Hawai`i Alternate Assessments

2023–2024 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.

September 2024

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

PREFACE

This report provides a technical summary of the 2023–2024 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 12 chapters that discuss all the evidence accrued about the technical quality of a testing system. This report is based on Hawai`i operational test data for the alternate assessments, covering all aspects of the technical qualities for the HSA-Alt outlined 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 *A State's Guide to the U.S. Department of Education's Assessment Peer Review* (U.S. Department of Education [USDE], 2018).

Chapter 1 provides an overview of the HSA-Alt, purposes and intended uses of the HSA-Alt scores, the testing population, and the content standards. Chapter 2 describes the HSA-Alt tests, content specifications in blueprints, and test assembly. Chapter 3 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 summarizes the field-test item analysis results and data review results from the spring 2024 test administration. Chapter 5 documents the test administration procedures, including test administrator training, test administration manual, accommodations, as well as prevention of disruptions in the Test Delivery System (TDS). Chapter 6 describes the scoring procedures used in producing scale scores and performance levels. Chapter 7 summarizes the results of the spring 2024 HSA-Alt test administration in ELA, mathematics, and science. This chapter summarizes the test-taking student population, their performance on the assessments, and the time spent in taking the assessments.

Chapter 8 provides validity evidence on the test blueprint coverage, cognitive lab, internal consistency, and relations to other variables. Chapter 9 provides evidence for the reliability of the HSA-Alt, including internal consistency reliability, standard errors of measurement (SEMs), and the reliability of performance-level classifications. Chapter 10 describes the procedures that the Hawai`i Department of Education (HIDOE) uses to identify and adopt performance standards for the HSA-Alt. Chapter 11 provides a description of the score reporting system and the interpretation of test scores. Chapter 12 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. THE HAWAI`I STATE ALTERNATE ASSESSMENTS

1.1 Overview

The Hawai`i Alternate Assessments (HSA-Alt) is 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 HSA-Alt is intended to support Hawai‘i’s broader efforts to promote access to the general education curriculum—including the knowledge, skills, and abilities defined in the academic content standards—for students with the most significant cognitive disabilities. As one component of the state’s comprehensive educational system, the assessment helps ensure that these students are included in Hawai‘i’s statewide assessment and educational accountability systems. Teachers and educators can use the results to identify potential gaps in student learning, evaluate the effectiveness of instructional approaches, and inform future educational planning—as one of multiple measures used to support student learning. The HSA-Alt is only for those students with documented significant cognitive disabilities and adaptive behavior deficits who require extensive support across multiple settings (e.g., home, school, community). Typically, this student population consists of about one percent of the total student population.

In 2018, Hawai`i Department of Education (HIDOE) began the transition to a new online, computer-adaptive test (CAT) for alternate assessment for students with significant cognitive disabilities. The new assessment was designed to assess students at each performance level (i.e., well below, approaches, meets, and exceeds) in grades 3–8 and 11 in English language arts (ELA) and mathematics, and in grades 5, 8, and 11 in science. Online operational field tests for ELA and mathematics were administered in spring 2019. A standard setting was convened in summer 2019 to set performance standards for ELA and mathematics. Online operational field tests for science were administered in spring 2021. Performance standards for science were set in summer 2021. The transition to CATs was fully implemented in all grade levels and subject areas beginning in spring 2022.

1.2 Purposes, Interpretations, and Intended Uses of HSA-Alt Scores

The purposes, interpretations, and intended uses of the HSA-Alt scores serve as the foundation for test design and development. They play a crucial role in the validation process, as any statements about validity are tied to specific interpretations and uses.

Purposes and Intended Uses

The purposes and intended uses of the HSA-Alt are to measure students’ academic performance and student’s progress in meeting the state alternate academic achievement standards in core content areas, including ELA, mathematics, and science.

To fulfill its intended purposes, the HSA-Alt provides an overall scale score and an associated performance level for each test. These performance levels are determined based on the performance standards established through a formal standard-setting process.

At the individual student level, the HSA-Alt test score can be used to estimate a student’s academic performance. The associated Performance Level, together with the Performance Level Descriptors, is able to indicate the knowledge and skills the student has attained in the assessed content area by the end of the academic year. Individual student scores and Performance Levels can be compared across students who take the same test. Additionally, scores can also be aggregated to estimate the average performance of specific groups or to compare the average performance between different groups, such as by school, district, or by gender.

Intended Test Users

Primary intended users of the HSA-Alt include:

- Students and parents can use the results to stay informed about the student's learning progress in school.
- Teachers and educators can use the results to identify potential gaps in student learning, evaluate the effectiveness of instructional approaches, and inform future educational planning—as one of multiple measures used to support student learning.
- Educational agencies, organizations, and governments can use the test data and results to monitor the educational improvement and make necessary changes to standards.

1.3 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 2024 *Test Administration Manual* at https://hsa-alt.alohahsap.org/resource-list/en/hsa-alt-test-administration-manuals-2023-2024.

The following are the participation guidelines for a Hawaiʻi student to take the HSA-Alt:

- 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.

1.4 Content Standards

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.

The HSA-Alt is aligned to the content standards for ELA, mathematics, and science, which are based on the HCCS. These content standards consist of Essence Statements, which serve as the foundation for the development of the HSA-Alt 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 Cambium Assessment, Inc. (CAI) staff review the HCCS and the NGSS and prioritize content and skills that are 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 PLDs for ELA, mathematics, and science. PLDs have been developed at the following four levels of complexity for each Essence Statement:

1. Exceeds Performance Level—Highest level of performance expectation for the alternate test
2. Meets Performance Level—Meets performance expectation for the alternate test
3. Approaches Performance Level—Approaches performance expectation for the alternate test
4. 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 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.

2. TEST DESIGN AND DEVELOPMENT

2.1 TEST DESCRIPTIONS

The HSA-Alt assesses three content areas: English language arts (ELA) and mathematics for students in grades 3-8, and 11, and science for students in grades 5, 8, and 11. In this technical report, a test is defined as each unique combination of content area and grade level. For example, ELA for grade 3 constitutes one test, mathematics for grade 11 constitutes another test.

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, refer to Section 5.5, Paper-Pencil Test Administration (via Online Fixed Form).

2.2 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. The Hawaii Department of Education (HIDOE) worked with their curriculum and special education departments to draft blueprints for each content area. Cambium Test Development Specialists reviewed the blueprints with HIDOE to clarify domain percentages and item amounts at each assessed level. Final 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 (https://hsa-alt.alohahsap.org). Each student is required to take 40 operational items from domains specific to each test subject.

2.3 TEST SEGMENTS

The online adaptive tests comprise two or three distinct test segments. Test segments are used by the Test Delivery System (TDS) to implement the Early Stopping Rule (ESR), and to implement separate field testing of field-test items shared among multiple states (see Section 3.1 for details) and Hawai`i-specific field-test items (items that are field tested only with Hawai`i students). The test segments are defined as follows:

- Segment 1 comprises eight operational items presented in an adaptive format. This segment is used by the TDS to enforce the ESR. If all eight items in Segment 1 are marked *No Response* (or NR), the system will end the test when the Test Administrator (TA) attempts to move to Segment 2 (item 9).
- Segment 2 comprises 32 operational items and 10 embedded field-test items (EFT) shared across multiple states, presented in an adaptive format. EFT 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.

- Segment 3 comprises between 1–10 Hawai`i-specific field-test items presented as a fixed form. The number of items in Segment 3 differs by grade and subject area. In spring 2024, only ELA grade 8 and mathematics grade 11 had Segment 3 Hawai`i-specific field-test items.

Fixed-form tests for all grades and subjects include only Segments 1 and 2 to allow for implementation of the ESR. Segment 1 comprises eight operational items as a fixed form; segment 2 comprises the remaining 32 operational items as a fixed form. No field-test items are included on the fixed-form tests.

2.4 Test Assembly

The computer-adaptive test (CAT) 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 comprises 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 test 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 level of the form is higher than that of the bank, 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 until 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 (ESR) segment (the first eight 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 segment with the remaining 32 operational items. Additionally, to reduce the reading load on the students, where possible, as many as three or four items linked to the same passage are consecutively placed 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 CAT.

3. ITEM DEVELOPMENT

3.1 MEMORANDUM OF UNDERSTANDING ON ITEM-SHARING INITIATIVE

The item development process for the alternate assessments is a collaborative effort among member states that have signed a Memorandum of Understanding (MOU) for item sharing in item development and field-testing. Each MOU member state retains ownership of the items they developed, but these items would be available for use by other MOU members. The number of items each state is responsible for developing is proportional to its alternate assessment population size. Given that the alternate assessment population in each state is small, the item-sharing initiative enables statistical and psychometric analysis based on combined data from all participating states. As a result, item parameter estimates are more stable compared to those derived from smaller sample sizes.

The MOU for the alternate assessments (MOU-Alt) was initiated in 2018 and originally signed by three states: Hawai`i, South Carolina, and Wyoming, covering English language arts (ELA), mathematics, and science. In early 2019, Idaho and Vermont joined the MOU for ELA, mathematics, and science. Montana and South Dakota joined in 2020, but only for science. Vermont exited the MOU in 2022.

In the 2023-24 academic year, there are six MOU member states: Hawai`i, Idaho, Montana, South Carolina, South Dakota, and Wyoming. South Dakota and Montana participate in the MOU for science only, while the other four states participate in all three subjects.

In addition to the items jointly developed by the MOU member states, each state may also develop items that are specifically aligned with its own content standards.

Each state in the MOU follows a similar process for developing and reviewing their items in collaboration with CAI. Items are developed by each state to fulfill their agreed-upon contribution to the MOU each school year. CAI requires Department of Education (DOE) staff in each participating state 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.

3.2 ITEM TYPES

The HSA-Alt 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 consecutively administered.

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 (TEIs) should be developed for the MOU shared field-test items. The item types included MS, 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.

3.3 Development of Crosswalk of State Alternate Content Standards

A crosswalk across all the individual state alternate content standards was completed for the first year of the MOU-Alt shared field test item development. This crosswalk has been updated as more states joined the MOU since 2018. Specifically, the content of the standards from each of the MOU states were reviewed and compared by special education and content experts at CAI to determine which standards are on-grade and overlapped across states. For example, CAI looked at all of the grade 3 mathematics standards for each MOU state and determined which standards contained common content. If standard A in the first state contained the same content as standard B in the next state, and standard C in the third state, then the three standards in the three states are common. When aligning items to standards in each state, with this crosswalk available, CAI knew instantly which standards items should be aligned to. The opposite is true as well. There were standards that did not have similar content to other states' on-grade standards, so items aligned to those standards were not aligned to other states.

The crosswalk was created by senior test development specialists in CAI and reviewed by the state Departments of Education. The crosswalk was based on each state's blueprint and included the common core standards and the general education and alternate content standards for each state. Each state has a unique set of alternate content standards as follows:

- Hawaii Essence Statements and Performance-Level Descriptors.
- Idaho Extended Content Standards Core Content Connectors.
- Montana Alternate Academic Achievement Standards and Performance-Level Descriptors.
- South Carolina Alternate Academic Achievement Standards and Performance-Level Descriptors.
- South Dakota Content Standards and Core Content Connectors.
- Wyoming Extended Standards and Instructional Achievement-Level Descriptors.

These content 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.

3.4 Item Development Plans

Once all individual state standards were aligned across all the states, item development plans (IDPs) were created for each state. These IDPs were based on identified areas where additional items were needed to ensure that all the MOU-Alt standards aligned on the crosswalk were addressed in the item-sharing pool. Items for each state-specific standard that were not aligned to the MOU-Alt crosswalk standards were created to meet the state's test blueprint, if the state decided to create additional items for their own state. These IDPs 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 IDP is created. As new states joined the MOU-Alt agreement, or when states changed their standards, the individual state standards were added to the crosswalk so that items from the state could be aligned across all the states.

IDP creation in school year (SY) 2023–2024 started with CAI content staff completing a pool analysis for Hawai`i and three other MOU-Alt member states for ELA and mathematics, and five other MOU-Alt states for science. CAI then added the results to a combined MOU-Alt crosswalk document. From here, CAI identified any essence statements for which Hawai`i 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 would affect all states' banks. For example, if two items were added to a particular grade 3 ELA essence statement to be developed in Hawai`i, the crosswalk indicated to which other states those two items could be shared to and aligned. Likewise, if another MOU-Alt member state had two items placed on their IDP at a grade 6 mathematics standard, the IDP indicated if those items could be shared and aligned to Hawai`i. The completed crosswalk document clearly showed the number of items to be developed for Hawai`i and contributed to the MOU by HIDOE, as well as the items to be developed for the MOU member states that will align to Hawai`i essence statements.

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

Once created, a senior-level CAI content team member reviewed the IDP. Once the IDP was approved by Senior Content, it was then sent to HIDOE for review and approval. If HIDOE requested changes, CAI content staff reviewed the plan, talked with HIDOE as necessary, and modified the IDP. The IDP was again reviewed by Senior Content staff and sent to HIDOE for final approval.

CAI used the IDP to author new items for initial batch delivery to the client. After newly written items passed the required seven stages of CAI internal reviews, which are 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.

3.5 Development of Item Specifications

The development of item specifications was informed by the crosswalk of state alternate content standards. The item specifications are for the MOU, not for individual states. For each common standard in the crosswalk, CAI examined the states' content extensions and Performance-Level Descriptor (PLD) or Achievement-Level Descriptor (ALD) documents to identify which extensions were aligned to that common standard. Each item specification included the General Education standard, followed by the state-specific alternate content standards that aligned with the General Education standard. The item specifications also included complexity statements and task demands. The language of the complexity statements and task demands were derived from each state's content standards, where applicable, and synthesized to drive items aligned to multiple states. Once completed, the item specifications were sent to each state for review to confirm alignment and overall approach.

The states' content extensions and PLDs or ALDs were further analyzed to cull relevant concepts, skills, and vocabulary. Based on MOU state 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 or ALDs were also analyzed to reveal state-specific and cross-state content limits in the content extensions. These were listed in the Content Limits section.

The analyses outlined were 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 or ALDs. 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.

3.6 Item Development Process

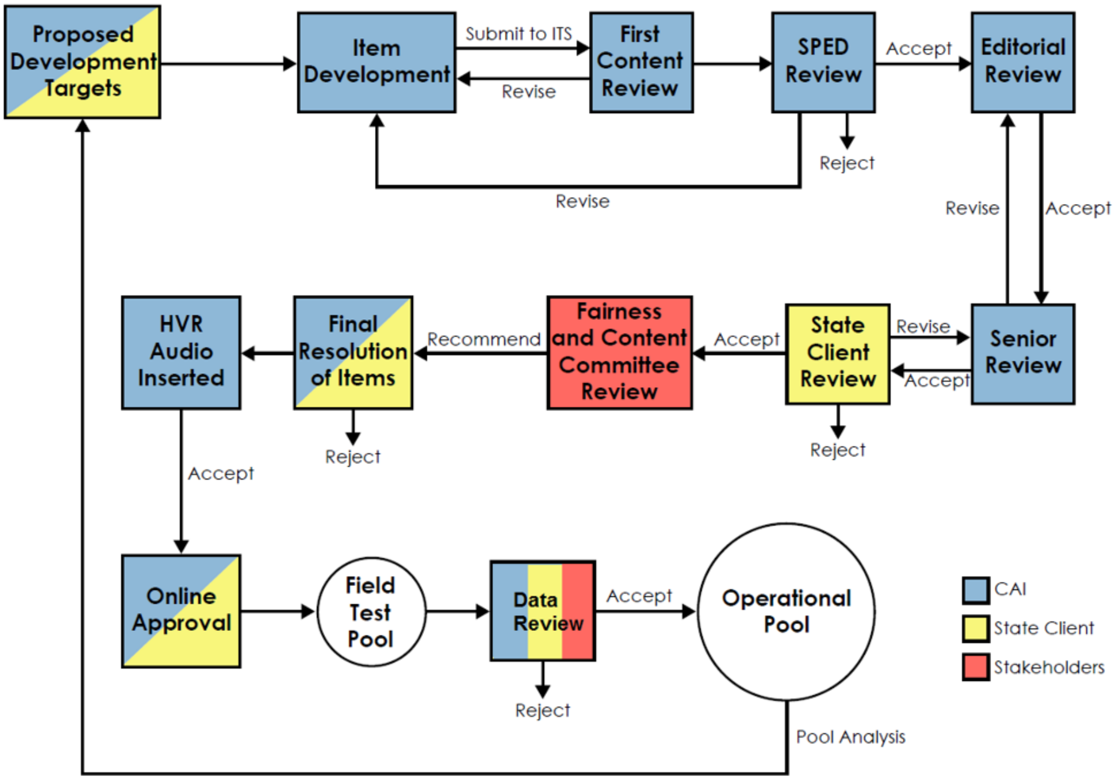
Items are developed by each of the states that joined the shared item development agreement. In each state, item development for each year begins in the spring. Prior to item development, item writers are trained on aspects of items that are unique to students with significant cognitive disabilities. Items are written by professional item writers with a background in education and expertise in the assigned content area and alternate assessments. A group of senior test-development specialists monitor and support the item development activities.

Items are written by CAI content staff or by third party item development vendors, in compliance with the item specifications and style guide documents to ensure items meet the expected alignment, complexity, and style criteria. The item specifications and style guide documents are created by CAI, and reviewed and approved by the department of education in each individual state. The item specifications are for the MOU, instead of for individual states. If a particular standard is only under one state, that standard is not included in the MOU item specifications. Rather, the state creates separate field-test slots for items associated with state-specific standards.

Item development 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 items are initially developed, they undergo a group review that includes content and senior reviewers, followed by an individual content review, where edits are made based on group reviews, and then a special education review. After the items are reviewed by the special education reviewer, they go through an editorial review. After editorial review, the items go back through a senior review, which is the last review step at CAI before the items are sent to each state for client review. At this step, the client may accept, recommend edits, or reject the items.

After the client comments are resolved, all accepted items are then submitted to a Content and Fairness Committee for review. At the same time the Content and Fairness Committee reviews the items, the other members of the MOU-Alt also review the items and provide feedback. After the Content and Fairness Committee makes its recommendations, the state and CAI convene a resolution meeting at which all of the comments from the Content and Fairness Committee and the other MOU-Alt states are reviewed. The state approves final edits to the items based on the Content and Fairness Committee and other state comments. The items then go through a final edit resolution. Lastly, CAI verifies that the items will appear on the test as expected through the platform review process. Figure 1 shows the full item development process.

Figure 1. Alternate Assessment Item Development Process



CAI Review

Items are reviewed at CAI at the following 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. Reviewers use the item specifications and a style guide to make sure the items fit all guidelines.

- CAI Internal Preliminary Review: Following group review, a preliminary review is conducted by a member of CAI's content team assigned to the Alternate Assessment. Items are revised to eliminate initial errors, meet content standards, and satisfy internal style and clarity expectations, as agreed on in the group review.

- CAI Internal Content Review: A second content review occurs after the preliminary review to further ensure changes based on the group review are implemented, and to revise items to address any errors and issues on content, alignment, clarity, and accessibility.

- 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 broad 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.

- Editorial 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, are free of typographical and technical errors (e.g., key check, spelling error check), and 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 compared to the blueprint and reviewed against the Essence Statements and the PLDs at all difficulty levels. 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 Hawai`i's contribution to shared MOU field testing in spring 2024 were developed during spring and summer 2023. During the state review process in this development cycle, all items were accepted by HIDOE assessment staff.

The Hawai`i-owned (Segment 3) items field-tested in spring 2024 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 science item.

Content and Fairness Committee Review

In each state, items owned and accepted by the state are prepared for review by a statewide Content and Fairness Committee convened for each content area in each state. The Content and Fairness Committee comprises 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. The review committee represents a diversity of gender, ethnicity, disability, race, and cultural subgroups across the state.

Table 1 presents a summary of the demographics of the committee members in Hawai`i who participated in the item content and fairness review process in summer 2023 for the spring 2024 field-test items.

Table 1. Content and Fairness Item Review Committee Participants

Subject Area Committee	**Participant Characteristics**	**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
	Black	1		
	Caucasian/White	1	5	2
	Chinese	1	1	
	Hawaiian		1	
	Hispanic		1	
	Indian			1
	Japanese	2		
	Middle Eastern			
	Multiracial (didn't specify)	1	1	
	Native American		1	
	Pacific Islander		1	1
	Portuguese	1		
Special Education	SPED Teacher	1	3	1
	Gen Ed Teacher	5	4	1
	Higher Education		2	2
	Other		1	0
Grade Level Taught	Elementary	2	4	
	Middle School	1	2	1
	High School	3	1	1
	College		2	2
	NA		1	
Parent of HI Student	Yes, currently	2	2	1
	Yes, previously	2	3	1
	No	2	5	2

Following revisions and state approval, items are brought to the Content and Fairness Committee for further review. At the beginning of each Content and Fairness Committee review 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 view the items in the same way that the student will view them.

The stakeholders in the committee review items and provide feedback to ensure that all accepted items are correct, meet bias and sensitivity guidelines, align to content standards, and abide by the principles of universal design. 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.

The common criteria used for item review are:

- Content accuracy and clarity
- Alignment to the content specifications
- Correct answer key and appropriate distractor(s) for each 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

3.7 FIELD TESTING

After going through various stages of reviews, items are moved into the field-test item pool, to be field tested in the following spring during the operational testing window. For example, the items developed in 2022-23 were field-tested in spring 2024; the items developed during the academic year of 2023-24 will be field-tested in spring 2025. These field-test items are embedded among operational items in the CATs. 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 and CATs 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.

After the operational test administration, CAI psychometricians perform both classical item analysis and Item Response Theory (IRT) analysis for all field-test items. Items are flagged based on predetermined statistical criteria. Details of the psychometric analysis and flagging rules on field-test items are presented in Section 4.

3.8 POST FIELD-TESTING ITEM DATA REVIEW

Following the psychometric analysis, items are categorized into four groups for further action:

- Items with a sample size of fewer than 50 are archived for future re-field-testing.
- Items with negative biserial/polyserial correlations are rejected after an additional key verification from CAI.
- Items not flagged by the statistics are reviewed through a Roman Voting process by the educator committees in the states.
- Items flagged by the statistics undergo an item data/content review (IDCR) process.

Roman Voting

Roman Voting is a new process implemented starting from the spring 2024 administration. In prior years, items not flagged by the statistics were automatically moved to the operational item pool without further reviewing.

The purpose of the Roman Voting process is to provide states and their educators with an additional opportunity to review items before they are used in future operational administrations. This process is carried out independently within each state. In Hawai`i, the Content and Fairness Committees first vote on whether to move items into the operational item pool. If the committees vote Yes, the items are added to the operational item pool without future review. If the committees vote No, they discuss as a group, and detailed notes on the reasoning behind the No vote are recorded and shared with HIDOE, who then makes the final decision on whether to include them in the operational item pool.

