Hawai`i Alternate Assessments

2022–2023 Technical Report

English Language Arts Grades 3–8, 11
Mathematics Grades 3–8, 11
Science Grades 5, 8, 11

Submitted to
Hawai`i Department of Education
by Cambium Assessment, Inc.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

1. OVERVIEW

This report provides a technical summary of the 2022–2023 Hawai`i State Alternate Assessments (HSA-Alt) in English Language Arts (ELA) and mathematics administered in grades 3–8 and 11, and in science administered in grades 5, 8, and 11. The purpose of this technical report is to document the evidence supporting the claims made for how HSA-Alt test scores may be interpreted. The report includes 10 chapters, including all the evidence accrued about the technical quality of a testing system. The data included in this report are based on Hawai`i data for the alternate assessments, including all aspects of the technical qualities for the HSA-Alt described in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) and the requirements of the U.S. Department of Education Peer Review of State Assessment Systems Non-Regulatory Guidance for States.

Chapter 2 documents the test administration procedures, including test administrator training, test administration manual, accommodations, as well as prevention of disruptions in the Test Delivery System. Chapter 3 summarizes the results of the spring 2023 HSA-Alt Assessments in ELA, mathematics, and science. These chapters summarize the test-taking student population, their performance on the assessments, and the time spent in taking the assessments. Chapter 4 describes the item-development process, specifically, the sequence of reviews that each item must pass through before being eligible for HSA-Alt test administration. Chapter 4 also summarizes the field-test item analyses, data review, and the procedures used to scale and calibrate the HSA-Alt for scoring and reporting. Chapter 5 provides validity evidence on the test blueprint coverage, cognitive lab, internal consistency, and relations to other variables.

Chapter 6 provides evidence for the reliability of the HSA-Alt, including internal consistency reliability, standard errors of measurement, and the reliability of performance-level classifications. Chapter 7 describes the scoring procedures used in producing scale scores and performance levels. Chapter 8 describes the procedures that the Hawai`i Department of Education uses to identify and adopt performance standards for the HSA-Alt Assessments. Chapter 9 provides a description of the score reporting system and the interpretation of test scores. Chapter 10 provides an overview of the quality assurance (QA) processes which ensure all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1.1 THE HAWAI`I STATE ALTERNATE ASSESSMENTS (HSA-ALT)

The Hawai`i State Alternate Assessment (HSA-Alt) is made up of assessments based on the Hawai`i Common Core Standards (HCCS) and Next Generation Science Standards (NGSS) and is designed for students with the most significant cognitive disabilities. The purpose of the HSA-Alt is to maximize student access to the general education curriculum including the knowledge, skills, and abilities across the academic content standards, including the knowledge, skills, and abilities across the academic content standards, for students with the most significant cognitive disabilities, ensure that all students with disabilities are included in Hawai`i's statewide assessments, and make certain that they are included in the educational accountability system. Assessment results can inform instruction in the classroom by providing data that guide decision making. The HSA-Alt is only for those students with documented significant cognitive disabilities and adaptive behavior deficits who require extensive support across multiple settings (such as home, school, and community). Typically, this student population consists of about one percent of the total student population.

In 2018, the Hawai`i Department of Education (HIDOE) began the transition to a new online computer-adaptive test for alternate assessment for students with significant cognitive disabilities. The new assessment is designed to assess students at each achievement level (well below, approaches, meets, and exceeds) in grades 3–8 and 11 in ELA and mathematics, and in grades 5, 8, and 11 in science, as an online fully computer-adaptive test (CAT). Online operational field tests for ELA and Math was administered in spring 2019. A standard setting was convened in summer 2019 to set achievement standards for ELA and mathematics. Online operational field tests for science were administered in spring 2021. Achievement standards for science were set in summer 2021. The transition to computer-adaptive testing (CAT) was fully implemented in all grade levels and subject areas beginning in spring 2020. In spring 2023, each student was administered a 40-item operational test with up to 20 field-test items in all subject areas.

1.2 Alternate Assessment Eligibility

Most students with disabilities are able to participate in the general state assessments with appropriate state test accommodations. However, for students with the most significant cognitive disabilities, it may be more appropriate to participate in the alternate assessments. Decisions concerning a student's participation in statewide assessments are made by each student's individualized education program (IEP) team. Guidance for IEP teams to inform decisions about which assessment is most appropriate for each student is provided in the Participation Guidelines from the spring 2023 *Test Administration Manual* at https://hsa-alt.alohahsap.org/resources/resources-2022-2023/hsa-alt-test-administration-manuals-and-test-coordinators-manual-2022-2023.

The participation guidelines for Hawai'i's students to take HSA-Alt assessments are summarized as:

- The student demonstrates significant cognitive disabilities that may be combined with limited adaptive skills, physical, or behavioral limitations.
- The student requires a highly specialized educational program with intensive modifications and supports in order to access grade level academic standards.
- The student's daily instruction is substantively different from that of their peers without disabilities and requires extensive, repeated individualized instruction and support, across multiple settings. The student requires intensive direct instruction in multiple contexts to accomplish the acquisition, application, and transfer of knowledge and skills.
- The student's difficulty with the demands of the general academic curriculum is not due to social, cultural, or environmental factors; expectation of poor performance; or excessive absences.

1.3 Content Specifications and Blueprints

The September 2018 U.S. Department of Education *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process* clearly indicates that content standards must specify what students are expected to know and be able to do. Standards should include coherent and rigorous content and encourage the use of advanced teaching pedagogy and research-based instructional practices.

1.3.1. Content Specifications

The HSA-Alt is aligned to the content specifications for ELA, mathematics, and science, which are based on the Hawaii Common Core State Standards. These content specifications consist of Essence Statements,

which serve as the foundation for the development of HSA-Alt assessment items and are incorporated into Performance-Level Descriptors (PLDs) at four levels of complexity.

Essence Statements in ELA, mathematics, and science are broad skill, knowledge, and ability statements that guide the item-writing process for each content area and provide teachers with the specificity needed to translate the HCCS and the NGSS into meaningful learning targets for students with significant cognitive disabilities.

To develop Essence Statements, HIDOE and CAI staff reviewed the HCCS and the NGSS and prioritized content and skills that were deemed most critical in the development of successful post-secondary outcomes for students with significant cognitive disabilities. This process meets the requirements of both the Individuals with Disabilities Education Act (IDEA) and the Every Student Succeeds Act (ESSA) to link alternate assessments to grade-level content standards, with the understanding that alternate assessments may include skills at lower levels of complexity.

Essence Statements for the HSA-Alt are then incorporated into the Hawai`i Alternate Assessment Performance-Level Descriptors (PLDs) for ELA, mathematics, and science. PLDs have been developed at four levels of complexity for each Essence Statement:

- Exceeds Performance Level — Highest level of performance expectation for the alternate test
- Meets Performance Level — Meets performance expectation for the alternate test
- Approaches Performance Level — Approaches performance expectation for the alternate test
- Well-Below Performance Level — Well-below performance expectation for the alternate test

PLDs reflect different entry points into the grade-level state standards for students with significant cognitive disabilities and serve the following three purposes: 1) to assist teachers in providing access to the academic standards for students with significant cognitive disabilities, 2) to assist assessment personnel in developing test items that are accessible for students with a range of skill levels, and 3) to be used by standard-setting committees in conjunction with Essence Statements to craft the Just Barely Statements, which describe what a student just barely scoring at the bottom of each performance level knows and can do, and the Reporting PLDs, which detail grade- and content-area-specific descriptions of exactly what students performing throughout the range of each performance level know and can do.

Students participating in the HSA-Alt also have communication skills ranging from symbolic or abstract, to concrete, to pre-symbolic. Accommodations may be provided to allow students to perceive and respond to test items in meaningful ways.

1.3.2. Test Blueprints

Content specifications are operationalized in test administration through test blueprints which specify content standards to be covered and the number of items to be tested in each content standard. Test blueprints are composed of well-balanced content standards required by the state. Due to the unique characteristics of the student population taking the alternate assessment, depth-of-knowledge (DOK) is not specified in test blueprints.

The HSA-Alt test blueprints at the domain level for all subjects and grades are posted on the state portal website at https://hsa-alt.alohahsap.org/resources/resources-2020-2021/hsa-alt-ela-mathematics-and-

science-test-design-and-blueprints. Each student is required to take 40 operational items from domains specific to each test subject.

1.3.3. Test Forms

The HSA-Alt is delivered to each student through either an online adaptive test format or an online fixed-form test format, also referred to as the paper-pencil test administration. The online adaptive version is the primary format for most students, while the online fixed-form version is used as an accommodation format for students who cannot fully access the online adaptive test. For details, see Section 2.5 - Paper-Pencil Test Administration (via Online Fixed Form with Printed Response Option Cards).

The online adaptive test is delivered on the CAI's test delivery platform using the standard computer adaptive testing algorithm utilized by CAI for all adaptive testing programs. During an adaptive test session, forty operational items that meet the blueprint requirements at the domain level and match the student's ability are selected from the subject and grade specific operational item pool.

The online fixed form is assembled in advance and comprised of 40 fixed operational items selected from the same operational item pool as the online adaptive tests. The average difficulty of each grade level item bank is calculated prior to form-building. This becomes the target difficulty level for each grade level fixed form. The assessment blueprint and the target difficulty level are used when constructing each fixed form. In general, the items on the fixed forms are arranged from lowest difficulty to highest difficulty. Once the form is created, the blueprint is checked and the average difficulty of the form is checked. If the difficulty of the form is higher than that of the bank, then items on the fixed form are replaced with less difficult items and the average difficulty is calculated again. This process continues until either the form difficulty is similar to the bank difficulty, or there are no additional items in the bank that will adhere to the blueprint and move the averages closer together. Items are selected to meet blueprint in both the Early Stopping Rule segment (the first 8 items in the test) and the remaining segment of 32 operational items. Other factors considered in form development include the avoidance of key runs, avoidance of extremely easy or extremely difficult items, and limit the number of items with low biserial. Some additional parameters have been established for building the ELA fixed forms. First, if a passage appears in the ESR segment, that same passage will not appear in the remaining 32 operational item segment. Additionally, to reduce the reading load on the students, where possible, as many as three or four items linked to the same passage are placed, consecutively, on the assessment.

Students taking the fixed form in a specific subject and grade see the same set of operational items. Since the fixed form version of the test is used as an accommodation for students who cannot fully access the online test, it does not include any items with access limitations. The online fixed form satisfies the same blueprint requirements and is representative of the item pool with respect to item difficulty. Scores of students taking the fixed form are comparable to scores of students taking the adaptive tests.

1.4 Memorandum of Understanding (MOU) on Item-Sharing Initiative

In 2018, Hawai`i, South Carolina, and Wyoming signed a Memorandum of Understanding (MOU) on item sharing in item development and field testing. Each participating state contributed a predetermined number of items proportional to their state's student population for the alternate assessment. In early 2019, Idaho and Vermont joined the collaborative item development and field testing MOU and participated in the spring 2019 field test. In spring 2020, Montana and South Dakota joined the MOU for science assessments. In 2022, Vermont exited the MOU. Because the total number of students in alternate assessments is very small in each state, field testing common items in all MOU states allowed for the calibration of items based

on the combined data across all states. In addition to the MOU shared item pool, each state also developed some items that aligned to the state's specific content standards or content specifications.

The item-sharing initiative is designed to implement an item development process that generates at least three times the number of items needed for each test administration for each grade and subject. With 40 operational items on the test, at least 40 x 3 (or 120) calibrated items in the pool are needed for a computer-adaptive test (CAT). The item-sharing initiative allowed for this item development. Each MOU member would own the items they developed, but their items would be available for use by the other MOU members. The number of items developed by each state would be proportional to the size of the alternate assessment population that would participate in the test.

Each state in the MOU follows a similar process for developing and reviewing their items in collaboration with CAI (see Chapter 4 Item Development for further details). Items are developed by each state to fulfill their agreed upon contribution to the shared MOU field test pool each school year. DOE staff in each participating state are required by CAI to review the items contributed by their partner MOU states for field testing each school year, and provide a state-specific alignment to their own state's content standards at the shared grade level for each item. Following yearly field-testing and data review, DOE staff in each participating state make a final determination on whether shared items are accepted for operational use by confirming the state-specific content alignment for each item.

2. TEST ADMINISTRATION

The spring 2023 testing window was open from February 21 through May 26, 2023, for online adaptive operational tests, and from February 22 through May 19, 2023, for the online fixed-form operational test. The online adaptive operational tests were the default method of administration. The online fixed form paired paper response option cards and test visuals with the digital presentation of the stimuli and items. The online fixed form was provided as a special paper-pencil test form accommodation for students who were unable to fully access the online tests, even with the available accommodations. In paper-pencil tests, one test administrator (TA) administered the assessment to one student at a time. In the online format, the student took the assessment with the TA's assistance, as needed.

The online adaptive tests consisted of 40 operational items selected based on item difficulty and student ability to meet the assessment blueprint, with up to 20 embedded field-test items. The online fixed-form tests for paper-pencil administration followed the same test design as the online adaptive test, but were limited to 40 operational items presented in a fixed form that met each test blueprint.

The online adaptive tests consisted of three distinct test segments. Test segments are used by the Test Delivery System to implement the Early Stopping Rule, and to implement separate field testing of Alternate Assessment MOU shared field test items versus Hawaii-only field test items (items that are field tested only with Hawaii students). The test segments are defined as follows:

1. Segment 1 consists of 8 operational items presented in an adaptive format. This segment is used by the Test Delivery System to enforce the Early Stopping Rule. If all 8 items in Segment 1 are marked "No response" (or NR), the system will end the test when the TA attempts to move to Segment 2 (item 9).
2. Segment 2 consists of 32 operational items and 10 MOU shared field test items, presented in an adaptive format. Field test items are interspersed with the operational items starting with item position 1 within Segment 2 (or position 9 over the entire test) and ending with item position 37 (or position 45 over the entire test). Once a student completes the final item in Segment 2 (item 50), the student has officially completed the operational test.
3. Segment 3 consists of between 1-10- Hawaii-only field test items presented as a fixed form. The number of items in segment 3 differs by grade and subject area. In Spring 2023 only the following subject areas and grades had Segment 3 Hawaii-only field test items: ELA grades 6 and 8; Math grades 6, 8 and 11; and Science grades 5 and 8.

Fixed form tests for all grades and subjects included only Segments 1 and 2 to allow for implementation of the Early Stopping Rule. Segment 1 consists of 8 operational items as a fixed form; segment 2 consists of the remaining 32 operational items as a fixed form. No MOU or Hawaii-only field test items are included on the Fixed Form tests.

2.1 TEST ADMINISTRATOR TRAINING

Test administrator training is critical in producing reliable and valid test scores. Comparability of test scores, whether between students and schools or across time for the same students, is based on standardization of test administration and test scoring rules. If TAs do not follow the same procedures, student performance cannot be compared meaningfully. HIDOE requires HSA-Alt Test Administrators to attend a yearly department-led test administrator training to ensure compliance with testing policies. Following the department-led training, all HSA-Alt Test Administrators are also required to complete the

online TA Certification Course, available via the Cambium system, before the online Test Delivery System allows the TA access to the TA Live Site to administer a test to students.

In January 2023, a series of in-person training sessions for the HSA-Alt 2022–2023 administration occurred at seven locations across the state, including training on the following topics:

- HSA-Alt Participation Guidelines
- HSA-Alt Essence Statements and Range Performance-Level Descriptors
- HSA-Alt Test Design
    - Early Stopping Rule
- HSA-Alt Universal Tools, Designated Supports, and Accommodations
    - Documentation and Verification
    - Guidelines for Translated Test Designated Support
    - Guidelines for Read-Aloud, Scribe, and Descriptions of Visuals Accommodations and the Translated Test Designated Support
- HSA-Alt Test Administration
    - Learner Characteristics Inventory
    - Hawai`i Observational Rating Assessment
    - Paper Form Administration
    - HSA-Alt Code of Ethics

At the end of each full-day training session, TAs were asked to evaluate the training session and provide feedback on ways to improve future training sessions. HIDOE uses this feedback to revise training materials, time allocation for the training, mode(s) of training to be used in the future and identifying areas where additional support for test administrators needs to be provided. In addition, all TAs needed to complete the online HSA-Alt TA Certification Course before being provided access to the live test site for live testing.

The online HSA-Alt TA Certification Course reviews the information provided during the yearly mandatory in-person or virtual training and requires, as a final step, TAs to affirm that they will uphold the HSA-Alt Code of Ethics. The specific responsibilities delineated in the HSA-Alt Code of Ethics are found below:

HSA-Alt Code of Ethics
Exhibit the highest degree of professional ethics.
• Plan for and include IEP-aligned accessibility supports during testing, including consideration of a student's familiar communication system. Students must receive all accommodations listed for state summative testing in their IEP during HSA-Alt testing.
Provide HSA-Alt students with online training test opportunities prior to testing. • Demonstrate tool use: the ear icon for reading and re-reading, as needed, the passage, question, and answer options, the double-headed arrow for expanding/collapsing the split screen to

HSA-Alt Code of Ethics
view/hide the full visual, and the "Next" arrow for finalizing answer selections and moving forward in the assessment. • Consider modeling metacognitive test-taking strategies for students: talking through the solution process, using scratch paper, concrete materials, or tools such as a calculator, eliminating one answer option, etc.
Follow all test security and test administration procedures: including the close supervision of all students during HSA-Alt testing to ensure that students receive: • The full audio delivery of stimulus, question, and answer options, • The expanded view of math and science visuals, and • Sufficient wait time and presentation repetition to maximize the elicitation of student response.

TAs who were unable to attend an in-person training session were required to complete the online certification course.

2.2 Test Administration Manuals

The 2023 *Test Administration Manual* (TAM) summarizes the HSA-Alt and provides guidelines for test administration. It includes the following:

- Overview of the background, purpose, and content specifications for HSA-Alt
- Assessment design
- Student inclusion and participation guidelines
- TA requirements
- Test delivery modes: online or online with fixed-form paper-pencil response cards and test visuals as a special accommodation
- Test administration procedures
- Test security guidelines

The TAM can be found at https://hsa-alt.alohahsap.org/resources/resources-2022-2023/hsa-alt-test-administration-manuals-and-test-coordinators-manual-2022-2023. .

For the convenience of TAs, specific directions are documented for the online system for adaptive and fixed-form test administration. The directions for online test administration can be found at https://hsa-alt.alohahsap.org/resources/resources-2022-2023/guide-to-navigating-the-online-hsa-alt-administration-and-quick-start-guide-2022-2023. .

A short guide for the use of paper response option cards and printed test visuals for students approved for the paper-pencil test accommodation were provided to TAs who administered the fixed-form tests to approved students. This guide can be found at https://hsa-alt.alohahsap.org/resources/resources-2022-2023/hsa-alt-spring-2023-instructions-for-use-of-printed-response-option-cards. .

There is no time limit besides the dates of the testing window for administering the HSA-Alt. If a student becomes fatigued, the TA can pause the assessment and restart it later within the testing window. Tests that are resumed, start up at the same point from where they were paused.

2.3 ACCOMMODATIONS

The HSA-Alt was designed following universal design principles that incorporate supports that a student might need to access the assessment (e.g., picture arrays, oral reading of passages, the use of a student's own receptive and expressive communication methods). The allowable accommodations listed in the next section provide students the opportunity to gain access to the items and make a response.

2.3.1. Allowable Accommodations

For the online and paper-pencil version (via online fixed form with printed response option cards), all items may be read and reread by the audio playback function in the Online Testing System. All items may further be orally presented after the teacher uses the online digital interface to present the test item the first time. Testing for either test form is not timed, may be completed over multiple sessions, and can stop at any point within the test form, as needed.

A variety of universal tools are available for the HSA-Alt assessment. A list of universal tools that are available is provided in the tables below. This list of universal tools is by no means exhaustive as students with significant cognitive disabilities vary widely in the type and amount of supports they may require. The list of universal tools found below contains examples of only some of the supports that a student who takes the HSA-Alt may need in order to demonstrate understanding. The same level of support needs to provided during the alternate assessment as are provided during customary classroom instruction. For example, if the students use the zoom when using computer devices, the same level of the zoom needs to be set for the students. If the students utilize the certain types of Graphic Organizers, the same types of Graphic Organizers needs to be used when administering the HSA-Alt.

Table 1. List of Available Universal Tools

Universal Tools	Description
Adjust the volume for listening passages (summative assessments)	All students can adjust the volume on their devices and/or headphones for the listening passages.
Adjusted visual or tactile field	Test administration display items or devices can be positioned to place the display and/or response options within the student's optimal field of vision and/or reach.
Altered setting	Provide for reduction in lighting, environmental sound or noise, visual stimuli or other features of the setting for students who are subject to sensory overstimulation. Provide for adaptive or special furniture or equipment for students who require it.

Universal Tools	Description
Audio Playback (summative assessments)	Text on summative assessment items is read aloud to the student via embedded audio files that includes audio playback of all items, passages/stimuli, and response options. Although test administration is designed primarily for one-to-one testing, some students who are able to navigate the test delivery system, independently, may be able to be tested in a small group setting. Therefore, these students need to either use headphones or be tested in a separate setting (see Separate Setting).
Breaks	Breaks may be given as often as necessary at the discretion of the test administrator to reduce cognitive fatigue when students experience heavy assessment demands.
Calculator (Embedded)	All students may access the online Desmos basic calculator tool available in the HSA-Alt mathematics tests.
Calculators (Hand-held)	Students who use a calculator during instruction may use the calculator during the administration of the assessment.
Color overlays (paper/pencil form only)	Color transparencies are placed over the paper-based answer option cards. This support also may be needed by some students with visual impairments or other print disabilities. Choice of color should be informed by evidence of those colors that meet the student's needs.
Expandable Passages and Stimuli	This tool provides a streamlined interface of the test stimulus window allowing items to be displayed full-screen. It is one of only three universal tools that can be set in TIDE; the default position for this tool in TIDE is *ON*.
Fidget tool	Allow/encourage movement and/or allow unrelated manipulative (e.g., fidget tools, rubber bands) in free hand to aid concentration. This tool may require a separate setting.
Graphic Organizers	Customary frames for organizing information used in language arts instruction such as: character, event, or story map; problem/solution, cause and effect, and sequence chain.
Highlight text	Highlight text with flashlight, pointer, highlight marker, or other means of focusing student's attention to the response options. Focusing attention must not prompt the student to the correct answer.
Magnification	Magnification allows increasing the size to a level not provided for by the embedded Zoom universal tool. This may include projection if testing is carried out in a separate setting. It may also include the use of a magnifying lens overlay.
Masking (paper/pencil form only)	Masking involves blocking off content on the paper answer option cards that is not of immediate need or that may be distracting to the student. Students are able to focus their attention on a specific part of the answer option card by masking.

Universal Tools	Description
No Response	If no response is indicated or recorded by the student, the TA will need to access the context menu for the item and select the "No Response" option for that item. This will mark the item as a "No Response" and the TA will be able to advance to the next test item for administration.
Noise Buffers	Ear mufflers, white noise, and/or other equipment used to block external sounds.
Refocusing prompts or gestures	TA may provide intermittent visual, tactile, physical, or auditory prompts for the purpose of refocusing the student's attention to the task at hand. The prompts must not provide any cues as to the correct response.
Repetition	Students may have all parts of an item presented to them as many times as necessary, including passages/stimuli, question stem, and response options; however, once the "Next" button is pressed, no item shall be redelivered. Hawai'i Department of Education HSA-Alt testing policies require students and Test Administrators to move on to the next item once the "Next" button is pressed. Students and Test Administrators shall not navigate back to earlier items in the assessment. Whatever answer was registered into the system when the "Next" button is pressed shall be the student's final answer. No test item should be re-presented and no student response should be changed after the "Next" button is pressed. Although this functionality is available, students and Test Administrators are required not to use it during HSA-Alt summative test administrations.
Scratch paper	Scratch paper to make notes, write computations, or record responses may be made available. Assistive technology devices, including low-tech assistive technology (Math Window), are permitted to make notes. The assistive technology device needs to be consistent with the student's IEP or 504 plan. Access to the Internet must be disabled on assistive technology devices. All scratch paper must be collected and disposed of at the end of each test session to maintain test security. Digital notes entered into an assistive device, if used, need to be deleted.
Separate Setting	Test location is altered so that the student is tested in a setting different from that made available for most students. The HSA-Alt is designed to be primarily administered in a one-to-one setting. Students who are easily distracted in the regular classroom setting, may need an alternate location to be able to take the assessment. Digitally delivered human voice recording (HVR) audio is a universal tool for these assessments, therefore students need to either use headphones or be tested in a separate setting. Allow students time to become familiar with the new testing location.
Suppress Score	Student test results are not shown on screen at the end of the test; for the HSA-Alt the default position for this universal tool is *OFF* with student results automatically shown on screen when the test is submitted.

Universal Tools	Description
Timing or Scheduling	Students can be tested during their optimal time of day. Scheduling should account for a student who requires frequent breaks and rest periods, over an extended time period.
Translated test directions	Students who have limited English language skills can receive test directions in another language if this support is provided by a bi-literate adult trained in the administration of the HSA-Alt.
Zoom	Students may make test questions, text, or graphics larger by clicking on the Zoom icon that has four levels of magnification; for the HSA-Alt the default position for this universal tool is *Level 1*.

For the spring 2023 HSA-ALT administration, there is one designated support, Translated Test, available for the HSA-Alt assessment. The Translated Test designated support allows a translator to provide full translation of all parts of the mathematics and science alternate tests. Translators are required to follow the specific guidelines found in Table 2 and must acknowledge understanding of these guidelines prior to testing by signing and submitting the *HSA-Alt Test Security and Confidentiality Form* to the school test coordinator who will then submit the form to the Assessment Section. For a description of the Translated Test designated support, see Table 2. Please note that the Translated Test designated support also requires the submittal and approval of the paper-pencil accommodation for a student.

Table 2. List of Available Designated Supports

Designated Supports	Description
Translated Test	Students who have limited English language skills and who use dual language supports in the classroom may have the **mathematics and science assessments** translated during alternate testing. Translation of the English Language Arts (ELA) assessment is not allowed. The translator must be a bi-literate adult trained in the administration of the HSA-Alt. Translators may translate the test directions, test items, and response options for these assessments. They must provide a full translation not deviating from the presented stimulus, item, and audio script. All translators must sign the *HSA-Alt Test Security and Confidentiality Form* found in Appendix M. The paper/pencil test kit is also required for the administration of a translated test. The use of a translator may result in the student needing additional overall time to complete the assessment.

Accommodations for the HSA-Alt need to be set in the Test Information Distribution Engine (TIDE) by the TA. The only accommodation requiring state approval and form submittal is the Paper-Pencil Test accommodation. In the TAM and during Test Administrator training, TAs are reminded of the importance of reviewing the student's IEP and accessibility supports available for HSA-Alt summative assessment to determine the most appropriate accessibility supports for the statewide assessment. TA training addressed the documenting of all accommodations and designated supports in the student IEP record. Test administration guidelines and the HSA-Alt Code of Ethics establish the requirement that students receive all accommodations listed in the student IEP during summative testing. Accommodations that were available for the HSA-Alt in spring 2023 are listed below.

Table 3. List of Available Accommodations

Accommodation	Description
Alternate Response Options	Students taking the HSA-Alt with TA assistance may respond using the mode of communication that they use during instruction. These response modes include, but are not limited to, an oral response, pointing, eye gaze, a response card, sign language, switches, or an augmentative communication device. Once the student has communicated a response, the TA may enter the student's response into the system. Consistent criteria must be used as the basis for student responses; i.e., TA cannot take orally provided answer on the first item and then switch response on the next.
American Sign Language (non-embedded)	Students who cannot hear the audio for the assessment, may have their TA repeat the audio script using American Sign Language. TAs must take care to precisely follow the audio script that is provided for each test item component: passage, stimulus, question, and answer option card descriptions.
Calculator	Students who have calculator use documented in their IEP and who regularly use a calculator during instruction must have the calculator available to them during the administration of the assessment. The difference between the Calculator Accommodation and the Calculator Universal Tool is the Calculator Accommodation is specifically listed as an accommodation in the student IEP.
Concrete materials	Students are provided with the customary concrete materials that are used for daily math instruction and assessment. These materials may include but are not limited to: base-10 blocks, counters, open number lines, pattern blocks, unifix cubes, etc. For the paper-and-pencil form concrete materials may also be substituted for response cards, if the presented objects are uniform in size and color and do not cue the student to the correct answer.

Accommodation	Description
Digital Math Manipulatives	Students are provided access to the virtual platform with digital math manipulatives such as unifix cubes, ten frames, fraction tiles, and number lines to use during the math assessment. Teachers may support in selecting the math manipulative the student selects for a presented problem. Teachers may not manipulate the digital math manipulatives for a student.
Multiplication Table	Students who need a multiplication table to solve math problems and who consistently use the table during instruction and assessment of math, may use a multiplication table on the assessment.
Paper/Pencil Test (summative assessments)	Some students with disabilities, such as visual impairment or blindness, and alternate-identified EL students who need language support may be better able to access the assessment with the paper/pencil version of the HSA-Alt. For students with sight limitations, the paper/pencil test version allows the teacher or test administrator to prepare tactilely-enhanced versions of the test visuals and answer options. For EL students who require the Translated Test Designated Support, the paper/pencil test form allows the test translator to preview and prepare full translations of the math and science assessments prior to test administration. If a student's IEP care coordinator determines a student would be best served by the paper/pencil version of the HSA-Alt, due to his or her specific needs, the student's Test Administrator will need to contact the school's Test Coordinator to order the paper/pencil test kit.

Accommodation	Description
Read Aloud (summative assessments)	The Read Aloud accommodation may be needed during the summative assessment for students who require a slower audio delivery speed than is currently available via the online platform. If this accommodation is provided to a student, the in-test audio must first be played for the student via the Test Delivery System with the TA listening carefully to the script as it is read aloud. The TA may then carefully reread or restate the passage, question, and/or answer option(s) exactly as read aloud by the in-test audio. TAs must not make any changes, additions or deletions, intonation, or emphases that might inadvertently lead a student to the correct response. All TAs who deliver the Read Aloud Accommodation during testing must follow the *HSA-Alt Guidelines for Read Aloud, Test Reader*. These guidelines can be found in Appendix D in this manual. After reading these guidelines TAs will need to complete and sign the *HSA-Alt Test Security and Confidentiality Form* found in Appendix M. This form upon completion should be given to the school's TC who will then submit the form to the Assessment Section. The Read Aloud accommodation is not required for the optional HSA-Alt Classroom Embedded Assessments (CEAs) because the CEAs, by design, have the teacher read all items to or with the student.
Reinforcement System	Students who receive a positive reinforcement system on a daily basis should receive this same support during summative testing. Reinforcement system support use must be documented in the IEP. Document this support in the Supplementary Aides and Services section on the Services page. (Follow student's Behavior Intervention Plan or Behavior Support Plan.) Positive reinforcement can be provided for continuing to focus and progress through the test <u>not</u> for correctly answering items.

Accommodation	Description
Scribe	Students either indicate their response or do not respond to a test item and the Test Administrator then enters a **[No Response]** or the student's indicated response into the data entry interface. Responses must be entered as directly observed or represented verbatim. If a TA anticipates that their student will be non-responsive during testing the Scribe accommodation should be requested so that the [**No Response**] option may be entered by the TA for items to which the student is non-responsive. The TA must follow the *HSA-Alt Scribing Protocol*. These guidelines can be found in Appendix E in this manual. After reading these guidelines TAs will need to complete and sign the *HSA-Alt Test Security and Confidentiality Form* found in Appendix M. This form upon completion should be given to the school's TC who will then submit the form to the Assessment Section.
Tactile sensitivity (paper/pencil form only)	Students are provided with tactilely enhanced visuals or answer options or analogous response options with enhanced/reduced features so as to increase access to test visuals and answer options, and/or to address specific tactile sensitivity: slippery, fuzzy, rough, etc.
Visual Descriptions	Students who are visually impaired may require TA description of charts and graphs in order to access the assessment materials. Descriptions provided must not cue students to the correct answer. Those TAs providing their students with a visual description of charts and graphs must follow the *HSA-Alt Visual Descriptions Protocol* found in Appendix F in this manual. After reading these guidelines TAs will need to complete and sign the *HSA-Alt Test Security and Confidentiality Form* found in Appendix M. This form upon completion should be given to the school's TC who will then submit the form to the Assessment Section.

Table 4, Table 5, and Table 6 present the number of students who were assigned specific accommodations in the 2022–2023 administration.

Table 4. Total Number of Students with Allowed Accommodations: ELA

Accommodations	Grade						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Alternate Response Options	5	5	4	2	4	4	4
American Sign Language (Non-Embedded)	0	0	0	0	2	1	1
Concrete Materials	2	8	5	3	11	9	3
Digital Math Manipulatives	2	4	0	0	1	0	0
Paper/Pencil Test	1	4	2	3	2	2	0
Read Aloud Stimuli	6	7	11	2	9	6	5
Reinforcement System	6	5	8	4	11	7	5
Scribe Items	6	9	9	4	15	10	8
Tactile Sensitivity	1	1	0	0	1	0	0
Visual Descriptions	2	3	2	1	1	3	3

Table 5. Total Number of Students with Allowed Accommodations: Mathematics

Accommodations	Grade						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Alternate Response Options	5	5	4	2	4	4	3
American Sign Language (Non-Embedded)	0	0	0	0	1	0	1
Calculator	1	0	1	0	4	3	6
Concrete Materials	2	8	5	3	11	9	2
Digital Math Manipulatives	2	4	0	0	1	0	0
Multiplication Table	1	5	1	0	4	1	0
Paper/Pencil Test	1	4	2	3	2	2	0
Read Aloud Stimuli	6	7	11	2	9	5	5
Reinforcement System	6	5	8	4	11	7	4
Scribe Items	6	11	9	4	15	11	7
Tactile Sensitivity	1	1	0	0	1	0	0
Visual Descriptions	2	3	2	1	1	3	3

Table 6. Total Number of Students with Allowed Accommodations: Science

Accommodations	Grade		
	5	**8**	**11**
Alternate Response Options	4	4	3
American Sign Language (Non-Embedded)	0	0	1
Calculator	1	3	6
Concrete Materials	5	9	2
Multiplication Table	1	1	0
Paper/Pencil Test	2	2	0
Read Aloud Stimuli	11	6	5
Reinforcement System	7	7	4
Scribe Items	9	10	6
Visual Descriptions	2	3	3

2.3.2. Stimulus and Response: Substitutions

The stimulus materials identified in each alternate assessment item are intended for students who have significant cognitive disabilities. In recognition of the need to occasionally depart from the standard stimulus and response materials, Table 7 shows suggested substitutions and alternatives that are based on the student's degree of vision, hearing, or physical mobility.

Table 7. Suggested Substitutions and Alternatives

Student Characteristic	The TA can adapt stimulus/response materials by doing the following:
Limited in reach or touch	Use iPad (or other device) in conjunction with switches or other assistive technology.
Limited in visual or tactile field	Position the iPad (or other device) level with the student's eyes and then move within the student's reach.
Apraxia/motor planning problems or sensory integration challenges	Rehearse movement needed for response; use an object for pointing; provide tactile and kinesthetic supports (e.g., pacing board).
	Provide frequent breaks; offer visual supports; allow/encourage movement; allow unrelated manipulative (e.g., rubber band in free hand) to aid concentration, supported seating, weighted vests, sensory diet before testing; reduce "noise" such as environmental sound, tactile and olfactory input, light.
Orthopedic impairment	Use assistive technology, visual cues, gestures (e.g., point to materials); change location to increase physical access; change location to access special equipment; offer adjustable-height desk, appropriate specialized seating, slant-top surface, assistive technology, extended time, multiple or frequent breaks.

2.3.3. Assistive Technology

Assistive technology (AT) that is documented in the student's IEP and used during regular instruction may be used to assist the student to access the HSA-Alt through the online TDS. Technology affords many ways to adapt student response on an iPad or computer. Any AT that does not unfairly provide advantage or disadvantage to a student may be used, including, but not limited to, the following:

- Screen magnifier or screen magnification software
- Arm support
- Mouth stick, head pointer with standard or alternative keyboard
- Voice output device, both single and multiple message
- Tactile/voice output measuring devices (e.g., clock, ruler)
- Overhead projector or whiteboard

Students who are eligible will take the HSA-Alt and will be able to access the assessment using the digital interface when provided the allowable supports. However, it is recognized that students with certain disabilities will still require access using the paper-pencil test version of the assessment.

Some students with disabilities may be better able to access the assessment with the paper-pencil version of the HSA-Alt. If a student's IEP care coordinator determines the student requires the paper-pencil version of the HSA-Alt, due to the nature of his or her disability or disabilities, the student's test administrator will need to contact the school's test coordinator. The school's test coordinator is responsible for submitting the paper-pencil accommodation verification request and submitting the paper-pencil test kit request form.

