.83
Gender	Female	20,627	8.83	44.46	34.98	11.73
	Male	22,583	15.25	47.12	29.53	8.10
Ethnicity	Hispanic or Latino Ethnicity	7,418	23.85	54.04	18.33	3.77
	Asian	723	10.37	37.07	36.79	15.77
	Native Hawaiian or Other Pacific Islander	591	21.66	57.19	18.61	2.54
	Black or African American	537	29.42	50.47	17.69	2.42
	American Indian or Alaska Native	328	20.12	60.06	17.38	2.44
	White	32,424	9.08	43.75	35.79	11.38
	Other	1,193	10.65	45.60	33.03	10.73
Limited English Proficiency	No	40,868	10.24	45.64	33.73	10.39
	Yes	2,346	46.12	49.40	4.35	0.13
Economic Disadvantage	No	32,255	8.96	43.79	35.63	11.62
	Yes	10,959	21.68	51.89	21.86	4.56
Special Education	No	39,060	8.99	45.41	34.83	10.77
	Yes	4,154	42.27	49.90	6.84	0.99

N-6. Reading Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	39,417	14.07	36.27	41.99	7.67
Gender	Female	19,003	9.71	36.50	45.07	8.71
	Male	20,408	18.14	36.04	39.12	6.70
Ethnicity	Hispanic or Latino Ethnicity	6,525	25.35	46.65	25.87	2.13
	Asian	683	11.42	31.04	47.73	9.81
	Native Hawaiian or Other Pacific Islander	493	26.98	49.70	22.11	1.22
	Black or African American	486	32.72	40.95	23.05	3.29
	American Indian or Alaska Native	273	26.01	48.35	24.54	1.10
	White	29,848	11.09	33.63	46.13	9.15
	Other	1,109	12.80	38.32	43.37	5.50
Limited English Proficiency	No	37,757	12.31	36.09	43.60	8.00
	Yes	1,660	54.16	40.24	5.48	0.12
Economic Disadvantage	No	30,246	11.40	33.89	45.74	8.97
	Yes	9,171	22.90	44.10	29.64	3.37
Special Education	No	35,941	10.98	35.68	45.00	8.34
	Yes	3,476	46.00	42.32	10.90	0.78

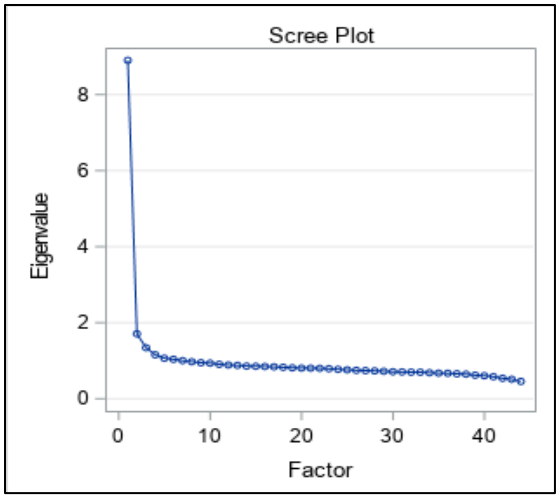
N-7. Science Grade 9 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	42,635	31.97	33.16	26.34	8.53
Gender	Female	20,328	29.92	36.10	27.63	6.35
	Male	22,303	33.84	30.49	25.16	10.51
Ethnicity	Hispanic or Latino Ethnicity	7,304	53.86	31.12	12.62	2.40
	Asian	713	26.37	32.40	30.15	11.08
	Native Hawaiian or Other Pacific Islander	590	54.24	33.73	11.19	0.85
	Black or African American	532	59.59	30.64	8.27	1.50
	American Indian or Alaska Native	325	51.69	35.69	10.77	1.85
	White	31,992	26.09	33.56	30.14	10.21
	Other	1,179	30.28	35.62	25.87	8.23
Limited English Proficiency	No	40,291	29.42	33.83	27.75	9.00
	Yes	2,344	75.85	21.67	2.13	0.34
Economic Disadvantage	No	31,824	26.47	33.73	29.72	10.08
	Yes	10,811	48.16	31.50	16.39	3.95
Special Education	No	38,548	28.08	34.10	28.53	9.29
	Yes	4,087	68.71	24.32	5.65	1.32