The results for the 2024 Roman Voting items are shown in the table below.

Table 2. Roman Voting Summary

Subject	Grade	# of Items Reviewed during RV	# of Items Discussed for Rejection during RV	# of Items Rejected by HIDOE After RV	# of Items Accepted during RV
ELA	3	20	0	0	20
	4	20	0	0	20
	5	24	0	0	24
	6	24	0	0	24
	7	25	3	3	22
	8	28	0	0	28
	11	21	0	0	21
Mathematics	3	11	0	0	11
	4	14	2	2	12
	5	14	2	0	14
	6	9	1	0	9
	7	7	0	0	7
	8	6	0	0	6
	11	2	0	0	2
Science	5	22	0	0	22
	8	16	0	0	16
	11	6	0	0	6
Total		**269**	**8**	**5**	**264**

Item Data/Content Review

Items flagged by the statistics are reviewed in IDCR meetings involving all MOU states. The MOU-Alt data review committee consists of staff across MOU states, CAI content specialists, special education specialists, and psychometricians. Before IDCR, CAI psychometricians provide a training to reviewers on how flagged statistics can be used to identify potential content flaws in items. During IDCR, the committee evaluates whether flagged items contain features that might result in undesirable statistics. They then decide whether to reject the item completely, accept it with modifications for further field testing, or accept it as is. Additionally, content experts from each state ensure that items from other states are only included if they align with the state's standards.

The IDCR process has two phases.

1. **Individual State Review:** In this initial phase, state staff or educators from each state independently review the items and decide whether to accept or reject them. After all states complete their reviews, the decisions are summarized into four categories:

   - Items that are accepted by all states.
   - Items that are rejected by all states.
   - Items that are rejected by the source state but accepted by at least one destination state.
   - Items that are accepted by the source state but rejected by at least one destination state.

   Items in the first category are added to the item pools of all states, while those in the second category are rejected from all state item pools. Items in the third or fourth categories, where there is disagreement among states, proceed to the second phase: group review.

2. **Group Review:** In this phase, all states participate in group IDCR meetings where they share their rationales for their decisions. After discussing and considering the perspectives of other states, states have the opportunity to revise their initial decisions from the individual state review.

Upon completion of the Roman Voting and IDCR process, all field-test items accepted by each state will be added to their operational item pool, ready for administration in the following year. Item data review results in the spring 2024 administration are presented in Section 4.

4. SUMMARY OF FIELD-TEST ITEM ANALYSIS IN SPRING 2024

The HSA-Alt spring 2024 field-test item pool included both MOU items, which are available to use by all member states, and Hawai`i-specific items that align only with the Hawaii alternate assessment content standards. Table 3 provides a summary of the number of MOU items by source state, the total number of MOU items, the number of items administered in Hawai`i, which includes MOU items and the number of Hawai`i-specific items. Out of a total of 778 MOU items, HIDOE approved 728 for field-testing. Additionally, Hawai`i field-tested 14 state-specific items.

Table 3. Number of Field-Test Items in Spring 2024

Subject	Grade	[1]MOU Items by Source State						[1]MOU Total	Items Administered in Hawai`i		
		HI	ID	MT	SC	SD	WY		MOU Items	Hawai`i Only	Total
ELA	3	5	8		32			45	43		43
	4	4	7		33		2	46	46		46
	5	5	7		34		4	50	48		48
	6	5	7		33		5	50	49		49
	7	5	7		33			45	45		45
	8	4	7		33		3	47	47	1	48
	11	4	7		34			45	45		45
Mathematics	3	7	6		10		2	25	25		25
	4	4	12		30		2	48	44		44
	5	3	6		41		2	52	49		49
	6	3	6		37		1	47	47		47
	7	7	6		21		3	37	35		35
	8	3	10		28		3	44	44		44
	11	6	4				2	12	12	13	25
Science	5	9	9	3	43	3	3	70	59		59
	8	17	6	3	42	3	5	76	68		68
	11	5	6	2	19	4	3	39	22		22
Total		**96**	**121**	**8**	**503**	**10**	**40**	**778**	**728**	**14**	**742**

[1]Number of items available for sharing across all MOU states.

4.1 FIELD-TEST ITEM ANALYSIS

After the close of the spring testing window, CAI psychometrics staff analyzed field-test data based on combined data from all MOU states, to prepare for item data review meetings and to promote of high-quality test items to operational item pools. Analysis of field-test items included the following:

- **Classical item analysis**, used to evaluate the relationship of each item to the overall scale and assess the quality of the distractors.
- **Item response theory (IRT) analysis**, used to assess how well items fit the measurement model and provide the statistical foundation for constructing operational forms, test scoring and reporting.
- **Differential item functioning (DIF) analysis**, used to identify items that may exhibit bias across subgroups.

4.1.1. Classical Item Analysis

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

Items that are extremely difficult or 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 polytomous items with 0–2 score points, item difficulty is calculated both as the item's mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of an item's mean score divided by the max score point possible). Items are flagged for review if the *p*-value or average proportion correct is less than one divided by the number of response options or greater than 0.95.

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 could differentiate between high- and low-achieving students. The discrimination index is calculated as the correlation between the item score and the student's IRT-based ability estimate. Items are flagged for subsequent reviews if the correlation for the keyed (correct) response is less than 0.20. For polytomous items, the mean total number correct score is computed for students scored within each possible score category; items are flagged for review if the mean total score for a lower score point is greater than the mean total score for a higher score point.

Distractor analysis for the dichotomously scored multiple-choice items is used to identify items with 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-based ability estimate, restricting the analysis to those students selecting either the target distractor or the keyed response. Items are flagged for subsequent reviews if the biserial correlation for the distractor response is greater than 0.05.

The flagging criteria based on classical item analysis are summarized in Table 4.

Table 4. Flagging Criteria Based on Classical Item Analysis

Analysis Type	Flagging Criteria
Item Difficulty	*p*-value (for dichotomous items) or average proportion correct (for polytomous items) is < 1/number of response options or > 0.95.
Item Discrimination	Biserial or polyserial correlation for the correct response is < 0.20.
Mean Score for Two-Point Items	Mean total score for a lower score point > Mean total score for a higher score point.
Distractor Analysis	Biserial correlation for any distractor response is > 0.05.

4.1.2. IRT Analysis

The Item Response Model

Traditional item response models assume a single underlying trait and 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 of

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

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 item's difficulty. 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 are calibrated using the partial credit model (PCM; Masters, 1982). Under Masters' partial credit model, the probability of getting a score of x_i on item i given ability θ can be written as

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

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

Item Calibration

Calibration is the process by which we estimate the statistical relationship between item responses and the underlying trait being measured. WINSTEPS is used to estimate the Rasch and Masters' partial credit model item parameters for the MOU-Alt. Winsteps, provided by Mesa Press, is publicly available software that utilizes a joint maximum likelihood estimation (JMLE) approach. This method simultaneously estimates both person and item parameters.

The Winsteps output, which includes item statistics, are reviewed. Item fit is evaluated via the mean square Infit and mean square Outfit statistics, which are based on weighted and unweighted standardized residuals for each item response. These residual statistics reflect the discrepancy between the observed item scores and predicted item scores according to the IRT model. The expected value for both fit statistics is 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.

Embedding randomly selected field-testing items among operational items in CATs results in a sparse data matrix. In this matrix, both operational and field-test items are calibrated concurrently for each grade and subject, with the parameter estimates of the operational items fixed. The operational items were previously calibrated and scaled to the existing MOU-Alt scale during the years they were used as field-test items. Consequently, the field-test item parameter estimates are also on the MOU-Alt scale. Completed records from all MOU states are included in the IRT analysis, with items not presented being treated as not administered.

4.1.3. DIF Analysis

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 are evaluated for DIF, and all items exhibiting DIF are flagged for further examination by CAI and the MOU states.

CAI conducts DIF analysis on all field-tested items to detect potential item bias among the following group comparisons:

- Female vs. Male
- African American vs. White
- Hispanic or Latino vs. White
- Severe and Moderate Intellectual Disability vs. Other
    - Severe and moderate intellectual disability is defined by each state based on their primary disability code.
    - For Hawai`i, the following disability categories are classified as severe/moderate disability: (1) Multiple Disabilities, (2) Intellectual Disability, and (3) 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 10 intervals to compute the generalized Mantel–Haenszel chi-square (*GMH* χ^2) DIF statistics. For dichotomous items, the analysis program computes the *GMH* χ^2 DIF statistic, the log-odds ratio, and the *MH*-delta (Δ_{MH}); for the polytomous items, the program computes the *GMH* χ^2 DIF statistic, the item score standard deviation (σ), and the standardized mean difference (*SMD*).

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 5. Items are also categorized as positive DIF (+A, +B, or +C), signifying that the item favors the focal group (e.g., African American/Black, Hispanic, female), or negative DIF (–A, –B, or –C), signifying that the item favors the reference group (e.g., White, male).

Table 5. DIF Classification Rules

Dichotomous Items	
Category	**Rule**
C	GMH_{X^2} is significant at .05 and $\lvert \Delta_{MH} \rvert$ >1.5
B	GMH_{X^2} is significant at .05 and $1 < \lvert \Delta_{MH} \rvert \leq 1.5$
A	GMH_{X^2} is not significant at .05 or $\lvert \Delta_{MH} \rvert \leq 1$
Polytomous Items	
Category	**Rule**
C	GMH_{X^2} is significant at .05 and $\lvert SMD \rvert / \; SD \; > .25$
B	GMH_{X^2} is significant at .05 and $.17 < \; \lvert SMD \rvert / \; SD \; \leq .25$
A	GMH_{X^2} is not significant at .05 or $\lvert SMD \rvert / \; SD \; \leq \; .17$

Items are flagged if their DIF statistics fall into the "C" category for any group and the sample size for both focal and reference groups are larger than or equal to 50. 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. 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).

All items flagged due to DIF are reviewed during the IDCR process. Reviewers are instructed to examine whether there are any content reasons that may have led to the item being flagged. Items that are determined to be biased are rejected and not included in the state's operational item pool.

4.2 Results of the Spring 2024 Field-Test Item Analysis

This section presents results from the classical item analysis, IRT analysis, and DIF analysis of items field tested in Hawai`i in spring 2024. Table 6 presents the average sample size and the sample size at various percentiles for the analysis. Table 7–Table 9 provide summaries of item statistics for ELA, mathematics, and science, respectively. For each item statistic (e.g., *p*-values), the percentiles are computed across items administered in Hawai`i in the corresponding subject and grade.

Table 10–Table 12 show *p*-value distributions by item type and number of response options in each grade for ELA, mathematics, and science, respectively. Table 13 provides the DIF analysis summary.

Table 6. Sample Size Distribution

Subject	Grade	# of Items	Average Sample Size	Sample Size in Percentiles								
				Min	5th	10th	25th	50th	75th	90th	95th	Max
ELA	3	43	167	133	133	150	161	165	175	189	191	192
	4	46	146	117	119	119	132	148	158	173	183	184
	5	48	146	120	124	126	137	147	157	161	163	167
	6	49	143	111	121	122	130	145	155	165	167	169
	7	45	154	114	126	134	143	156	166	171	175	178
	8	48	143	96	119	121	128	144	152	168	177	180
	11	45	160	127	137	141	151	160	170	174	188	195
	Overall	324	151	96	121	125	138	152	164	173	180	195
Mathematics	3	25	300	246	276	276	293	301	313	320	326	327
	4	44	140	120	121	125	133	141	148	154	156	161
	5	49	141	109	117	119	133	140	151	161	164	168
	6	47	152	131	136	139	145	152	161	165	168	173
	7	35	189	163	167	170	180	189	200	205	210	213
	8	44	154	125	127	135	147	155	164	168	174	179
	11	25	374	118	118	118	118	118	698	711	714	720
	Overall	269	188	109	118	127	140	153	176	298	326	720
Science	5	59	265	106	158	251	260	269	281	293	295	298
	8	68	241	100	211	225	233	242	252	264	266	273
	11	22	256	142	217	229	246	269	277	281	286	290
	Overall	149	252	100	211	227	240	256	271	282	290	298

Table 7. Summary of Item Analysis for ELA

Grade	# of Items	Statistics	Min	P10	P25	P50	P75	P90	Max
3	43	*p*-value	0.26	0.35	0.41	0.48	0.57	0.63	0.66
		Biserial/Polyserial	-0.17	0.01	0.10	0.25	0.38	0.48	0.70
		Step Difficulty	-1.13	-0.95	-0.74	-0.42	0.06	0.29	0.78
		Infit	0.84	0.91	0.96	1.02	1.09	1.18	1.22
		Outfit	0.81	0.89	0.94	1.02	1.13	1.19	1.32
4	46	*p*-value	0.25	0.32	0.40	0.48	0.55	0.61	0.64
		Biserial/Polyserial	-0.06	0.02	0.16	0.26	0.38	0.45	0.53
		Step Difficulty	-0.97	-0.92	-0.63	-0.24	0.04	0.39	0.79
		Infit	0.92	0.93	0.99	1.03	1.08	1.14	1.21
		Outfit	0.89	0.93	0.96	1.05	1.12	1.18	1.26
5	48	*p*-value	0.24	0.36	0.47	0.53	0.60	0.65	0.78
		Biserial/Polyserial	-0.04	0.11	0.24	0.31	0.42	0.55	0.65
		Step Difficulty	-1.72	-1.07	-0.88	-0.58	-0.32	0.24	0.77
		Infit	0.85	0.92	0.97	1.01	1.06	1.14	1.22
		Outfit	0.84	0.88	0.94	1.01	1.07	1.16	1.31
6	49	*p*-value	0.24	0.33	0.39	0.46	0.55	0.62	0.66
		Biserial/Polyserial	-0.13	0.07	0.19	0.31	0.42	0.52	0.72
		Step Difficulty	-1.00	-0.92	-0.57	-0.26	0.17	0.49	1.03
		Infit	0.76	0.89	0.97	1.03	1.08	1.19	1.31
		Outfit	0.73	0.85	0.97	1.02	1.11	1.19	2.73
7	45	*p*-value	0.29	0.31	0.40	0.50	0.55	0.61	0.68
		Biserial/Polyserial	0.05	0.09	0.23	0.32	0.41	0.52	0.71
		Step Difficulty	-1.23	-0.76	-0.54	-0.23	0.24	0.57	0.76
		Infit	0.85	0.91	0.97	1.03	1.10	1.13	1.18
		Outfit	0.80	0.88	0.96	1.03	1.11	1.17	1.22
8	48	*p*-value	0.24	0.32	0.42	0.51	0.59	0.68	0.73
		Biserial/Polyserial	-0.12	0.02	0.17	0.33	0.41	0.55	0.67
		Step Difficulty	-1.42	-1.12	-0.85	-0.45	-0.09	0.58	0.72
		Infit	0.85	0.87	0.96	1.01	1.11	1.17	1.28
		Outfit	0.76	0.83	0.93	1.01	1.13	1.22	1.35
11	45	*p*-value	0.25	0.38	0.44	0.48	0.57	0.69	0.74
		Biserial/Polyserial	-0.09	0.20	0.27	0.36	0.50	0.62	0.71
		Step Difficulty	-1.48	-1.21	-0.64	-0.19	-0.03	0.18	1.01
		Infit	0.82	0.87	0.95	1.03	1.08	1.14	1.39
		Outfit	0.78	0.81	0.94	1.01	1.09	1.22	1.50

Table 8. Summary of Item Analysis for Mathematics

Grade	# of Items	Statistics	Min	P10	P25	P50	P75	P90	Max
3	25	*p*-value	0.24	0.27	0.35	0.44	0.59	0.63	0.67
		Biserial/Polyserial	0.01	0.10	0.19	0.27	0.42	0.49	0.56
		Step Difficulty	-1.30	-1.12	-0.92	-0.27	0.12	0.56	0.76
		Infit	0.91	0.93	0.96	1.02	1.05	1.13	1.16
		Outfit	0.85	0.89	0.94	1.03	1.08	1.17	1.29
4	44	*p*-value	0.24	0.28	0.33	0.41	0.49	0.63	0.69
		Biserial/Polyserial	-0.18	-0.03	0.11	0.18	0.35	0.51	0.71
		Step Difficulty	-1.48	-1.13	-0.56	-0.18	0.17	0.44	0.68
		Infit	0.83	0.89	0.96	1.04	1.10	1.16	1.24
		Outfit	0.79	0.87	0.95	1.05	1.11	1.27	1.43
5	49	*p*-value	0.26	0.29	0.33	0.40	0.49	0.58	0.64
		Biserial/Polyserial	-0.10	0.00	0.07	0.22	0.36	0.58	0.74
		Step Difficulty	-1.14	-0.82	-0.42	-0.11	0.24	0.42	0.66
		Infit	0.85	0.90	0.96	1.02	1.09	1.13	1.22
		Outfit	0.83	0.87	0.97	1.06	1.13	1.18	1.27
6	47	*p*-value	0.22	0.29	0.32	0.39	0.50	0.55	0.75
		Biserial/Polyserial	-0.29	-0.10	0.01	0.14	0.29	0.43	0.62
		Step Difficulty	-1.83	-1.03	-0.63	-0.17	0.13	0.35	0.65
		Infit	0.89	0.93	0.99	1.06	1.12	1.16	1.28
		Outfit	0.80	0.93	0.99	1.07	1.18	1.24	1.69
7	35	*p*-value	0.21	0.26	0.33	0.43	0.53	0.59	0.63
		Biserial/Polyserial	-0.37	-0.08	0.04	0.15	0.25	0.30	0.35
		Step Difficulty	-1.29	-1.04	-0.75	-0.42	0.13	0.50	0.76
		Infit	0.95	0.97	0.99	1.04	1.09	1.13	1.25
		Outfit	0.92	0.96	0.98	1.06	1.12	1.20	1.56
8	44	*p*-value	0.21	0.25	0.31	0.37	0.44	0.50	0.68
		Biserial/Polyserial	-0.28	-0.16	-0.06	0.05	0.20	0.36	0.55
		Step Difficulty	-1.59	-0.67	-0.44	-0.19	0.17	0.49	0.74
		Infit	0.88	0.95	0.98	1.06	1.09	1.13	1.16
		Outfit	0.84	0.93	1.01	1.07	1.13	1.18	1.22
11	25	*p*-value	0.28	0.30	0.34	0.36	0.46	0.53	0.85
		Biserial/Polyserial	-0.35	-0.18	-0.11	0.15	0.33	0.40	0.52
		Step Difficulty	-2.31	-0.75	-0.45	0.04	0.16	0.37	0.45
		Infit	0.90	0.93	0.97	1.04	1.07	1.14	1.19
		Outfit	0.88	0.92	0.95	1.05	1.09	1.21	1.35

Table 9. Summary of Item Analysis for Science

Grade	# of Items	Statistics	Min	P10	P25	P50	P75	P90	Max
5	59	*p*-value	0.18	0.30	0.35	0.43	0.53	0.61	0.74
		Biserial/Polyserial	-0.13	0.07	0.13	0.26	0.45	0.52	0.69
		Step Difficulty	-1.81	-0.99	-0.61	-0.19	0.21	0.49	1.10
		Infit	0.79	0.91	0.96	1.05	1.10	1.14	1.25
		Outfit	0.75	0.88	0.95	1.03	1.13	1.19	1.38
8	68	*p*-value	0.17	0.27	0.32	0.39	0.50	0.58	0.72
		Biserial/Polyserial	-0.10	0.03	0.08	0.17	0.28	0.40	0.49
		Step Difficulty	-1.43	-0.79	-0.41	0.04	0.37	0.67	1.28
		Infit	0.88	0.93	0.98	1.03	1.08	1.10	1.25
		Outfit	0.81	0.90	0.97	1.03	1.08	1.16	1.30
11	22	*p*-value	0.30	0.35	0.40	0.47	0.59	0.61	0.62
		Biserial/Polyserial	0.02	0.05	0.10	0.33	0.47	0.52	0.57
		Step Difficulty	-0.99	-0.93	-0.80	-0.23	0.05	0.30	0.48
		Infit	0.86	0.90	0.93	1.00	1.10	1.16	1.17
		Outfit	0.84	0.87	0.92	0.99	1.16	1.19	1.26

Table 10. *p*-value by Item Type and Number of Response Options for ELA

Grade	Item Type	Number of Response Options	# of Items	Percentage	Min	P10	P25	P50	P75	P90	Max
3	multipleChoice	2	20	46.50%	0.41	0.45	0.50	0.55	0.62	0.65	0.66
	multipleChoice	3	23	53.50%	0.26	0.33	0.36	0.41	0.48	0.57	0.61
		Total	43	100.00%	0.26	0.35	0.41	0.48	0.57	0.63	0.66
4	multipleChoice	2	24	52.20%	0.43	0.45	0.49	0.54	0.60	0.63	0.64
	multipleChoice	3	22	47.80%	0.25	0.28	0.33	0.40	0.44	0.52	0.63
		Total	46	100.00%	0.25	0.32	0.40	0.48	0.55	0.61	0.64
5	multipleChoice	2	31	64.60%	0.45	0.5	0.51	0.59	0.63	0.66	0.78
	multipleChoice	3	17	35.40%	0.24	0.30	0.36	0.43	0.49	0.57	0.59
		Total	48	100.00%	0.24	0.36	0.47	0.53	0.60	0.65	0.78
6	multipleChoice	2	16	32.70%	0.39	0.46	0.53	0.55	0.62	0.65	0.66
	multipleChoice	3	33	67.30%	0.24	0.29	0.35	0.43	0.50	0.56	0.65
		Total	49	100.00%	0.24	0.33	0.39	0.46	0.55	0.62	0.66
7	multipleChoice	2	14	31.10%	0.51	0.52	0.54	0.56	0.63	0.68	0.68
	multipleChoice	3	31	68.90%	0.29	0.31	0.34	0.41	0.51	0.54	0.58
		Total	45	100.00%	0.29	0.31	0.40	0.5	0.55	0.61	0.68
8	multipleChoice	2	22	45.80%	0.47	0.48	0.53	0.59	0.65	0.68	0.73
	multipleChoice	3	26	54.20%	0.24	0.28	0.38	0.42	0.50	0.55	0.58
		Total	48	100.00%	0.24	0.32	0.42	0.51	0.59	0.68	0.73
11	multipleChoice	2	16	35.60%	0.47	0.48	0.53	0.58	0.67	0.70	0.74
	multipleChoice	3	29	64.40%	0.25	0.31	0.42	0.46	0.50	0.59	0.69
		Total	45	100.00%	0.25	0.38	0.44	0.48	0.57	0.69	0.74