2.4 ONLINE ADMINISTRATION

Before Student Testing

For each student who took the online alternate assessment, the student's teacher completed the Learner Characteristic Inventory (LCI) and the Hawai`i Observational Rating Assessment (HIORA) surveys. These teacher surveys were completed before the students took any content-area tests. On the surveys, teachers provided student ratings based upon their perception of the student's characteristics, knowledge, skills, abilities, and transition readiness. While the LCI is a national standardized inventory, the HIORA is a Hawai`i-specific add-on to this. Hawai`i uses the LCI to gather information about alternate-identified students' characteristics in the state. The HIORA was created to gather additional information from the teacher on the student's understanding of grade-level content in each subject (ELA, math, and science) and the student's readiness for transition. Hawai`i instituted the HIORA content ratings of performance in 2018–2019 and the ratings of transition readiness in 2021-2022. The HIORA is grade-specific and references the tiered performance expectations found in the HSA-Alt Range Performance-Level Descriptors (PLDs) and the National Technical Assistance Center on Transition (NTACT) Success Predictors. The LCI and HIORA are completed by the student's teacher for each student.

During Student Testing

During test administration, the student or TA touches the button bearing an ear icon for the stimulus, question, and response option portion of each item to be read aloud. The read aloud script is a recorded human voice. The speed of narration was comparable to the average speed of narration when teachers read to students. Students responded to each item by clicking on one of the response options presented, or the TA could click on the student's selected response option for them. Students could change their answer selection as needed, however, once the *Next* button was selected, the assessment moved on to the next item. The online system automatically stored item responses when students touched on their selected-response option and then clicked or tapped the "Next" button.

For all test items, if no response was indicated or recorded by the student, the TA accessed the context menu for the item and select the "No Response" option. This marked the item as a "No Response" item, and the TA was able to advance to the next test item for administration.

In spring 2023, an Early Stopping Rule was available for students who were non-responsive to the first eight items on each content-area test. Students and TAs were required to follow the administration guidelines put in place by the HIDOE Assessment Section. The Early Stopping Rule was instituted for a student's test if all of the following conditions were met:

1. The student did not respond to the first eight items in the assessment.
2. The eight items were administered across two different sessions on two different days.

3. The "No Response" option was selected for the student by the TA using the context menu for each of the eight items.
4. The TA confirmed that the student was provided with sufficient response time and appropriate communication and accessibility supports during testing.
5. The required Test Session Observer (someone other than the TA) verified that they were present during testing and did not observe the student respond to the questions that they were presented, and that the TA administered the assessment with fidelity. The Test Session Observer was required to be present for a minimum of four of the eight questions in a content area.

When the first three conditions are met, the online TDS automatically stop the student's test. The TA and the Test Session Observer were then required to complete conditions 4 and 5 by submitting the signed *Early Stopping Rule Verification Form*. This form was submitted by fax or email to the Assessment Section.

2.5 PAPER-PENCIL TEST ADMINISTRATION (VIA ONLINE FIXED FORM WITH PRINTED RESPONSE OPTION CARDS)

In spring 2023, students who required a paper-pencil accommodation were administered a fixed-form test via the Online Testing System alongside printed response option cards and test visuals, which the TA placed in front of the student while the student listened to the human voice recording via the Online Testing System. Test administrators completed and submitted the Learner Characteristic Inventory (LCI), which investigates the learning characteristics of students participating in alternate assessments based on alternate achievement standards, and the Hawai`i Observational Rating Assessment (HIORA), a grade-level aligned evaluation of student knowledge and skills in ELA, mathematics, and science and an appraisal of student readiness for transition. During administration, the student's item responses were entered directly by the TA into the Online Testing System after the student indicated their response option via the printed response option cards. No access limited items were included on the fixed form tests for paper-pencil administration. The number of students who received the fixed form test in spring 2023 can be found in Table 8.

2.6 TEST SECURITY

The Test Security Guidelines, embedded in the *Test Administration Manual*, indicated that photocopying any printed testing materials was strictly prohibited. Printed response cards and printed test visuals are secure materials. School test coordinators were responsible for receiving, accounting for, and returning all test materials to CAI. If CAI did not receive the returned test materials within the scheduled time frame, CAI would make enough effort to be sure that all secure materials were returned. Any known violations of test security were to be reported immediately.

2.6.1. Student-Level Testing Confidentiality

The online adaptive and fixed-forms tests are administered through secure websites. All of the secure websites enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA) and other federal laws. Secure transmission and password-protected access are the basic features of the current system and ensure authorized data access. All aspects of the system, including item development and review, test delivery, and reporting, are secured by password-protected logins. The systems use role-based security models that ensure that users may access only the data to which they are entitled and may edit data only in accordance with their user rights.

FERPA prohibits public disclosure of student information or test results. To comply with the secure standards, student names and IDs are communicated via a secure file transfer system. Student login information is associated with particular tests they are assigned. If information must be sent via email or fax, only the SSID number, not the student's name, is included. A student cannot take a test under another student's ID.

Student login information is entered only at the beginning of a test, after an authorized TA creates and manages the test session, and the TA reviews and approves a test (and its settings) for the student. Only authorized users can make changes to the test registration system. Test materials and reports are carefully protected so that student names and test results cannot be identified and accessed by unauthorized individuals.

All test takers, including home-schooled students, must be enrolled or registered at their testing schools in order to take the online or paper-pencil tests. Student enrollment information, including demographic data, is generated by the Hawai`i Department of Education (HIDOE) and uploaded nightly via a secured file transfer site to the online testing system.

Only staff with the administrative roles of complex area superintendent (CAS), complex staff (CS), school-level test coordinator (TC), teacher (TE), or HIDOE staff can view students' scores. CASs and CSs have access to all scores within their district. TCs have access to all scores within their school. TEs have access to scores within their classrooms. Parents receive ONLY a printed copy of their children's online score reports if the school or teacher provides one.

2.6.2. System Security

The objective of system security is to ensure that all data are protected and accessed correctly by the appropriate user groups. It is about protecting data and maintaining data and system integrity, as intended, including ensuring that all personal information is secured, that transferred data (whether sent or received) is not altered in any way, that the data source is known, and that any service can be performed ONLY by a specific, designated user.

Password Protection: All access points by different roles—at the state, complex area, complex, school principal, and school staff levels—require a password to log in to the system. Newly added TCs and TAs receive separate passwords through their personal email addresses assigned by the school. All new users receive updated passwords on a yearly basis.

Secure Browser: A role of the Technology Coordinator is to ensure that the CAI Secure Browser is properly installed on the testing device (iPads, Chromebooks, or other devices) used for the administration of the online assessments. Developed by the testing contractor, the Secure Browser prevents students from accessing other computers or Internet applications and from copying test information. It suppresses access to commonly used browsers such as Chrome and Firefox and prevents students from searching for answers on the Internet or communicating with other students. Assessments can be accessed only through the Secure Browser and not through other Internet browsers.

Testing personnel are reminded in the online training and user manuals that assessments should be administered in an appropriate environment.

2.7 PREVENTION OF AND RECOVERY FROM DISRUPTIONS IN TEST DELIVERY SYSTEM

CAI is continuously improving our ability to protect our systems from interruptions. CAI's Test Delivery System (TDS) is designed to ensure that student responses are captured accurately and stored on more than one server in case of a failure. Our architecture, described here, is designed to recover from a failure of any component with little interruption. Each system is redundant, and crucial student response data are transferred to a different data center each night.

CAI has developed a unique monitoring system that is sensitive to changes in server performance. Most monitoring systems provide warnings when something is going wrong. In addition to general warnings of malfunction, our monitoring system also provides warnings when any given server is performing differently from its performance over the few hours prior, or differently than the other servers performing the same jobs. Subtle changes in performance often precede actual failure by hours or days, allowing us to detect potential problems, investigate them, and mitigate them before a failure. On multiple occasions, this has enabled us to make adjustments and replace equipment before any problems occurred.

CAI has also implemented an escalation procedure that enables us to alert clients within minutes of any disruption. Our emergency alert system notifies by text message our executive and technical staff, who then immediately join a call to understand the problem.

The next section describes CAI system architecture and how it recovers from device failures, Internet interruptions, and other problems.

2.7.1. High-Level System Architecture

CAI system architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stakes testing program. The general approach is pragmatic and well supported by the architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. The CAI system is designed to ensure that the testing results and experience can respond robustly to such inevitable failures. Thus, CAI's TDS is designed to protect data integrity and prevent student data loss at every point in the process.

Key elements of the testing system, including the data integrity processes at work at each point in the system, are described in the paragraphs that follow. Fault tolerance and automated recovery are built into every component of the system.

Student Machine

Student responses are conveyed to our servers in real time, as students respond. Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually 30–90 seconds), the system will prevent the student from doing any more work until connectivity is restored. The student is offered the choice of asking the system to try again or pausing the test and returning later. For example:

- If connectivity is lost and restored within the designated time period, the student may be unaware of the momentary interruption.
- If connectivity cannot be silently restored, the student is prevented from testing and given the option of logging out or retrying to save.

- If the system fails completely, upon logging back in to the system, the student returns to the item at which the failure occurred.

In short, data integrity is preserved by confirmed saves to our servers and the prevention of further testing if confirmation is not received.

Test Delivery Satellites

The test delivery satellites communicate with the student machines to deliver items and receive responses. Each satellite is a collection of web and database servers. Each satellite is equipped with a redundant array of independent disks (RAID) system to mitigate the risk of disk failure. Each response is stored on multiple independent disks.

One server serves as a backup hub for every four satellites. This server continually monitors and stores all changed student response data from the satellites, creating an additional copy of the real-time data. In the unlikely event of failure, data are completely protected. Satellites are automatically monitored, and upon failure, they are removed from service. Real-time student data are immediately recoverable from the satellite, backup hub, or hub (described in this section), with backup copies remaining on the drive arrays of the disabled satellite.

If a satellite fails, students will exit the system. The automatic recovery system enables them to log in again within seconds or minutes of the failure without data loss. This process is managed by the hub. Data will remain on the satellites until the satellite receives notice from the demographic and history servers that the data are safely stored on those disks.

Hub

Hub servers are redundant clusters of database servers with RAID drive systems. Hub servers continuously gather data from the test delivery satellites and their mini-hubs and store that data as described earlier. This real-time backup copy remains on the hub until the hub receives a notification from the demographic and history servers that the data have reached the designated storage location.

Demographic and History Servers

The demographic and history servers store student data for the duration of the testing window. They are clustered database servers, also with RAID subsystems, providing redundant capability to prevent data loss in the event of server or disk failure. At the normal conclusion of a test, these servers receive completed tests from the test delivery satellites. Upon successful completion of the storage of information, these servers notify the hub and satellites that it is safe to delete student data.

Quality Assurance System

The quality assurance (QA) system gathers data, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QA system, and any anomalies (such as unscored or missing items, unexpected test lengths, or other unlikely issues) are flagged, and a notification immediately goes out to our psychometricians and project team.

Database of Record

The Database of Record (DOR) is the final storage location for student data. These clustered database servers with RAID systems hold the completed student results.

2.7.2. Automated Backup and Recovery

Every system is backed up nightly. Industry-standard backup and recovery procedures are in place to ensure the safety, security, and integrity of all data. This set of systems and processes is designed to provide complete data integrity and prevent the loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered real-time backup processes prevent the loss of student data, even in the unlikely event of system failure.

2.7.3. Other Disruption Prevention and Recovery

These testing systems are designed to be extremely fault tolerant. The system can withstand failure of any component with little to no interruption. This robustness is achieved through redundancy. Key redundant systems include the following:

- The system's hosting provider has redundant power generators that can continue to operate for up to 60 hours without refueling. With multiple refueling contracts in place, these generators can operate indefinitely.
- The hosting provider has multiple redundancies in the flow of information to and from our data centers by partnering with nine different network providers. Each fiber carrier must enter the data center at separate physical points, protecting the data center from a complete service failure caused by an unlikely network cable cut.
- On the network level, we have redundant firewalls and load balancers throughout the environment.
- The system uses redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. Should a catastrophic event occur, CAI is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or needs to be rerun.

CAI's TDS is hosted in an industry-leading facility, with redundant power, cooling, state-of-the-art security, and other features that protect the system from failure. The system itself is redundant at every component, and the unique design ensures that in the event of failure, data are always stored in at least two locations. The engineering that led to this system protects student responses from loss.

3. SUMMARY OF SPRING 2023 OPERATIONAL TEST ADMINISTRATION

3.1 STUDENT PARTICIPATION

The HSA-Alt was administered by subject and grade level. All students in grades 3–8 and 11 were assessed in English language arts (ELA) and mathematics. Students in grades 5, 8, and 11 were also assessed in science. For a test to be considered participated, or attempted for scoring, a student must respond to at least one item, or a "No Response" was recorded by the TA to at least one item.

Table 8 presents the total number of students who attempted the online adaptive and online fixed-form HSA-Alt tests by subject and grade. Table 9 presents the alternate assessment participation rate, computed as the number of students taking the HSA-Alt divided by the total number of students in the state taking the general education summative tests and the HSA-Alt. Table 10 presents the total number and percentage of students who participated in the HSA-Alt by subgroup. Table 11 presents the total number and percentage of students who participated in the HSA-Alt in each Individuals with Disabilities Education Act (IDEA) disability category, and by subgroup. Table 12 through Table 15 provide the total number of students who participated in the HSA-Alt by subgroup and IDEA category for each grade.

Table 8. Number of Attempted Students

Subject	Grade	Online Adaptive				Online Fixed-Form				Total
		Completed	ESR*	Incomplete	Total	Completed	ESR*	Incomplete	Total	
ELA	3	111	10	3	124	1			1	125
	4	114	11	2	127	1	1		2	129
	5	108	7	5	120	1			1	121
	6	133	10	5	148	1			1	149
	7	107	6	5	118					118
	8	88	3	1	92	1			1	93
	11	87	2	6	95	1			1	96
Math	3	111	10	2	123	1			1	124
	4	113	10	2	125	1	1		2	127
	5	109	9	3	121	1			1	122
	6	132	12	4	148	1			1	149
	7	108	6	3	117					117
	8	87	4	1	92	1			1	93
	11	88	1	3	92	1			1	93
Science	5	108	7	3	118	1			1	119
	8	84	4	1	89	1			1	90
	11	85	3		88	1			1	89

* Early-stopped records.

Table 9. Overall Alternate Assessment Participation Rate

Subject	Grade	Number of HSA-Alt Test Participants	Number of Hawai`i State Summative Test Participants	Overall Hawai`i State Alternate Assessment Participation Rate (%)[1]
ELA	3	125	12,913	0.96%
	4	129	13,087	0.98%
	5	121	13,032	0.92%
	6	149	12,762	1.15%
	7	118	12,467	0.94%
	8	93	9,843	0.94%
	11	96	10,518	0.90%
	Overall	**831**	**84,622**	**0.97%**
Mathematics	3	124	12,972	0.95%
	4	127	13,119	0.96%
	5	122	13,081	0.92%
	6	149	12,801	1.15%
	7	117	12,538	0.92%
	8	93	9,932	0.93%
	11	93	10,551	0.87%
	Overall	**825**	**84,994**	**0.96%**
Science	5	119	13,089	0.90%
	8	90	10,027	0.89%
	11	89	12,220	0.72%
	Overall	**298**	**35,336**	**0.84%**

[1]The U.S. Department of Education (US DOE) looks at the overall participation rates in each subject with all grades combined. All three subject areas were at or below the 1.0% cap when the US DOE rounding rule was applied.

Table 10. Number of Participated Students by Subgroups

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
ELA							
All	125	129	121	149	118	93	96
Female	33	47	41	48	29	22	31
Male	92	82	80	101	89	71	65
Asian	34	32	40	32	27	19	22
Native Hawaiian or Pacific Islander	37	35	29	42	31	26	37
White	9	9	10	11	10	5	11
Hispanic or Latino	20	24	23	34	29	21	11
American Indian or Alaska Native				1			
Black or African American	3	2	4			1	3
Two or More Races	22	27	15	29	21	21	12
Migrant	2			1	1		
Disadvantaged	68	61	57	69	63	50	45
ELL	22	24	25	27	21	16	22
Mathematics							
All	124	127	122	149	117	93	93
Female	32	46	41	48	29	23	29
Male	92	81	81	101	88	70	64
Asian	34	32	40	32	26	20	22
Native Hawaiian or Pacific Islander	36	34	30	42	31	26	36
White	9	9	10	11	10	5	10
Hispanic or Latino	20	23	23	34	29	20	11
American Indian or Alaska Native				1			
Black or African American	3	2	4			1	2
Two or More Races	22	27	15	29	21	21	12
Migrant	2			1	1		
Disadvantaged	68	60	57	69	63	50	42
ELL	22	24	25	27	21	17	21
Science							
All			119			90	89
Female			41			22	29
Male			78			68	60
Asian			40			19	21
Native Hawaiian or Pacific Islander			28			25	36
White			10			4	9
Hispanic or Latino			22			20	11
American Indian or Alaska Native							
Black or African American			4			1	2
Two or More Races			15			21	10
Migrant							
Disadvantaged			56			49	40
ELL			25			15	21

Table 11. Number of Participated Students by Subgroups and Disability Category - Overall

Subgroup Category	ASD	DD6	DF	ED	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Number of Students												
All	305	3	2	7	198	216	4	69	15	2	10	3
Female	52	2		2	81	81	1	24	4		3	2
Male	253	1	2	5	117	135	3	45	11	2	7	1
Asian	77			1	37	73		14	3		1	1
Native Hawaiian or Pacific Islander	72		1	3	75	55	1	20	5	2	4	1
White	28				13	15	1	8				
Hispanic or Latino	62			1	40	37		13	6		3	
American Indian or Alaska Native					1							
Black or African American	7			1	3	1	1					
Two or More Races	59	3	1	1	29	35	1	14	1		2	1
Migrant				1	3							
Disadvantaged	133	3		3	123	90	2	38	12	2	7	1
ELL	57		1	1	49	34	1	11	4		1	
Percentage of Students by Subgroup Conditional on Each IDEA Category												
Female	17%	67%		29%	41%	38%	25%	35%	27%		30%	67%
Male	83%	33%	100%	71%	59%	63%	75%	65%	73%	100%	70%	33%
Asian	25%			14%	19%	34%		20%	20%		10%	33%
Native Hawaiian or Pacific Islander	24%		50%	43%	38%	25%	25%	29%	33%	100%	40%	33%
White	9%				7%	7%	25%	12%				
Hispanic or Latino	20%			14%	20%	17%		19%	40%		30%	
American Indian or Alaska Native					1%							
Black or African American	2%			14%	2%	0%	25%					
Two or More Races	19%	100%	50%	14%	15%	16%	25%	20%	7%		20%	33%
Migrant				14%	2%							
Disadvantaged	44%	100%		43%	62%	42%	50%	55%	80%	100%	70%	33%
ELL	19%		50%	14%	25%	16%	25%	16%	27%		10%	

Note. ASD= Autism Spectrum Disorder; DD6= Developmental Delay (Age 6–8); DF=Deaf; ED= Emotional Disability; HH= Hard of Hearing; ID= Intellectual Disability; MD= Multiple Disabilities; OD= Orthopedic Disability; OHD= Other Health Disability; SLD= Specific Learning Disability; SOL= Speech-Language Disability; TBI= Traumatic Brain Injury; VDB= Visual Disability Including Blindness.

Table 12. Number of Participated Students by Subgroups and Disability Category (Grades 3–4)

Group	ASD	DD6	DF	ED	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 3												
All Students	55	3		1	22	28	1	10	4		1	
Female	9	2		1	6	8		7				
Male	46	1			16	20	1	3	4		1	
Asian	17				3	11		2	1			
Native Hawaiian or Pacific Islander	15				10	5	1	4	2			
White	4				1	2		2				
Hispanic or Latino	8				6	3		2	1			
American Indian or Alaska Native												
Black or African American	2			1								
Two or More Races	9	3			2	7					1	
Migrant					2							
Disadvantaged	26	3			13	15	1	6	4			
ELL	13				2	3	1	1	2			
Grade 4												
All Students	49				33	32		10	1	2	1	1
Female	11				16	17		1	1			1
Male	38				17	15		9		2	1	
Asian	13				5	13		1				
Native Hawaiian or Pacific Islander	9				12	7		5		2		
White	3				3	3						
Hispanic or Latino	12				7	2		1	1		1	
American Indian or Alaska Native												
Black or African American	2											
Two or More Races	10				6	7		3				1
Migrant												
Disadvantaged	14				22	14		6	1	2	1	1
ELL	7				9	5		3				

Note. ASD= Autism Spectrum Disorder; DD6= Developmental Delay (Age 6–8); DF=Deaf; ED= Emotional Disability; HH= Hard of Hearing; ID= Intellectual Disability; MD= Multiple Disabilities; OD= Orthopedic Disability; OHD= Other Health Disability; SLD= Specific Learning Disability; SOL= Speech-Language Disability; TBI= Traumatic Brain Injury; VDB= Visual Disability Including Blindness.

Table 13. Number of Participated Students by Subgroups and Disability Category (Grades 5–6)

Group	ASD	DD6	DF	ED	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 5												
All Students	53			1	23	30	1	11	1		1	1
Female	13				13	8		4	1		1	1
Male	40			1	10	22	1	7				
Asian	18				6	9		6	1			
Native Hawaiian or Pacific Islander	12			1	5	8		2			1	1
White	5				2	2		1				
Hispanic or Latino	6				9	7		1				
American Indian or Alaska Native												
Black or African American	2				1		1					
Two or More Races	10					4		1				
Migrant												
Disadvantaged	26				17	9		3	1		1	
ELL	13				6	5		1				
Grade 6												
All Students	57			1	33	38		15	3		2	
Female	9				17	18		4				
Male	48			1	16	20		11	3		2	
Asian	10				7	13		2				
Native Hawaiian or Pacific Islander	14			1	11	8		5	2		1	
White	5				2	3		1				
Hispanic or Latino	16				6	8		3	1			
American Indian or Alaska Native					1							
Black or African American												
Two or More Races	12				6	6		4			1	
Migrant				1								
Disadvantaged	25				18	16		9	1			
ELL	9				9	5		4				

Note. ASD= Autism Spectrum Disorder; DD6= Developmental Delay (Age 6–8); DF=Deaf; ED= Emotional Disability; HH= Hard of Hearing; ID= Intellectual Disability; MD= Multiple Disabilities; OD= Orthopedic Disability; OHD= Other Health Disability; SLD= Specific Learning Disability; SOL= Speech-Language Disability; TBI= Traumatic Brain Injury; VDB= Visual Disability Including Blindness.

Table 14. Number of Participated Students by Subgroups and Disability Category (Grades 7–8)

Group	ASD	DD6	DF	ED	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 7												
All Students	39		2		25	36		10	3		3	
Female	4				8	12		3	1		1	
Male	35		2		17	24		7	2		2	
Asian	10				5	11		1				
Native Hawaiian or Pacific Islander	8		1		9	10		1			2	
White	4				3	2		1				
Hispanic or Latino	11				3	8		4	2		1	
American Indian or Alaska Native												
Black or African American												
Two or More Races	6		1		5	5		3	1			
Migrant					1							
Disadvantaged	22				19	11		5	3		3	
ELL	6		1		7	7						
Grade 8												
All Students	32			1	29	23		6	1		1	1
Female	3			1	8	7		2	1		1	
Male	29				21	16		4				1
Asian	6				6	5		2				1
Native Hawaiian or Pacific Islander	8			1	9	7		1				
White	4				1							
Hispanic or Latino	5				5	8		1	1		1	
American Indian or Alaska Native												
Black or African American						1						
Two or More Races	9				8	2		2				
Migrant												
Disadvantaged	13			1	17	15		4			1	
ELL	6			1	5	3		2				

Note. ASD= Autism Spectrum Disorder; DD6= Developmental Delay (Age 6–8); DF=Deaf; ED= Emotional Disability; HH= Hard of Hearing; ID= Intellectual Disability; MD= Multiple Disabilities; OD= Orthopedic Disability; OHD= Other Health Disability; SLD= Specific Learning Disability; SOL= Speech-Language Disability; TBI= Traumatic Brain Injury; VDB= Visual Disability Including Blindness.

Table 15. Number of Participated Students by Subgroups and Disability Category (Grade 11)

Group	ASD	DD6	DF	ED	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 11												
All Students	20			3	33	29	2	7	2		1	
Female	3				13	11	1	3				
Male	17			3	20	18	1	4	2		1	
Asian	3			1	5	11			1		1	
Native Hawaiian or Pacific Islander	6				19	10		2	1			
White	3				1	3	1	3				
Hispanic or Latino	4			1	4	1		1				
American Indian or Alaska Native												
Black or African American	1				2							
Two or More Races	3			1	2	4	1	1				
Migrant												
Disadvantaged	7			2	17	10	1	5	2		1	
ELL	3				11	6			2		1	

Note. ASD= Autism Spectrum Disorder; DD6= Developmental Delay (Age 6–8); DF=Deaf; ED= Emotional Disability; HH= Hard of Hearing; ID= Intellectual Disability; MD= Multiple Disabilities; OD= Orthopedic Disability; OHD= Other Health Disability; SLD= Specific Learning Disability; SOL= Speech-Language Disability; TBI= Traumatic Brain Injury; VDB= Visual Disability Including Blindness.

3.2 SUMMARY OF OVERALL STUDENT PERFORMANCE

Table 16 through Table 24 present a summary of the spring 2023 HSA-Alt test results for all students and by subgroup, including the average and the standard deviation of scale scores, the percentage of students in each performance level, and the percentage of proficient (Meets + Exceeds) students. The results are based on the students who meet attemptedness requirements for scoring and reporting of the HSA-Alt.

Table 16. Student Performance by Grade and Subgroup – ELA (Grades 3–4)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 3								
All Students	125	275.68	70.77	45	16	21	18	39
Female	33	273.66	70.20	45	15	21	18	39
Male	92	276.40	71.34	45	16	21	18	39
Asian	34	262.64	77.21	59	15	12	15	26
Native Hawaiian or Pacific Islander	37	284.86	73.01	35	16	19	30	49
White	9	304.82	31.60	22	33	33	11	44
Hispanic or Latino	20	279.73	54.62	45	15	25	15	40
American Indian or Alaska Native								
Black or African American	3*							
Two or More Races	22	259.46	81.74	55	14	18	14	32
Migrant	2*							
Disadvantaged	68	291.05	61.39	35	21	19	25	44
ELL	22	282.00	54.85	50	18	18	14	32
Grade 4								
All Students	129	279.42	64.31	36	22	19	22	41
Female	47	257.53	78.91	43	30	17	11	28
Male	82	291.96	50.61	33	18	21	28	49
Asian	32	275.83	60.49	47	22	16	16	31
Native Hawaiian or Pacific Islander	35	289.63	40.49	34	26	29	11	40
White	9	297.95	24.39	33	11	22	33	56
Hispanic or Latino	24	272.60	82.62	33	17	25	25	50
American Indian or Alaska Native								
Black or African American	2*							
Two or More Races	27	275.20	77.27	30	30	7	33	41
Migrant								
Disadvantaged	61	280.80	64.30	31	25	21	23	44
ELL	24	290.95	45.58	29	21	38	13	50

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 17. Student Performance by Grade and Subgroup – ELA (Grades 5–7)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 5								
All Students	121	277.98	59.19	38	24	28	10	38
Female	41	274.20	59.76	46	22	27	5	32
Male	80	279.91	59.18	34	25	29	13	41
Asian	40	270.63	58.13	50	20	20	10	30
Native Hawaiian or Pacific Islander	29	291.99	49.42	31	21	31	17	48
White	10	268.76	67.13	40	30	20	10	30
Hispanic or Latino	23	263.23	78.12	26	39	30	4	35
American Indian or Alaska Native								
Black or African American	4*							
Two or More Races	15	290.98	42.35	47	13	33	7	40
Migrant								
Disadvantaged	57	281.17	58.70	35	19	33	12	46
ELL	25	287.90	45.69	28	28	32	12	44
Grade 6								
All Students	149	278.63	69.27	40	20	22	17	40
Female	48	275.11	75.17	40	23	17	21	38
Male	101	280.30	66.62	41	19	25	16	41
Asian	32	249.62	73.24	56	22	19	3	22
Native Hawaiian or Pacific Islander	42	291.32	52.78	40	17	21	21	43
White	11	298.19	75.85	36	9	9	45	55
Hispanic or Latino	34	275.54	81.25	29	32	24	15	38
American Indian or Alaska Native	1*							
Black or African American								
Two or More Races	29	284.54	60.42	38	14	31	17	48
Migrant	1*							
Disadvantaged	69	286.97	57.51	38	16	28	19	46
ELL	27	277.61	60.82	44	22	22	11	33
Grade 7								
All Students	118	277.06	63.49	41	22	20	17	37
Female	29	273.84	56.63	38	28	24	10	34
Male	89	278.11	65.84	42	20	19	19	38
Asian	27	259.58	64.52	63	19	15	4	19
Native Hawaiian or Pacific Islander	31	279.49	57.42	39	32	13	16	29
White	10	257.32	66.36	50	30	10	10	20
Hispanic or Latino	29	281.34	81.99	34	7	28	31	59
American Indian or Alaska Native								
Black or African American								
Two or More Races	21	299.43	25.79	19	29	33	19	52
Migrant	1*							
Disadvantaged	63	286.23	54.77	35	19	27	19	46
ELL	21	259.54	73.81	48	33	14	5	19

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 18. Student Performance by Grade and Subgroup – ELA (Grades 8 and 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 8								
All Students	93	276.30	49.96	52	17	22	10	31
Female	22	270.70	64.23	50	5	36	9	45
Male	71	278.03	45.06	52	21	17	10	27
Asian	19	276.53	38.06	53	26	11	11	21
Native Hawaiian or Pacific Islander	26	278.39	36.43	62	8	23	8	31
White	5*							
Hispanic or Latino	21	258.73	69.86	62	14	14	10	24
American Indian or Alaska Native								
Black or African American	1*							
Two or More Races	21	288.70	52.74	29	29	29	14	43
Migrant								
Disadvantaged	50	272.20	53.96	62	12	16	10	26
ELL	16	274.29	35.17	69	13	6	13	19
Grade 11								
All Students	96	288.31	52.79	34	29	13	24	36
Female	31	274.47	51.33	39	39	10	13	23
Male	65	294.91	52.58	32	25	14	29	43
Asian	22	297.51	47.37	36	23	14	27	41
Native Hawaiian or Pacific Islander	37	275.59	58.70	38	35	11	16	27
White	11	289.52	61.92	36	27	9	27	36
Hispanic or Latino	11	316.87	40.43	18	18	9	55	64
American Indian or Alaska Native								
Black or African American	3*							
Two or More Races	12	295.08	32.02	25	33	25	17	42
Migrant								
Disadvantaged	45	287.45	60.78	38	22	11	29	40
ELL	22	297.45	37.58	27	32	14	27	41

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 19. Student Performance by Grade and Subgroup – Mathematics (Grades 3–5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient ^
Grade 3								
All Students	124	275.60	65.74	42	23	15	21	35
Female	32	274.99	56.85	50	19	6	25	31
Male	92	275.81	68.85	39	24	17	20	37
Asian	34	268.55	78.58	47	24	12	18	29
Native Hawaiian or Pacific Islander	36	280.88	60.76	33	19	28	19	47
White	9	289.62	41.83	44	22	0	33	33
Hispanic or Latino	20	285.71	39.11	55	25	0	20	20
American Indian or Alaska Native								
Black or African American	3*							
Two or More Races	22	262.33	83.13	41	14	18	27	45
Migrant	2*							
Disadvantaged	68	284.04	56.54	38	22	21	19	40
ELL	22	281.15	45.47	41	18	32	9	41
Grade 4								
All Students	127	273.22	68.34	46	17	23	14	37
Female	46	253.14	81.17	48	22	26	4	30
Male	81	284.61	57.34	44	15	21	20	41
Asian	32	265.17	69.06	47	25	19	9	28
Native Hawaiian or Pacific Islander	34	277.36	48.42	47	26	15	12	26
White	9	303.41	38.13	22	22	33	22	56
Hispanic or Latino	23	277.78	78.64	43	9	30	17	48
American Indian or Alaska Native								
Black or African American	2*							
Two or More Races	27	265.91	78.62	52	4	30	15	44
Migrant								
Disadvantaged	60	274.02	67.11	43	18	27	12	38
ELL	24	276.17	50.84	50	21	25	4	29
Grade 5								
All Students	122	282.51	65.61	43	14	22	21	43
Female	41	283.21	61.73	39	15	24	22	46
Male	81	282.15	67.86	44	14	21	21	42
Asian	40	284.88	63.4	40	13	23	25	48
Native Hawaiian or Pacific Islander	30	284.07	55.99	43	13	30	13	43
White	10	257.30	78.14	50	10	30	10	40
Hispanic or Latino	23	262.86	79.93	43	26	17	13	30
American Indian or Alaska Native								
Black or African American	4*							
Two or More Races	15	313.57	54.74	47	7	7	40	47
Migrant								
Disadvantaged	57	290.48	68.66	35	14	25	26	51
ELL	25	294.49	52.50	36	20	28	16	44

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 20. Student Performance by Grade and Subgroup – Mathematics (Grades 6–8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 6								
All Students	149	258.15	74.11	52	19	18	10	28
Female	48	244.86	71.42	63	17	17	4	21
Male	101	264.47	74.87	48	21	19	13	32
Asian	32	226.58	69.69	72	22	6	0	6
Native Hawaiian or Pacific Islander	42	275.32	69.56	48	12	26	14	40
White	11	284.03	69.52	27	27	18	27	45
Hispanic or Latino	34	258.04	83.60	53	18	18	12	29
American Indian or Alaska Native	1*							
Black or African American								
Two or More Races	29	257.62	68.85	48	24	21	7	28
Migrant	1*							
Disadvantaged	69	267.74	63.01	43	25	26	6	32
ELL	27	268.45	68.78	44	22	19	15	33
Grade 7								
All Students	117	267.83	64.19	44	31	12	14	26
Female	29	257.70	49.69	55	38	7	0	7
Male	88	271.17	68.22	40	28	14	18	32
Asian	26	254.7	68.35	54	31	12	4	15
Native Hawaiian or Pacific Islander	31	269.89	63.93	48	29	3	19	23
White	10	240.94	56.93	70	20	10	0	10
Hispanic or Latino	29	267.80	75.06	34	31	21	14	34
American Indian or Alaska Native								
Black or African American								
Two or More Races	21	293.89	35.77	24	38	14	24	38
Migrant	1*							
Disadvantaged	63	273.93	55.45	43	32	11	14	25
ELL	21	254.16	73.89	48	38	0	14	14
Grade 8								
All Students	93	273.76	51.73	52	17	12	19	31
Female	23	264.97	65.40	52	17	13	17	30
Male	70	276.65	46.59	51	17	11	20	31
Asian	20	279.08	31.85	55	20	15	10	25
Native Hawaiian or Pacific Islander	26	268.97	37.95	65	8	12	15	27
White	5*							
Hispanic or Latino	20	258.23	75.06	45	25	15	15	30
American Indian or Alaska Native								
Black or African American	1*							
Two or More Races	21	290.97	54.95	33	19	10	38	48
Migrant								
Disadvantaged	50	269.16	55.89	54	16	12	18	30
ELL	17	283.51	40.50	53	12	12	24	35

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 21. Student Performance by Grade and Subgroup – Mathematics (Grade 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 11								
All Students	93	281.23	45.48	39	25	22	15	37
Female	29	279.63	44.43	34	38	10	17	28
Male	64	281.96	46.27	41	19	27	14	41
Asian	22	292.54	35.39	36	23	23	18	41
Native Hawaiian or Pacific Islander	36	271.88	49.35	44	25	22	8	31
White	10	262.82	64.00	50	20	20	10	30
Hispanic or Latino	11	298.79	24.74	27	27	18	27	45
American Indian or Alaska Native								
Black or African American	2*							
Two or More Races	12	291.54	40.48	25	25	25	25	50
Migrant								
Disadvantaged	42	278.69	50.46	40	31	14	14	29
ELL	21	294.56	25.94	24	24	38	14	52

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 22. Student Performance by Grade and Subgroup – Science (Grade 5)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 5								
All Students	119	270.09	66.30	46	17	25	12	37
Female	41	265.34	65.70	54	17	17	12	29
Male	78	272.58	66.90	42	17	29	12	41
Asian	40	265.94	64.53	50	15	28	8	35
Native Hawaiian or Pacific Islander	28	273.62	59.25	43	18	29	11	39
White	10	249.15	70.67	50	20	20	10	30
Hispanic or Latino	22	253.10	74.13	45	32	14	9	23
American Indian or Alaska Native								
Black or African American	4*							
Two or More Races	15	298.46	63.54	47	0	33	20	53
Migrant								
Disadvantaged	56	278.68	67.26	43	18	20	20	39
ELL	25	267.34	55.58	48	16	32	4	36

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 23. Student Performance by Grade and Subgroup – Science (Grade 8)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 8								
All Students	90	268.97	53.45	41	31	19	9	28
Female	22	253.20	66.57	50	23	23	5	27
Male	68	274.08	47.94	38	34	18	10	28
Asian	19	279.82	47.22	47	21	11	21	32
Native Hawaiian or Pacific Islander	25	269.98	33.35	48	28	20	4	24
White	4*							
Hispanic or Latino	20	246.26	71.21	45	30	25	0	25
American Indian or Alaska Native								
Black or African American	1*							
Two or More Races	21	276.69	53.56	29	43	19	10	29
Migrant								
Disadvantaged	49	265.14	56.91	47	27	16	10	27
ELL	15	281.85	43.46	40	33	7	20	27

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

Table 24. Student Performance by Grade and Subgroup – Science (Grade 11)

Group	Number Tested	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient^
Grade 11								
All Students	89	285.82	58.35	30	31	20	18	38
Female	29	275.14	51.73	31	52	3	14	17
Male	60	290.99	61.04	30	22	28	20	48
Asian	21	288.76	50.99	33	29	10	29	38
Native Hawaiian or Pacific Islander	36	275.88	57.44	25	44	22	8	31
White	9	317.90	80.07	33	11	11	44	56
Hispanic or Latino	11	308.65	43.22	18	18	36	27	64
American Indian or Alaska Native								
Black or African American	2*							
Two or More Races	10	268.46	64.67	40	30	30	0	30
Migrant								
Disadvantaged	40	285.48	65.64	30	33	8	30	38
ELL	21	293.54	48.99	24	29	19	29	48

*To protect individual student confidentiality, results are not reported for 5 or fewer students.
^% Proficient is the sum of % Meets and % Exceeds.