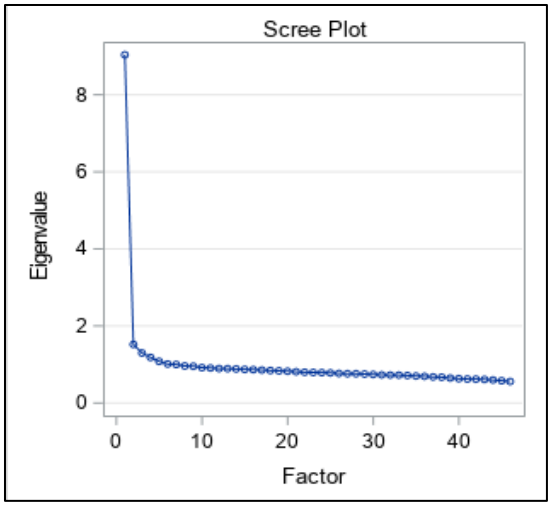
N-8. Science Grade 10 Performance Level Distribution

	Test Group	*N*	Below Proficient	Approaching Proficient	Proficient	Highly Proficient
All	Students Scored	39,067	31.89	31.68	29.68	6.75
Gender	Female	18,800	30.46	34.28	30.26	4.99
	Male	20,261	33.23	29.25	29.14	8.38
Ethnicity	Hispanic or Latino Ethnicity	6,494	51.15	31.14	15.89	1.82
	Asian	676	27.51	25.59	32.99	13.91
	Native Hawaiian or Other Pacific Islander	502	53.98	33.07	11.75	1.20
	Black or African American	480	58.75	28.13	11.46	1.67
	American Indian or Alaska Native	269	51.67	32.71	13.75	1.86
	White	29,551	26.77	31.92	33.40	7.90
	Other	1,095	31.78	32.69	29.13	6.39
Limited English Proficiency	No	37,381	30.02	32.08	30.85	7.05
	Yes	1,686	73.55	22.66	3.68	0.12
Economic Disadvantage	No	29,934	27.67	31.60	32.89	7.85
	Yes	9,133	45.75	31.93	19.17	3.15
Special Education	No	35,666	28.42	32.39	31.89	7.29
	Yes	3,401	68.27	24.20	6.47	1.06

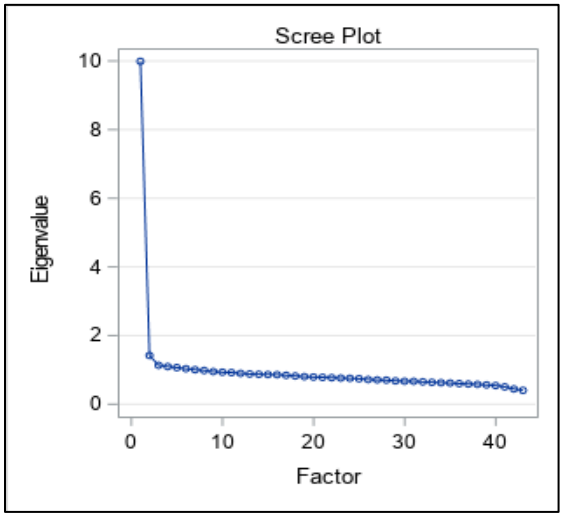
Appendix O: Principal Components Scree Plots



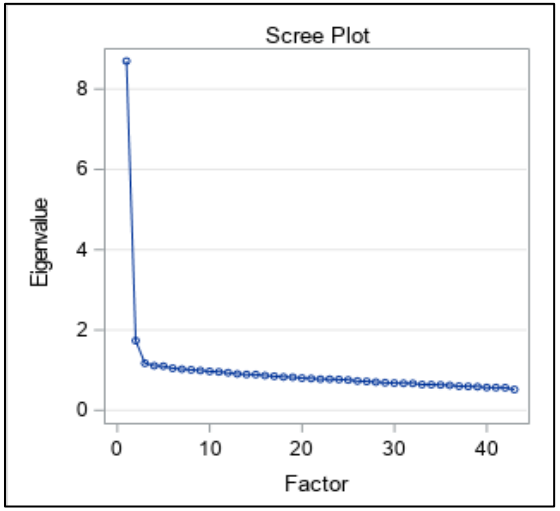
O-1. English Grade 9 Principal Components Scree Plot