Table 11. *p*-value by Item Type and Number of Response Options for Mathematics

Grade	Item Type	Number of Response Options	# of Items	Percentage	Min	P10	P25	P50	P75	P90	Max
3	multipleChoice	2	10	40.00%	0.44	0.48	0.54	0.60	0.62	0.65	0.67
	multipleChoice	3	15	60.00%	0.24	0.27	0.28	0.37	0.42	0.53	0.66
		Total	25	100.00%	0.24	0.27	0.35	0.44	0.59	0.63	0.67
4	multipleChoice	2	11	25.00%	0.44	0.46	0.47	0.58	0.66	0.69	0.69
	multipleChoice	3	33	75.00%	0.24	0.27	0.32	0.36	0.42	0.48	0.65
		Total	44	100.00%	0.24	0.28	0.33	0.41	0.49	0.63	0.69
5	multipleChoice	2	13	26.50%	0.38	0.44	0.45	0.54	0.58	0.61	0.64
	multipleChoice	3	36	73.50%	0.26	0.28	0.31	0.37	0.43	0.52	0.64
		Total	49	100.00%	0.26	0.29	0.33	0.40	0.49	0.58	0.64
6	multipleChoice	2	7	14.90%	0.38	0.38	0.52	0.55	0.62	0.70	0.70
	multipleChoice	3	40	85.10%	0.22	0.28	0.32	0.37	0.41	0.50	0.75
		Total	47	100.00%	0.22	0.29	0.32	0.39	0.50	0.55	0.75
7	multipleChoice	2	16	45.70%	0.40	0.46	0.47	0.54	0.58	0.63	0.63
	multipleChoice	3	19	54.30%	0.21	0.22	0.28	0.34	0.41	0.48	0.50
		Total	35	100.00%	0.21	0.26	0.33	0.43	0.53	0.59	0.63
8	multipleChoice	2	9	20.50%	0.39	0.39	0.41	0.50	0.52	0.68	0.68
	multipleChoice	3	35	79.50%	0.21	0.25	0.28	0.35	0.43	0.45	0.49
		Total	44	100.00%	0.21	0.25	0.31	0.37	0.44	0.50	0.68
11	multipleChoice	2	4	16.00%	0.51	0.51	0.52	0.54	0.7	0.85	0.85
	multipleChoice	3	21	84.00%	0.28	0.3	0.32	0.35	0.41	0.46	0.52
		Total	25	100.00%	0.28	0.3	0.34	0.36	0.46	0.53	0.85

Table 12. *p*-value by Item Type and Number of Response Options for Science

Grade	Item Type	Number of Response Options	# of Items	Percentage	Min	P10	P25	P50	P75	P90	Max
5	multipleChoice	2	15	25.40%	0.41	0.49	0.50	0.59	0.69	0.74	0.74
	multipleChoice	3	44	74.60%	0.18	0.28	0.31	0.40	0.48	0.53	0.57
		Total	59	100.00%	0.18	0.30	0.35	0.43	0.53	0.61	0.74
8	multipleChoice	2	16	23.50%	0.45	0.46	0.49	0.55	0.60	0.71	0.72
	multipleChoice	3	52	76.50%	0.17	0.26	0.30	0.35	0.43	0.49	0.67
		Total	68	100.00%	0.17	0.27	0.32	0.39	0.50	0.58	0.72
11	multipleChoice	2	7	31.80%	0.47	0.47	0.49	0.57	0.61	0.62	0.62
	multipleChoice	3	15	68.20%	0.30	0.33	0.35	0.42	0.52	0.61	0.62
		Total	22	100.00%	0.30	0.35	0.40	0.47	0.59	0.61	0.62

Table 13. 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								**ELA**							
3	24	15	8				1	3	1	1					
4	14	7	7					4	4		3			1	
5	21	7	13			1		5							
6	22	11	10			1		6	3	1	2				
7	27	12	14			1		7							
8	26	15	11					8							
11	37	18	15			2	2	11	4	2	2				
Mathematics								**Mathematics**							
3	25	10	15					3	25	14	10			1	
4	9	3	5			1		4							
5	13	5	8					5							
6	33	13	18			2		6	4		4				
7	34	17	16			1		7							
8	30	15	15					8							
11	12	7	5					11	12	4	6		2		
Science								**Science**							
5	58	22	34			1	1	5	56	20	34			1	1
8	67	34	31			1	1	8	67	31	33			1	2
11	22	11	10	1				11	15	8	7				

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**							
3								3	3	3					
4								4	2	1	1				
5								5	16	9	7				
6								6	18	6	12				
7								7	44	19	24				1
8								8	42	20	21			1	
11								11	38	21	15				2
Mathematics								**Mathematics**							
3	7	5	2					3	25	8	15				2
4								4	2		2				
5								5	11	4	6			1	
6								6	26	10	14			1	1
7								7	35	13	21				1
8								8	42	21	18			2	1
11	12	6	6					11	25	13	11			1	
Science								**Science**							
5								5	54	28	26				
8								8	67	33	28			4	2
11								11	21	10	10				1

Note. This table only includes items with sample size > = 50 in both the focal and reference groups.

4.3 Item Data Review Results

After the psychometric analysis was completed, 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 item data review and were not seen by the item data review committees in each MOU state. Items flagged for undesired statistics were reviewed by the Hawai`i stakeholder IDCR committee. Additionally, the Hawai`i IDCR committee also participated in the Roman Voting process and reviewed all items with desired statistics that were not flagged. The Hawai`i IDCR committee included special education teachers, content-area experts, advocates, and community members who worked with individuals with significant cognitive disabilities. Table 14 presents a summary of the demographics of the committee members who participated in the item data review process for the spring 2024 field-test items in summer 2024.

Table 14. Item Data/Content Review Committee Participants

Subject Area Committee	Participant Characteristics	ELA	Mathematics	Science
Total Participants		**8**	**7**	**6**
Island	Oahu	6	2	4
	Maui	1	1	2
	Hawai'i	1	4	
Gender	Female	6	5	5
	Male	2	2	1
Ethnicity (self-reported categories)	Asian			1
	Black	2		1
	Caucasian/White	1	6	1
	Japanese	1		
	Middle Eastern			1
	Multiracial (didn't specify)	1	1	2
	Native American		1	
	Pacific Islander	1		
	Declined	1		
Special Education	SPED Teacher	7	3	4
	Gen Ed Teacher	1	1	
	Higher Education		1	
	Other		2	2
Grade Level Taught	Elementary	7	4	
	Middle School		3	2
	High School	1		2
	College		1	1
	NA		2	1
Parent of HI Student	Yes, currently	2	2	
	Yes, previously	2	2	1
	No	4	3	5

Table 15 presents a summary of post field-testing item review results for all items field-tested in Hawai`i in spring 2024. Out of the 742 items field-tested, 76 items had negative biserials/polyserials that were rejected without further review. HIDOE and their IDCR committee reviewed the remaining items, rejecting those that did not align with state content standards, were deemed inappropriate for Hawai`i, or had content flaws as indicated by statistical analysis. Ultimately, 599 field-test items passed the review and were added to the HSA-Alt operational item pool. This included an average of 39 items per grade for ELA, 27 for mathematics, and 45 for science.

Table 15. Summary of Post Field-Testing Item Review

Subject	Grade	Total # of Items Administered in Hawai`i	# of Items with n < 50	# of Items with biserial < 0	# of Items Rejected After Review	# of Items Eligible for Operational Use
ELA	3	43	0	4	3	36
	4	46	0	3	8	35
	5	48	0	1	3	44
	6	49	0	3	2	44
	7	45	0	0	6	39
	8	48	0	3	1	44
	11	45	0	1	10	34
Mathematics	3	25	0	0	0	25
	4	44	0	5	4	35
	5	49	0	5	4	40
	6	47	0	10	13	24
	7	35	0	7	3	25
	8	44	0	19	2	23
	11	25	0	8	0	17
Science	5	59	0	3	1	55
	8	68	0	4	3	61
	11	22	0	0	4	18
Total		**742**	**0**	**76**	**67**	**599**

5. TEST ADMINISTRATION

The spring 2024 testing window was open February 20–May 30, 2024 for online adaptive operational tests, and from February 20–May 23, 2024 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 comprised 40 operational items selected based on item difficulty and student ability to meet the assessment blueprint, with up to 20 embedded field-test (EFT) 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.

5.1 TEST ADMINISTRATOR TRAINING

TA 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 meaningfully compared. HIDOE requires HSA-Alt TAs to attend a yearly department-led TA training to ensure compliance with testing policies. Following the department-led training, all HSA-Alt TAs are also required to complete the online TA Certification Course (available via the CAI portal) before the online TDS allows the TA access to the TA Live Site to administer a test to students.

In January 2024, a series of in-person training sessions for the HSA-Alt 2023–2024 administration occurred at seven locations across the state. These sessions included training on the following topics:

- HSA-Alt Participation Guidelines
- HSA-Alt Essence Statements and Range PLDs
- HSA-Alt Test Design
    - ESR
- 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 (LCI)
    - 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 used this feedback to revise training materials; revise time allocation for the training, mode(s) of training to be used in the future; and identify areas where additional support for TAs needed 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.

As a final step, the online HSA-Alt TA Certification Course reviewed the information provided during the yearly mandatory in-person or virtual training and required TAs to affirm that they would uphold the HSA-Alt Code of Ethics. The specific responsibilities delineated in the HSA-Alt Code of Ethics are illustrated as follows.

- 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 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 following:
  - The full audio delivery of stimulus, question, and answer options.
  - The expanded view of mathematics and science visuals.
  - 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.

5.2 Test Administration Manuals

The *2024 Test Administration Manual* (TAM) summarizes the HSA-Alt and provides guidelines for test administration. It includes the following topics:

- 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 on https://hsa-alt.alohahsap.org/resources#refine=2024-2025.

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 on the state portal.

A short guide for the use of printed materials 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 on the state portal.

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.

5.3 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 Section 5.3.1, Allowable Accommodations, provide students the opportunity to gain access to the items and make a response.

5.3.1. 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. A list of universal tools that are available is provided in Table 16. 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 in the following table 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 provided during the alternate assessment 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 those students on the testing device. 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 16. 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.
Audio Playback (summative assessments)	Text on summative assessment items is read aloud to the student via embedded audio files that include 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 TDS 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 (refer to the Separate Setting accommodation).
Breaks	Breaks may be given as often as necessary at the discretion of the TA 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.
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 in full screen. It is one of only three universal tools that can be set in the Test Information Distribution Engine (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 includes: 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.

Universal Tools	Description
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.
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. This requires the Scribe Accommodation.
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 re-delivered. HIDOE HSA-Alt testing policies require students and TAs to move on to the next item once the “Next” button is pressed. Students and TAs 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 TAs 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 Section 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.

Universal Tools	Description
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.
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 2024 HSA-Alt administration, there is one designated support, Translated Test, available for the HSA-Alt. 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 17 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, refer to Table 17. Please note that the Translated Test designated support also requires the submittal and approval of the paper-pencil accommodation for a student.

Table 17. List of Available Designated Supports

Designated Supports	**Description**
Translated Test	This is a linguistic support that is available for students with limited English language skills and who use dual-language supports in the classroom. Dual-language translation provides the full translation for mathematics and science assessments. The Translated Test accommodation is not provided for the ELA test. 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. Translators 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*. **The Paper-Pencil Test Accommodation (fixed form) is also required for the administration of a translated test.**

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 TA 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 addresses 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 2024 are listed in Table 18.

Table 18. 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 an orally provided answer on the first item and then switch response on the next).
American Sign Language (non-embedded)	Test items are translated into American Sign Language (ASL). Some students who are deaf or hard of hearing and who use ASL may need this accommodation. TAs must precisely follow the audio script that is provided for the test item component, including passage, stimulus, question, and answer option card descriptions to translate using ASL. The translator should translate all the words on the test without adding more information or explanation than provided in the item.
Calculator (hand-held)	Students who use a calculator during instruction may use the calculator during the administration of the assessment.
Concrete materials	Students are provided with the customary concrete materials that are used for daily mathematics 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-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.
Digital Mathematics Manipulatives	Students are provided access to the virtual platform with digital mathematics manipulatives, such as unifix cubes, ten frames, fraction tiles, and number lines to use during the mathematics assessment. Teachers may support in selecting the mathematics manipulative that the student selects for a presented problem. Teachers may not manipulate the digital mathematics manipulatives for a student.
Multiplication Table	Students who need a multiplication table to solve mathematics problems and who consistently use the table during instruction and assessment of mathematics may use a multiplication table on the assessment.

Accommodation	Description
Paper Response Card (summative assessments)	Students select the answer using Paper Response Cards that are identical to the options presented in the online testing system. Then, the TA enters the students' responses into the online testing system. Some students with disabilities, such as visual impairment or blindness, are advised to use Paper Response Card Accommodation. The Paper Response Card Accommodation allows the teacher or TA to prepare tactilely enhanced versions of the test visuals and answer options. Students can be provided with tactilely enhanced visuals, 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 (e.g., slippery, fuzzy, rough) If a student’s IEP team determines that a student needs Paper Response Cards to access the assessment due to his or her specific needs, the *Paper Response Card/Paper and Pencil Test Accommodation Request Form* needs to be submitted for verification and approval. Students using the Paper Response Card Accommodation will take the fixed-form test.
Paper-Pencil Test (summative assessments)	The Paper-Pencil Test accommodation provides printed test item booklets for students who cannot access the assessment through the online testing system due to their sensitivity to electronic devices. Students will indicate their answers on the paper test booklet provided. TAs should read aloud provided scripts for all components of the assessment, and enter the student’s answers into the online testing system. The Paper-Pencil Test accommodation is for only a small number of students who are not able to interact with the computer because of their disabilities as indicated in their IEP. The *Paper Response Card/Paper and Pencil Test Accommodation Request Form* needs to be submitted to the Assessment Section for verification. The Paper-Pencil Test accommodation is recommended for alternate-identified English learner (EL) students who need the Translated Test Designated support. This allows the test translator to preview and prepare full translations of the mathematics and science assessments prior to test administration. Students using the Paper-Pencil Test accommodation will take the fixed-form test.

Accommodation	Description
Read Aloud (summative assessments)	The item is read aloud to the student by a trained and qualified human reader. The Read Aloud accommodation may be needed during the Summative assessment for students who are not able to follow embedded HVR in the online testing system and requires a slower audio delivery speed than is currently available via the online platform. The TA should first play the audio. If this accommodation is provided to a student, the in-test audio must first be played for the student through the TDS and carefully reread with the TA carefully listening 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. The TA must not add more information or explanation or 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*. After reading these guidelines, TAs will need to complete and sign the *HSA-Alt Test Security and Confidentiality Form*. Once completed, this form should be given to the school's test coordinator, 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 student's 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, not for correctly answering items.
Scribe	Students either indicate their response or do not respond to a test item, and the TA then enters a **[No Response]** or the student's indicated response into the Data Entry Interface (DEI). 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 TA can enter the **[No Response]** option 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.

Table 19–Table 21 present the number of students who were assigned specific accommodations.

Table 19. Total Number of Students with Allowed Accommodations: ELA

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Alternate Response Options	3	4	6	2	2	5	2
American Sign Language (Non-Embedded)	0	0	1	0	0	1	0
Concrete Materials	3	2	6	2	4	9	7
Paper Response Card	0	1	2	0	0	0	0
Paper-Pencil Test	0	0	2	1	3	1	1
Read Aloud Stimuli	5	6	6	8	2	6	1
Reinforcement System	6	7	4	8	5	9	11
Scribe Items	3	6	9	7	4	16	3
Translated Test	0	0	1	0	0	1	1

Table 20. Total Number of Students with Allowed Accommodations: Mathematics

Accommodations	**Grade**						
	3	**4**	**5**	**6**	**7**	**8**	**11**
Alternate Response Options	4	4	6	2	2	5	3
American Sign Language (Non-Embedded)	0	0	1	0	0	1	0
Calculator	2	2	0	1	1	1	3
Concrete Materials	3	2	6	2	4	9	7
Digital Math Manipulatives	4	2	5	0	1	2	2
Multiplication Table	1	1	4	1	1	4	1
Paper Response Card	0	1	2	0	0	0	0
Paper-Pencil Test	0	0	2	1	2	1	1
Read Aloud Stimuli	4	5	6	8	1	6	1
Reinforcement System	6	7	4	8	5	9	11
Scribe Items	3	5	9	7	3	16	4
Translated Test	0	0	1	0	0	1	1

Table 21. Total Number of Students with Allowed Accommodations: Science

Accommodations	**Grade**		
	5	**8**	**11**
Alternate Response Options	4	5	3
American Sign Language (Non-Embedded)	1	1	0
Concrete Materials	3	9	7
Calculator	0	1	3
Paper Response Card	2	0	0
Multiplication Table	2	4	1
Paper-Pencil Test	2	1	1
Read Aloud Stimuli	3	6	1
Reinforcement System	1	9	11
Scribe Items	6	16	4
Translated Test	1	1	1

5.3.2. 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 22 shows suggested substitutions and alternatives that are based on the student's degree of vision, hearing, or physical mobility.

Table 22. 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 (AT).
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 AT, 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, AT, extended time, and multiple or frequent breaks.

5.3.3. 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 in accessing the HSA-Alt through the TDS. Technology affords many ways to adapt student responses 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 that 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 TA 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.

5.4 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. 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, mathematics, 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 clicks 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 human voice recording (HVR). The speed of narration is comparable to the average speed of narration when teachers read to students. Students respond to each item by clicking one of the response options presented, or the TA can click the student's selected response option for them. Students can change their answer selection as needed, however, once the *Next* button is selected, the assessment moved on to the next item. The online system automatically stores item responses when students click their selected-response option and then select the "Next" button.

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

In spring 2024, an ESR 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 ESR was instituted for a student's test if all of the following five 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 stops the student's test. The TA and the Test Session Observer are 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.

5.5 PAPER-PENCIL TEST ADMINISTRATION (VIA ONLINE FIXED FORM)

In spring 2024, students who required a paper-pencil accommodation were administered a fixed-form test via one of two options:

1. 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 HVR via the online testing system
2. A printed student test booklet, alongside a printed TA script booklet; the TA read aloud provided scripts for all components of the assessment, marked the student's response in the student test booklet, and entered the student's answers into the online testing system at the completion of paper-based testing

TAs completed and submitted the LCI, which investigated 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. No access-limited items were included on the fixed-form tests for the paper-pencil administration. The number of students who received the fixed-form test in spring 2024 can be found in Table 25.

5.6 TEST SECURITY

The Test Security Guidelines, embedded in the *Test Administration Manual*, indicate that photocopying any printed testing materials is strictly prohibited. Printed response cards and printed test visuals are secure materials. School test coordinators are 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 will make enough effort to be sure that all secure materials are returned. Any known violations of test security are to be immediately reported.

5.6.1. 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 the particular tests they are assigned. If information must be sent via email or fax, only the Statewide Student Identifier (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, teacher, or HIDOE staff can view students' scores. CASs and CSs have access to all scores within their district. Test coordinators have access to all scores within their school. Teachers 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.

5.6.2. System Security

The objective of system security is to ensure that all data are protected and correctly accessed 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) are 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 test coordinators and teachers 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.

5.7 Prevention of and Recovery from Disruptions in Test Delivery System

CAI is continuously improving our ability to protect our systems from interruptions. CAI's TDS is designed to ensure that student responses are accurately captured 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 sends out text messages to notify 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.

5.7.1. 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 robustly respond 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 asynchronously saved, 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. Depending on the situation, the student is presented with the following situations:

- 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.

5.7.2. 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.

5.7.3. 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 nightly incremental backups. 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.

6. SCORING

For the HSA-Alt, 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.

6.1 ITEM SCORING RULES

For multiple-choice items scored dichotomously, students receive one point for selecting the correct response option and zero points for any incorrect response options. For multi-select items with two correct response options, students earn two points for selecting both options, one point for selecting only one, and zero points for selecting none. If the Test Administrator (TA) marks an item as *No Response* (NR), the student receives zero points.

6.2 ATTEMPTEDNESS RULES FOR SCORING

When a student logs in to the test administration system and is presented with one item, they are considered to have participated if they provide a valid response to that item. A valid response includes either marking one or more response options or an NR marked by the TA on the item. Participated students are counted as attempted.

Scores are generated only for attempted tests. Detailed scoring rules are as follows (refer to Section 2.3 for the description of test segments):

- If a student answers 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, which consists solely of field-test items, the test will be scored without penalty.
- If a student does not complete Segments 1 and 2 but generates five or more valid responses with at least one non-*NR* response, the student is scored with penalty. The penalty is the theta estimate based on responded items minus one conditional standard error of measurement (SEM) for the estimated theta value.
- If a student generates 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 (LOSS). The SEM and theta score will be set to BLANK.
- If a student has *NR*s on all eight items in Segment 1 (Early Stopping Rule [ESR]), the test will end and the student is given the LOSS. The SEM and theta score will be set to BLANK.

Table 25 in Section 7.1, Student Participation, lists the number of "completed" tests without scoring penalty, the number of "Incomplete" tests (second and third bullets in the above list), and the number of ESR students.

6.3 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

The item response model (IRT) used to generate student scores employs the Rach model for dichotomous items and the Partial Credit Model (PCM) for polytomous items. The HSA-ALT tests 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 jth person's score pattern for I items is

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

where $b_i = (b_{i,1}, \dots, b_{i,m_i})$ for the ith 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 $p_{ij}(z_{ij}|\theta_j, b_{i,1}, \dots, b_{i,m_i})$ takes either the form of the Rasch model for items with one point or the form based on the PCM for items with two or more points.

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

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

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.,

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

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:

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

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

6.4 Scoring All Correct and All Incorrect Cases

Using the MLE method, a test where no items are answered correctly (i.e., all incorrect) would receive a theta estimate of negative infinity, and a test where all items are answered correctly (i.e., all correct) would receive a theta estimate of positive infinity. To obtain real-valued theta score estimates for these extreme cases, 0.3 is added to an item score among the administered operational items for the all-incorrect case, and 0.3 is subtracted from an item score for the all-correct case.

6.5 Rules for Transforming Theta Scores to Scale Scores

The student's performance in each test is summarized in an overall test score referred to as a *scale score*. Student theta scores, which are based on the number of items answered correctly and the difficulty of those items, are converted into scale scores. This conversion involves a linear transformation using the formula $SS = a * \theta \ + b$, where a is the transformation slope and b is the transformation intercept. Table 23 presents the scaling slope and intercept for each test. The final scale scores are rounded to the nearest integer.

Standard errors of the MLEs are converted to the scale score metric using the following formula:

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

where SE_{ss} is the standard error of the ability estimate on the scale score metric, SE_{θ} is the standard error of the ability estimate on the theta score metric, and a is the transformation slope used to convert theta scores into scale scores.

Table 23. 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

6.6 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, the minimum and maximum scale scores are set at 100 and 500, respectively. Overall scale scores below 100 are truncated to 100, and those above 500 are truncated to 500. The standard error for LOSS and HOSS is calculated using the estimated theta scores derived from the responded items.