3.3 TEST-TAKING TIME

The HSA-Alt assessments are not timed and are either administered one-on-one or in a dyad or triad grouping with the test administrator assisting in the test administration, as needed and supervising during the testing process to ensure all test components are delivered to each student. The time spent on each item may vary among individual students, which may provide useful information about student testing behaviors and motivation, for example. Since the length of a test session can be monitored by TAs who are knowledgeable about their students, additional time for students who need it can be arranged.

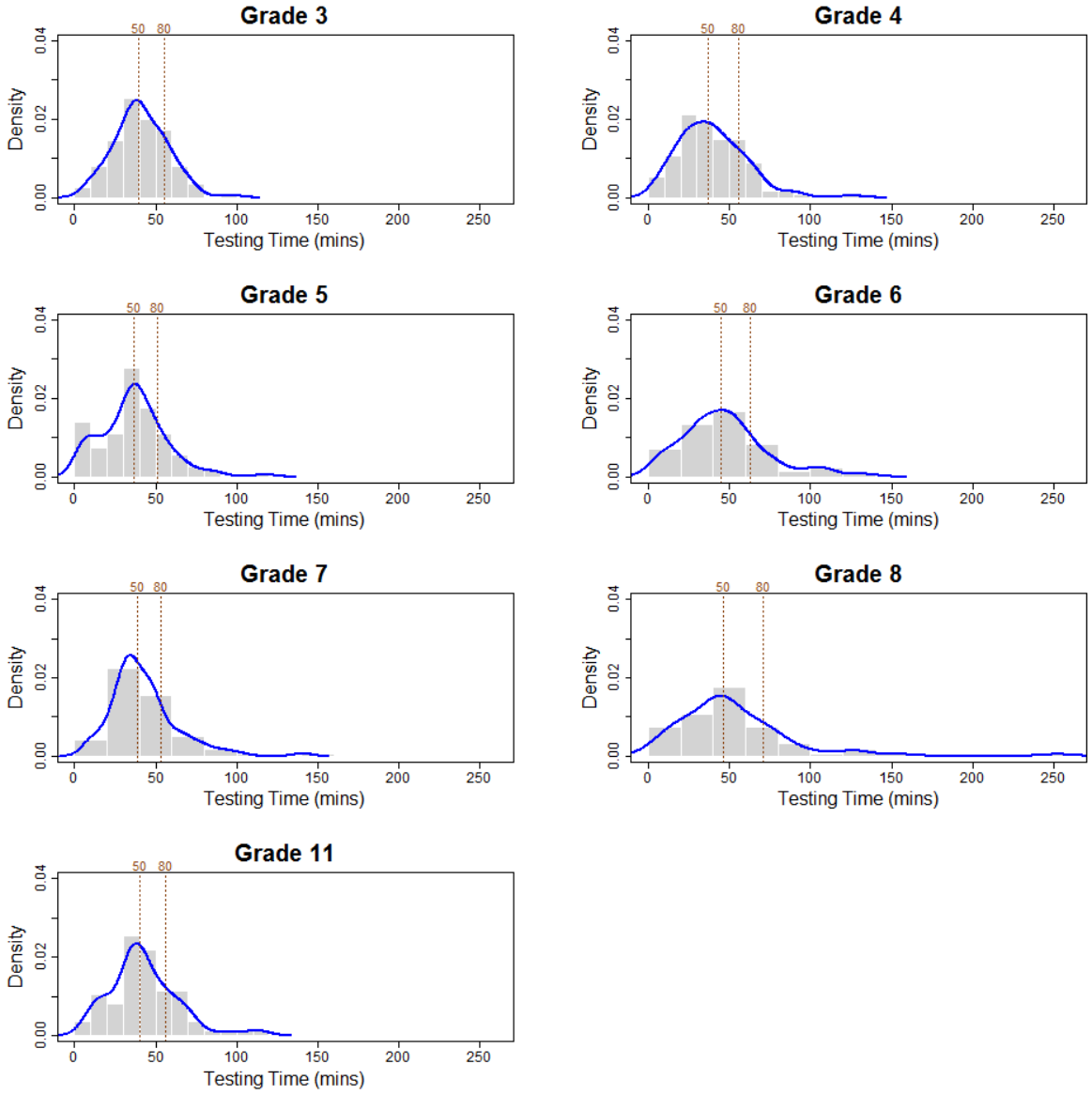
In the Test Delivery System (TDS), item response time is captured as the item page time (the time that a student spends on each item page) in milliseconds. Discrete items appear on the screen one item at a time, and items associated with a stimulus appear on the screen together with the page time measured as the total time spent on all associated items. In this case, the page time for each item is the average time for all the items associated with the stimulus. For each student, the total testing time for the test was the sum of the page time for all items.

Table 25 presents the 2023 TDS time (the average testing time, the median testing time, and the testing time at various percentiles for students who completed the online adaptive tests.) The distribution of TDS testing time is also provided in Figure 1–Figure 3.

Table 25. Test-Taking Time

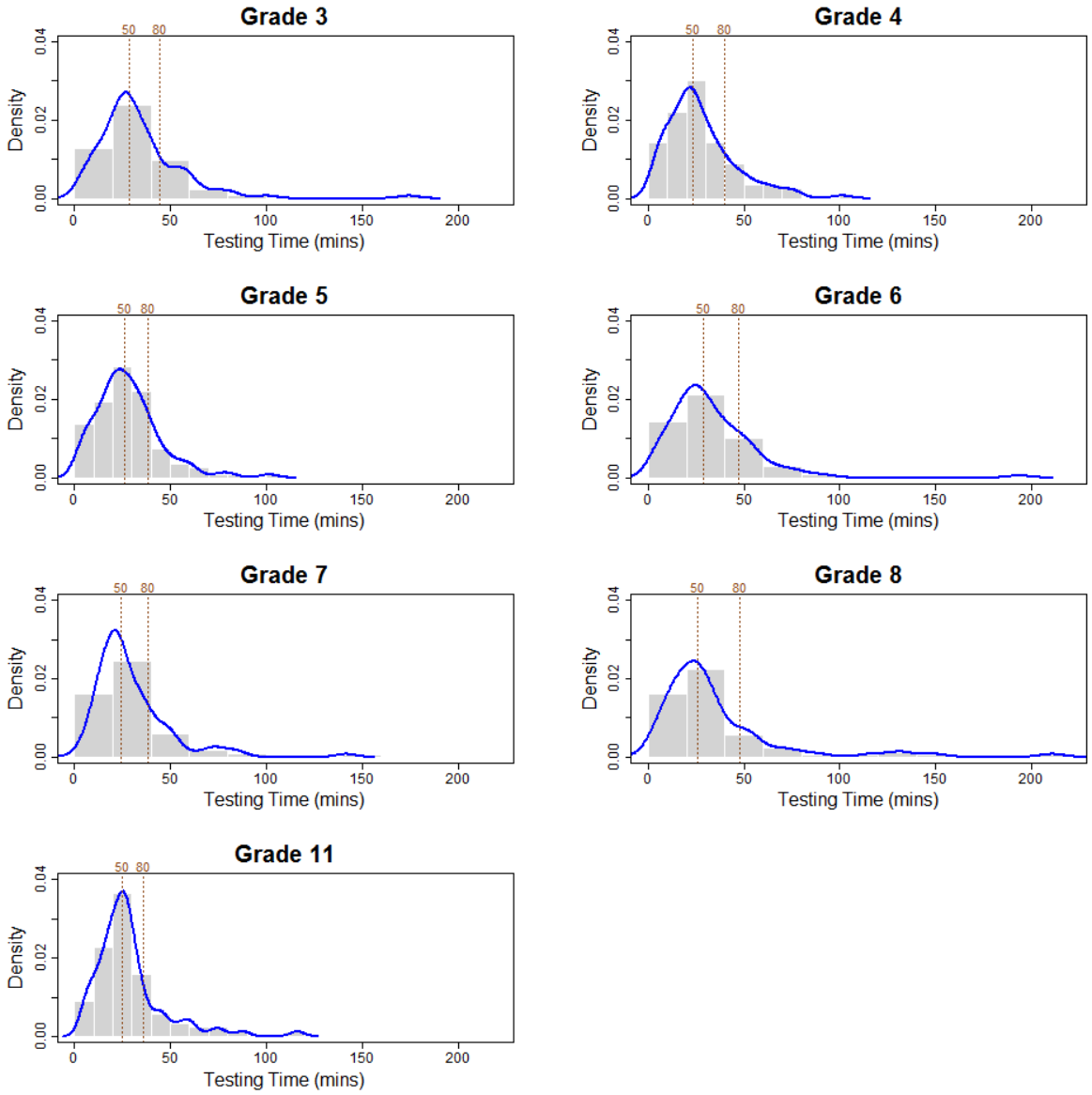
Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)			
			Min	25th	75th	Max
ELA						
3	0:41	0:40	0:06	0:31	0:52	1:37
4	0:40	0:37	0:04	0:24	0:52	2:05
5	0:37	0:37	0:03	0:22	0:47	1:57
6	0:47	0:43	0:02	0:30	0:59	2:16
7	0:43	0:38	0:07	0:31	0:51	2:20
8	0:52	0:45	0:03	0:32	1:05	4:12
11	0:42	0:40	0:05	0:32	0:53	1:56
Mathematics						
3	0:33	0:28	0:03	0:20	0:41	2:54
4	0:27	0:23	0:03	0:16	0:35	1:41
5	0:28	0:26	0:02	0:17	0:36	1:41
6	0:33	0:28	0:02	0:19	0:42	3:13
7	0:29	0:24	0:02	0:17	0:36	2:21
8	0:35	0:26	0:02	0:16	0:40	3:31
11	0:28	0:25	0:04	0:19	0:32	1:56
Science						
5	0:32	0:30	0:03	0:22	0:40	1:38
8	0:36	0:28	0:02	0:18	0:42	2:54
11	0:24	0:21	0:02	0:15	0:31	2:26

Figure 1. Distribution of Testing Time - ELA



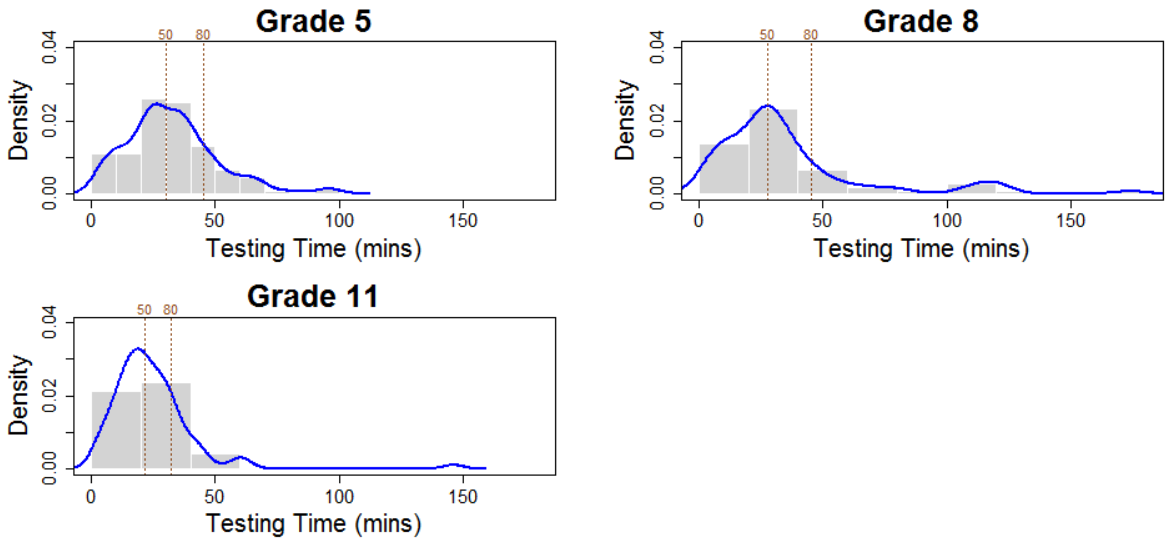
Note: 50 is median; 80 is 80th percentile

Figure 2. Distribution of Testing Time - Mathematics



Note: 50 is median; 80 is 80th percentile

Figure 3. Distribution of Testing Time - Science



Note: 50 is median; 80 is 80th percentile

3.4 Distribution of Student Ability and Item Difficulty for HSA-Alt Item Pool

Figure 4–Figure 6 display the empirical distribution of the Hawai`i student overall scores in theta in the spring 2023 administration and the distribution of the item difficulty parameter estimates in the 2023 item pool. The student ability distributions in ELA, mathematics, and science tests are based on the completed test results from both the adaptive and fixed form tests. These charts provide a visual presentation on whether the difficulty levels of the items in the pool cover the ability range of the population being assessed. They can also inform a direction for future item development. For example, in some mathematics tests, more easier items are needed in the item pool that target students with lower academic achievement.

Table 26 presents the correlations between students' final estimated theta scores and the average test form difficulty for each subject and grade. This analysis only included students who completed the online adaptive tests. The high correlations for all tests, ranging from 0.70 in grade 7 mathematics to 0.92 in grade 5 science, provide evidence that the adaptive algorithm works as expected and items are selected to match students' abilities.

Figure 4. Student Ability and Item Difficulty Distributions for ELA

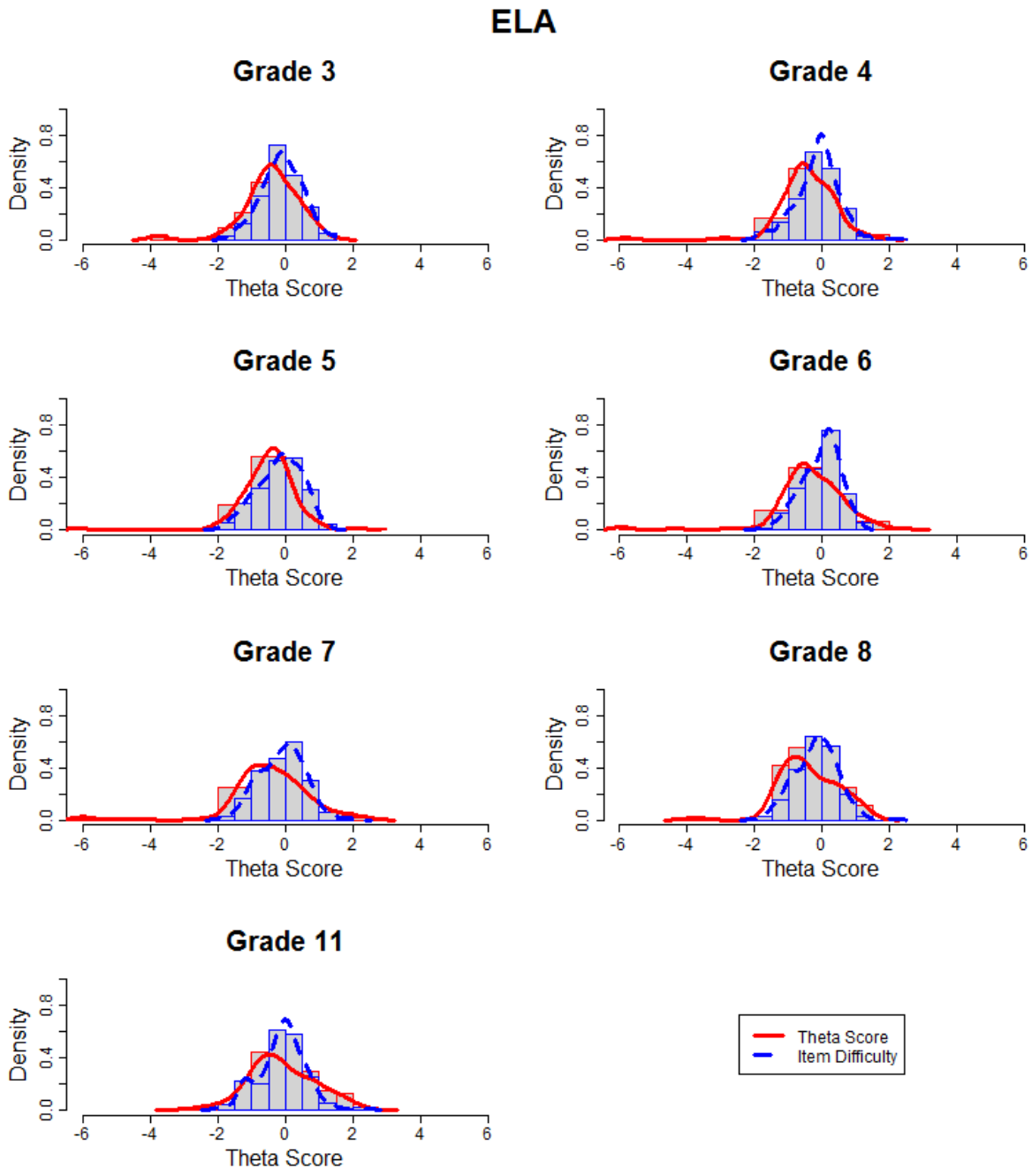


Figure 5. Student Ability and Item Difficulty Distributions for Mathematics

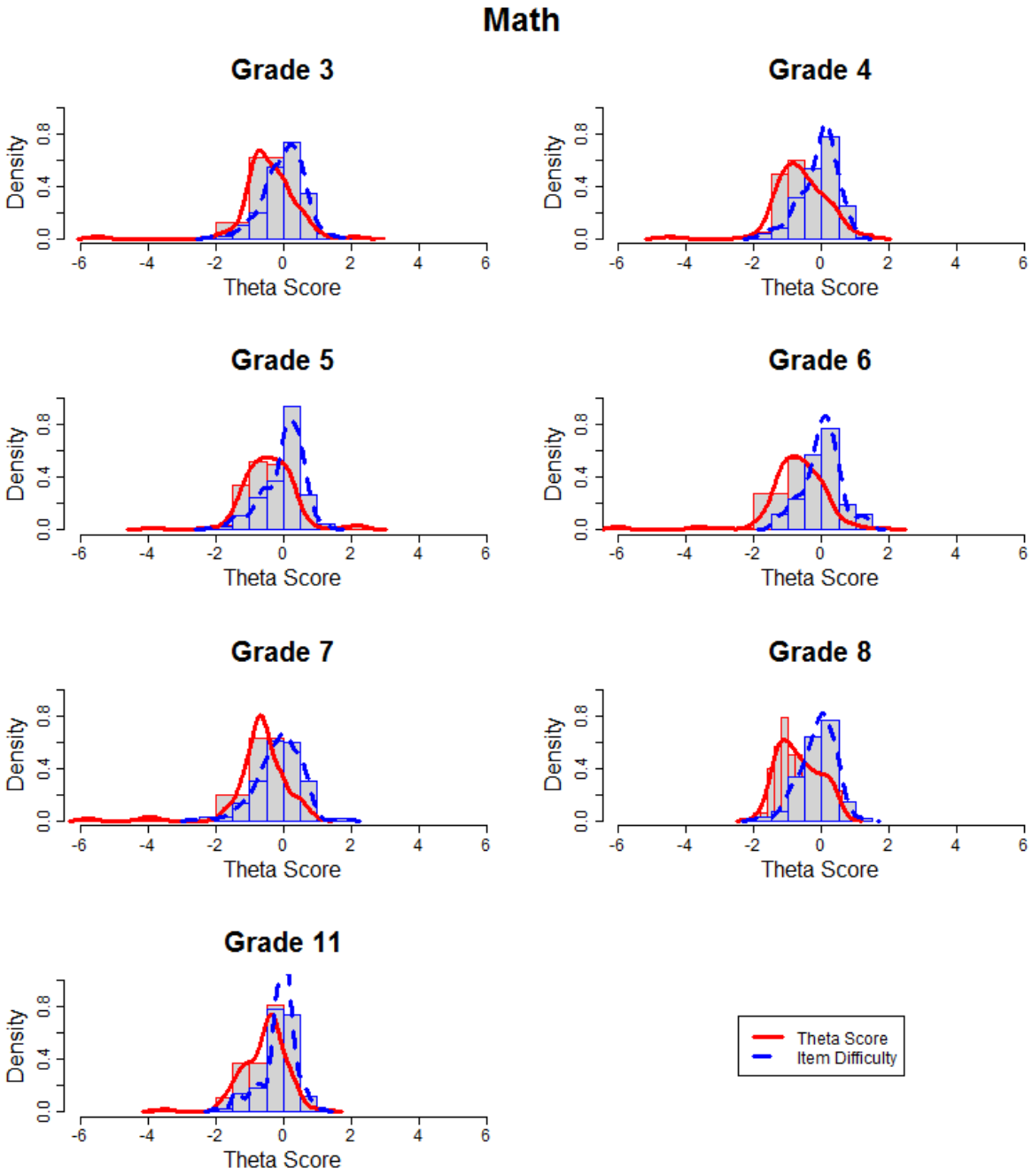

Math
Grade 3
Grade 4
Grade 5
Grade 6
Grade 7
Grade 8
Grade 11
Density
Theta Score
Theta Score
Item Difficulty

Figure 6. Student Ability and Item Difficulty Distributions for Science

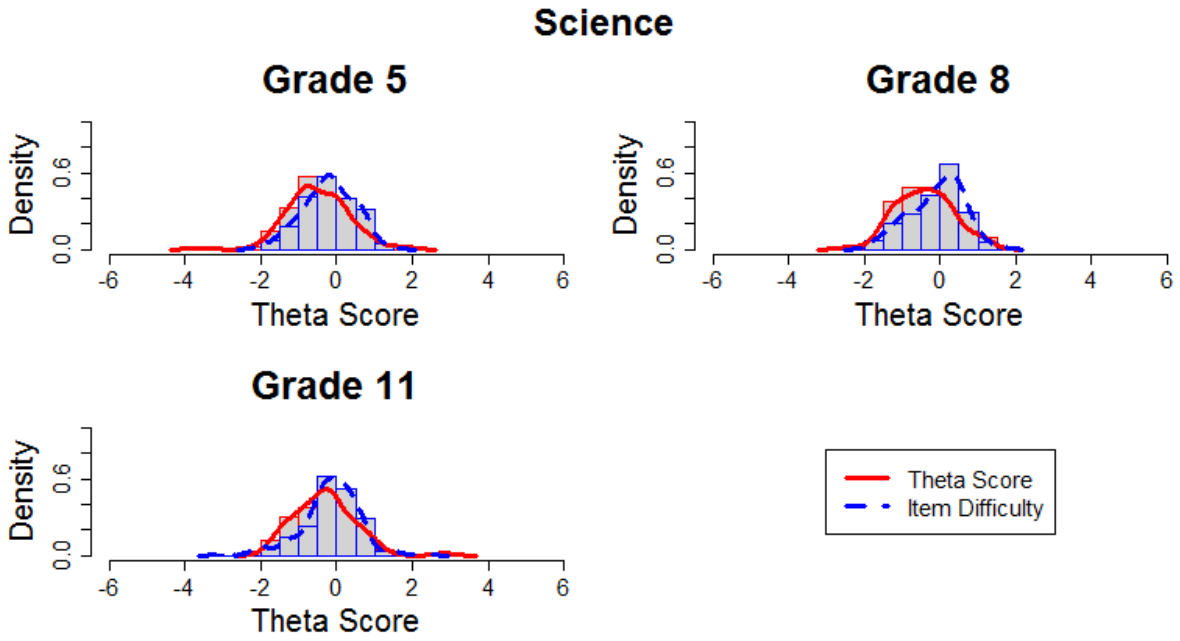


Table 26. Correlation Between Student Ability Scores and Average Test Form Difficulty

Subject	Grade	N	Correlation
ELA	3	111	0.87
	4	114	0.79
	5	108	0.90
	6	133	0.87
	7	107	0.89
	8	88	0.85
	11	87	0.88
Mathematics	3	111	0.74
	4	113	0.83
	5	109	0.88
	6	132	0.81
	7	108	0.70
	8	87	0.89
	11	88	0.89
Science	5	108	0.92
	8	84	0.90
	11	85	0.91

4. ITEM DEVELOPMENT

4.1 ITEM DEVELOPMENT FOR THE MOU-ALT

A Memorandum of Understanding (MOU) on item sharing in item development and field testing was initiated in 2018 and signed among states who would like to join the agreement. Each participating state contributed a predetermined number of items proportional to their state's student population for the alternate assessment. In 2018, Hawai`i, South Carolina, and Wyoming joined the MOU for ELA, mathematics, and science; in 2019, Idaho and Vermont joined the MOU for ELA, mathematics and science; in 2020, Montana and South Dakota joined the MOU for science. In 2022, Vermont exited the MOU.

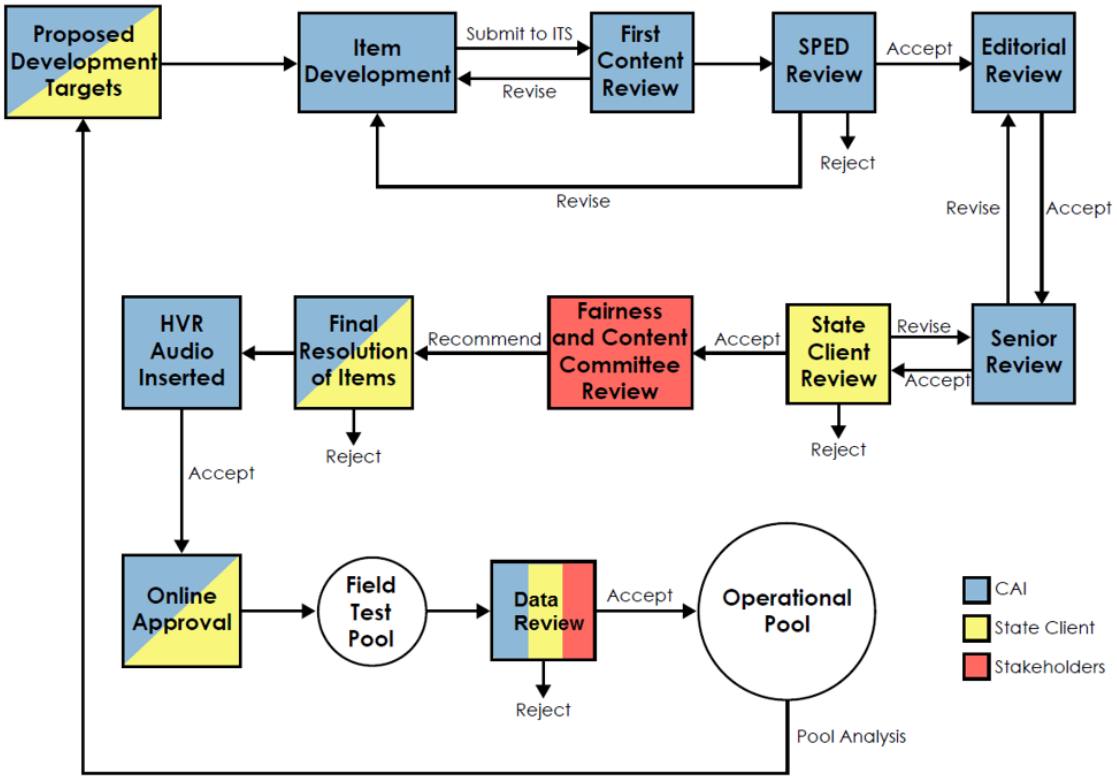
For the first year of the alternate assessment MOU shared field test item development, a crosswalk across all the individual state alternate assessment standards was completed. Test items from each of the original three states could then be aligned across states. Once all individual state items were aligned across all three states, item development plans were created for each. These item development plans were based on identified areas where additional items were needed to ensure that all the MOU standards aligned on the crosswalk were addressed in the shared field test pool and items for each state-specific standard or content specification that was not aligned to the MOU crosswalk standards were created to meet the state's test blueprint. These item development plans guided the development of the new items to be field tested across states. Each year, following data review of the field-test items, an item pool analysis is conducted and a new item development plan is created. As new states joined the MOU Alternate Assessment agreement, or in cases where states changed their standards, the individual state standards were added to the crosswalk so the items from the state could be aligned across all the states.

Starting in 2017, items were developed each year for the state-shared MOU Alternate Assessment field test pool. All the items were developed by a group of professional item writers that included both experienced item writers with a background in education and expertise in the assigned content area and specialists in alternate assessments with experience in teaching students with significant cognitive disabilities. Prior to item development, item writers were trained on aspects of items that would be unique to students with significant cognitive disabilities. A group of senior test-development specialists monitored and supported item development activities.

The development process begins with establishing CAI's proposed development targets and working with the individual states to edit the development targets and accept a final plan. The CAI content team then starts item development. After the initial round of development, a group of reviewers, including content and senior reviewers, review each item. This is followed by an individual content review, where edits are made based on comments from the group reviews, and then each item is reviewed by a special education expert. After items are reviewed by the special education reviewer, the items go through an editorial review. After editorial review, the items go back through a senior review as the last step of review at CAI before the items are sent to each state for the client review. At this step, the client accepts the items, recommends edits, or rejects the items. After client comments are resolved, all accepted items are then taken to a stakeholder Content and Fairness Committee review. After the Content and Fairness Committee makes its recommendations, the states and CAI go through a final edit resolution. The items then go through an approval step in which CAI verifies that the items will appear on the test as expected. Items are then moved into the field-test item pool and are field tested. After the testing window, all the field-test items are analyzed. Items with sample size smaller than 50 are archived and will be field-tested again in future years; items with a negative biserial/polyserial correlation are first verified by CAI content specialists to ensure the items are not mis-keyed before they are rejected from the item bank; items with sample size larger than

50 and a positive biserial/polybiserial correlation are reviewed in an item data review meeting, first with Hawai`i teachers and CAI, then with CAI and the states. This is followed with a stakeholder item data review meeting. At either item data review meeting, items may be accepted and/or rejected by each individual state. The accepted items are then moved into the operational item pools for each individual state. Figure 7 below outlines the test development process.

Figure 7. Alternate Assessment Item Development Process



4.1.1. Item Type and Scoring Rubrics

The Hawai'i Alternate Assessment item pool has multiple-choice (MC) items and multi-select (MS) items. The MC items have two to four options with one key. The MS items have up to five options with two keys. For MC items, if the key is selected, the student will receive one point; otherwise, the student receives zero points. For MS items, if a student selects two keys, they earn two points; if the student selects only one key, they earn one point; otherwise, the student earns zero points. Each item measures a specific content standard. Items were written to a variety of difficulty levels. The final item difficulties are determined through field testing.

Items can be stand-alone, grouped in short passages with two to three items, or grouped in long passages with four or more items. The test administration algorithm ensures that items within a passage are always administered consecutively.

Starting in late spring 2018, cognitive labs (cog labs) were conducted in each of the original three states to determine if certain types of technology-enhanced items should be developed for the MOU shared field test items. The item types included multi-select, equation editor, table match, and animation. Neither equation

editor nor table match proved to be a successful item type for this population of students, and therefore, states will not develop any more of these item formats. MS items were successful for high-functioning middle school and high school students and will continue to be developed for this segment of the Alternate Assessment population. Animations were successful in Hawai`i across all grade levels, and these item formats were developed and field tested beginning in spring 2022.

4.1.2 Item Development Procedure and Item Reviews

4.1.2.1 Item Development Procedure

Items were developed by each of the states that belong to the shared item development agreement. The priority for item development each year of the MOU agreement has been to increase the volume of the bank in areas that have had few or no items. In each state, item development for each year begins in the spring with creation and approval of an item development plan (IDP) for each state.

Item development plan creation in SY 2022-2023 started with CAI content staff completing a pool analysis for Hawaii and three other Alt MOU member states (for ELA and math, and five other Alt MOU states for science). CAI then added the results to a combined Alt MOU crosswalk document. From here, CAI identified any essence statements for which Hawaii has only a few or no items. Once this was completed for all states, items were added to the IDP making it easy to see how developed items will affect all states' banks. For example, if two items were added to a particular grade 3 ELA essence statement to be developed in Hawaii, the crosswalk indicated to which other states those two items could be shared to and aligned. Likewise, if another Alt MOU member state had 2 items placed on their IDP at a grade 6 math standard, the IDP indicated if those items could be shared and aligned to Hawaii. The completed crosswalk document clearly shows the number of items to be developed for Hawaii and contributed to the MOU by HIDOE, as well as the items to be developed for the MOU member states that will align to Hawaii essence statements.

Additionally, CAI psychometricians provide guidance during the development of the IDP based on the need to ensure that the item pool is sufficient to meet the test blueprint.

Once created, a senior level CAI content team member reviews the IDP. Once the IDP is approved by Senior Content, it is then sent to HIDOE for review and approval. If HIDOE requests changes, CAI content staff reviews, talks with HIDOE as necessary, and modifies the IDP. The IDP is again reviewed by Senior Content staff and sent to HIDOE for final approval.

CAI used the item development plan to author new items for initial batch delivery to the client. After newly-written items passed the required four stages of CAI internal reviews, described at length in the following sections, items were then presented to the state for department review and acceptance. Following a state's item approval, the other sharing state partners were notified that the items were ready for review and to receive comments. During this review step, states could also verify whether the items aligned to their own state standards. Any comments regarding item content and suggested revisions were sent to the state that owned the items, and it was that state's determination whether these comments should be acted upon.

In each state, items owned by the state that were accepted by the state were prepared for review by a state-wide Content and Fairness Committee convened for each content area in each state. The Content and Fairness Committee was composed of stakeholders from around the state, including special educators, general educators, complex-level staff with expertise in special education, and university professors with expertise in special education. These stakeholders reviewed items and provided feedback to ensure that all accepted items are correct and free from bias and sensitivity issues. Most importantly, these educators made

sure that this population of students would be able to understand the language used in the items and that the included visuals and audio directions would aid and not distract students.

Following committee reviews in these states, the accepted items were then shared across the five state item banks for field testing.

4.1.2.2 Item Reviews

Draft items are reviewed at various stages within CAI, followed by a review from the state staff and the state special education and general education teachers.

CAI Review: Items are reviewed at CAI at various levels.

- CAI Internal Group Review: Prior to making any changes to draft items, content and senior reviewers meet to discuss items and determine revisions to content, alignment, and style.
- CAI Internal Preliminary Review: Following internal group review, the internal preliminary review is conducted by a member of CAI's content team assigned to the Alternate Assessments. Items are revised, as agreed upon in the group review, to eliminate initial errors, meet content standards, and meet internal style and clarity expectations.
- CAI Internal Content Review: A second content review occurs after the preliminary review to further ensure changes based on the group review, and to revise items further, as necessary, to address any content, alignment, clarity, accessibility, and errors.
- Special Education Review: At this stage, items are reviewed by a CAI special education expert. The expert reviews and revises the items to ensure that they not only meet the content standards but are also as accessible as possible to students across a wide spectrum of cognitive and physical disabilities. When appropriate, the special education expert designates items as "Access Limited," meaning that a task is inappropriate to administer to students with a specific physical disability (e.g., blindness). If revisions are required, the special education reviewer will send items back to the content reviewer to implement changes.
- Edit Review: After the special education reviewer approves items, they send them through an editorial review. At this stage, a CAI content editor reviews each item to verify that the language used conforms to the standard editorial and style conventions outlined in the item-development style guide.
- Senior Review: At this stage, a CAI senior content specialist reviews all items to ensure that they meet the content standards, they are free of typographical and technical errors (e.g., key check, spelling error check), and the previously requested edits are in place.
- CAI Batch Review: This is the last step in the CAI internal review process and is designed as a final quality control check to ensure the items are ready for state review.

State Review: At this level, items are compared to the extended and prioritized standards, state standards, and state content specifications. The items are also reviewed against the Essence Statements and the Performance-Level Descriptors at all difficulty levels and compared to the blueprint. At this stage, state staff review each item and make the following decisions:

- Accept without modification ("Accept as Appears")

- Request minor revisions ("Accept as Revised")
- Request substantial changes and resubmit for a second HIDOE review ("Revise and Resubmit")
- Reject entirely (e.g., failure to meet content standards, inappropriateness for the targeted grade, general lack of clarity)

The items developed for Hawaii's contribution to shared MOU field testing in Spring 2023 were developed during spring and summer 2022. During the state review process in this development cycle, all items were accepted by HIDOE assessment staff.

The Hawaii-owned (Segment 3) items field tested in Spring 2023 were developed during the SY 2019-2020 development cycle. During the state review process in that development cycle, all items were accepted by HIDOE assessment staff except for one (1) science item.