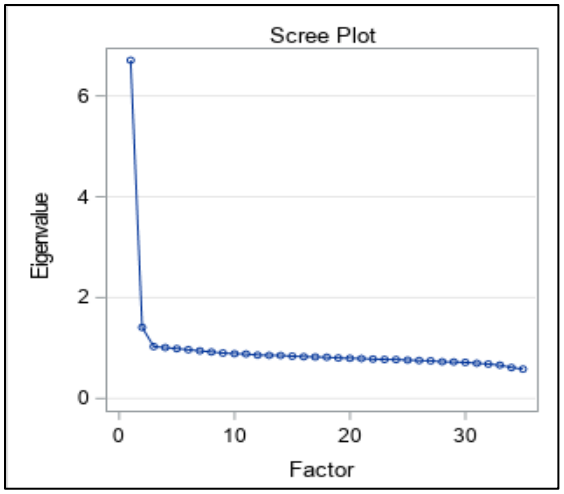
O-2. English Grade 10 Principal Components Scree Plot



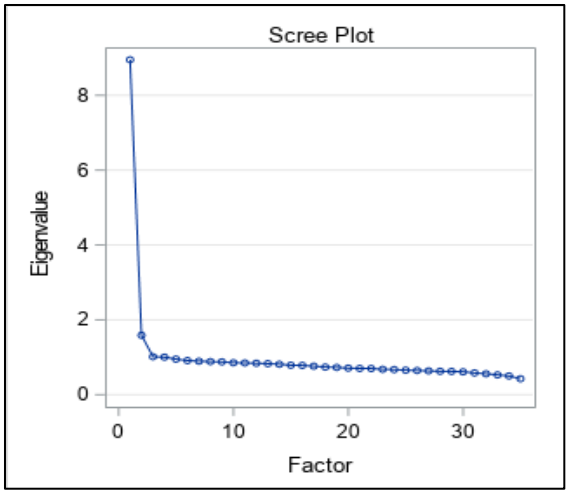
O-3. Math Grade 9 Principal Components Scree Plot



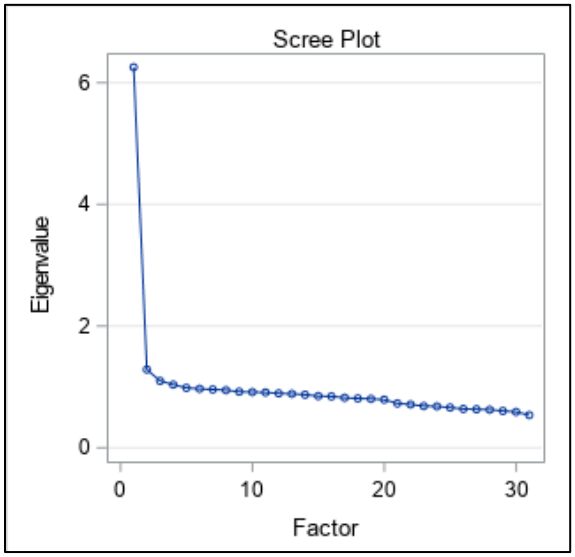
O-4. Math Grade 10 Principal Components Scree Plot



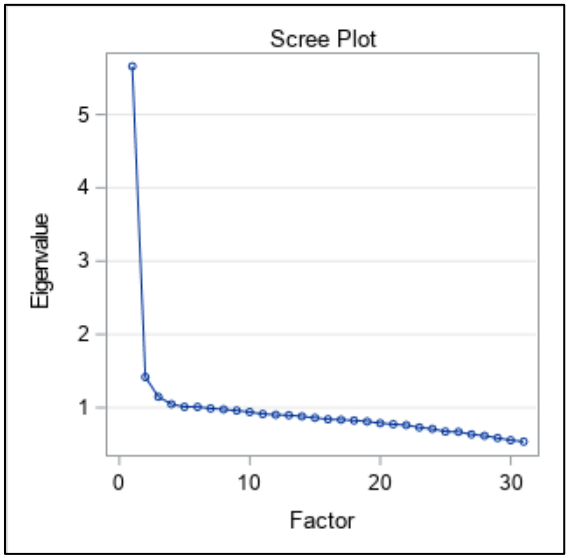
O-5. Reading Grade 9 Principal Components Scree Plot