6.7 Performance Levels

The scale scores are mapped into four performance levels. Table 24 provides the range of scale scores corresponding to each performance level by subject and grade.

Table 24. 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. SUMMARY OF SPRING 2024 OPERATIONAL TEST ADMINISTRATION

7.1 Student Participation

The HSA-Alt was administered by subject and grade level. All students in grades 3–8 and 11 were assessed in 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 the Test Administrator (TA) marked *No Response* to at least one item.

Table 25 displays the total number of students who attempted the online adaptive and online fixed-form HSA-Alt tests by subject and grade. The "Completed" column shows the number of students who finished the test, while the "Incomplete" column shows the number of students who did not. The "ESR" column (Early Stopping Rule) shows the number of students who did not respond to any items in the first segment of test and exited early. The ESR is available for non-responsive students. Annual HSA-Alt test administration training provides detailed guidance on ESR eligibility criteria, the verification process, and the implications of non-verification, which may include invalidation of the test score and the need for the student to retake the assessment. HIDOE reinforces appropriate ESR administration through Assessment News and Office Hours.

Table 25. Number of Attempted Students

Subject	Grade	Online Adaptive				Online Fixed-Form				Total
		Completed	ESR*	Incomplete	Total	Completed	ESR*	Incomplete	Total	
ELA	3	123	4	5	132					132
	4	112	9	3	124	1			1	125
	5	101	15	0	116	1	1		2	118
	6	109	3	4	116					116
	7	130	11	5	146					146
	8	96	5	2	103					103
	11	119	3	0	122					122
Math	3	123	5	2	130					130
	4	111	9	2	122	1			1	123
	5	102	14	0	116	1		1	2	118
	6	108	3	4	115					115
	7	129	10	5	144					144
	8	94	5	4	103					103
	11	118	3	1	122					122
Science	5	95	13	1	109	1	1		2	111
	8	94	6	1	101					101
	11	118	3	0	121					121

* Early Stopping Rule.

Table 26 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 27 presents the total number and percentage of students who participated in the HSA-Alt by subgroup. Table 28 presents the total number and percentage of students who participated in the HSA-Alt in each disability category classified under Individuals with Disabilities Education Act (IDEA), and by subgroup. Table 29–Table 32 provide the total number of students who participated in the HSA-Alt by subgroup and IDEA category for each grade.

Table 26. Overall Alternate Assessment Participation Rate

Subject	Grade	Number of HSA-Alt Participants	Number of Hawai`i State Summative Test Participants	Overall Hawai`i State Alternate Assessment Participation Rate (%)[1]
ELA	3	132	12,256	1.07%
	4	125	12,785	0.97%
	5	118	13,141	0.89%
	6	116	12,400	0.93%
	7	146	12,167	1.19%
	8	103	12,202	0.84%
	11	122	10,884	1.11%
	Overall	**862**	**85,835**	**0.99%**
Mathematics	3	130	12,317	1.04%
	4	123	12,831	0.95%
	5	118	13,189	0.89%
	6	115	12,479	0.91%
	7	144	12,257	1.16%
	8	103	12,270	0.83%
	11	122	10893	1.11%
	Overall	**855**	**86,236**	**0.98%**
Science	5	111	13,243	0.83%
	8	101	12,370	0.81%
	11	121	11,882	1.01%
	Overall	**333**	**37,495**	**0.88%**

[1] The U.S. Department of Education (USDE) looks at the overall participation rates in each subject with all grades combined. All three subject areas were around 1.0%.

Table 27. Number of Participated Students by Subgroup

Group	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
ELA							
All	132	125	118	116	146	103	122
Female	34	33	44	36	50	27	43
Male	98	92	74	80	96	76	79
Asian/Pacific Islander	30	34	30	39	33	24	34
Hawaiian Pacific Islander	38	34	29	29	36	30	38
White	11	9	8	9	12	9	11
Hispanic	19	21	26	22	34	25	22
American Indian/Alaska Native							1
African American	2	4	2	3			1
Multi-Racial	32	23	23	14	31	15	15
Migrant	1					1	2
Disadvantaged	50	64	56	56	69	51	60
ELL	22	24	22	24	25	21	25
Mathematics							
All	130	123	118	115	144	103	122
Female	34	32	44	36	49	27	44
Male	96	91	74	79	95	76	78
Asian/Pacific Islander	30	34	30	39	33	24	35
Hawaiian Pacific Islander	37	33	29	29	34	30	38
White	11	9	8	9	12	9	10
Hispanic	19	21	26	22	34	25	22
American Indian/Alaska Native							1
African American	2	3	2	3			1
Multi-Racial	31	23	23	13	31	15	15
Migrant	1					1	2
Disadvantaged	48	63	56	56	68	51	60
ELL	20	24	22	24	25	21	24
Science							
All			111			101	121
Female			41			25	43
Male			70			76	78
Asian/Pacific Islander			29			24	34
Hawaiian Pacific Islander			28			29	38
White			6			9	10
Hispanic			24			24	22
American Indian/Alaska Native							1
African American			2				1
Multi-Racial			22			15	15
Migrant						1	2
Disadvantaged			51			49	59
ELL			21			20	25

Table 28. Number of Participated Students by Subgroup and Disability Category—Overall

Subgroup Category	ASD	DD6	DF	ED	HH	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Number of Students													
All	347	1	1	6	1	196	231	6	60	6	1	5	2
Female	74	1		1	1	76	89	1	21	1		2	1
Male	273		1	5		120	142	5	39	5	1	3	1
Asian/Pacific Islander	101					39	72		10	1			1
Hawaiian Pacific Islander	75		1	2		66	62	3	20	2	1	2	
White	28					15	19		7				
Hispanic	64	1		1		48	39	1	10	3		2	
American Indian/Alaska Native	1												
African American	8					1		1	2				
Multi-Racial	70			3		27	39	1	11			1	1
Migrant	2					2							
Disadvantaged	130	1		4		126	98	3	34	3	1	5	1
ELL	66		1	1	1	46	32	2	11	2			1
Percentage of Students by Subgroup Conditional on Each IDEA Category													
Female	21%	100%		17%	100%	39%	39%	17%	35%	17%		40%	50%
Male	79%		100%	83%		61%	61%	83%	65%	83%	100%	60%	50%
Asian/Pacific Islander	29%				100%	20%	31%		17%	17%			50%
Hawaiian Pacific Islander	22%		100%	33%		34%	27%	50%	33%	33%	100%	40%	
White	8%					8%	8%		12%				
Hispanic	18%	100%		17%		24%	17%	17%	17%	50%		40%	
American Indian/Alaska Native	0%												
African American	2%					1%		17%	3%				
Multi-Racial	20%			50%		14%	17%	17%	18%			20%	50%
Migrant	1%					1%							
Disadvantaged	37%	100%		67%		64%	42%	50%	57%	50%	100%	100%	50%
ELL	19%		100%	17%	100%	23%	14%	33%	18%	33%			50%

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 29. Number of Participated Students by Subgroup and Disability Category (Grades 3–4)

Group	ASD	DD6	DF	ED	HH	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 3													
All Students	78	1		2		24	19	2	6				
Female	15	1				7	6	1	4				
Male	63			2		17	13	1	2				
Asian/Pacific Islander	22					1	6		1				
Hawaiian Pacific Islander	16					10	8	1	3				
White	6					4	1						
Hispanic	13	1				3			2				
American Indian/Alaska Native													
African American	2												
Multi-Racial	19			2		6	4	1					
Migrant	1												
Disadvantaged	21	1				14	7	2	5				
ELL	14					4	1		3				
Grade 4													
All Students	52					29	30	2	11	1			
Female	10					7	8		8				
Male	42					22	22	2	3	1			
Asian/Pacific Islander	17					4	11		1	1			
Hawaiian Pacific Islander	12					10	7	1	4				
White	2					2	2		3				
Hispanic	7					9	3	1	1				
American Indian/Alaska Native													
African American	2					1			1				
Multi-Racial	12					3	7		1				
Migrant													
Disadvantaged	22					22	15		4	1			
ELL	11					5	5	1	1	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.

Table 30. Number of Participated Students by Subgroup and Disability Category (Grades 5–6)

Group	ASD	DD6	DF	ED	HH	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 5													
All Students	43					29	34		8	1	1	1	1
Female	12					12	18		1				1
Male	31					17	16		7	1	1	1	
Asian/Pacific Islander	12					6	11		1				
Hawaiian Pacific Islander	6					11	6		5		1		
White	3					2	3						
Hispanic	12					7	5			1		1	
American Indian/Alaska Native													
African American	2												
Multi-Racial	8					3	9		2				1
Migrant													
Disadvantaged	13					19	17		4		1	1	1
ELL	6					7	6		2	1			
Grade 6													
All Students	52					20	36	1	6			1	
Female	12					11	11		1			1	
Male	40					9	25	1	5				
Asian/Pacific Islander	19					5	11		4				
Hawaiian Pacific Islander	13					5	10					1	
White	5					1	2		1				
Hispanic	5					9	8						
American Indian/Alaska Native													
African American	2							1					
Multi-Racial	8						5		1				
Migrant													
Disadvantaged	24					15	13		3			1	
ELL	14					4	4		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 31. Number of Participated Students by Subgroup and Disability Category (Grades 7–8)

Group	ASD	DD6	DF	ED	HH	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 7													
All Students	54					32	45		13	2			
Female	9					14	22		5				
Male	45					18	23		8	2			
Asian/Pacific Islander	9					6	16		2				
Hawaiian Pacific Islander	13					10	9		3	1			
White	5					2	4		1				
Hispanic	15					7	8		3	1			
American Indian/Alaska Native													
African American													
Multi-Racial	12					7	8		4				
Migrant													
Disadvantaged	21					17	21		9	1			
ELL	8					8	6		3				
Grade 8													
All Students	29		1	2		26	32		10	1		2	
Female	6			1		8	9		2			1	
Male	23		1	1		18	23		8	1		1	
Asian/Pacific Islander	9					4	10		1				
Hawaiian Pacific Islander	7		1	2		9	8		2			1	
White	4					2	2		1				
Hispanic	6					6	8		3	1		1	
American Indian/Alaska Native													
African American													
Multi-Racial	3					5	4		3				
Migrant						1							
Disadvantaged	13			2		19	9		6			2	
ELL	5		1	1		9	5						

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 32. Number of Participated Students by Subgroup and Disability Category (Grade 11)

Group	ASD	DD6	DF	ED	HH	ID	MD	OD	OHD	SLD	SOL	TBI	VDB
Grade 11													
All Students	39			2	1	36	35	1	6	1		1	1
Female	10				1	17	15			1			
Male	29			2		19	20	1	6			1	1
Asian/Pacific Islander	13				1	13	7						1
Hawaiian Pacific Islander	8					11	14	1	3	1			
White	3					2	5		1				
Hispanic	6			1		7	7		1				
American Indian/Alaska Native	1												
African American									1				
Multi-Racial	8			1		3	2					1	
Migrant	1					1							
Disadvantaged	16			2		20	16	1	3	1		1	
ELL	8				1	9	5	1					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.

7.2 SUMMARY OF STUDENT PERFORMANCE

Table 33–Table 41 present a summary of the spring 2024 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 33. 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	132	278.79	57.17	52	10	27	11	39
Female	34	274.75	73.46	44	6	35	15	50
Male	98	280.19	50.69	54	11	24	10	35
Asian/Pacific Islander	30	283.82	59.10	50	10	27	13	40
Hawaiian Pacific Islander	38	268.71	64.24	55	8	29	8	37
White	11	285.37	52.00	36	18	27	18	45
Hispanic	19	294.97	59.02	42	0	32	26	58
American Indian/Alaska Native								
African American	2*							
Multi-Racial	32	274.36	47.80	59	16	22	3	25
Migrant	1*							
Disadvantaged	50	285.15	53.76	44	10	30	16	46
ELL	22	293.53	41.06	41	9	36	14	50
Grade 4								
All Students	125	285.36	60.97	36	17	26	22	47
Female	33	296.57	44.06	30	18	27	24	52
Male	92	281.34	65.73	38	16	25	21	46
Asian/Pacific Islander	34	279.74	70.15	35	15	35	15	50
Hawaiian Pacific Islander	34	287.99	57.51	38	24	6	32	38
White	9	310.56	22.65	11	11	44	33	78
Hispanic	21	292.16	53.60	38	5	43	14	57
American Indian/Alaska Native								
African American	4*							
Multi-Racial	23	269.82	71.30	48	22	13	17	30
Migrant								
Disadvantaged	64	289.42	56.88	31	22	25	22	47
ELL	24	289.33	62.68	29	17	33	21	54

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 34. 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	118	261.00	74.80	50	25	14	10	25
Female	44	241.51	82.25	57	27	11	5	16
Male	74	272.59	67.95	46	24	16	14	30
Asian/Pacific Islander	30	264.97	51.61	57	33	7	3	10
Hawaiian Pacific Islander	29	286.78	59.55	31	28	24	17	41
White	8	262.63	66.97	38	63	0	0	0
Hispanic	26	242.36	94.32	62	19	4	15	19
American Indian/Alaska Native								
African American	2*							
Multi-Racial	23	248.26	85.90	57	9	26	9	35
Migrant								
Disadvantaged	56	255.88	78.68	52	21	14	13	27
ELL	22	277.61	62.15	32	50	9	9	18
Grade 6								
All Students	116	280.51	56.00	40	23	25	12	37
Female	36	282.34	47.75	36	28	25	11	36
Male	80	279.69	59.60	41	21	25	13	38
Asian/Pacific Islander	39	272.89	57.08	46	26	23	5	28
Hawaiian Pacific Islander	29	286.01	49.12	31	24	34	10	45
White	9	295.48	43.70	33	22	0	44	44
Hispanic	22	271.10	75.15	41	14	32	14	45
American Indian/Alaska Native								
African American	3*							
Multi-Racial	14	296.60	43.24	36	29	21	14	36
Migrant								
Disadvantaged	56	291.52	53.75	32	21	30	16	46
ELL	24	277.99	51.51	42	25	21	13	33
Grade 7								
All Students	146	270.91	63.19	47	27	14	12	26
Female	50	264.18	72.87	54	24	12	10	22
Male	96	274.41	57.61	44	28	16	13	28
Asian/Pacific Islander	33	259.28	58.70	64	21	9	6	15
Hawaiian Pacific Islander	36	277.16	37.00	42	39	14	6	19
White	12	265.84	82.96	42	17	25	17	42
Hispanic	34	265.22	76.20	44	29	6	21	26
American Indian/Alaska Native								
African American								
Multi-Racial	31	284.22	68.47	42	19	26	13	39
Migrant								
Disadvantaged	69	282.60	58.97	39	29	14	17	32
ELL	25	271.01	44.48	52	32	8	8	16

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 35. 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	103	267.18	56.99	48	29	17	6	23
Female	27	270.19	58.23	44	33	15	7	22
Male	76	266.11	56.90	49	28	18	5	24
Asian/Pacific Islander	24	255.89	53.87	58	33	4	4	8
Hawaiian Pacific Islander	30	263.41	41.34	57	30	13	0	13
White	9	254.87	54.05	78	0	22	0	22
Hispanic	25	273.46	84.45	32	24	24	20	44
American Indian/Alaska Native								
African American								
Multi-Racial	15	289.72	22.27	20	47	33	0	33
Migrant	1*							
Disadvantaged	51	273.32	53.83	41	29	24	6	29
ELL	21	251.05	55.28	57	38	5	0	5
Grade 11								
All Students	122	275.74	45.64	44	28	19	9	28
Female	43	273.20	40.27	42	35	19	5	23
Male	79	277.13	48.50	46	24	19	11	30
Asian/Pacific Islander	34	279.53	38.28	47	21	21	12	32
Hawaiian Pacific Islander	38	281.81	43.90	34	34	21	11	32
White	11	275.76	42.67	55	27	9	9	18
Hispanic	22	268.35	54.69	55	18	23	5	27
American Indian/Alaska Native	1*							
African American	1*							
Multi-Racial	15	263.66	56.54	40	40	13	7	20
Migrant	2*							
Disadvantaged	60	279.28	43.45	43	28	17	12	28
ELL	25	281.08	44.03	52	12	24	12	36

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 36. 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	130	282.47	58.93	41	21	17	22	38
Female	34	272.94	68.58	44	21	18	18	35
Male	96	285.85	55.12	40	21	17	23	40
Asian/Pacific Islander	30	274.04	69.12	47	17	20	17	37
Hawaiian Pacific Islander	37	275.92	65.77	54	8	19	19	38
White	11	295.29	40.99	18	36	9	36	45
Hispanic	19	286.82	64.33	37	16	21	26	47
American Indian/Alaska Native								
African American	2*							
Multi-Racial	31	287.88	40.77	32	39	13	16	29
Migrant	1*							
Disadvantaged	48	286.91	55.57	35	17	25	23	48
ELL	20	297.50	33.58	35	15	35	15	50
Grade 4								
All Students	123	280.09	68.82	44	19	23	15	37
Female	32	289.39	53.35	44	19	19	19	38
Male	91	276.82	73.47	44	19	24	13	37
Asian/Pacific Islander	34	284.75	79.74	32	21	32	15	47
Hawaiian Pacific Islander	33	278.70	66.28	45	18	21	15	36
White	9	306.57	44.59	44	11	22	22	44
Hispanic	21	283.90	59.91	52	14	14	19	33
American Indian/Alaska Native								
African American	3*							
Multi-Racial	23	263.97	75.04	48	22	22	9	30
Migrant								
Disadvantaged	63	283.83	65.17	44	16	24	16	40
ELL	24	291.60	70.67	33	21	25	21	46
Grade 5								
All Students	118	263.99	70.37	58	9	25	7	32
Female	44	244.97	79.30	70	7	18	5	23
Male	74	275.30	62.33	51	11	30	8	38
Asian/Pacific Islander	30	275.53	51.43	60	10	30	0	30
Hawaiian Pacific Islander	29	285.45	43.70	52	14	24	10	34
White	8	266.45	74.05	50	0	50	0	50
Hispanic	26	249.33	89.51	62	4	23	12	35
American Indian/Alaska Native								
African American	2*							
Multi-Racial	23	240.32	81.24	65	13	17	4	22
Migrant								
Disadvantaged	56	259.67	71.18	59	13	23	5	29
ELL	22	274.74	62.74	59	9	27	5	32

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 37. 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	115	266.79	61.89	52	19	20	9	29
Female	36	262.13	47.10	58	17	22	3	25
Male	79	268.91	67.73	49	20	19	11	30
Asian/Pacific Islander	39	267.61	60.75	56	13	21	10	31
Hawaiian Pacific Islander	29	264.15	59.95	55	14	24	7	31
White	9	289.98	44.57	22	22	44	11	56
Hispanic	22	251.77	70.13	50	36	9	5	14
American Indian/Alaska Native								
African American	3*							
Multi-Racial	13	290.48	65.97	46	23	15	15	31
Migrant								
Disadvantaged	56	276.14	64.68	46	21	20	13	32
ELL	24	265.23	63.49	54	21	17	8	25
Grade 7								
All Students	144	259.54	68.44	49	24	16	12	28
Female	49	251.53	73.87	53	24	12	10	22
Male	95	263.68	65.48	46	23	18	13	31
Asian/Pacific Islander	33	240.28	63.10	70	15	9	6	15
Hawaiian Pacific Islander	34	273.81	52.49	35	35	12	18	29
White	12	259.11	87.29	33	25	25	17	42
Hispanic	34	254.54	77.79	53	15	24	9	32
American Indian/Alaska Native								
African American								
Multi-Racial	31	270.06	69.39	42	29	16	13	29
Migrant								
Disadvantaged	68	271.07	63.09	43	22	22	13	35
ELL	25	255.04	50.46	60	24	12	4	16
Grade 8								
All Students	103	265.29	58.85	56	23	11	10	20
Female	27	262.16	60.70	67	22	4	7	11
Male	76	266.40	58.55	53	24	13	11	24
Asian/Pacific Islander	24	267.32	64.73	58	25	8	8	17
Hawaiian Pacific Islander	30	265.77	46.56	63	17	7	13	20
White	9	253.62	59.43	67	22	11	0	11
Hispanic	25	254.93	79.49	52	24	8	16	24
American Indian/Alaska Native								
African American								
Multi-Racial	15	285.35	17.05	40	33	27	0	27
Migrant	1*							
Disadvantaged	51	267.45	54.37	53	24	14	10	24
ELL	21	251.71	60.84	71	19	0	10	10

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 38. 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	122	281.57	43.99	43	22	21	13	34
Female	44	277.34	36.71	52	30	14	5	18
Male	78	283.96	47.67	38	18	26	18	44
Asian/Pacific Islander	35	284.08	43.57	51	20	9	20	29
Hawaiian Pacific Islander	38	277.27	42.60	39	26	29	5	34
White	10	276.10	18.71	60	20	20	0	20
Hispanic	22	279.18	46.82	45	18	27	9	36
American Indian/Alaska Native	1*							
African American	1*							
Multi-Racial	15	292.09	59.68	27	20	20	33	53
Migrant	2*							
Disadvantaged	60	281.39	40.67	40	23	27	10	37
ELL	24	286.18	41.88	42	21	21	17	38

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 39. 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	111	254.92	75.07	50	23	17	9	26
Female	41	225.03	79.05	66	22	10	2	12
Male	70	272.42	67.23	41	24	21	13	34
Asian/Pacific Islander	29	254.13	54.74	59	28	14	0	14
Hawaiian Pacific Islander	28	277.72	51.20	39	29	25	7	32
White	6	244.77	75.39	50	33	17	0	17
Hispanic	24	252.95	97.81	42	21	13	25	38
American Indian/Alaska Native								
African American	2*							
Multi-Racial	22	234.16	86.53	64	14	18	5	23
Migrant								
Disadvantaged	51	252.56	75.66	55	20	20	6	25
ELL	21	260.55	64.17	57	29	5	10	14

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 40. 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	101	261.39	60.08	47	36	10	8	18
Female	25	263.69	53.91	48	36	12	4	16
Male	76	260.63	62.29	46	36	9	9	18
Asian/Pacific Islander	24	248.51	61.09	58	33	4	4	8
Hawaiian Pacific Islander	29	266.34	39.97	45	41	3	10	14
White	9	243.02	57.63	56	44	0	0	0
Hispanic	24	261.33	88.46	42	21	21	17	38
American Indian/Alaska Native								
African American								
Multi-Racial	15	283.55	24.87	33	47	20	0	20
Migrant	1*							
Disadvantaged	49	267.56	54.29	41	37	12	10	22
ELL	20	251.40	49.67	45	50	0	5	5

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

Table 41. 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	121	272.69	52.72	48	25	13	14	27
Female	43	272.35	46.91	44	30	14	12	26
Male	78	272.88	55.95	50	22	13	15	28
Asian/Pacific Islander	34	278.30	47.22	50	21	12	18	29
Hawaiian Pacific Islander	38	275.52	52.76	39	34	13	13	26
White	10	268.69	33.54	50	30	20	0	20
Hispanic	22	267.58	58.55	50	23	14	14	27
American Indian/Alaska Native	1*							
African American	1*							
Multi-Racial	15	266.24	70.88	53	13	13	20	33
Migrant	2*							
Disadvantaged	59	276.71	51.43	46	24	14	17	31
ELL	25	281.64	53.26	56	16	4	24	28

*To protect individual student confidentiality, results are not reported for five or fewer students.