Content and Fairness Committee Review: Following revisions and state approval, items are brought to the Content and Fairness Committee for further review. The review committee includes special educators, general educators, complex-level staff with expertise in special education, and university professors with expertise in special education. The review committees represent a diversity of gender, ethnicity, disability, race, and cultural subgroups across the state. During the review meeting, each item is reviewed to ensure that it meets bias and sensitivity guidelines, is aligned to content standards, and is determined to abide by the principles of universal design.

The common criteria used for item review are:

- Content accuracy and clarity
- Alignment to the content specifications
- Appropriate scoring rubrics
- Correct answer key and appropriate distractor(s) for each multiple-choice (MC) item
- Appropriate item format for item content
- Precision and clarity of wording in directions and items
- Appropriate graphics for color-blindness issues and standardized font size
- Accessibility for students with vision impairment
- Appropriate, fair, and unbiased content

At the beginning of each meeting, a CAI item development specialist provides a training session to ensure that the committee members understand the expectations and are familiar with the training materials that encompass the pertinent content and bias guidelines. Because the MOU shared items are used in each state for its online assessment, the committee members conduct the review online to see the item just as the student will see it.

The items developed for Hawaii's contribution to shared MOU field testing in Spring 2023 were developed during spring and summer 2022. During the educator content and fairness committee review process in that development cycle, all items were accepted by the review committee.

The Hawaii-owned items field tested in Spring 2023 were developed during the SY 2019-2020 development cycle. During the educator content and fairness committee review process in that development cycle, all items were accepted by the review committee.

.

Table 27 presents a summary of the demographics of the committee members who participated in the item content and fairness review process for the Spring 2023 field test items.

Table 27. Content and Fairness Item Review Committee Participants

Subject Area Committee		ELA	Math	Science
Total Participants		**10**	**9**	**9**
Island	Oahu	7	6	5
	Maui	1		
	Hawai'i	2	3	4
Gender	Female	9	7	7
	Male	1	2	2
Ethnicity (self-reported categories)	Asian	2	2	2
	Black		1	
	Caucasian/White	5	5	6
	Chinese	2	1	1
	Filipino			1
	Hawaiian	1	1	2
	Hispanic			1
	Japanese	1		
	Middle Eastern	1		
	Multiracial (didn't specify)	1		
	Native American			1
	Pacific Islander		1	
	Portuguese	1		
Special Education	SPED Teacher	2	3	1
	Gen Ed Teacher	6	2	4
	Higher Education	1	3	2
	Other	1	1	2
Grade Level Taught	Elementary	2	4	3
	Middle School		1	1
	High School	4		1
	College	2	3	2
	NA	2	1	2
Parent of HI Student	Yes, currently	1	2	2
	Yes, previously	2	4	2
	No	7	3	5

4.1.2. Development of Crosswalk and State Alternate Achievement Standards

Before item development began, the alternate achievement standards for each state were compared in a crosswalk created by senior test development specialists. The crosswalk was based on each state's blueprint and included the common core standards and the general education and alternate achievement standards for each state. Each state has a unique set of alternate achievement standards as follows:

- Hawai'i Essence Statements and Performance-Level Descriptors
- Idaho Extended Content Standards Core Content Connectors
- Montana Content Standards for Science
- South Carolina Prioritized Standards and Performance-Level Descriptors
- South Dakota Science Standards and Core Content Connectors
- Wyoming Content Extended Standards and Instructional Achievement-Level Descriptors

These achievement standards were examined to determine how they aligned to the general education standards and to each other. This revealed the standards to which items could be developed to meet the needs of each of the states.

The crosswalk then informed the development of item specifications. Each item specification included the Common Core or General Education standard, followed by the state-specific alternate achievement standards that align to the general education standard. The item specifications also included complexity statements and task demands. The language of the complexity statements and task demands was derived from each state's achievement standards, where applicable, and synthesized in an effort to guide the development of items that aligned to multiple states. Once completed, the item specifications were sent to each state for review to confirm alignment and overall approach.

The content extensions of the MOU states were internally examined to create a content extension crosswalk between the states. For each common standard, CAI examined the states' content extensions and PLD documents to identify which extensions were aligned to that Common Core standard. This crosswalk was used as the basis for the structure of the MOU item specifications, informing the "Common Core Standard" and "Content Extensions by State" sections of the MOU item specifications.

The states' content extensions and PLDs were further analyzed to cull relevant concepts, skills, and vocabulary. Based on the MOU states' feedback, these were compiled and displayed in the form of a Complexity matrix and a Vocabulary matrix, revealing which concepts, skills, and vocabulary were relevant to each state. The intent was to provide an "at-a-glance" perspective on content extension overlap across the states. The Complexity and Vocabulary matrices were subdivided into three categories of cognitive complexity: Low, Moderate, and High. The states' content extensions and PLDs were also analyzed to reveal state specific and cross-state content limits in the content extensions. These were listed in the Content Limits section.

All the above analysis was then used to create a numbered list of task demands describing the essential tasks students were expected to perform based on the language of the content extensions and PLDs. Additionally, these task demands were annotated with information regarding complexity and any special exceptions for individual states. A sample items section was added to the list of task demands. Each sample item was annotated with information regarding complexity and special state exceptions. Each sample item also refers to the numbered list of task demands as a reference.

4.2 FIELD TESTING

Items that survived the Content and Fairness Committee review were field tested in the spring 2023 test administration and embedded among operational items in the online adaptive tests. The operational items in the Alternate Assessment tests were administered online using a computer adaptive test (CAT) design. Tests were assembled using CAI's adaptive testing algorithm. The adaptive item selection algorithm selects items based on their content value that meets blueprint and information value that matches students' ability.

Embedding field-test items among operational items yields item parameter estimates that capture all the contextual effects contributing to item difficulty in operational test administrations. Field testing in an operational setting is beneficial in the context of a pre-equating model for scoring and reporting test results. Because the test administration context remains the same as subsequent operational test administration, item parameter estimates are more stable over time than they may be when obtained through stand-alone field testing.

Following the spring 2023 test administrations, all field-test items were calibrated anchoring on the operational items for each grade and subject, and thus, placed on the same scale as the existing operational items. Items field tested in spring 2023 were not used for scoring in the same year.

The spring 2023 field-test item pool consists of the items shared across MOU-Alt states and the unique items within each state. The items shared across MOU states were administered in all MOU states while the state-only items were administered in that state only. The number of MOU field-test items and Hawai`i-only items in the spring 2023 administration are summarized in Table 28.

Table 28. Summary of 2023 Field-Test Item Pool

Subject	Grade	State Only	MOU						
		HI	HI	ID	MT	SC	SD	WY	MOU Total
ELA	3		4	8		18		2	32
	4		4	6		18		4	32
	5		4	7		18		2	31
	6	2	4	6		17		2	29
	7		5	6		17		2	30
	8	10	4	7		17		2	30
	11		4	6		17		2	29
Mathematics	3		4	7		13		2	26
	4		4	8		12		3	27
	5		5	7		13		2	27
	6	4	6	8		12		2	28
	7		5	7		13		2	27
	8	3	4	7		13		3	27
	11	10	4	6		11		2	23
Science	5	7	22	10	5		7	30	74
	8	10	21	6	2		6	4	39
	11		8	7	3	1	7	20	46

4.2.1. Item Statistics

Following the close of the spring testing window, CAI psychometrics staff analyzed field-test data in preparation for item data/content review meetings and promotion of high-quality test items to operational item pools. Analysis of field-test items employed both the classical test and the item response theory (IRT) calibrations. Item analyses were conducted based on the combined data across MOU-Alt states.

The classical test approach is designed to evaluate the relationship of each item to the overall scale, evaluate the quality of the distractors, and identify items that may exhibit bias across subgroups (differential item functioning [DIF] analyses). The IRT item analyses allow examination of the fit of data to the measurement model and provide the statistical foundation for operational form construction and test scoring and reporting. Items are flagged if analyses indicate resulting values are out of range. Flagged items are reviewed by Hawai`i stakeholders and CAI and MOU-Alt states staff. Items that pass through the statistical review by CAI and the MOU-Alt states are accepted for future operational use.

4.2.2. Classical Statistics

Classical item analyses ensured that the field-test items function as intended with respect to the MOU-Alt states' underlying scales. CAI's analysis program computed the required item and test statistics for each dichotomous and polytomous items to check the integrity of the item and to verify the appropriateness of the difficulty level of the item. Key statistics that are computed and examined include item difficulty, item discrimination, and distractor analysis.

Items that are either extremely difficult or extremely easy are flagged for review but not necessarily rejected if they align with the test and content specifications. For dichotomous items, the proportion of test takers in the sample selecting the correct answer (*p*-value) is computed, as well as those selecting the incorrect responses. For items with 0–2 score points, item difficulty is calculated both as the item's mean score and as the average proportion correct (analogous to *p*-value and indicating the ratio of an item's mean score divided by the maximum score point possible). Items are flagged for review if the *p*-value was less than .25 or greater than .95.

The item discrimination index indicates the extent to which each item differentiates between those test takers who possess the skills being measured and those who do not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for dichotomous items was calculated as the correlation between the item score and the student's IRT-based ability estimate. For polytomous items, we computed the mean total number correct for student scoring within each of the possible score categories. Items were flagged for subsequent reviews if the biserial correlation for the keyed (correct) response is less than .20.

Distractor analysis for the dichotomous items was used to identify items that had marginal distractors or ambiguous correct responses. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. The biserial correlation for distractors is the correlation between the item score, treating the target distractor as the correct response, and the student's IRT ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items were flagged for subsequent reviews if the biserial correlation for the distractor response is greater than .05.

Table 29 summarizes all the flagging criteria based on the classical item analysis.

Table 29. Flagging Criteria Based on Classical Item Analysis

Analysis Type	Flagging Criteria
Item Discrimination	Biserial or polyserial correlation for the correct response is < 0.20.
Distractor Analysis	Point biserial correlation for any distractor response is > 0.05.
Item Difficulty	The proportion of students (*p*-value) is < 0.25 or > 0.95.
Mean Score for 2-Point items	Mean total score for a lower score point > Mean total score for a higher score point

4.2.3. Item Response Theory Statistics

Rasch and Masters' Partial Credit Model are used to estimate the item response theory (IRT) model parameters for dichotomously and polytomously scored items, respectively. The Winsteps output showing the item statistics resulting from the anchored estimation of parameters for field-test items in the operational tests were reviewed. Item fit is evaluated via the mean square Infit and mean square Outfit statistics reported by Winsteps, which are based on weighted and unweighted standardized residuals for each item response, respectively. These residual statistics indicate the discrepancy between observed item responses and the predicted item responses based on the IRT model. Both fit statistics have an expected value of 1. Values substantially greater than 1 indicate model underfit, while values substantially less than 1 indicate model overfit (Linacre, 2004). Items are flagged if Infit or Outfit values are less than 0.5 or greater than 2.0.

4.2.4. Analysis of Differential Item Functioning

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because sometimes it is a clue that an item contains a cultural or other bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in low-income areas are less likely to offer geometry classes, students at those schools might perform more poorly on geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but the curriculum. However, DIF can indicate bias, so all field-tested items were evaluated for DIF, and all items exhibiting DIF were flagged for further examination by CAI and the MOU-Alt states.

CAI conducts DIF analysis on all field-tested items to detect potential item bias across major ethnic and gender groups. For the MOU-Alt, DIF is investigated among the following group comparisons:

- Female vs. Male
- African-American vs. White
- Hispanic or Latino vs. White
- Severe and Moderate Mental Disability vs. Other. Severe and moderate Intellectual disability is defined by each state based on their primary disability code. For Hawai`i, the following four disability categories were classified as severe/moderate intellectual disability: Intellectual Disability, Multiple Disabilities, Intellectual Disability, and Traumatic Brain Injury.

CAI uses a generalized Mantel-Haenszel (MH) procedure to evaluate DIF. The generalizations include (1) adaptation to polytomous items, and (2) improved variance estimators to render the test statistics valid under complex sample designs. Because students within a district, school, and classroom are more similar

than would be expected in a simple random sample of students statewide, the information provided by students within a school is not independent, so that standard errors based on the assumption of simple random samples are underestimated. We compute design-consistent standard errors that reflect the clustered nature of educational systems. While clustering is mitigated through random administration of large numbers of embedded field-test items, design effects in student samples are rarely reduced to the level of a simple random sample.

The ability distribution is divided into a configurable number of intervals to compute the Mantel-Haenszel chi-square (*MH* χ^2) DIF statistics. The analysis program computes the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio, and the MH-delta ($\Delta_{\text{hat } MH}$) for the dichotomous items; the MH chi-square, the standardized mean difference (*SMD*), and the standard error of the *SMD* for the polytomous items.

Items are classified into three categories (A, B, or C), ranging from no evidence of DIF to severe DIF according to the DIF classification convention listed in Table 30. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, female), or negative DIF (i.e., –A, –B, or –C), signifying that the item favors the reference group (e.g., White or male). Items are flagged if their DIF statistics fall into the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. DIF classification rules are presented in Table 30. Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used when evaluating DIF classifications for items where focal or reference groups are less than 200 students (Mazor, Clauser, & Hambleton, 1992; Camilli & Shepard, 1994; Muniz, Hambleton, & Xing, 2001; Sireci & Rios, 2013).

Table 30. DIF Classification Rules

Dichotomous Items	
Category	**Rule**
C	MH_{X^2} is significant and $\lvert\hat{\Delta}_{MH}\rvert \geq 1.5$
B	MH_{X^2} is significant and $1 \leq \lvert\hat{\Delta}_{MH}\rvert < 1.5$
A	MH_{X^2} is not significant or $\lvert\hat{\Delta}_{MH}\rvert < 1$
Polytomous Items	
Category	**Rule**
C	MH_{X^2} is significant and $\lvert SMD\rvert / \lvert SD\rvert > .25$
B	MH_{X^2} is significant and $.17 < \lvert SMD\rvert / \lvert SD\rvert \leq .25$
A	MH_{X^2} is not significant or $\lvert SMD\rvert / \lvert SD\rvert \leq .17$

4.2.5. Summary of Item Statistics

This section presents a summary of results from classical item analysis and IRT item calibration analysis of the 2023 MOU-Alt field-test items conducted after the testing window closes. Table 31 presents the average sample size and the sample size at various percentiles for the MOU items. Table 32—Table 34 provide summaries of item statistics for all MOU items administered in each grade for ELA, mathematics, and science, respectively. For each item statistic, e.g., *p*-values, the percentiles are computed across items

in the corresponding subject and grade. The column "Total MOU Items" shows the number of MOU field-test items that were used in the computation. Table 35 to

Table 37 show *p*-value distributions by item type and number of response options in each grade for ELA, mathematics, and science, respectively. Table 36 provides the DIF analysis summary.

Table 31. Sample Size Distribution – MOU Items

Subject	Grade	Total MOU Items	Average Sample Size	Sample Size in Percentiles								
				Min	5th	10th	25th	50th	75th	90th	95th	Max
ELA	3	32	214	179	188	192	199	219	226	231	237	237
	4	32	232	200	200	214	221	235	243	250	258	258
	5	31	240	201	204	216	231	241	258	262	267	267
	6	29	262	238	242	242	249	261	270	280	297	301
	7	30	230	191	209	213	218	230	240	254	256	266
	8	30	240	74	217	224	242	244	255	262	267	272
	HS	29	264	203	223	226	261	267	272	282	288	290
	Overall	213	240	74	199	213	224	241	258	270	275	301
Mathematics	3	26	263	217	227	243	252	267	275	281	283	286
	4	27	274	189	220	231	259	277	296	301	306	322
	5	27	275	219	232	235	246	284	298	313	319	326
	6	28	271	183	252	252	265	272	279	297	297	298
	7	27	254	183	215	217	249	257	271	278	280	292
	8	27	268	212	229	232	243	265	291	305	315	321
	HS	23	331	259	267	301	318	337	347	357	363	379
	Overall	185	275	183	227	235	255	272	294	322	339	379
Science	ES	74	88	13	23	33	53	71	94	188	201	218
	MS	39	218	21	41	66	92	115	374	394	395	399
	HS	46	180	27	33	35	63	119	314	356	369	386
	Overall	159	146	13	27	35	58	91	203	367	383	399

Note. ES=Elementary School (grades 3-5); MS=Middle School (grades 6-8); HS=High School (grades 9 – 12).

Table 32. Summary of Item Analysis Results for MOU-Alt ELA

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
3	32	*p*-value	0.31	0.43	0.47	0.54	0.59	0.62	0.65
		Biserial/Polyserial	-0.07	0.03	0.16	0.31	0.43	0.52	0.7
		Step Difficulty	-1.03	-0.84	-0.74	-0.52	-0.16	0.07	0.51
		Infit	0.82	0.91	0.94	1.01	1.1	1.17	1.23
		Outfit	0.78	0.86	0.93	1.02	1.11	1.18	1.29
4	32	*p*-value	0.28	0.35	0.42	0.48	0.56	0.61	0.75
		Biserial/Polyserial	0.05	0.13	0.2	0.31	0.46	0.51	0.63
		Step Difficulty	-1.55	-0.92	-0.65	-0.25	0.01	0.29	0.62
		Infit	0.85	0.91	0.96	1.02	1.07	1.11	1.15
		Outfit	0.81	0.88	0.95	1.01	1.09	1.16	1.23
5	31	*p*-value	0.34	0.4	0.49	0.54	0.61	0.72	0.75
		Biserial/Polyserial	-0.17	0.15	0.21	0.36	0.47	0.53	0.69
		Step Difficulty	-1.5	-1.38	-0.8	-0.58	-0.31	-0.01	0.36
		Infit	0.86	0.91	0.94	1	1.07	1.09	1.25
		Outfit	0.81	0.85	0.9	0.99	1.09	1.13	1.41
6	29	*p*-value	0.3	0.38	0.48	0.54	0.63	0.69	0.71
		Biserial/Polyserial	0.1	0.13	0.24	0.3	0.46	0.54	0.68
		Step Difficulty	-1.24	-1.12	-0.85	-0.33	-0.09	0.36	0.77
		Infit	0.86	0.91	0.94	1.03	1.09	1.15	1.17
		Outfit	0.77	0.85	0.93	1.06	1.12	1.28	1.39
7	30	*p*-value	0.25	0.4	0.47	0.54	0.64	0.68	0.7
		Biserial/Polyserial	0.02	0.06	0.14	0.25	0.41	0.48	0.6
		Step Difficulty	-1.31	-1.26	-0.96	-0.54	-0.23	0.11	0.92
		Infit	0.88	0.95	0.96	1.07	1.12	1.16	1.21
		Outfit	0.84	0.87	0.96	1.08	1.16	1.23	1.24
8	30	*p*-value	0.19	0.39	0.47	0.6	0.64	0.71	0.79
		Biserial/Polyserial	-0.24	0.07	0.24	0.31	0.5	0.61	0.64
		Step Difficulty	-1.76	-1.39	-1.03	-0.81	-0.2	0.26	1.34
		Infit	0.84	0.87	0.94	1	1.07	1.19	1.27
		Outfit	0.75	0.81	0.88	0.97	1.13	1.28	1.66
HS	29	*p*-value	0.37	0.42	0.52	0.61	0.67	0.78	0.83
		Biserial/Polyserial	0.12	0.17	0.35	0.46	0.58	0.65	0.68
		Step Difficulty	-1.95	-1.57	-1	-0.63	-0.3	0.08	0.4
		Infit	0.84	0.88	0.89	0.95	1.03	1.12	1.18
		Outfit	0.69	0.77	0.84	0.95	1.02	1.13	1.21

Note. HS=High School (grades 9 – 12).

Table 33. Summary of Item Analysis Results for MOU-Alt Mathematics

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
3	26	*p*-value	0.28	0.32	0.37	0.45	0.57	0.62	0.64
		Biserial/Polyserial	0.08	0.1	0.18	0.24	0.32	0.38	0.51
		Step Difficulty	-1.09	-1.01	-0.82	-0.22	0.13	0.35	0.58
		Infit	0.92	0.98	1	1.04	1.08	1.11	1.18
		Outfit	0.9	0.96	0.99	1.05	1.09	1.17	1.26
4	27	*p*-value	0.27	0.3	0.33	0.47	0.54	0.61	0.74
		Biserial/Polyserial	-0.06	0.07	0.11	0.25	0.4	0.51	0.62
		Step Difficulty	-1.73	-1	-0.79	-0.4	0.17	0.42	0.48
		Infit	0.86	0.93	0.96	1.03	1.09	1.13	1.19
		Outfit	0.82	0.9	0.94	1.03	1.15	1.2	1.24
5	27	*p*-value	0.25	0.29	0.33	0.46	0.6	0.64	0.7
		Biserial/Polyserial	-0.17	-0.14	0.07	0.2	0.35	0.4	0.64
		Step Difficulty	-1.28	-1.13	-0.93	-0.29	0.29	0.5	0.77
		Infit	0.86	0.93	0.97	1.03	1.08	1.14	1.19
		Outfit	0.81	0.93	0.96	1.02	1.09	1.15	1.22
6	28	*p*-value	0.27	0.29	0.34	0.44	0.58	0.65	0.67
		Biserial/Polyserial	-0.26	-0.12	0.01	0.13	0.34	0.41	0.41
		Step Difficulty	-1.39	-1.26	-1.02	-0.3	0.1	0.31	0.48
		Infit	0.94	0.95	0.98	1.06	1.11	1.16	1.26
		Outfit	0.93	0.94	0.96	1.07	1.12	1.28	1.39
7	27	*p*-value	0.23	0.28	0.32	0.37	0.5	0.67	0.69
		Biserial/Polyserial	-0.16	-0.11	0.07	0.2	0.3	0.37	0.49
		Step Difficulty	-1.56	-1.38	-0.64	-0.05	0.11	0.36	0.64
		Infit	0.93	0.95	1	1.03	1.09	1.14	1.17
		Outfit	0.91	0.96	0.99	1.03	1.11	1.22	1.26
8	27	*p*-value	0.23	0.27	0.31	0.45	0.56	0.63	0.76
		Biserial/Polyserial	-0.23	-0.12	-0.04	0.07	0.16	0.21	0.25
		Step Difficulty	-1.85	-1.16	-0.92	-0.4	0.2	0.44	0.53
		Infit	0.98	1	1.02	1.04	1.09	1.11	1.15
		Outfit	0.95	0.98	1.02	1.05	1.12	1.14	1.18
HS	23	*p*-value	0.29	0.36	0.46	0.55	0.6	0.64	0.78
		Biserial/Polyserial	-0.17	-0.03	0.14	0.19	0.29	0.41	0.49
		Step Difficulty	-1.97	-1.15	-1.02	-0.79	-0.41	0.05	0.41
		Infit	0.89	0.92	0.99	1.03	1.07	1.14	1.2
		Outfit	0.81	0.9	0.96	1.02	1.08	1.15	1.33

Note. HS=High School (grades 9 – 12).

Table 34. Summary of Item Analysis Results for MOU-Alt Science

Grade	Total MOU Items	Statistics	Min	P10	P25	P50	P75	P90	Max
ES	74	*p*-value	0.13	0.26	0.35	0.45	0.56	0.69	0.81
		Biserial/Polyserial	-0.5	-0.03	0.14	0.3	0.53	0.63	0.94
		Step Difficulty	-1.98	-1.14	-0.55	-0.07	0.41	0.73	1.52
		Infit	0.7	0.86	0.92	1.03	1.13	1.23	1.51
		Outfit	0.52	0.79	0.89	1.04	1.16	1.26	2.86
MS	39	*p*-value	0.27	0.34	0.4	0.52	0.61	0.73	0.8
		Biserial/Polyserial	-0.01	0.11	0.18	0.29	0.49	0.57	0.83
		Step Difficulty	-1.89	-1.45	-0.86	-0.34	0.17	0.47	0.94
		Infit	0.85	0.87	0.93	1	1.09	1.14	1.2
		Outfit	0.64	0.79	0.87	0.99	1.1	1.16	1.18
HS	46	*p*-value	0.21	0.3	0.39	0.49	0.6	0.67	0.79
		Biserial/Polyserial	-0.2	-0.09	0.11	0.28	0.4	0.6	0.82
		Step Difficulty	-1.55	-1.19	-0.73	-0.25	0.27	0.79	1.54
		Infit	0.79	0.85	0.96	1.02	1.12	1.25	1.34
		Outfit	0.63	0.8	0.92	1.04	1.14	1.32	1.59

Note. ES=Elementary School (grades 3 – 5); MS= Middle School (grades 6-8); HS=High School (grades 9-12).

Table 35. p-value by Item Type/Number of Response Options for MOU-Alt ELA

Grade	Item Type	Number of Response Options	N	Percentage	Min	P10	P25	P50	P75	P90	Max
3	multipleChoice	2	26	81.30%	0.38	0.46	0.5	0.55	0.6	0.63	0.65
	multipleChoice	3	6	18.80%	0.31	0.31	0.38	0.43	0.43	0.51	0.51
		Total	32	100.00%	0.31	0.43	0.47	0.54	0.59	0.62	0.65
4	multipleChoice	2	7	21.90%	0.52	0.52	0.53	0.61	0.68	0.75	0.75
	multipleChoice	3	25	78.10%	0.28	0.35	0.38	0.44	0.51	0.56	0.6
		Total	32	100.00%	0.28	0.35	0.42	0.48	0.56	0.61	0.75
5	multipleChoice	2	19	61.30%	0.49	0.51	0.54	0.59	0.7	0.74	0.75
	multipleChoice	3	12	38.70%	0.34	0.39	0.4	0.46	0.52	0.57	0.59
		Total	31	100.00%	0.34	0.4	0.49	0.54	0.61	0.72	0.75
6	multipleChoice	2	16	55.20%	0.51	0.51	0.55	0.61	0.67	0.7	0.71
	multipleChoice	3	13	44.80%	0.3	0.36	0.41	0.48	0.5	0.6	0.68
		Total	29	100.00%	0.3	0.38	0.48	0.54	0.63	0.69	0.71
7	multipleChoice	2	22	73.30%	0.42	0.43	0.53	0.58	0.66	0.68	0.7
	multipleChoice	3	8	26.70%	0.25	0.25	0.38	0.49	0.5	0.53	0.53
		Total	30	100.00%	0.25	0.4	0.47	0.54	0.64	0.68	0.7
8	multipleChoice	2	16	53.30%	0.45	0.47	0.56	0.64	0.69	0.72	0.79
	multipleChoice	3	14	46.70%	0.19	0.29	0.39	0.5	0.63	0.7	0.71
		Total	30	100.00%	0.19	0.39	0.47	0.6	0.64	0.71	0.79
HS	multipleChoice	2	17	58.60%	0.49	0.5	0.58	0.62	0.72	0.78	0.83
	multipleChoice	3	12	41.40%	0.37	0.41	0.43	0.54	0.62	0.63	0.71
		Total	29	100.00%	0.37	0.42	0.52	0.61	0.67	0.78	0.83

Note. HS=High School (grades 9-12).

Table 36. *p*-value by Item Type/Number of Response Options for MOU-Alt Mathematics

Grade	Item Type	Number of Response Options	N	Percentage	Min	P10	P25	P50	P75	P90	Max
3	multipleChoice	2	11	42.30%	0.44	0.47	0.49	0.59	0.62	0.64	0.64
	multipleChoice	3	15	57.70%	0.28	0.29	0.33	0.37	0.43	0.46	0.5
		Total	26	100.00%	0.28	0.32	0.37	0.45	0.57	0.62	0.64
4	multipleChoice	2	12	44.40%	0.47	0.47	0.52	0.55	0.6	0.64	0.74
	multipleChoice	3	15	55.60%	0.27	0.28	0.32	0.34	0.4	0.47	0.54
		Total	27	100.00%	0.27	0.3	0.33	0.47	0.54	0.61	0.74
5	multipleChoice	2	12	44.40%	0.46	0.51	0.55	0.61	0.64	0.66	0.7
	multipleChoice	3	15	55.60%	0.25	0.27	0.31	0.36	0.41	0.47	0.5
		Total	27	100.00%	0.25	0.29	0.33	0.46	0.6	0.64	0.7
6	multipleChoice	2	12	42.90%	0.37	0.54	0.56	0.60	0.65	0.66	0.67
	multipleChoice	3	16	57.10%	0.27	0.29	0.30	0.37	0.42	0.48	0.52
		Total	28	100.00%	0.27	0.29	0.34	0.44	0.58	0.65	0.67
7	multipleChoice	2	7	25.90%	0.43	0.43	0.54	0.55	0.69	0.69	0.69
	multipleChoice	3	20	74.10%	0.23	0.27	0.31	0.36	0.37	0.46	0.5
		Total	27	100.00%	0.23	0.28	0.32	0.37	0.5	0.67	0.69
8	multipleChoice	2	15	55.60%	0.39	0.43	0.45	0.55	0.59	0.71	0.76
	multipleChoice	3	12	44.40%	0.23	0.26	0.27	0.3	0.33	0.39	0.47
		Total	27	100.00%	0.23	0.27	0.31	0.45	0.56	0.63	0.76
HS	multipleChoice	2	18	78.30%	0.44	0.49	0.54	0.56	0.61	0.67	0.78
	multipleChoice	3	5	21.70%	0.29	0.29	0.3	0.36	0.42	0.46	0.46
		Total	23	100.00%	0.29	0.36	0.46	0.55	0.6	0.64	0.78

Note. HS=High School (grades 9-12).

Table 37. *p*-value by Item Type/Number of Response Options for MOU-Alt Science

Grade	Item Type	Number of Response Options	N	Percentage	Min	P10	P25	P50	P75	P90	Max
ES	multipleChoice	2	9	12.20%	0.41	0.41	0.59	0.65	0.72	0.81	0.81
	multipleChoice	3	60	81.10%	0.13	0.27	0.34	0.43	0.53	0.59	0.79
	multipleChoice	4	3	4.10%	0.25	0.25	0.25	0.25	0.27	0.27	0.27
	multipleSelect	5	2	2.70%	0.35	0.35	0.35	0.52	0.69	0.69	0.69
		Total	74	100.00%	0.13	0.26	0.35	0.45	0.56	0.69	0.81
MS	multipleChoice	2	11	28.20%	0.29	0.59	0.59	0.64	0.73	0.79	0.8
	multipleChoice	3	27	69.20%	0.27	0.34	0.36	0.43	0.55	0.61	0.8
	multipleChoice	4	1	2.60%	0.55	0.55	0.55	0.55	0.55	0.55	0.55
		Total	39	100.00%	0.27	0.34	0.4	0.52	0.61	0.73	0.8
HS	multipleChoice	2	16	34.80%	0.43	0.49	0.51	0.61	0.67	0.71	0.79
	multipleChoice	3	29	63.00%	0.21	0.26	0.33	0.42	0.5	0.64	0.68
	multipleSelect	5	1	2.20%	0.45	0.45	0.45	0.45	0.45	0.45	0.45
		Total	46	100.00%	0.21	0.3	0.39	0.49	0.6	0.67	0.79

Note. ES=Elementary School (grades 3 – 5); MS= Middle School (grades 6-8); HS=High School (grades 9-12).

Table 38. Number of Items in Each DIF Classification Category

Female vs Male								**African American vs. White**							
Subject/Grade	Total	+A	-A	+B	-B	+C	-C	Subject/Grade	Total	+A	-A	+B	-B	+C	-C
ELA	32	18	14					**ELA**	14	6	5			2	1
3	32	12	18			2		3	18	5	13				
4	31	17	14					4	31	17	14				
5	29	14	13		1		1	5	21	9	11				1
6	30	14	14			2		6	13	6	7				
7	29	18	10				1	7	27	10	17				
8	29	15	13			1		8	29	9	16				4
HS	26	13	12		1			HS	26	11	14			1	
Mathematics	27	12	15					**Mathematics**	25	14	11				
3	27	12	15					3	27	18	9				
4	28	14	14					4	24	8	14			1	1
5	27	13	14					5	22	8	12				2
6	27	13	14					6	27	9	18				
7	23	13	8		2			7	23	10	13				
8	15	9	6					8	2	1	1				
HS	18	7	10				1	HS	18	5	12		1		
Science	22	11	11					**Science**	19	10	9				
ES	32	18	14					ES	14	6	5			2	1
MS	32	12	18			2		MS	18	5	13				
HS	31	17	14					HS	31	17	14				
Hispanic vs. White								**Severe/Moderate Disability vs.Other**							
Subject/Grade	Total	+A	-A	+B	-B	+C	-C	Subject/Grade	Total	+A	-A	+B	-B	+C	-C
ELA								**ELA**	1						1
3								3	26	16	10				
4								4	29	9	19				1
5								5	29	12	15				2
6	6	2	3				1	6	30	14	15				1
7								7	29	11	17				1
8								8	29	9	18				2
HS								HS	14	9	5				
Mathematics								**Mathematics**	24	12	12				
3	6	5	1					3	26	9	16				1
4	1						1	4	28	12	13			1	2
5	4	2	2					5	26	9	16				1
6	6	4	2					6	27	15	12				
7								7	23	14	5			1	3
8								8	7	4	3				
HS	1			1				HS	18	5	11			1	1
Science								**Science**	19	7	12				
ES								ES	1						1
MS	4		4					MS	26	16	10				
HS	2	1	1					HS	29	9	19				1

Note. This table only includes items with sample size >=50 in both the focal and reference groups. ES=Elementary School (grades 3 – 5); MS= Middle School (grades 6-8); HS=High School (grades 9-12).

4.2.6. Item Data Review

MOU-Alt Shared Item

Items flagged for undesired statistics were reviewed in the MOU-Alt and Hawai`i stakeholder item data review committees. CAI flagged and removed the items with the sample size less than 50 or negative biserial/polyserial correlations for the key. These items were removed from the item pool before data review, and were not seen by the data review committees. In addition to the flagged items, the Hawai`i stakeholder item data review committee also reviewed all items with desired statistics that were not flagged.

The Hawai`i stakeholder item data/content review (IDCR) committee included special education teachers, content-area experts, advocates, and community members who work with individuals with significant cognitive disabilities. The MOU-Alt data review committee consisted of staff across MOU states, and CAI content specialists, special education specialists, and psychometricians. During the meetings, the committees were charged with identifying any defects that might have led to the undesired statistics of the items and then asked to render a decision on the items. Committees could choose to reject the item completely, accept the item with modifications for further field testing, or accept the item without any changes. Items accepted without modification were included in the Hawai`i state operational item pool.

Table 39 presents a summary of the MOU-Alt data review results.

Table 39. Summary of Item Data/Content Review: MOU-Alt Item Pool

Subject	Grade	Total Number of MOU Items	Items with N < 50	Items with biserial < 0	Total Number of Reviewed Items for IDR	Items Rejected by IDR Committee
ELA	3	32	0	3	10	0
	4	32	0	0	13	0
	5	31	0	1	7	0
	6	29	0	0	9	0
	7	30	0	0	13	1
	8	30	0	2	6	0
	HS	29	0	0	9	0
Mathematics	3	26	0	0	11	0
	4	27	0	2	10	0
	5	27	0	3	12	0
	6	28	0	7	13	0
	7	27	0	4	12	0
	8	27	0	11	10	0
	HS	23	0	3	11	0
Science	5	74	14	5	20	1
	8	39	3	0	14	2
	HS	46	8	6	8	0

Hawai`i Item Pool

In addition to the MOU Alt items, the Hawai`i stakeholder item data/content review committee also examined items field-tested only in Hawai`i, developed by CAI specifically for use in Hawai`i's item pool. Committee members reviewed all flagged Hawai`i-only items for their subject area alongside review of the shared MOU items. Table 40 presents a summary of all items field tested in Hawai`i.

Table 40. Summary of Item Data/Content Review: Hawai`i Item Pool

Grade	Total # of Items Tested	# Items with n < 50	# Items with biserial < 0	# Items Rejected	# Items Eligible for Operational Use
ELA					
3	32	0	3	2	27
4	32	0	0	3	29
5	31	0	1	2	28
6	31	0	0	5	26
7	29	0	0	1	28
8	40	0	2	1	37
11	28	0	0	2	26
Mathematics					
3	25	0	0	2	23
4	25	0	1	1	23
5	26	0	3	3	20
6	31	0	8	3	20
7	25	0	4	3	18
8	30	0	11	1	18
11	33	0	7	3	23
Science					
5	62	4	5	9	44
8	43	1	3	0	39
11	15	0	3	3	9

Table 41 presents a summary of the demographics of the committee members who participated in the item data review process for the Spring 2023 field test items in summer 2023.

Table 41. Item Data/Content Review Committee Participants

Subject Area Committee		ELA	Mathematics	Science
Total Participants		**6**	**10**	**4**
Island	Oahu	3	7	1
	Maui	2		1
	Hawai'i	1	3	2
Gender	Female	5	9	3
	Male	1	1	1
Ethnicity (self-reported categories)	Asian	1	2	1
	Caucasian/White	2	5	2
	Chinese	1	1	
	Hawaiian		1	
	Hispanic		1	
	Japanese	2		
	Multiracial (didn't specify)		1	
	Native American		1	
	Pacific Islander		1	1
	Portuguese	1		
	Other	1		
	Declined			
Special Education	SPED Teacher	1	3	1
	Gen Ed Teacher	5	4	1
	Higher Education		2	2
	Other		1	
Grade Level Taught	Elementary	2	4	
	Middle School	1	2	1
	High School	3	1	1
	College		1	2
	NA		2	
Parent of HI Student	Yes, currently	2	2	1
	Yes, previously	2	3	1
	No	2	5	2

4.3 SCALING AND EQUATING

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. Traditional item response models assume a single underlying trait and assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

$$L(Z) = \prod_{j=1}^{n} P(z|\theta),$$

where *Z* represents the pattern of item responses, and θ represents a student's true proficiency.