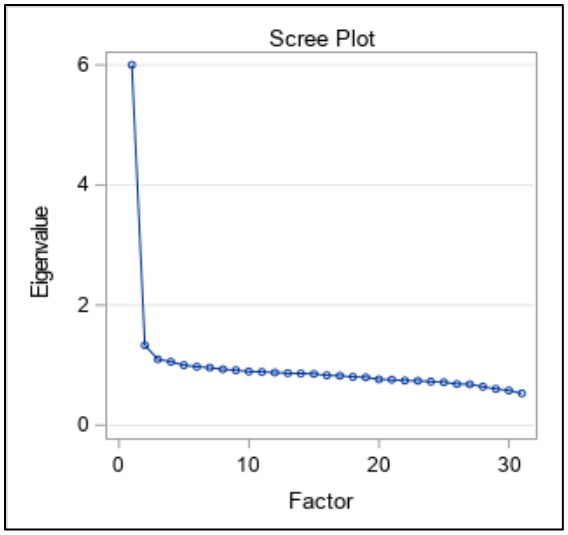
O-6. Reading Grade 10 Principal Components Scree Plot



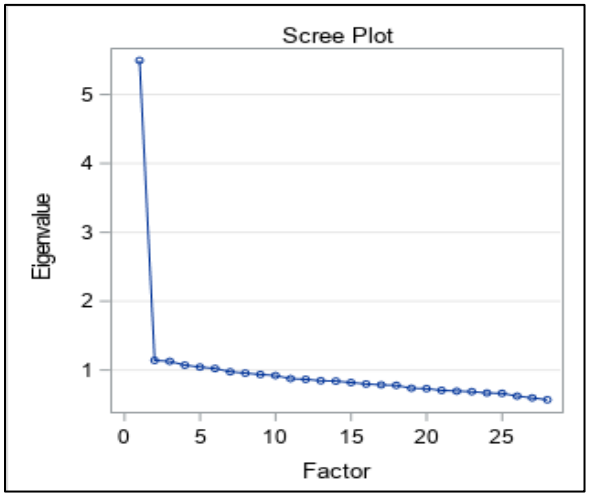
O-7. Science Grade 9 Form 1 Principal Components Scree Plot



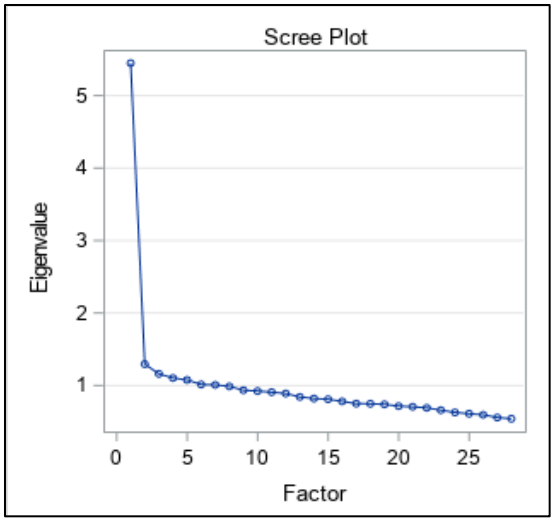
O-8. Science Grade 9 Form 2 Principal Components Scree Plot



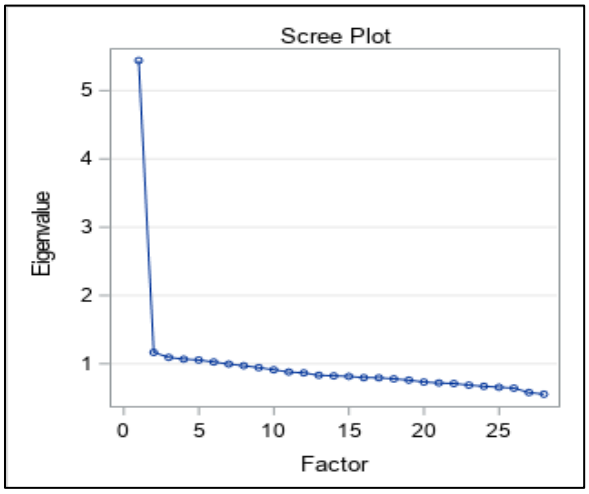
O-9. Science Grade 9 Form 3 Principal Components Scree Plot