^% Proficient is the sum of % Meets and % Exceeds.

7.3 TEST-TAKING TIME

The HSA-Alt 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 that 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 test administrators who are knowledgeable about their students, additional time for students who need it can be arranged.

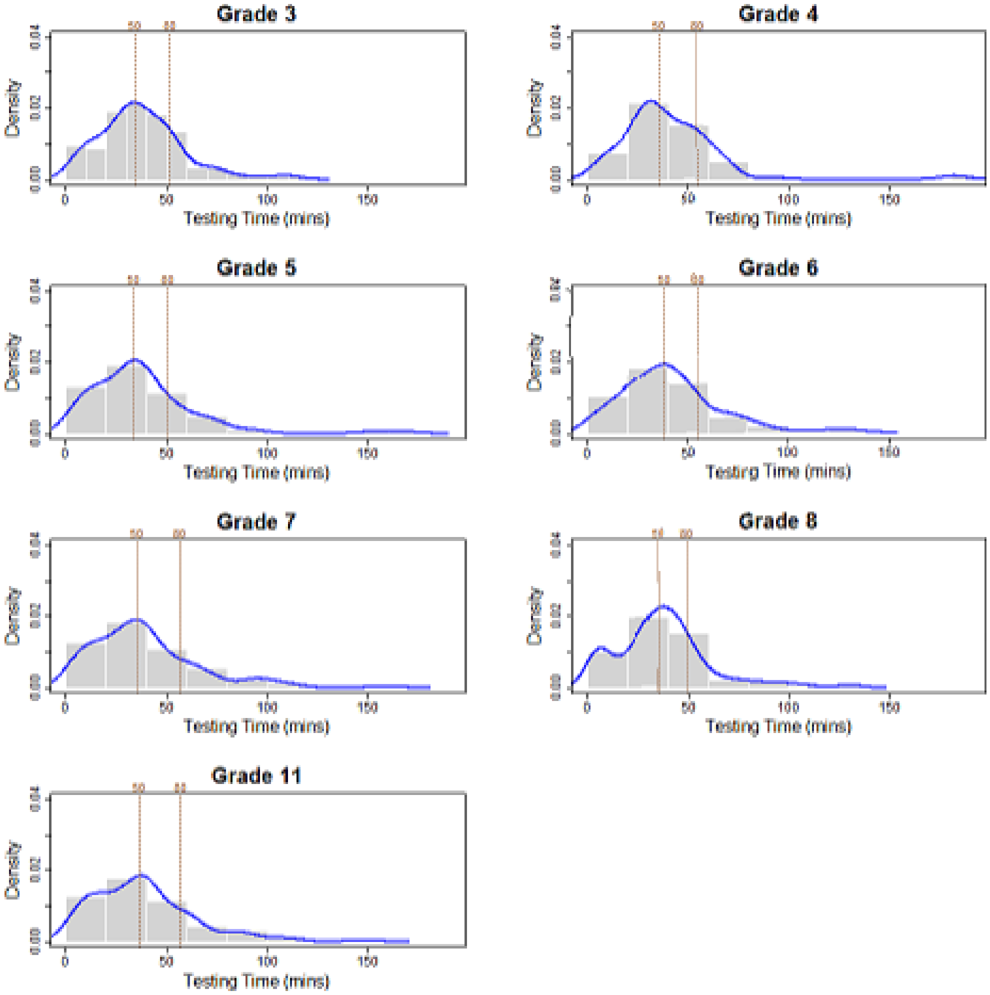
In the Test Delivery System (TDS), item response time is captured as the item page time (i.e., the time that a student spends on each item page) in milliseconds. Discrete items appear on the screen one item at a time, and items associated with a stimulus appear on the screen together with the page time 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 was the sum of the page time for all items. Students who meet the ESR criteria are not included in the analysis. The results are based on students who meet attemptedness requirements for scoring and reporting of the HSA-Alt.

Table 42 presents the 2024 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 2–Figure 4. Students who meet the ESR criteria are not included in the analysis. The results are based on students who meet attemptedness requirements for scoring and reporting of the HSA-Alt.

Table 42. Test-Taking Time

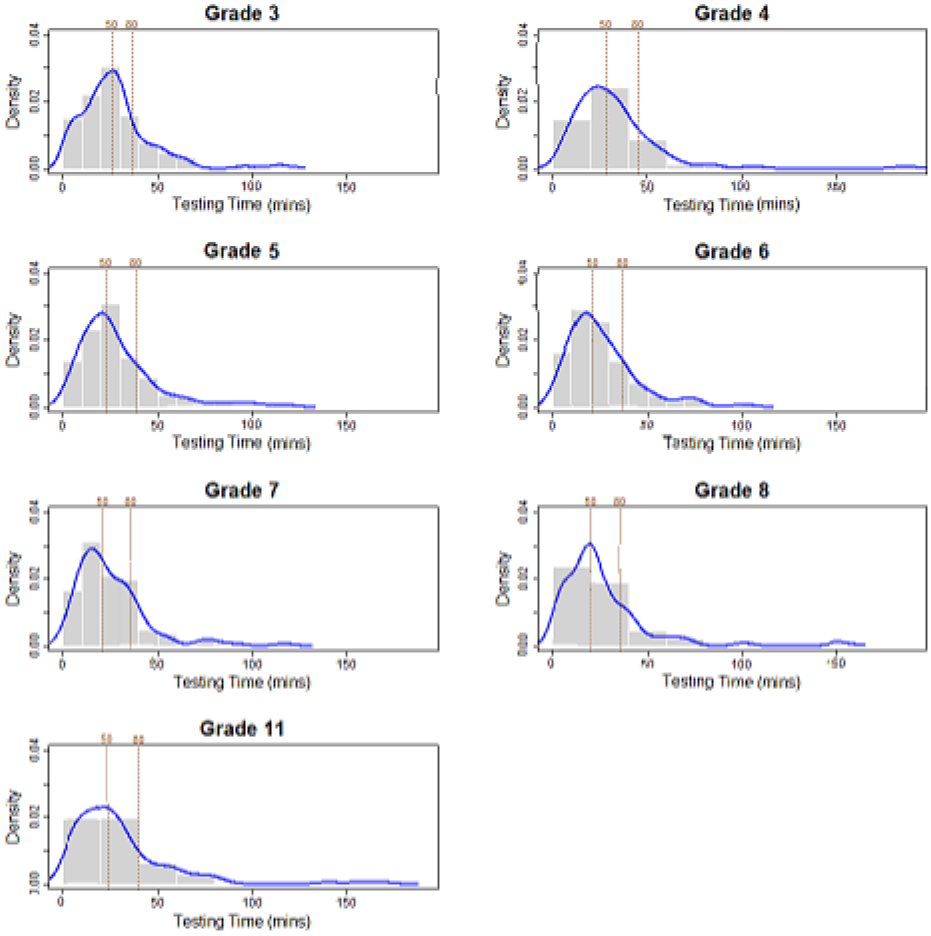
Subject	Grade	Average Testing Time (hh:mm)	Median Testing Time (hh:mm)	Testing Time in Percentiles (hh:mm)			
				Min	25th	75th	Max
ELA	3	00:37	00:35	00:02	00:23	00:49	01:50
	4	00:41	00:36	00:04	00:27	00:52	03:03
	5	00:39	00:34	00:02	00:25	00:47	02:48
	6	00:41	00:38	00:01	00:26	00:51	02:11
	7	00:41	00:36	00:02	00:23	00:50	02:38
	8	00:37	00:35	00:03	00:24	00:46	02:09
	11	00:38	00:36	00:01	00:19	00:50	02:27
Mathematics	3	00:26	00:25	00:02	00:15	00:32	01:54
	4	00:32	00:27	00:03	00:18	00:37	03:06
	5	00:28	00:23	00:01	00:17	00:35	01:56
	6	00:26	00:22	00:01	00:14	00:34	01:40
	7	00:25	00:21	00:01	00:14	00:32	01:56
	8	00:25	00:20	00:01	00:14	00:32	02:30
	11	00:30	00:23	00:02	00:12	00:37	02:48
Science	5	00:30	00:26	00:01	00:20	00:36	01:37
	8	00:25	00:22	00:02	00:15	00:30	01:20
	11	00:22	00:18	00:01	00:10	00:27	02:12

Figure 2. Distribution of Testing Time—ELA



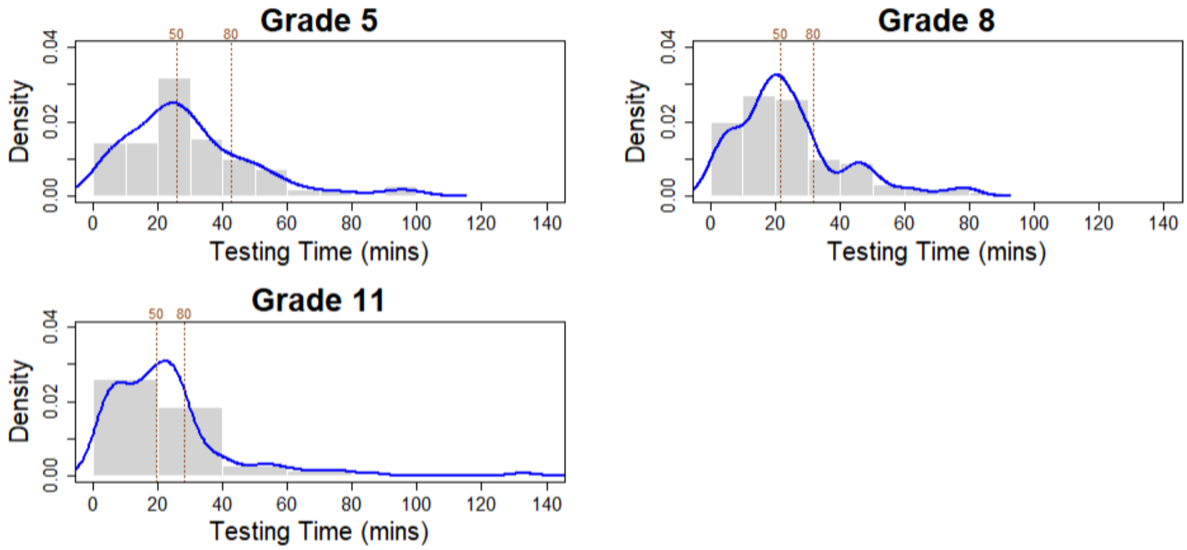
Note: 50 is median; 80 is 80th percentile

Figure 3. Distribution of Testing Time—Mathematics



Note: 50 is median; 80 is 80th percentile

Figure 4. Distribution of Testing Time—Science



Note: 50 is median; 80 is 80th percentile

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

Figure 5–Figure 7 display the empirical distribution of Hawai`i students' ability scores on the theta metric from the spring 2024 test administration and the distribution of item difficulty parameter estimates in the 2024 item pool. The student ability distributions were based on results from the completed test results from both the adaptive and fixed-form tests. These charts visually assess whether the difficulty levels of items in the pool cover the ability range of the assessed population and can guide future item development. For example, some mathematics tests may require additional easier items to better address students with lower academic achievement.

Table 43 presents the correlations between students' final estimated theta scores and the average test form difficulty for each subject and grade, based solely on students who completed the online adaptive tests. The strong correlations, ranging from 0.71 in grade 5 mathematics to 0.93 in grade 11 science, demonstrate that the adaptive algorithm functioned as intended and effectively matched items to students' abilities.

Figure 5. Student Ability and Item Difficulty Distributions for ELA

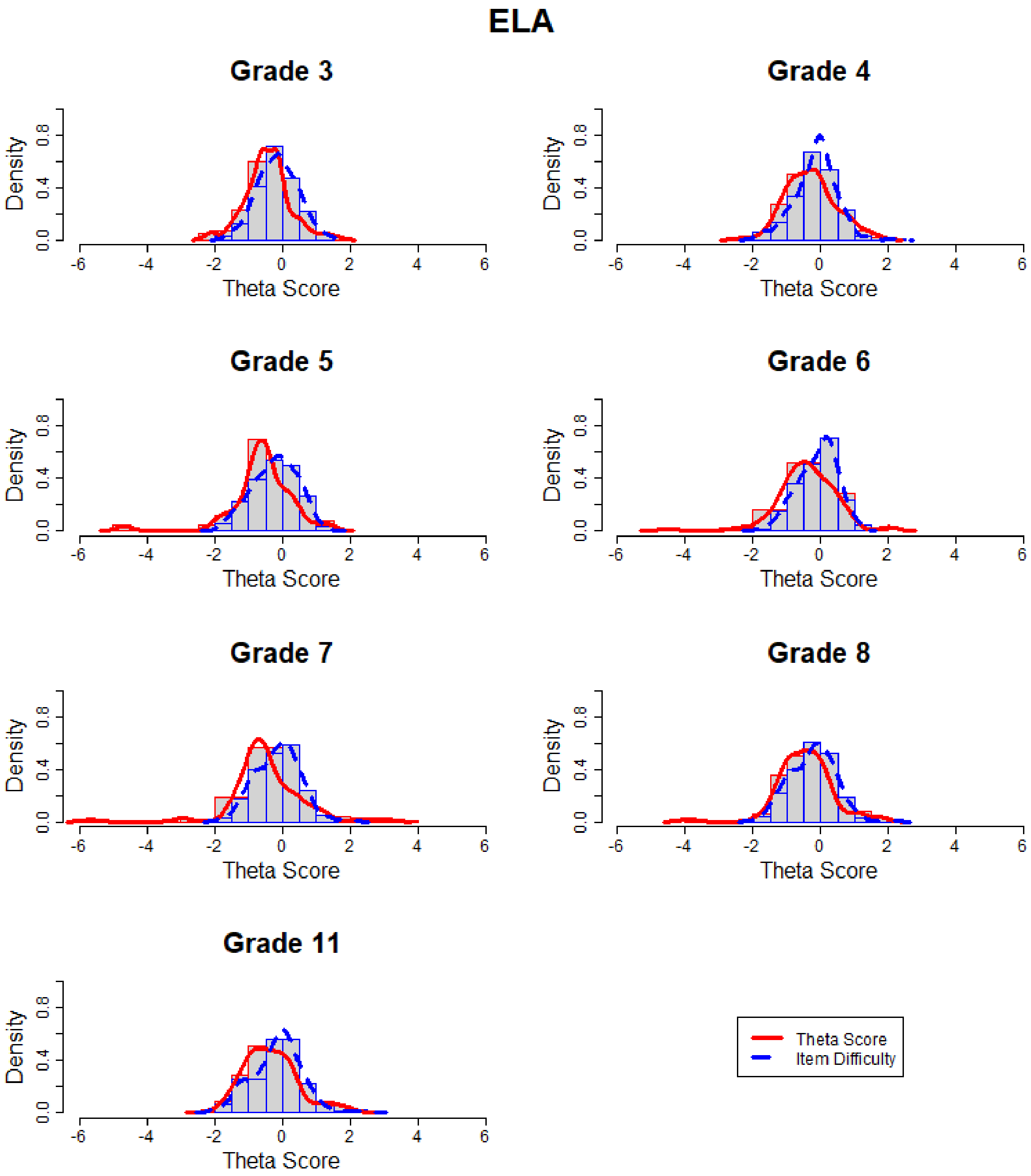


Figure 6. Student Ability and Item Difficulty Distributions for Mathematics

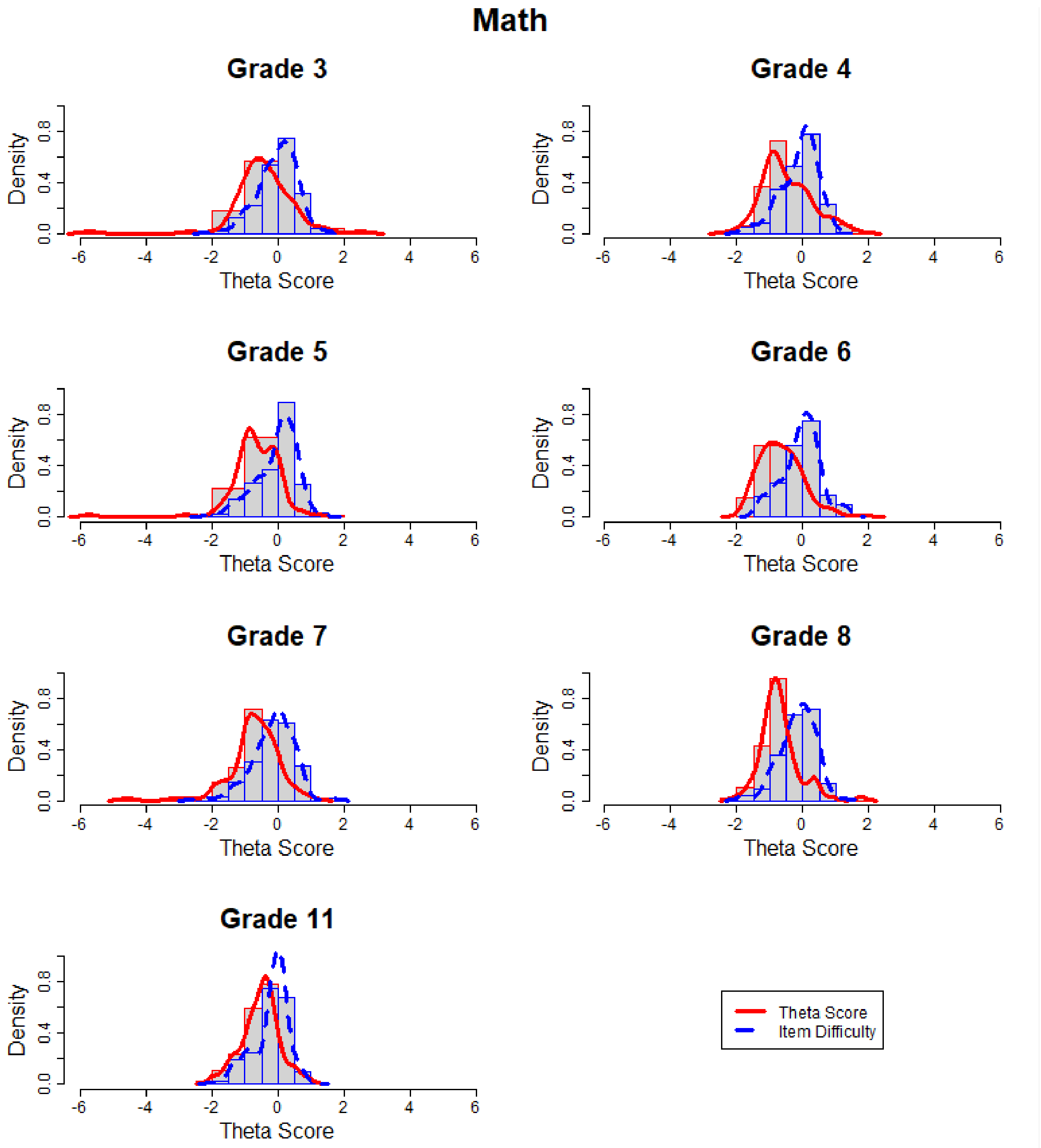


Figure 7. Student Ability and Item Difficulty Distributions for Science

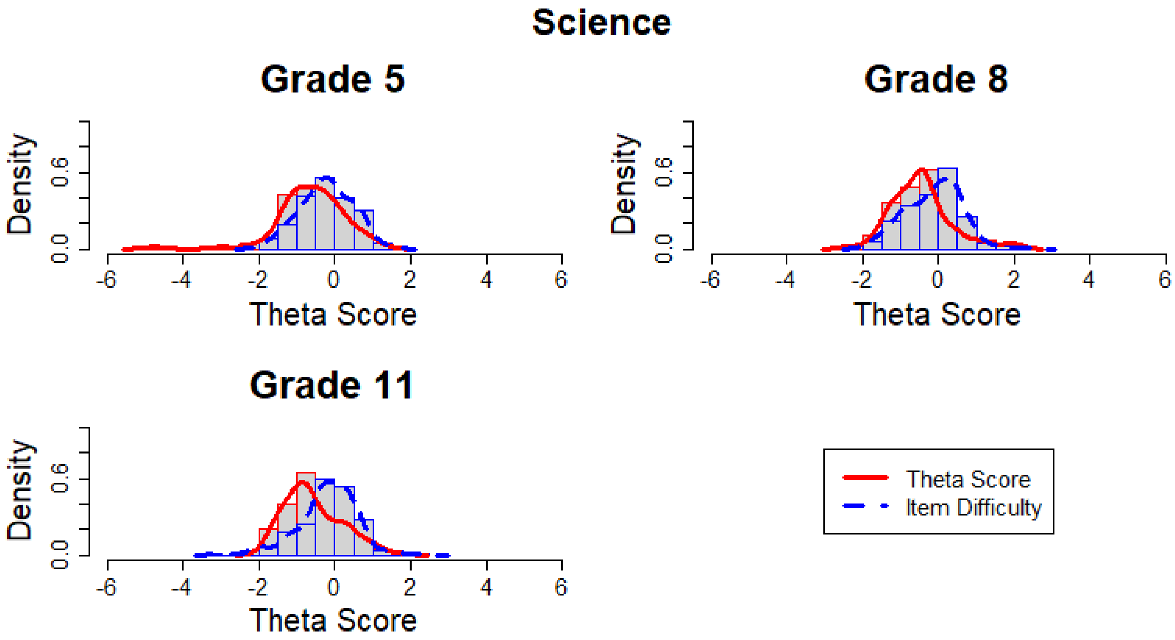


Table 43. Correlation Between Student Ability Scores and Average Test Form Difficulty

Subject	Grade	N	Correlation
ELA	3	123	0.90
	4	112	0.90
	5	101	0.79
	6	109	0.84
	7	130	0.80
	8	96	0.85
	11	119	0.86
Mathematics	3	123	0.80
	4	111	0.90
	5	102	0.71
	6	108	0.89
	7	129	0.80
	8	94	0.85
	11	118	0.88
Science	5	95	0.87
	8	94	0.88
	11	118	0.93

8. VALIDITY

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [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 logically starts 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.

The HSA-Alt was created with answering fundamental questions such as, what are the purposes of the assessment? Who are the intended users and what are the intended uses? Section 1.2 in this technical report illustrates that the purposes and intended uses of the HSA-Alt are to measure students' academic performance and student's progress in meeting the state alternate academic achievement standards in core content areas including ELA, mathematics, and science. The validation process and validity argument for the HSA-Alt, documented in this chapter, are established around these uses.