Traditional item response models differ only in the form of the function *P*(*Z*). The one-parameter model (1PL; also known as the Rasch model), is used to calibrate MOU-Alt items that are scored either right or wrong, and takes the form

$$P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)},$$

where b_i is the difficulty parameter for item i.

The *b* parameter is often called the *location* or *difficulty* parameter; the greater the value of *b*, the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered-response categories (i.e., partial credit items), MOU-Alt items were calibrated using the Rasch family Masters' (1982) partial credit model. Under Masters' partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

$$P(X_i = x_i|\theta) = \frac{\exp\sum_{k=0}^{x_i}(\theta - b_{ki})}{\sum_{l=0}^{m_i}\exp\sum_{k=0}^{l}(\theta - b_{ki})},$$

with the constraint that $\sum_{k=0}^{0}(\theta - b_{ki}) \equiv 0$. b_{ki} is item location parameter for category k of item i.

4.3.1. Item Calibration

The field-test items were calibrated by anchoring on the operational item parameters under the CAT test design. Through this anchoring process, field-test item parameter estimates were placed on the same MOU scale as the operational items. These operational item parameters will be used as a reference scale to calibrate new items in the following years.

Winsteps was used to estimate Rasch and Masters' partial credit model item parameters for the MOU-Alt. Winsteps is a publicly available software program from Mesa Press. Winsteps employs a joint maximum likelihood approach towards estimation (JMLE), which jointly estimates the person and item parameters. The Rasch model estimates the parameters for student responses to dichotomous (0/1 point) items. Masters'

(1982) partial credit model, an extension of the one parameter Rasch model which allows for partial credit to be given on items, estimates the responses for polytomous items. All completed records are included in IRT analysis.

5. VALIDITY

According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014; hereafter referred to as the Standards), "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p.11). Statements about validity should refer to particular interpretations for specified uses, and thus, the validation process starts logically with well-articulated statements on intended uses of test scores. Arguments of logic, theoretical, and empirical evidence are then provided to support the intended uses.

This chapter will first present the statements on intended uses of the HSA-Alt test scores, followed by various sources of evidence validating the interpretation of test scores for the intended uses.

5.1 INTENDED USES AND INTERPRETATIONS OF THE HSA-ALT SCORES

Development and design of the HSA-Alt assessments are reflected in a theory of action that begins by answering fundamental questions about the purpose, uses, interpretations, and outcomes of the test and integrates evidence comprised of theoretical, logical, and empirical components.

The intended uses of the HSA-Alt score include:

- measuring students' academic achievements and progress in core content areas taught in school,
- measuring achievement and progress toward meeting the state performance standards, and
- monitoring the education system and make necessary improvement to meet federal accountability requirements.

Intended test users include students and parents who would like to be informed of the students' learning progress in school; teachers and other educators in school who can use testing results to guide in-class instruction and identify students who need more help; educational agencies, organizations, and governments who monitor the education system and make necessary changes in standards.

In realizing the uses, HSA-Alt provides an overall scale score and an associated performance level for each test taken. The performance level is determined based on the performance standards that are set through a formal standard setting process. Validity evidence on measuring achievement and progress toward meeting the state performance standards is documented separately in greater detail in the standard setting technical report. Chapter 8 in this technical report provides a high-level overview on the standard setting procedure and results.

5.2 SOURCES OF VALIDITY EVIDENCE

According to the Standards (AERA, APA, & NCME, 2014), "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which exiting evidence and theory support the intended interpretation of test scores for specific uses" (p. 21). Validity of an intended interpretation of test scores relies on all the evidence accrued about the technical quality of a testing system, including test development and construction procedures, test score reliability, accurate scaling and equating, procedures for setting meaningful performance standards, standardized test administration and scoring procedures, and attention to fairness for all test takers. The appropriateness and usefulness of the HSA-Alt depends on the assessments meeting the relevant standards of validity.

Providing sufficient and solid validity evidence is also required of the State to meet federal peer review requirements. In the guidance provided by the United States Department of Education for assessing peer review process (U.S. Department of Education, 2018), the requirements related to validity are represented by three critical elements.

Validity evidence for the HSA-Alt are gathered from the following four sources, as outlined in the Standards. The particular critical element in the peer review guidance corresponding to each source is included in the parenthesis.

- Evidence based on test content (Critical Element 3.1- Overall Validity, Including Validity Based on Content)
- Evidence based on response processes (Critical Element 3.2- Validity Based on Cognitive Process/Linguistic Processes)
- Evidence based on internal structure (Critical Element 3.3- Validity Based on Internal Structure)
- Evidence based on relations to other variables (Critical Element 3.4- Validity Based on Relations to Other Variables)

Evidence on test content validity is provided with both theorical and empirical evidence related to content specifications, test specifications, blueprints, item and test development process, administration process, and scoring. Evidence on response processes is gathered by conducting cognitive laboratory studies of student response to items. Evidence on internal structure is examined in the results of intercorrelations among content strand scores. Evidence on relations to other variables is provided with the correlations between test scores and Learner Characteristics Inventory (LCI) and Hawai`i Observational Rating Assessment (HIORA) questions.

5.2.1. Evidence Based on Test Content

Content evidence for validity is based on the appropriateness of test content and the procedures used to create test content, which should be well aligned with the required state-wide standards implemented in daily instruction at school by teachers. This evidence is based on the justification for and connections among several factors listed below:

- Content specifications
- Test blueprints
- Item development
- Test administration conditions, and
- Item and test scoring

These resources are developed by content and measurement experts and are consistent with state standards. Collectively, they help connect the assessment results to learning and instruction. The descriptions of the evidence, most of which are documented in early chapters, are summarized as follows.

Content Specifications

Content standards and content specification is the starting point for test development. The HSA-Alt is developed based on the Hawai'i Common Core Standards and designed for students with the most significant cognitive disabilities. The purpose of the HSA-Alt is to maximize access of this student population to the general education curriculum, ensure that all students with disabilities are included in the

statewide assessments, and make certain that they are included in the educational accountability system. See Section 1.3 – Content Specification in this technical report for details.

Test Blueprints

Test blueprints specify the content standards to be covered in the test, and the minimum and maximum number of items in each content domain. The goal is to ensure the test has a balanced representation of items from each content standard.

For the HSA-Alt in all three subjects, each student receives 40 operational items, 10 field-test items from the MOU pool, and 1-10 field-test items from the Hawai`i-specific item pool. Only operational items contribute to student scores. In the adaptive algorithm used on the operational items, item selection takes place in two stages: blueprint satisfaction and match-to-ability.

The blueprint match rates are provided for the operational tests. Table 42 —Table 58 present the percentages of administered tests aligned with the test blueprint constraints for ELA, mathematics, and science. The blueprint match rates are based on the completed online adaptive tests only. The adaptive algorithm selected items for all tests according to the blueprint requirements (100% blueprint match) at the overall strand level.

Table 42. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 3 ELA

Strand	Benchmark	Grade 3					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.3.1	0	1	100	0	2	100
	L.3.2	0	1	100	0	2	100
	L.3.3	0	1	100	0	2	100
	L.3.4	0	1	100	0	2	86
	L.3.5	0	1	100	0	2	100
	L.3.6	0	1	100	0	2	100
Reading – Informational (RI)	Overall	3	3	100	8	9	100
	RI.3.1	0	1	100	0	2	100
	RI.3.2	0	1	100	0	2	100
	RI.3.3	0	1	100	0	2	99
	RI.3.4	0	1	100	0	2	100
	RI.3.5	0	1	100	0	2	100
	RI.3.6	0	1	100	0	2	100
	RI.3.7	0	1	100	0	2	99
	RI.3.8	0	1	100	0	2	100
	RI.3.9	0	1	100	0	2	100
Reading – Literature (RL)	Overall	3	3	100	8	9	100
	RL.3.1	0	1	100	0	2	89
	RL.3.2	0	1	100	0	2	96
	RL.3.3	0	1	100	0	2	100
	RL.3.4	0	1	100	0	2	100
	RL.3.6	0	1	100	0	2	100
	RL.3.7	0	1	100	0	2	100
	RL.3.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.3.1	0	1	100	0	2	96
	W.3.2	0	1	100	0	2	98
	W.3.3	0	1	100	0	2	100
	W.3.7	0	1	100	0	2	100
	W.3.8	0	1	100	0	2	100

Table 43. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 4 ELA

Strand	Benchmark	Grade 4					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.4.1	0	1	100	0	2	100
	L.4.2	0	1	100	0	2	100
	L.4.3	0	1	100	0	2	100
	L.4.4	0	1	100	0	2	98
	L.4.5	0	1	100	0	2	100
	L.4.6	0	1	100	0	2	100
Reading – Informational (RI)	Overall	3	3	100	8	9	100
	RI.4.1	0	1	100	0	2	94
	RI.4.2	0	1	100	0	2	100
	RI.4.3	0	1	100	0	2	100
	RI.4.4	0	1	100	0	2	99
	RI.4.5	0	1	100	0	2	100
	RI.4.6	0	1	100	0	2	100
	RI.4.7	0	1	100	0	2	98
	RI.4.8	0	1	100	0	2	100
	RI.4.9	0	1	100	0	2	100
Reading – Literature (RL)	Overall	3	3	100	8	9	100
	RL.4.1	0	1	100	0	2	73
	RL.4.2	0	1	100	0	2	100
	RL.4.3	0	1	100	0	2	96
	RL.4.4	0	1	100	0	2	100
	RL.4.6	0	1	100	0	2	99
	RL.4.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.4.1	0	1	100	0	2	100
	W.4.2	0	1	100	0	2	95
	W.4.3	0	1	100	0	2	100
	W.4.4	0	1	100	0	2	100
	W.4.7	0	1	100	0	2	100
	W.4.8	0	1	100	0	2	93

Table 44. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 5 ELA

Strand	Benchmark	Grade 5					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.5.2	0	1	100	0	2	100
	L.5.3	0	1	100	0	2	99
	L.5.4	0	1	100	0	2	100
	L.5.5	0	1	100	0	2	100
	L.5.6	0	1	100	0	2	99
Reading – Informational (RI)	Overall	3	3	100	8	9	100
	RI.5.1	0	1	100	0	2	100
	RI.5.2	0	1	100	0	2	70
	RI.5.3	0	1	100	0	2	100
	RI.5.4	0	1	100	0	2	100
	RI.5.5	0	1	100	0	2	100
	RI.5.6	0	1	100	0	2	100
	RI.5.7	0	1	100	0	2	100
	RI.5.8	0	1	100	0	2	100
	RI.5.9	0	1	100	0	2	100
Reading – Literature (RL)	Overall	3	3	100	8	9	100
	RL.5.1	0	1	100	0	2	100
	RL.5.2	0	1	100	0	2	89
	RL.5.3	0	1	100	0	2	100
	RL.5.4	0	1	100	0	2	99
	RL.5.6	0	1	100	0	2	98
	RL.5.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.5.1	0	1	100	0	2	100
	W.5.2	0	1	100	0	2	99
	W.5.3	0	1	100	0	2	99
	W.5.4	0	1	100	0	2	100
	W.5.7	0	1	100	0	2	100
	W.5.8	0	1	100	0	2	100

Table 45. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 6 ELA

Strand	Benchmark	Grade 6					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.6.1	0	1	100	0	2	100
	L.6.2	0	1	100	0	2	100
	L.6.4	0	1	100	0	2	100
	L.6.5	0	1	100	0	2	100
	L.6.6	0	1	100	0	2	100
Reading – Informational (RI)	Overall	3	3	100	8	10	100
	RI.6.1	0	1	100	0	2	83
	RI.6.2	0	1	100	0	2	100
	RI.6.3	0	1	100	0	2	100
	RI.6.4	0	1	100	0	2	100
	RI.6.5	0	1	100	0	2	100
	RI.6.6	0	1	100	0	2	100
	RI.6.8	0	1	100	0	2	100
	RI.6.9	0	1	100	0	2	100
Reading – Literature (RL)	Overall	3	3	100	7	9	100
	RL.6.1	0	1	100	0	2	100
	RL.6.2	0	1	100	0	2	100
	RL.6.3	0	1	100	0	2	100
	RL.6.4	0	1	100	0	2	98
	RL.6.6	0	1	100	0	2	100
	RL.6.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.6.1	0	1	100	0	2	100
	W.6.2	0	1	100	0	2	100
	W.6.3	0	1	100	0	2	100
	W.6.4	0	1	100	0	2	100
	W.6.7	0	1	100	0	2	100
	W.6.8	0	1	100	0	2	100

Table 46. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 7 ELA

Strand	Benchmark	Grade 7					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.7.1	0	1	100	0	2	100
	L.7.2	0	1	100	0	2	100
	L.7.4	0	1	100	0	2	99
	L.7.5	0	1	100	0	2	99
	L.7.6	0	1	100	0	2	94
Reading – Informational (RI)	Overall	3	3	100	8	10	100
	RI.7.1	0	1	100	0	2	95
	RI.7.2	0	1	100	0	2	100
	RI.7.3	0	1	100	0	2	100
	RI.7.4	0	1	100	0	2	100
	RI.7.5	0	1	100	0	2	99
	RI.7.6	0	1	100	0	2	100
	RI.7.8	0	1	100	0	2	100
	RI.7.9	0	1	100	0	2	99
Reading – Literature (RL)	Overall	3	3	100	7	9	100
	RL.7.1	0	1	100	0	2	97
	RL.7.2	0	1	100	0	2	100
	RL.7.3	0	1	100	0	2	96
	RL.7.4	0	1	100	0	2	96
	RL.7.6	0	1	100	0	2	100
	RL.7.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.7.1	0	1	100	0	2	100
	W.7.2	0	1	100	0	2	100
	W.7.3	0	1	100	0	2	100
	W.7.4	0	1	100	0	2	100
	W.7.7	0	1	100	0	2	98
	W.7.8	0	1	100	0	2	100

Table 47. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 8 ELA

Strand	Benchmark	Grade 8					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.8.1	0	1	100	0	2	100
	L.8.2	0	1	100	0	2	100
	L.8.4	0	1	100	0	2	100
	L.8.5	0	1	100	0	2	100
	L.8.6	0	1	100	0	2	100
Reading – Informational (RI)	Overall	3	3	100	8	10	100
	RI.8.1	0	1	100	0	2	99
	RI.8.2	0	1	100	0	2	100
	RI.8.3	0	1	100	0	2	100
	RI.8.4	0	1	100	0	2	100
	RI.8.5	0	1	100	0	2	100
	RI.8.6	0	1	100	0	2	100
	RI.8.8	0	1	100	0	2	100
	RI.8.9	0	1	100	0	2	100
Reading – Literature (RL)	Overall	3	3	100	7	9	100
	RL.8.1	0	1	100	0	2	100
	RL.8.2	0	1	100	0	2	100
	RL.8.3	0	1	100	0	2	100
	RL.8.4	0	1	100	0	2	100
	RL.8.6	0	1	100	0	2	100
	RL.8.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.8.1	0	1	100	0	2	100
	W.8.2	0	1	100	0	2	100
	W.8.3	0	1	100	0	2	100
	W.8.4	0	1	100	0	2	100
	W.8.7	0	1	100	0	2	100
	W.8.8	0	1	100	0	2	100

Table 48. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 11 ELA

Strand	Benchmark	Grade 11					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Language (L)	Overall	1	1	100	7	9	100
	L.11-12.1	0	1	100	0	2	100
	L.11-12.2	0	1	100	0	2	100
	L.11-12.4	0	1	100	0	2	100
	L.11-12.5	0	1	100	0	2	100
	L.11-12.6	0	1	100	0	2	100
Reading – Informational (RI)	Overall	3	3	100	10	12	100
	RI.11-12.1	0	1	100	0	3	100
	RI.11-12.2	0	1	100	0	3	95
	RI.11-12.3	0	1	100	0	3	95
	RI.11-12.4	0	1	100	0	3	100
	RI.11-12.6	0	1	100	0	3	83
	RI.11-12.8	0	1	100	0	3	100
	RI.11-12.9	0	1	100	0	3	100
Reading – Literature (RL)	Overall	3	3	100	6	8	100
	RL.11-12.1	0	1	100	0	2	98
	RL.11-12.2	0	1	100	0	2	100
	RL.11-12.3	0	1	100	0	2	100
	RL.11-12.4	0	1	100	0	2	93
	RL.11-12.6	0	1	100	0	2	100
	RL.11-12.9	0	1	100	0	2	100
Writing (W)	Overall	1	1	100	7	9	100
	W.11-12.1	0	1	100	0	2	95
	W.11-12.2	0	1	100	0	2	98
	W.11-12.3	0	1	100	0	2	100
	W.11-12.4	0	1	100	0	2	100
	W.11-12.7	0	1	100	0	2	100
	W.11-12.8	0	1	100	0	2	100

Table 49. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 3 Mathematics

Strand	Benchmark	Grade 3					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Geometry (G)	Overall	1	1	100	2	3	100
	3.G.A.1	0	1	100	0	2	100
	3.G.A.2	0	1	100	0	2	100
Measurement and Data (MD)	Overall	2	2	100	7	8	100
	3.MD.A.1	0	1	100	0	2	100
	3.MD.A.2	0	1	100	0	2	100
	3.MD.B.3	0	1	100	0	1	100
	3.MD.B.4	0	1	100	0	1	100
	3.MD.C.6	0	1	100	0	2	100
	3.MD.C.7d	0	1	100	0	2	100
	3.MD.D.8	0	1	100	0	1	100
Number and Operations in Base Ten (NBT)	Overall	1	1	100	3	4	100
	3.NBT.A.1	0	1	100	0	2	100
	3.NBT.A.2	0	1	100	0	2	100
	3.NBT.A.3	0	1	100	0	2	100
Numbers and Operations – Fractions (NF)	Overall	2	2	100	6	7	100
	3.NF.A.1	0	1	100	0	3	100
	3.NF.A.2a	0	1	100	0	3	100
	3.NF.A.3	0	1	100	0	3	100
Operations and Algebraic Thinking (OA)	Overall	2	2	100	10	11	100
	3.OA.A.1	0	1	100	0	2	100
	3.OA.A.2	0	1	100	0	2	100
	3.OA.A.3	0	1	100	0	2	100
	3.OA.A.4	0	1	100	0	2	100
	3.OA.B.5	0	1	100	0	2	100
	3.OA.B.6	0	1	100	0	2	100
	3.OA.C.7	0	1	100	0	2	100
	3.OA.D.8	0	1	100	0	2	100
	3.OA.D.9	0	1	100	0	2	100

Table 50. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 4 Mathematics

Strand	Benchmark	Grade 4					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Geometry (G)	Overall	1	1	100	2	3	100
	4.G.A.1	0	1	100	0	1	100
	4.G.A.2	0	1	100	0	1	100
	4.G.A.3	0	1	100	0	1	100
Measurement and Data (MD)	Overall	1	1	100	4	5	100
	4.MD.A.1	0	1	100	0	1	100
	4.MD.A.2	0	1	100	0	1	100
	4.MD.A.3	0	1	100	0	1	100
	4.MD.B.4	0	1	100	0	1	100
	4.MD.C.6	0	1	100	0	1	100
	4.MD.C.7	0	1	100	0	1	100
Number and Operations in Base Ten (NBT)	Overall	2	2	100	7	8	100
	4.NBT.A.1	0	1	100	0	2	100
	4.NBT.A.2	0	1	100	0	2	100
	4.NBT.A.3	0	1	100	0	2	100
	4.NBT.B.4	0	1	100	0	2	100
	4.NBT.B.5	0	1	100	0	2	100
	4.NBT.B.6	0	1	100	0	2	100
Numbers and Operations – Fractions (NF)	Overall	3	3	100	11	13	100
	4.NF.A.1	0	1	100	0	2	100
	4.NF.A.2	0	1	100	0	2	100
	4.NF.B.3b	0	1	100	0	2	100
	4.NF.B.3c	0	1	100	0	2	100
	4.NF.B.3d	0	1	100	0	2	100
	4.NF.B.4c	0	1	100	0	2	100
	4.NF.C.5	0	1	100	0	2	100
	4.NF.C.6	0	1	100	0	2	100
	4.NF.C.7	0	1	100	0	2	100
Operations and Algebraic Thinking (OA)	Overall	1	1	100	6	7	100
	4.OA.A.1	0	1	100	0	2	100
	4.OA.A.2	0	1	100	0	2	100
	4.OA.A.3	0	1	100	0	2	100
	4.OA.B.4	0	1	100	0	1	100
	4.OA.C.5	0	1	100	0	1	100

Table 51. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 5 Mathematics

Strand	Benchmark	Grade 5					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Geometry (G)	Overall	1	1	100	4	5	100
	5.G.A.1	0	1	100	0	2	100
	5.G.A.2	0	1	100	0	2	100
	5.G.B.3	0	1	100	0	2	100
	5.G.B.4	0	1	100	0	2	100
Measurement and Data (MD)	Overall	1	1	100	4	5	100
	5.MD.A.1	0	1	100	0	1	100
	5.MD.B.2	0	1	100	0	1	100
	5.MD.C.4	0	1	100	0	2	100
	5.MD.C.5a	0	1	100	0	2	100
	5.MD.C.5b	0	1	100	0	2	100
	5.MD.C.5c	0	1	100	0	2	100
Number and Operations in Base Ten (NBT)	Overall	2	2	100	8	9	100
	5.NBT.A.1	0	1	100	0	2	100
	5.NBT.A.2	0	1	100	0	2	100
	5.NBT.A.3a	0	1	100	0	2	100
	5.NBT.A.3b	0	1	100	0	2	100
	5.NBT.A.4	0	1	100	0	2	100
	5.NBT.B.5	0	1	100	0	2	100
	5.NBT.B.6	0	1	100	0	2	100
	5.NBT.B.7	0	1	100	0	2	100
Numbers and Operations – Fractions (NF)	Overall	3	3	100	9	11	100
	5.NF.A.1	0	1	100	0	2	99
	5.NF.A.2	0	1	100	0	2	100
	5.NF.B.3	0	1	100	0	2	100
	5.NF.B.4	0	1	100	0	2	100
	5.NF.B.4b	0	1	100	0	2	100
	5.NF.B.6	0	1	100	0	2	100
	5.NF.B.7	0	1	100	0	2	100
Operations and Algebraic Thinking (OA)	Overall	1	1	100	3	4	100
	5.OA.A.1	0	1	100	0	2	100
	5.OA.A.2	0	1	100	0	2	100
	5.OA.B.3	0	1	100	0	2	100

Table 52. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 6 Mathematics

Strand	Benchmark	Grade 6					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Expressions and Equations (EE)	Overall	2	2	100	7	8	100
	6.EE.A.1	0	1	100	0	2	100
	6.EE.A.2a	0	1	100	0	2	100
	6.EE.A.2b	0	1	100	0	2	100
	6.EE.A.2c	0	1	100	0	2	100
	6.EE.A.3	0	1	100	0	2	100
	6.EE.B.5	0	1	100	0	2	100
Geometry (G)	Overall	1	1	100	5	6	100
	6.G.A.1	0	1	100	0	2	100
	6.G.A.2	0	1	100	0	2	100
	6.G.A.3	0	1	100	0	2	100
	6.G.A.4	0	1	100	0	2	100
The Number System (NS)	Overall	2	2	100	7	8	100
	6.NS.A.1	0	1	100	0	2	100
	6.NS.B.2	0	1	100	0	2	100
	6.NS.B.3	0	1	100	0	2	100
	6.NS.B.4	0	1	100	0	2	100
	6.NS.C.6	0	1	100	0	2	100
	6.NS.C.7b	0	1	100	0	2	100
	6.NS.C.7c	0	1	100	0	2	100
	6.NS.C.8	0	1	100	0	2	100
Ratios and Proportional Relationships (RP)	Overall	1	1	100	5	6	100
	6.RP.A.1	0	1	100	0	2	100
	6.RP.A.3a	0	1	100	0	2	100
	6.RP.A.3b	0	1	100	0	2	100
	6.RP.A.3c	0	1	100	0	2	100
	6.RP.A.3d	0	1	100	0	2	100
Statistics and Probability (SP)	Overall	2	2	100	6	8	100
	6.SP.A.1	0	1	100	0	3	100
	6.SP.A.2	0	1	100	0	3	100
	6.SP.B.4	0	1	100	0	3	100
	6.SP.B.5	0	1	100	0	3	100

Table 53. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 7 Mathematics

Strand	Benchmark	Grade 7					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Expressions and Equations (EE)	Overall	1	1	100	4	5	100
	7.EE.A.1	0	1	100	0	3	100
	7.EE.B.3	0	1	100	0	3	100
Geometry (G)	Overall	1	1	100	6	7	100
	7.G.A.1	0	1	100	0	2	100
	7.G.A.2	0	1	100	0	2	100
	7.G.A.3	0	1	100	0	2	100
	7.G.B.4	0	1	100	0	2	100
	7.G.B.5	0	1	100	0	2	100
	7.G.B.6	0	1	100	0	2	100
The Number System (NS)	Overall	2	2	100	9	10	100
	7.NS.A.1	0	1	100	0	3	100
	7.NS.A.1b	0	1	100	0	3	100
	7.NS.A.2	0	1	100	0	3	100
	7.NS.A.2c	0	1	100	0	3	100
	7.NS.A.2d	0	1	100	0	3	100
Ratios and Proportional Relationships (RP)	Overall	2	2	100	5	6	100
	7.RP.A.1	0	1	100	0	2	100
	7.RP.A.2a	0	1	100	0	2	100
	7.RP.A.2b	0	1	100	0	2	100
	7.RP.A.3	0	1	100	0	2	100
Statistics and Probability (SP)	Overall	2	2	100	6	10	100
	7.SP.A.1	0	1	100	0	2	100
	7.SP.A.2	0	1	100	0	2	100
	7.SP.B.3	0	1	100	0	2	100
	7.SP.B.4	0	1	100	0	2	100
	7.SP.C.5	0	1	100	0	2	100
	7.SP.C.8	0	1	100	0	2	100

Table 54. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 8 Mathematics

Strand	Benchmark	Grade 8					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Expressions and Equations (EE)	Overall	2	2	100	10	11	100
	8.EE.A.1	0	1	100	0	2	100
	8.EE.A.2	0	1	100	0	2	100
	8.EE.A.3	0	1	100	0	2	100
	8.EE.A.4	0	1	100	0	2	100
	8.EE.B.5	0	1	100	0	2	100
	8.EE.B.6	0	1	100	0	2	100
	8.EE.C.7	0	1	100	0	2	100
	8.EE.C.7b	0	1	100	0	2	100
	8.EE.C.8a	0	1	100	0	2	100
Functions (F)	Overall	1	1	100	6	7	100
	8.F.A.1	0	1	100	0	2	100
	8.F.A.2	0	1	100	0	2	100
	8.F.A.3	0	1	100	0	2	100
	8.F.B.4	0	1	100	0	2	100
	8.F.B.5	0	1	100	0	2	100
Geometry (G)	Overall	3	3	100	9	11	100
	8.G.A.1	0	1	100	0	2	100
	8.G.A.2	0	1	100	0	2	100
	8.G.A.3	0	1	100	0	2	100
	8.G.A.4	0	1	100	0	2	100
	8.G.A.5	0	1	100	0	2	100
	8.G.B.7	0	1	100	0	2	100
	8.G.B.8	0	1	100	0	2	100
	8.G.C.9	0	1	100	0	1	100
The Number System (NS)	Overall	1	1	100	1	2	100
	8.NS.A.1	0	1	100	0	1	100
	8.NS.A.2	0	1	100	0	1	100
Statistics and Probability (SP)	Overall	1	1	100	2	3	100
	8.SP.A.2	0	1	100	0	1	100
	8.SP.A.3	0	1	100	0	1	100

Table 55. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 11 Mathematics

Strand	Benchmark	Grade 11					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Algebra (A)	Overall	2	2	100	12	15	100
	HS.A.SSE.A.1a	0	1	100	0	2	100
	HS.A.SSE.A.2	0	1	100	0	2	100
	HS.A.SSE.B.3	0	1	100	0	2	100
	HS.A.APR.A.1	0	1	100	0	2	100
	HS.A.CED.A.1	0	1	100	0	2	100
	HS.A.CED.A.2	0	1	100	0	2	100
	HS.A.CED.A.3	0	1	100	0	2	100
	HS.A.REI.A.1	0	1	100	0	2	100
	HS.A.REI.B.3	0	1	100	0	2	100
	HS.A.REI.C.5	0	1	100	0	2	100
	HS.A.REI.C.6	0	1	100	0	2	100
	HS.A.REI.D.10	0	1	100	0	1	100
	HS.A.REI.D.12	0	1	100	0	1	100
Functions (F)	Overall	2	2	100	7	8	100
	HS.F.BF.A.2	0	1	100	0	2	100
	HS.F.IF.A.1	0	1	100	0	2	100
	HS.F.IF.B.4	0	1	100	0	2	100
	HS.F.IF.B.6	0	1	100	0	2	100
	HS.F.LE.A.1	0	1	100	0	2	100
	HS.F.LE.A.2	0	1	100	0	2	100
	HS.F.LE.B.5	0	1	100	0	2	100
	HS.F.BF.A.2	0	1	100	0	2	100
	HS.F.IF.A.1	0	1	100	0	2	100
Geometry (G)	Overall	2	2	100	7	9	100
	HS.G.CO.A.1	0	1	100	0	2	100
	HS.G.CO.A.3	0	1	100	0	2	100
	HS.G.CO.C.10	0	1	100	0	2	100
	HS.G.CO.C.9	0	1	100	0	2	100
	HS.G.GMD.A.3	0	1	100	0	1	100
	HS.G.GMD.B.4	0	1	100	0	1	100
	HS.G.GPE.B.7	0	1	100	0	2	100
	HS.G.MG.A.1	0	1	100	0	1	100
	HS.G.SRT.B.5	0	1	100	0	2	100
Number and Quantity (N)	Overall	1	1	100	4	5	100
	HS.N.RN.A.1	0	1	100	0	2	100
	HS.N.RN.A.2	0	1	100	0	2	100
	HS.N.Q.A.2	0	1	100	0	2	100
	HS.N.Q.A.3	0	1	100	0	2	100
Statistics and Probability (S)	Overall	1	1	100	1	2	100
	HS.S.ID.A.1	0	1	100	0	1	100
	HS.S.ID.A.2	0	1	100	0	2	100
	HS.S.ID.C.7	0	1	100	0	2	100
	HS.S.CP.B.6	0	1	100	0	1	100

Table 56. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 5 Science

Strand	Benchmark	Grade 5					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Physical Science (PS)	Overall	3	3	100	9	12	100
	PS1	0	1	100	2	4	100
	5-PS|PS1|5-PS1-1	0	1	100	0	2	100
	5-PS|PS1|5-PS1-2	0	1	100	0	2	100
	5-PS|PS1|5-PS1-3	0	1	100	0	2	100
	5-PS|PS1|5-PS1-4	0	1	100	0	2	100
	PS2	0	1	100	2	4	100
	3-PS|PS2|3-PS2-1	0	1	100	0	2	100
	3-PS|PS2|3-PS2-2	0	1	100	0	2	100
	3-PS|PS2|3-PS2-3	0	1	100	0	2	100
	5-PS|PS2|5-PS2-1	0	1	100	0	2	100
	PS3	0	1	100	2	4	100
	4-PS|PS3|4-PS3-1	0	1	100	0	2	100
	4-PS|PS3|4-PS3-2	0	1	100	0	2	100
	4-PS|PS3|4-PS3-3	0	1	100	0	2	100
	4-PS|PS3|4-PS3-4	0	1	100	0	1	100
	5-PS|PS3|5-PS3-1	0	1	100	0	2	100
	PS4	0	1	100	0	2	100
	4-PS|PS4|4-PS4-1	0	1	100	0	2	100
	4-PS|PS4|4-PS4-2	0	1	100	0	2	100
Life Science (LS)	Overall	2	2	100	10	13	100
	LS1	0	1	100	2	4	100
	3-LS|LS1|3-LS1-1	0	1	100	0	2	100
	4-LS|LS1|4-LS1-1	0	1	100	0	2	100
	4-LS|LS1|4-LS1-2	0	1	100	0	2	100
	5-LS|LS1|5-LS1-1	0	1	100	0	2	100
	LS2	0	1	100	2	2	100
	3-LS|LS2|3-LS2-1	0	1	100	0	2	100
	5-LS|LS2|5-LS2-1	0	1	100	0	2	100
	LS3	0	1	100	2	2	100
	3-LS|LS3|3-LS3-1	0	1	100	0	2	100
	3-LS|LS3|3-LS3-2	0	1	100	0	2	100
	LS4	0	1	100	2	4	100
	3-LS|LS4|3-LS4-1	0	1	100	0	2	100
	3-LS|LS4|3-LS4-2	0	1	100	0	2	100
	3-LS|LS4|3-LS4-3	0	1	100	0	2	100
	3-LS|LS4|3-LS4-4	0	1	100	0	2	100
Earth and Space Science (ESS)	Overall	3	3	100	9	12	100
	ESS1	1	1	100	2	5	100
	4-ESS|ESS1|4-ESS1-1	0	1	100	0	2	100
	5-ESS|ESS1|5-ESS1-1	0	1	100	0	2	100
	5-ESS|ESS1|5-ESS1-2	0	1	100	0	2	100
	ESS2	1	1	100	2	5	100
	3-ESS|ESS2|3-ESS2-1	0	1	100	0	2	100

Strand	Benchmark	Grade 5					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Earth and Space Science (ESS) (cont.)	3-ESS|ESS2|3-ESS2-2	0	1	100	0	2	100
	4-ESS|ESS2|4-ESS2-1	0	1	100	0	2	100
	4-ESS|ESS2|4-ESS2-2	0	1	100	0	2	100
	5-ESS|ESS2|5-ESS2-1	0	1	100	0	2	100
	5-ESS|ESS2|5-ESS2-2	0	1	100	0	2	100
	ESS3	1	1	100	2	5	100
	3-ESS|ESS3|3-ESS3-1	0	1	100	0	2	100
	4-ESS|ESS3|4-ESS3-2	0	1	100	0	2	100
	4-ESS|ESS3|4-ESS3-1	0	1	100	0	2	100
	5-ESS|ESS3|5-ESS3-1	0	1	100	0	2	100

Table 57. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 8 Science

Strand	Benchmark	Grade 8					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Physical Science (PS)	Overall	3	3	100	9	12	100
	PS1	0	1	100	2	4	100
	MS-PS|PS1|MS-PS1-1	0	1	100	0	2	100
	MS-PS|PS1|MS-PS1-2	0	1	100	0	2	100
	MS-PS|PS1|MS-PS1-3	0	1	100	0	2	100
	MS-PS|PS1|MS-PS1-4	0	1	100	0	2	100
	MS-PS|PS1|MS-PS1-6	0	1	100	0	2	100
	PS2	0	1	100	2	4	100
	MS-PS|PS2|MS-PS2-1	0	1	100	0	2	100
	MS-PS|PS2|MS-PS2-2	0	1	100	0	2	100
	MS-PS|PS2|MS-PS2-3	0	1	100	0	2	100
	MS-PS|PS2|MS-PS2-4	0	1	100	0	2	100
	MS-PS|PS2|MS-PS2-5	0	1	100	0	2	100
	PS3	0	1	100	2	4	100
	MS-PS|PS3|MS-PS3-1	0	1	100	0	2	100
	MS-PS|PS3|MS-PS3-3	0	1	100	0	2	100
	MS-PS|PS3|MS-PS3-4	0	1	100	0	2	100
	MS-PS|PS3|MS-PS3-5	0	1	100	0	2	100
	PS4	0	1	100	2	3	100
	MS-PS|PS4|MS-PS4-1	0	1	100	0	2	100
	MS-PS|PS4|MS-PS4-2	0	1	100	0	2	100
	MS-PS|PS4|MS-PS4-3	0	1	100	0	2	100
Life Science (LS)	Overall	3	3	100	9	12	100
	LS1	0	1	100	2	5	100
	MS-LS|LS1|MS-LS1-1	0	1	100	0	2	100
	MS-LS|LS1|MS-LS1-2	0	1	100	0	2	100

Strand	Benchmark	Grade 8					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Life Science (LS) (cont.)	MS-LS|LS1|MS-LS1-3	0	1	100	0	2	100
	MS-LS|LS1|MS-LS1-4	0	1	100	0	2	100
	MS-LS|LS1|MS-LS1-5	0	1	100	0	2	100
	MS-LS|LS1|MS-LS1-6	0	1	100	0	2	100
	MS-LS|LS1|MS-LS1-7	0	1	100	0	2	100
	MS-LS|LS1|MS-LS1-8	0	1	100	0	2	100
	LS2	0	1	100	2	4	100
	MS-LS|LS2|MS-LS2-1	0	1	100	0	2	100
	MS-LS|LS2|MS-LS2-2	0	1	100	0	2	100
	MS-LS|LS2|MS-LS2-3	0	1	100	0	2	100
	MS-LS|LS2|MS-LS2-4	0	1	100	0	2	100
	LS3	0	1	100	1	2	100
	MS-LS|LS3|MS-LS3-1	0	1	100	0	2	100
	MS-LS|LS3|MS-LS3-2	0	1	100	0	2	100
	LS4	0	1	100	0	2	100
	MS-LS|LS4|MS-LS4-1	0	1	100	2	4	100
	MS-LS|LS4|MS-LS4-2	0	1	100	0	2	100
	MS-LS|LS4|MS-LS4-4	0	1	100	0	2	100
	MS-LS|LS4|MS-LS4-5	0	1	100	0	2	100
	MS-LS|LS4|MS-LS4-6	0	1	100	0	2	100
Earth and Space Science (ESS)	Overall	2	2	100	10	13	100
	ESS1	0	1	100	2	4	100
	MS-ESS|ESS1|MS-ESS1-1	0	1	100	0	2	100
	MS-ESS|ESS1|MS-ESS1-2	0	1	100	0	2	100
	MS-ESS|ESS1|MS-ESS1-3	0	1	100	0	2	100
	MS-ESS|ESS1|MS-ESS1-4	0	1	100	0	2	100
	ESS2	0	1	100	4	6	100
	MS-ESS|ESS2|MS-ESS2-1	0	1	100	0	2	100
	MS-ESS|ESS2|MS-ESS2-2	0	1	100	0	2	100
	MS-ESS|ESS2|MS-ESS2-3	0	1	100	0	2	100
	MS-ESS|ESS2|MS-ESS2-4	0	1	100	0	2	100
	MS-ESS|ESS2|MS-ESS2-5	0	1	100	0	2	100
	MS-ESS|ESS2|MS-ESS2-6	0	1	100	0	2	100
	ESS3	0	1	100	2	4	100
	MS-ESS|ESS3|MS-ESS3-1	0	1	100	0	2	100
	MS-ESS|ESS3|MS-ESS3-2	0	1	100	0	2	100
	MS-ESS|ESS3|MS-ESS3-3	0	1	100	0	2	100
	MS-ESS|ESS3|MS-ESS3-4	0	1	100	0	2	100
	MS-ESS|ESS3|MS-ESS3-5	0	1	100	0	2	100

Table 58. Percentage of Administered Tests Meeting Blueprint Requirements in Grade 11 Science

Strand	Benchmark	Grade 11					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Life Science (LS)	Overall	8	8	100	32	32	100
	LS1	2	2	100	10	13	100
	HS-LS|LS1|HS-LS1-1	0	1	100	0	2	100
	HS-LS|LS1|HS-LS1-2	0	1	100	0	2	100
	HS-LS|LS1|HS-LS1-3	0	1	100	0	2	100
	HS-LS|LS1|HS-LS1-4	0	1	100	0	2	100
	HS-LS|LS1|HS-LS1-5	0	1	100	0	2	100
	HS-LS|LS1|HS-LS1-6	0	1	100	0	2	100
	HS-LS|LS1|HS-LS1-7	0	1	100	0	2	100
	LS2-ESS2-ESS3	3	3	100	9	12	100
	HS-LS|LS2|HS-LS2-1	0	1	100	0	2	100
	HS-LS|LS2|HS-LS2-2	0	1	100	0	2	100
	HS-LS|LS2|HS-LS2-4	0	1	100	0	2	100
	HS-LS|LS2|HS-LS2-5	0	1	100	0	2	100
	HS-LS|LS2|HS-LS2-6	0	1	100	0	2	100
	HS-LS|LS2|HS-LS2-7	0	1	100	0	2	100
	HS-LS|LS2|HS-LS2-8	0	1	100	0	2	100
	HS-ESS|ESS2|HS-ESS2-6	0	1	100	0	2	100
	HS-ESS|ESS3|HS-ESS3-3	0	1	100	0	2	100
	LS3-LS4-ESS2	3	3	100	9	12	100
	HS-LS|LS3|HS-LS3-1	0	1	100	0	2	100
	HS-LS|LS3|HS-LS3-2	0	1	100	0	2	100
	HS-LS|LS4|HS-LS4-1	0	1	100	0	2	100
	HS-LS|LS4|HS-LS4-2	0	1	100	0	2	100
	HS-LS|LS4|HS-LS4-3	0	1	100	0	2	100
	HS-LS|LS4|HS-LS4-4	0	1	100	0	2	100
	HS-LS|LS4|HS-LS4-5	0	1	100	0	2	100
	HS-LS|LS4|HS-LS4-6	0	1	100	0	2	100
	HS-ESS|ESS2|HS-ESS2-7	0	1	100	0	2	100

Item Development

Chapter 4 – Item Development, provides detailed description on how items are developed. The number and type of items to be developed are based on an evaluation of content needs and available sample size for field testing that can result in reliable statics. Item writers are carefully chosen and well trained to follow standardized procedures and template when creating items. All items undergo rigorous multiple rounds of internal and external reviews from the content and fairness perspective before they are field-tested in an operational context. After field-testing, item analysis is conducted to examine whether items perform as expected. All items are reviewed by special education teachers and content experts in the state before they are moved to the final operational item pool.