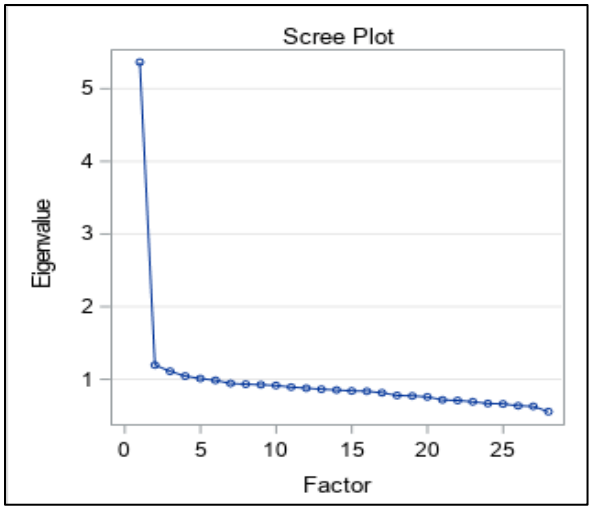
O-11. Science Grade 10 Form 2 Principal Components Scree Plot



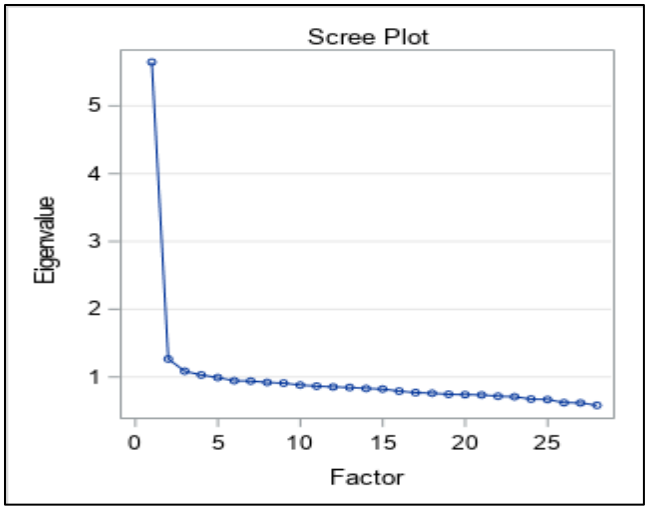
O-10. Science Grade 10 Form 1 Principal Components Scree Plot



O-12. Science Grade 10 Form 3 Principal Components Scree Plot



O-13. Science Grade 10 Form 4 Principal Components Scree Plot



O-14. Science Grade 10 Form 5 Principal Components Scree Plot

Appendix P: Subscore Correlations

P-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	0.73	1.00		
	Knowledge of Language	0.71	0.51	1.00	
	Production of Writing	0.97	0.63	0.62	1.00
10	Total	1.00			
	Conventions of Standard English	0.81	1.00		
	Knowledge of Language	0.72	0.56	1.00	
	Production of Writing	0.97	0.69	0.62	1.00

P-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	0.94	1.00		
	Craft and Structure	0.87	0.71	1.00	
	Integration of Knowledge and Ideas	0.54	0.44	0.42	1.00
10	Total	1.00			
	Key Ideas	0.90	1.00		
	Craft and Structure	0.89	0.70	1	
	Integration of Knowledge and Ideas	0.71	0.56	0.58	1.00

P-3. Math Correlations of Total Score and Subscores

Grade	Subdomain	Math Total	Number and Quantity	Algebra	Functions	Geometry	Statistics and Probability
9	Total	1.00	—				
	Algebra	0.82	—	1.00			
	Functions	0.81	—	0.64	1.00		
	Geometry	0.81	—	0.61	0.61	1.00	
	Statistics and Probability	0.78	—	0.61	0.59	0.58	1.00
10	Total	1.00	1.00				
	Number and Quantity	0.59	0.59	1.00			
	Algebra	0.78	0.78	0.52	1.00		
	Functions	0.74	0.74	0.46	0.59	1.00	
	Geometry	0.81	0.81	0.48	0.62	0.60	1.00
	Statistics and Probability	0.59	0.59	0.33	0.43	0.42	0.46

P-4. Science Correlations of Total Score and Subscores

Grade	Subdomain	Science Total	Gathering & Investigating	Developing Models	Using Mathematical Thinking	Construct Explanation
9	Total	1.00				
	Gathering & Investigating	0.78	1.00			
	Developing Models	0.71	0.49	1.00		
	Using Mathematical Thinking	0.71	0.50	0.40	1.00	
	Construct Explanation	0.60	0.42	0.35	0.35	1.00
10	Total	1.00				
	Gathering & Investigating	0.63	1.00			
	Developing Models	0.23	0.13	1.00		
	Using Mathematical Thinking	0.64	0.38	0.12	1.00	
	Construct Explanation	0.86	0.49	0.16	0.46	1.00