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses (p. 21; AERA, APA, & NCME, 2014). 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.

The state is also required to provide sufficient and solid validity evidence 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 Critical Element #3.

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.

1. Evidence based on test content (Critical Element 3.1—Overall Validity, Including Validity Based on Content)
2. Evidence based on response processes (Critical Element 3.2—Validity Based on Cognitive Process/Linguistic Processes)
3. Evidence based on internal structure (Critical Element 3.3—Validity Based on Internal Structure)
4. 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 standards, 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.

8.1 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 statewide standards implemented in daily instruction at school by teachers. This evidence is based on the justification for and connections among the following factors:

- Content standards
- Test blueprints
- Item development
- Test administration conditions
- 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 in this section.

8.1.1. Content Standards

The HSA-Alt is developed based on the Hawai`i Common Core Standards (HCCS) 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. The Hawaii alternate content standards, aligned with HCCS, were designed to make the standards more accessible to students with significant cognitive disabilities while maintaining the rigor and high expectations of the HCCS. These standards ensure that this student population are provided with multiple ways to learn and demonstrate knowledge. Refer to Section 1.4, in this technical report for details.

8.1.2. Test Blueprints

Content specifications in test blueprints specify the content standards to be covered in the test and the minimum and maximum number of items from each content domain and sub-standards under these domains. The goal is to ensure that 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 (i.e., field-test items have no impact on student scores). In the adaptive algorithm used on the operational items, item selection takes place in two stages: (1) blueprint satisfaction and (2) match-to-ability.

The blueprint match rates are provided for the operational tests. Table 44–Table 60 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 44. 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	99
	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	93
	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	94
	RI.3.2	0	1	100	0	2	100
	RI.3.3	0	1	100	0	2	100
	RI.3.4	0	1	100	0	2	100
	RI.3.5	0	1	100	0	2	97
	RI.3.6	0	1	100	0	2	100
	RI.3.7	0	1	100	0	2	88
	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	100
	RL.3.2	0	1	100	0	2	98
	RL.3.3	0	1	100	0	2	98
	RL.3.4	0	1	100	0	2	96
	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	98
	W.3.2	0	1	100	0	2	97
	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 45. 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	99
	RI.4.4	0	1	100	0	2	97
	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	81
	RI.4.8	0	1	100	0	2	98
	RI.4.9	0	1	100	0	2	99
Reading—Literature (RL)	Overall	3	3	100	8	9	100
	RL.4.1	0	1	100	0	2	52
	RL.4.2	0	1	100	0	2	99
	RL.4.3	0	1	100	0	2	80
	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	13
Writing (W)	Overall	1	1	100	7	9	100
	W.4.1	0	1	100	0	2	99
	W.4.2	0	1	100	0	2	96
	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	95

Table 46. 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.5.1	0	1	100	0	2	100
	L.5.2	0	1	100	0	2	98
	L.5.3	0	1	100	0	2	100
	L.5.4	0	1	100	0	2	100
	L.5.5	0	1	100	0	2	97
	L.5.6	0	1	100	0	2	100
Reading—Informational (RI)	Overall	3	3	100	8	9	100
	RI.5.1	0	1	100	0	2	64
	RI.5.2	0	1	100	0	2	66
	RI.5.3	0	1	100	0	2	98
	RI.5.4	0	1	100	0	2	99
	RI.5.5	0	1	100	0	2	98
	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	98
	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	59
	RL.5.2	0	1	100	0	2	100
	RL.5.3	0	1	100	0	2	98
	RL.5.4	0	1	100	0	2	58
	RL.5.6	0	1	100	0	2	86
	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	93
	W.5.3	0	1	100	0	2	100
	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 47. 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	76
	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	99
	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	96
Reading—Literature (RL)	Overall	3	3	100	7	9	100
	RL.6.1	0	1	100	0	2	88
	RL.6.2	0	1	100	0	2	99
	RL.6.3	0	1	100	0	2	100
	RL.6.4	0	1	100	0	2	100
	RL.6.6	0	1	100	0	2	98
	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	99
	W.6.7	0	1	100	0	2	100
	W.6.8	0	1	100	0	2	100

Table 48. 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	95
	L.7.5	0	1	100	0	2	96
	L.7.6	0	1	100	0	2	96
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	98
	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	100
Reading—Literature (RL)	Overall	3	3	100	7	9	100
	RL.7.1	0	1	100	0	2	83
	RL.7.2	0	1	100	0	2	96
	RL.7.3	0	1	100	0	2	95
	RL.7.4	0	1	100	0	2	97
	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	99
	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	92
	W.7.8	0	1	100	0	2	100

Table 49. 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	100
	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	99
	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	99
	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 50. 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	99
	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	2	100
	RI.11-12.2	0	1	100	0	2	99
	RI.11-12.3	0	1	100	0	2	91
	RI.11-12.4	0	1	100	0	2	100
	RI.11-12.6	0	1	100	0	2	87
	RI.11-12.8	0	1	100	0	2	100
	RI.11-12.9	0	1	100	0	2	100
Reading—Literature (RL)	Overall	3	3	100	6	8	100
	RL.11-12.1	0	1	100	0	2	95
	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	94
	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	92
	W.11-12.2	0	1	100	0	2	92
	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	99

Table 51. 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	99
	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	95
	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 52. 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 53. 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	100
	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 54. 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 55. 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	8	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 56. 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
	8.SP.A.4	0	1	100	0	1	100

Table 57. 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.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
	HS.A.SSE.A.1a	0	1	100	0	2	100
	HS.A.SSE.B.3	0	1	100	0	2	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.5	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
Geometry (G)	Overall	2	2	100	7	9	100
	HS.G.C.A.2	0	1	100	0	1	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.B.6	0	1	100	0	2	100
	HS.G.CO.C.10	0	1	100	0	2	100
	HS.G.CO.C.11	0	1	100	0	1	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.5	0	1	100	0	2	100
	HS.G.GPE.B.7	0	1	100	0	2	100
	HS.G.MG.A.1	0	1	100	0	2	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.Q.A.2	0	1	100	0	2	100
	HS.N.Q.A.3	0	1	100	0	2	100
	HS.N.RN.A.1	0	1	100	0	2	100
	HS.N.RN.A.2	0	1	100	0	2	100
Statistics and	Overall	1	1	100	1	2	100
	HS.S.CP.B.6	0	1	100	0	1	100
	HS.S.ID.A.1	0	1	100	0	1	100

Strand	Benchmark	Grade 11					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
Probability (S)	HS.S.ID.A.2	0	1	100	0	2	100
	HS.S.ID.C.7	0	1	100	0	2	100

Table 58. 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	2	100
	5-PS|PS3|5-PS3-1	0	1	100	0	2	100
	PS4	0	1	100	1	2	88
	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

Strand	Benchmark	Grade 5					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
	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
	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 59. 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

Strand	Benchmark	Grade 8					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
	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
	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	2	4	100
	MS-LS|LS4|MS-LS4-1	0	1	100	0	2	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

Strand	Benchmark	Grade 8					
		Segment 1			Segment 2		
		Minimum Required Items	Maximum Required Items	% BP Match	Minimum Required Items	Maximum Required Items	% BP Match
	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 60. 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
	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	14	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	14	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

8.1.3. Item Development

Chapter 3, Item Development, provides a 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 templates when creating items. All items undergo multiple rigorous rounds of internal and external reviews from the content and fairness perspective before they are field-tested in an operational context. Items are created, edited, and reviewed amongst content and special education assessment experts who work together to produce the product that is sent to HIDOE for final review, edits, and approval. Item writing teams hold multiple on-going training and feedback sessions so all item writers and reviewers can learn best practices, client preferences, and continue to improve the quality of items. 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.

8.1.4. 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 test administrators (TAs) do not follow the same procedures, student performance cannot be meaningfully compared. For the 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). Refer to Chapter 5, Test Administration, for details.

8.1.5. Item and Test Scoring

Item and test scores are critical elements. All interpretations are established around students' test results. Every effort is made to ensure absolute accuracy on item and test scores. Section 12.3, Quality Assurance in Test Scoring, provides a detailed description on quality control and monitoring procedures implemented within CAI to assure accurate scores are generated and reported.

8.2 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. For students with the most significant cognitive disabilities, the Every Student Succeeds Act (ESSA) places a one-percent maximum 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 an 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.

Cog lab studies conducted in Hawai`i in spring 2019 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. 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. A brief description of the cog lab studies is provided below.

Study Sample

Students with significant cognitive disabilities at all grade levels and at each of three cognitive levels (low, moderate, and high ability) were included, with four-to-five students 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 Individuals with Disabilities Education Act (IDEA) disability categories with the greatest number of students in the state's significantly cognitively disabled student population, intellectual disability, autism, and multiple disabilities.

Items Selected

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 Next Generation Science Standards (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. Some 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 successfully respond to items that were at a cognitive level that came close to matching the grade-level standard expectations.

Data Collection

The data for these studies were obtained from three sources: (1) student behaviors while responding to each item, (2) student oral responses to questions that asked them to reflect on how they answered each item, and (3) 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). 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.

Findings

The evidence and insights gained from the cog lab studies supported Hawai‘i’s validity argument that the HSA-Alt is eliciting the intended cognitive response inherent in the grade-level CCSS and NGSS as mediated by the HSA-Alt Range PLDs. Students were challenged by many of the items but were able to apply some of the skills that they had learned in the classroom to answer test items successfully. Insights gained through the critical analysis of off-target student responses resulted in several completed and planned initiatives. An updated style guide and test specifications that included the consideration of language complexity, vocabulary, and audio and visual supports were created by the multi-state collaborative.

8.3 Evidence Based on Internal Structure

The measurement and reporting model used in the HSA-Alt 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 dissattenuated 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 61–Table 63 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. When the correlations are high, it suggests that the content strands within the subject essentially measure the same construct.

Table 61. Correlations Among Strand Scores for ELA

Grade	Strand	Observed & Disattenuated Correlation		
		Strand 1	**Strand 2**	**Strand 3**
3	Strand 1: Language	**0.41**	0.87	0.65
	Strand 2: RI and RL	0.45	**0.64**	0.80
	Strand 3: Writing	0.27	0.42	**0.42**
4	Strand 1: Language	**0.51**	0.93	0.82
	Strand 2: RI and RL	0.54	**0.68**	0.86
	Strand 3: Writing	0.41	0.50	**0.50**
5	Strand 1: Language	**0.44**	1.00	1.00
	Strand 2: RI and RL	0.61	**0.72**	1.00
	Strand 3: Writing	0.46	0.64	**0.45**
6	Strand 1: Language	**0.61**	0.91	0.93
	Strand 2: RI and RL	0.58	**0.68**	0.99
	Strand 3: Writing	0.51	0.57	**0.50**
7	Strand 1: Language	**0.59**	1.00	0.92
	Strand 2: RI and RL	0.69	**0.76**	0.84
	Strand 3: Writing	0.54	0.56	**0.58**
8	Strand 1: Language	**0.52**	0.84	0.96
	Strand 2: RI and RL	0.52	**0.73**	1.00
	Strand 3: Writing	0.42	0.52	**0.37**
11	Strand 1: Language	**0.62**	0.90	0.95
	Strand 2: RI and RL	0.59	**0.68**	0.87
	Strand 3: Writing	0.52	0.50	**0.48**

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

Table 62. Correlations Among Strand Scores for Mathematics

Grade	**Strand**	**Observed & Disattenuated Correlation**		
		Strand 1	**Strand 2**	**Strand 3**
3	Strand 1: Measurement and Data & Geometry	**0.67**	0.88	1.00
	Strand 2: Number and Operations - Fractions	0.50	**0.49**	0.92
	Strand 3: OA & NBT	0.70	0.50	**0.60**
4	Strand 1: Measurement and Data & Geometry	**0.52**	0.68	0.95
	Strand 2: Number and Operations - Fractions	0.37	**0.58**	0.81
	Strand 3: OA & NBT	0.55	0.49	**0.65**
5	Strand 1: Measurement and Data & Geometry	**0.44**	0.93	1.00
	Strand 2: Number and Operations - Fractions	0.44	**0.51**	0.55
	Strand 3: OA & NBT	0.61	0.30	**0.60**
6	Strand 1: NS & EE	**0.54**	1.00	0.77
	Strand 2: RP & G	0.55	**0.52**	0.91
	Strand 3: Statistics and Probability	0.32	0.37	**0.31**
7	Strand 1: NS & EE	**0.45**	1.00	1.00
	Strand 2: RP & G	0.63	**0.58**	1.00
	Strand 3: Statistics and Probability	0.53	0.49	**0.28**
8	Strand 1: Functions & Statistics and Probability	**0.21**	0.88	1.00
	Strand 2: Geometry	0.26	**0.42**	-0.10
	Strand 3: NS & EE	0.38	-0.05	**0.68**
11	Strand 1: Functions & Statistics and Probability	**0.48**	0.97	0.90
	Strand 2: Geometry	0.44	**0.43**	0.80
	Strand 3: Number Quantity & Algebra	0.39	0.33	**0.39**

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 63. Correlations Among Strand Scores for Science

Grade	**Strand**	**Observed & Disattenuated Correlation**		
		Strand 1	**Strand 2**	**Strand 3**
5	Strand 1: Earth & Space Science	**0.49**	1.00	1.00
	Strand 2: Life Science	0.66	**0.58**	1.00
	Strand 3: Physical Science	0.66	0.71	**0.61**
8	Strand 1: Earth & Space Science	**0.57**	1.00	1.00
	Strand 2: Life Science	0.67	**0.63**	1.00
	Strand 3: Physical Science	0.63	0.65	**0.64**
11	Strand 1: Life Science	**0.43**	0.99	1.00
	Strand 2: Ecosystems: Interactions, Energy and Dynamics	0.49	**0.58**	0.95
	Strand 3: Heredity and Biological Evolution	0.57	0.60	**0.70**

The between-subject correlations are presented in Table 64. 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 0.37 in grade 11 between mathematics and science to the highest of 0.87 in grade 5 between ELA and mathematics.

Table 64. Correlations Among Subject Scale Scores

Grade	Subject	ELA	Mathematics	Science
3	ELA	**0.75**	0.77	
	Mathematics	0.59	**0.79**	
	Science			
4	ELA	**0.79**	0.83	
	Mathematics	0.66	**0.80**	
	Science			
5	ELA	**0.84**	0.87	0.81
	Mathematics	0.68	**0.72**	0.85
	Science	0.67	0.65	**0.81**
6	ELA	**0.83**	0.65	
	Mathematics	0.51	**0.74**	
	Science			
7	ELA	**0.85**	0.80	
	Mathematics	0.64	**0.75**	
	Science			
8	ELA	**0.81**	0.66	0.82
	Mathematics	0.49	**0.68**	0.75
	Science	0.67	0.56	**0.83**
11	ELA	**0.82**	0.47	0.83
	Mathematics	0.35	**0.67**	0.37
	Science	0.68	0.27	**0.81**

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 65 presents the comparison of between-subject disattenuated correlations with the average disattenuated between-strand correlations within the same subject for each grade. For both ELA and science, the average between-strand correlations within each subject are either equal to or higher than the corresponding between-subject correlations. For mathematics, the same pattern is observed in grades 3, 6, 7, and 11. The largest difference happens in grade 8 mathematics where the average between-strand correlation of 0.59 is smaller than the correlation of 0.66 between ELA and mathematics, and 0.75 between mathematics and science, probably due to the low correlation between strands of Geometry and NS & EE (-0.10).

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 65. 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	0.77			0.77	0.93	
4	0.83			0.87	0.81	
5	0.87	0.81	0.85	1.00	0.83	1.00
6	0.65			0.94	0.89	
7	0.80			0.92	1.00	
8	0.66	0.82	0.75	0.93	0.59	1.00
11	0.47	0.83	0.37	0.91	0.89	0.98

8.4 Evidence Based on Relations to Other Variables

The peer review guide (U.S. Department of Education, 2018) lays out the expectation that "the state's assessment scores are related as expected with other variables." This can be demonstrated through the results of a correlational study between assessment results or student test scores and variables related to test takers. HIDOE and 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 are related to variables of student behaviors that might directly impact student performance on the alternate assessment; 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 in this section following a discussion of the purpose and questions extracted from the LCI, and the purpose and questions from the HIORA.

8.4.1. Learner Characteristics Inventory

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 USDE. "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 comprises 22 questions that 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 following are the 22 questions:

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 if other than English
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 students who are classified as having significant cognitive disabilities. 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 the following:

- 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. The following are the three levels of descriptors:

1. The first, or highest-level, descriptor states that the student uses symbolic language to communicate.
2. The second, or middle-level, descriptor states that the student uses intentional communication but not at a symbolic level.
3. 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 symbolically communicate 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 the following four levels of descriptors:

1. The first, or highest, descriptor states that the student can independently follow one-to-two step directions presented through words without additional cues.
2. The second descriptor states that the student can follow one-to-two step directions with additional cues.
3. 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.
4. 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 independently respond 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 has the following four descriptive statements:

1. The first, or highest, states that the student can initiate and sustain social interactions.
2. The second descriptor describes the student as responding but not initiating social interactions.
3. The third descriptor defines a student who alerts to others.
4. 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 than 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
- 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 "student's 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
- 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.

8.4.2. Hawai`i Observational Rating Assessment

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 PLDs were 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 the following four rating levels to choose from:

1. Minimal Understanding
2. Partial or Inconsistent Understanding
3. Adequate Understanding
4. 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 in this section.

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 the following:

- 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 the following:

- 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 the following:

- 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)

The following four success predictors are for students in 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.
    - 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 or 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)

The following three success predictors are for students in grades 7–8 and 11:

1. 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.

2. 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.

3. 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?

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

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

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

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

  - o Less than three months
  - o Six months to three months
  - o One year to seven months
  - o More than one year

8.4.3. 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 66, 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 ability to engage with others and students’ mathematics scores (-0.05) in grade 8; the highest correlations was between the composite and students’ ELA scores (0.51) in grade 5.

A teacher’s description of a student’s ability level, as required when completing the LCI, does moderately correlate 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 66. 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	123	0.46	0.32	0.23	0.28	0.35	0.41
4	112	0.41	0.15	0.20	0.32	0.41	0.37
5	101	0.51	0.35	0.31	0.39	0.46	0.43
6	109	0.36	0.31	0.17	0.28	0.25	0.29
7	130	0.41	0.32	0.23	0.20	0.37	0.40
8	96	0.40	0.20	0.31	0.12	0.36	0.45
11	119	0.46	0.30	0.35	0.18	0.49	0.40
Mathematics							
3	123	0.46	0.23	0.16	0.21	0.43	0.49
4	111	0.38	0.19	0.18	0.18	0.36	0.44
5	102	0.42	0.28	0.23	0.17	0.49	0.40
6	108	0.21	0.06	0.02	0.04	0.23	0.27
7	129	0.46	0.30	0.28	0.25	0.41	0.44
8	94	0.26	0.24	0.24	-0.05	0.27	0.27
11	118	0.34	0.28	0.25	-0.03	0.37	0.38
Science							
5	95	0.42	0.41	0.22	0.23	0.35	0.37
8	94	0.32	0.27	0.28	-0.02	0.33	0.31
11	118	0.45	0.27	0.41	0.21	0.48	0.35

Table 67 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 7 and 11 with all values larger than 0.3, except for item 7.

Table 68 represents the correlation between a teacher rating of each HIORA question and a student's overall scale score in mathematics. Items 1–5 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, and instruction time in the remaining 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. The correlations are the highest in grade 3 with all values larger than 0.3 except for items 1 and 7; the correlations are the lowest in grades 8 and 11, ranging from -0.03 to 0.19.

Table 69 represents the correlation between teacher rating of each HIORA question and student's overall scale score in science. Items 1–4 are questions related to different science content areas, and item 5 is the question related to instruction time. In general, correlations in science tend to be lower than that in ELA as well.

Table 67. Correlation Between HIORA and ELA Scale Score

	HIORA Question						
Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3	0.44	0.15	0.41	0.22	0.14	0.22	0.16
4	0.28	0.21	0.37	0.31	0.26	0.46	0.21
5	0.25	0.16	0.20	0.29	0.26	0.21	0.19
6	0.30	0.13	0.22	0.23	0.19	0.30	0.21
7	0.36	0.33	0.38	0.40	0.37	0.35	0.29
8	0.27	0.24	0.24	0.15	0.27	0.18	0.20
11	0.47	0.45	0.43	0.51	0.50	0.33	0.19

Table 68. Correlation Between HIORA and Mathematics Scale Score

	HIORA Question						
Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3	0.25	0.42	0.33	0.33	0.33	0.30	0.15
4	0.21	0.47	0.29	0.32	0.13		0.11
5	0.11	0.16	0.19	0.15	0.29	0.20	0.11
6	0.12	0.09	0.06	0.06	0.06		-0.02
7	0.27	0.30	0.23	0.17	0.29		0.16
8	-0.01	0.11	0.16	0.02	0.01	0.19	0.03
11	0.14	0.17	0.04	-0.03	-0.02		0.10

Table 69. Correlation Between HIORA and Science Scale Score

	HIORA Question				
Grade	**1**	**2**	**3**	**4**	**5**
5	0	0.07	0.17	0	0.09
8	0.18	0.04	0.22	0.15	0.37
11	0.28	0.09	0.17	0.02	0.13

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 skills 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.

8.5 SUMMARY

This chapter 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 ongoing process and validity evidence will continue to be accumulated and evaluated as more relevant data become available.

9. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard errors of measurement (SEMs). In Classical Test Theory (CTT), *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, SEM, and classification accuracy and consistency for each performance standard.

9.1 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 (CSEM), 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 CSEM 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 CATs, administered items vary among all students, so the SEM also can vary among students, which yields CSEM. The average CSEM 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 CSEM, the greater the accuracy of test scores.

Table 70 presents the marginal reliability coefficients and the average CSEM for the total scale scores, based on all completed tests, excluding the Early Stopping Rule (ESR) records.

Table 70. Marginal Reliability for ELA, Mathematics, and Science

Subject	Grade	Number of Operational Items	Marginal Reliability	Scale Score Mean	Scale Score SD	Average CSEM	Ratio of CSEM over SD
ELA	3	40	0.75	289.88	38.54	19.24	0.50
	4	40	0.79	302.88	25.63	11.66	0.45
	5	40	0.84	286.23	41.81	16.95	0.41
	6	40	0.83	288.63	41.98	17.35	0.41
	7	40	0.85	287.79	39.19	15.19	0.39
	8	40	0.81	279.37	35.74	15.40	0.43
	11	40	0.82	280.17	36.48	15.62	0.43
Mathematics	3	40	0.79	292.46	42.29	19.28	0.46
	4	40	0.80	297.97	41.07	18.54	0.45
	5	40	0.72	287.44	34.76	18.41	0.53
	6	40	0.74	276.82	48.52	24.76	0.51
	7	40	0.75	274.60	50.38	25.09	0.50
	8	40	0.68	279.41	35.31	20.11	0.57
	11	40	0.67	287.03	32.28	18.49	0.57
Science	5	40	0.81	276.95	49.89	21.57	0.43
	8	40	0.83	273.41	42.12	17.38	0.41
	11	40	0.81	277.08	45.45	19.82	0.44

9.2 STANDARD ERROR CURVES

Figure 8–Figure 10 present plots of the CSEM 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

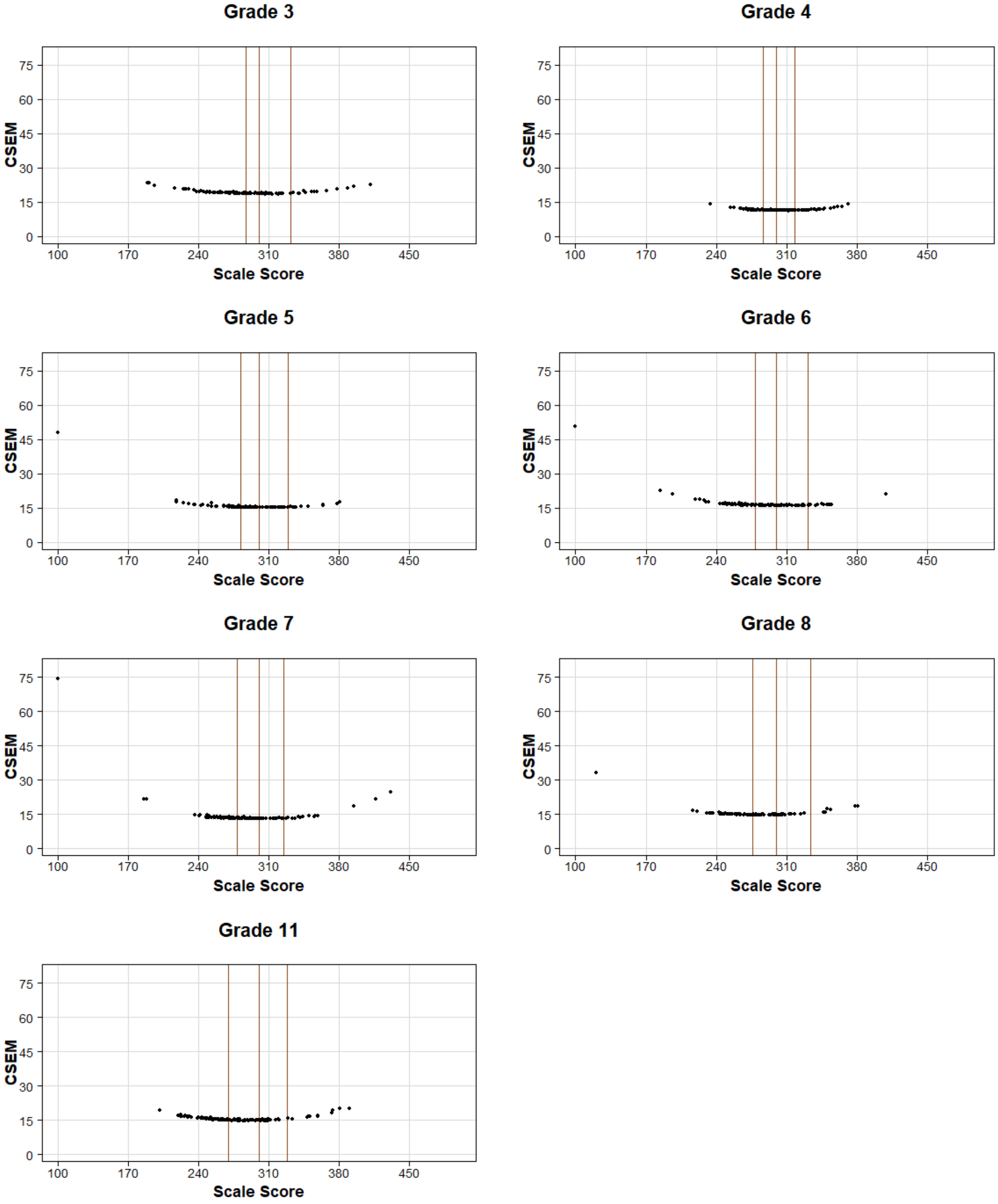
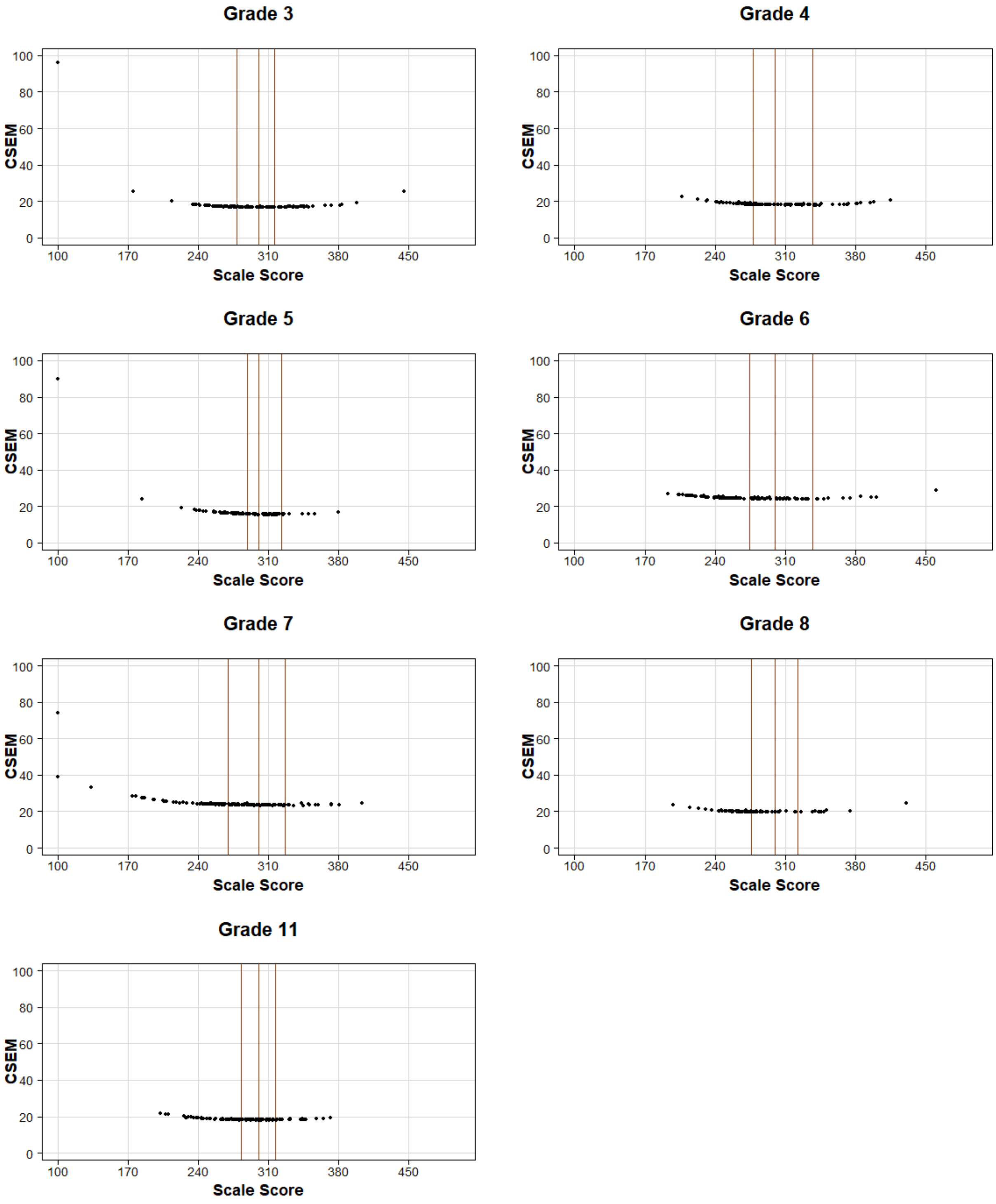


Figure 9. Conditional Standard Error of Measurement for Mathematics



Grade 3

Grade 4

Grade 5

Grade 6

Grade 7

Grade 8

Grade 11

Figure 10. Conditional Standard Error of Measurement for Science

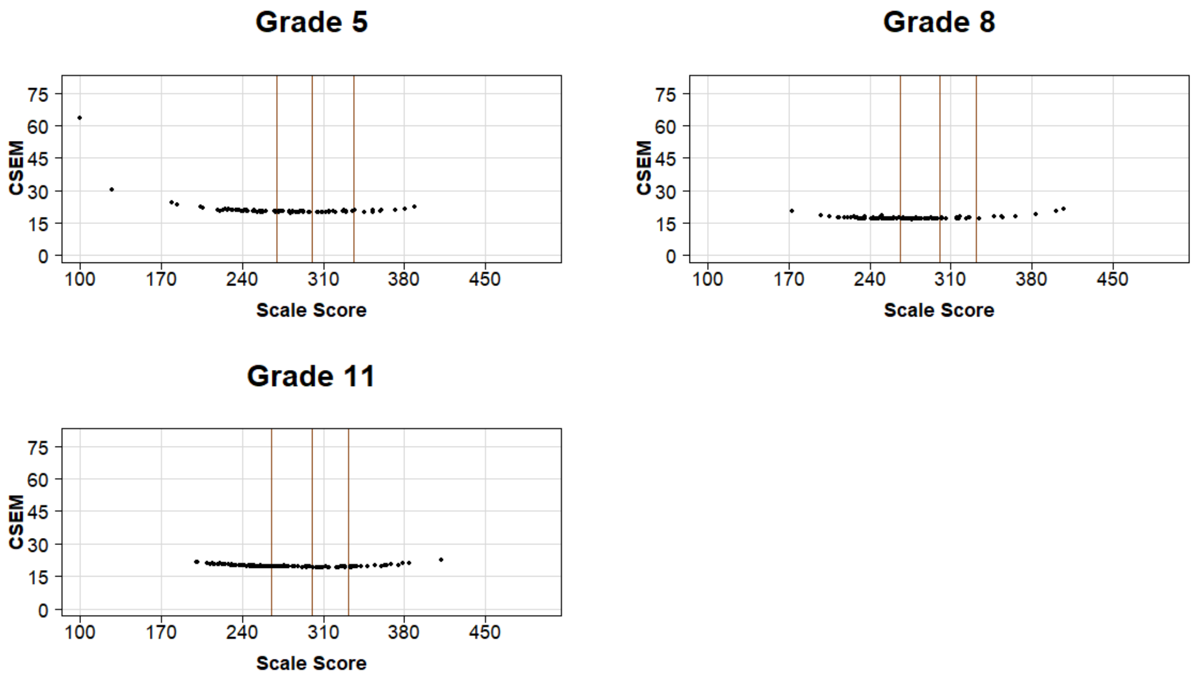


Table 71 presents the average CSEM for scores in each performance level. As shown in Figure 8–Figure 10, the average CSEMs in Approaches and Meets are similar, but slightly larger in Well Below and Exceeds, which can be expected for tests with extreme scores.

Table 71. Average Conditional Standard Error of Measurement by Performance Level

Subject	Grade	Well Below	Approaches	Meets	Exceeds	Average CSEM
ELA	3	19.46	18.85	18.74	19.88	19.24
	4	11.83	11.45	11.38	11.92	11.66
	5	18.72	15.42	15.33	16.01	16.95
	6	18.72	16.32	16.22	17.21	17.35
	7	17.15	13.19	13.06	15.41	15.19
	8	15.78	14.79	14.90	17.15	15.40
	11	15.82	15.01	15.06	17.52	15.62
Mathematics	3	22.45	16.87	16.87	17.52	19.28
	4	18.95	18.26	18.12	18.57	18.54
	5	20.52	15.74	15.66	15.84	18.41
	6	25.11	24.36	24.18	25.06	24.76
	7	26.89	23.70	23.47	23.63	25.09
	8	20.27	19.79	19.79	20.42	20.11
	11	18.83	18.21	18.13	18.50	18.49
Science	5	23.07	20.13	20.21	20.88	21.57
	8	17.43	17.03	17.30	18.75	17.38
	11	20.10	19.46	19.29	20.04	19.82

9.3 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 in this section. The true score is an expected value of the test score with a measurement error.

For the ith 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 ith 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 directly estimate the previously mentioned probabilities 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 score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut score. If a student with estimated theta is below the cut score, a probability of at or above the cut score is an estimate of the chance that this student is misclassified as below the cut score, 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 score (i.e., *cut*), the probability of the *i*th student being classified as at or above the cut score given the item scores $\mathbf{z}_i = \left(z_{i1}, \cdots, z_{iJ}\right)$ and item parameters $\mathbf{b} = \left(\mathbf{b}_1, \cdots, \mathbf{b}_J\right)$ 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; $\mathbf{b}_j = \left(b_j\right)$ 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 score when the *i*th student's performance level, pl_i, is below the cut score. 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 score is estimated by

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

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

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

and the overall classification accuracy for the cut score 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 independently administered twice 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 score, respectively, based on observed scores and hypothetical scores from an equivalent test form.

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

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

the classification consistency (*CC*) for the below the cut score 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 72 shows classification accuracy and consistency indexes for the spring 2024 HSA-Alt tests. Accuracy classifications are slightly higher than the consistency classifications for 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 72. Classification Accuracy and Consistency for Performance Standards

Subject	**Grade**	**Accuracy**			**Consistency**		
		Approaches	**Meets**	**Exceeds**	**Approaches**	**Meets**	**Exceeds**
ELA	3	0.83	0.84	0.94	0.77	0.77	0.90
	4	0.87	0.86	0.90	0.82	0.81	0.86
	5	0.83	0.87	0.94	0.77	0.82	0.91
	6	0.87	0.88	0.92	0.81	0.83	0.89
	7	0.84	0.91	0.95	0.78	0.86	0.93
	8	0.86	0.88	0.98	0.80	0.84	0.96
	11	0.88	0.89	0.97	0.82	0.84	0.95
Mathematics	3	0.85	0.87	0.90	0.80	0.82	0.86
	4	0.83	0.89	0.92	0.78	0.84	0.89
	5	0.85	0.87	0.90	0.78	0.80	0.86
	6	0.86	0.87	0.94	0.79	0.82	0.91
	7	0.84	0.86	0.91	0.77	0.80	0.87
	8	0.79	0.88	0.95	0.72	0.83	0.92
	11	0.83	0.82	0.89	0.76	0.76	0.84
Science	5	0.87	0.89	0.94	0.82	0.84	0.91
	8	0.85	0.92	0.97	0.79	0.88	0.95
	11	0.85	0.92	0.93	0.80	0.89	0.91

9.4 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 completed CATs only because the content of the items that were not administered in the incomplete CATs is unknown. Table 73–Table 75 show the reliability coefficients, scale score mean, scale score standard deviation (SD), and average CSEM for each strand.

Table 73. 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	10	0.41	292.55	60.39	45.25
	Reading—Informational & Literature	22	24	0.64	290.59	43.12	25.71
	Writing	8	9	0.42	282.27	61.14	45.29
4	Language	8	9	0.51	304.49	40.39	27.17
	Reading—Informational & Literature	22	23	0.68	302.97	28.14	15.98
	Writing	8	10	0.50	301.77	37.41	25.93
5	Language	8	10	0.44	287.40	49.83	36.32
	Reading—Informational & Literature	22	24	0.72	286.30	46.48	22.84
	Writing	8	10	0.45	285.22	48.57	34.93
6	Language	8	10	0.61	289.14	64.51	38.60
	Reading—Informational & Literature	21	24	0.68	289.28	43.64	23.85
	Writing	8	10	0.50	284.32	53.12	36.96
7	Language	8	10	0.59	286.31	52.98	32.43
	Reading—Informational & Literature	21	24	0.76	290.16	40.41	18.94
	Writing	8	10	0.58	283.48	51.65	31.94
8	Language	8	10	0.52	274.73	51.65	34.74
	Reading—Informational & Literature	21	23	0.73	281.15	40.91	21.13
	Writing	8	10	0.37	278.67	42.04	33.14
11	Language	8	9	0.62	281.85	67.43	39.76
	Reading—Informational & Literature	22	24	0.68	280.87	37.96	21.32
	Writing	8	10	0.48	276.67	44.74	32.04

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

Table 74. 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 TAC, 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 75. 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 TAC, scores for strands are not reported.

10. PERFORMANCE STANDARDS

Standard-setting workshops were held to establish performance standards (i.e., cut scores) for the HSA-Alt tests. The initial/original workshops were held during the first operational year. Later, if any updates were made to the test, follow-up confirmation standard-setting workshops were conducted to ensure that these changes did not impact the performance standards originally set during the initial workshop.

Table 76 outlines the original standard-setting workshops for the HSA-Alt and whether a confirmation standard-setting workshop was held for each subject.

Table 76. Original and Confirmation Standard-Setting Workshops of the HSA-Alt

Subject	Original Standard-Setting Workshop	Confirmation Standard-Setting Workshop
ELA	Summer of 2019	n/a
Mathematics	Summer of 2019	Summer of 2023 (In response to Changes to essence statements)
Science	Summer of 2021	Summer of 2023 (In response to Changes to essence statements)

Details of the original and confirmation standard-setting workshops for the HSA-Alt are described as follows.

Original Standard-Setting Workshops

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 tests. 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 science tests. 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 ordered-item booklets (OIBs) and edited some of the Performance-Level Descriptors (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 HSA-Alt in mathematics and science.

After reviewing changes in the Range PLDs, creating Threshold PLDs, and reviewing OIBs of the HSA-Alt mathematics and science tests, panelists came to a consensus for all grades in mathematics (3–8 and 11) and science (5, 8, and 11), that the existing performance standards still accurately classified 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 HSA-Alt confirmation standard-setting technical report (2023).

10.1 Standard-Setting Procedures

Hawai`i uses 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 OIB for each test. Each OIB contains a set of items that meet the test blueprint. The panelists also review the corresponding Hawai`i content standards and HSA-Alt Essence Statements and Range PLDs for each test. With this information in mind, the panelists select pages in the OIB that best represent the cut scores on the test. The Bookmark standard-setting process is described in a standard-setting plan submitted to HIDOE. The plan is reviewed by the Hawai`i Technical Advisory Committee (TAC) and approved by HIDOE prior to the workshop.

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

10.2 Performance-Level Descriptors

HSA-Alt item development is 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 (HCCS) 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 HCCS to the performance expectations for the test (Essence Statements). The following are the three types of PLDs used within the HSA-Alt program:

1. **Policy PLDs.** Policy PLDs provide a 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 *Meets* describes what students know and can do at that level all the way to just below the *Exceeds* 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 use the Essence Statements, Range PLDs, and Just Barely PLDs during the standard-setting workshop.

10.3 Recommended Performance Standards

Panelists are tasked with recommending three performance standards (Approaches, Meets, and Exceeds) that resulted in four performance levels (Well below, Approaches, Meets, and Exceeds). Table 77 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 77. 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.

11. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that include 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 are immediately generated for ELA and mathematics; starting in spring 2022, online score reports are immediately generated for science. Because the performance score report is updated each time a student completes a test, authorized users (e.g., school principals, teachers) can access timely information on students' performance scores to evaluate the effectiveness of instructional approaches and inform future educational planning.. 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.

11.1 Centralized Reporting System for Students and Educators

11.1.1. 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 (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 HSA-Alt has been designed with stakeholders who are not technical measurement experts in mind, with the intention to make score reports 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 performance 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 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 78 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 78. 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 PLDs • 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 test results by any of the subgroups. Average scale score and performance levels will be calculated at n ≥ 2. Table 79 presents the types of subgroups and subgroup categories provided in the CRS.

Table 79. Types of Subgroups

Subgroup	Subgroup Category
Gender	• Male • Female
English Learner (ELL	• ELL • Not ELL
*Disability	• With Disability • No Disability
Migrant Status	• Migrant • Not Migrant
Disadvantaged	• Disadvantaged • Not Disadvantaged
Ethnicity	• American Indian/Alaska Native • Asian/Pacific Islander • African American • Hispanic • Hawaiian Pacific Islander • White • Multi-Racial

* Available in CRS as a standard filter but not applicable to Alt students.

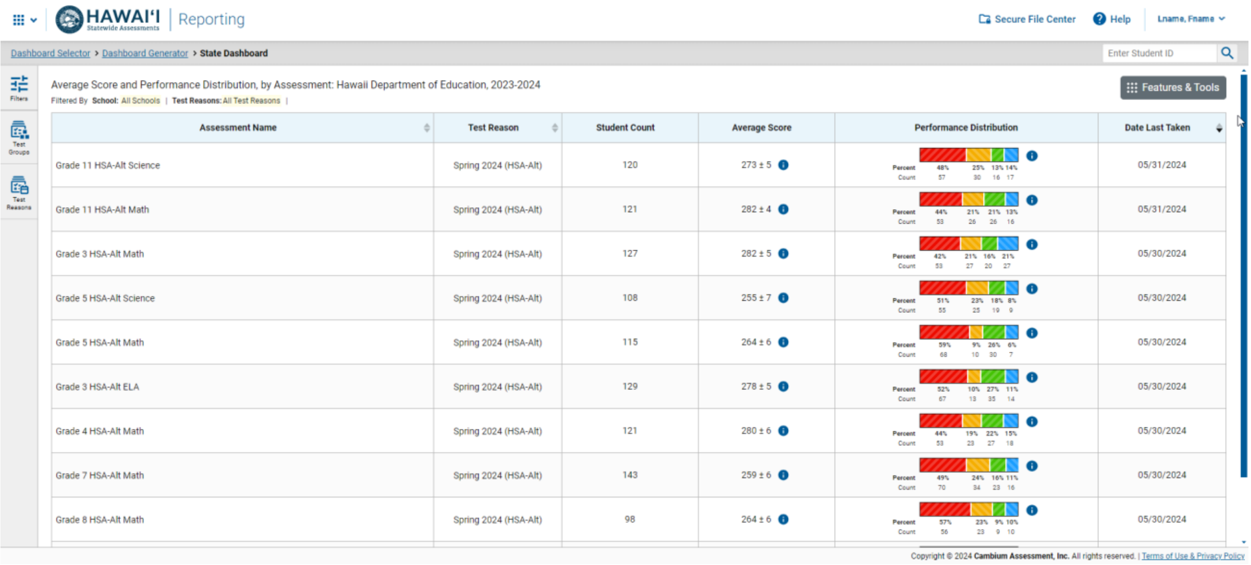
11.1.2. Centralized Reporting System

11.1.2.1 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 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



Average Score and Performance Distribution, by Assessment: Hawaii Department of Education, 2023-2024

Filtered By **School:** All Schools | **Test Reasons:** All Test Reasons |

Assessment Name	Test Reason	Student Count	Average Score	Performance Distribution	Date Last Taken
Grade 11 HSA-Alt Science	Spring 2024 (HSA-Alt)	120	273 ± 5	Percent 48% 25% 13% 14% Count 57 30 16 17	05/31/2024
Grade 11 HSA-Alt Math	Spring 2024 (HSA-Alt)	121	282 ± 4	Percent 44% 21% 21% 13% Count 53 26 26 16	05/31/2024
Grade 3 HSA-Alt Math	Spring 2024 (HSA-Alt)	127	282 ± 5	Percent 42% 21% 16% 21% Count 53 27 20 27	05/30/2024
Grade 5 HSA-Alt Science	Spring 2024 (HSA-Alt)	108	255 ± 7	Percent 51% 23% 18% 8% Count 55 25 19 9	05/30/2024
Grade 5 HSA-Alt Math	Spring 2024 (HSA-Alt)	115	264 ± 6	Percent 59% 9% 26% 6% Count 68 10 30 7	05/30/2024
Grade 3 HSA-Alt ELA	Spring 2024 (HSA-Alt)	129	278 ± 5	Percent 52% 10% 27% 11% Count 67 13 35 14	05/30/2024
Grade 4 HSA-Alt Math	Spring 2024 (HSA-Alt)	121	280 ± 6	Percent 44% 19% 22% 15% Count 53 23 27 18	05/30/2024
Grade 7 HSA-Alt Math	Spring 2024 (HSA-Alt)	143	259 ± 6	Percent 49% 24% 16% 11% Count 70 34 23 16	05/30/2024
Grade 8 HSA-Alt Math	Spring 2024 (HSA-Alt)	98	264 ± 6	Percent 57% 23% 9% 10% Count 56 23 9 10	05/30/2024

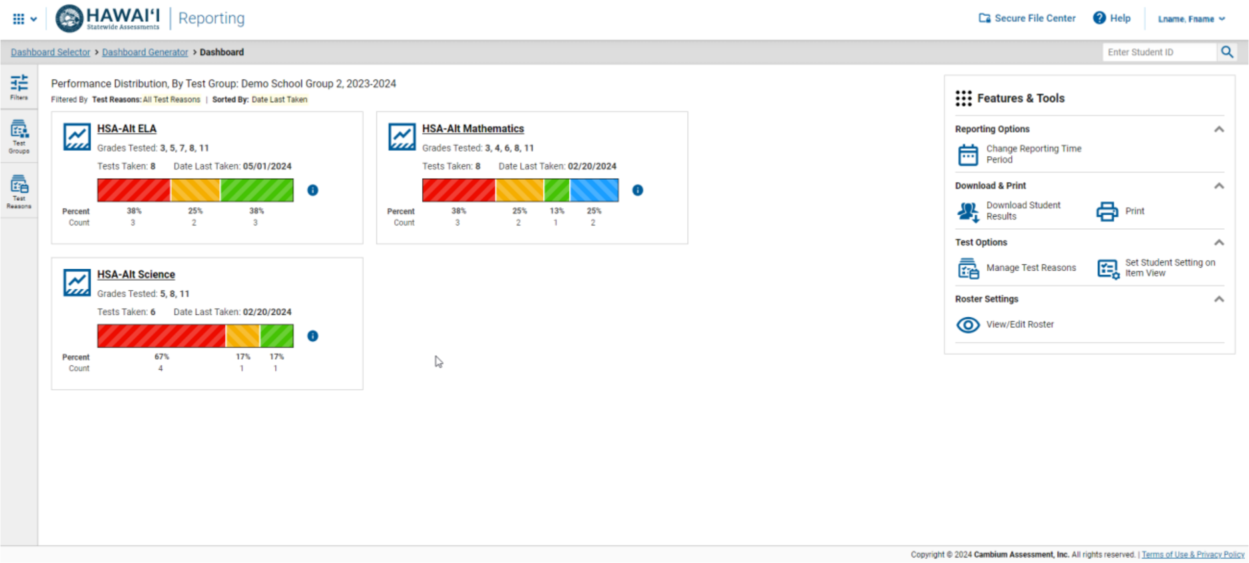
Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

The four performance 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 the subject group to view individual test results for the selected test group. Once the user clicks 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 the HSA-Alt at the complex-area level.

Exhibit 2. Dashboard: Complex-Area Level



Performance Distribution, By Test Group: Demo School Group 2, 2023-2024
Filtered By **Test Reasons:** All Test Reasons | **Sorted By:** Date Last Taken

HSA-Alt ELA
Grades Tested: **3, 5, 7, 8, 11**
Tests Taken: **8** Date Last Taken: **05/01/2024**

Percent	38%	25%	38%
Count	3	2	3

HSA-Alt Mathematics
Grades Tested: **3, 4, 6, 8, 11**
Tests Taken: **8** Date Last Taken: **02/20/2024**

Percent	38%	25%	13%	25%
Count	3	2	1	2

HSA-Alt Science
Grades Tested: **5, 8, 11**
Tests Taken: **6** Date Last Taken: **02/20/2024**

Percent	67%	17%	17%
Count	4	1	1

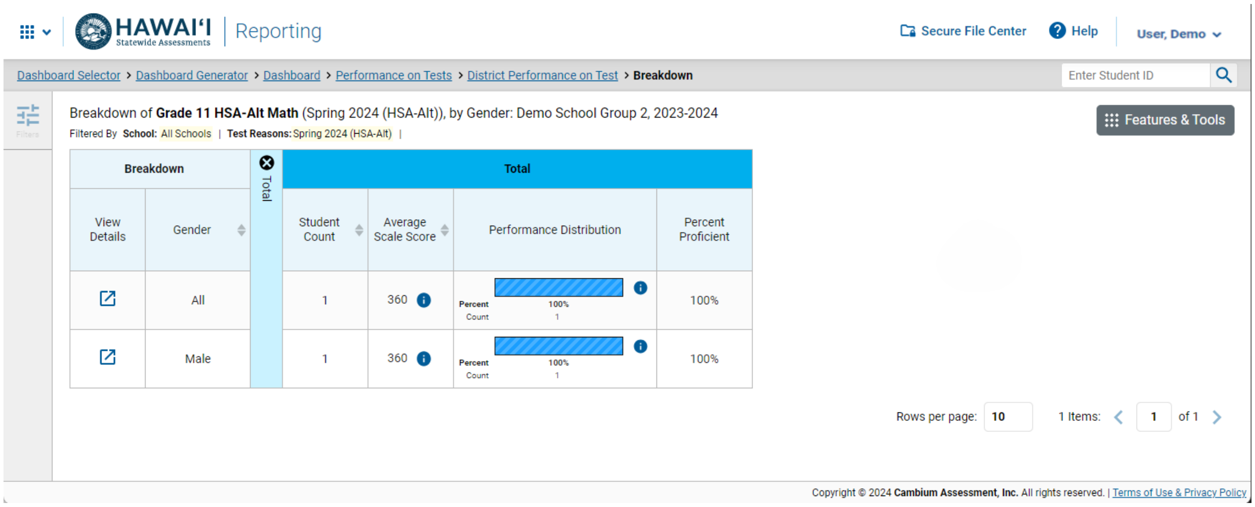
Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

11.1.2.2 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 performance level. The summaries are also presented for students overall and by subgroup. Exhibit 3 presents an example of subject summary results for grade 5 mathematics with gender breakdowns at the complex-area level.

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



Dashboard Selector > Dashboard Generator > Dashboard > Performance on Tests > District Performance on Test > **Breakdown**

Breakdown of **Grade 11 HSA-Alt Math** (Spring 2024 (HSA-Alt)), by Gender: Demo School Group 2, 2023-2024

Filtered By **School:** All Schools | **Test Reasons:** Spring 2024 (HSA-Alt) |

Breakdown		Total	Total			
View Details	Gender		Student Count	Average Scale Score	Performance Distribution	Percent Proficient
	All		1	360	Percent 100% Count 1	100%
	Male		1	360	Percent 100% Count 1	100%

Copyright © 2024 **Cambium Assessment, Inc.** All rights reserved. | Terms of Use & Privacy Policy

11.1.2.3 Student Detail Page

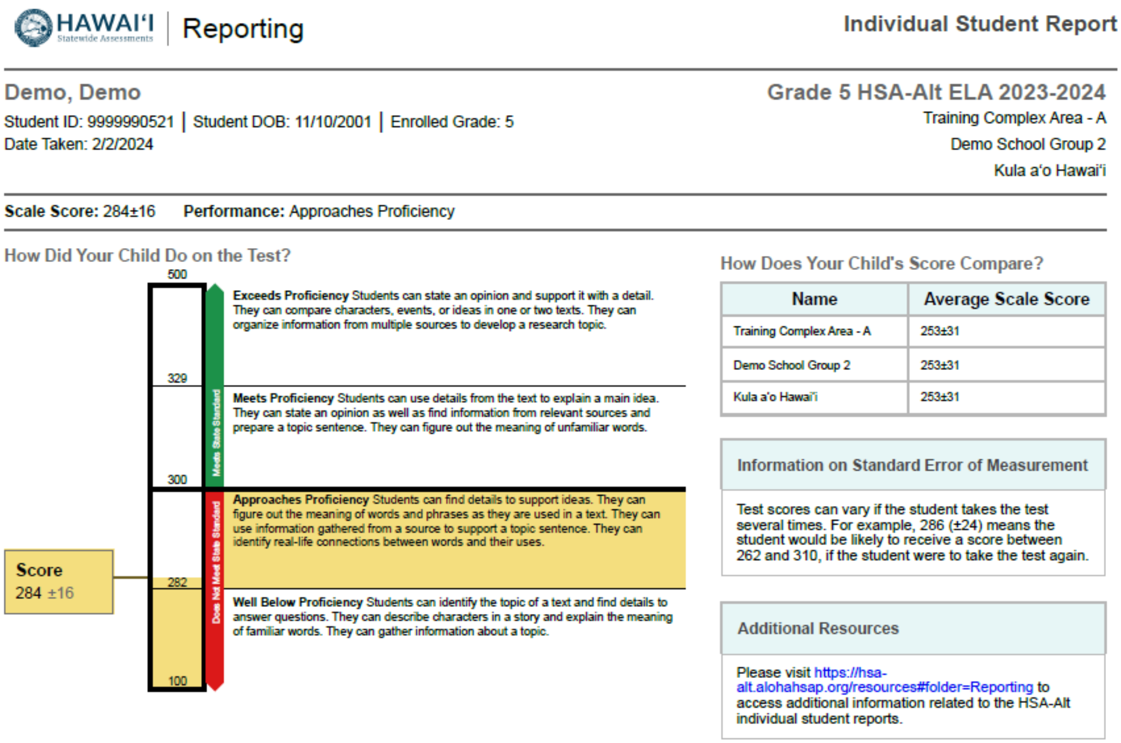
When a student completes a test, an online score report appears in the individual student report (ISR) in the CRS. The ISR shows individual student performance on the test. In each subject area, the ISR provides (1) the scale score and SEM; (2) performance level for overall test; and (4) average scale 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, PLDs with cut scores at each performance level are provided. This defines the content-area knowledge, skills, and processes that test takers at the performance 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 SEM 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 ISRs.

Exhibit 4. Student Detail Page for HSA-Alt ELA



HAWAIʻI Statewide Assessments | Reporting

Individual Student Report

Demo, Demo

Student ID: 9999990521 | Student DOB: 11/10/2001 | Enrolled Grade: 5
Date Taken: 2/2/2024

Grade 5 HSA-Alt ELA 2023-2024
Training Complex Area - A
Demo School Group 2
Kula a'o Hawai'i

Scale Score: 284±16 **Performance:** Approaches Proficiency

How Did Your Child Do on the Test?

Exceeds Proficiency Students can state an opinion and support it with a detail. They can compare characters, events, or ideas in one or two texts. They can organize information from multiple sources to develop a research topic.

Meets Proficiency Students can use details from the text to explain a main idea. They can state an opinion as well as find information from relevant sources and prepare a topic sentence. They can figure out the meaning of unfamiliar words.

Approaches Proficiency Students can find details to support ideas. They can figure out the meaning of words and phrases as they are used in a text. They can use information gathered from a source to support a topic sentence. They can identify real-life connections between words and their uses.

Well Below Proficiency Students can identify the topic of a text and find details to answer questions. They can describe characters in a story and explain the meaning of familiar words. They can gather information about a topic.

How Does Your Child's Score Compare?

Name	Average Scale Score
Training Complex Area - A	253±31
Demo School Group 2	253±31
Kula a'o Hawai'i	253±31

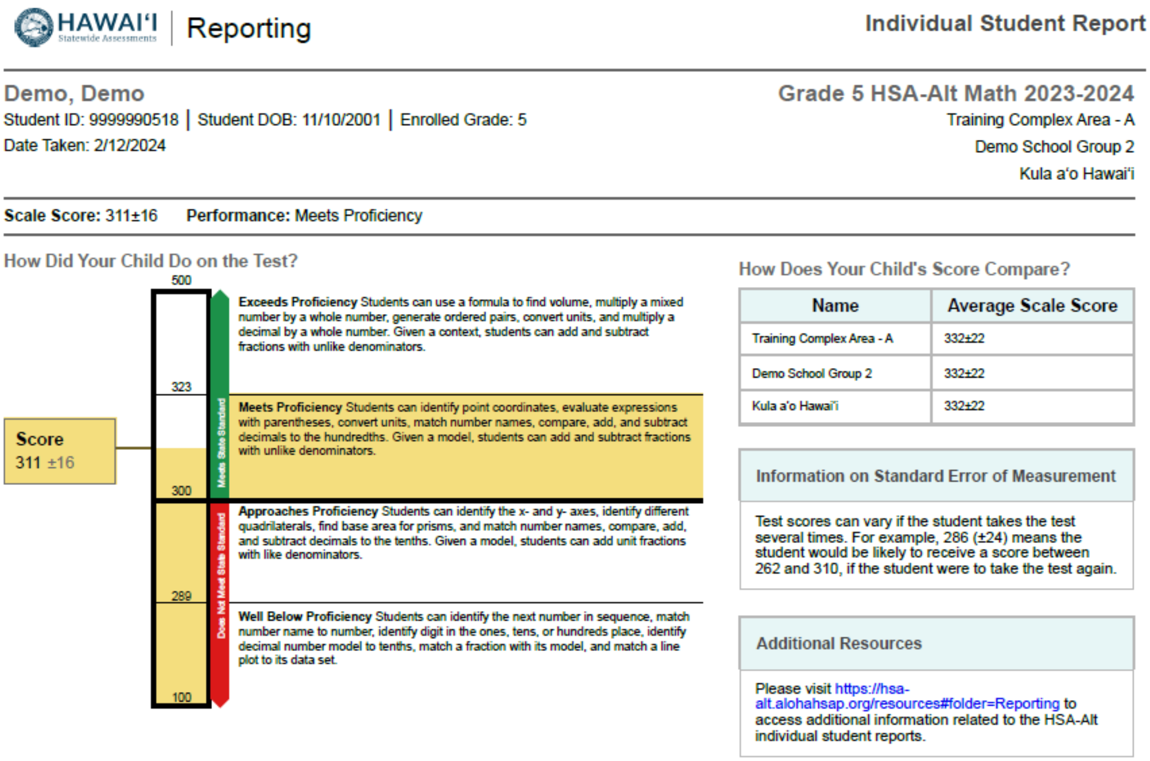
Information on Standard Error of Measurement

Test scores can vary if the student takes the test several times. For example, 286 (±24) means the student would be likely to receive a score between 262 and 310, if the student were to take the test again.

Additional Resources

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

Exhibit 5. Student Detail Page for HSA-Alt Mathematics



HAWAIʻI Statewide Assessments | Reporting

Individual Student Report

Demo, Demo

Student ID: 9999990518 | Student DOB: 11/10/2001 | Enrolled Grade: 5
Date Taken: 2/12/2024

Grade 5 HSA-Alt Math 2023-2024
Training Complex Area - A
Demo School Group 2
Kula a'o Hawai'i

Scale Score: 311±16 **Performance:** Meets Proficiency

How Did Your Child Do on the Test?

Exceeds Proficiency Students can use a formula to find volume, multiply a mixed number by a whole number, generate ordered pairs, convert units, and multiply a decimal by a whole number. Given a context, students can add and subtract fractions with unlike denominators.

Meets Proficiency Students can identify point coordinates, evaluate expressions with parentheses, convert units, match number names, compare, add, and subtract decimals to the hundredths. Given a model, students can add and subtract fractions with unlike denominators.

Approaches Proficiency Students can identify the x- and y- axes, identify different quadrilaterals, find base area for prisms, and match number names, compare, add, and subtract decimals to the tenths. Given a model, students can add unit fractions with like denominators.

Well Below Proficiency Students can identify the next number in sequence, match number name to number, identify digit in the ones, tens, or hundreds place, identify decimal number model to tenths, match a fraction with its model, and match a line plot to its data set.

How Does Your Child's Score Compare?

Name	Average Scale Score
Training Complex Area - A	332±22
Demo School Group 2	332±22
Kula a'o Hawai'i	332±22

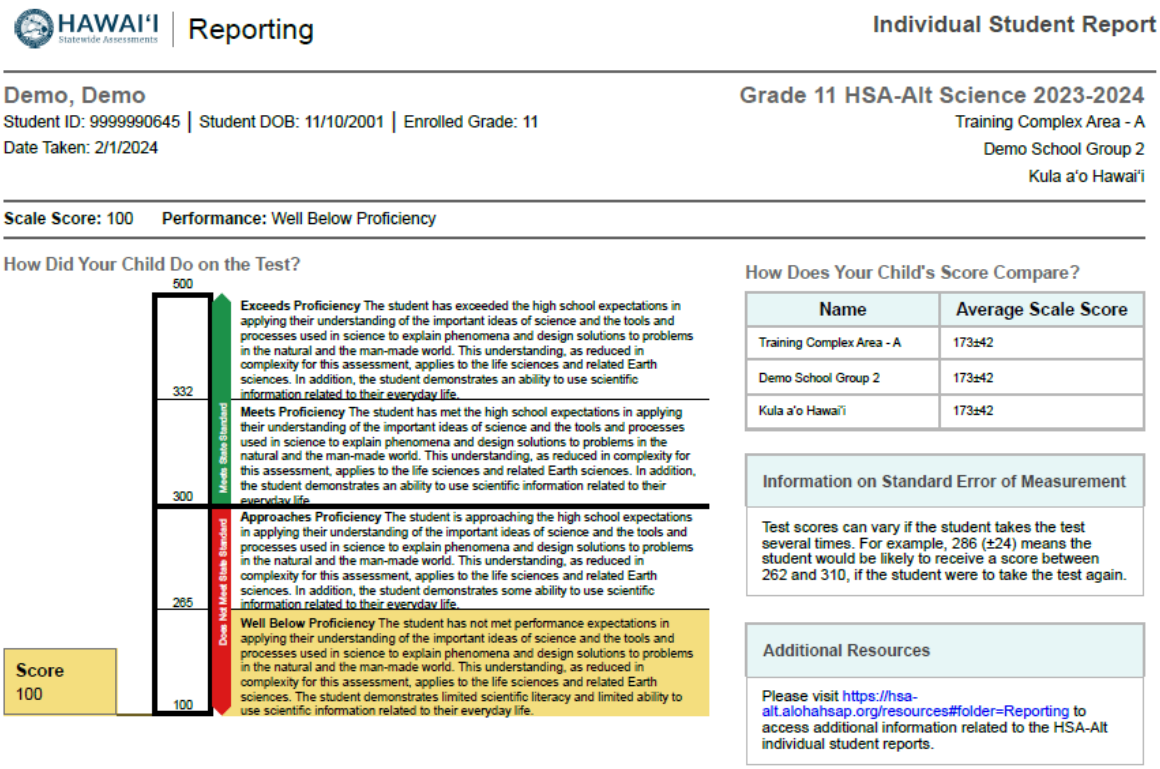
Information on Standard Error of Measurement

Test scores can vary if the student takes the test several times. For example, 286 (±24) means the student would be likely to receive a score between 262 and 310, if the student were to take the test again.

Additional Resources

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

Exhibit 6. Student Detail Page for HSA-Alt Science



Reporting

Individual Student Report

Demo, Demo

Student ID: 9999990645 | Student DOB: 11/10/2001 | Enrolled Grade: 11

Date Taken: 2/1/2024

Grade 11 HSA-Alt Science 2023-2024

Training Complex Area - A

Demo School Group 2

Kula a'o Hawai'i

Scale Score: 100 **Performance:** Well Below Proficiency

How Did Your Child Do on the Test?

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.

Score
100

How Does Your Child's Score Compare?

Name	Average Scale Score
Training Complex Area - A	173±42
Demo School Group 2	173±42
Kula a'o Hawai'i	173±42

Information on Standard Error of Measurement

Test scores can vary if the student takes the test several times. For example, 286 (±24) means the student would be likely to receive a score between 262 and 310, if the student were to take the test again.

Additional Resources

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

11.1.3. 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.

11.1.4. 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. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and PLDs.

11.1.5. 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 will 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.

11.1.6. Performance Level

Performance levels are proficiency categories on a test that students fall into based on their scale scores. For the HSA-Alt, 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 refer to Section 10, 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.

11.1.7. Aggregated Score

Student scale scores are aggregated at roster, teacher, school, complex, complex-area, and state levels to represent how a group of students perform 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 perform overall.

11.2 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. The ISR may provide more useful information for a student's learning and teaching, as it considers the diverse needs of the student's significant cognitive disability/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 perform 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 reported scale scores 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.

12. QUALITY CONTROL PROCEDURES

Quality control 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 thoroughly tested before, during, and after the testing window opens.

12.1 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 that are 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. 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 independently checked, following the scoring rules detailed in the scoring specifications.

12.1.1. 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 across platforms. For the HSA-Alt, there are two commonly used layouts: (1) the stimulus and item response options/response area are displayed side by side, where stimulus and response options have independent scroll bars; and (2) the item stem and responses appear 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.

A team conducts Platform Review; the team leader projects the item as it is 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.

12.1.2. 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 HIDOE with an opportunity to interact with the exact test that the students will use.

12.2 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 Quality Monitor (QM) System. The QM System 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 QM System 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 HIDOE. CAI staff ensures that data in the extract files match the DOR before delivering them to HIDOE.

12.3 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 testing 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 instantly know 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 the schools or districts 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 quality assurance reports can be generated on any desired schedule. Item analysis and blueprint match reports are frequently evaluated 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; reports can also be generated based on all items in the pool.

Table 80 presents an overview of the quality assurance reports. No significant QA issues were flagged during the spring 2024 administration.

Table 80. 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

12.4 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 QM 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 quality assurance 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 QM 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 procedure: 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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment**.** National Academy Press.

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.