Test Administration Conditions

Standardized test administration is critical in producing reliable and valid test scores. Comparability of test scores, whether between students and schools or across time for the same students, is based on standardization of test administration and test scoring rules. If TAs do not follow the same procedures, student performance cannot be compared meaningfully. For HSA-Alt, TAs are required to complete an online TA Certification Course before they can administer the HSA-Alt to their students. The guidelines for test administration are summarized in the Test Administration Manual (TAM). See Chapter 2 – Test Administration for details.

Item and Test Scoring

Item and test scores are probably the most critical element. All interpretations are established around students' test results. Every effort are made to ensure absolute accuracy on item and test scores. Section 10.3 Assurance in Test Scoring, provides detailed description on quality control and monitoring procedures implemented within CAI to assure accurate scores are generated and reported.

5.2.2. Evidence Based on Response Processes

Cognitive laboratory (cog lab) studies document validity evidence to show that the assessments tap the intended cognitive processes appropriate for each grade level as represented in the state's Alternate Assessment performance expectations. Cognitive lab studies conducted in Hawai`i explored student performance on items that linked to the state standards and aligned with the HSA-Alt Essence Statement expectations for student knowledge, skills, and abilities. The results of these studies demonstrated students' application of their knowledge and skills.

For students with the most significant cognitive disabilities, Every Student Succeeds Act (ESSA) places a one percent threshold on their participation in a state's alternate assessment. The students who participate in the alternate assessments for students with significant cognitive disabilities represent a variety of disability categories and demonstrate many concomitant learning difficulties. Students in this population can exhibit difficulties in attending to stimuli; committing information to working, short-term, or long-term memory; generalizing learning to familiar and novel environments; meta-cognition; or self-regulating behaviors. Furthermore, students with significant cognitive disabilities may also demonstrate significant communication and/or sensory deficits; limited fine or gross motor abilities; specialized health care needs; or inability to synthesize learned skills. Students with significant cognitive disabilities require multiple opportunities to engage with academic content and daily activities, as well as multiple ways to express and represent their knowledge.

Students with significant cognitive disabilities at all grade levels and at each of three cognitive levels (low ability, moderate ability, and high ability) were included in the study, 4–5 student per grade. The estimation of low-, moderate-, or high-ability level was determined either by the student's score on the previous year's alternate assessment administration or teacher recommendation. In addition to the grade-level and ability-level considerations, the students selected for this study represented the IDEA disability categories with the greatest number of students in the state's significantly cognitively disabled student population, intellectual disability, autism spectrum, and multiple disabilities.

Items from the state's item bank were selected for this study based on their closeness of fit to the cognitive demands of the standard the item was intended to assess. For each ELA, mathematics, and science item for each grade level, the CAI, state content experts, and a state stakeholder panel agreed on the item's linkage/alignment to the HCCS or NGSS HSA-Alt Essence Statements/HSA-Alt Range PLDs and the

thinking process that the student would most likely engage in to answer the question. Five items for each content area and grade level were selected for these studies. Each student at a grade level answered the same five items for ELA, mathematics, and science. Many of the items chosen for the cog lab were based on standards that had higher cognitive demands (cognitive demand does not equal Depth of Knowledge [DOK]). This was done to examine if the students could respond successfully to items that were at a cognitive level that came close to matching the grade-level standard expectations.

The data for these studies were obtained from three sources: student behaviors while responding to each item, student oral responses to questions that asked them to reflect on how they answered each item, and teacher observations about the student's behaviors during the cog lab, typical behaviors during instruction, and previous content exposure. Teacher insight into the student's response and assumed cognitive processing was an integral component of the study given that the limited communication and limited mobility of many students in the alternate population. Non-verbal students, if able, were provided with the opportunity to respond via communication board, Yes/No keys, or eye gaze. As a result, several different methods were used to document student response and thinking processes. The students were video recorded as they interacted with the computer-delivered items so that the researchers could return to the video to verify the student's responses and analyze the student's interaction with and response to the testing interface. The student's teacher and two observers entered each student's behaviors and oral responses to prompts on a data collection protocol as the student took each item. Following the delivery of each item, the teacher was interviewed by the study researcher(s) and notes and inferences on the student's actions and response were recorded. In Hawai`i's cog lab, student responses to items that matched the cognitive demands and skills included in the aligned standard were collected. The same was true for all states that participated in the MOU. A full description of Hawai`i's study and a discussion of the results are documented in the Hawai`i State Cognitive Lab Study Report, which is available upon request submitted to HIDOE.

5.2.3. Evidence Based on Internal Structure

The measurement and reporting model used in the HSA-Alt assessments assumes a single underlying latent trait, with achievement reported as a total scale score and an associated performance level for each subject and grade. There are also content domains/strands specified in the blueprints for each test, though the strand scores are neither reported at the individual student level nor at any aggregate level. The evidence on the internal structure is examined based on the correlations among content strand scores within the same subject and correlations between subjects.

Both observed and disattenuated (correction for attenuation) correlations are computed. The correction for attenuation indicates what the correlation would be if the construct could be measured with perfect reliability and corrected (adjusted) for measurement error estimates. The observed correlation between two claim scores with measurement errors can be corrected for attenuation as $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}}*\sqrt{r_{yy}}}$, where $r_{x'y'}$ is the correlation between *x* and *y* corrected for attenuation, r_{xy} is the observed correlation between *x* and *y*, r_{xx} is the reliability coefficient for *x*, and r_{yy} is the reliability coefficient for *y*. Since the reliability estimates are typically less than 1, the disttenuated correlations are higher than the observed correlations. Disattenuated correlations greater than 1 are set to 1.

The correlations among content strand scores are presented in Table 59—Table 61 for ELA, mathematics, and science, respectively. The observed correlations are presented below the diagonal, the disattenuated

correlations are presented above the diagonal, and the reliabilities of strand scores (bolded) are on the diagonal.

The correlation analyses are based on completed tests only. The number of items in each strand varies across students taking online adaptive tests and the strand scores are less reliable than the overall test score. As shown, the disattenuated correlations are the highest among strands in science, followed by ELA and mathematics.

Table 59. Correlations Among Strand Scores for ELA

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	**Strand 2**	**Strand 3**
3	Strand 1: Language	0.47	1.00	1.00
	Strand 2: RI and RL	0.60	0.71	1.00
	Strand 3: Writing	0.44	0.52	0.28
4	Strand 1: Language	0.54	0.91	0.94
	Strand 2: RI and RL	0.55	0.68	1.00
	Strand 3: Writing	0.48	0.58	0.49
5	Strand 1: Language	0.43	1.00	0.79
	Strand 2: RI and RL	0.52	0.59	1.00
	Strand 3: Writing	0.34	0.51	0.43
6	Strand 1: Language	0.47	1.00	0.98
	Strand 2: RI and RL	0.64	0.76	0.95
	Strand 3: Writing	0.46	0.57	0.46
7	Strand 1: Language	0.61	1.00	1.00
	Strand 2: RI and RL	0.73	0.77	1.00
	Strand 3: Writing	0.74	0.68	0.57
8	Strand 1: Language	0.55	0.86	0.61
	Strand 2: RI and RL	0.54	0.72	1.00
	Strand 3: Writing	0.28	0.54	0.39
11	Strand 1: Language	0.61	0.98	1.00
	Strand 2: RI and RL	0.66	0.75	1.00
	Strand 3: Writing	0.60	0.72	0.57

Note. RI=Reading – Informational; RL=Reading – Literature.

Table 60. Correlations Among Strand Scores for Mathematics

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	Strand 2	Strand 3
3	Strand 1: Measurement and Data & Geometry	0.63	0.78	0.88
	Strand 2: Number and Operations - Fractions	0.35	0.32	1.00
	Strand 3: OA & NBT	0.53	0.50	0.59
4	Strand 1: Measurement and Data & Geometry	0.55	1.00	0.85
	Strand 2: Number and Operations - Fractions	0.59	0.48	1.00
	Strand 3: OA & NBT	0.48	0.59	0.56
5	Strand 1: Measurement and Data & Geometry	0.55	0.78	1.00
	Strand 2: Number and Operations - Fractions	0.37	0.41	0.77
	Strand 3: OA & NBT	0.60	0.40	0.66
6	Strand 1: NS & EE	0.66	1.00	0.87
	Strand 2: RP & G	0.62	0.55	1.00
	Strand 3: Statistics and Probability	0.41	0.45	0.34
7	Strand 1: NS & EE	0.66	1.00	0.87
	Strand 2: RP & G	0.62	0.55	1.00
	Strand 3: Statistics and Probability	0.41	0.45	0.34
8	Strand 1: Functions & Statistics and Probability	0.47	0.88	0.47
	Strand 2: Geometry	0.45	0.55	0.12
	Strand 3: NS & EE	0.26	0.07	0.63
11	Strand 1: Functions & Statistics and Probability	0.45	0.54	0.97
	Strand 2: Geometry	0.22	0.38	0.90
	Strand 3: Number Quantity & Algebra	0.45	0.38	0.48

Note. OA & NBT=Operations and Algebraic Thinking & Number and Operations in Base Ten; RP & G=Ratios and Proportional Relationships & Geometry; NS & EE=The Number System & Expressions and Equations.

Table 61. Correlations Among Strand Scores for Science

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	Strand 2	Strand 3
5	Strand 1: Earth & Space Science	0.63	0.83	0.94
	Strand 2: Life Science	0.50	0.58	0.94
	Strand 3: Physical Science	0.61	0.58	0.66
8	Strand 1: Earth & Space Science	0.53	1.00	0.93
	Strand 2: Life Science	0.63	0.59	1.00
	Strand 3: Physical Science	0.53	0.65	0.61
11	Strand 1: Life Science	0.50	0.97	0.91
	Strand 2: Ecosystems: Interactions, Energy and Dynamics	0.56	0.66	0.77
	Strand 3: Heredity and Biological Evolution	0.54	0.52	0.70

The between-subject correlations are presented in Table 62. The observed correlations are presented below the diagonal, the disattenuated correlations are presented above the diagonal, and the reliabilities of subject scores (bolded) are on the diagonal. Disattenuated correlations among the three subjects range from the

lowest of 0.52 in grade 11 between mathematics and science to the highest of 0.92 in grade 6 between ELA and mathematics.

Table 62. Correlations Among Subject Scale Scores

Grade	Subject	ELA	Mathematics	Science
3	ELA	0.80	0.82	
	Mathematics	0.63	0.73	
	Science			
4	ELA	0.81	0.78	
	Mathematics	0.62	0.78	
	Science			
5	ELA	0.75	0.58	0.65
	Mathematics	0.44	0.78	0.66
	Science	0.51	0.53	0.82
6	ELA	0.83	0.92	
	Mathematics	0.74	0.78	
	Science			
7	ELA	0.87	0.78	
	Mathematics	0.59	0.65	
	Science			
8	ELA	0.81	0.74	0.77
	Mathematics	0.56	0.70	0.81
	Science	0.62	0.61	0.81
11	ELA	0.86	0.68	0.90
	Mathematics	0.52	0.68	0.52
	Science	0.76	0.39	0.83

Each subject test is designed and developed to measure a specific construct. Although it is expected to see decently high correlations between subjects, the between-strand correlations within the same subject are expected to be higher since they measure the same construct. Table 63 presents the comparison of between-subject disattenuated correlations with the average disattenuated between-strand correlations within the same subject for grade. For example, the correlation between grade 3 ELA and mathematics subject tests is 0.82, smaller than the average between-strand correlation of 1.00 in ELA and 0.89 in mathematics; the correlation between grade 5 ELA and science tests is 0.56, smaller than the average between-strand correlation of 0.93 in ELA and 0.90 in science. This pattern is observed for all subjects and grades except for grade 8 – the subject score correlation of 0.74 between ELA and mathematics and 0.81 between mathematics and science are both greater than the average between-strand correlation in mathematics (0.49), probably due to the low correlation between strands of Geometry and NS & EE (0.12), as well as between Functions & Statistics & Probability and NS & EE (0.47).

In summary, higher between-strand correlations provide validity evidence related to internal structure and indicate that the relationships among test items and test components conform to the construct on which the proposed test score interpretation are based.

Table 63. Disattenuated Between-Subject Correlations and Average Between-Strand Correlations

Grade	Between-Subject Correlations			Average Between-Strand Correlations		
	ELA vs Mathematics	ELA vs Science	Mathematics vs Science	ELA	Mathematics	Science
3	0.82			1.00	0.89	
4	0.78			0.95	0.95	
5	0.58	0.65	0.66	0.93	0.85	0.90
6	0.92			0.98	0.96	
7	0.78			1.00	0.96	
8	0.74	0.77	0.81	0.82	0.49	0.98
11	0.68	0.90	0.52	0.99	0.80	0.88

5.2.4. Evidence Based on Relations to Other Variables

In the U.S. Department of Education's document of "A State's Guide to the U.S. Department of Education's Assessment Peer Review Process", the Department lists as "adequate validity evidence that the state's assessment scores are related as expected with other variables" the results of a correlational study between assessments results or student test scores and variables related to test takers. The Hawai`i Department of Education (HIDOE) and Cambium Assessment, Inc. (CAI) implemented a study that required all teachers of students with severe cognitive disabilities who took the HSA-Alt to complete the Learner Characteristics Inventory (LCI) and the Hawai`i Observational Rating Assessment (HIORA) for each student who took the assessments. CAI then analyzed the results and ran a correlational study. Several of the LCI questions related to variables of student behaviors that might directly impact student performance on the alternate assessment and all of the grade-specific teacher rating questions of student skills and knowledge in a content area were used. The results of this study are discussed below, following a discussion of the purpose and questions extracted from the LCI, and the purpose and questions from the HIORA.

5.2.4.1 Learner Characteristics Inventory (LCI)

The LCI was developed by a committee of experts brought together by the National Center and State Collaborative (NCSC) project across all of the 18 core partner states. NCSC is funded through a four-year General Supervision Enhancement Grant (GSEG) from the Office of Special Education Programs at the U.S. Department of Education. "Its purpose is to create a system of high quality supports and resources for educators who work with students with the most significant cognitive disabilities" (Towles-Reeves, E., Kearns, J., Flowers, C., Hart, L., Kerbel, A., Kieinert, H., Quenemoen, R., & Thurlow, M., 2012, p. 1). According to these experts, the LCI was based on the work of Pellegrino, Chudowsky, & Glaser, 2001, who defined three pillars on which every assessment must rest: "a model of how students represent knowledge and develop competence in the subject domain, tasks or situations that allow one to observe students' performance, and an interpretation method for drawing inferences from the performance evidence thus obtained" (p. 2).

The final version of the LCI consists of 22 questions which a teacher answers about each student. These characteristics, taken together across all students participating in an alternate assessment across the state, help states understand the characteristics of their population of alternate assessment test takers. The questions include the following:

1. Student's grade
2. Student's age in years
3. Student's demonstration of significant cognitive disabilities
4. Student's requirement of a highly specialized educational program
5. Student's daily instruction
6. Student's difficulty with the demands of the general academic curriculum
7. Student's primary IDEA disability label
8. Student's secondary IDEA disability label
9. Student's primary language is other than English or not
10. Student's primary language
11. Student's primary classroom setting
12. Student's expressive communication skills
13. Student's use of an augmentative communication system
14. Student's use of an augmentative communication system (specify)
15. Student's receptive language skills
16. Student's vision
17. Student's hearing
18. Student's motor skills
19. Student's ability to engage with others
20. Student's health/attendance issues
21. Student's reading skills
22. Student's mathematics skills

The LCI provides a description of the state's significant cognitive disability classified students. The LCI is designed to be a descriptive instrument for the states to define this population of students and to then develop participation guidelines for their states' alternate assessments.

While reviewing the results of the Hawai`i LCI administration, it was observed that several of these questions did yield evidence relevant to the academic performance of these students. These questions include:

- Student's expressive communication skills
- Student's receptive language skills
- Student's ability to engage with others
- Student's reading skills
- Student's mathematics skills

The student's **expressive communication skills** question asks teachers to describe the student's oral/written or augmentative communication. Three levels of descriptors are defined:

- The first, or highest-level, descriptor states that the student uses symbolic language to communicate.
- The second, or middle-level, descriptor states that the student uses intentional communication but not at a symbolic level.
- The third, or lowest-level, descriptor states that the student communicates predominately through cries, facial expressions, change in muscle tone, or other indicators.

Students who communicate symbolically would be able to respond to items on the assessment and be more successful on an assessment that requires the use of symbolic communication; students with limited or no

symbolic communication skills would do less well on an assessment that relied on symbolic communication. The LCI "expressive communication skills" question would therefore predict, at a broad level, the student's final score on an assessment.

The student's **receptive language skills** include four levels of descriptors.

- The first, or highest, descriptor states that the student can independently follow 1–2 step directions presented through words without additional cues.
- The second descriptor states that the student can follow 1–2 step directions with additional cues.
- The third descriptor states that the student is receptive and alerts to sensory input from another person, but the student requires actual physical assistance to follow simple directions.
- The fourth, or lowest, descriptor states that the student demonstrates an uncertain response to sensory stimuli.

On an academic assessment, a student must be able to respond independently to directions, and students who are able to do so will receive a higher score on an assessment than those who cannot. Therefore, the receptive language descriptors do relate to a student's performance on a symbolic-language based assessment.

The student's **engagement descriptor** also has four descriptive statements.

- The first, or highest, states that the student can initiate and sustain social interactions.
- The second descriptor describes the student as responding but not initiating social interactions.
- The third descriptor defines a student who alerts to others.
- The fourth, or lowest, descriptor defines a student who does not alert to others.

An academic assessment situation is a social interaction, and the computer audio voice reads the questions and options to the student; students who enter into social interactions with others—even if they do not initiate the interaction, as this is not necessary on an assessment—would have more of a chance of success on an assessment then students who do not enter into social interactions with others.

The student's **reading skills** descriptor directly relates to the student's reading ability as well as the student's ability to understand all instruction in the content areas, as much of the instruction requires the student to read; even if the instruction does not require reading letters and words, it may include numbers and operation signs. The reading descriptors progress as follows:

- Reads fluently with critical understanding;
- Reads fluently with literal understanding;
- Reads basic sight words;
- is aware of text; and
- Demonstrates no observable awareness of print.

Students who can read critically will do better on an assessment than students who only read with literal understanding, and students who read with literal understanding will do better on an assessment than students who only read sight words. These descriptors seem to have the potential of being predictive of high and low scores on an academic assessment.

The **mathematics skills** descriptors relate to mathematics instruction and assessment, as well as any other content areas such as science or the reading of graphs and charts that require the use of mathematics or an understanding of numerical values. The mathematics descriptors progress as follows:

- Applies computation procedures to solve real-life or routine word problems;
- Does computational procedures with or without a calculator;
- Counts to at least 10 with 1:1 correspondence;
- Counts by rote to 5; and
- Demonstrates no observable awareness or use of numbers.

A student who can apply computational procedures to real-life problems will do better on an assessment than a student who can only do computation procedures, and a student who can do computational procedures will do better than a student who counts to 10 with 1:1 correspondence. Just as with the reading descriptors, the mathematics descriptors also have the potential of being predictive of high and low scores on an academic assessment.

5.2.4.2 Hawai`i Observational Rating Assessment (HIORA)

The HIORA was developed in two stages by HIDOE content experts. In the first stage, the descriptions of skills, knowledge, and understanding expected of students with significant cognitive disabilities were developed in a two-year process within the state based upon educator, content area, and special education professional input. The HSA-Alt Range Performance-Level Descriptors are the culmination of that work. The HSA-Alt Range PLDs describe what constituted an appropriate reduction of the general education standards for students who took the alternate form of the assessment. Four levels of test performance expectations were established in the HSA-Alt Range PLDs. These expectations for performance were distilled into sets of six questions for ELA and mathematics, and four questions for science. Each set of questions was specifically designed for one grade level. Each HIORA question provided teachers with four rating levels to choose from:

- Minimal Understanding,
- Partial or Inconsistent Understanding,
- Adequate Understanding, and
- Thorough Understanding.

Teachers were charged with selecting what seemed to them to be the most fitting description of student performance for their student given a description of student skills and knowledge for a content area and grade. A grade-level sample question for each content area is shown directly below.

Example HIORA Question – Grade 3 English Language Arts

In the Reading Literature domain, can the student answer literal questions related to something concrete (i.e., tangible, sensory) found in a literary text? For this skill, the student demonstrates:

- Minimal Understanding
- Partial or Inconsistent Understanding
- Adequate Understanding
- Thorough Understanding

Example HIORA Question – Grade 3 Mathematics

In the Operations and Algebraic Thinking domain, can the student represent and solve multiplication and division problems involving equal groups, area, and arrays? For this skill, the student demonstrates:

- Minimal Understanding
- Partial or Inconsistent Understanding
- Adequate Understanding
- Thorough Understanding

Example HIORA Question – Grade 5 Science

In the life science domain, can the student describe: how organisms vary in their traits; ways in which plants, animals, and environments of the past are similar or different from plants, animals, and environments of today; how internal and external structures support the survival, growth, behavior, and reproduction of plants and animals; where the energy in food comes from and what it is used for; how matter cycles through ecosystems; and, what happens to organisms when their environment changes. In these areas, the student demonstrates:

- Minimal Understanding
- Partial or Inconsistent Understanding
- Adequate Understanding
- Thorough Understanding

The HIORA rating of student skills was collected under the assumption that students who were rated by teachers as having minimal, partial, or inconsistent understanding of the skills and knowledge being tested on the alternate summative form, would perform at a lower level than students who received teacher ratings of adequate or thorough understanding of those same skills. This assumption was then tested through a correlative comparison in which the teacher ratings within each content area were transformed to ordinal numbers one to four, averaged and then compared to the student's overall performance rating in the content area.

In the second stage of HIORA development, the state borrowed the Transition Success Predictors from the National Technical Assistance Center on Transition (NTACT) to craft grade-specific questions for teachers to provide a response. Teachers used these questions in the second part of the HIORA to rate student readiness for transition.

HIORA NTACT Success Predictors – Part One (Grades 3-8 and 11)

1. Was the student included in general education instruction during this school year? Select as many as apply.

    - The student was not included in any general education instruction.
    - The student was included in ELA instruction.
    - The student was included in mathematics instruction. o The student was included in science instruction.
    - The student was included in social studies instruction.

2. How would you rate the student's ability to interact with others? Select one.

    - The student has difficulty interacting with people, both familiar and unfamiliar persons.

- The student has difficulty interacting with unfamiliar people but is able to interact with people he/she knows.
- The student generally interacts well with both familiar and unfamiliar people.

3. How would you rate the student's ability to interact with others in unfamiliar situations? Select one.

    - The student does not interact well with others in both familiar and unfamiliar social situations.
    - The student has difficulty interacting well with others in new social situations but interacts well with others in known social situations.
    - The student generally interacts well with others in both familiar and unfamiliar social situations.

4. How would you rate the student's parents' educational expectations for the student? Select one.

    - Insufficient information to report.
    - None to minimal expectations.
    - Low expectations; the student can achieve more than is expected.
    - Reasonable expectations for the student's educational achievement.
    - Higher expectations than the student will be able to achieve.

HIORA NTACT Success Predictors – Part Two (Grades 7-8 and 11)

5. What type of career skills instruction has the student received? Select all that apply.

    - The student did not receive instruction in career choices.
    - The student received instruction in career choices.
    - The student received social skill instruction required for his/her career choices.
    - The student received instruction in the specific reading skills required for his/her possible career choices.
    - The student received instruction in the specific writing skills required for his/her possible career choices.
    - The student received instruction in the specific mathematics skills required for his/her possible career choices.

6. Did the student have some work experience this year? Select one.

    - I do not know.
    - The student has had no work experience, paid or unpaid.
    - The student had unpaid work experience.
    - The student had paid work experience.

7. If the student had either paid or unpaid work experience, please answer the three questions below.

    - Was the student successful in his/her work experience?

- I do not know.
- The student was unsuccessful in his/her work experience.
- The student was successful in his/her work experience.

- What educational skills did the student's work experience require? Select as many as apply.

  - I do not know.
  - The student's work experience required the use of reading skills.
  - The student's work experience required the use of writing skills.
  - The student's work experience required the use of mathematics skills.
  - The students work experience required the use of science skills.

- How long did the student's work experience last? Select one.

  - Less than 3 months
  - 6 months to 3 months o One year to 7 months o More than one year

5.2.4.1 Correlations with LCI and HIORA Descriptors

The LCI descriptors on Expressive Language, Receptive Language, Engagement, Reading Skills, Mathematics Skills, and a composite score by adding five LCI descriptors were correlated with the HSA-Alt scores in ELA, mathematics, and science.

As shown in Table 64, both Reading and Mathematics Skills tend to have higher correlations with the test scores than the other three descriptors. Combining all the descriptors together into a composite yields a higher correlation with student total test scores for all three content areas. The lowest correlation was between the Expressive Communication Skills and students' ELA scores (0.07) in grade 5; the highest correlations was between the Reading Skills and students' ELA scores (0.58) in grade 11.

A teacher's description of a student's ability level, as required when completing the LCI, does correlate moderately with students' overall scores on the HSA-Alt. It provides supporting validity evidence of the HSA-Alt in relation to other relevant measures. The assessment itself reflects the range of student skills in an academic content area that are positively and moderately correlated with their teachers' independent judgment of the students' skills.

Table 64. Correlation Between LCI Descriptors and the HSA-Alt Total Score

Grade	N	Composite	Expressive Communication Skills	Receptive Language Skills	Ability to Engage with Others	Reading Skills	Mathematics Skills
ELA							
3	111	0.34	0.19	0.28	0.29	0.22	0.26
4	114	0.48	0.33	0.33	0.38	0.40	0.45
5	108	0.40	0.07	0.26	0.23	0.43	0.32
6	133	0.42	0.29	0.29	0.25	0.41	0.34
7	107	0.55	0.32	0.47	0.41	0.40	0.42
8	88	0.39	0.34	0.18	0.32	0.37	0.21
11	87	0.57	0.20	0.40	0.39	0.58	0.56
Mathematics							
3	111	0.57	0.36	0.45	0.32	0.48	0.50
4	113	0.40	0.28	0.24	0.15	0.40	0.44
5	109	0.30	0.09	0.08	0.16	0.29	0.34
6	132	0.42	0.21	0.20	0.23	0.46	0.43
7	108	0.55	0.32	0.43	0.43	0.39	0.47
8	87	0.33	0.22	-0.07	0.20	0.43	0.26
11	88	0.53	0.28	0.36	0.29	0.56	0.51
Science							
5	108	0.44	0.26	0.37	0.23	0.33	0.36
8	84	0.30	0.23	0.08	0.20	0.28	0.28
11	85	0.46	0.23	0.31	0.37	0.48	0.41

Table 65 represents the correlation between teacher rating of each HIORA question and student's overall scale score in ELA. In all grades, Items 1 and 2 are the questions related to reading literature, Items 3 and 4 are the questions related to reading informational text, Item 5 is the question related to writing, Item 6 is the question related to language content, and Item 7 is the question related to instruction time. The correlations seem to be higher in grades 11 with all values larger than 0.3 except for item 7.

Table 66 represents the correlation between teacher rating of each HIORA question and student's overall scale score in mathematics. Items 1 to 5 are the questions related to different mathematics content areas across all grades. Item 6 is the question related to geometry in grades 3 and 5, instruction time in the rest of grades. Item 7 is the question related to instruction time in grade 3 and 5. In general, correlations in mathematics tend to be lower than that in ELA.

Table 67 represents the correlation between teacher rating of each HIORA question and student's overall scale score in science. Items 1 to 4 are questions related to different science content areas, and Item 5 is the question related to instruction time. The correlations in grade 5 are the highest among all three grades ranging from 0.20 to 0.31.

Table 65. Correlation Between HIORA and ELA Scale Score

	HIORA Question						
Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3	0.39	0.26	0.35	0.27	0.19	0.18	0.17
4	0.20	0.21	0.19	0.28	0.12	0.16	0.07
5	0.33	0.14	0.20	0.19	0.21	0.07	0.26
6	0.32	0.36	0.38	0.31	0.30	0.34	0.22
7	0.24	0.24	0.23	0.24	0.21	0.16	-0.10
8	0.24	0.24	0.23	0.24	0.21	0.16	-0.10
11	0.43	0.40	0.35	0.38	0.43	0.43	0.16

Table 66. Correlation Between HIORA and Mathematics Scale Score

	HIORA Question						
Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3	0.24	0.48	0.17	0.37	0.25	0.23	0.18
4	0.28	0.41	0.09	0.20	0.18		0.20
5	0.21	0.25	0.35	0.36	0.39	0.40	0.30
6	0.25	0.19	0.15	0.25	-0.01		0.20
7	0.08	0.13	0.08	0.13	0.15		0.17
8	0.25	0.31	0.23	0.29	0.25	-0.04	0.27
11	0.13	0.17	0.15	0.13	0.10		0.23

Table 67. Correlation Between HIORA and Science Scale Score

	HIORA Question				
Grade	**1**	**2**	**3**	**4**	**5**
5	0.31	0.28	0.27	0.30	0.20
8	0.11	0.04	0.06	0	0.04
11	-0.08	0.18	0	0	0.10

In general, relatively weak correlations are observed between teachers' ratings in the HIORA and the test results than the correlations in LCI. This could be due to several different factors. First, teachers may have misinterpreted the descriptions of students' knowledge and skills in a HIORA question. The use of multiple measures and descriptions of skill embedded within a single content-area question may have confused teachers and led to inconsistent interpretations and ratings. Second, there may be a lack of referents for teachers to compare with. All but the most veteran teachers may have an adequate background to compare

and evaluate student performance on content- and grade-specific skills, with the small customary class size for this population.

5.3 SUMMARY

This chapter presented the statements on intended purposes and uses of the HSA-Alt test scores and summarized various sources of theoretical and empirical evidence that can inform validity arguments related to using and interpreting HSA-Alt scores. The focus was on how four sources of validity evidence support uses and interpretations of test scores. Validation is an on-going process and validity evidence will continue to be accumulated and evaluated as more relevant data become available.

6. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard errors of measurement. In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming the error variance is the same for all scores. Within the item response theory (IRT) framework, measurement error varies conditioning on ability. The amount of precision in estimating achievement can be determined by the test information, which describes the amount of information provided by the test at each score point along the ability continuum. Test information is a value that is inversely related to the measurement error of the test; the larger the measurement error, the less test information is being provided.

Each item in the computer-adaptive test (CAT) was selected based on content values that meet the blueprint and information values that match students' ability. The reliability estimates of the HSA-Alt is provided with marginal reliability, standard error of measurement (SEM), and classification accuracy and consistency for each performance standard.

6.1 MARGINAL RELIABILITY

Marginal reliability was computed for the scale scores, taking into account the varying measurement errors across the ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional SEM, estimated at different points on the ability scale, for all students.

The marginal reliability ($\bar{\rho}$) is defined as

$$\bar{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^{N} CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i$ is the conditional SEM of the scale score for student i; and σ^2 is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

Another way to examine test reliability is with SEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that makes up the test. In computer-adaptive testing, items administered vary among all students, so the SEM also can vary among students, which yields conditional SEM. The average conditional SEM can be computed as

$$Average\ CSEM = \sigma\sqrt{1-\bar{\rho}} = \sqrt{\sum_{i=1}^{N} CSEM_i^2/N}.$$

The smaller the value of average conditional SEM, the greater the accuracy of test scores.

Table 68 presents the marginal reliability coefficients and the average conditional SEM for the total scale scores, based on all attempted tests, excluding the early stopped tests (Early Stopping Rule), tests with responses for 1–4 items only, and tests with 1–7 "No Responses". For the tests with 1–4 item responses and 1–7 "No Response," the lowest scale score was reported, but the SEM associated with the scale score was left blank.

Table 68. Marginal Reliability for ELA, Mathematics, and Science

Grade	Number of Operational Items	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
ELA					
3	40	0.80	294.77	43.52	19.66
4	40	0.81	297.90	30.12	13.11
5	40	0.75	292.77	31.43	15.71
6	40	0.83	297.43	42.11	17.22
7	40	0.87	292.43	39.20	14.09
8	40	0.81	284.19	35.34	15.60
11	40	0.86	297.31	42.50	15.82
Mathematics					
3	40	0.73	293.58	37.23	19.41
4	40	0.78	292.04	40.53	19.18
5	40	0.78	300.32	34.33	16.27
6	40	0.78	276.74	54.32	25.69
7	40	0.65	279.92	46.92	27.57
8	40	0.70	282.16	37.00	20.15
11	40	0.68	287.60	32.73	18.59
Science					
5	40	0.82	283.41	49.74	20.94
8	40	0.81	277.63	39.85	17.36
11	40	0.83	292.38	47.86	19.87

6.2 STANDARD ERROR CURVES

Figure 8–Figure 10 present plots of the conditional SEM of scale scores. The vertical lines indicate the cut scores for Approaches, Meets, and Exceeds. For each student's test, the item selection algorithm selected items that matched student ability and met the test blueprint requirement.

Overall, the standard error curves suggest that students are measured with a similar precision across the range of score distribution, except for a few outliers with extremely low score.

Figure 8. Conditional Standard Error of Measurement for ELA

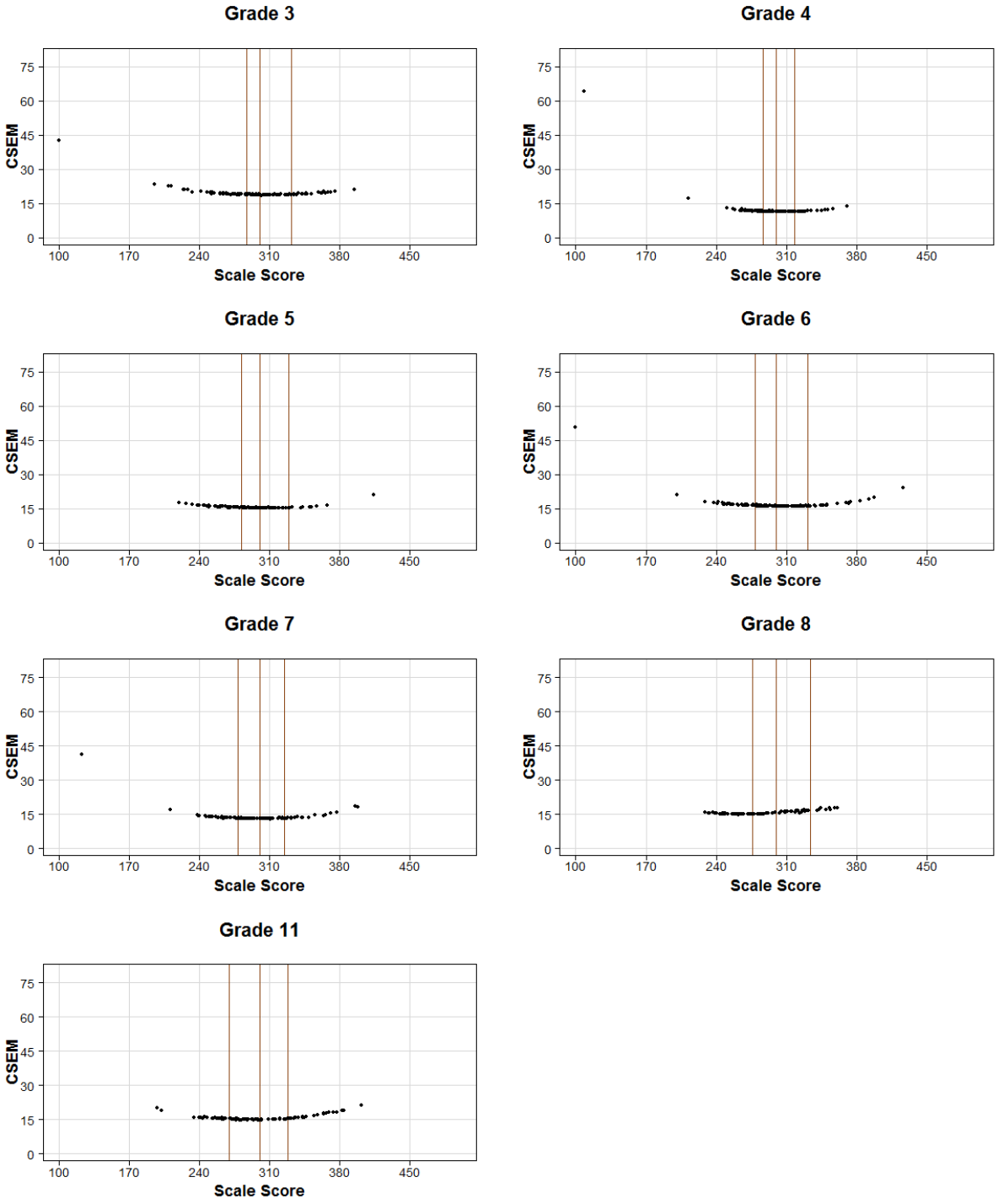


Figure 9. Conditional Standard Error of Measurement for Mathematics

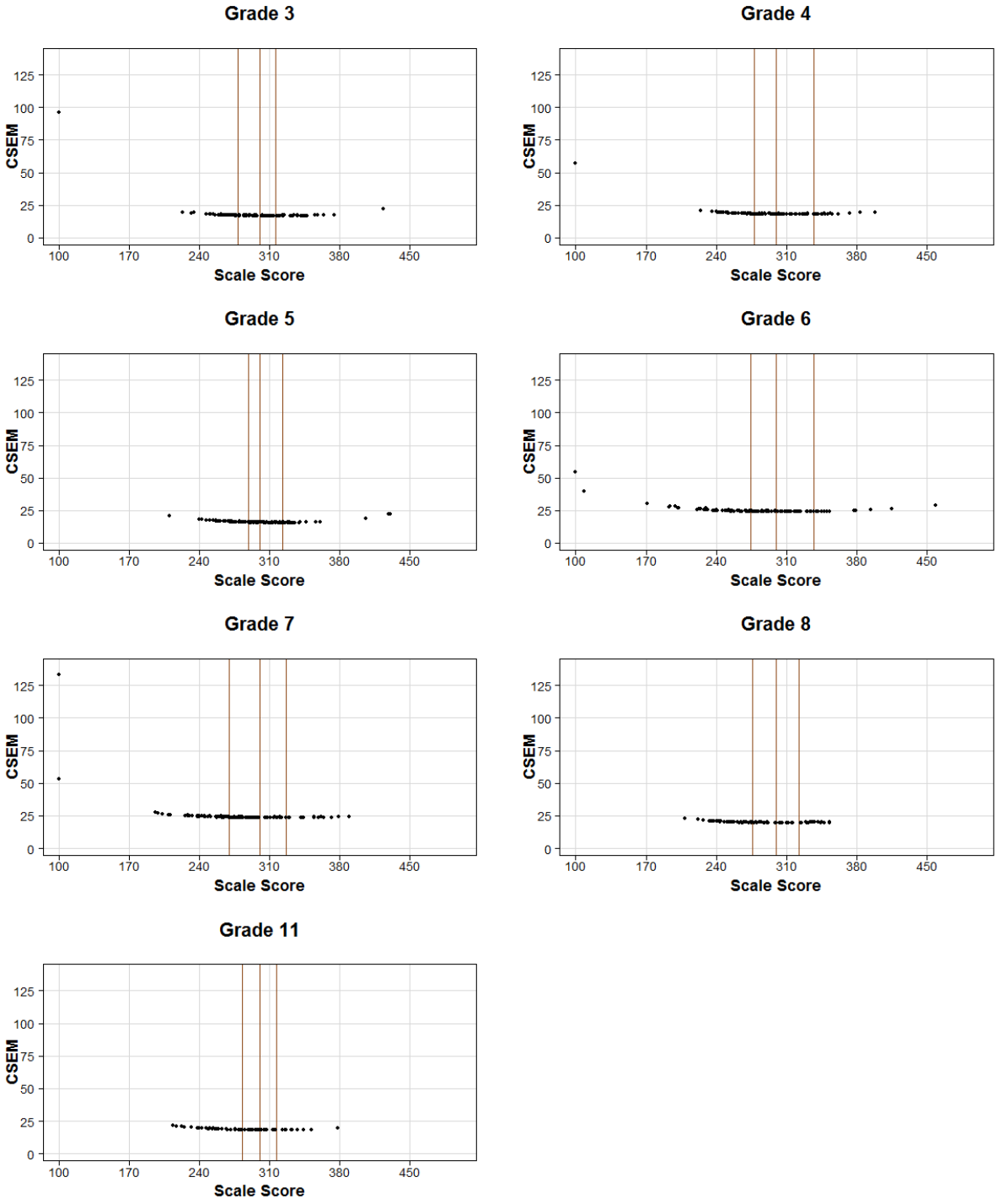


Figure 10. Conditional Standard Error of Measurement for Science

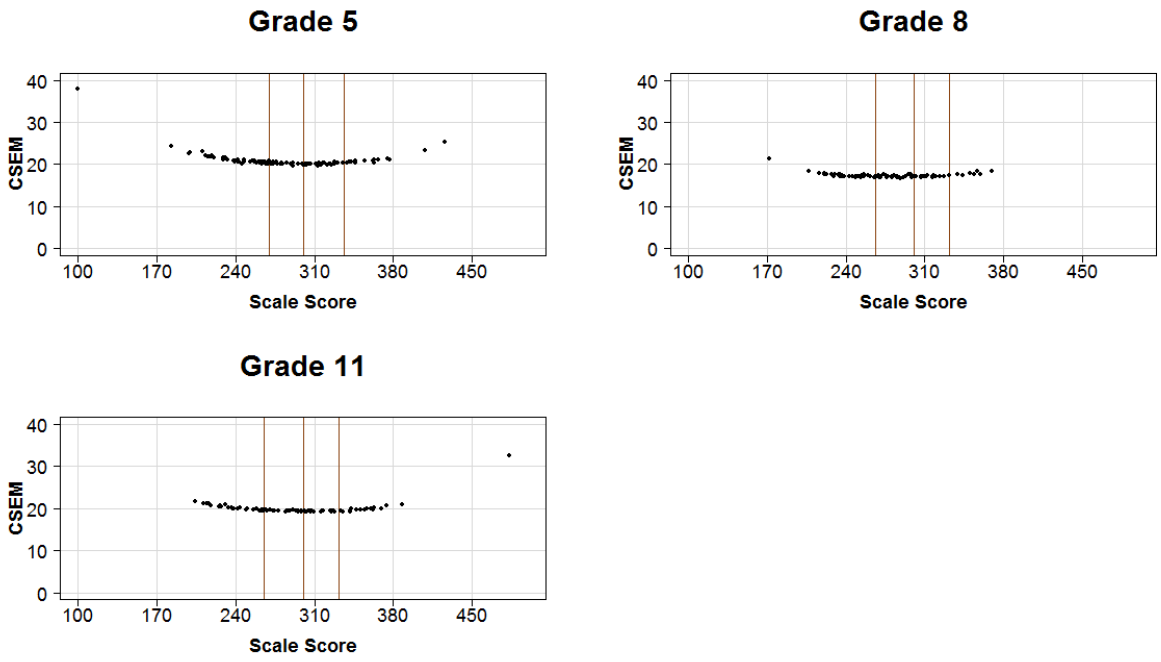


Table 69 presents the average conditional SEM for scores in each performance level. As shown in Figure 8 – Figure 10, the average conditional SEMs in Approaches and Meets are similar, but slightly larger in Well Below and Exceeds, which can be expected for tests with extreme scores.

Table 69. Average Conditional Standard Error of Measurement by Performance Level

Grade	Well Below	Approaches	Meets	Exceeds	Average CSEM
ELA					
3	20.56	18.92	18.84	19.46	19.66
4	16.20	11.49	11.40	11.71	13.11
5	16.03	15.50	15.36	16.19	15.71
6	18.45	16.32	16.16	17.37	17.22
7	15.18	13.17	13.09	14.34	14.09
8	15.13	15.21	16.15	17.16	15.60
11	15.95	15.01	15.06	16.96	15.82
Mathematics					
3	23.16	16.99	16.88	17.20	19.41
4	20.57	18.21	18.11	18.33	19.18
5	16.75	15.83	15.70	16.42	16.27
6	27.12	24.31	24.19	24.89	25.69
7	32.56	23.73	23.64	23.72	27.57
8	20.47	19.81	19.75	19.89	20.15
11	19.29	18.13	18.13	18.33	18.59
Science					
5	21.66	20.26	20.10	21.37	20.94
8	17.56	17.12	17.16	17.83	17.36
11	20.24	19.37	19.28	20.79	19.87

6.3 RELIABILITY OF PERFORMANCE CLASSIFICATION

When student performance is reported with performance levels, a reliability of performance classification is computed in terms of the probabilities of accurate and consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014). The indexes consider the accuracy and consistency of classifications.

For a fixed-form test, the accuracy and consistency of classifications are estimated on a single form's test scores from a single test administration based on the true-score distribution estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Wingersky, 1979; Subkoviak, 1976; Livingston & Lewis, 1995). For the CAT, because the adaptive testing algorithm constructs a test form unique to each student, the classification indexes are computed based on all sets of items administered across students using an IRT-based method (Guo, 2006).

The classification index can be examined in terms of the classification accuracy and the classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. Classification consistency refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made on the basis of an alternate form (another set of adaptively administered items given the same ability), that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test forms.

In reality, the true ability is unknown and students do not take an alternate, equivalent form; therefore, the classification accuracy and the classification consistency are estimated based on students' item scores, the item parameters, and the assumed underlying latent ability distribution as described below. The true score is an expected value of the test score with a measurement error.

For the *i*th student, the student's estimated ability is $\hat{\theta}_i$ with SEM of $se(\hat{\theta}_i)$, and the estimated ability is distributed as $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, assuming a normal distribution where θ_i is the unknown true ability of the *i*th student. The probability of the true score at performance level *l* based on the cut scores c_{l-1} and c_l is estimated as

$$p_{il} = p(c_{l-1} \le \theta_i < c_l) = p\left(\frac{c_{l-1} - \hat{\theta}_i}{se(\hat{\theta}_i)} \le \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} < \frac{c_l - \hat{\theta}_i}{se(\hat{\theta}_i)}\right) = p\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)} < \frac{\hat{\theta}_i - \theta_i}{se(\hat{\theta}_i)} \le \frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right)$$
$$= \Phi\left(\frac{\hat{\theta}_i - c_{l-1}}{se(\hat{\theta}_i)}\right) - \Phi\left(\frac{\hat{\theta}_i - c_l}{se(\hat{\theta}_i)}\right).$$

Instead of assuming a normal distribution of $\hat{\theta}_i \sim N\left(\theta_i, se^2(\hat{\theta}_i)\right)$, we can estimate the above probabilities directly using the likelihood function.

The likelihood function of theta, given a student's item scores, represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point. If a student with estimated theta is below the cut point, a probability of at or above the cut point is an estimate of the chance that this student is misclassified as below the cut, and 1 minus that probability is the estimate of the chance that the student is correctly classified as below the cut score. Using this logic, we can define various classification probabilities.

If we are interested in only the classification at each cut, *cut*, the probability of the *i*th student being classified as at or above the cut given the item scores $\mathbf{z}_i = (z_{i1}, \cdots, z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1, \cdots, \mathbf{b}_J)$ with *J* administered items, can be estimated as

$$p_i = P(\theta_i \geq cut|\mathbf{z}, \mathbf{b}) = \frac{\int_{cut}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta},$$

where the likelihood function based on Rasch IRT models is

$$L(\theta|\mathbf{z}_i, \mathbf{b}) = \prod_{j\in \mathrm{d}} \left(\frac{Exp\left(z_{ij}(\theta - b_j)\right)}{1+Exp(\theta - b_j)} \right) \prod_{j\in \mathrm{p}} \left(\frac{Exp\left(z_{ij}\theta - \sum_{k=1}^{z_{ij}} b_{ik}\right)}{1+\sum_{m=1}^{K_j} Exp(\sum_{k=1}^{m}(\theta - b_{jk}))} \right),$$

where d stands for dichotomous and p stands for polytomous items; $\mathbf{b}_j = (b_j)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (b_{j1}, \ldots, b_{jK_j})$ if the *j*th item is a polytomous item.

Classification Accuracy

Using p_i, we can construct a 2×2 table as

$$\begin{pmatrix} n_{a11} & n_{a12} \\ n_{a21} & n_{a22} \end{pmatrix}$$

where $n_{a11} = \sum_{pl_i=\text{below}}(1 - p_i)$, which is the expected number of students below the cut when the *i*th student's performance level, pl_i, is below the cut. Similarly we can define $n_{a12} = \sum_{pl_i=\text{below}} p_i$, $n_{a21} = \sum_{pl_i=\text{at or above}}(1 - p_i)$, and $n_{a22} = \sum_{pl_i=\text{at or above}} p_i$. In the above table, the row represents the observed level and the column represents the expected level.

The classification accuracy (*CA*) for the at or above the cut is estimated by

$$CA_{\text{at or above}} = \frac{n_{a22}}{n_{a21}+n_{a22}},$$

the classification accuracy (*CA*) for the below the cut is estimated by

$$CA_{\text{below}} = \frac{n_{a11}}{n_{a11}+n_{a12}},$$

and the overall classification accuracy for the cut is estimated by

$$CA = \frac{n_{a22}+n_{a11}}{n_{a21}+n_{a22}+n_{a11}+n_{a12}}.$$

Classification Consistency

Using p_i, which is similar to accuracy, we can construct another 2×2 table by assuming the test is administered twice independently to the same student group, hence we have

$$\begin{pmatrix} n_{c11} & n_{c12} \\ n_{c21} & n_{c22} \end{pmatrix}$$

where $n_{c11} = \sum_{i=1}^{N}(1 - p_i)(1 - p_i)$, $n_{c12} = \sum_{i=1}^{N}(1 - p_i)p_i$, $n_{c21} = \sum_{i=1}^{N} p_i(1 - p_i)$, and $n_{c22} = \sum_{i=1}^{N} p_i p_i$. In each of the above four equations, the first and the second probabilities are the probabilities of the *i*th student being classified at either below or at or above the cut, respectively, based on observed scores and hypothetical scores from an equivalent test form.

The classification consistency (*CC*) for the at or above the cut is estimated by

$$CC_{\text{at or above}} = \frac{n_{c22}}{n_{c21}+n_{c22}},$$

the classification consistency (*CC*) for the below the cut is estimated by

$$CC_{\text{below}} = \frac{n_{c11}}{n_{c11}+n_{c12}},$$

and the overall classification consistency is

$$CC = \frac{n_{c22}+n_{c11}}{n_{c21}+n_{c22}+n_{c11}+n_{c12}}.$$

The analysis of the classification index is performed based on overall scale scores.

Table 70 shows classification accuracy and consistency indexes for the spring 2023 HSA-Alt tests. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, but the accuracy index assumes only a single test score and a true score, which does not include measurement error.

Table 70. Classification Accuracy and Consistency for Performance Standards

Grade	Accuracy			Consistency		
	Approaches	Meets	Exceeds	Approaches	Meets	Exceeds
ELA						
3	0.85	0.86	0.90	0.80	0.80	0.87
4	0.86	0.86	0.90	0.81	0.80	0.85
5	0.86	0.84	0.93	0.80	0.78	0.90
6	0.86	0.89	0.92	0.81	0.83	0.88
7	0.89	0.90	0.93	0.84	0.86	0.90
8	0.88	0.92	0.94	0.83	0.88	0.91
11	0.88	0.91	0.92	0.84	0.87	0.90
Mathematics						
3	0.83	0.86	0.89	0.78	0.80	0.84
4	0.85	0.88	0.92	0.79	0.83	0.89
5	0.86	0.85	0.87	0.80	0.80	0.83
6	0.85	0.88	0.92	0.80	0.82	0.89
7	0.81	0.86	0.91	0.75	0.80	0.88
8	0.87	0.89	0.91	0.81	0.85	0.87
11	0.85	0.81	0.88	0.78	0.76	0.82
Science						
5	0.87	0.88	0.93	0.82	0.84	0.90
8	0.88	0.88	0.95	0.84	0.84	0.93
11	0.89	0.86	0.92	0.85	0.82	0.88

6.4 RELIABILITY OF CONTENT STRAND SCORES

For the HSA-Alt, although only the overall score is reported, the marginal reliability coefficients and the measurement errors are also computed for strand scores. The reliability coefficients were computed based on the complete tests only because the content of the items that were not administered in the incomplete tests is unknown. Table 71—Table 73 show the reliability coefficients, scale score means and standard deviations, and average CSEM for each strand.

Table 71. Marginal Reliability Coefficients for Content Strand Scores - ELA

Grade	Strand*	Number of Items Specified in Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Language	8	9	0.46	295.39	62.82	45.38
	Reading – Informational & Literature	22	24	0.70	294.13	48.48	26.20
	Writing	8	10	0.27	294.33	51.94	43.78
4	Language	8	9	0.54	303.65	43.09	28.00
	Reading – Informational & Literature	22	24	0.68	299.42	29.82	16.29
	Writing	8	10	0.49	290.18	38.30	26.44
5	Language	8	10	0.43	295.94	47.18	35.50
	Reading – Informational & Literature	22	24	0.59	294.54	32.61	20.93
	Writing	8	10	0.43	283.72	48.92	36.50
6	Language	8	10	0.47	304.81	52.21	36.94
	Reading – Informational & Literature	21	24	0.76	293.11	48.95	23.75
	Writing	8	10	0.46	299.19	50.25	36.22
7	Language	8	10	0.61	295.13	50.89	30.81
	Reading – Informational & Literature	21	23	0.77	293.12	39.43	18.94
	Writing	8	10	0.57	289.62	47.23	29.93
8	Language	8	10	0.55	283.06	55.48	36.26
	Reading – Informational & Literature	21	23	0.72	289.24	40.73	21.42
	Writing	8	10	0.39	278.82	45.18	34.75
11	Language	8	9	0.61	302.68	66.34	39.71
	Reading – Informational & Literature	22	24	0.75	299.26	42.83	21.24
	Writing	8	10	0.57	288.74	52.40	33.52

Note. Based on this data and recommendation of the HIDOE Technical Advisory Committee, scores for strands are not reported.

Table 72. Marginal Reliability Coefficients for Content Strand Scores - Mathematics

Grade	Strand*	Number of Items Specified in Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
3	Measurement and Data & Geometry	13	14	0.61	289.49	51.23	31.07
	Numbers and Operations – Fractions	8	9	0.32	290.54	50.76	40.88
	OA & NBT	17	18	0.57	296.56	42.37	26.85
4	Measurement and Data & Geometry	10	10	0.55	300.84	58.81	39.04
	Numbers and Operations – Fractions	14	14	0.47	291.96	46.03	32.74
	OA & NBT	16	16	0.53	286.15	45.82	30.67
5	Measurement and Data & Geometry	11	12	0.55	301.61	45.37	30.09
	Numbers and Operations – Fractions	12	13	0.41	301.43	40.95	31.30
	OA & NBT	15	16	0.66	297.06	45.73	26.62
6	NS & EE	18	20	0.60	274.83	65.11	39.20
	RP & G	12	14	0.54	278.57	66.79	45.07
	Statistics and Probability	8	9	0.26	272.20	70.15	59.11
7	NS & EE	16	18	0.44	283.15	54.70	39.56
	RP & G	14	16	0.46	273.54	58.53	41.87
	Statistics and Probability	8	9	0.29	282.47	68.63	55.94
8	Functions & Statistics and Probability	11	12	0.47	286.50	55.81	40.18
	Geometry	12	14	0.55	282.61	54.29	36.21
	NS & EE	15	16	0.61	272.15	60.47	36.24
11	Functions & Statistics and Probability	12	12	0.45	289.99	47.66	35.28
	Geometry	9	9	0.38	280.08	52.52	41.24
	Number Quantity & Algebra	19	19	0.48	288.48	38.11	27.48

Note. Based on this data and recommendation of the HIDOE Technical Advisory Committee, scores for strands are not reported. OA & NBT=Operations and Algebraic Thinking & Number and Operations in Base Ten; RP & G=Ratios and Proportional Relationships & Geometry; NS & EE=The Number System & Expressions and Equations.

Table 73. Marginal Reliability Coefficients for Content Strand Scores - Science

Grade	Strand*	Number of Items Specified in Blueprint		Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM
		Min	Max				
5	Earth & Space Science	13	14	0.63	288.08	61.24	37.06
	Life Science	12	13	0.53	280.71	57.66	38.33
	Physical Science	13	14	0.66	279.82	64.66	37.43
8	Earth & Space Science	13	13	0.53	279.95	44.77	30.70
	Life Science	13	14	0.59	277.29	49.52	31.48
	Physical Science	13	14	0.61	275.32	48.81	30.43
11	Life Science	13	13	0.50	291.73	50.28	35.47
	Ecosystems: Interactions, Energy and Dynamics	13	14	0.66	294.63	62.70	36.15
	Heredity and Biological Evolution	13	14	0.66	289.40	62.52	35.35

Note. Based on this data and recommendation of the HIDOE Technical Advisory Committee, scores for strands are not reported.

7. SCORING

For the HSA-Alt assessments, each student receives an overall scale score and an overall performance level. No subscores are reported. This section describes the rules used in generating overall scores.

7.1 ATTEMPTEDNESS RULES FOR SCORING

If a student logged in to the test administration system, was presented one item, and a valid response was entered for that first item, the student is counted as participated. A valid response is recorded when the student marks on one or more response options, or the TA marks *No Response* (NR) on the item. Participated students are counted as attempted.

Scores are generated only for tests with Test Attempted = Y. Please see Section 2 on Test Administration for more information on the test segments.

- If a student answered all items in Segments 1 and 2, the test will be scored without penalty.

  Note: If a student completes Segments 1 and 2 but does not complete all of Segment 3, the test will be scored without penalty.

- If a student did not complete Segments 1 and 2, but generated five or more valid responses with at least one non-NR response, the student is scored with penalty. The penalty is the theta estimate minus one conditional SEM for the estimated theta value.

- If a student generated at least one but fewer than five valid responses or consecutive NR responses for items within Segment 1, the student is given the lowest obtainable scale score. The SEM and theta score will be set to BLANK.

- If a student has NRs on all eight items in segment 1 (early stopped record; ESR), the test will end and the student is given the lowest obtainable scale score. The SEM and theta score will be set to BLANK.

Table 8 in Section 3.1 lists the number of "completed" tests without scoring penalty, the number of "Incomplete" tests (2nd and 3rd bullets), and the number of ESR students.

7.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The HSA-Alt assessments are scored using maximum likelihood estimation (MLE). The likelihood function for generating the MLEs is based on a mixture of item score points.

Indexing items by *i*, the likelihood function based on the *j*th person's score pattern for *I* items is

$$L_j\left(\theta_j|\boldsymbol{z}_j, b_1, \ldots b_k\right) = \prod_{i=1}^{I} p_{ij}\left(z_{ij}|\theta_j, b_{i,1}, \ldots b_{i,m_i}\right),$$

where $b_i' = (b_{i,1}, \ldots, b_{i,m_i})$ for the *i*th item's step parameters, m_i is the maximum possible score of this item, z_{ij} is the observed item score for the person *j*, and *k* indexes the step of the item *i*.

Depending on the item score points, the probability $p_{ij}(z_{ij}|\theta_j, b_i, \ldots, b_{i,m_i})$ takes either the form of the Rasch model for items with one point or the form based on the partial credit model (PCM) for items with two or more points.

In the case of items with one score point, we have $m_i = 1$,

$$p_{ij}(z_{ij}|\theta_j, b_{i,1}) = \begin{cases} \dfrac{exp\left((\theta_j - b_{i,1})\right)}{1 + exp\left((\theta_j - b_{i,1})\right)}, & if\ z_{ij} = 1 \\ \dfrac{1}{1 + exp\left((\theta_j - b_{i,1})\right)}, & if\ z_{ij} = 0 \end{cases}$$

in the case of items with two or more points,

$$p_{ij}(z_{ij}|\theta_j, b_{i,1}, \dots b_{i,m_i}) = \begin{cases} \dfrac{\exp\left(\sum_{k=1}^{z_{ij}}(\theta_j - b_{i,k})\right)}{s_{ij}(\theta_j, b_{i,1,\dots}b_{i,m_i})}, & if\ z_{ij} > 0 \\ \dfrac{1}{s_{ij}(\theta_j, b_{i,1,\dots}b_{i,m_i})}, & if\ z_{ij} = 0 \end{cases}$$

where $s_{ij}(\theta_j, b_{i,1}, \dots, b_{i,m_i}) = 1 + \sum_{l=1}^{m_i} \exp\left(\sum_{k=1}^{l}(\theta_j - b_{i,k})\right)$.

The MLE theta is then estimated by finding the value of theta that maximizes the loglikelihood, i.e.,

$\hat{\theta}_j = \text{argmax} \log\left(L_j(\theta_j|\mathbf{z}_j, \mathbf{b}_1, \dots, \mathbf{b}_I)\right)$.

Standard Error of Measurement

With MLE, the standard error (SE) for student *j* is:

$$SE(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

where $I(\theta_j)$ is the test information for student *j*, calculated as:

$$I(\theta_j) = \sum_{i=1}^{I}\left(\frac{\sum_{l=1}^{m_i} l^2 Exp\left(\sum_{k=1}^{l}(\theta_j - b_{i,k})\right)}{s_{ij}(\theta_j, b_{i,1}, \dots, b_{i,m_i})} - \left(\frac{\sum_{l=1}^{m_i} lExp\left(\sum_{k=1}^{l}(\theta_j - b_{i,k})\right)}{s_{ij}(\theta_j, b_{i,1}, \dots, b_{i,m_i})}\right)^2\right),$$

where m_i is the maximum possible score point (starting from 0) for the *i*th item.

7.3 Rules for Transforming Theta to Scale Scores

The scale score is the linear transformation of the IRT ability estimate using the scaling constants *a* and *b*, $SS = a * \theta + b$, where *a* is the slope and *b* is the intercept.

Table 74 provides the linear transformation constants, intercept, and slope values with four decimals. For the score reports, the scale scores computed applying the slope and intercept for individual students will be rounded to the nearest integer.

Table 74. Scaling Constants

Subject	Grade	Slope (a)	Intercept (b)
ELA	3	58.2226	315.2557
	4	34.9890	313.1294
	5	47.1900	313.4609
	6	49.9795	308.6650
	7	40.4259	305.903
	8	45.6364	299.7642
	11	46.5888	296.4862
Mathematics	3	52.2253	313.5599
	4	56.2908	325.0816
	5	48.9529	319.7003
	6	74.9348	325.9483
	7	72.7005	324.0774
	8	61.1726	322.9731
	11	56.3914	316.6731
Science	5	62.3787	312.6114
	8	53.1189	298.1127
	11	60.3206	311.5589

Standard errors of the MLEs are transformed to be placed onto the reporting scale. This transformation is:

$$SE_{ss} = a * SE_{\theta},$$

where SE_{ss} is the standard error of the ability estimate on the reporting scale, SE_{θ} is the standard error of the ability estimate on the θ scale, and a is the slope of the scaling constant that transforms θ into the reporting scale. The scale scores are mapped into four performance levels. Table 75 provides the range of scale scores in each performance level by grade and subject.

Table 75. Range of Scale Scores by Performance Level

Subject	Grade	Well Below	Approaches	Meets	Exceeds
ELA	3	100-286	287-299	300-331	332-500
	4	100-286	287-299	300-317	318-500
	5	100-281	282-299	300-328	329-500
	6	100-278	279-299	300-330	331-500
	7	100-277	278-299	300-324	325-500
	8	100-275	276-299	300-333	334-500
	11	100-269	270-299	300-327	328-500
Mathematics	3	100-277	278-299	300-315	316-500
	4	100-277	278-299	300-336	337-500
	5	100-288	289-299	300-322	323-500
	6	100-273	274-299	300-336	337-500
	7	100-269	270-299	300-325	326-500
	8	100-275	276-299	300-321	322-500
	11	100-282	283-299	300-316	317-500
Science	5	100-269	270-299	300-335	336-500
	8	100-265	266-299	300-331	332-500
	11	100-264	265-299	300-331	332-500

7.4 LOWEST/HIGHEST OBTAINABLE SCALE SCORES (LOSS/HOSS)

Extremely unreliable student ability estimates are truncated to the lowest obtainable scale score (LOSS) or the highest obtainable scale score (HOSS). For the HSA-Alt assessments, the minimum and maximum scale scores are set at 100 and 500. For the overall scale scores, scale scores lower than 100 or higher than 500 are truncated to 100 or 500. The standard errors for LOSS and HOSS are computed using the estimated theta scores based on the responded items.

7.5 SCORING ALL CORRECT AND ALL INCORRECT CASES

With item response theory (IRT) maximum likelihood (ML) ability estimation methods, 0 and perfect scores are assigned the ability of minus and plus infinity. All incorrect tests are scored by adding 0.3 to an item score among the administered operational items for a test. All correct tests are scored by subtracting 0.3 from an item score among the administered operational items for a student.

8. PERFORMANCE STANDARDS

In the summer of 2019, following the close of the testing window, the American Institutes for Research (AIR) (now CAI) convened panels of Hawai`i educators to recommend performance standards on each of the HSA-Alt ELA and mathematics assessments. From July 9–11, 2019, AIR, under contract to HIDOE, invited a panel of 54 teachers and administrators to recommend performance standards (new cut scores) for the test. HIDOE recruited a broadly representative panel ensuring that a diverse range of perspectives informed the standard-setting process. Panelists included special-education teachers, curriculum specialists, education administrators, and other stakeholders. The panel was also broadly representative of Hawai`i's special education teacher population in terms of gender, race/ethnicity, and regional composition. HIDOE designated the most knowledgeable and experienced panelists at the workshop as table leaders.

In the summer of 2021, following the close of the testing window, CAI convened panels of Hawai`i educators to recommend performance standards on each of the HSA-Alt science assessments. On July 15–16, 2021, CAI, under contract to HIDOE, invited a panel of 21 teachers and administrators to recommend performance standards (new cut scores) for the test. HIDOE recruited a broadly representative panel ensuring that a diverse range of perspectives informed the standard-setting process. Panelists included special education teachers, curriculum specialists, education administrators, and other stakeholders. The panel was also broadly representative of Hawai`i's special education teacher population in terms of gender, race/ethnicity, and regional composition. HIDOE designated the most knowledgeable and experienced panelists at the workshop as table leaders.

Confirmation Standard Setting Workshops for Mathematics and Science

After the original ELA and mathematics standard setting in 2019, and the science standard setting in 2021, WebbAlign conducted an alignment study for mathematics and science and recommended changes to HIDOE's essence statements. Based on WebbAlign's recommendations, HIDOE changed their essence statements to include more detailed, actionable language that reflects the claims being measured in their assessments. HIDOE also chose to reject some items from the mathematics and science item pools that were included on the standard setting OIBs and edited some of the PLDs.

To determine whether the location of the performance standards adopted in 2019 for mathematics and 2021 for science continue to validly describe students' levels of proficiency with respect to these changes, HIDOE conducted a workshop in July 2023 designed to re-evaluate the appropriateness of the performance standards adopted for the Hawai`i Alternate Assessments (HSA-Alt) in mathematics and science.

After reviewing changes in the range performance level descriptors (PLDs), creating threshold PLDs, and reviewing ordered item booklets of the Hawai`i Alternate Assessments, panelists came to a consensus for all grades in mathematics (3–8, 11) and science (5, 8, and 11), that the existing performance standards still accurately classify students as belonging in the performance levels based on the PLDs.

This section of the technical report briefly describes the procedures used by educators to recommend standards and resulting performance standards. Details of the panels, procedures, and outcomes are documented in the "Hawai`i Alternate Assessments Standard Setting" technical reports for ELA and mathematics (2019) and science (2021), and the Hawai`i Alternate Assessments confirmation standard setting technical report (2023).

8.1 STANDARD-SETTING PROCEDURES

Hawai'i used the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001), which is the most common procedure used throughout the country. In this process, the panelists review items ordered by difficulty in an ordered-item booklet (OIB) for each test. Each OIB contains a set of items that meet the test blueprint. The panelists also reviewed the corresponding Hawai`i content standards and HSA-Alt Essence Statements and Range Performance Level Descriptors (PLDs) for each test. With this information in mind, the panelists selected pages in the OIB that best represent the cut scores on the test. The Bookmark standard-setting process was described in a standard-setting plan submitted to HIDOE. The plan was reviewed by the Hawai`i Technical Advisory Committee and approved by HIDOE prior to the workshop.

The standard-setting workshop was held over three days. The first day was devoted to training and review of materials, and the last two days were devoted to two rounds of standard setting. At the end of the activity, the panelists completed a survey that evaluated the workshop.

8.2 PERFORMANCE-LEVEL DESCRIPTORS

HSA-Alt item development was based on the HSA-Alt Essence Statements for ELA, mathematics, and science. These Essence Statements are an extension of the Hawai`i Common Core Standards and provide a full description of content to be targeted and tested for students with significant cognitive disabilities. Based on the general education content standards, the HSA-Alt Essence Statements preserve the core of the grade-level expectations, but may modify the scope or complexity of the general education standards or take the form of introductory or prerequisite skills to the grade-level standards.

A prerequisite to standard setting is to determine the nature of the categories into which students are classified. These categories, or performance levels, are associated with PLDs. PLDs link the Hawai`i Common Core Standards to the performance expectations for the test (Essence Statements). There are three types of PLDs used within the HSA-Alt program:

1. **Policy PLDs.** Brief description of the policy goals of each performance level that do not vary across grade or content.

2. **Range PLDs.** Range PLDs describe what students should know and be able to do at different proficiency levels. For example, the range PLD for Meeting Proficiency describes what students know and can do at that level all the way to just below the Exceeding Proficiency cut score. This document also contains the HSA-Alt Essence Statements which are the basis for the HSA-Alt.

3. **"Just Barely" PLDs.** Sometimes called *threshold* or Target PLDs, Just Barely PLDs are created during the standard-setting workshop and are used for standard setting only. The "Just Barely" PLDs describe what a student "Just Barely" scoring at the bottom of each performance level knows and can do.

The standard-setting panelists used the Essence Statements, Range PLDs, and "Just Barely" PLDs during the standard-setting workshop.

8.3 RECOMMENDED PERFORMANCE STANDARDS

Panelists were tasked with recommending three performance standards (Approaches, Meets, and Exceeds) that resulted in four performance levels (Well below, Approaches, Meets, and Exceeds). Table 76 presents the performance standard associated with panelist-recommended OIB page numbers in scale scores, as well as the percentage of students classified as meeting or exceeding each standard based on the 2019 HSA-Alt results (for ELA and mathematics) and 2021 HSA-Alt results (for science).

Table 76. Final Recommended Performance Standards for HSA-Alt

Grade	Cut Scores			Impact Data			*Impact Data (Include ESR)			Benchmark Data
	Approaches	Meets	Exceeds	Approaches	Meets	Exceeds	Approaches	Meets	Exceeds	Proficient
ELA										
3	287	300	332	75%	57%	25%	69%	53%	24%	49%
4	287	300	318	80%	54%	30%	75%	50%	28%	49%
5	282	300	329	75%	54%	25%	72%	52%	24%	55%
6	279	300	331	80%	49%	25%	78%	48%	25%	50%
7	278	300	325	76%	51%	26%	74%	50%	26%	49%
8	276	300	334	71%	45%	20%	67%	42%	19%	50%
11	270	300	328	71%	36%	17%	69%	35%	17%	57%
Mathematics										
3	278	300	316	80%	54%	27%	75%	51%	25%	53%
4	278	300	337	80%	53%	19%	73%	48%	17%	47%
5	289	300	323	71%	52%	23%	69%	50%	22%	43%
6	274	300	337	71%	45%	15%	70%	44%	15%	40%
7	270	300	326	72%	42%	24%	70%	40%	23%	37%
8	276	300	322	74%	42%	18%	70%	40%	17%	38%
11	283	300	317	67%	36%	17%	66%	35%	17%	31%
Science										
5	270	300	336	60%	39%	12%				37%
8	266	300	332	64%	37%	14%				33%
11	265	300	332	64%	38%	14%				34%

* Conducted only for ELA and Mathematics in spring 2019.

9. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that includes information describing student performance for students, parents, educators, and other stakeholders. The online score reports are generally produced immediately after students complete the tests. Starting in spring 2021, online score reports were generated immediately for ELA and mathematics; starting in spring 2022, online score reports were generated immediately for science. Because the performance score report is updated each time a student completes a test, authorized users (e.g., school principals, teachers) can have available information on students' performance scores quickly and use them to improve student learning. In addition to individual students' score reports, the CRS also produces aggregate score reports by class, school, complex, complex area, and state. The timely accessibility of aggregate score reports could help users to monitor students' performance in each grade by subject area and evaluate the effectiveness of instructional strategies; it can also inform the adoption of strategies to improve student learning and teaching and inform professional development for educators and curriculum decisions for the state over time.

This section describes the types of scores reported in the CRS and a description of the ways to interpret and use these scores in detail.

9.1 CENTRALIZED REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

9.1.1. Types of Online Score Reports

The CRS is designed to help educators and students answer questions about how students have performed on English language arts/literacy (ELA), mathematics, and science assessments. The CRS is the online tool that provides educators and other stakeholders with timely, relevant score reports. The CRS for the Hawai`i Alternate Assessment has been designed with stakeholders who are not technical measurement experts in mind, with the intention to make score reports to be easy to read and understand for a non-technical audience. This is achieved by using simple language so that users can quickly understand assessment results and make inferences about student achievement. The CRS is also designed to present student performance in a uniform format. For example, similar colors are used for groups of similar elements, such as achievement levels, throughout the design. This design strategy allows readers to compare similar elements and to avoid comparing dissimilar ones.

Once authorized users log in to the CRS, the dashboard page shows overall test results for all tests that the students have taken grouped by test family (e.g., grade 5 science, grade 6 ELA). Once the user clicks on the test family that he or she wants to further explore, it will take the user to the detailed dashboard, where the results are shown by test (e.g., grade 3 ELA). Additionally, when authorized state-level users log in to the CRS and select "State View," the CRS generates a summary of student performance data for a test across the entire state.

Generally, the CRS provides two categories of online score reports: (1) aggregate score reports, and (2) student score reports. Table 77 summarizes the types of online score reports available at the aggregate level and the individual student level. Detailed information about the online score reports and instructions on how to navigate the online score reporting system can be found in the *Centralized Reporting System User Guide*, located via a help button on the CRS.

Table 77. Types of Online Score Reports by Level of Aggregation

Level of Aggregation	Types of Online Score Reports
State Complex Area Complex School Teacher Roster	• Number of students tested and percentage of proficient students (for overall students and by subgroup) • Average scale score (for overall students and by subgroup) • Percentage of students at each performance level • On-demand student roster report
Student	• Total scale score and Standard Error of Measurement • Performance level for overall score with Performance-Level Descriptors • Average scale scores for individual schools, complexes, complex areas, and states

Aggregate score reports at a selected aggregate level are provided for overall students and by subgroup. Users can see student assessment results by any of the subgroups. Average scale score and performance levels will be calculated at n≥2. Table 78 presents the types of subgroups and subgroup categories provided in the CRS.

Table 78. Types of Subgroups

Subgroup	Subgroup Category
Gender	Male Female
ELL	ELL Not ELL
Disability	With Disability No Disability
Migrant Status	Migrant Not Migrant
Disadvantaged	Disadvantaged Not Disadvantaged
Ethnicity	American Indian or Alaska Native Asian Black or African American Hispanic or Latino Native Hawaiian or Pacific Islander White Two or More Races

9.1.2. Centralized Reporting System

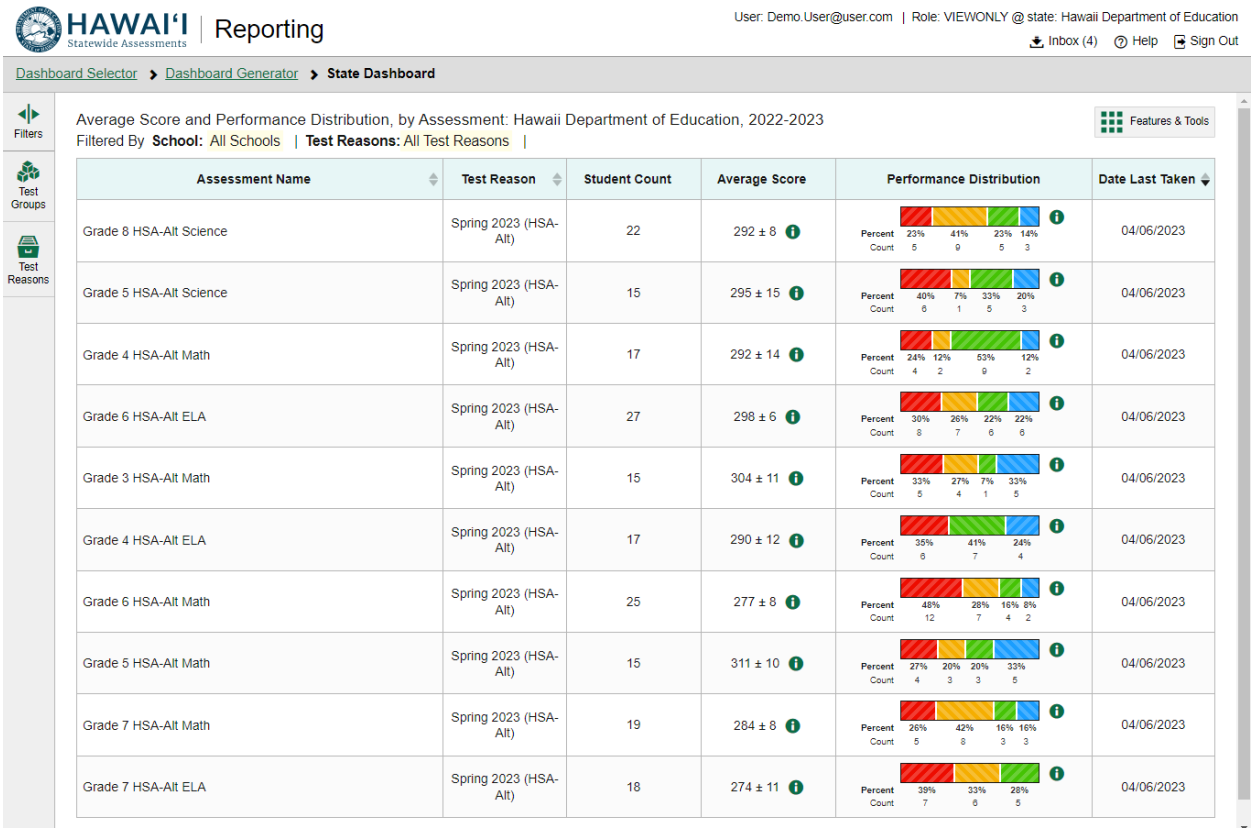
9.1.2.1 Dashboard

The first page users see when they log in to the CRS contains summaries of student performance by test family (i.e., HSA-Alt ELA). Complex personnel see complex summaries, school personnel see school

summaries, and teachers see summaries of their students. State personnel and complex-area personnel would need to select the specific complex in order to view the aggregate results.

The dashboard summarizes students' performance by test family, including (1) the number of students tested, (2) the grades of the students who have tested, and (3) the percentage and counts of students at each performance level. Exhibit 1 presents a sample dashboard page at the state level.

Exhibit 1. Dashboard: State Level

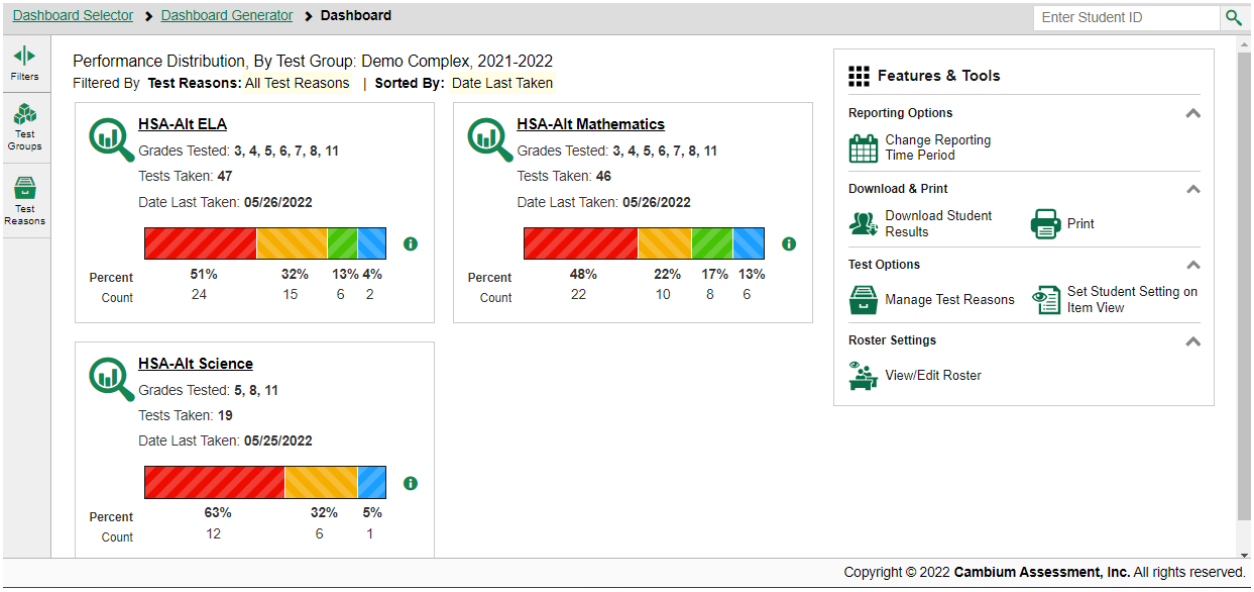


The four performance achievement levels are color-coded in the performance distribution bar as follows:

1. Red is the percentage of "Well Below" students.
2. Orange is the percentage of "Approaches" students.
3. Green is the percentage of "Meets" students.
4. Blue is the percentage of "Exceeds" students.

Educators can click on the subject group to view individual test results for the selected test group. Once the user clicks on the test family that he or she wants to explore further, the detailed dashboard page will appear. The detailed dashboard summarizes students' performance by test, including (1) the number of students tested, (2) average score and standard error of the means, and (3) the percentage and counts of students at each performance level. Exhibit 2 presents a sample detailed dashboard page for HSA-Alt Assessments at the complex level.

Exhibit 2. Dashboard: Complex-Area Level

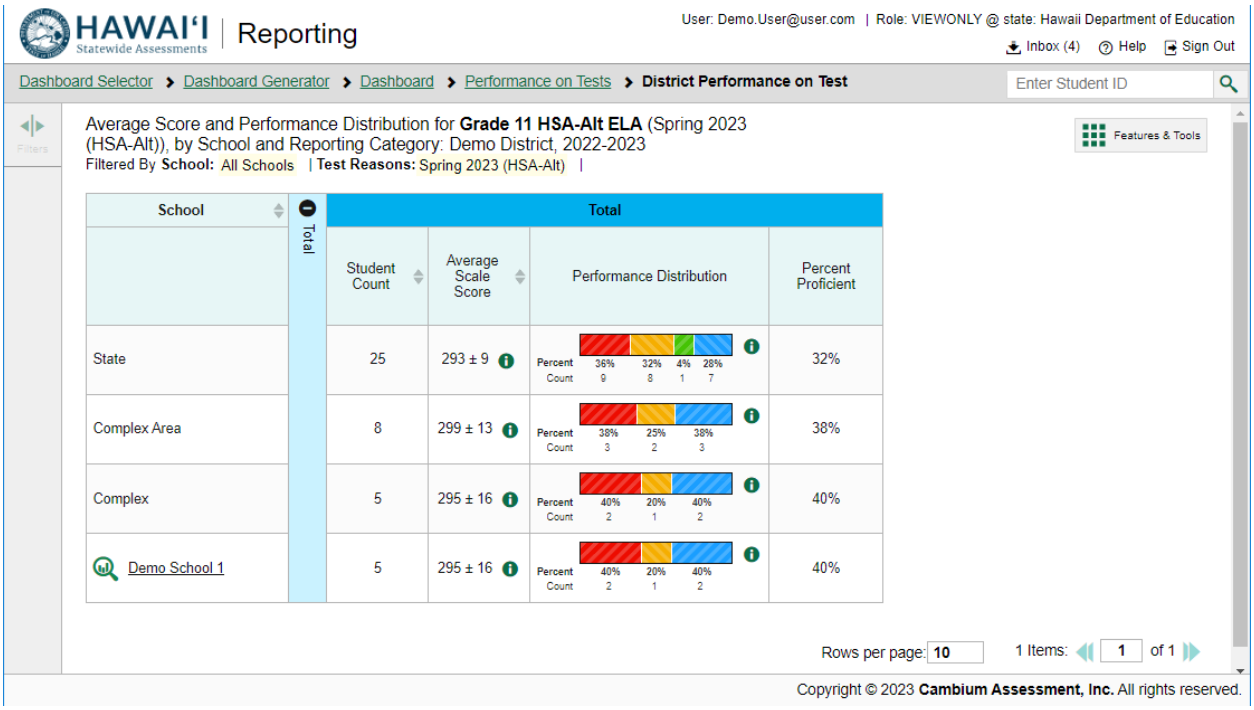


9.1.2.2 Subject Detail Page

Detailed summaries of student performance for each grade in a subject area for a selected aggregate level are presented when users select a specific assessment name. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for the state and the aggregate unit above the selected aggregate. For example, if a school is selected, the summary results of the state, complex area, and complex of the school are provided above the school summary results as well, so that school performance can be compared with the aggregate levels.

The aggregated subject summary report provides the summaries on a specific subject area, including (1) the number of students tested, (2) the average scale score and standard error associated with the average scale score, (3) the percentage of proficient students, and (4) the percentage and counts of students in each achievement level. The summaries are also presented for students overall and by subgroup. Exhibit 3 presents an example of subject summary results for grade 5 math with gender breakdowns at the complex level.

Exhibit 3. Subject Detail Page for HSA-Alt ELA by Gender: Complex-Area Level



9.1.2.3 Student Detail Page

When a student completes a test, an online score report appears in the individual student report in the CRS. The individual student report shows individual student performance on the test. In each subject area, the individual student report provides (1) the scale score and standard error of measurement (SEM); (2) achievement level for overall test; and (4) average scale scores for student's state, complex area, complex, and school.

The student's name, scale score with the SEM and performance level are shown at the top of the page. In the middle section, the student's performance is described in detail using a barrel chart. In the barrel chart, the student's scale score is presented with the SEM using a "±" sign. SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered multiple times. Furthermore, in the barrel chart, Performance-Level Descriptors with cut scores at each achievement level are provided. This defines the content-area knowledge, skills, and processes that test takers at the achievement level are expected to possess.

Underneath, average scale scores and standard errors of the average scale scores for state, complex area, complex, and school are displayed so that student achievement can be compared with the above aggregate levels. It should be noted that the "±" next to the student's scale score is the standard error of measurement of the scale score, whereas the "±" next to the average scale scores for aggregate levels represents the standard error of the average scale scores.

On the following page, the trend of student performance over time is displayed. Exhibit 4, 5, and 6 present examples of individual student reports.

Exhibit 4. Student Detail Page for HSA-Alt ELA

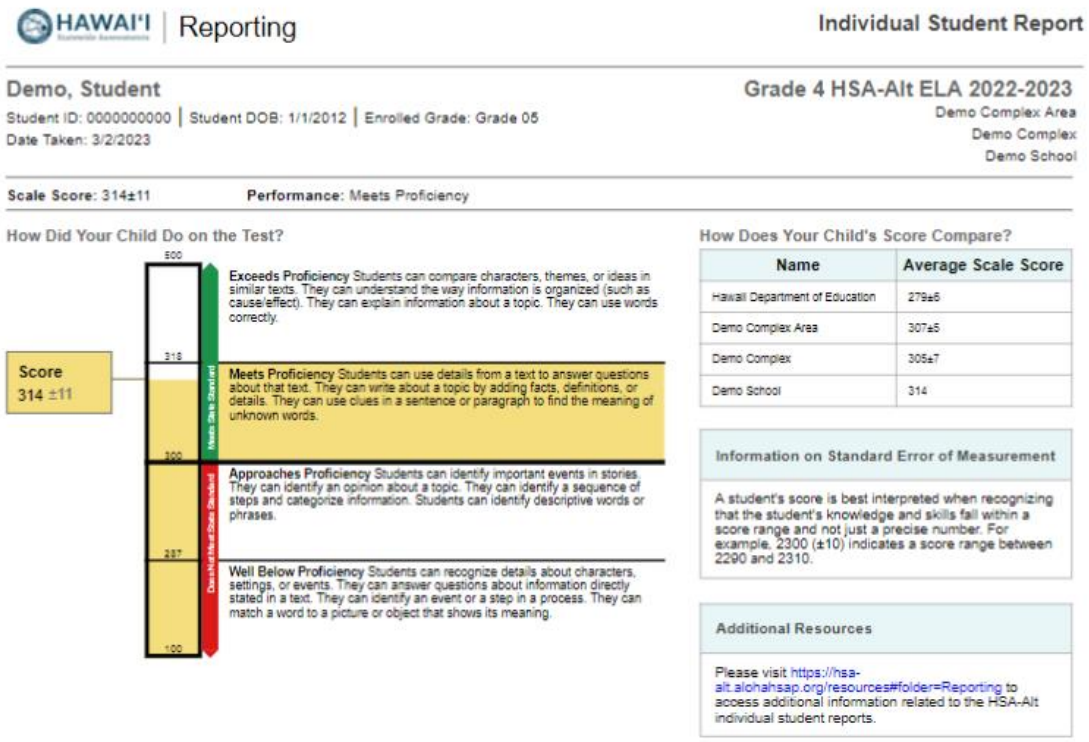


Exhibit 5. Student Detail Page for HSA-Alt Mathematics

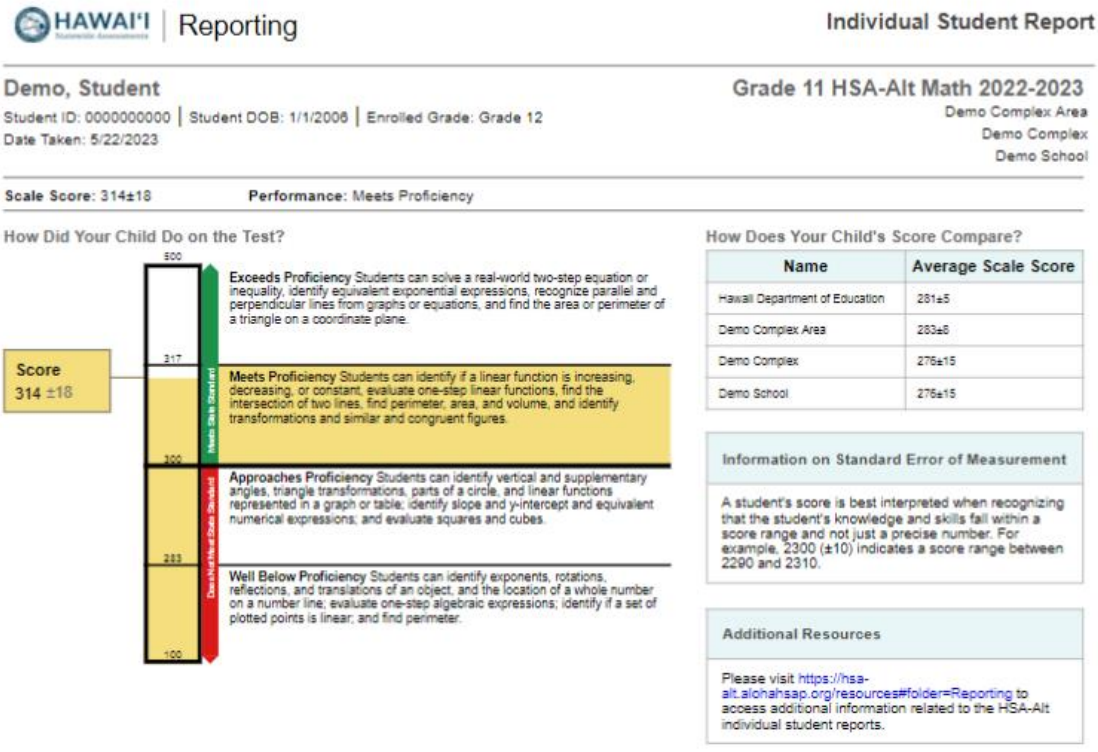
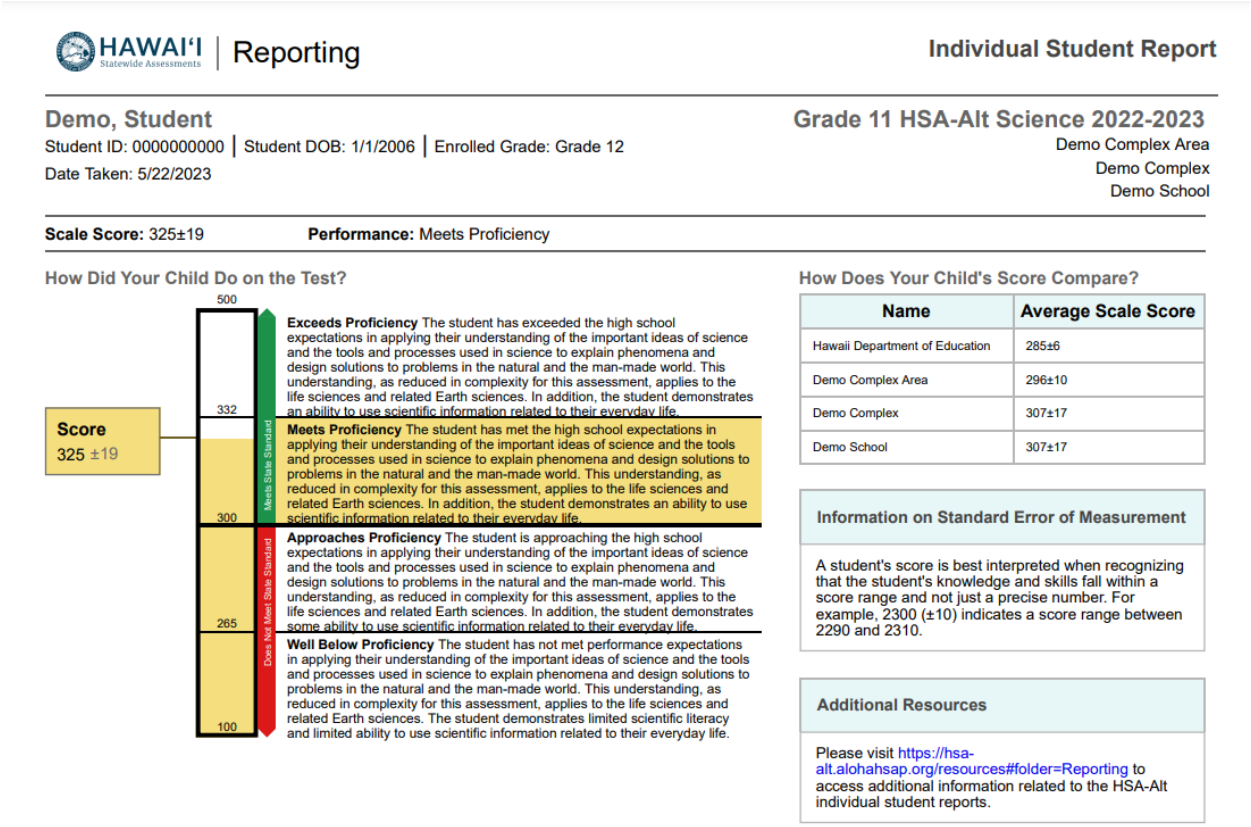


Exhibit 6. Student Detail Page for HSA-Alt Science



HAWAI'I Statewide Assessments | Reporting

Individual Student Report

Demo, Student
Student ID: 0000000000 | Student DOB: 1/1/2006 | Enrolled Grade: Grade 12
Date Taken: 5/22/2023

Grade 11 HSA-Alt Science 2022-2023
Demo Complex Area
Demo Complex
Demo School

Scale Score: 325±19 **Performance:** Meets Proficiency

How Did Your Child Do on the Test?

500
332
300
265
100
Score 325 ±19
Meets State Standard
Does Not Meet State Standard

Exceeds Proficiency The student has exceeded the high school expectations in applying their understanding of the important ideas of science and the tools and processes used in science to explain phenomena and design solutions to problems in the natural and the man-made world. This understanding, as reduced in complexity for this assessment, applies to the life sciences and related Earth sciences. In addition, the student demonstrates an ability to use scientific information related to their everyday life.

Meets Proficiency The student has met the high school expectations in applying their understanding of the important ideas of science and the tools and processes used in science to explain phenomena and design solutions to problems in the natural and the man-made world. This understanding, as reduced in complexity for this assessment, applies to the life sciences and related Earth sciences. In addition, the student demonstrates an ability to use scientific information related to their everyday life.

Approaches Proficiency The student is approaching the high school expectations in applying their understanding of the important ideas of science and the tools and processes used in science to explain phenomena and design solutions to problems in the natural and the man-made world. This understanding, as reduced in complexity for this assessment, applies to the life sciences and related Earth sciences. In addition, the student demonstrates some ability to use scientific information related to their everyday life.

Well Below Proficiency The student has not met performance expectations in applying their understanding of the important ideas of science and the tools and processes used in science to explain phenomena and design solutions to problems in the natural and the man-made world. This understanding, as reduced in complexity for this assessment, applies to the life sciences and related Earth sciences. The student demonstrates limited scientific literacy and limited ability to use scientific information related to their everyday life.

How Does Your Child's Score Compare?

Name	Average Scale Score
Hawaii Department of Education	285±6
Demo Complex Area	296±10
Demo Complex	307±17
Demo School	307±17

Information on Standard Error of Measurement

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and not just a precise number. For example, 2300 (±10) indicates a score range between 2290 and 2310.

Additional Resources

Please visit https://hsa-alt.alohahsap.org/resources#folder=Reporting to access additional information related to the HSA-Alt individual student reports.

9.1.3. Interpretation of Reported Scores

A student's performance on a test is reported in a scale score and on a performance level for the overall test. Students' scores and performance levels are summarized at the aggregate levels. The next section describes how to interpret these scores.

9.1.4. Scale Score

A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of the students' knowledge and skills. The scale score is the transformed score from a theta score estimated based on mathematical models. Low scale scores can be interpreted to mean that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high-scale scores can be interpreted to mean that the student has proficient knowledge and skills measured by the test. Scale scores can be used to measure student growth across school years. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and PLDs.

9.1.5. Standard Error of Measurement

A scale score (observed score on any test) is an estimate of the true score. If a student takes a similar test multiple times, the resulting scale score would vary across administrations, being sometimes a little higher,

a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered multiple times. When interpreting scale scores, it is recommended to consider the range of scale scores incorporating the SEM of the scale score.

The "±" next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one SEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, "312 ± 18" indicates that if a student were tested again, he or she would likely receive a score between 294 and 330. SEM can be different for the same scale score, depending on how closely the administered items match the student's ability.

9.1.6. Performance Level

Performance levels are proficiency categories on a test that students fall into based on their scale scores. For the HSA-Alt Assessments, scale scores are mapped into four performance levels (i.e., Well Below Proficiency, Approaches Proficiency, Meets Proficiency, Exceeds Proficiency) using three performance standards (i.e., cut scores). These four performance levels are identified and set by educators during the Standard Setting process described in the previous chapter. Please see Chapter 8: Performance Standards for more details on the development of the four performance levels used in the online student reports.

PLDs are a description of the content area knowledge and skills that test takers at each performance level are expected to possess. Thus, performance levels can be interpreted based on the PLDs.

9.1.7. Aggregated Score

Student scale scores are aggregated at roster, teacher, school, complex, complex area, and state levels to represent how a group of students performed on a test. When students' scale scores are aggregated, the aggregated scale scores can be interpreted as an estimate of the knowledge and skills that a group of students possesses. Given that student scale scores are estimates, the aggregated scale scores are also estimates and are subject to measures of uncertainty. In addition to the aggregated scale scores, the percentage of students in each performance level for the overall test is reported at the aggregate level to represent how a group of students performed overall.

9.2 Appropriate Uses for Scores and Reports

Assessment results can provide information about individual students' achievement on the test. Overall, these results tell what students know and are able to do in certain subject areas. Additionally, assessment results can be used to identify students' relative strengths and weaknesses in certain content areas.

Assessment results for student achievement on the test can be used to help teachers or schools make decisions on how to support student learning. Aggregate score reports provide a summary of the average overall scale score of all students at that aggregate level. The aggregate score reports may be used to monitor the trends of the student proficiency or subgroup proficiency, or planning the professional development for teachers. Individual student report may provide more useful information for student's learning and teaching considering the diverse needs of the students with the significant cognitive disabilities.

In addition, assessment results can be used to compare student performance among different students and among different groups. Teachers can evaluate how their students performed compared with students in other schools, complexes, complex areas, and the state overall.

Although assessment results provide valuable information to understand student performance, these scores and reports should be used with caution. It is important to note that scale scores reported are estimates of true scores and, therefore, do not represent a precise measure of student performance. A student's scale score is associated with measurement error, and, thus, users need to consider measurement error when using student scores to make decisions about student achievement. Moreover, although student scores may be used to help make important decisions about students' placement and retention, or teachers' instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student achievement, such as classroom assessment and teacher evaluation, should be considered when making decisions about student learning.

10. QUALITY CONTROL PROCEDURES

Quality assurance (QA) procedures are enforced through all stages of alternate assessment development, administration, and scoring and reporting of results. CAI uses a series of quality control steps to ensure the error-free production of score reports. The quality of the information produced in the Test Delivery System (TDS) is tested thoroughly before, during, and after the testing window opens.

10.1 OPERATIONAL TEST CONFIGURATION

For the operational test, a test configuration file is the key file that contains all specifications for the item selection algorithm and the scoring algorithm, such as the test blueprint specification, slopes and intercepts for theta-to-scale score transformation, cut scores, and the item information (e.g., answer keys, item attributes, item parameters, passage information). The accuracy of the information in the configuration file is independently checked and confirmed numerous times by multiple staff members before the testing window opens.

To verify the accuracy of the scoring engine, we use simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the population. The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability distribution.

Simulations are generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a wide range of student response patterns. The results of simulated test administrations are used to configure and evaluate the adequacy of the item selection algorithm used to administer the HSA-Alt assessments. The purpose of the simulations is to configure the algorithm to optimize item selection to meet blueprint specifications as well as to check the score accuracy. The scores in the simulated data file are checked independently, following the scoring rules specified in the scoring specifications.

10.1.1. Platform Review

CAI's TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems such as Windows, Linux, and iOS to ensure that the item looks consistent in all of them. For HSA-Alt assessment, there are two commonly used layouts: one has the stimulus and item response options/response area displayed side by side, where stimulus and response options have independent scroll bars; the other has the item stem and responses on the full screen.

Platform review is a process during which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on various platforms that are significantly different from one another.

Platform review is conducted by a team. The team leader projects the item as it was web approved in the Item Tracking System (ITS), and team members, each using a different platform, look at the same item to confirm that it is rendered as expected.

10.1.2. User Acceptance Testing and Final Review

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to user acceptance testing (UAT). UAT of the TDS serves as both a software evaluation and a content approval role. The UAT period provides the HIDOE with an opportunity to interact with the exact test that the students will use.

10.2 Quality Assurance in Data Preparation

CAI's TDS has a real-time quality-monitoring component built in. After a test is administered to a student, the TDS passes the resulting data to our QA system. QA conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item, keys for multiple-choice items, score points in each item, total number of field-test items and operation items, and that the test record contains no data from items that have been invalidated.

Data pass directly from the Quality Monitoring System (QMS) to the Database of Record (DOR), which serves as the repository for all test information and from which all test information for reporting is pulled. The Data Extract Generator (DEG) is the tool that is used to pull data from the DOR for delivery to the HIDOE. CAI staff ensures that data in the extract files match the DOR before delivering it to the HIDOE.

10.3 Quality Assurance in Test Scoring

To monitor the performance of the TDS during the test administration window, CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic, state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and CAI contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. The applications log not only errors and exceptions, but also latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem. In addition, latency data, such as data about how long it takes to load, view, or respond to an item, are captured for each assessed student. All of this information is logged, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

A series of quality assurance reports, such as blueprint match rate, item exposure rate, and item statistics, can also be generated at any time during the online assessment window for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved.

Blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to the simulation results. The QA reports can be generated on any desired schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to the blueprint and that items are performing as anticipated.

The item statistics analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. This report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial

correlation. The report is configurable and can be produced so that only items with statistics falling outside of a specified range are flagged for reporting or to generate reports based on all items in the pool.

Table 79 presents an overview of the QA reports.

Table 79. Overview of Quality Assurance Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items)
Blueprint Match Rates	To monitor unexpectedly low blueprint match rates	Early detection of unexpected blueprint match issue
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (highly unused items/passages)	Early detection of any oversight in the blueprint specification

10.3.1. Score Report Quality Check

Online Report Quality Assurance

Scores for online assessments are assigned by automated systems in real time. During operational testing, actual item responses are compared to expected item responses (given the item response theory [IRT] parameters), which can detect miskeyed items, item score distribution, or other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed to the Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the CRS until it passes all of the QA system's validation checks.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage, 1994.

Guo, F. (2006). Expected Classification Accuracy using the Latent Distribution. *Practical, Assessment, Research & Evaluation, 11*(6).

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing, *Journal of Educational Measurement, 13*(4), 253–264.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, 5(1), 95–110.

Livingston, S. A., & Lewis, C. (1995), Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement*, 32: 179–197.

Livingston, S. A. & Wingersky, M. S. (1979), Assessing the Reliability of Tests Used to Make Pass/Fail Decisions. *Journal of Educational Measurement*, 16: 247–260.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, 52(2), 443–451.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark Procedfigure 8ure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards*. Mahwah, NJ: Lawrence Erlbaum.

Muniz, J. Hambleton, R. & Xing, D. (2001). Small Sample Studies to Detect Flaws in Item Translations, *International Journal of Testing*, 1:2, 115–135

Sireci, S. G., & Rios, J. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19(2-3), 170–187.

Subkoviak, M. J. (1976). Estimating Reliability From a Single Administration of a Criterion-Referenced Test*. *Journal of Educational Measurement*, 13: 265–276.

Towles-Reeves, E., Kearns, J., Flowers, C., Hart, L., Kerbel, A., Kleinert, H., Quenemoen, R., & Thurlow, M. (2012). Learner characteristics inventory project report (A product of the NCSC validity evaluation). Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

U.S. Department of Education (2018). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* Retrieved from https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf.