Hawai‘i State End-of-Course Exams 2022–2023 Technical Report

Algebra 1 and Algebra 2

DRAFT

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

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

APPENDICES

1. OVERVIEW

Hawai‘i is a governing state of the Smarter Balanced Assessment Consortium (SBAC) and is committed to implementing the Smarter Balanced assessments that are aligned to the Common Core State Standards (CCSS) for English language arts/literacy (ELA/L) and mathematics beginning in school year (SY) 2015–2016. In June 2010, the Hawai‘i Board of Education adopted the CCSS. These standards are rigorous college and career readiness (CCR) standards in ELA/L and mathematics. The Hawai‘i Department of Education (HIDOE) has also received federal approval by the U.S. Department of Education (ED) for a new Strive Hawai‘i (Strive HI) Performance System. Strive HI replaces many of the requirements of the No Child Left Behind (NCLB) Act, strengthening the educational standards to an expectation that all high school students will be college and career ready when they graduate. Consistent with HIDOE's expectations for CCR standards, the *Meets Proficiency* standard for each End-of-Course (EOC) exam was developed to match the CCR standards.

EOC exams are statewide summative tests administered at the end of a course. Starting in SY2015–2016, Algebra 1 and Algebra 2 EOC exams were classified as optional for public school students who are enrolled in the corresponding course.

The purpose of an EOC exam is to measure students' proficiency in course content standards, inform instruction, and standardize course expectations (as required by the Race to the Top grant and NCLB). The EOC exams measure student proficiency in the standards and benchmarks assigned to each course. The Algebra 1 and Algebra 2 EOC exams measure student proficiency in the CCSS. Teachers may use EOC exam results as one factor of up to 15% when determining a student's final grade for a course.

EOC exams are administered online using the same system as the Hawai‘i State Science (NGSS) Assessments for Science and EOC for Biology 1. These tests are administered at the end of instruction during the last three weeks of the fall testing window and the last five weeks of the spring testing window for the course. Students can take the test once within the three-week testing window. (A second opportunity is allowed on a case-by-case basis.)

Each exam has approximately 43–45 questions aligned to the content standards assigned to the course. The exams are untimed, and tests may be paused by either the student or the test administrator (TA) and completed later within the testing window. A student whose exam is paused for more than 20 minutes will not be allowed to review questions answered during the previous test session.

The EOC exams include selected-response and constructed-response questions that are machine-scored. When a student completes an exam, the student's score is displayed on the monitor and access to the student's score report is available to the student's teacher via the Centralized Reporting System (CRS) already in use. Starting in 2013–2014, all EOC exams were administered adaptively. Schools with a block schedule administered the exams in the fall and spring, and schools with a year-long schedule administered the exams only in the spring. Schools that held summer school courses in Algebra 1 and Algebra 2 had the option to administer EOC exams to their attending students.

In the 2019–2020 school year, the ED granted a waiver from testing requirements due to the COVID-19 pandemic (https://www2.ed.gov/policy/gen/guid/secletter/200320.html). In 2020–2021, the ED did not grant waivers for standardized testing, but did waive certain accountability requirements (e.g., mandatory high participation rates) for the 2020–2021 school year due to the impacts of the pandemic in many states, resulting in lower participation rates than in previous years.

The American Institutes for Research delivered the Algebra 1 and Algebra 2 EOC exams through the 2018–2019 school year. Starting with the 2020–2021 school year, Cambium Assessment (CAI) delivered and scored the EOC exams and produced score reports.

This technical report describes and summarizes the test development, test administration, and statistical and psychometric analyses that are performed on the 2022–2023 Algebra 1 and Algebra 2 EOC exams. The report includes the following sections:

- *Test Development* summarizes test specifications, test blueprints, and the item development process.
- *Test Administration* describes test administration features, TA training, security procedures, test accommodations, and the data forensics program.
- *Summary of Simulation Studies* summarizes the adaptive algorithm and the simulation results.
- *Maintenance of the Item Bank* includes information about item release, item calibration, and scaling.
- *Summary of 2022–2023 Operational Test Administration* summarizes student performance overall and by subgroup, and the student ability-operational item difficulty distribution.
- *Validity* provides the validity evidence of the 2022–2023 EOC exams.
- *Reliability* provides the reliability evidence of the 2022–2023 EOC exams.
- *Scoring* summarizes the scoring rules used in generating student test scores.
- *Reporting and Interpreting Scores* outlines the features of the ORS that stakeholders can use to help them understand and appropriately use the results of the EOC exams and describes how to interpret the reported scores.
- *Quality Control Procedures* summarizes the quality control procedures that are enforced before, during, and after the testing window.

2. TEST DEVELOPMENT

2.1 TEST SPECIFICATIONS

The Hawai‘i EOC test specifications represent the information provided in the CCSS for the Algebra 1 and Algebra 2 exams. Test specifications provide guidelines for item writers on the range of content that may be tested and how items must be written. These specifications lead to test blueprints that outline test design and the number of questions to be tested in each score reporting category.

2.2 TARGET BLUEPRINTS

Blueprints specify a range of items to be administered in each reporting category for the specific CCSS benchmarks assigned to each course. The target blueprints include the requirements for the total test length and the minimum and maximum number of operational items for each score reporting category that each test must include. Allowing a range in the number of required items gives the computer-adaptive testing (CAT) algorithm flexibility to select items that match the test blueprints as well as the ability of the student.

To ensure that the computer-adaptive EOC exams accurately reflect the content included in the CCSS benchmarks, the test blueprints require that the knowledge and skills specified in the CCSS for each reporting category be assessed on each exam. In each test, at least 50% of the CCSS standards benchmarks are assessed within each reporting category. In the aggregate, however, all the standards and benchmarks specified in the test blueprints are assessed. Providing the student performance on all benchmarks at an aggregate level is very beneficial for instructional purposes.

In addition to specifying the number of items to be administered at each reporting category, the blueprints also specify how many of a certain type of item should be administered. There are selected response items (i.e. multiple-choice and multi-select items), as well as machine-scored constructed-response (MSCR) items. These MSCR items may require the student to type an open-ended response composed of alpha numeric characters. There are also graphical MSCR items, which may require the student to draw or move images around to construct his or her response.

Each item is aligned to one of Norman Webb’s Depth of Knowledge (DOK) levels. These DOK levels represent the intended cognitive complexity for each item. The levels range from 1 to 4 as follows:

- Level 1 represents rote demonstration of understanding and is usually referred to as the “recall” level.
- Level 2 requires demonstration of skill and concepts or basic reasoning. Level 2 items may require a student to make a basic inference or apply a specific skill to solve a well-posed problem.
- Level 3 requires strategic thinking and complex reasoning. Items at this level are usually more unique and require application of skills through critiquing and explaining thoughts.
- Level 4 is called “extended thinking” or “reasoning” and usually requires a student to gather information, analyze the information, and apply the knowledge. Items at this level may require conducting an experiment of some sort over time.

Because the range of these levels should be assessed for each student exam, there are ranges for each of these levels documented on the blueprints for each course’s assessment.

The blueprints were initially drafted by a CAI assessment specialist in collaboration with HIDOE Student Assessment Section specialists. Content specialists from HIDOE's Office of Curriculum and Instructional Design (OCID) also had an opportunity to review and revise the blueprints. OCID provided a draft course framework that was used to help finalize the blueprints.

Tables 1–2 show the test blueprint requirements specified in the Test Delivery System (TDS) for the 2022–2023 operational tests. Each exam must include items within the range of the minimum and maximum number of items for the total exam and the score reporting categories.

Table 1. Number of Test Items Assessing Each Score Reporting Category in Algebra 1

Reporting Category/ Additional Constraints	Number of Standards	Total Number of Items	
		Min	Max
Total Test	27	43	43
Algebraic Concepts and Procedures	15	21	23
Modeling and Problem Solving	12	20	22
Additional Constraints			
Total MSCR Items		8	12
Total SR Items		28	38
DOK 1		4	9
DOK 2		33	39
DOK 3		2	3

Table 2. Number of Test Items Assessing Each Score Reporting Category in Algebra 2

Reporting Category/ Additional Constraints	Number of Standards	Total Number of Items	
		Min	Max
Total Test	57	45	45
Algebraic Concepts and Procedures	38	29	31
Modeling and Problem Solving	19	14	16
Additional Constraints			
Total MSCR Items		8	12
Total SR Items		31	39
DOK 1		2	6
DOK 2		35	41
DOK 3		2	4

2.3 SCORE REPORTING CATEGORIES

The Hawai‘i EOC exams are designed to assess the following reporting categories, which reflect the knowledge and skill expectations outlined in the CCSS.

Algebra 1

The Algebra 1 EOC exam is designed to assess the following reporting categories (standards):

- *Algebraic Concepts and Procedures:* Identify, apply, and solve linear and quadratic functions; describe the operations used to solve linear equations and inequalities and systems of equations and inequalities; and determine zeroes of quadratic functions.
- *Modeling and Problem Solving:* Create linear and quadratic equations and inequalities to model a variety of situations; interpret characteristics of graphical representations; define appropriate quantities for modeling; and fit a function to a data set.

Algebra 2

The Algebra 2 EOC exam is designed to assess the following reporting categories (standards):

- *Algebraic Concepts and Procedures:* Identify, apply, and solve polynomial, rational, radical, exponential, and logarithmic functions; describe the operations used to solve these functions; and understand the relationship between solving equations and graphing them.
- *Modeling and Problem Solving:* Create polynomial, rational, exponential, and logarithmic equations to model a variety of situations; interpret characteristics of graphical representations; interpret parameters within a context; and fit a function to a data set.

2.4 ITEM SPECIFICATIONS

The item specifications contain information about items used to assess each CCSS benchmark. The specifications are used by item writers and reviewers to ensure consistency in item development. Information about calculator usage, appropriate item types, rubric score points, suitable DOK, and content limits are presented for each CCSS benchmark to be assessed by each EOC exam.

CAI assessment specialists used their understanding of the CCSS, along with information provided by HIDOE's OCID, about each of the courses to create the detailed specifications. Once a draft was created, the specifications were reviewed by the Assessment Section of the HIDOE with input from OCID content specialists.

Once the draft of the item specifications was complete, item development began. Many times during the item development process (writing and reviewing), additional pertinent information was revealed. CAI and HIDOE worked together to update the specifications to reflect any relevant information that may clarify the items being developed for each CCSS benchmark.

Item Development Procedures

All items developed for the EOC exams were written and reviewed using the principles of universal design (UD). To provide equal access to the assessments for all students, including those with disabilities such as limited vision or learning disabilities, item writers used these principles when writing and reviewing items. Although some concepts may have to be tested using complex graphics, every effort is made to consider UD when writing and reviewing test items.

The five principles of UD that CAI test development specialists refer to when writing and reviewing items for EOC exams are listed in Table 3.

Table 3. Principles of Universal Design Applicable to Item Writing and Reviewing

Principles	Attributes
1. Flexible Use	Provide equal availability for access to the item. Make the design of the items appealing and accessible to all.
2. Simple and Intuitive	Eliminate unnecessary complexity, particularly in language and visuals.
3. Perceptible Information	Provide adequate contrast between the essential information and surrounding information. Eliminate any extraneous information.
4. Tolerance for Error	Maintain the cognitive complexity being measured by eliminating unnecessary clutter that may artificially raise the complexity of the item.
5. Low Physical Effort	Eliminate the need for excessive writing and unnecessary calculations.

Implementing Universal Design Principles

All the test developers at CAI are trained to write items that are accessible to all students based on the principles of UD. Additionally, they are required to pass a certification examination that certifies their ability to implement CAI's Language Accessibility Guidelines in the items they are developing. Each item presented to the Hawai‘i review committees is reviewed by three CAI content experts, as well as an editor. At each of these reviews, every item is checked for language accessibility and adherence to UD principles.

These are the Language Accessibility Guidelines used by CAI when writing and reviewing items.

Language should be as direct, clear, and inclusive as possible. The following should be avoided or used with care:

- Passive construction
- Idioms
- Multiple subordinate clauses
- Pronouns with unclear antecedents
- Words with multiple meanings
- Nonstandard grammar
- Dialect
- Jargon

2.5 Development and Review Process for New Items

2.5.1 Development of New Items

For the EOC exams, new items were developed according to the blueprint and item development plan. All items were developed originally by CAI content specialists. CAI staff used the content specification guides to create items that matched each CCSS benchmark. Then, these items were reviewed internally by content, editorial, and senior content specialists. Each item went through an extensive five-step review process: preliminary review, content 1 review, edit review, content 2 review, and batch review. Each step required either a content expert or an assessment production editor to review the item. Items were reviewed for alignment to the CCSS benchmarks and for basic item construction. The CAI content and assessment staff

discussed and revised items as needed. A different person reviewed the item at each review level. Approved items were then sent to HIDOE for review.

Following the completion of the CAI and HIDOE internal reviews, the items were reviewed by a Hawai‘i committee that combined the Fairness and Sensitivity review with the content review. This committee is composed of teachers and educators from across Hawai‘i. The fairness review identifies any potential item biases or stereotypes. Content review determines if the items are properly aligned to the CCSS benchmarks, accurately measure intended content, and are grade-level appropriate. Items are modified based on the review comments from the committee. Items the committee deems to have fatal flaws are rejected prior to field testing.

After the field test is completed, members of the rubric validation committee review a sample of the responses provided to each MSCR item and either approve the scoring rubric or suggest a revised score based on their interpretation of the item task and rubric. A sample of responses is chosen to find possible scoring inconsistencies. A portion of the sample represents tests that received an item score higher than the overall score, and another portion is chosen based on tests that received an item score lower than the overall score.

CAI staff used the item specifications to train qualified item writers, each of whom had prior item-writing experience. The item writers were trained previously at CAI item-writing workshops or had previous training on writing selected-response and constructed-response items. A CAI content-area assessment specialist worked with the item writers to explain the purpose of the assessment, review measurement practices in item writing, and interpret the meaning of the CCSS benchmarks as illustrated by the test and item specification documents. Sample item stems in the test/item specification documents served as models for the writers to use in creating items to match the standards. To ensure that the items tapped the range of difficulty and taxonomic levels required by HIDOE, item writers used a method based on Webb’s cognitive demands (Webb, 2002) to develop item types that incorporate a variety of cognitive processing levels from “recall” to “strategic thinking.” Eligible DOK levels are indicated in the test and item specification documents. Item writing and passage selection are guided by the following principles for each of the item types. When writing selected-response items, item writers are trained to develop items that

- have one correct response option;
- contain plausible distractors that represent feasible misunderstandings of the content;
- represent the range of cognitive complexities and include challenging items for students performing at all levels;
- are appropriate for students in the assigned grade in terms of reading level, vocabulary, interest, and experience;
- are embedded in a real-world context;
- do not provide answers or hints to other items in the set or test;
- are in the form of questions or sentences that require completion;
- use clear language and are not worded in the negative unless doing so provides substantial advantages in item construction;
- are free from absolute wording, such as “always” and “never,” and have qualifying words (e.g., least, most, except) printed in small caps for emphasis; and

- are free of ethnic, gender, political, and religious bias.

Algebra 1/Algebra 2

The item writers also consider the DOK levels while writing test items. When determining these levels, content experts make judgment calls, taking the following characteristics into account.

DOK 1: Recall

- Recall information, such as a fact, definition, term, or simple procedure.
- Perform a simple algorithm.
- Apply a formula.

DOK 2: Skill/Concept

- Carry out experimental procedures.
- Make observations and collect data.
- Classify, organize, and compare data.
- Organize and display data in tables, graphs, and charts.

DOK 3: Strategic Thinking

- Draw conclusions from observations.
- Cite evidence and develop a logical argument for concepts.
- Explain phenomena in terms of concepts.
- Use concepts to solve problems.

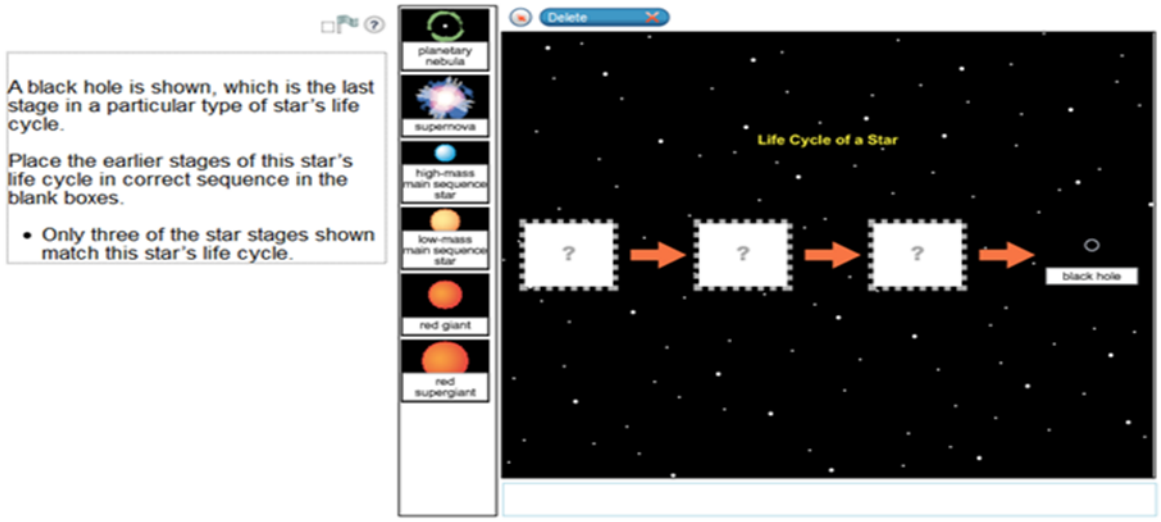
2.5.2 Developing Machine-Scorable Constructed-Response Items

One of the important features of the online EOC exams is the administration of MSCR items. Various types of MSCR items were developed, including graphical response items (GI) and equation response (EQ) items. The GIs require a student to place objects or move objects around in the answer space. The student can also plot points, draw lines, and draw shapes. The EQ items allow students to create equations and expressions using their keyboard and/or an online keypad. The development process for these items follows a typical procedure for human-scored constructed-response items, with content experts and graphic artists working together to create items. Throughout the development process, each item is associated with a rubric described in English. Using online tools designed for this purpose, test developers operationalize the human-readable rubric in declarative, machine-readable form.

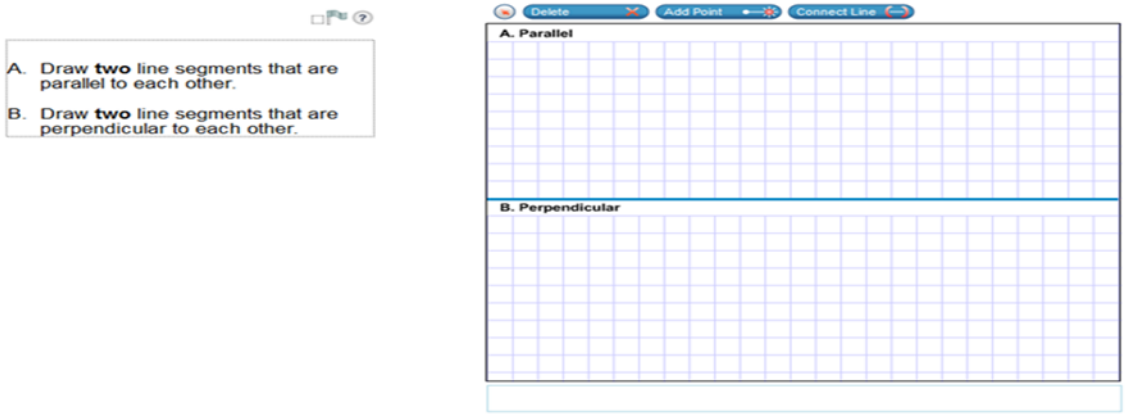
2.5.2.1. Graphical Response Items

The GIs require a student to place objects or move objects around in the answer space. The student can also plot points and draw lines and shapes. GIs allow assessing a high level of complexity that usually cannot be achieved with selected-response items. GIs are rendered online only. The two basic types of GIs are shown in the screen captures on the following pages. They include the following:

- In a drag-and-drop item, the student is given a choice of images, housed in the palette or preplaced in the answer space, and can drag and drop those images on the answer space to show his or her answer. The following screen capture shows one such example.

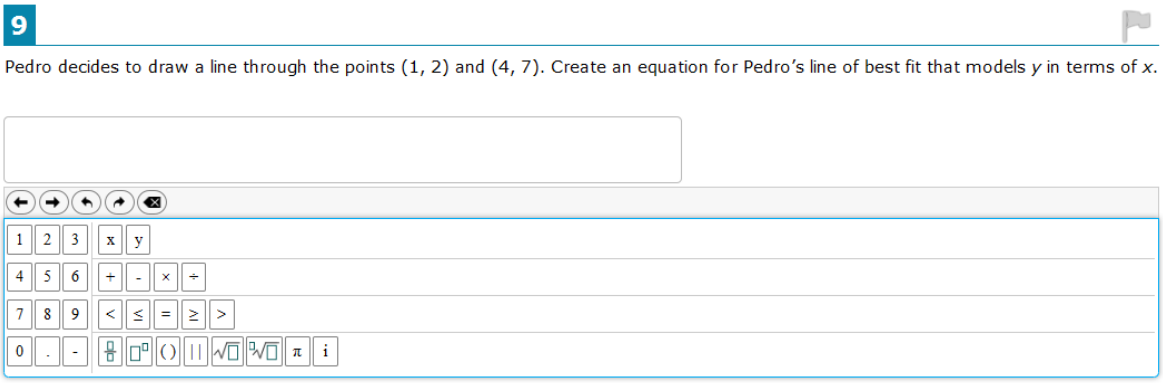


- In a drawing item, the student is given the option to plot points and/or draw lines. An item might require the student to plot points or draw lines on a coordinate grid. Additionally, the student may use the connect line tool to draw shapes within the answer space, as the following screen capture indicates.



2.5.2.2. Equation Response Items

EQ items require students to enter an equation, expression, or numerical value using an online keypad or their keyboard. The following screen capture provides an example.



2.5.3 Developing Selected-Response Items

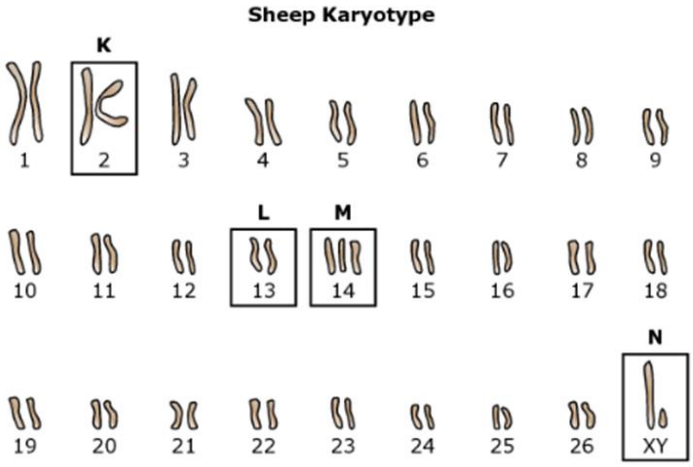
Various types of selected-response items were developed: evidence-based selected-response items (EBSR), editing task choice (ETC) items, and multi-select (MS) items. An EBSR item has two parts working together as a single item. Part A requires the student to identify a fact or recall a specific bit of information, and in Part B, the student selects supporting evidence for his or her answer in Part A. The supporting evidence in Part B is typically taken from information in the stem of the item or from the associated stimulus. An ET choice item allows the student to choose from options to replace given text. The given text is crossed out and replaced with the student's choice. An ET choice inline item allows the student to choose text to fill in or complete a sentence to construct an explanation. A MS item requires the student to evaluate each of the five to eight options and select all the correct responses. MS items can identify the specific number of correct responses the student needs to select.

2.5.3.1. Evidence-Based Selected-Response Items

EBSR items have two parts working together as a single item. Part A requires the student to identify a fact or recall a specific piece of information, and Part B asks the student to select supporting evidence for his or her answer in Part A. The supporting evidence in Part B is typically taken from information in the item stem or from the associated stimulus. An example of an EBSR item is shown below.

Part A

A wild sheep is observed to have a number of physical ailments. A geneticist takes a cell sample from the sheep and produces the karyotype shown to investigate the cause of these ailments. The geneticist labels four parts of the karyotype as K, L, M, and N.



What is the genetic cause of the sheep's physical ailments, based on information shown in the karyotype?

Ⓐ The sheep did not receive any dominant alleles from its parents.

Ⓑ The sheep did not receive the correct combination of chromosomes from its parents.

Ⓒ The sheep is infected by a virus that is removing essential genes from its DNA.

Ⓓ The sheep experienced a new selection pressure, which caused multiple insertion mutations.

Part B

Which part of the karyotype provides evidence to support your choice in part A?

Ⓐ K

Ⓑ L

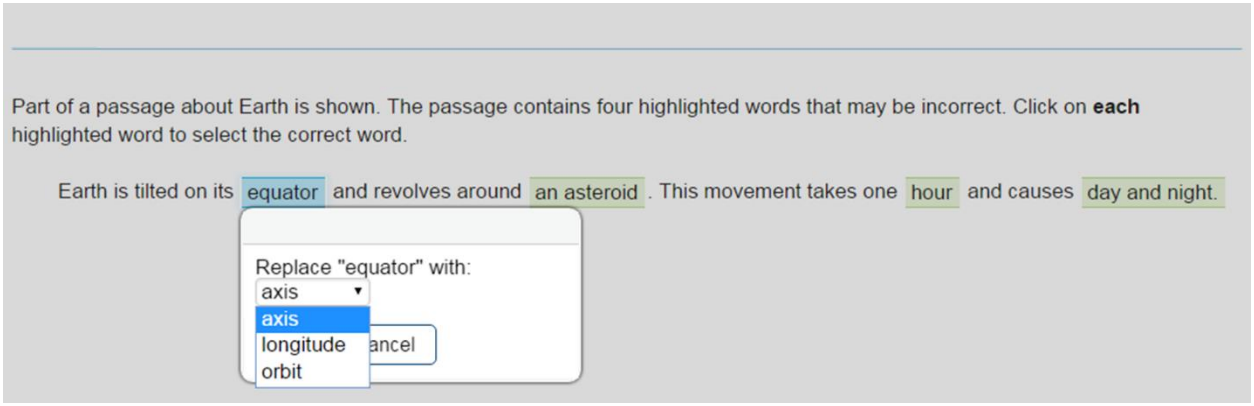
Ⓒ M

Ⓓ N

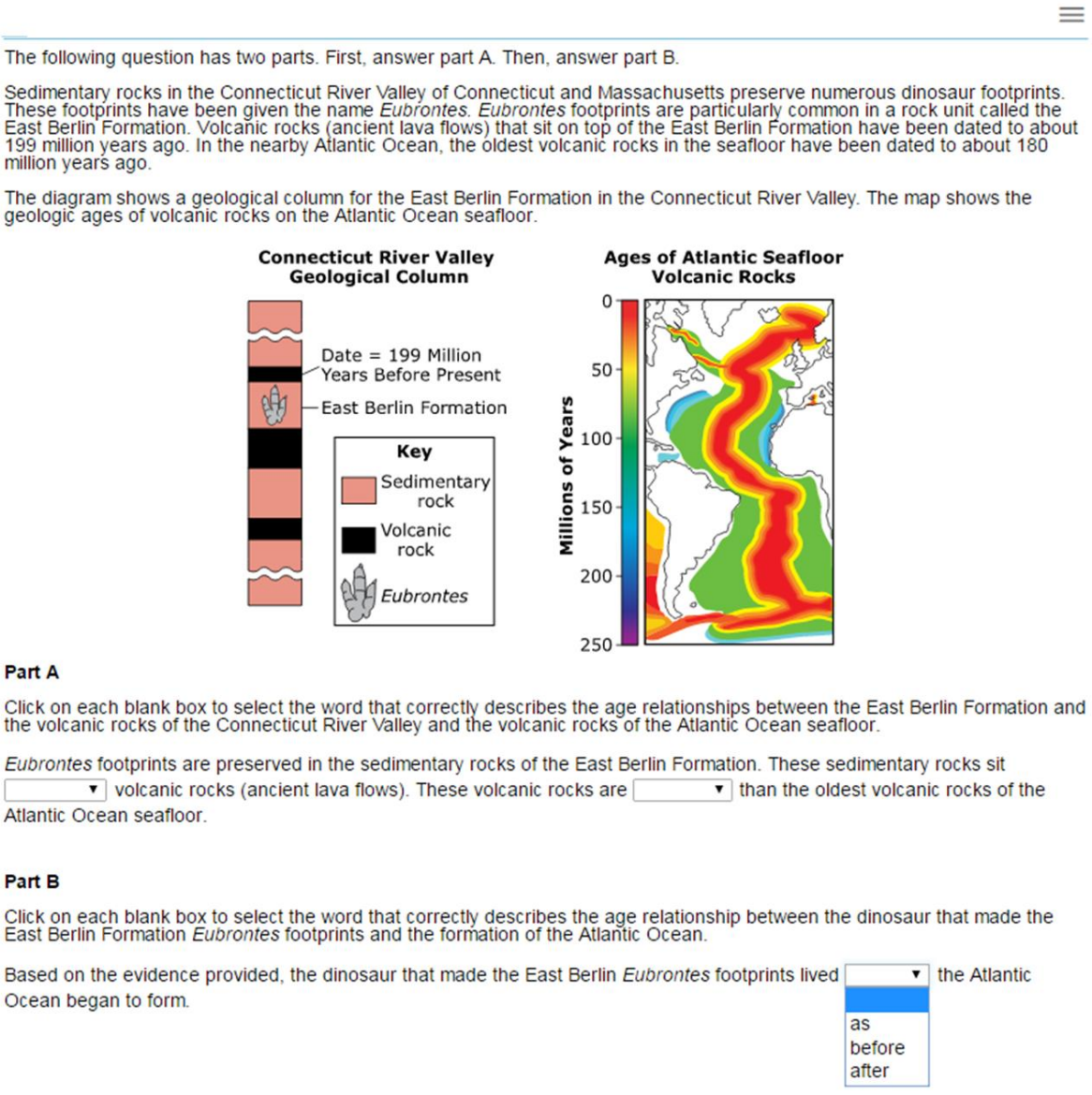
2.5.3.2. Editing Task Choice Items

An ETC item is similar in format to an ET item. An ET item allows the student to correct errors by typing in text to replace certain text within a sentence. An ET choice item allows the student to choose from options to replace given text. The given text is crossed out and replaced with the student's choice. An ET choice inline item allows the student to choose text to fill in or complete a sentence to construct an explanation. Examples of ETC items are shown below.

- ET Choice Item

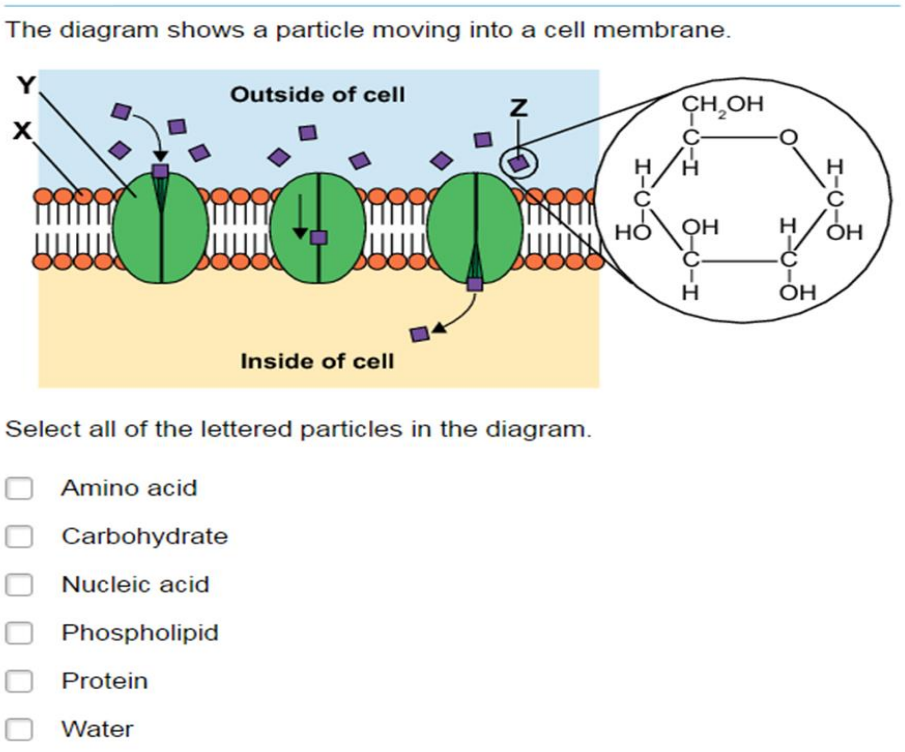


- ET Choice Inline Item



2.5.3.3. Multi-Select Items

A MS item requires the student to evaluate each of the five to eight options and select all the correct responses. MS items may or may not identify the specific number of correct responses the student needs to select. An example of a MS item is shown below.



2.5.4 Department Item Review and Approval

Once the newly developed items were reviewed and approved internally, they were submitted to HIDOE for review. CAI made HIDOE's revisions to the items before the Content and Fairness Advisory Committee (CFAC) reviewed the items. The items that were field tested in each administration had been reviewed by the CFAC. The CFAC is made up of expert representatives, including HIDOE reading, mathematics, and science curriculum staff and Hawai'i educators, including special education (SPED) teachers and English language (EL) teachers. This item review consisted of a short training, after which the reviewers reviewed each item independently and discussed issues or potential problems and solutions. The items were accepted with no changes, accepted with approved changes, or rejected from the item pool.

2.5.5 Committee Review of Item Pool

After a general introductory session, the CFAC was divided into subgroups by content area and grade to learn how to conduct an item review. After a PowerPoint training, the subgroups began reviewing each item. The reviews started as a group effort. However, once the committee members felt confident in their task, they began reviewing the items independently. After a predetermined set of items was reviewed independently, the group came back together to discuss concerns and solutions, eventually agreeing on the outcome for each item.

The discussion centered on the alignment of the item to the CCSS benchmarks, the alignment to the DOK level, the grade-level appropriateness, and the readability of each item. The CFAC used the CCSS benchmarks to review the content that each item measured.

During the CFAC item review meeting, members also reviewed all the items using the language accessibility and bias and sensitivity (LABS) guidelines. CAI leaders outlined the purpose of this review, discussed the guidelines, and worked through a few of the items with the group as a practice so that the committee members knew what to look for as they completed the reviews on their own.

3. TEST ADMINISTRATION

3.1 TESTING WINDOWS

The EOC exams were administered online via the CAI Secure Browser used for administration of the Hawai'i State Assessments (HSA) at the end of course instruction (see testing windows in Table 4). For the online exams, schools schedule their testing dates according to the number of students tested within this testing window. Students are allowed one opportunity for each EOC exam during a semester. If the student fails the course during the semester, the student may retake the course and the EOC exam during a later semester.

Table 4. 2022–2023 EOC Testing Windows

Tests	Start Date	End Date	Item Selection
Fall 2022 Algebra 1 and Algebra 2 Exams	11/21/2022	12/16/2022	Adaptive
Spring 2023 Algebra 1 and Algebra 2 Exams	4/24/2023	5/26/2023	Adaptive
Summer 2023 Algebra 1 and Algebra 2 Exams	6/16/2023	7/14/2023	Adaptive

3.2 TEST OPTIONS AND ADMINISTRATIVE ROLES

The Hawai'i State EOC exams are administered entirely online, either in person or remotely. To ensure standardized administration conditions, test administrators (TAs) follow procedures outlined in the test administration manual (TAM). TAs must review the TAM prior to the beginning of testing, ensure that the testing room is prepared for testing (e.g., removing certain classroom posters, arranging desks), complete an online TA Certification Course, and establish makeup procedures for any students who are absent on the day(s) of testing.

TAs follow required administration procedures and directions. They read the boxed directions aloud to students verbatim to ensure standardized administration conditions for all exams.

3.2.1 Administrative Roles

The key personnel involved with test administration are school principals, test coordinators (TCs), TAs, and technology coordinators. Proctors may also assist TAs during test administration if more than 25 students are assigned to one TA. The main responsibilities of these key personnel are described in this section. More detailed descriptions can be found in the *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023*, provided online at this URL: https://eoc.alohahsap.org/resources/resources-2022-2023/hsa-science-ngss-and-eoc-exams-test-administration-manual-2022-2023.

School Principal

The school principal is held accountable for ensuring that online testing is conducted in accordance with test security and other policies and procedures established by HIDOE. The school principal is responsible for creating or approving the testing schedule and procedures for the school and resolving testing problems as needed. The school principal is also responsible for designating a school employee, either himself/herself or another staff member, to act as the official TC, entering the TC contact information into the online testing system, and updating the information throughout the year.

Technology Coordinator

The primary responsibility of the technology coordinator is to ensure that the school's hardware and software meet the requirements for the online exams. The technology coordinator is expected to understand the basic functionality of the Online EOC exams, install the Secure Browser for online testing on each computer prior to testing, and work with the TAs to coordinate the technical details for testing. For further details on the secure browser used for testing and other hardware and software requirements, please refer to the *Online Technology Guide* *and the* *Technology Requirements Training Module (Non-Narrated) 2022-2023*, provided online at the EOC page of the AlohaHSAP.org portal.

Test Coordinator

The role of a TC is to coordinate the testing activities at the school level. TCs are responsible for identifying TAs and ensuring that the assigned TAs are properly trained and certified. The TC(s) at each school must also work with the technology coordinator to ensure that there are sufficient hardware and software resources to support testing, create the test schedule, disseminate information about testing to other staff and parents, monitor testing progress during the testing window to ensure that all students participate as required, and report major testing problems to the principal.

Schools may have more than one TC. Although many qualified school staff members can serve in the capacity of a TC, it is recommended that a TC be a person with non-instructional or limited instructional duties.

Test Administrator

TCs identify the TAs, and all TAs must pass the online TA Certification Course. TAs are responsible for administering the exams to the students.

TAs are expected to

- review the appropriate manuals and user guides on how to administer the exams;
- practice administering the test through the TA training website before conducting the first test session;
- prepare the testing environment and ensure that students have the necessary test materials, including scratch paper, keyboard shortcut handouts, and pencils, as appropriate;
- administer each exam according to the *Directions for Administration* for the online versions;
- report testing irregularities; and
- shred scratch paper and paper handouts that students write on during testing in a secure manner immediately after each test session.

Proctors

Proctors are recommended when more than 25 students will test with one TA. TCs work with the school principal to identify proctors to assist the TAs in administering the exams. Staff members eligible to serve as proctors include educational assistants, part-time teachers, and project teachers. The role of a proctor is to walk around the testing room, monitor student behavior, and inform the TA if any student(s) becomes ill, is disruptive, or appears to be cheating.

TCs are responsible for ensuring that the proctors are familiar with test security procedures and student confidentiality requirements before they are allowed to assist a TA in a testing room. Additionally, all proctors are required to sign an *Acknowledgment Form for Proctors and Skills Trainers*, listed as Appendix V in the *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023*, provided online at https://eoc.alohahsap.org/resources/resources-2022-2023/hsa-science-ngss-and-eoc-exams-test-administration-manual-2022-2023.

3.2.2 Online Administration

Online EOC exam testing allows schools to choose testing dates and to test students in intervals rather than in one continuous period. With the online EOC exams, schools do not need to handle test booklets and address the storage and security problems inherent in large shipments of materials to a school site.

Starting with the 2020–2021 school year, a new feature was developed within the universally used TDS that allowed tests to be administered remotely by a TA to student's who remained at home. It was a school-level decision to allow students to test remotely in cases when a parent or guardian refused to bring a student onto campus but insisted on the student to be tested. These new features allowed a TA to pre-schedule a testing session, have online video chats with a group of students, and enabled TAs to video-monitor a group of students while in a testing session. To help TAs understand how to use these additional features, an additional *Remote Testing TA Certification Course* was developed and was required to be taken by all TAs that were to administer a remote testing session. Also, before a student was eligible for a remote administration, a parent or guardian must give written consent to the school to administer a remote test, which would contain video and audio for the TA to view the student. The TC at the school would then identify the positive consent to remote testing within the Test Information Distribution Engine (TIDE) system. Additional resources were developed for TAs to understand the requirements for remote testing and posted to the state portal at https://eoc.alohahsap.org/resources/resources-2022-2023/remote-summative-test-administration-2022-2023.

To start a test session, the TA must first log in to the TA Interface of the online testing system using his or her own computer. A session ID is generated when the test session is created. Students who are taking the exam with the TA then enter their Statewide Student Identifier (SSID), first name, and the session ID to log in to the Student Interface using computers provided by the school. The TA then verifies that the students are taking the appropriate exam and are provided with their appropriate exam accommodations, such as testing in a small group. Students can begin testing only after the TA confirms that the students are taking the appropriate exam and approves them to be tested. The TA needs to read the *Directions for Administration* in the *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023* aloud to the students and walk them through the login process.

Once an exam is started, the student must answer all test questions on a page before proceeding to the next page; students are not allowed to skip questions. The online testing system lets a student review and edit answers as long as the student is in the same test session and the test session has not been paused for more than 20 minutes.

In the online testing system, an exam can be started in one test session and completed in another session. However, the exam must be completed within the applicable EOC testing window, or the exam opportunity will expire.

Test sessions are not timed; therefore, students can use as much time as they need to complete an assessment. TAs can pause a single student's exam or all the exams during a test session (e.g., to give students a break).

It is up to the TA to determine an appropriate stopping point; however, to ensure the integrity of the exams, tests cannot be paused for more than 20 minutes. If that happens, the student can continue the same exam opportunity but must do so in a new test session. In the new test session, answers provided in the previous session are unavailable for review or editing.

If in-person testing is occurring, the TA should remain in the room at all times during a test session to monitor student testing, or if testing remotely, continually use the video feature to monitor the student at all times. Once the test session ends, the TA must make sure that each student has successfully logged out of the system and collect and securely shred any handouts or pieces of scratch paper that were used by students during the exam.

3.3 Allowable Resources for Online Testing

During testing, students may use specified tools and resources, including the Algebra 2 mathematics reference sheet, graph paper, and scratch paper. For the Algebra 1 and Algebra 2 EOC exams, a pop-up scientific calculator is available in the online system during the first portion of the exam, and a pop-up scientific/graphing/regression calculator is available during the second portion of the exam. A pop-up Algebra 2 mathematics reference sheet is also available in the online system. TAs can also print out the reference sheet for students. Students may use blank scratch paper and response aids (e.g., adaptive pencils, key guards, skins). Table 5 identifies resources that may be provided to students during the exams.

Table 5. Allowable Resources for 2022–2023 Online EOC Exams

Algebra 1 and Algebra 2 EOC Exams

▪ Calculators:
  o For Algebra 1 and Algebra 2 EOC exams, two pop-up calculators are available in the online system. Students may *not* use handheld calculators. For the first segment in the Algebra 1 and Algebra 2 EOC exams, a pop-up scientific calculator is available. For the second segment, a pop-up scientific/graphing/regression calculator is available.
▪ Mathematics Reference Sheets:
  o Algebra 2 EOC exam pop-up Mathematics Reference Sheets are available in the online system. These sheets may also be copied and handed out to students. Mathematics Reference Sheets for the Algebra 2 EOC exam are available at https://eoc.alohahsap.org/resources/resources-2020-2021/algebra-2-mathematics-reference-sheet.

▪ Headphones are required for students with the text-to-speech designated support.
▪ Pen or pencil.
▪ Blank scratch paper or graph paper for Algebra 1 and Algebra 2 EOC exams (must be securely shredded immediately after a test session if written on by students). The blank scratch paper can be used to take notes about test questions or work problems using mathematics calculations and drawings.
▪ Masks or barriers to prevent students from looking at others' computers.
▪ Posters offering students encouragement or inspiration without any specific content from the Common Core State Standards related to the Algebra 1 and Algebra 2 EOC exams, such as
  o "Believe in Yourself"
  o "Set Your Goals High"
▪ Handout of keyboard shortcuts (online testing system navigation symbols). These may also be posted in larger sizes on a wall if desired.

3.4 TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

TCs oversee all aspects of testing at their schools and serve as the main point of contact, and TAs administer the online EOC exams. The online TA Certification Course, webinars, user guides, manuals, and training sites are used to train TCs and TAs in the online testing requirements and the mechanics of starting, pausing, and ending a test session. Training materials for test administration are provided online. Multiple online training opportunities are offered to the key staff through the Internet.

3.4.1 TA Certification Course

All school personnel who serve as TAs must complete an online Summative TA Certification Course. This web-based course takes about 30–45 minutes and covers information on testing policies and the steps for administering a test session in the online system. The course is interactive, requiring participants to practice starting test sessions. Throughout the training and at the end of the course, participants must answer multiple-choice questions about the information provided. Staff members who meet the requirements to serve as a TA and who pass this certification course receive a certificate of completion and then appear in the online testing system as qualified TAs who are authorized to administer the exams in all content areas. A second TA Certification Course of about 20 minutes long was an added requirement for TAs who would be administering tests in a remote format. TAs that were administering remote tests were required to take both TA certification courses.

3.4.2 Modules

Seven training modules were offered. The first module provided information on administering a test using Speech-to-Text (STT) software. The second module explained how to navigate the CRS, including how to retrieve student results. The third module assisted TCs and TAs in understanding the Secure Browser that students use to take the online test. The fourth module helped users understand the TA Live Site used in online testing. The fifth module helped technology coordinators prepare for the administration of online tests. The sixth module provided TAs with information on administering online tests to students using braille. The seventh module explained how to navigate TIDE, including how to manage student testing information such as test settings and accommodations. All modules were provided as PowerPoint presentations.

Ten short training tutorials were also offered that covered various aspects of the CRS. Topics included: Viewing Claim Detail Reports, Creating Rosters, Defining the Student Population, Downloading Individual Student Reports, Downloading Student Data Files, Printing Reports, Viewing and Editing Rosters, Viewing Claim Level Detail Reports, Viewing Item Detail Reports, Viewing Reports by Demographic Subgroups, and Viewing Target Reports. All tutorials were provided as MP4 video files.

3.4.3 Webinars

Two webinar presentations were offered. The first webinar was for new test coordinators and focused on how to navigate TIDE, including instructions on managing student information and monitoring test progress, setting up student testing sessions, and discussed accessibility and supports available to students during testing, including universal tools, designated supports, and accommodations. The second webinar was for all test coordinators and further described how to navigate TIDE, TDS, CRS, and instructions on managing student testing.

The length of each of these webinars is about one hour. The interactive nature of these training webinars allows the participants to ask questions during and after the presentation. The webinar is recorded, and a streaming video of the webinar is made available to all Hawai'i school personnel a few days after each live webinar on the Hawai'i Statewide Assessment Program (HSAP) portal website at https://eoc.alohahsap.org/resources#folder=Webinars.

3.4.4 Training Sites

About four weeks before the first online EOC exam testing window begins, TAs can practice administering exams and starting and ending test sessions on the TA training site, and students can practice taking an online exam on the student practice and training site. The practice tests mirror the corresponding content exams and contain approximately 40–50 items in each content area. The practice tests are designed to give students and TAs opportunities to quickly familiarize themselves with the software and navigational tools that they will use on the exams. A combined training test containing 5–10 test items is available for Algebra 1 and Algebra 2. A student can log in directly to the training site as a guest without a TA-generated test session ID, or they can log in through a training test session created by the TA in the TA training site. Items in the student training test include all item types that are included in the operational item pool (e.g., multiple -choice items, grid items, and natural-language items).

3.4.5 Manuals and User Guides

In addition to the online training and resources, a series of manuals and user guides are available on the HSAP portal, https://eoc.alohahsap.org/resources#school_year_sm=2022-2023.

The *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023* identifies the procedures to be followed before, during, and after test administration and includes clear procedures for properly collecting and destroying student test materials between and after test sessions to ensure security. This manual also provides eligibility requirements for student participation in the EOC exams, forms related to test security, contact information for the HSAP Help Desk, and student accommodations information. The *Assistive Technology Manual 2022–2023* includes information about supported operating systems and required hardware and software for braille testing. It provides information on how to configure Job Access With Speech (JAWS), how to navigate an online test with JAWS, and how to administer a test to a student requiring braille.

To standardize test administration conditions, all TAs are required to follow the procedures outlined in this manual, which includes explicit directions for administration, for each test session.

The *Secure Browser Installation Manual 2022–2023, Quick Guide for Setting Up Your Online Testing Technology 2022-2023, Technology Requirements Training Module (Non-Narrated) 2022-2023* and the *Operating System Support Plan for Test Delivery System 2022–2023* summarize the technology coordinator role, the Online HSAP applications, the hardware and the software requirements for the EOC exams, and information about the secure browsers. The *Test Information Distribution Engine (TIDE) User Guide 2022–2023* is a manual for TCs and other users that discusses the TIDE application within the Online HSAP system, which allows TCs to manage user role assignments, verify the TA's certification status, set student accommodations for testing, create and review testing incident requests, create and edit rosters, and update school contact information.

The *Guide to Navigating Online HSAP Administrations 2022–2023* is a software guide on how to use the online system applications, including the test administration and student testing sites. The *Centralized*

Reporting System User Guide 2022–2023 is a user guide that provides instructions on how to generate reports to see which students have or have not completed assessments and how to generate reports with student score information. All manuals and user guides pertaining to 2022–2023 online testing are available on the HSAP portal during the three- to four-week testing window.

3.5 TEST SECURITY

All test items, test materials, and student-level testing information are secure materials for online exams. This section describes student confidentiality, test security, and policies on testing improprieties.

3.5.1 Student-Level Testing Confidentiality

The Family Educational Rights and Privacy Act (FERPA) prohibits the public disclosure of student information or test results. The following are examples of prohibited practices:

- Giving out login information (username and/or password) either to other authorized TIDE users or to unauthorized individuals
- Sending a student's name and SSID number together in an email message (if information must be sent via email or fax, include only the SSID number, not the student's name)
- Having students log in and test under another student's SSID number

Student test materials and reports should not be exposed in such a manner that student names can be identified with student results, except by authorized individuals with an educational need to know.

All students must be enrolled or registered at their testing schools to take the online exams. Student enrollment information, including demographic data, is generated using an HIDOE file and uploaded nightly to the online testing system during the testing period via a secure file-transfer site.

Students log in to the online EOC exams using their legal first name, SSID number, and a Test Session ID. Only students can log in to an online test session. TAs, proctors, and other personnel are not permitted to log in to the online EOC system on behalf of students, though they are permitted to assist students who need help logging in.

After a test session, only staff in the administrative roles of school principals, TCs, and teachers can view their students' scores. TAs and proctors do not have access to student scores.

3.5.2 System Security

The objective of system security is to ensure that all data are kept protected and accessed appropriately by the right user groups. It is about protecting and maintaining data and system integrity as intended, including ensuring that all personal information is secured, transferred data (whether sent or received) are not altered in any way, the data source is known, and any service can be performed only by a specific, designated user. The importance of maintaining test security and the integrity of test items is stressed throughout the online TA Certification Course, webinar trainings, user guides, and manuals. Features in the testing system also protect test security.

3.5.2.1. System Built-In Test Security

A Hierarchy of Control

As described in Section 3.1, Testing Windows, principals, technology coordinators, TCs, TAs, and teachers have well-defined roles and access to the testing system. Principals are responsible for selecting and entering the TC's information into TIDE, and the TC is responsible for entering TA and teacher information into TIDE. Throughout the year, the TC is also expected to delete information in TIDE for any staff members who have transferred to other schools, resigned, or no longer serve as TAs or teachers.

Password Protection

All access points by different roles—at the state level, complex area level, school principal level, and school staff level—require a password to log in to the system. Newly added TCs, TAs, and teachers receive separate passwords through their personal email addresses assigned by the school.

Secure Browser

A key role of the technology coordinator is to ensure that the Secure Browser is properly installed on the computers used for the administration of the online exams. The Secure Browser, developed by the testing contractor, CAI, prevents students from accessing other computers or Internet applications and copying test information. The Secure Browser suppresses access to commonly used browsers, such as Internet Explorer and Firefox, and prevents students from searching for answers on the Internet or communicating with other students. The online EOC exams can be accessed only through the Secure Browser and not by other Internet browsers.

3.5.3 Security of the Testing Environment

The school principal, technology coordinator, TC, teachers, and TAs work together to determine appropriate testing schedules based on the number of computers available, the number of students in each EOC course, and the average amount of time needed to complete each exam.

TCs are reminded in the online training and user manuals that exams should be administered in testing rooms that do not crowd students. Good lighting, ventilation, and freedom from noise and interruptions are important factors to consider when selecting testing rooms.

TCs and TAs must establish procedures to maintain a quiet environment during each test session, recognizing that some students may finish quicker than others. If students are allowed to leave the testing room when they finish, TAs must explain the procedures for leaving without disrupting others and tell students where they are expected to report once they leave. If students are expected to remain in the testing room until the end of the session, TAs are encouraged to prepare some quiet work for students to do after they finish the exam.

If a student needs to leave the room for a brief time, the TA is required to pause the student's exam. If the pause lasts longer than 20 minutes, the student can continue with the rest of the exam in a new test session, but the system will not allow the student to return to the answers provided prior to the pause. This measure was implemented to prevent students from using the time to look up answers.

Room Preparation

The room should be prepared prior to the start of the test session. Any information displayed on bulletin boards, chalkboards, or charts which students might use to help answer test questions should be removed or covered. This applies to rubrics, vocabulary charts, student work, posters, graphs, content-area strategies charts, etc. TA and student cell phones must be turned off and stored out of sight in the testing room. TAs are encouraged to minimize access to the testing rooms by posting signs in halls and entrances to promote optimum testing conditions; they should also post "TESTING—DO NOT DISTURB" signs on the doors of testing rooms.

Seating Arrangements

TAs should provide adequate spacing between students' seats. Students should be seated in such a way that they will not be tempted to look at the answers of others. Because the online EOC exams are adaptive, it is unlikely that students will see the same test questions as other students; however, students should be discouraged from communicating with one another through appropriate seating arrangements.

After the Test

The TA must walk through the classroom at the end of a test session to pick up any scratch paper that students used and any papers that display students' SSID numbers and names together. These materials should be securely shredded or stored in a locked area immediately. The printed questions for any EOC exam provided for a student who is allowed to use this accommodation in an individual setting must also be shredded immediately after a test session ends.

3.5.4 Test Security Violations

School Personnel

Everyone who administers or proctors the exams is responsible for understanding the security procedures for administering the exams. Prohibited practices, as detailed in the *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023*, include but are not limited to

- reproducing, photographing, or recording any information from secure online exams;
- providing access to or disclosing any information from secure online exams to anyone before the exam; and
- reviewing, discussing, or analyzing secure test questions or student responses with anyone during or after exam administration.

During testing, school personnel, and other adults are prohibited from

- providing or allowing the translation of test questions or directions for any students beyond the accommodations;
- providing or allowing the use of accommodations or resources that were not in a student's Individualized Education Program (IEP) or have not received prior approval for the individual student;
- omitting portions of directions that must be read to the students;

- explaining test questions or providing nonverbal clues to students;
- allowing students to leave the testing room prior to ending or pausing their test session;
- displaying content- or process-related information beyond the allowable materials such as keyboard shortcuts and Mathematics Reference Sheets;
- explaining or reviewing test-taking strategies with the students immediately prior to a testing session;
- altering student responses or encouraging a student to alter responses; and
- using a student's SSID number to log in to the online testing system.

Students

All students are reminded that the exams are secure materials. This reminder is included in the *Directions for Administration* and should be read aloud verbatim by TAs at the beginning of each test session.

Students are prohibited from disclosing any information from secure exam materials to anyone, including other students or unauthorized adults such as parents or guardians, other relatives, or friends. Scratch paper and authorized paper handouts that schools may provide to students are allowed to be used during test sessions. However, any scratch paper or handouts that students write on during test sessions must be collected and securely shredded immediately after each test session.

During testing, students are prohibited from

- sharing content or procedural information, including discussing test questions or directions during the test administration;
- translating test questions or directions for other students;
- talking to other students;
- passing papers or sharing materials;
- using electronic communication tools, such as cell phones, to photograph or share information;
- altering the response(s) of another student or encouraging another student to alter his or her response;
- using unapproved resources for information to answer test questions;
- accessing the Internet; and
- using another student's SSID number to log in to the online testing system.

Cheating

During the test session, the TA and proctor should walk around the room and monitor student behavior. If a student is found cheating (e.g., communicating with another student during the test session), he or she should be removed from the testing room immediately. In these instances, the TA should immediately pause the student's exam and notify the school principal and TC. The principal or authorized designee is required to immediately inform the HIDOE Assessment Section and contact the student's parent(s) or guardian(s) to inform them of their child's cheating and the associated consequences. The TA is also required to fill out the Testing Incident Report Form in Appendix P of the *Hawai'i State Science (NGSS) Assessments and*

End-of-Course Exams Test Administration Manual 2022–2023, which is available from the HSAP portal website at https://eoc.alohahsap.org/resources/resources-2022-2023/hsa-science-ngss-and-eoc-exams-test-administration-manual-2022-2023.

3.5.4.1. Student Illness, Disruptiveness, and Other Testing Incidents

If a student becomes ill while taking an exam, the TA should pause the student's exam and allow the student to complete the exam later (within the applicable EOC testing window).

If a student becomes disruptive, the TA should pause the student's exam and remove him or her from the testing room immediately. The student can be given another opportunity to complete the exam at a later time. The TA must contact the student's parent(s) or guardian(s) immediately to inform them of the student's disruption during testing and the associated consequences. The TA is also required to fill out the Testing Incident Report Form in Appendix P of the *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023*.

Other testing incidents include major disruptions such as a fire drill, a school-wide power outage, or a natural disaster that could impact either test security or test validity. During an event such as a fire drill or other evacuation, safety is the top priority. Detailed instructions on how to pause and restart a test session in these circumstances are provided in the *Hawai'i State Science (NGSS) Assessments and End-of-Course Exams Test Administration Manual 2022–2023*.

3.5.4.2. Reporting Testing Incidents

All school staff members are required to report testing incidents to the school principal. Testing incidents that do not involve the TC should also be reported immediately to the TC.

School principals who witness, are informed of, or suspect the possibility of a testing incident that could potentially impact the integrity of the exams, data, or results are required to immediately contact their Complex Area Superintendent and HIDOE's Assessment Section. The Strategy, Innovation, and Performance Office Director informs the state superintendent of all reported testing incidents that could impact the integrity of the assessments, data, and results.

3.5.4.3. Consequences of Testing Improprieties

If testing incidents occur during the administration of an online EOC exam, HIDOE personnel communicate with the school principal and TC to verify the facts associated with the alleged testing incident. Upon investigation, the HIDOE personnel may invalidate the impacted exams. HIDOE employees can be held personally responsible for any violation of copyright laws or breach in test security.

3.6 Online Testing Features and Testing Accommodations

3.6.1 Online Testing Universal Tools for All Students

In 2022–2023, the following universal tools were available for all students to access. For specific information on how to access and use these universal tools, refer to the *Guide to Navigating Online HSAP Administrations 2022–2023*, provided at https://eoc.alohahsap.org/resources/resources-2021-2022/guide-to-navigating-the-online-hsap-administration-2021-2022.

Embedded Universal Tools

Breaks (Pause): With this tool, a student can pause an assessment or exam and return to the test question he or she was working on (however, if the assessment or exam is paused for 20 minutes or more, a student will not be allowed to return to previously answered test questions).

Calculator: With this tool, an embedded on-screen digital calculator is accessible for calculator-allowed items when students click the calculator button. For the first segment in the Algebra 1 and Algebra 2 EOC exams, a pop-up scientific calculator is available. For the second segment, a pop-up scientific/graphing/regression calculator is available.

Digital Notepad: With this tool, a student can make notes about an item. The digital notepad is item-specific and is available through the end of a test segment. Notes are not saved when a student moves on to the next segment or after a break of more than 20 minutes.

Expandable Passages and/or Stimuli: With this tool, students can expand each stimulus so that it takes up a larger portion of the screen.

Highlighter: This tool is used to mark desired text, test questions, item answers, or parts of these with a color. An enhanced highlighting feature allows multiple color options. Highlighted text remains available throughout each test segment. This tool is not available while the Line Reader tool is in use.

Keyboard Navigation: With this tool, students can navigate throughout text by using a keyboard.

Line Reader: Students use an onscreen universal tool to assist in reading by raising and lowering the tool for each line of text on the screen. If the enhanced line reader mode is enabled, all content except for the line in focus is grayed out for greater emphasis. This tool is not available while the Highlighter tool is in use.

Mark for Review: With this tool, students can mark questions they have answered to review them later (however, if an exam is paused for more than 20 minutes, students will not be allowed to return to marked test questions that were previously answered). Students taking the Algebra 1 and Algebra 2 EOC exams, both of which have two segments, only will be able to review items in the segment they are currently working on.

Strikethrough: With this tool, students can cross out text in answer options using the strikethrough function. If an answer option is an image, a strikethrough line will not appear, but the image will be grayed out.

Zoom: With this tool, students can make test questions, text, or graphics larger by clicking on the zoom icon, which has four levels of magnification. The default font size for all exams is 14-point. When using the zoom feature, a student only changes the size of text and graphics on the screen. Additionally, the print size may be preset in TIDE or set immediately prior to the start of a test session for a student. The print size levels are as follows:

- Level 1 (default × 1.5 = 21-point font)
- Level 2 (default × 1.75 = 24.5-point font)
- Level 3 (default × 2.5 = 35-point font)
- Level 4 (default × 3.0 = 42-point font)

Non-Embedded Universal Tools

Breaks: With this tool, breaks may be given at predetermined intervals or after completion of sections of the assessment for students taking a paper-pencil test. Sometimes students can take breaks when individually needed to reduce cognitive fatigue when they experience heavy assessment demands. The use of this universal tool may result in the student needing additional overall time to complete the assessment.

Scratch Paper: With this tool, students can use scratch paper to make notes, write computations, or record responses. Only plain paper or lined paper is appropriate for ELA/L. Graph paper is required beginning in 6th grade and can be used on all mathematics assessments. A student may use an assistive technology device for scratch paper as long as the device is consistent with the student's IEP and acceptable to the state.

3.6.2 Designated Supports

Designated supports are access features that are available for use by a student for whom a need has been indicated by an educator or team of educators, a parent/guardian, or a student. A consistent process needs to be used to determine which embedded and non-embedded designated supports are needed by a student for an exam. Educators who make these decisions for an identified student must have a clear understanding of the process for ensuring that this student is currently using the feature during classroom instruction and is given an opportunity to practice using any variation in the feature that will be provided during the administration of an exam.

Embedded Designated Supports

Color Contrast: This support enables a student to adjust screen background or font color based on his or her needs or preferences. This may include reversing the colors for the entire interface or choosing a font or background color.

Masking: This support enables a student to block off content that is not of immediate need or that may be distracting. A student is able to focus his or her attention on a specific part of the test item using masking.

Mouse Pointer: This support allows the mouse pointer to be set to a larger size or different color. A TA sets the size and color of the mouse pointer prior to testing.

Streamline: This support provides a streamlined interface of the test in an alternative, simplified format in which the items are displayed below the stimuli.

Text-to-Speech: This support enables a student who is a struggling reader to listen to instructions, stimuli, and/or items using text-to-speech (TTS) technology, which requires headphones.

Turn Off Any Universal Tools: With this support, a TA may disable any universal tools that might be distracting, that students do not need to use, or that students are unable to use.

Non-Embedded Designated Supports

100s Number Table: This support is a table listing numbers from 1–100 and is a non-embedded accommodation only for students with visual processing or spatial perception needs as documented in their IEP or Section 504 Plan. This table may be printed only for students approved for this accommodation.

Abacus: This tool may be used in place of scratch paper for students who typically use an abacus.

Amplification: Students may use this tool to adjust the volume control beyond the computer's built-in settings using headphones or other non-embedded devices.

Bilingual Dictionary: This support is a bilingual/dual-language word-to-word dictionary and is a language support. A bilingual/dual-language word-to-word dictionary can be provided for the EOC exams.

Calculator: This support is a non-embedded calculator for students needing a special calculator, such as a braille calculator or a talking calculator, which is currently unavailable in the assessment platform.

Color Contrast: With this support, test content of online items may be printed with different colors.

Color Overlays: With this support, a student who meets the criteria for the Print-on-Demand accommodation may place color transparencies over the printed stimuli, items, and answer options in an EOC exam if the color transparencies are used during classroom instruction.

Magnification: With this support, a student may adjust the size of specific areas or objects on the screen (e.g., text, formulas, tables, graphics, and navigation buttons) with an assistive listening device, including projection on a closed-circuit television. Magnification allows students to increase the size to a level not provided by the universal zoom tool.

Medical Supports: With this support, students may have access to an electronic device for medical purposes (e.g., Glucose Monitor). The device may include a cell phone, and should only support the student during testing for medical reasons.

Multiplication Table: This support is a single-digit (1–9) multiplication table and is a non-embedded accommodation only for students with a documented and persistent calculation disability. This table may be printed only for students approved for this accommodation.

Noise Buffers: With this support, a student may wear equipment (e.g., ear mufflers) or use white noise to block external sounds and must wear headphones unless tested individually in a separate setting.

Read Aloud: With this support, students who are struggling readers may have all or portions of an assessment or exam read aloud (e.g., stimuli and/or items) by a trained and qualified TA human reader who independently reviews the *Read Aloud Guidelines* and signs a training verification form included in the guidelines document on the portal at https://eoc.alohahsap.org/resources/resources-2020-2021/test-administration-guidelines-for-read-aloud-test-reader-sy-22-23-prior.

Scribe: With this support, a student who has documented significant motor or processing difficulties or who has had a recent injury, such as a broken hand or arm, that makes it difficult to produce responses may dictate his or her responses to a trained and qualified TA human scribe who records the responses verbatim. The scribe must independently review the *Scribing Guidelines* and sign a training verification form included in the Guidelines document on the portal at https://eoc.alohahsap.org/resources/resources-2020-2021/test-administration-scribing-protocol-sy-22-23-prior.

Separate Setting: With this support, the test location may be altered so that the student is tested in a setting different from the setting made available for most students. The following are four previous settings that were accommodations:

- Read Aloud to Self
- Being Seated Near TA
- Being Tested Individually
- Being Tested in a Small Group

Simplified Test Directions: With this support, students who need additional support understanding the test directions may be provided with Simplified Test Directions as a designated support. This could include students with difficulties in auditory processing, short-term memory, attention, or decoding. This designated support may require testing in a separate setting to avoid distracting other test takers.

Translated Student Interface Messages: With this support, a bilingual adult may read aloud a PDF file of directions translated in each of the currently supported languages.

Table 6 shows the number of students in this test administration who were provided designated supports.

Table 6. Number of Students with Allowed Designated Supports in 2022–2023

Designated Supports	Test	
	Algebra 1	Algebra 2
Embedded Designated Supports		
Color Contrast		
Masking		
Mouse Pointer		
Streamline		
Text-to-Speech: Instructions, Stimuli, and Items	45	4
Text-to-Speech: Items	3	
Text-to-Speech: Stimuli and Items	10	
Non-Embedded Designated Supports		
100s Number Table		
Abacus		
Amplification		
Bilingual Dictionary	1	
Calculator		1
Color Contrast		
Color Overlays		
Magnification		
Medical Supports	1	
Multiplication Table		
Noise Buffers		
Read Aloud Items		
Read Aloud Stimuli		
Read Aloud Stimuli and Items	1	
Scribe		
Separate Setting	1	
Simplified Test Directions		
Translated Student Interface Messages	1	

3.6.3 Accommodations

An accommodation may be provided for an English language learner (ELL), Individuals with Disabilities Education Act (IDEA)-eligible, or Section 504 Plan student. An accommodation is a practice or procedure in presentation, response, setting, timing, or scheduling that, when used in testing, provides equal access to all students. State-approved accommodations do not compromise the learning expectations, constructs, grade-level standards, or measured outcome of the assessment.

In the 2022–2023 administration, accommodations were granted based on the needs of individual students, not to a group of students or an entire class without investigation of need. Table 7 lists four accommodations that were available in 2022–2023. There were no students who were provided with these accommodations in 2022–2023. TCs were required to submit the Accommodations Verification Form to the HIDOE Assessment Section for verification of student need for the accommodation and, if necessary, set the accommodation in TIDE.

Table 7. Allowable Accommodations in 2022–2023

Available Non-Embedded Accommodations
Alternate Response Options: Students with some physical disabilities, including both fine motor and gross motor skills, may need to use adapted keyboards, StickyKeys, MouseKeys, FilterKeys, adapted mouse, touch screen, head wand, and switches.
Math Manipulatives: This accommodation allows eligible students with IEPs and 504 Plans to represent their understanding of mathematical concepts using visual and tactile concrete materials. This list of approved math manipulatives that may be provided on-site are: Algebra Tiles (recommended for grade 6 and above), Base Ten Blocks, Colored Tiles, Geoblocks Set, Geoboards and Geobands, Multi-Link Cubes, Pop Cubes, or Similar Cubes, Multi-Sensory Learning (MSL) Kit, One-Inch Blocks, Pattern Blocks, Transparent Sheets, and Two-Color Counters. Up to four manipulatives may be selected for a student; other accommodations not listed can be requested for verification.
Print-on-Demand: A student may request printed copies of individual test items and stimuli based on a documented need. A TC must request this accommodation for a student using the Appendix Q form in the TAM. It will then be preset for an approved student by the Assessment Section. TCs cannot set this accommodation in TIDE.
Speech-to-Text: Voice recognition allows students to use their voices as input devices to the computer to dictate responses or give commands (e.g., opening application programs, pulling down menus, and saving work). Voice recognition software generally can recognize speech up to 160 words per minute. Students may use their own assistive technology devices.

3.7 Data Forensics Program

The validity of test scores depends on the integrity of the test administration. Any irregularities in test administration could cast doubt on the validity of the inferences based on those test scores. Multiple facets ensure that tests are administered properly, including clear test administration policies, effective TA training, and tools to identify possible irregularities in test administrations.

For online administrations, a set of quality assurance (QA) reports is generated during and after the testing window. One of the QA reports focuses on flagging possible testing anomalies. Testing anomalies are analyzed by examining changes in student performance from year to year, test-taking time, item response patterns using a person-fit index, and item response change analyses. For the EOC exams, the score changes

are not examined because students have only one opportunity for an EOC exam when they are enrolled in the course.

Analyses are performed at the student level and summarized for each aggregate unit, including the testing session, TA, and school. The flagging criteria used for these analyses are described in the following section and are configurable by an authorized user. When the aggregate unit size is small, the aggregate unit is flagged if the percentage of flagged students is greater than 50% in the analysis. The default small aggregate unit size is five or fewer students but this value is configurable. For each aggregate unit, small groups are identified based on the number of tests included in the aggregate unit from that analysis. Thus, a small unit identified in one analysis may not be a small unit in another analysis. The QA reports are provided to state clients to monitor testing anomalies throughout the testing window.

3.7.1 Changes in Student Performance

Score changes are examined across opportunities within a year using a regression model. For within-year comparisons, the most recent opportunity is regressed on previous performance (second-most-recent score), controlling for the number of days between two scores, to identify performance gains or losses that are substantially greater than might reasonably be expected. Score comparison among past and current years is not possible because no previous-year performance is available for the EOC exam.

A large score gain or loss in student scores between administration years is detected by examining the residuals for outliers. The residuals are computed as the observed value minus the regression model's predicted value. The studentized residuals are computed to detect unusual residuals. An unusual increase or decrease in student scores between administration years is flagged when the absolute value of the studentized residual is greater than 3.

The residuals of students are also aggregated for a testing session, TA, and school. The system flags any unusual changes in an aggregate performance between administrations and/or years based on the average of the residuals in the aggregate unit (e.g., testing session, TA, school). For each aggregate unit, a t value is computed and flagged when |t| is greater than 3,

$$t = \frac{\sum_{i=1}^{n} \hat{e}_i \,/n}{\sqrt{\frac{s^2}{n} + \frac{\sum_{i=1}^{n} \sigma^2(1-h_{ii})}{n^2}}},$$

where *s* is the standard deviation of residuals in an aggregate unit; *n* is the number of students in an aggregate unit (e.g., testing session, TA, school), σ^2 is the MSE from the regression, and $\hat{e}_i$ is the residual for the *i*th student.

The total variance of residuals in the denominator is estimated in two components, conditioned on true residual e_i, $var\big(E(\hat{e}_i|e_i)\big) = s^2$ and $E\big(var(\hat{e}_i|e_i)\big) = \sigma^2(1-h_{ii})$. Following the law of total variance (Billingsley, 1995, p. 456),

$$var(\hat{e}_i) = var\big(E(\hat{e}_i|e_i)\big) + E\big(var(\hat{e}_i|e_i)\big) = s^2 + \sigma^2(1-h_{ii}), \text{ hence,}$$

$$var\left(\frac{\sum_{i=1}^{n}\hat{e}_i}{n}\right) = \frac{\sum_{i=1}^{n}\left(s^2+\sigma^2(1-h_{ii})\right)}{n^2} = \frac{s^2}{n} + \frac{\sum_{i=1}^{n}\left(\sigma^2(1-h_{ii})\right)}{n^2}.$$

3.7.2 Test-Taking Time

The summative assessments are not timed, and thus, individual test-taking times may vary across students. However, unusual test-taking times such as excessively shorter or longer test-taking times may indicate irregularities in test administration. An example of an unusual test-taking time is a test record for an individual who scores very well on the test even though the average time spent is far less than that required of students statewide. If students already know the answers to the questions, the test-taking time may be much shorter than the test-taking time for those who have no prior knowledge of the item content. Conversely, if a TA helps students by coaching them to change their responses during the test, the testing time could be longer than expected.

The state average testing time and standard deviation are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the test-taking time is different from the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

3.7.3 Inconsistent Item Response Pattern (Person Fit)

In item response theory (IRT) models, person-fit measurement is used to identify test takers whose response patterns are improbable given an IRT model. If a test has psychometric integrity, little irregularity will be seen in the item responses of the individual who responds to the items fairly and honestly.

If a test taker has prior knowledge of some test items (or is provided answers during the exam), he or she will respond correctly to those items at a higher probability than indicated by his or her ability as estimated across all items. In this case, the person-fit index will be large for the student. We note, however, that if a student has prior knowledge of the entire test content, this will not be detected based on the person-fit index, although the item response time index might flag such a student.

The person-fit index is based on all item responses in a test. An unlikely response to a single test question may not result in a flagged person-fit index. Of course, not all unlikely patterns indicate cheating, as in the case of a student who is able to guess a significant number of correct answers. Therefore, the evidence of person-fit index should be evaluated along with other testing irregularities to determine possible testing irregularities. The number of flagged students is summarized for every testing session, TA, and school.

The person-fit index is computed using a standardized log-likelihood statistic. Following Drasgow, Levine, and Williams (1985) and Sotaridona, Pornell, and Vallejo (2003), an aberrant response pattern is defined as a deviation from the expected item score model. Snijders (2001) showed that the distribution of l_z is asymptotically normal (i.e., with an increasing number of administered items). Even at shorter test lengths of 8 or 15 items, the "asymptotic error probabilities are quite reasonable for nominal Type I error probabilities of 0.10 and 0.05" (Snijders, 2001).

Sotaridona et al. (2003) report promising results of using l_z for systematic flagging of aberrant response patterns. Students with $l_z$values less than -3 are flagged. Aggregate units are flagged with *t* less than -3,

$$t = \frac{\text{Average } l_z \text{ values}}{\sqrt{s^2/n}},$$

where *s* = standard deviation of $l_z$values in an aggregate unit and *n* = number of students in an aggregate unit.

3.7.4 Item-Response Change

Students are allowed to revisit items as many times as they wish within a session and may also mark items to be revisited prior to completing the session. However, excessively high rates of response change, especially high rates of item score increases (i.e., response changes from wrong to right), may indicate irregularities in test administration. For example, TAs could review students' responses and either coach them to modify their responses or keep the session active and change responses themselves.

To identify irregular patterns of response change, we examine the item score for the final response to each item and the penultimate response if one exists, and then count the number of instances in which the item score increases.

The average and standard deviation of positive item score changes are computed based on all students available when the analysis was performed. Students and aggregate units are flagged if the number of positive item score changes is larger than the state average by three standard deviations or more, although the flagging criteria can be adjusted by an authorized user.

3.8 Prevention and Recovery of Disruptions in the Test Delivery System

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

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

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

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

3.8.1 High-Level System Architecture

Our architecture provides the redundancy, robustness, and reliability required by a large-scale, high-stake testing program. Our general approach, which has been adopted by Smarter Balanced as a standard policy, is pragmatic and well-supported by our architecture.

Any system built around an expectation of flawless performance of computers or networks within schools and districts is bound to fail. Our system is designed to ensure that the testing results and experience are

able to respond robustly to such inevitable failures. Thus, CAI's TDS is designed to protect data integrity and prevent student data loss at every point in the process.

The key elements of the testing system, including the data integrity processes at work at each point in the system, are described in the following sections. 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. Long responses, such as essays, are saved automatically at configurable intervals (usually set to one minute) so that student work is not at risk of being lost during testing.

Responses are saved asynchronously, with a background process on the student machine waiting for confirmation of successfully stored data on the server. If confirmation is not received within the designated time (usually set to 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. The following are examples of what can happen after connectivity is lost/the system fails:

- 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 options of logging out or retrying the save.
- If the system fails completely, upon logging back into 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 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) systems 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 continually gather data from the test delivery satellites and their mini-hubs and store that data as described in this section. This real-time backup copy remains on the hub until it receives 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, that provide 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 the information, these servers notify the hub and satellites that it is safe to delete student data.

Quality Assurance System

The QA system gathers data used to detect cheating, 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 immediate notification goes out to our psychometricians and project team.

Database of Record

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

3.8.2 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 loss of student data. Redundant systems at every point, real-time data integrity protection and checks, and well-considered, real-time backup processes prevent loss of student data, even in the unlikely event of system failure.

3.8.3 Other Disruption Prevention and Recovery

We have designed our system to be extremely fault-tolerant. The system can withstand failure of any component with little or no interruption of service. One way that we achieve this robustness is through redundancy. Key redundant systems are as follows:

- Our hosting provider has redundant power generators that can operate for up to 60 hours without refueling. With the multiple refueling contracts that are in place, these generators can operate indefinitely.
- Our 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.
- We use redundant power and switching within all of our server cabinets.
- Data are protected by nightly backups. We complete a full weekly backup and incremental backups nightly. Should a catastrophic event occur, CAI is able to reconstruct real-time data using the data retained on the TDS satellites and hubs.
- The server backup agents send alerts to notify system administration staff in the event of a backup error, at which time they will inspect the error to determine whether the backup was successful or if they will need to rerun the backup.

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 the student responses from loss.

4. SUMMARY OF SIMULATION STUDIES

Prior to the operational testing window, CAI conducts simulations to evaluate and ensure the implementation and quality of the adaptive item-selection algorithm and the scoring algorithm. The simulation tool enables us to manipulate key blueprint and configuration settings to match the blueprint and minimize measurement error and to maximize the number of different assessments seen by students.

4.1 SUMMARY OF ADAPTIVE ALGORITHM

For the online End-of-Course (EOC) exams, item selection rules ensure that each student receives an assessment representing an adequate sample of the domain with appropriate difficulty. The algorithm maximizes the information for each student and allows for certain constraints to be set, ensuring that the items selected represent the required content distribution. The test delivery system (TDS) ensures that students are not exposed to the same items or passages in subsequent assessments if they attempt multiple opportunities for the same content area.

Items selected for each student depend on the student's performance on previously selected items. The accuracy of the student responses to items determines the next items and passages that the student will see. Therefore, each student is presented with a set of items that most accurately aligns with his or her proficiency level based on grade-level content. Higher performance is followed by more difficult items, and lower performance is followed by less difficult items until test length constraints are met.

The adaptive algorithm selects the items to administer on each student's assessment to meet the following three objectives:

1. Match the test specifications (test blueprints).
2. Accurately classify test takers' proficiency in each content strand (or reporting category).
3. Minimize the measurement error by administering an assessment with items targeted to a student's ability.

For the first opportunity, the algorithm starts each assessment with an item of average difficulty near the average ability of students at the specific content area because no prior information about the test taker is available. All test takers in each EOC exam are assumed to have the same initial ability. Subsequent items are selected for administration by the algorithm based on student responses. For the subsequent opportunities, if a student takes the test more than once, the algorithm starts each assessment with the item or item set that best matches the student's estimated ability in the previous opportunity.

After the first item is administered, the algorithm identifies the best item to administer using the following criteria.

Match to the Blueprint

The algorithm first selects items to maximize fit to the test blueprint. Blueprints specify a range of items to be administered in each strand (reporting category) for each assessment, with a collection of constraint sets. A *constraint set* is a set of exhaustive, mutually exclusive classifications of items. For example, if a content area consists of four content strands, and each item measures one—and only one—of the strands, the content -strand classifications constitute a constraint set.

During item selection, the algorithm rewards strands that have not yet reached the minimum number of items. For example, if the measurement content strand requires that an assessment contain either eight or nine items, measurement is the constrained feature. At any point in time, the minimum constraint on some features may have already been satisfied, though others may not have been. Other features may be approaching the maximum defined by the constraint. The value measure must reward items that have not yet met minimum constraints and penalize items that would exceed the maximum constraints. The algorithm stops administering items when the specified assessment length is met.

Increased Precision

The adaptive algorithm can derive quickly and efficiently very precise estimates of student achievement. To increase the diagnostic value of score reports, the algorithm also seeks to increase the likelihood that a student's strand score will be clearly above or below the proficient-level performance standard. Thus, when selecting items from within each strand, the algorithm also values items that increase the likelihood function that a student's strand score is above or below the proficiency cut score. After identifying eligible items that meet the blueprint, the algorithm selects items that maximize the precision with which proficiency is assessed for each strand (reporting category) by selecting the best-fitting item from the available items within the targeted strand.

Match to Student Ability

In addition to rewarding items that match the blueprint, the adaptive algorithm also places greater value on items that maximize assessment information near the student's estimated ability, ensuring the most precise estimate of student ability possible, given the constraints of the item pool and satisfaction of the blueprint match requirement. After each response is submitted, the algorithm recalculates a score. As more answers are provided, the estimate becomes more precise, and the difficulty of the items selected for administration more closely aligns to the student's ability level. Higher performance (i.e., answering items correctly) is followed by more difficult items, and lower performance (i.e., answering items incorrectly) is followed by less difficult items. When the assessment is completed, the algorithm scores the overall assessment and each content strand.

The algorithm allows previously answered items to be changed, but it does not allow items to be skipped. Item selection requires iteratively updating the estimate of the overall and strand ability estimates after each item is answered. When a previously answered item is changed, the proficiency estimate is adjusted to account for the changed responses when the next new item is selected. Although an update of the ability estimates is performed at each iteration, the overall and strand scores are recalculated using all data at the end of the assessment for the final score.

The online EOC TDS administers assessments with items representing the breadth and depth identified in the test specifications and content standards. Because the assessment adapts to each student's performance while maintaining an accurate representation of the required knowledge and skills in content breadth and depth, the online EOC exam results provide precise estimates of each student's true performance level across the range of proficiency.

4.2 TESTING PLAN

The testing of the adaptive item-selection algorithm begins by generating a sample of test takers with true thetas from a normal distribution, with (μ,σ) for each grade and subject where μ and σ represent mean and

standard deviation of the normal distribution. The parameters for the normal distribution are based on student scores in the 2021–2022 operational tests.

Each simulated test taker is administered one test opportunity. In the first test opportunity for the simulation, the initial ability (prior ability) used to initiate the test by choosing the first few items is drawn from a uniform distribution within the range of true theta plus or minus 1. The starting theta is used to initiate the test by choosing the first few items.

Table 8 provides the means and standard deviations (SDs) used to generate a sample of student abilities in the simulation for EOC exams.

Table 8. Mean and Standard Deviation Used in Simulation

EOC Exams	Mean	SD
Algebra 1	−0.446	1.200
Algebra 2	−0.572	0.989

4.3 STATISTICAL SUMMARIES

The statistics computed include the statistical bias of the estimated theta parameter (statistical bias refers to if test scores systematically underestimate or overestimate the student's true ability); mean squared error (MSE); significance of the bias; average standard error of the estimated theta; and the standard error at the 5th, 25th, 75th, and 95th percentiles.

The computational details of each statistic is

$$bias = N^{-1}\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i),$$

$$MSE = N^{-1}\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i)^2,$$

where θ_i is the true theta and $\hat{\theta}_i$ is the estimated theta for individual *i*. For the variance of the bias, a first-order Taylor series is used as

$$var(bias) = \sigma^2 * g'(\hat{\theta}_i)^2 = \frac{1}{N(N-1)}\sum_{i=1}^{N}(\theta_i - \bar{\hat{\theta}}_i)^2,$$

where, $\bar{\hat{\theta}}_i$ is an average of the estimated thetas.

Significance of the bias is then tested as

$$z = bias/\sqrt{var(bias)}.$$

A *p*-value for the significance of the bias is reported from this *z* test.

The average standard error of the estimated theta is computed as

$$\boldsymbol{mean(se)} = \sqrt{\boldsymbol{N^{-1}\sum_{i=1}^{N} se(\hat{\theta})^2}},$$

where $se(\hat{\theta}_i)$ is the standard error of the estimated theta (θ) for individual *i*.

To determine the percentage of students' estimated theta falling outside the 95% and 99% confidence interval coverage, a *t*-statistic is performed as follows:

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)}$$

where $\hat{\theta}_i$ is the estimated theta for individual *i*, and θ_i is the true theta for individual *i*. The percentage of students' estimated theta falling outside the coverage is determined by comparing the absolute value of the *t*-statistic to a critical value of 1.96 for the 95% coverage and to 2.58 for the 99% coverage.

4.4 SUMMARY STATISTICS ON TEST BLUEPRINTS

In the adaptive item-selection algorithm, item selection takes place in two discrete stages: blueprint satisfaction and match to ability. Table 9 shows the percentages of simulated test forms that met test specifications exactly for each EOC exam. The table shows that all simulated test forms conform to the test specifications 100% in all subjects.

Table 9. Simulation: Blueprint Match Rate for Algebra 1 and Algebra 2

BP Constraints	Segment	Algebra 1			Algebra 2		
		Min	Max	% BP Match	Min	Max	% BP Match
Algebraic Concepts and Procedures	1	4	4	100.00	5	5	100.00
Modeling and Problem Solving	1	-	-	-	-	-	-
DOK 1	1	0	1	100.00	0	1	100.00
DOK 2	1	2	4	100.00	4	5	100.00
DOK 3	1	0	1	100.00	-	-	–
Selected Response	1	2	4	100.00	3	5	100.00
MSCR	1	0	2	100.00	0	2	100.00
Algebraic Concepts and Procedures	2	17	19	100.00	24	26	100.00
Modeling and Problem Solving	2	20	22	100.00	14	16	100.00
DOK 1	2	4	8	100.00	2	5	100.00
DOK 2	2	31	35	100.00	31	36	100.00
DOK 3	2	2	2	100.00	2	4	100.00
Selected Response	2	26	34	100.00	28	34	100.00
MSCR	2	8	10	100.00	8	10	100.00

4.5 SUMMARY STATISTICS ON ABILITY ESTIMATION

Each simulated test includes an initial ability, a true score, and an ability estimate based on the adaptive test administration. Table 10 shows statistical summaries of the ability estimation including mean of the biases, which is the average of the biases of estimated abilities (true ability – estimated ability) across all students and the *p*-value for the significance of the estimated bias reported from the *z*-test, providing the evidence needed to demonstrate that the true score is adequately recovered in the observed score. Table 10 also provides the percentages of students' estimated theta falling outside the 95% coverage and 99% coverage. The mean bias of the estimated abilities is very small and statistically insignificant in all EOC exams. The percentage of students' estimated theta falling outside the 95% and 99% confidence interval coverage is as expected within 5% and 1%, respectively.

Table 10. Bias of the Estimated Abilities for Simulated Tests

Subject	Bias	*p*-value	95% Coverage	99% Coverage
Algebra 1	-0.002	0.873	4.6	0.5
Algebra 2	0.004	0.666	4.1	0.6

Table 11 shows the mean standard error of the ability estimate across 1,000 simulated test administrations, as well as the standard error across the ability distribution. The average standard errors of the estimated abilities are similar across the ability ranges in all EOC exams, with a slightly larger standard error at the 5th percentile, indicating a shortage of easy items to better match the low-ability students.

The summary statistics of the estimated abilities show that for all test takers in all EOC exams, the item-selection algorithm chooses items that are optimized conditioned on each test taker's ability. Essentially, this shows that the test-taker ability estimates generated based on the items chosen are optimal in the sense that the final score for each test taker always recovers the true score within expected statistical limits. In other words, given that we know the true score for each test taker in a simulation, these results show that the true score is virtually always recovered—an indication that the algorithm is working exactly as expected for a computer-adaptive test (CAT).

Overall, these diagnostics on the item-selection algorithm provide evidence that scores are comparable with respect to the targeted content, and scores at various ranges of the score distribution are measured with precision.

Table 11. Standard Errors of the Estimated Abilities for Simulated Tests

Subject	Average Standard Error	MSE	SE at 5th Percentile	SE at Bottom Quartile	SE at Top Quartile	SE at 95th Percentile
Algebra 1	0.336	0.117	0.425	0.330	0.292	0.340
Algebra 2	0.323	0.099	0.387	0.323	0.301	0.305

Table 12 provides the correlations between true and estimated abilities and between estimated ability and average item difficulty (average item difficulty for each simulated test). The correlations between estimated ability and true score, reliability indexes, are high, indicating that the adaptive test administrations reliably estimate student ability. The correlations are also high between the estimated ability and the average difficulty (form difficulty) of the test administered to each student. The higher the correlations, the more adaptive the assessment. The high correlations demonstrate that the algorithm efficiently adapted to student ability.

Table 12. Correlations Between True Ability and Estimated Ability and Between Estimated Ability and Average Item Difficulty for Simulated Test

Subject	True Ability and Estimated Ability	Estimated Ability and Average Item Difficulty
Algebra 1	0.964	0.906
Algebra 2	0.954	0.915

4.6 ITEM EXPOSURE

The item exposure rate for each item was calculated by dividing the total number of test administrations in which an item appears by the total number of tests administered. Then, we reported the distribution of the item exposure rate (r) in six bins. The bins are $r = 0\%$ (unused), $0\% < r \leq 20\%$, $20\% < r \leq 40\%$, $40\% < r \leq 60\%$, $60\% < r \leq 80\%$, and $80\% < r \leq 100\%$. If global item exposure is minimal, we would expect the largest portion of items to appear in the $0\% < r \leq 20\%$ bin, an indication that most of the items appear on a very small percentage of the test forms.

Table 13 presents the percentages of items that fall into each exposure bin by EOC exam. The distribution of exposure rates is as expected, given the number of items in the blueprint constraints. Almost all items are administered in 20% or less test administrations.

Table 13. Percentage of Items by Exposure Rate

Subject	Total Items	Exposure Rate					
		Unused	0%–20%	21%–40%	41%–60%	61%–80%	81%–100%
Algebra 1	374	0%	87.43%	12.30%	0.27%	0%	0%
Algebra 2	417	0.72%	81.53%	15.35%	2.40%	0%	0%

5. MAINTENANCE OF THE ITEM BANK

5.1 ITEM RELEASE AND RETIREMENT POLICIES

Each year, Hawai'i releases a few items per content area. The released items selected include selected-response and machine-scorable construct-response (MSCR) items with a range of Depth of Knowledge (DOK) levels and item difficulties. All released items are posted on the Training Tests and Practice Tests Site at https://hsapt.tds.cambiumast.com/student. As the item pool gets larger, HIDOE plans to retire items that have become overexposed or outdated and replace them with new items.

5.2 FIELD-TESTING

5.2.1 Administration

HIDOE uses an embedded "operational" field-test design to augment items across content standards and benchmarks in the item pool. Each operational test embeds seven field-test items in the Algebra 1 EOC Exam, and nine field-test items in the Algebra 2 EOC Exam. All field-test items are aligned to the HCPS III.

Before the field-test items are embedded, all field-test items are reviewed by the CFAC prior to administration. Items are reviewed for (1) alignment to HCPS III and (2) potential bias, including language that might be disadvantageous to a group, be considered offensive to members of a particular group, or present obstacles to a group because of factors unrelated to content and processes specified in the standards. Only the items approved by the CFAC and HIDOE are embedded in the operational assessments.

Students are exposed to different field-test items in each opportunity. In each year, the field-test item development plan is based on a thorough review of the current operational item pool. The field-test items are developed to increase the number of items for all benchmarks covering a full range of item difficulties as well as to supplement the content standards and benchmarks, and MSCR item types that need to be increased. The field-test items were not counted toward student scores. Table 17 presents the number of field-test items embedded in SY2022-2023.

The Algebra 1 and Algebra 2 EOC Exam item pools include several types of field-test items which are detailed in Section 2.5.2 and Section 2.5.3.

Table 14. Number of Embedded Field-Test Items in EOC Exams

Test	EQ	ETC	GI	Total
Algebra 1	24	5	3	32
Algebra 2	16	2	-	18
Total	40	7	3	50

5.2.2 Sample Selection and Item Administration Selection Algorithm

CAI's field-test sampling algorithm is designed to yield an efficient, scientifically sound, representative random sample. The field-test item administered to a student is selected randomly from among those that have been *least frequently administered*, and this produces a similar sample size for all items with subgroup

compositions similar to the population for each item. This is a very powerful sample design that will yield a representative sample.

The algorithm employed by CAI's field-test engine ensures that

- efficient samples are used for all items by randomly selecting from among the least administered items to ensure that resulting item statistics (and DIF statistics) are maximally efficient across all items;
- position effects are averaged out by randomly administering test items across test positions; and
- more robust linkages exist among items, because each item is linked with every other item across hundreds of unique test "forms" to a degree not manageable through a set of fixed forms.

For pragmatic considerations related to system performance, the field-test algorithm limits the item selection to test start-up, instead of selecting each item in real time, as with the adaptive algorithm. Therefore, the field-test engine assigns all items at the beginning of each test administration.

Upon test start-up, the algorithm selects a series of items in the following iterative sequence:

1. Identify all the items that were least frequently administered
2. Randomly select an item with equal probability
3. Return to step 1 and continue if the requisite number of items has not been met

The first step initiates a random sequence. Note that for the first student, all items may be selected with equal probability. Subsequent selections depend on this first selection because the item is sampled without replacement until the entire pool has been administered, at which point the whole set becomes eligible for sampling again. This dependence is analogous to the dependence that occurs when test books are spiraled within classrooms—the probability of a student being assigned a particular book depends on the availability of books left to be handed out. In both cases the temporal dimension is incidental: the end result is a random distribution of items (or books) within the classroom.

We can see that the probability of administering an item from the pool is constant across individuals in the population. For a single grade and content area, let *n* represent the number of students, *k* represent the number of field-test items on the test, and *m* represent the total number of items in a pool. The expected number of times that any single item *j* will be administered can be calculated by $n_j = \frac{nk}{m}$. The corresponding probability that a given student *i* will receive item *j* is therefore $p_i(j) = \frac{n_j}{n} = \frac{k}{m}$.

From this we see that

- a random sample of students receives each item; and
- for any given item, the students are sampled with equal probability.

This design is both randomized, ensuring representation of the population and the validity of estimates of sampling error, and efficient.

The field-test algorithm also leads to randomization of item position and the context in which items appear. Field-testing each item in many positions and contexts should render the resulting statistics more robust to these factors.

5.3 EMBEDDED FIELD-TEST ITEM ANALYSES OVERVIEW

5.3.1 Rubric Validation

Prior to the analysis of the online field-test items, the rubrics for the MSCR items went through a validation process to refine the machine-scored rubrics. The rubric validation process is analogous to rangefinding for human-scored items, checking the validity of scoring rubrics as well as the scoring technology. The rubric validation process used a committee of content area experts (three or four teachers on each committee) to review student responses and propose changes to the machine-scored rubric.

During this process, a committee of content area experts reviewed a sample of student responses and proposes modifications to the rubric. CAI implemented these changes and reviews the resulting changes in scores. Only items that survive the rubric validation process are included in the operational pool.

CAI and HIDOE evaluated the impact of the revised rubrics on the scores of individual responses for the 2022–2023 embedded field-test items, and a final determination was made about changes to the rubrics. The committee found that a small number of items simply did not work and recommended to HIDOE that they be rejected. The rejected items were excluded from the online item pool.

Table 15 presents the number of MSCR items reviewed at the rubric validation committee and the number of items rejected by the reviewers. No field-test items were rejected.

Table 15. SY2022–2023 Summary Results of Rubric Validation Committee

Test	# of Reviewed Items	# of Items Rejected
	EQ	EQ
Algebra 1	2	0
Algebra 2	2	0

5.3.2 Field-Test Item Analyses

Once the scoring rubrics for all MSCR items are validated, all MSCR items are rescored using the final rubrics, and the final data file is extracted for the item analyses. The item analyses include classical item statistics and item calibrations using the Rasch-family IRT models. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (i.e., item discrimination) and to identify items that may exhibit a bias across subgroups (differential item functioning [DIF] analyses).

Item Difficulty

Items that are either extremely difficult or extremely easy are flagged for review but are not necessarily rejected if the item discrimination index is not flagged. For one point items, the proportion of examinees in the sample selecting the correct answer (*p*-values), as well as those selecting incorrect responses, is computed. For items with two or more points, item difficulty is calculated both as the item's mean score and as the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score

divided by the number of points possible). Items are flagged for review if the *p*-value is less than .25 or greater than .95.

Items with two or more points are flagged if the proportion of students in any score-point category is greater than .95. A very high proportion of students in any single score-point category may suggest that the other score points are not useful or, if the score point is in the minimum or maximum score-point category, that the item may not be grade-appropriate. Items with two or more points are also flagged if the average IRT-based ability estimate of students in a score-point category is lower than the average IRT-based ability estimate of students in the next lower score-point category. For example, if students who receive three points on a constructed-response item score, on average, lower on the total test than students who receive only two points on the item, the item is flagged. This situation may indicate that the scoring rubric is flawed.

Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those examinees who possess the skills being measured and those who do not. In general, the higher the value, the better the item is able to 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 (biserial correlations for multiple-choice items and polyserial correlations for constructed-response items). Multiple-choice items are flagged for subsequent reviews if the correlation for the item is less than .20 for the keyed (correct) response and greater than .05 for distractors. For constructed-response items, items are flagged if the polyserial correlation is less than .20.

Differential Item Functioning (DIF)

DIF analyses are designed to determine whether students at similar levels of ability have different probabilities of answering the same item correctly (or of receiving higher scores in the case of the items with two or more points), based on a group membership. In some cases, DIF may indicate item bias. However, a Fairness and Sensitivity Committee must review all items classified as DIF to determine whether an item is unfair to members of various student subgroup populations.

CAI conducted DIF analyses on all items included in the field test to detect potential item bias for subgroups. The performance on each item by focal group members (Hawaiian students, Filipino students, Japanese students, female students, English language learners, students with disabilities, and disadvantaged students) was compared with the performance of the appropriate reference group (white students, male students, not ELL students, not disability students, or not disadvantaged students), resulting in 10 sets of comparisons: Hawaiian/white, Filipino/white, Japanese/white, Hawaiian/Filipino, Hawaiian/Japanese, Filipino/Japanese, female/male, ELL/not ELL, disability/not disability, and disadvantaged/not disadvantaged. The purpose of these analyses was to identify items that may have favored students in one group (focal group) over students of similar ability in another group (reference group).

The procedures that CAI selected for detecting DIF were the Mantel-Haenszel (MH) chi-square for dichotomously scored items and Mantel's chi-square for polytomously scored items. CAI calculated the Mantel-Haenszel statistic (MH D-DIF) for multiple-choice items (Holland & Thayer, 1988) and the standardized mean difference (SMD) for polytomously scored items (Zwick, Donoghue, & Grima, 1993) to measure the degree and magnitude of DIF. Total scores for each student on the test were used as the ability-matching variable. The total score was divided into five intervals to compute the MH chi-square DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computed the MH chi-square value, the log-odds ratio, the standard error of the log-odds ratio,

and the MH-delta for the MC items, as well as the MH chi-square, the standardized mean difference (SMD), and the standard error of the SMD for the polytomously scored items. The purification method described by Holland and Thayer (1988) was included in the DIF procedure. Items were classified into three categories (A, B, or C) ranging from no DIF to mild DIF to severe DIF according to the DIF classification convention. Items were also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group, or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group. Table 16 details the DIF classification rules.

Table 16. DIF Classification Rules

Δ metric	
Category	*Rule*
C	$GMH\chi^2$ is significant at .05 and $\lvert\Delta_{MH}\rvert > 1.5$
B	$GMH\chi^2$ is significant at .05 and $1 < \lvert\Delta_{MH}\rvert \leq 1.5$
A	$GMH\chi^2$ is not significant at .05 or $\lvert\Delta_{MH}\rvert \leq 1$
SMD metric	
Category	*Rule*
C	$GMH\chi^2$ is significant at .05 and $\frac{\lvert SMD\rvert}{\sigma} > .25$
B	$GMH\chi^2$ is significant at .05 and $.17 < \frac{\lvert SMD\rvert}{\sigma} \leq .25$
A	$GMH\chi^2$is not significant at .05 or $\frac{\lvert SMD\rvert}{\sigma} \leq .17$

Items are flagged if their DIF statistics fall into the C category for any group. A DIF classification of C means that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. These items are flagged regardless of whether the DIF statistic favored the focal or referent group.

5.3.3 Field-Test Item Data Review Committee Meetings and Results

Although all field-test items had already been reviewed and approved by the CFAC, statistically flagged are further reviewed by the Content Data Review Committee and the Fairness Data Review Committee. The Content Data Review Committee consisted of HIDOE curriculum and assessment specialists as well as a few content area teachers. The Fairness Data Review Committee included community members, teachers, and HIDOE content area experts.

Table 17 summarizes the number of flagged items under the flag criteria described above. "%Flagged" indicates the percentage of flagged items among the total number of embedded field-test items. "%Rejected" indicates the percentage of rejected items among the total flagged items. The item difficulties and item discrimination correlations for all field-test items are included in Appendix B. The field-test item parameters and infit/outfit indices are provided in Appendix C. The DIF statistics for all field-test items are included in Appendix D.

Content Data Review Committee Meeting

The Content Data Review Committee reviewed the items flagged for item difficulty, item discrimination, and item fit index. Committee members examined the items for any indication that the items might have caused statistical flags. For each rejected item, the committee provided the content reason for the rejection.

The HIDOE and CAI content specialists reviewed the reasons and incorporated the reasons into the future item development process.

Fairness Data Review Committee Meeting

After data were collected, the DIF statistics were produced for the statistical review. A psychometric definition of the term *test fairness* is the degree to which an item performs differently for one group of examinees than for another group of equally able examinees. DIF refers to statistical properties of an item in two equally able groups and is subject to later interpretation and judgment. Once an item is flagged for a significant DIF, judgment should be used to decide whether the difference in difficulty shown by the DIF index is unfairly related to group membership.

The DIF statistics should be seen not as indicators of bias or unfairness but as indicators of the relative strengths and weaknesses of the two groups being compared when the overall ability that the test is intended to measure has been controlled. Items may show DIF because some concepts may be less likely to be covered in schools in low-income areas and because a lack of opportunity to learn can be manifested in these schools.

Table 17. SY2022–2023 Number of Flagged and Rejected Items in EOC Exams

Test	# of EFTs	Number of Flagged Items				Number of Rejected Items			
		Content Committee	Fairness Committee	Total Flagged Items	% Flagged	Content Committee	Fairness Committee	Total Rejected Items	% Rejected
Algebra 1	32	11	5	15	46.9%	3	2	4	26.7%
Algebra 2	18	5	2	7	38.9%	2	-	2	28.6%

5.4 ITEM CALIBRATION AND SCALING

5.4.1 Methodology

The EOC exam items are calibrated using the one-parameter Rasch model (Rasch, 1980; Wright & Stone, 1979) for selected-response items, scored dichotomously and the Rasch partial-credit model (Masters, 1982) for constructed-response items, scored polytomously. Calibrating mixed item types from different assessment modes (i.e., dichotomously and polytomously scored items) requires the use of a polytomous model, which allows the number of score categories (typically score points on a scoring rubric) to vary across assessment modes. The Rasch partial credit model (Wright & Masters, 1982) can accommodate the mixing of dichotomous and polytomous items.

The Winsteps software program (Linacre, 2011) is used in the item calibration. Winsteps employs a joint maximum likelihood approach to estimation (JMLE), which estimates the item and person parameters simultaneously. This estimation method is subject to small statistical biases, which increase as the length of the scale decreases. This estimation bias is corrected through the use of the Winsteps feature STBIAS=Y.

Under the Rasch model, the probability of a correct response conditional on ability is

$$\mathrm{p}\left(x_i = 1 \middle| \theta_j\right) = \frac{1}{1 + \exp\left[-\left(\hat{\theta}_j - b_{i,1}\right)\right]} \qquad (2)$$

where b_i is the location or difficulty parameter for the *i*th item, and x_i is the binary reponse to the *i*th item (where 1 = correct). The generalization for polytomous items in the partial credit model is

$$\mathrm{p}(\theta_j|x_i) = \frac{\exp\sum_{k=1}^{x_{ij}}(\hat{\theta}_j - \delta_{i,k})}{1 + \sum_{l=1}^{m_i}\left[\exp\sum_{k=1}^{l}(\hat{\theta}_j - \delta_{i,k})\right]} \qquad (3)$$

where the notation is the same as Equation (2) other than $\delta_{i,k}$, which is the k^{th} step for the *i*th item. Note that in the case of a dichotomous response item, the Masters' model reduces to the Rasch model.

5.4.2 Item Calibration

The online field-test design produces the field-test data in a sparse data matrix. The online field-test items in the sparse data matrix were concurrently calibrated fixing the pre-calibrated operational item parameters, placing the field-test items on the operational scale. All EOC items were calibrated using the Winsteps software program, version 5.2.2.0. Table 18 presents the average overall item difficulties by item type.

Table 18. SY2022–2023 Average Item Difficulty for Field-Test Items in EOC Exams

Test	Item Type			
	EQ	**ETC**	**GI**	**Total**
Algebra 1	2.07	0.25	0.98	1.69
Algebra 2	1.61	0.59	-	1.50

5.4.3 Item Fit Index

The item fit index is examined using the infit and outfit statistics. The infit statistic is more sensitive to the overall pattern of responses, less influenced by outliers, and more sensitive to patterns of observations by persons on items that are roughly targeted for them. The outfit statistic is highly influenced by a few outliers (very unexpected observations) by persons on items that are relatively very easy or very hard for them.

Table 19 presents the item fit summary for field-test items. The items are flagged for infit and outfit values larger than 2.0 or smaller than 0.5.

Table 19. SY2022–2023 Summary of Infit and Outfit Values for Field-Test Items in EOC Exams

Test	Number of FT Items	Infit				Outfit			
		< 0.5	**0.5–1.5**	**1.5–2.0**	**> 2.0**	**< 0.5**	**0.5–1.5**	**1.5–2.0**	**> 2.0**
Algebra 1	32		32			7	22	1	2
Algebra 2	18		18			4	14		

5.4.4 Item Dependency

IRT requires that the items in a test be locally independent once overall test performance is considered. Statistical independence in data occurs when student success on one item is not influenced by success on another. Local independence specifies that the score of one item has no influence on another once the underlying student ability has been accounted for (i.e., conditioned out). That is, when a pair of items are

locally independent, the conditional probability, given the student's ability level, θ, of obtaining any pair of scores on these items, is the product of the probabilities for the separate items as shown here:

$$P(X_1 = x_1 \text{ and } X_2 = x_2|\theta) = P(X_1 = x_1 |\theta)\ P(X_2 = x_2|\theta).$$

The traditional discrete items are usually carefully designed to be independent of one another, are not chained, and theoretically could be placed in any order without affecting the item difficulty. Yen (1984) introduced the Q_3 statistic as a measure of Local Item Dependency (LID). The Q_3 statistic is the correlation between performance on two items after overall test performance is considered. Winsteps produces a residual correlation matrix that corresponds to the Q_3 statistic. *Residual* is the deviation between the student's observed and expected item performances, given the student's ability level. High correlation of residuals for two items (or persons) indicates that they may not be locally independent, either because they duplicate some feature in each other or because they both incorporate some other shared dimension. Yen suggested Q_3 values $\geq$ .20 as an indication of LID. None of the field-test items were flagged by Yen's Q_3 criteria.

6. SUMMARY OF 2022-2023 OPERATIONAL TEST ADMINISTRATION

6.1 STUDENT POPULATION

All students (including retained students) currently enrolled in an Algebra 1 or Algebra 2 course at public schools or public charter schools in Hawai‘i have the option of taking the corresponding EOC exam. The HIDOE statewide student database is used to verify the courses in which each student is enrolled and student demographic information, such as the categories of gender, federal ethnic categories, English language learner, lunch program participation (disadvantaged), disability status, and migrant status.

The demographic compositions for the students who took the 2022–2023 EOC exams are shown in Table 20.

Table 20. Number of Students in 2022–2023 EOC Exams

Group	Algebra 1	Algebra 2
All Students	5,643	2,201
Female	2,658	1,056
Male	2,812	1,048
African American	76	28
American Indian/Alaskan Native	9	4
Asian/Pacific Islander	1,692	755
Hispanic	1,063	352
Hawai‘i Pacific Islander	1,096	374
White	590	195
Multi-Racial	944	396
English Language Learner	465	65
Disadvantaged	2,309	591
Disability	438	78
Migrant	19	8

6.2 OVERALL STUDENT PERFORMANCE

The 2022–2023 state summary results for the average scale scores and the percentage of students in each performance level overall and by subgroup are shown in Tables 21 and 22. Additionally, Table 23 provides the summary of student performance across grades, while Table 24 offers information on test administrations from 2014–2015 to 2022–2023. The 2019–2020 performance is not included because the testing was canceled due to the COVID-19 pandemic.

Table 21. Algebra 1 Percentage of Students in Performance Levels for Overall and by Subgroups

Group	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient
All Students	286.99	51.38	34	31	17	19	36
Gender							
Female	288.60	48.53	32	31	17	19	37
Male	289.00	52.46	33	30	17	20	37
Ethnicity							
African American	287.53	48.78	37	28	20	16	36
AmerIndian/Alaskan	–	–	–	–	–	–	–
Asian/Pacific Islander	304.37	51.95	20	30	21	29	50
Hispanic	277.79	44.35	39	34	15	12	27
Hawai‘i Pacific Islander	263.57	39.11	52	31	11	6	17
White	305.78	54.10	22	28	20	31	51
Multi-Racial	292.28	49.68	29	31	19	21	40
ELL Program							
ELL	253.27	40.09	62	27	7	5	11
Not ELL	290.01	51.20	31	31	18	20	38
Lunch Program							
Disadvantaged	270.86	46.46	46	31	12	11	23
Not Disadvantaged	298.15	51.66	25	30	20	25	45
Disability							
Disability	243.97	35.75	76	17	4	3	7
Not Disability	290.60	50.85	30	32	18	20	38
Migrant							
Migrant	303.22	30.07	5	42	32	21	53
Not Migrant	288.75	50.64	32	31	17	20	37

Note: The percentage of each performance level may not add up to 100% due to rounding.

"–" means that the data was suppressed due to small sample size, n < 10.

Table 22. Algebra 2 Percentage of Students in Performance Levels for Overall and by Subgroups

Group	Scale Score Mean	Scale Score SD	% Well Below	% Approaches	% Meets	% Exceeds	% Proficient
All Students	281.58	53.58	41	26	18	15	33
Gender							
Female	282.99	50.24	39	27	20	14	34
Male	284.79	55.70	40	26	17	16	34
Ethnicity							
African American	268.71	44.98	54	21	14	11	25
AmerIndian/Alaskan	–	–	–	–	–	–	–
Asian/Pacific Islander	299.29	57.71	28	26	22	24	45
Hispanic	271.89	47.87	51	28	11	11	22
Hawai'i Pacific Islander	260.84	39.57	58	25	14	4	17
White	284.42	49.68	35	30	22	14	35
Multi-Racial	287.88	51.16	35	27	22	16	38
ELL Program							
ELL	236.65	26.93	86	11	3	0	3
Not ELL	282.95	53.61	40	27	18	15	34
Lunch Program							
Disadvantaged	262.43	46.66	58	21	13	7	20
Not Disadvantaged	288.61	54.25	35	28	20	17	37
Disability							
Disability	240.38	38.13	77	18	3	3	5
Not Disability	283.10	53.47	40	26	19	15	34
Migrant							
Migrant	–	–	–	–	–	–	–
Not Migrant	284.08	53.00	39	27	19	15	34

Note: The percentage of each performance level may not add up to 100% due to rounding.
"–" means that the data was suppressed due to small sample size, n < 10.

Table 23. Percentage of Proficient Students Across Grades

Test	Grade	N	Scale Score Mean	Scale Score SD	% Proficient
Algebra 1	6	1*			
	7	211	329.77	49.87	73
	8	1,303	326.47	54.97	67
	9	3,460	275.70	41.65	27
	10	557	254.80	36.39	10
	11	95	255.36	31.14	8
	12	16	246.52	46.37	13
Algebra 2	9	141	336.75	57.90	74
	10	1,039	286.97	58.29	38
	11	816	271.03	40.42	23
	12	205	258.33	40.12	17

Note: ***** Data suppressed due to the small sample size, $n < 10$.

Table 24. Percentage of Proficient Students Across Test Administrations

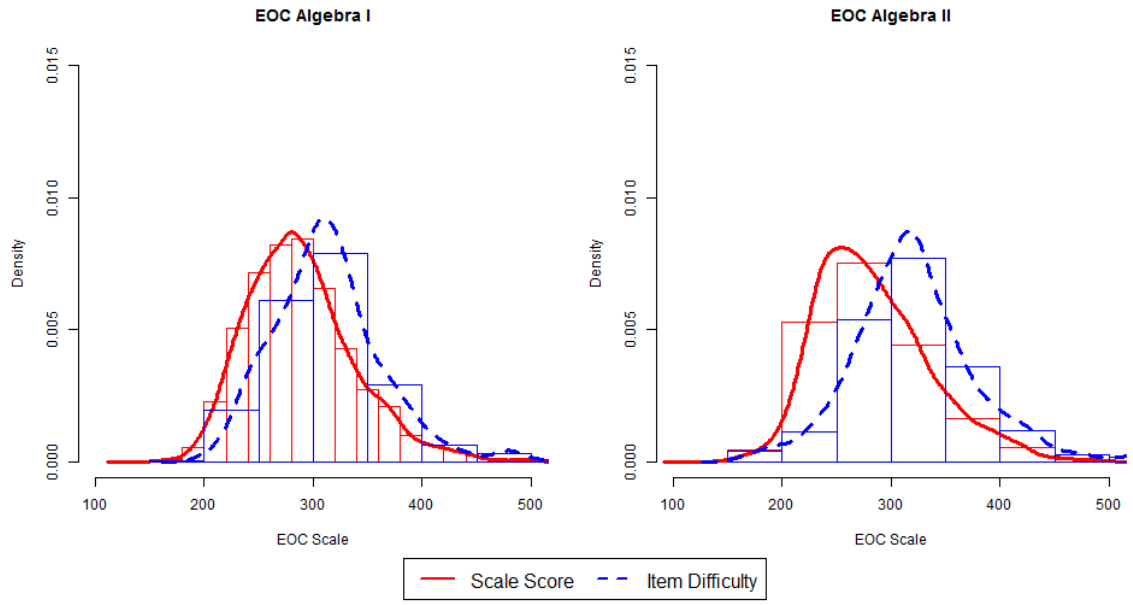
Subject	Year	N	Mean	SD	% Proficient
Algebra 1	2014–2015	8,239	293.49	47.15	42
	2015–2016	6,332	298.06	51.14	45
	2016–2017	5,927	302.78	51.78	49
	2017–2018	5,721	307.50	51.86	53
	2018–2019	7,627	297.15	52.05	45
	2020–2021	1,688	302.95	50.13	48
	2021-2022	5,444	288.18	53.61	36
	2022-2023	5,643	286.99	51.38	36
Algebra 2	2014–2015	7,586	284.23	44.06	33
	2015–2016	4,100	289.51	44.91	38
	2016–2017	2,990	299.57	47.63	47
	2017–2018	2,792	302.61	48.70	49
	2018–2019	3,405	300.55	50.35	45
	2020–2021	571	297.57	55.24	43
	2021-2022	1,967	287.58	50.67	36
	2022-2023	2,201	281.58	53.58	33

Note: There was no testing in 2019–2020 due to the COVID-19 pandemic.

6.3 Student Ability–Item Difficulty Distribution for the 2022–2023 Operational Item Pool

Figure 1 shows the empirical distribution of the student scaled scores in the 2022–2023 administration and the distribution of the EOC exam item difficulty parameters in the operational pool. The student ability distribution is shifted to the left in Algebra 1 and Algebra 2, indicating that the pool includes a larger number of difficult items than the ability of students in the tested population requires.

Figure 1. Student Ability–Item Difficulty Distribution for Algebra 1 and Algebra 2



7. VALIDITY

According to 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), validity refers to the degree to which evidence and theory support the interpretations of test scores as described by the intended uses of assessments. The 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 General Summative Assessments depends on the assessments meeting the relevant standards of validity.

Validity evidence provided in this chapter is as follows:

- Test Content
- Internal Structure
- Relations to Other Variables (External Structure)

Evidence on test content validity is provided with the item alignment to the HCPS III and CCSS, and the blueprint match rates for the delivered tests. Evidence on internal structure is examined in the results of intercorrelations among reporting category scores. Evidence on external structure is examined in the relationships among Smarter Balanced ELA/L and mathematics scores and EOC scores.

Some of the evidence on standardized test administration, scoring procedures, and attention to fairness for all test takers is provided in other chapters.

7.1 EVIDENCE ON TEST CONTENT

7.1.1 Alignment of EOC Item Banks to the HCPS III and the CCSS

As a criterion-referenced system of tests, the meaning of test scores is, in part, appropriately evaluated by the degree to which test content is aligned with the HCPS III and CCSS. Alignment of item contents to the HCPS III and the CCSS is achieved through a highly iterative test development process that includes HIDOE, CAI, and two committees comprising Hawai‘i educators and other stakeholders. The evidence of content validity is also provided in Section 2, Test Development.

7.1.2 Fidelity to Test Blueprints

The statistical information of content distribution is summarized in the blueprint match rate for all tests. Blueprints specify a range of items to be administered in each strand (reporting category), by item type (selected-response items and machine-scored constructed-response [MSCR] items) and Depth of Knowledge (DOK). Tables 25 and 26 show the percentages of tests aligned with the test specifications for Algebra 1 and Algebra 2 by subgroup. In all EOC tests, all adaptively delivered tests met the test blueprint.

The content distribution of each test was the same for all students (e.g., general education students, EL students, students with disabilities) indicating the validity and comparability of all tests across all students. The high blueprint-match rates for assessments indicate that all assessments are equivalent in content

coverage and produce comparable scores using the item parameters from the same item pool, ensuring the comparability of assessments in content and scores.

Table 25. Blueprint Match Rate in 2022–2023 EOC Algebra 1 by Subgroup

Blueprint Constraints	Segment	Min	Max	General Education	ELL	Disability
Overall	1	4	4	100.00%	100.00%	100.00%
Algebraic Concepts and Procedures	1	4	4	100.00%	100.00%	100.00%
DOK 1	1	0	1	100.00%	100.00%	100.00%
DOK 2	1	2	4	100.00%	100.00%	100.00%
DOK 3	1	0	1	100.00%	100.00%	100.00%
SR	1	2	4	100.00%	100.00%	100.00%
MSCR	1	0	2	100.00%	100.00%	100.00%
Overall	2	39	39	100.00%	100.00%	100.00%
Algebraic Concepts and Procedures	2	17	19	100.00%	100.00%	100.00%
Modeling and Problem Solving	2	20	22	100.00%	100.00%	100.00%
DOK 1	2	4	8	100.00%	100.00%	100.00%
DOK 2	2	31	35	100.00%	100.00%	100.00%
DOK 3	2	2	2	100.00%	100.00%	100.00%
SR	2	26	34	100.00%	100.00%	100.00%
MSCR	2	8	10	100.00%	100.00%	100.00%

Table 26. Blueprint Match Rate in 2022–2023 EOC Algebra 2 by Subgroup

Blueprint Constraints	Segment	Min	Max	General Education	ELL	Disability
Overall	1	5	5	100.00%	100.00%	100.00%
Algebraic Concepts and Procedures	1	5	5	100.00%	100.00%	100.00%
DOK 1	1	0	1	100.00%	100.00%	100.00%
DOK 2	1	4	5	100.00%	100.00%	100.00%
SR	1	3	5	100.00%	100.00%	100.00%
MSCR	1	0	2	100.00%	100.00%	100.00%
Overall	2	40	40	100.00%	100.00%	100.00%
Algebraic Concepts and Procedures	2	24	26	100.00%	100.00%	100.00%
Modeling and Problem Solving	2	14	16	100.00%	100.00%	100.00%
DOK 1	2	2	5	100.00%	100.00%	100.00%
DOK 2	2	31	36	100.00%	100.00%	100.00%
DOK 3	2	2	4	100.00%	100.00%	100.00%
SR	2	28	34	100.00%	100.00%	100.00%
MSCR	2	8	10	100.00%	100.00%	100.00%

7.1.3 Benchmark or Standard Coverage

Table 27 summarizes the number of unique benchmarks or standards administered in each delivered test. The table includes the number of benchmarks or standards specified in the blueprints, and the mean and the range of the number of benchmarks administered to students. The test blueprints do not require each test to include items for every benchmark; however, all delivered tests covered almost all benchmarks in Algebra 1 and Algebra 2. The computer-adaptive test (CAT) delivers a test covering more standards or benchmarks with more precision than a fixed-form test.

Table 27. Distribution of Standards and Benchmarks Covered in Each Delivered Test

EOC Subject	Subgroup	Number of CCSS or HCPS III Covered in Blueprint	Average	Min	Max
Algebra 1	All	27	25.6	23	27
	General	27	25.6	23	27
	ELL	27	25.5	24	27
	Disability	27	25.6	24	27
Algebra 2	All	57	40.4	36	45
	General	57	40.4	36	45
	ELL	57	40.0	37	43
	Disability	57	40.6	37	45

7.2 Evidence on Internal Structure

The measurement and reporting model used in Hawai'i assumes a single underlying latent trait, with achievement reported as a total score, as well as scores for each reporting category measured. The evidence on the internal structure is examined based on the correlations among reporting category scores.

The observed and attenuated correlations among reporting category scores are shown in Table 28. The correction for attenuation indicates what the correlation would be if reporting category scores could be measured with perfect reliability. The observed correlation between two reporting category scores with measurement errors can be corrected for attenuation as $r_{x'y'} = r_{xy} / SQRT(r_{xx} * r_{yy})$, 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*.

When corrected for attenuation, the correlations among reporting scores are quite high, indicating that the assessments measure a common underlying construct.

Table 28. Correlations Among Reporting Category Scores for Algebra 1 and Algebra 2

EOC Exam	Reporting Categories	Observed Correlation		Disattenuated Correlation	
		ACP	MPS	ACP	MPS
Algebra 1	Algebraic Concepts and Procedures (ACP)	1		1	
	Modeling and Problem Solving (MPS)	0.80	1	0.95	1
Algebra 2	Algebraic Concepts and Procedures (ACP)	1		1	
	Modeling and Problem Solving (MPS)	0.77	1	0.98	1

7.3 EVIDENCE ON RELATIONS TO OTHER VARIABLES

Validity evidence based on relationships to other variables can address a variety of questions. At its core, this type of validity addresses the relationships between test scores and variables of interest that are derived outside the testing system. One type of validity evidence based on relations to other variables is evidence for convergent and discriminant validity. Evidence for convergent validity is based on the degree to which test scores correlate with other measures of the same attribute (i.e., scores from two tests measuring the same attribute should be correlated). Conversely, when test scores are not correlated with measures of construct irrelevant attributes, evidence is obtained for discriminant validity.

Evidence for convergent and discriminant validity is determined by examining the patterns of correlations among Hawai‘i’s course-specific statewide assessments and performance on the Smarter Balanced ELA/L and mathematics assessments. Observed correlations between alternate indicators of student achievement of course objectives, such as Hawai‘i’s statewide assessment scores, should be limited only by the unreliability of the measures.

When both assessments measure student achievement in common subject areas, as with, for example, test scores based on statewide assessments in Algebra, the correlations between test scores are expected to be substantially correlated. Additionally, the magnitude of observed correlations among test scores in different subject areas is expected to be lower than correlations among test scores in a common subject area. It is important to note, however, that test scores across subject areas and test systems are nevertheless expected to be highly correlated. This is because even though subject-area test scores measure different academic content domains, student achievement across subject areas is influenced by factors both internal (e.g., general intelligence) and external (e.g., socioeconomic status) to the student that contribute to student achievement across all academic subject areas. Therefore, student test scores across subject areas are highly intercorrelated. Although we certainly do expect correlations between test scores across subject areas to be lower than correlations between test scores within a subject area, we nevertheless expect test scores across subject areas to be quite high.

Table 29 provides the correlations between the EOC exam scores and the Smarter Balanced ELA/L and mathematics test scores. As expected, the magnitude of observed correlations among test scores in different subject areas was lower than correlations among test scores in a common subject area, which is evidence for convergent and discriminant validity. The correlation coefficients among the test scores are moderate, with higher correlations among scores with common or similar traits. Algebra 1 and Algebra 2 scores are correlated higher with Smarter Balanced mathematics scores than with Smarter Balanced ELA/L scores.

Table 29. Correlations Between EOC Scores with Other Test Scores

EOC Exams	Smarter ELA/L		Smarter Mathematics	
	N	Correlation	N	Correlation
Algebra 1	1,583	0.64	1,589	0.84
Algebra 2	808	0.46	809	0.68

7.4 EVIDENCE OF COMPARABILITY

The same precision across the range of ability for subgroups and the same content distribution for all tests and subgroups indicate the comparability of test forms among students. An adaptive testing algorithm constructs a test form unique to each student, targeting the student’s level of ability and how well the test matches the test blueprints. Consequently, the test forms will not be statistically parallel (e.g., equal test

difficulty). However, test scores should be comparable, and each test form should measure the same content, albeit with a different set of test items.

7.5 FAIRNESS AND ACCESSIBILITY

7.5.1 Fairness in Content

The principles of universal design (UD) of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. UD removes barriers to access for the widest range of students possible. The following seven principles of UD are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Test development specialists receive extensive training on the principles of UD and apply the principles to the development of all test materials, including tasks, items, and manipulatives. In the review process, adherence to the principles of UD is verified.

7.5.2 Statistical Fairness in Item Statistics

All field-test items were reviewed before being included in the item pool to be field tested. They were also analyzed for fairness to all students. When new items are developed, the Content and Fairness Advisory Committee (CFAC) reviews the items using the *CAI Guidelines for Language Accessibility, Bias, and Sensitivity* (Appendix A). After the field-test item analyses, the items flagged with the C category for any group in the differential item functioning (DIF) statistics were reviewed for any indications that they might have caused a significant DIF.

The DIF analyses were performed for the following groups:

- Hawai‘ian/White
- Filipino/White
- Japanese/White
- Hawai‘ian/Filipino
- Hawai‘ian/Japanese
- Filipino/Japanese
- Female/Male
- ELL/not ELL

- Students with Disability/Students without Disability
- Disadvantaged/Not Disadvantaged

The purpose of these analyses is to identify items that may have favored students in one group (focal group) over students of similar ability in another group (reference group).

8. RELIABILITY

Reliability refers to the consistency in test scores. Reliability is evaluated in terms of the standard error of measurement (SEM). In classical test theory, reliability is defined as the ratio of the true score variance to the observed score variance, assuming that the error variance is the same for all scores. Within the item response theory (IRT) framework, measurement error varies conditioning on ability. The level 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 the inverse of the measurement error of the test; the larger the measurement error, the less test information is being provided. In CATs, because selected items vary among students, the measurement error can vary for the same ability depending on the selected items for each student.

The reliability evidence of the EOC exams is provided with marginal reliability, SEM, and decision accuracy and consistency at each performance level.

8.1 MARGINAL RELIABILITY

For reliability, *marginal reliability* was computed for 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 standard errors of measurement, estimated at different points on the ability scale, for all students.

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

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

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

Another way to examine test reliability is with the SEM. In IRT, SEM is estimated as a function of test information provided by a given set of items that make up the test. In CATs, the items administered vary across 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 value of average CSEM indicates the greater accuracy of test scores.

Table 30 shows the marginal reliability coefficients and the average SEM for the total scale scores for overall and by subgroups.

Table 30. Marginal Reliability for Algebra 1 and Algebra 2

Subgroup	Algebra 1				Algebra 2			
	MR	SS	SD	CSEM	MR	SS	SD	CSEM
All Students	0.91	286.99	51.38	15.42	0.90	281.58	53.58	17.29
Female	0.91	288.60	48.53	14.82	0.89	282.99	50.24	16.63
Male	0.92	289.00	52.46	15.05	0.91	284.79	55.70	16.78
African American	0.91	287.53	48.78	14.75	0.86	268.71	44.98	16.91
American Indian/Alaskan Native	–	–	–	–	–	–	–	–
Asian/Pacific Islander	0.92	304.37	51.95	14.85	0.92	299.29	57.71	16.57
Hispanic	0.89	277.79	44.35	14.86	0.88	271.89	47.87	16.92
Hawai'i Pacific Islander	0.85	263.57	39.11	15.17	0.81	260.84	39.57	17.10
White	0.92	305.78	54.10	15.03	0.89	284.42	49.68	16.51
Multi-Racial	0.91	292.28	49.68	14.87	0.90	287.88	51.16	16.50
English Language Learner	0.83	253.27	40.09	16.62	0.52	236.65	26.93	18.61
Disadvantaged	0.88	270.86	46.46	15.91	0.85	262.43	46.66	18.20
Disability	0.78	243.97	35.75	16.58	0.77	240.38	38.13	18.30
Migrant	0.80	303.22	30.07	13.56	–	–	–	–

Note: "–" means that the data was suppressed due to small sample size, n < 10.

8.2 STANDARD ERROR OF MEASUREMENT

Table 31 provides the average CSEM within each performance level and the average CSEMs at each cut score. Consistent with the simulation results in Section 4.5, Summary Statistics on Ability Estimation, the largest standard error is shown in the "Well Below" performance level in all EOC exams. However, average CSEMs are very similar at all cut scores.

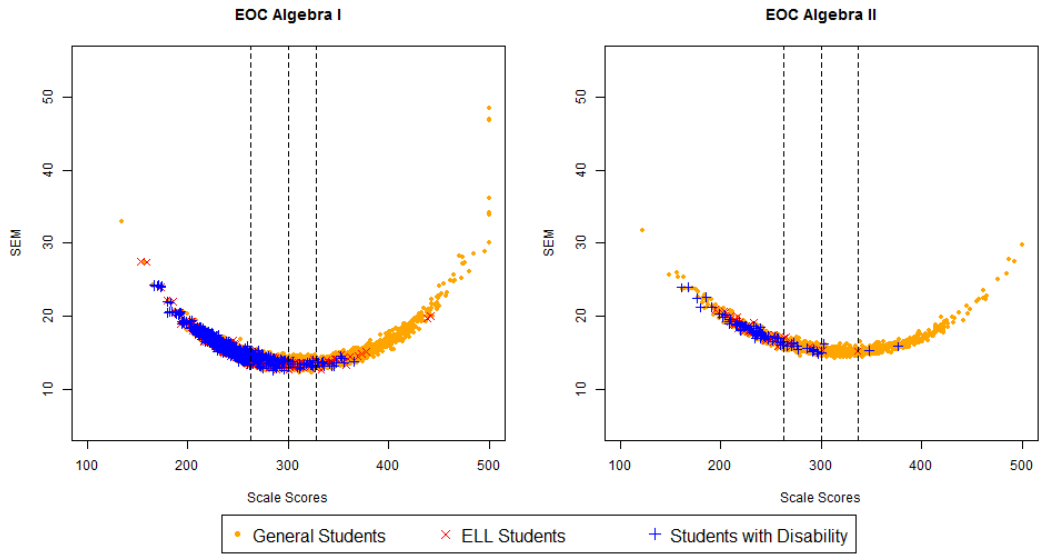
Table 31. Average Conditional Standard Error of Measurement
by Performance Level and at Each Performance-Level Cut Score

Test	Well Below	Approaches	Meets	Exceeds	Total	Approaches Cut	Meets Cut	Exceeds Cut
Algebra 1	17.21	14.01	13.50	15.81	15.42	14.52	13.78	13.53
Algebra 2	19.04	16.05	15.21	16.59	17.29	17.08	15.39	15.05

Figure 2 plots the CSEM across the range of ability by subgroups for the Algebra 1 and Algebra 2 scores obtained in 2022–2023. The item-selection algorithm selected the items efficiently, matching to each student's ability while also matching to the test blueprints, with the same precision across the range of abilities for all students (e.g., general education students, EL students, students with disabilities). The "general education students" subgroup excludes EL students and students with disabilities from the total number of students who received a family report in each grade and content area. The vertical lines indicate the cut scores for "Approaches," "Meets," and "Exceeds."

Overall, the standard error curves suggest that students are measured with a very high degree of precision, given that the standard errors are consistently low. However, larger standard errors are observed at the two ends of the score distribution. Content experts use this information to consider how to further target and populate item pools. The standard errors across score points are also the same across subgroups, indicating the same precision on score points.

Figure 2. Conditional Standard Error of Measurement by Subgroup



8.3 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

When student performance is reported in terms of performance levels, a reliability of performance classification is computed in terms of the probabilities of consistent classification of students as specified in Standard 2.16 in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014). This index considers the consistency of classifications for the percentage of test takers that would, hypothetically, be classified in the same category on an alternate, equivalent form.

For a fixed-form test, the consistency of classifications is estimated on a single-form test score 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 CATs, because the adaptive testing algorithm constructs a test form unique to each student, targeting the student's level of ability while meeting test blueprint requirements, the consistency of classifications is based on all sets of items administered across students using an IRT -based method (Guo, 2006).

The classification index can be examined for the decision accuracy and the decision consistency. *Decision accuracy* refers to the agreement between the classifications based on the form actually administered and the classifications that would be made based on the test takers' true scores if their true scores could somehow be known. *Decision consistency* refers to the agreement between the classifications based on the form (adaptively administered items) actually taken and the classifications that would be made based on 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, true ability is unknown, and students do not take an alternate, equivalent form; therefore, the classification accuracy and consistency are estimated based on students' item scores and the item parameters, and the assumed underlying latent ability distribution. The true score is an expected value of the test score with a measurement error.

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

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

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

The likelihood function of theta, given a student's item scores, represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut point (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut score. If a student with estimated theta is below the cut score, the 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.

The probability of the *i*th student being classified at performance level *l* ($l = 1,2,\cdots,L$) based on the cut scores cut_{l-1} and cut_l, given the student's item scores $\mathbf{z}_i = (z_{i1},\cdots,z_{iJ})$ and item parameters $\mathbf{b} = (\mathbf{b}_1,\cdots,\mathbf{b}_J)$, using the *J* administered items, can be estimated as

$$p_{il} = P(cut_{l-1} \le \theta_i < cut_l|\mathbf{z},\mathbf{b}) = \frac{\int_{cut_{l-1}}^{cut_l} L(\theta|\mathbf{z},\mathbf{b})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\mathbf{z},\mathbf{b})d\theta},$$

where the likelihood function, based on general IRT models, is

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

where, d stands for dichotomous and p stands for polytomous items, $\mathbf{b}_j = (a_j, b_j, c_j)$ if the *j*th item is a dichotomous item, and $\mathbf{b}_j = (a_j, b_{j1}, \ldots, b_{jK_i})$ if the *j*th item is a polytomous item, a_j is the item's discrimination parameter (for Rasch model, $a_j = 1$), c_j is the guessing parameter (for Rasch and 2PL models, $c_j = 0$), *D* is 1.7 for non-Rasch models and 1 for Rasch model. For level 1, $cut_0 = -\infty$, and for level *L*, $cut_L = \infty$.

Classification Accuracy

Using p_{il}, we can construct a $L \times L$ table as

$$\begin{pmatrix} n_{a11} & \cdots & n_{a1L} \\ \vdots & \vdots & \vdots \\ n_{aL1} & \cdots & n_{aLL} \end{pmatrix}$$

where $n_{alm} = \sum_{pl_i=l} p_{im}$, pl_i is the *i*th student's performance level. In the above table, the row represents the observed level, and the column represents the expected level.

Based on the above table, the classification accuracy (*CA*) for cut_l $(l = 1, \cdots, L-1)$ is estimated by

$$CA_{cut_l} = \frac{\sum_{k,m=1}^{l} n_{akm} + \sum_{k,m=l+1}^{L} n_{akm}}{N},$$

and the overall classification accuracy is estimated by

$$CA = \frac{\sum_{l=1}^{L} n_{all}}{N},$$

where *N* is the total number of students.

For classification accuracy, the false positive (*FP*) for $cut_l (l = 1, \cdots, L-1)$ is estimated by

$$FP_{cut_l} = \frac{\sum_{m=1}^{l} \sum_{k=l+1}^{L} n_{akm}}{N},$$

and the false negative (*FN*) for $cut_l (l = 1, \cdots, L-1)$ is estimated by

$$FN_{cut_l} = \frac{\sum_{k=1}^{l} \sum_{m=l+1}^{L} n_{akm}}{N}.$$

The overall false positive is estimated by

$$FP = \frac{\sum_{m=1}^{L} \sum_{k=m+1}^{L} n_{akm}}{N}.$$

The overall false negative is estimated by

$$FN = \frac{\sum_{k=1}^{L} \sum_{m=k+1}^{L} n_{akm}}{N}.$$

Classification Consistency

Using p_{il}, similar to accuracy, we can construct another $L \times L$ table by assuming the test is administered twice independently to the same student group; hence, we have

$$\begin{pmatrix} n_{c11} & \cdots & n_{c1L} \\ \vdots & \vdots & \vdots \\ n_{cL1} & \cdots & n_{cLL} \end{pmatrix}$$

where $n_{clm} = \sum_{i=1}^{N} p_{il} p_{im}$. p_{il} and p_{im} are the probabilities of the *i*th student being classified at achievement level *l* and *m*, respectively, based on observed scores and hypothetical scores from an equivalent test form.

The classification consistency (*CC*) at level l $(l = 1, \cdots, L)$ is estimated by

$$CC_l = \frac{n_{cll}}{\sum_{m=1}^{L} n_{clm}},$$

and the overall classification consistency is

$$CC = \frac{\sum_{l=1}^{L} n_{cll}}{N}.$$

Cohen's Coefficient Kappa Index

The probability of classification accuracy by chance, p_{ca}, is the sum of the marginal probabilities of classifications into the same level based on observed and expected classifications; hence, for cut_l ($l = 1, \cdots, L-1$), this is estimated by

$$p_{cal} = p_{cal1} + p_{cal2},$$

where

$$p_{cal1} = \left(\frac{\sum_{k,m=1}^{l} n_{akm}}{N} + \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{akm}}{N}\right)\left(\frac{\sum_{k,m=1}^{l} n_{akm}}{N} + \frac{\sum_{k=1}^{l}\sum_{m=l+1}^{L} n_{akm}}{N}\right),$$

$$p_{cal2} = \left(\frac{\sum_{k,m=l+1}^{L} n_{akm}}{N} + \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{akm}}{N}\right)\left(\frac{\sum_{k,m=l+1}^{L} n_{akm}}{N} + \frac{\sum_{k=1}^{l}\sum_{m=l+1}^{L} n_{akm}}{N}\right).$$

For the overall classification accuracy, the chance probability is estimated by

$$p_{ca} = \sum_{l=1}^{L}\left(\frac{\sum_{m=1}^{L} n_{alm}}{N}\right)\left(\frac{\sum_{m=1}^{L} n_{aml}}{N}\right),$$

and Cohen's coefficient kappa (Cohen, 1960) is estimated by $\frac{CA_{cut_l} - p_{cal}}{1 - p_{cal}}$ for the classification accuracy at cut_l, and $\frac{CA - p_{ca}}{1 - p_{ca}}$ for the overall classification accuracy.

Similarly, the same calculations can be conducted for classification consistency. Hence, for cut_l ($l = 1, \cdots, L-1$), the chance probability is estimated by

$$p_{ccl} = p_{ccl1} + p_{ccl2},$$

where

$$p_{ccl1} = \left(\frac{\sum_{k,m=1}^{l} n_{ckm}}{N} + \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{ckm}}{N}\right)\left(\frac{\sum_{k,m=1}^{l} n_{ckm}}{N} + \frac{\sum_{k=1}^{l}\sum_{m=l+1}^{L} n_{ckm}}{N}\right),$$

$$p_{ccl2} = \left(\frac{\sum_{k,m=l+1}^{L} n_{ckm}}{N} + \frac{\sum_{m=1}^{l}\sum_{k=l+1}^{L} n_{ckm}}{N}\right)\left(\frac{\sum_{k,m=l+1}^{L} n_{ckm}}{N} + \frac{\sum_{k=1}^{l}\sum_{m=l+1}^{L} n_{ckm}}{N}\right).$$

For the overall classification consistency, the chance probability is estimated by

$$p_{cc} = \sum_{l=1}^{L}\left(\frac{\sum_{m=1}^{L} n_{clm}}{N}\right)\left(\frac{\sum_{m=1}^{L} n_{cml}}{N}\right),$$

and Cohen's coefficient kappa is estimated by $\frac{CC_{cut_l} - p_{ccl}}{1 - p_{ccl}}$ for the classification consistency at cut_l, and $\frac{CC - p_{cc}}{1 - p_{cc}}$ for the overall classification consistency.

Table 32 shows the classification accuracy and consistency indexes. Accuracy classifications are slightly higher (1–4%) than the consistency classifications in all performance levels. The consistency classification rate can be lower because the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score. The classification index ranged from 88% to 94% for accuracy, and from 84% to 93% for consistency across all EOC exams. The better the test is targeted to the student's ability, the higher the classification index.

Table 32. Classification Accuracy and Consistency Indexes for Performance Levels

Test	Performance Level	Accuracy		Consistency	
		% Accuracy	Kappa	% Consistency	Kappa
Algebra 1	Approaches	88	0.73	85	0.67
	Meets	88	0.74	86	0.69
	Exceeds	93	0.77	92	0.73
Algebra 2	Approaches	88	0.74	84	0.66
	Meets	90	0.78	87	0.72
	Exceeds	94	0.79	93	0.74

8.4 REPORTING CATEGORY RELIABILITY

Table 33 shows the marginal reliability coefficients and the measurement errors computed for the reporting categories. Because the precision of scores in reporting categories is not sufficient to report scores, the scores in each reporting category are reported using one of the three performance categories: Meets or Exceeds, Near, or Does Not Meet. The classification rules are detailed in Section 9.6, Rules for Calculating Strengths and Weaknesses for Reporting Categories.

Table 33. Marginal Reliability Coefficients for Reporting Categories

Test	Reporting Categories	Number of Items Specified in Test Blueprint		Marginal Reliability	N	Mean	SD	SEM
		Min	Max					
Algebra 1	Algebraic Concepts and Procedures	21	23	0.85	5,643	291.11	55.01	21.25
	Modeling and Problem Solving	20	22	0.82	5,643	285.34	53.18	22.26
Algebra 2	Algebraic Concepts and Procedures	29	31	0.85	2,201	283.66	54.13	20.75
	Modeling and Problem Solving	14	16	0.77	2,201	282.96	62.11	30.10

9. SCORING

9.1 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

A student's score for an adaptive assessment depends on two factors: the number of items the student answers correctly and the difficulty of those items. In the adaptive assessment, each time a student answers an item, that item is scored, and the selection of subsequent items is based on how the student performed on earlier items. If a student answers items correctly, the adaptive system assigns the student items of higher difficulty. If a student answers items incorrectly, he or she will receive items of lower difficulty. Each time a student takes an assessment, the online TDS administers the test with items representing the breadth and depth identified in the test specifications and content standards, covers the full range of content and DOK included in the standards, and determines the extent to which the adaptive system adjusts the difficulty level of the items.

When a student is administered the first opportunity for a content-area assessment, the first few items match an average Hawai'i student score because no previous score exists. When a student uses the second opportunity for the same content-area assessment, the score from the first test is used by the adaptive system to assign the first few items at a difficulty level that is related to the student's previous score. As the student answers additional items, the adaptive system continues to assign higher- or lower-difficulty items based on whether the student is answering the items correctly or incorrectly. The system functions in the same way for the third testing opportunity.

Because the test adapts to each student's performance while maintaining accurate representation of the required grade-level knowledge and skills in content breadth and depth, the online results provide precise estimates of each student's true performance level across the range of proficiency.

Test items are selected from the pre-calibrated item bank using a Rasch model to best match the ability level of each student. Student ability estimates are obtained by indexing items by *i*. The likelihood function based on the *j*th person's score pattern is

$$L_j\left(\theta|z_j, b'_1, \ldots b'_{k_j}\right) = \prod_{i=1}^{k_j} p_i(z_{ji}|\theta, b_{i,1}, \ldots, b_{i,m_i}) \qquad (4)$$

where $b'_1 = (b_{i,1}, \ldots, b_{i,m_i})$ is the parameter vector of the *i*th item, m_i is the maximum possible score of the item, and the product is computed over only the k_j items presented to student *j*. Depending on the item type, the probability $p_i(z_{ji}|\theta, b_{i,1}, \ldots, b_{i,m_i})$ takes either the form based on the one-parameter Rasch model of the dichotomously scored items (in which case, we only have $b_{i,1}$, which can be simply written as b_i), or the form based on Masters' partial credit model for the polytomous items.

In case of dichotomously scored items, we have

$$p_i(z_{ji}|\theta, b_i) = \begin{Bmatrix} \dfrac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} = p_i, if\ z_{ji} = 1 \\ \dfrac{1}{1 + \exp(\theta - b_i)} = 1 - p_i, if\ z_{ji} = 0 \end{Bmatrix}$$

and in case of polytomous items,

$$p_i(z_{ji}|\theta, b_{i,1}, \dots, b_{i,m_i}) = \begin{Bmatrix} \frac{\exp\left(\sum_{r=1}^{z_{ji}}(\theta - b_{i,r})\right)}{s_i(\theta, b_{i,1}, \dots, b_{i,m_i})}, if\ z_{ji} > 0 \\ \frac{1}{s_i(\theta, b_{i,1}, \dots, b_{i,m_i})}, if\ z_{ji} = 0 \end{Bmatrix}$$

where $s_i\left(\theta, b_{i,1}, \dots, b_{i,m_i}\right) = 1 + \sum_{l=1}^{m_i} \exp\left(\sum_{r=1}^{l}(\theta - b_{i,r}\right)$.

The log likelihood function is

$$l_i\left(\theta|z_j, b_{i,1}, \dots, b_{i,m_i}\right) = \log(L_j(\theta|z_j, b_{i,1}, \dots, b_{i,m_i})) = \sum_{i=1}^{k} \log\left(p_i(\, z_{ij}|\theta, b_{i,1}, \dots, b_{i,m_i})\right). \quad (5)$$

The ability θ is estimated by maximizing the log likelihood function defined in Equation (5), and the SEM is approximated by the square root of the inverse of the Fisher information evaluated at the maximum likelihood estimate (MLE) of θ.

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

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

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

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

where m_i is the maximum possible score point (starting from 0) for the *i*th item. The $SE(\theta_j)$ is calculated based on only the answered item(s).

The algorithm allows previously answered items to be changed; however, it does not allow items to be skipped. Item selection requires iteratively updating the estimate of the overall and strand ability estimates after each item is answered. When a previously answered item is changed, the proficiency estimate is adjusted to account for the changed responses when the next new item is selected. When the update of the ability estimates is performed at each iteration, the overall and strand scores are recalculated using all data at the end of the test for the final score.

9.2 Rules for Transforming Theta to Scale Scores

The student's performance in each content area test is summarized in an overall test score referred to as a *scale score*. The number of items a student answers correctly and the difficulty of the items presented are used to estimate students' abilities (i.e., theta scores) and then statistically transform the theta scores to scale scores so that scores from different sets of items can be meaningfully compared. The scale scores represent a linear transformation of the ability estimates (theta scores) using the formula, $SS = a * \hat{\theta} + b$. The scaling constants *a* and *b* are determined by the *Meets Proficiency* standard set at a scale score of 300 and the scale score standard deviation at 40, using the formula, $SS = 300 + 40(\hat{\theta} - \theta_c)/\sigma_{\hat{\theta}})$, where the theta ($\hat{\theta}$) represents any level of student ability on the operational pool. The theta cut score (θ_c) represents the theta that the panelists determined for the *Meets Proficiency* standard cut score from the ordered-item

booklet. The standard deviation of theta ($\sigma_{\hat{\theta}}$) represents the standard deviation of all the thetas, or logit values. Table 34 provides the parameters used for the linear transformation. The scale scores are truncated so that the lowest possible scale score is 100 and the highest possible scale score is 500.

Table 34. Intercept and Slope for the Theta-to-Scale Score Linear Transformation

Test	SD (Observed Theta)	Meets Cut	Slope (a)	Intercept (b)
Algebra 1	0.89032	-0.18380	44.92764	308.25769
Algebra 2	0.78002	-0.33054	51.28057	316.95053

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

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

where SE_{ss} is the standard error of the ability estimate on the reporting scale, SE_{θ} is the standard error of the ability estimate on the Θ scale, and a is the slope of the scaling constant that transforms Θ to the reporting scale.

The scale scores are mapped into four performance levels using three performance standards (i.e., cut scores). Table 35 provides three performance standards for each grade and content area.

Table 35. Performance Standards for Algebra 1 and Algebra 2

Test	Performance Standards		
	Approaches	Meets	Exceeds
Algebra 1	263	300	328
Algebra 2	263	300	337

9.3 Lowest/Highest Obtainable Scores

Although student ability is estimated more precisely in an adaptive test than in a fixed-form test, especially for high- and low-performing students, if the item pool does not include enough easy or difficult items to measure low- and high-performing students, the standard error could be large in the lower and higher ends of the ability range. It was decided to truncate extreme unreliable student ability estimates, 100 and 500 for the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) in a scale score metric in all grades and content areas.

9.4 Scoring All Correct and All Incorrect Cases

In IRT, maximum-likelihood ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. In such cases, MLEs are generated by adding ±0.5 to the zero or perfect raw scores, respectively, and maximizing conditional on the adjusted raw score.

9.5 Attemptedness Rule

A test is scored and reported if five or more operational items are attempted. In each opportunity, students are instructed to respond to all items and submit the test by clicking the "submit" button. An incomplete opportunity is an opportunity that expired because the student did not submit the test. The student might

have responded to all items, but if the test was not submitted, the opportunity is incomplete. The rules for scoring the incomplete tests are as follows:

- An incomplete opportunity with five or more attempted operational items receives an overall score but NOT subscores (strand score or subscore for each reporting category).
- The overall score for an incomplete opportunity is the student's theta based on the five or more attempted operational items minus one SEM.

9.6 RULES FOR CALCULATING STRENGTHS AND WEAKNESSES FOR REPORTING CATEGORIES

In addition to the overall scale score, relative strength and weakness at the reporting category level is produced in three proficiency classifications. The ability estimates for the reporting categories are on the same scale as the total score; hence, the same cut score of the *Meets Proficiency* standard is used to judge student performance on each reporting category. For each reporting category, a 68% confidence interval of the reporting category ability score (θ), $\theta \pm 1$ SE(θ) is computed. The ability scores are categorized into three classifications referenced to the *Meets Proficiency* standard cut score (θ_c) as follows:

- *Meets or Exceeds Proficiency* (code = 3): $if\,(\theta - SE(\theta)) \geq \theta_c$
- *Near Proficiency* (code = 2): $if\big(\theta - SE(\theta)\big) < \theta_c \leq (\theta + SE(\theta))$
- *Does Not Meet Proficiency* (code = 1): $if\big(\theta + SE(\theta)\big) < \theta_c$

9.7 BENCHMARK SCORES

The benchmark-level reports are impossible to produce for a fixed-form test because the number of items included per benchmark is too few to produce a reliable score at the benchmark level. A typical fixed-form test includes only one or two items per benchmark. Even when aggregated, these data only narrowly reflect the benchmark because they reflect only one or two ways of measuring the benchmark. An adaptive test, however, offers a tremendous opportunity for benchmark-level data at the class, school, and complex area level. With an adequate item pool, a class of 20 students might respond to 10 or 15 different items measuring any given benchmark. A benchmark score is an aggregate of the differences in student overall proficiency and the differences in the difficulty of the items measuring a benchmark in a class, school, or complex area. Benchmark scores are computed for attempted tests. Benchmark scores are computed within each reporting category.

Benchmark scores are computed as follows:

Defining $p_{ij} = p(z_{ij} = 1)$ represents the probability that student j responds correctly to item i (z_{ij} represents the *j*th student's score on the *i*th item). For items with one score point, the Rasch model was used to calculate the expected score on item *i* for student *j* with estimated ability θ as

$$E\big(z_{ij}\big) = \frac{\exp(\hat{\theta}_j - b_{i,1})}{1+\exp(\hat{\theta}_j - b_{i,1})}.$$

For items with two or more score points, using the partial credit model, the expected score for student *j* with estimated ability $\hat{\theta}_j$ on an item *i* with a maximum possible score of m_i is calculated as

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{l\exp(\sum_{k=1}^{l}(\hat{\theta}_j - b_{i,k}))}{1+\sum_{l=1}^{m_i}\exp(\sum_{k=1}^{l}(\hat{\theta}_j - b_{i,k}))}.$$

For each item *i*, the residual between observed and expected score for each student is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

Residuals are summed for items within a benchmark. The sum of residuals is divided by the total number of points possible for items within the benchmark, *B*

$$\delta_{jB} = \frac{\sum_{i\in T}\delta_{ji}}{\sum_{i\in T}K_i}.$$

For an aggregate unit, a benchmark score is computed by averaging individual student scores for the benchmark across students of different abilities that received different items measuring the same benchmark at different levels of difficulty,

$$\bar{\delta}_{Bg} = \frac{1}{n_g}\sum_{j\in g}\delta_{jT}, \text{ and } se(\bar{\delta}_{Bg}) = \sqrt{\frac{1}{n_g(n_g-1)}\sum_{j\in g}(\delta_{jB} - \bar{\delta}_{Bg})^2},$$

where n_g is the number of students who responded to any of the items that belong to the benchmark *T* for an aggregate unit *g*. If a student did not happen to see any items on a particular benchmark, the student is not included in the n_g count for the aggregate.

A statistically significant difference from zero in these aggregates is evidence that a roster, teacher, school, or complex area is more effective (if $\bar{\delta}_{Tg}$ is positive) or less effective (negative $\bar{\delta}_{Tg}$) in teaching a given benchmark.

In the aggregate, a benchmark performance is reported as a group of students that performs better, worse, or as expected on this benchmark. In some cases, insufficient information will be available as well, and will be indicated where applicable.

For benchmark-level strengths/weakness, report the following:

- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is better than on the rest of the test.
- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is worse than on the rest of the test.
- Otherwise, performance is similar to performance on the test as a whole.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

10. REPORTING AND INTERPRETING SCORES

The Centralized Reporting System (CRS) generates a set of online score reports that includes information describing student performance for students, parents, educators, and other stakeholders. The online score reports are produced immediately after students complete the tests. Because the performance score report is updated each time a student completes a test, authorized users (e.g., school principals, teachers) can have quickly available information on students' performance scores and use them to improve student learning. In addition to individual student score reports, the CRS produces aggregate score reports for teachers, schools, complex areas, and states. The timely accessibility of aggregate score reports help users monitor student performance in each subject, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year. Additionally, the CRS provides participation data that help monitor student participation rate.

This section describes the types of scores reported in the CRS, as well as the ways to interpret and use these scores in detail.

10.1 CENTRALIZED REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

10.1.1 Types of Score Reports

The CRS is designed to help educators, students, and parents answer questions regarding how well students have performed in Algebra 1 and Algebra 2. The CRS is the online tool that provides educators and other stakeholders with timely, relevant score reports and guide stakeholders to make valid, actionable interpretations of student assessment results. The CRS is designed with stakeholders (such as teachers, parents, and students) who are not technical measurement experts in mind. It ensures that test results are presented in a way that is easy to read and understand by using simple language so that users can quickly understand assessment results and make valid inferences about student performance.

The CRS is also designed to present student performance in a uniform format. For example, throughout the design, similar colors are used for groups of similar elements, such as performance levels. This strategy allows readers to easily compare similar elements and to avoid comparing dissimilar elements.

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. Once the user clicks on the test family that he or she wants to explore further, it will take the user to the detailed dashboard, where the results are shown. Additionally, when authorized state-level users login 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 36 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 36. 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 students Proficient (for overall students and by subgroup) • Average scale score and standard error of average scale score on the overall test and claim (for overall students and by subgroup) • Percentage of students at each performance level on the overall test (for overall students and by subgroup) • Performance category in each target (for overall students)[1]
Student	• Total scale score and standard error of measurement • Performance level for overall score with Performance-Level Descriptors (PLDs) • Average scale scores and standard errors of average scale scores for individual complex, complex areas, and states

Notes:
[1] Performance category in each target is provided for all aggregate levels except for state.

Aggregate score reports at a selected aggregate level are provided for overall students and by subgroup. Users can see student assessment results by any of the subgroups. Table 37 presents the types of subgroups and subgroup categories provided in the CRS.

Table 37. Types of Subgroups

Subgroup	Subgroup Category
Gender	Male Female
ELL	ELL Not ELL
Disability	01 - Autism 02 - Deaf-Blindness 03 - Deafness 04 - Developmental Delay (Age 3–5) 05 - Developmental Delay (Age 6–8) 06 - Emotional Disturbance 07 - Hearing Impaired 08 - Mental Retardation 09 - Multiple Disability 10 - Orthopedic Impairment 11 - Other Health Impairment 12 - Specific Learning Disability 13 - Speech/Language Impairment 14 - Traumatic Brain Injury 15 - Visual Impairment including Blindness 16 - Autism Spectrum Disorder 17 - Other Health Disability 18 - Speech or Language Disability 19 - Intellectual Disability 20 - Visual Disability Incl Blindness 21 - Hard of Hearing 22 - Orthopedic Disability Missing

Subgroup	Subgroup Category
Migrant Status	Migrant Not Migrant
Disadvantaged	C D F R Missing 1 2 3 E
Ethnicity	American Indian/Alaskan Native Asian/Pacific Islander African American Hispanic Hawai‘i Pacific Islander White Multi-Racial
Enrolled Grade	Grade 3 Grade 4 Grade 5 Grade 6 Grade 7 Grade 8 Grade 9 Grade 10 Grade 11 Grade 12 Grade 31 Grade 00 Grade 32 Grade 33 Grade 34

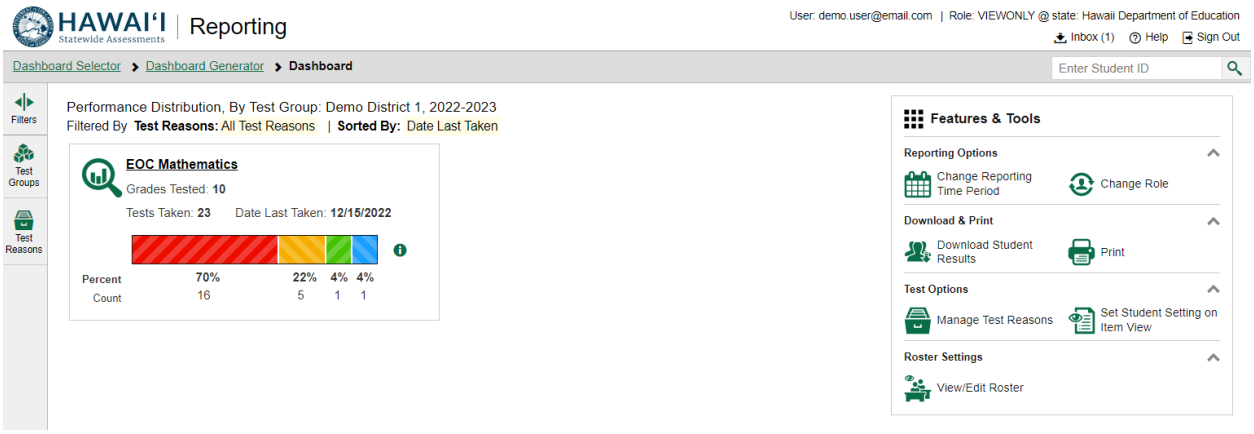
10.1.2 Centralized Reporting System

10.1.2.1. Dashboard

The first page users see when they log onto the CRS contains summaries of student performance by test family. Complex personnel see complex summaries, school personnel see school summaries, and teachers see summaries of their students. State personnel and complex area personnel would need to select the specific complex in order to view the aggregate results.

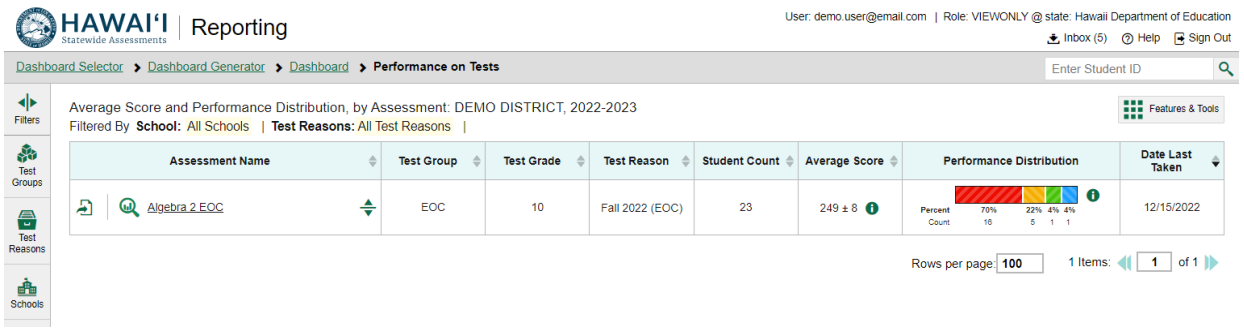
The dashboard summarizes students' performance by test family, including the number of students tested, the grades of the students who have tested, and the percentage and counts of students at each performance level. Exhibit 1 presents a sampled dashboard page at the district level.

Exhibit 1. Dashboard



Educators can select the subject group to view individual test results for the selected test group. Once the user selects 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 the number of students tested, average score and standard error of the means, and the percentage and counts of students at each performance level. Exhibit 2 presents a sampled detailed dashboard page for Algebra 2 EOC at the complex level.

Exhibit 2. Detailed Dashboard: Complex Level

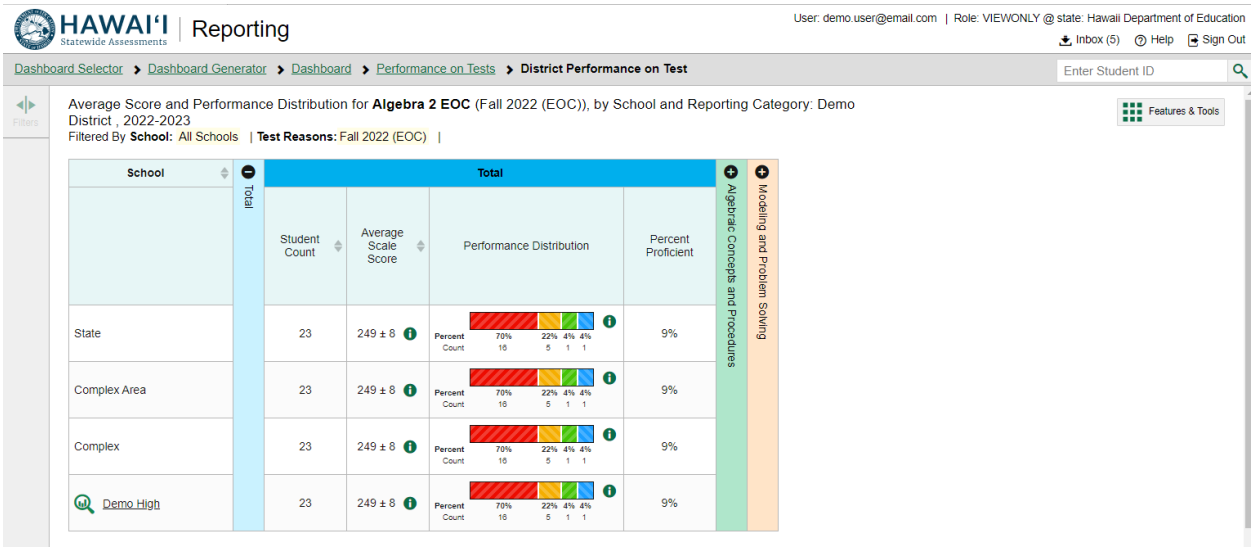


10.1.2.2. Subject Summary Results

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 the number of students tested, the average scale score and standard error associated with the average scale score, the percentage of proficient students, and 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 a subject summary results for Algebra 2 EOC at the complex level.

Exhibit 3. Subject Summary Results for Algebra 2 EOC: Complex Level

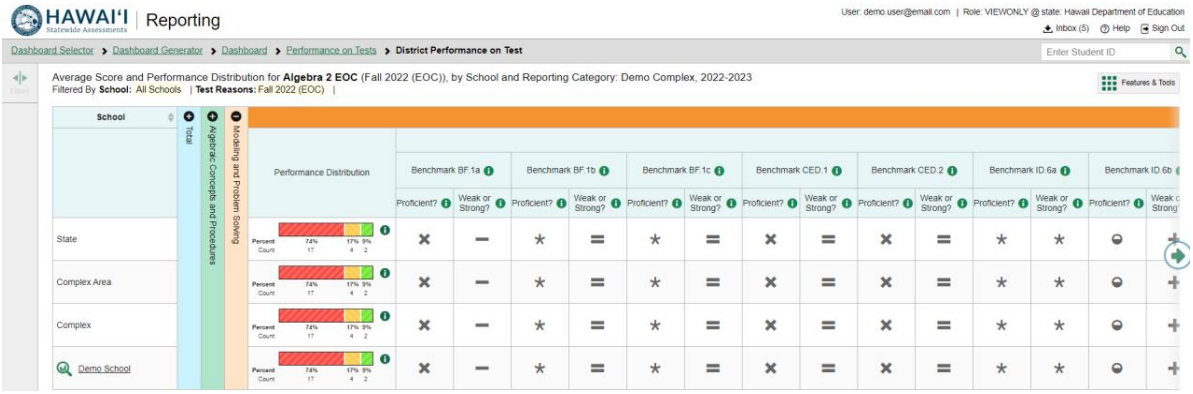


10.1.2.3. Performance Distribution Results

Aggregated performance distribution results are also available on the same report page as the subject level results. The performance distribution provides aggregate summaries on student count and percentage of students in each performance level for a particular grade and subject.

Like the subject level results, the performance distribution presents the summary results for the selected aggregate unit and the summary results for the state and aggregate unit above the selected aggregate. Also, the performance level results can be presented for overall students and by subgroup. Exhibit 4 presents an example of performance distribution results for Algebra 2 EOC at a complex level.

Exhibit 4. Performance Distribution Results for Algebra 2 EOC: Complex Level



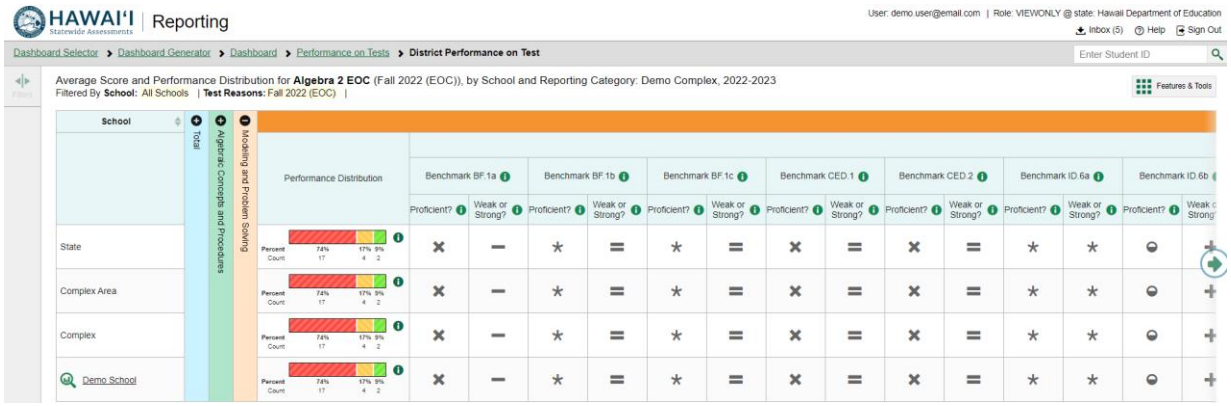
10.1.2.4. Benchmark-Level Results

The benchmark-level results provide the aggregate summaries on student performance in each benchmark. The benchmark-level results provide the strength or weakness indicators in each benchmark that are computed in two ways (i.e., performance relative to proficiency, performance relative to the test as a whole). In the benchmark level, strengths and weaknesses are reported for groups of students based on whether

there is a statistically significant difference between that group's performance on each benchmark and the group's performance on the rest of the test. A benchmark-level result also includes group performance relative to the expected performance of a student at the proficient cut score.

Exhibit 5 presents an example of benchmark-level results for Algebra 2 EOC at the complex level.

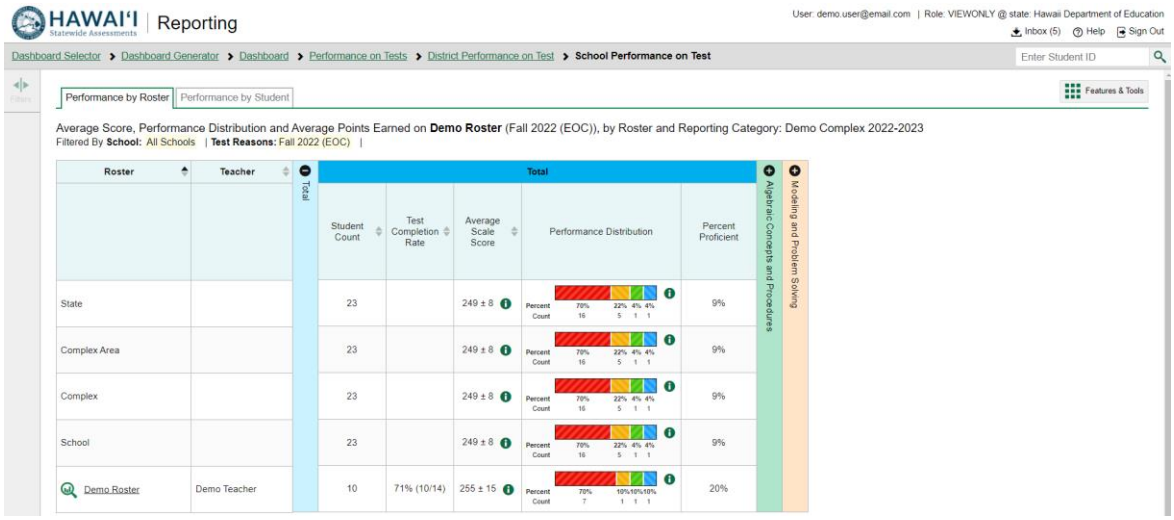
Exhibit 5. Benchmark-Level Results for Algebra 2 EOC: Complex Level



10.1.2.5. Roster Performance Report

Class, teacher, and school performance rosters provide users with performance data for a group of students belonging to a system-defined or user-defined class. The report includes the student's overall subject scale scores with SEM and the performance level. Exhibit 6 shows a sample roster performance report for the EOC assessment.

Exhibit 6. Roster Performance Report for Algebra 2 EOC



10.1.2.6. Individual Student Report

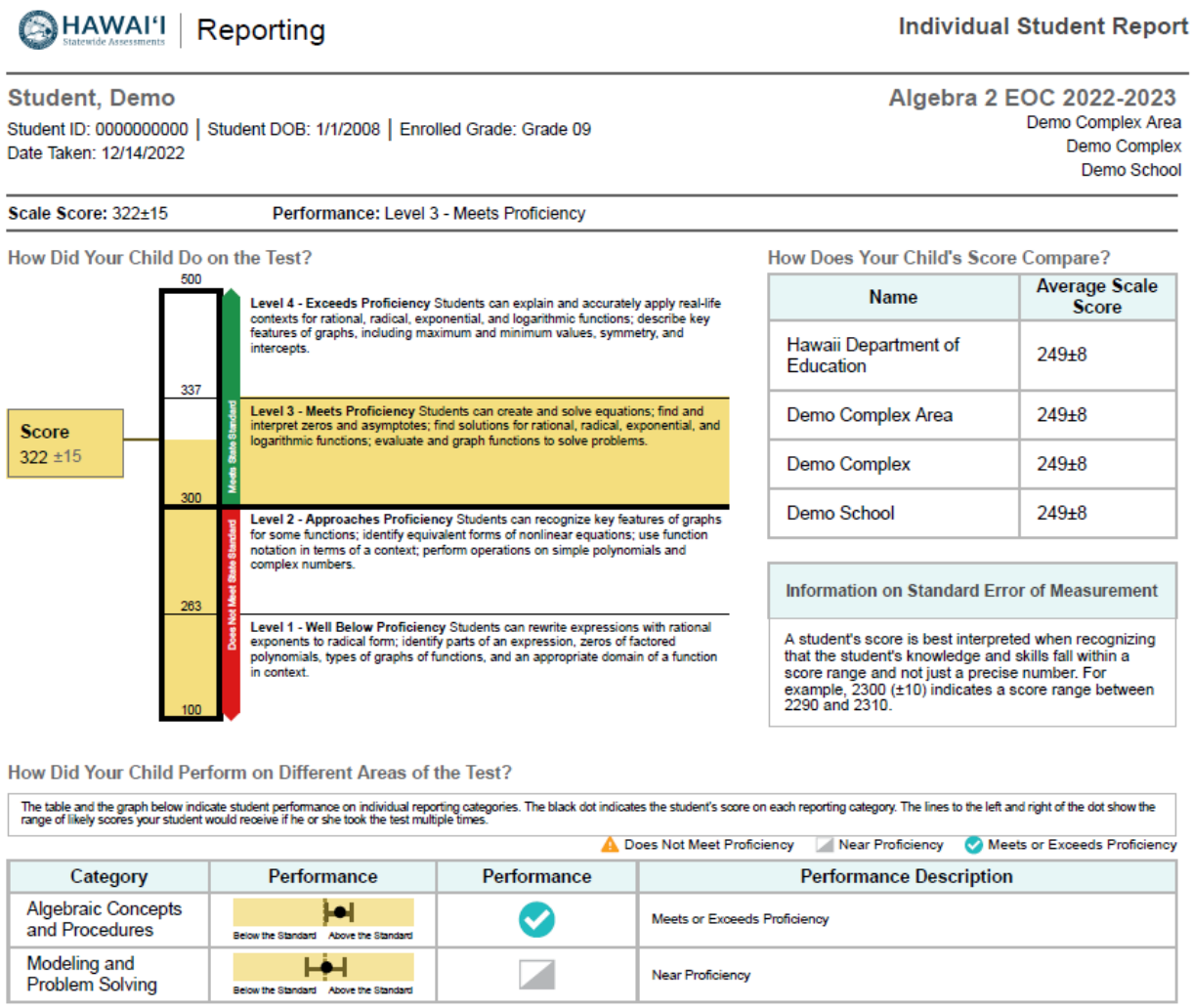
The student's name, scale score with the SEM, and performance level are shown at the top of the page. In the middle section, the student's performance is described in detail using a barrel chart. In the barrel chart, the student's scale score is presented with the SEM using a "±" sign. SEM represents the precision of the

scale score, or the range in which the student would likely score if a similar test were administered multiple times. Furthermore, in the barrel chart, Performance-Level Descriptors (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.

Next to the barrel chart, average scale scores and standard errors of the average scale scores for state, complex area, complex, and school are displayed so that student performance can be compared with the 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.

Exhibit 7 presents example of individual student report for Algebra 2 EOC.

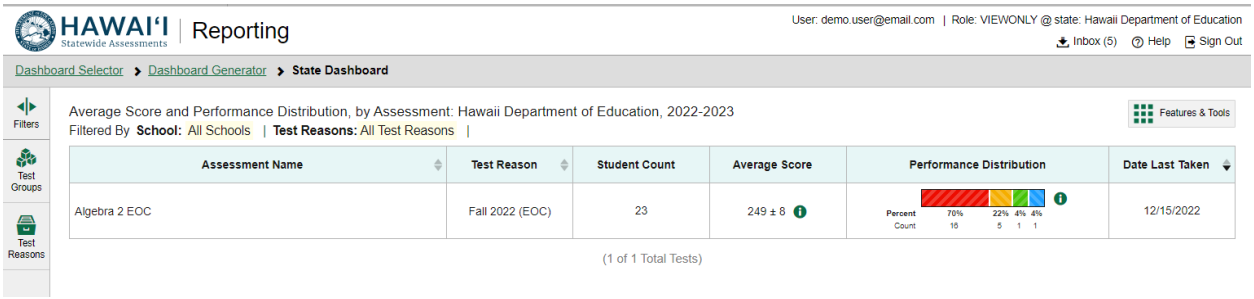
Exhibit 7. Individual Student Report for Algebra 2 EOC.



10.1.2.7. State-Level Summary

The CRS provides a state dashboard for authorized state-level users to track student performance for a test across the entire state. Users can specify the test and administration year to display in the report. Exhibit 8 presents a sample state dashboard page.

Exhibit 8. State Dashboard



10.2 Interpretation of Reported Scores

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

10.2.1 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 student knowledge and skills measured. The scale score is the transformed score from a theta score, which is estimated based on mathematical models. Low scale scores can be interpreted to mean that the student does not possess sufficient knowledge and skills measured by the test. Conversely, high scale scores can be interpreted to mean that the student has proficient knowledge and skills measured by the test. Scale scores can be used to measure student growth across school years. The interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and PLDs.

10.2.2 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 several times, the resulting scale score would vary across administrations, sometimes being 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 was administered several 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, 340 ± 10 indicates that if a student was tested again, it is likely that the student would receive a score between 330 and 350. SEM can differ for the same scale score, depending on how closely the administered items match the student's ability.

10.2.3 Performance Level

Performance levels are proficiency categories on a test that students fall into based on their scale scores. 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). PLDs describe 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 PLDs. For the performance level of *Approaches Proficiency* in Algebra 1, for instance, PLDs are described as follows: "Students can factor simple quadratic expressions; transform a basic quadratic equation to an equivalent form; graph systems of linear equations; identify either the slope or the *y*-intercept of a linear function for a scatter plot."

10.2.4 Performance Levels for Reporting Categories

Student performance in each reporting category is reported at three performance levels: (1) *Does Not Meet Proficiency*, (2) *Near Proficiency*, and (3) *Meets or Exceeds Proficiency*. Unlike the performance level for overall test, student performance on each of reporting categories is evaluated with respect to the *Meets Proficiency* standard. Performance at either *Does Not Meet Proficiency* or *Meets or Exceeds Proficiency* can be interpreted to mean that student performance is clearly above or below the *Meets Proficiency* cut score for a specific reporting category. Students performing at *Near Proficiency* can be interpreted as meaning that students' performance does not provide enough information to tell whether students reached the *Meets Proficiency* mark for the specific reporting category.

10.2.5 Benchmark-Level Report

In addition to the reporting category-level reports, teachers and educators ask for additional reports on student performance for instructional needs. Benchmark-level reports are produced for the aggregate units only, not for individual students, because each student is administered too few items in a benchmark to produce a reliable score for each benchmark.

CAI reports relative strength and weakness scores for each benchmark within a reporting category. The strengths and weaknesses report are generated for aggregate units of classroom, school, and complex area, and provides information about how a group of students in a class, school, or complex area performed on each benchmark, relative to their performance on the test as a whole. For each benchmark, the observed performance on items is compared with expected performance, based on the overall ability estimate. At the aggregate level, when observed performance within a benchmark is greater than expected performance, the reporting unit (e.g., class, school, complex area) shows a relative strength in that benchmark. Conversely, when observed performance within a benchmark is below the level expected based on overall performance, the reporting unit shows a relative weakness in that benchmark.

The benchmark performance shows how a group of students performed on each benchmark, relative to their overall subject performance on a test. The performance on benchmark is mapped into three performance levels: (1) *Performance is better than on the rest of the test as a whole*, (2) *Performance similar to the test as a whole*, and (3) *Performance is worse than on the rest of the test as a whole*. The *Performance is worse than on the rest of the test as a whole* does not imply a lack of performance. Instead, it can be interpreted to mean that student performance on that benchmark was below their performance across all other benchmarks combined. Although performance level for benchmarks provides some evidence to help address student strengths and weaknesses, they should not be over-interpreted because student performance on each benchmark is based on relatively few items, especially for a small group.

10.2.6 Aggregated Score

Student scale scores are aggregated at the roster, teacher, school, complex, complex-area, and state levels to represent how a group of students performs on a test. When student scale scores are aggregated, the 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, therefore, are subject to measures of uncertainty. In addition to the aggregated scale scores, the percentages of students in each performance level for overall and by reporting category are reported at the aggregate level to represent how well a group of students performs overall and by reporting category.

10.3 Appropriate Uses for Scores and Reports

Assessment results on student test performance can be used to help teachers or schools make decisions on how to support students' learning. Aggregate score reports for the teacher and school levels provide information on the strengths and weaknesses of their students and can be utilized to improve teaching and student learning. For example, a group of students can perform very well overall, but it is possible that they will not perform as well on several benchmarks compared to their overall performance. In this case, teachers or schools can identify strengths and weaknesses of their students through the group performance by reporting category and can benchmark and promote instruction on a specific reporting category or benchmark area at which student performance is below overall performance. Furthermore, by narrowing the student performance result by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from a disadvantaged subgroup. For example, teachers can see student assessment results by EL status and observe that EL students are struggling with the *Algebraic Concepts and Procedures* reporting category in Algebra 1. Teachers can then provide additional instructions for these students to enhance their performance in the reporting category for *Algebraic Concepts and Procedures*.

Additionally, assessment results can be used to compare students' performance among different students and groups. Teachers can evaluate how their students perform compared with other students in schools, complexes, and complex areas both overall and by reporting category. Although all students are administered different sets of items in each CAT, scale scores are comparable across students.

Although assessment results provide valuable information to understand student performance, scores and reports should be interpreted with caution. It is important to note that reported scale scores are estimates of true scores and, therefore, do not represent the precise measure of student performance. A student's scale score is associated with measurement error, so users need to consider measurement error when using student scores to make decisions about student performance. Moreover, though student scores may be used to help make important decisions about student 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 performance, such as classroom assessment and teacher evaluation, should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to consider group size. The smaller the group, the larger the measurement error related to these aggregate data, thus requiring a more cautious interpretation.

11. QUALITY CONTROL PROCEDURES

Quality assurance (QA) procedures are enforced throughout all stages of HSA test development, administration, and scoring and reporting. CAI implements a series of quality-control steps to ensure the error-free production of score reports in both the online and paper-pencil formats. The quality of the information produced in the Test Delivery System (TDS) is tested thoroughly before, during, and after the testing window.

11.1 ADAPTIVE TEST CONFIGURATION

For CATs, a test configuration file 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 item information (i.e., cut scores, answer keys, item attributes, item parameters, passage information). The accuracy of the information in the configuration file is checked and confirmed numerous times independently by multiple staff members prior to the testing window.

To verify the accuracy of the scoring engine, CAI uses simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the Hawai‘i student population for EOC exams. The ability of each simulated student is used to generate a sequence of item response scores consistent with the underlying ability distribution. These simulations provide a rigorous test of the adaptive algorithm for adaptively administered tests. They also provide a check of form distributions (if administering multiple test forms) and test scores in fixed-form tests.

Simulations are generated using the production-item selection and scoring engine to ensure that the verification of the scoring engine is based on a very 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 EOC exams. The purpose of the simulations is to configure the adaptive algorithm to optimize item selection to meet blueprint specifications while targeting test information to student ability as well as checking the score accuracy. The simulated data are used to check whether the scoring specifications were applied accurately. The scores in the simulated data file are checked independently following the scoring rules specified in the scoring specifications.

11.1.1 Platform Review

CAI’s TDS supports a variety of item layouts. Each item goes through an extensive platform review on different operating systems, like Windows, Linux, and iOS. to ensure that the item’s appearance is consistent in all layouts. Some of the layouts have the stimulus and item-response options/response area displayed side by side. In each of these layouts, both the stimulus and response options have independent scroll bars.

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

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

11.1.2 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). The UAT of the TDS serves both as a software evaluation and content approval role. The UAT period gives the state an opportunity to interact with the exact test with which the students will interact.

11.2 QUALITY ASSURANCE IN DATA PREPARATION

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

Data pass directly from the Quality Monitor (QM) System to the DOR, which serves as the repository for all test score information, and from which all test information for reporting is pulled. The Data Extract Generator (DEG) is the tool used to pull data from the DOR for delivery to HIDOE. CAI staff ensure that extract file data matches the DOR prior to delivery to HIDOE.

11.3 QUALITY ASSURANCE REPORTS

To monitor the performance of the online TDS during the testing window, CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and CAI contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first sign that trouble may be ahead. Applications log not only errors and exceptions, but latency (timing) information for critical database calls. This information enables us to know instantly whether the system is performing as designed, or if it is starting to slow down or experience a problem. Additionally, latency data (i.e., 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 complex areas experiencing unusual slowdowns, often before the schools or complex areas notice.

A series of QA reports can also be generated at any time during the online assessment window, such as blueprint match rate, item exposure rate, and item statistics, for early detection of any unexpected issues. Any deviations from the expected outcome are flagged, investigated, and resolved. In addition to these statistics, a cheating analysis report is produced to flag any unlikely patterns of behavior in a testing session, as discussed in Section 3.8, Prevention and Recovery of Disruptions in the Test Delivery System.

For example, the item statistics analysis report allows psychometricians to ensure that items are performing as intended and serves as an empirical key check throughout the operational testing window. 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 a specified range are flagged for reporting or to generate reports based on all items in the pool.

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

Table 38 presents an overview of the QA reports.

Table 38. 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 and scoring errors for constructed-response, performance, or technology-enhanced items)
Blueprint Match Rates	To monitor unexpected 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 (high unused items/passages)	Early detection of any oversight in the blueprint specification
Cheating Analysis	To monitor testing irregularities	Early detection of testing irregularities

11.4 SCORE REPORT QUALITY CHECK

For the 2022–2023 EOC exams, two types of score reports were produced: (1) online reports and (2) printed reports (family reports only).

11.4.1 Online Report Quality Assurance

Scores for online assessments are assigned by automated systems in real time. For machine-scored portions of assessments, the machine rubrics are created and reviewed, along with the items, then validated and finalized during rubric validation following field testing. The review process "locks down" the item and rubric when the item is approved for web display (Web Approval). 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, and other scoring problems. Potential issues are automatically flagged in reports available to our psychometricians.

The human-scoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are married with the machine-scored items by CAI's Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are checked by our QA system. The integrated scores are sent to our test-scoring system, a mature, well-tested, real-time system that applies client-specific scoring rules and assigns scores from the calibrated items, including calculated performance-level indicators, subscale scores, and other features, which then pass automatically to the reporting system and DOR. The scoring system is tested extensively, including physical checks of scored tests and large-scale simulations, prior to deployment to ensure that point estimates and standard errors are correct.

Every test undergoes a series of validation checks. Once the QA system signs off, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the "official" record is stored. Only after scores have passed the QA checks and are uploaded to the DOR are they passed to the Centralized Reporting System (CRS), which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the CRS until it passes all of the QA system's validation checks. All of these processes take milliseconds to complete so that within less than a second of handscores being received by CAI and passing QA validation checks, the composite score is available in the CRS.

11.4.2 Paper Report Quality Assurance

Statistical Programming

The family reports contain custom programming and require rigorous QA processes to ensure their accuracy. All custom programming is guided by detailed and precise specifications in our reporting specifications document. Upon the approval of the specifications, analytic rules are programmed, and each program is tested extensively on test decks and real data from other programs. The final programs are reviewed by two senior statisticians and one senior programmer to ensure that they implement agreed-upon procedures. Custom programming is implemented independently by two statistical programming teams working from the specifications. Only when the output from both teams matches exactly are the scripts released for production. Quality control, however, does not stop there.

Much of the statistical processing is repeated. CAI has implemented a structured software development process to ensure that the repeated tasks are implemented correctly and identically each time. CAI writes small programs (called macros) that take specified data as input and produce data sets containing derived variables as output. Approximately 30 such macros reside in our library for the grades 3–8 and 11 program score reports. Each macro is extensively tested and stored in a central development server. Once a macro is tested and stored, changes to the macro must be approved by the director of score reporting and the director of psychometrics, as well as by the project directors for affected projects.

Each change is followed by a complete retesting with the entire collection of scenarios on which the macro was originally tested. The main statistical program is made up mostly of calls to various macros, including macros that read in and verify the data and conversion tables and macros that do the many complex calculations. This program is developed and tested using artificial data generated to test both typical and extreme cases. Additionally, the program goes through a rigorous code review by a senior statistician.

Display Programming

The paper report development process uses graphical programming, which takes place in a Xerox-developed programming language called Variable Data Intelligent PostScript Printware (VIPP) and allows virtually infinite control of the visual appearance of the reports. After CAI designers create backgrounds, our VIPP programmers write code that indicates where to place all variable information (e.g., data, graphics, text) on the reports. The VIPP code is tested using both artificial and real data. CAI's data generation utilities can read the output layout specifications and generate artificial data for direct input into the VIPP programs. This allows program testing to begin before statistical programming is complete. In later stages, artificial data are generated according to the input layout and run through the psychometric process and the score reporting statistical programs, and the output is formatted as VIPP input. This enables us to test the entire system. Programmed output goes through multiple stages of review and revision by graphics editors and the CAI Score Reporting Team to ensure that design elements are accurately

reproduced and data are correctly displayed. Once we receive final data and VIPP programs, the team reviews proof that contains actual data based on our standard QA documentation. Additionally, we compare data independently calculated by CAI psychometricians with data on the reports. A large sample of reports is reviewed by several CAI staff members to ensure that all data are correctly placed on reports.

This rigorous review is typically conducted over several days and takes place in a secure location in the CAI building. All reports containing actual data are stored in a locked storage area. Prior to printing the reports, CAI provides a live data file and individual student reports (ISRs) with sample complex areas for HIDOE staff review. CAI works closely with HIDOE to resolve questions and correct any problems. The reports are not delivered unless HIDOE approves the sample reports and data file.

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.

Billingsley, P. (1995). *Probability and measure*. New York: John Wiley & Sons, p. 456.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

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. (2011). WINSTEPS Rasch-Model computer program. Chicago: MESA Press.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 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*(4), 247–260.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342.

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some applications of item response theory to testing. *The Philippine Statistician, 52*(1–4), 81–92.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*(4), 265–276.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large-scale assessments. *NCEO Synthesis Report* (44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.unm.edu/NCEO/OnlinePubs/Synthesis44.html.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. Unpublished.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. (1979). *Best test design. Rasch measurement.* Chicago: MESA Press.

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

APPENDICES

`Appendix A: Language Accessibility, Bias, and Sensitivity Guidelines

1. STEREOTYPING

Testing materials should not present people stereotyped according to the following characteristics:

- Age
- Disability
- Gender
- Race/Ethnicity
- Sexual Orientation

2. SENSITIVE OR CONTROVERSIAL SUBJECTS

Controversial or potentially distressing subjects should be avoided or treated sensitively. For example, a passage discussing the historical importance of a battle is acceptable, whereas a graphic description of a battle would not be. Controversial subjects include the following:

- Death and Disease
- Gambling*
- Politics (Current)
- Race Relations
- Religion
- Sexuality
- Superstition
- War

*References to gambling should be avoided in Mathematics items related to probability.

3. ADVICE

Testing materials should not advocate specific lifestyles or behaviors except in the most general or universally agreed-upon ways. For example, a recipe for a healthful fruit snack is acceptable, but a passage recommending a specific diet is not. The following are categories of advice to be avoided completely:

- Religion
- Sexual Preference

4. DANGEROUS ACTIVITIES

Care should be taken not to present dangerous activities in such a way as to make them seem appealing or acceptable.

5. POPULATION DIVERSITY, REPRESENTATIVENESS, AND ETHNOCENTRISM

Testing materials should

- reflect the diversity of the testing population;
- use stimulus materials (such as works of literature) produced by members of minority communities;
- use personal names from different ethnic origin communities;
- use pictures of people from different ethnic origin communities; and
- avoid ethnocentrism (the attitude that all people should share a particular group's language, beliefs, culture, or religion).

6. DIFFERENTIAL FAMILIARITY: ELITISM AND DIF

Specialized concepts and terminology extraneous to the core content of test questions should be avoided. This caveat applies to terminology from the following fields:

- Construction
- Finance
- Sports
- Law
- Machinery
- Military Topics
- Politics
- Science
- Technology
- Agriculture

7. LANGUAGE ACCESSIBILITY

Language should be as direct, clear, and inclusive as possible. The following should be avoided or used with care:

- Passive Constructions
- Idioms
- Multiple Subordinate Clauses
- Pronouns with Unclear Antecedents
- Multiple-Meaning Words
- Nonstandard Grammar
- Dialect
- Jargon

8. GRAPHICS

All of the relevant foregoing standards apply to graphics.

Appendix B: Field-Test Items: Classical Item Statistics

Table B-1. Field-Test Items: Classical Item Statistics for EOC Algebra 1

Item Number	Benchmark	Type	Points	Average Score	Adjusted Polyserial / Biserial	Proportion by Score Point in CR and Response Option in MC					Distractor Adjusted Biserial for MC/ Mean Score for CR				
						0	A/1	B/2	C/3	D/4	0	A/1	B/2	C/3	D/4
23625	A-CED.1	EQ	1	0.05	0.75	0.95	0.05				280.24	384.84			
23626	A-CED.4	EQ	1	0.19	0.73	0.81	0.19				274.28	350.67			
23627	A-REI.12	EQ	1	0.14	0.63	0.86	0.14				280.05	349.00			
23628	A-REI.4	EQ	1	0.03	0.67	0.97	0.03				285.63	401.39			
23633	A-SSE.3	EQ	1	0.14	0.64	0.86	0.14				278.07	347.03			
23635	F-IF.6	EQ	1	0.04	0.27	0.96	0.04				286.22	324.44			
23636	F-IF.7	EQ	1	0.08	0.70	0.92	0.08				281.93	368.40			
23638	F-IF.9	EQ	1	0.14	0.77	0.86	0.14				277.76	364.44			
23639	F-IF.9	EQ	1	0.02	0.79	0.98	0.02				285.63	414.14			
23641	A-REI.1	EQ	2	0.55	0.75	0.32	0.26	0.42			251.07	278.28	322.98		
23644	F-BF.1	EQ	1	0.56	0.79	0.44	0.56				254.63	313.18			
23646	F-IF.6	EQ	1	0.05	0.08	0.95	0.05				287.93	296.90			
23651	F-IF.4	EQ	1	0.29	0.63	0.71	0.29				273.29	333.12			
23652	A-SSE.3	ETC	1	0.02	0.49	0.98	0.02				286.11	363.50			
23660	F-IF.2	EQ	1	0.21	0.72	0.79	0.21				274.47	341.63			
23661	F-IF.5	EQ	1	0.03	0.69	0.97	0.03				284.77	400.45			
23662	F-IF.5	EQ	1	0.02	0.71	0.98	0.02				288.28	415.68			
23668	F-LE.5	EQ	1	0.21	0.77	0.79	0.21				271.94	344.62			
23670	A-CED.2	EQ	1	0.53	0.71	0.47	0.53				258.67	313.49			
23671	F-IF.4	EQ	1	0.31	0.30	0.69	0.31				282.31	308.25			
23676	F-LE.5	EQ	1	0.17	0.65	0.83	0.17				279.92	350.00			
23677	S-ID.6	EQ	1	0.01	0.21	0.99	0.01				288.65	319.81			
23678	A-REI.11	EQ	1	0.16	0.71	0.84	0.16				276.09	351.16			
23679	A-SSE.3	EQ	1	0.17	0.71	0.83	0.17				274.07	345.76			
23643	N-Q.2	ETC	1	0.42	0.48	0.58	0.42				271.94	312.37			
23647	S-ID.6	ETC	1	0.09	0.38	0.91	0.09				286.19	331.32			
23648	A-SSE.1	ETC	1	0.21	0.66	0.79	0.21				277.23	340.61			
23649	A-SSE.1	ETC	2	0.71	0.55	0.13	0.31	0.56			247.45	271.52	306.88		
23655	F-LE.5	ETC	2	0.57	0.69	0.21	0.43	0.36			245.72	280.45	326.80		

Item Number	Benchmark	Type	Points	Average Score	Adjusted Polyserial / Biserial	Proportion by Score Point in CR and Response Option in MC					Distractor Adjusted Biserial for MC/ Mean Score for CR				
						0	A/1	B/2	C/3	D/4	0	A/1	B/2	C/3	D/4
23624	A-APR.3	GI	1	0.14	0.67	0.86	0.14				278.34	352.08			
23642	A-REI.12	GI	2	0.32	0.52	0.48	0.41	0.11			271.69	289.30	361.92		
2001162	A-REI.12	GI	1	0.32	0.69	0.68	0.32				271.69	331.29			

Note: EQ = equation response item; ETC = editing task choice item; GI = grid item

Table B-2. Field-Test Items: Classical Item Statistics for EOC Algebra 2

Item Number	Benchmark	Type	Points	Average Score	Adjusted Polyserial / Biserial	Proportion by Score Point in CR and Response Option in MC					Distractor Adjusted Biserial for MC/ Mean Score for CR				
						0	A/1	B/2	C/3	D/4	0	A/1	B/2	C/3	D/4
23629	A-REI.4	EQ	1	0.03	0.68	0.97	0.03				282.29	391.68			
23631	A-SSE.2	EQ	1	0.39	0.70	0.61	0.39				260.48	321.97			
23632	A-SSE.3	EQ	1	0.08	0.62	0.92	0.08				278.42	362.44			
23634	A-SSE.3	EQ	1	0.05	0.71	0.95	0.05				281.32	379.88			
23640	A-CED.4	EQ	1	0.16	0.65	0.84	0.16				273.04	344.31			
23645	F-BF.5	EQ	1	0.07	0.58	0.93	0.07				279.50	362.70			
23653	F-BF.1	EQ	1	0.21	0.65	0.79	0.21				270.10	340.23			
23654	A-REI.2	EQ	1	0.21	0.45	0.79	0.21				276.17	322.81			
23658	A-SSE.2	EQ	1	0.33	0.69	0.67	0.33				260.96	323.75			
23663	F-IF.7	EQ	1	0.04	0.68	0.96	0.04				277.91	386.70			
23664	F-IF.8	EQ	1	0.09	0.70	0.91	0.09				273.00	369.73			
23665	F-LE.2	EQ	1	0.11	0.70	0.89	0.11				271.86	360.48			
23669	A-REI.1	EQ	1	0.50	0.77	0.50	0.50				251.40	314.52			
23672	A-SSE.3	EQ	1	0.04	0.35	0.96	0.04				281.95	331.86			
23674	A-APR.4	EQ	1	0.30	0.73	0.70	0.30				263.15	332.18			
23680	F-LE.5	EQ	1	0.15	0.57	0.85	0.15				272.21	336.35			
23630	A-SSE.1	ETC	1	0.23	0.47	0.77	0.23				273.06	319.16			
23673	F-IF.7	ETC	1	0.29	0.64	0.71	0.29				265.60	327.82			

Note: EQ = equation response item; ETC = editing task choice item

Appendix C: Field-Test Items: Item Parameters

Table C-1. Field-Test Items: Item Parameters for EOC Algebra 1

Item Number	Benchmark	Type	Step 1	Step 2	Infit Value	Outfit Value
23625	A-CED.1	EQ	2.97719		0.81	0.33
23626	A-CED.4	EQ	1.40379		0.81	0.58
23627	A-REI.12	EQ	1.82328		0.92	0.64
23628	A-REI.4	EQ	3.92760		0.86	0.29
23633	A-SSE.3	EQ	1.77680		0.88	0.67
23635	F-IF.6	EQ	3.38569		1.17	1.62
23636	F-IF.7	EQ	2.58160		0.84	0.43
23638	F-IF.9	EQ	1.85791		0.75	0.50
23639	F-IF.9	EQ	4.01232		0.78	0.19
23641	A-REI.1	EQ	-0.74690	-0.68120	0.79	0.74
23644	F-BF.1	EQ	-0.82034		0.75	0.69
23646	F-IF.6	EQ	3.02234		1.18	2.38
23651	F-IF.4	EQ	0.72032		0.89	0.83
23652	A-SSE.3	EQ	4.31506		1.01	0.55
23660	F-IF.2	EQ	1.17634		0.82	0.65
23661	F-IF.5	EQ	3.77558		0.83	0.42
23662	F-IF.5	EQ	4.28653		0.84	0.19
23668	F-LE.5	EQ	1.13330		0.77	0.56
23670	A-CED.2	EQ	-0.65862		0.81	0.83
23671	F-IF.4	EQ	0.57066		1.22	1.31
23676	F-LE.5	EQ	1.63171		0.89	0.67
23677	S-ID.6	EQ	4.53561		1.08	2.10
23678	A-REI.11	EQ	1.58557		0.82	0.59
23679	A-SSE.3	EQ	1.45518		0.80	0.61
23643	N-Q.2	ETC	-0.05472		1.04	1.08
23647	S-ID.6	ETC	2.46491		1.10	1.30
23648	A-SSE.1	ETC	1.22210		0.86	0.70
23649	A-SSE.1	ETC	-2.01354	-1.07823	0.98	1.14
23655	F-LE.5	ETC	-1.69893	0.01349	0.85	0.87

Item Number	Benchmark	Type	Step 1	Step 2	Infit Value	Outfit Value
23624	A-APR.3	GI	1.78210		0.85	0.63
23642	A-REI.12	GI	-0.46145	1.69747	1.07	1.06
2001162	A-REI.12	GI	0.52645		0.84	0.74

Note: EQ = equation response item; ETC = editing task choice item; GI = grid item

Table C-2. Field-Test Items: Item Parameters for EOC Algebra 2

Item Number	Benchmark	Type	Step 1	Infit Value	Outfit Value
23629	A-REI.4	EQ	3.54562	0.85	0.35
23631	A-SSE.2	EQ	-0.10866	0.80	0.74
23632	A-SSE.3	EQ	2.30806	0.89	0.56
23634	A-SSE.3	EQ	2.83251	0.84	0.37
23640	A-CED.4	EQ	1.34451	0.88	0.66
23645	F-BF.5	EQ	2.46353	0.88	0.74
23653	F-BF.1	EQ	0.98004	0.86	0.69
23654	A-REI.2	EQ	1.03213	1.05	1.05
23658	A-SSE.2	EQ	0.11484	0.81	0.74
23663	F-IF.7	EQ	3.20667	0.85	0.35
23664	F-IF.8	EQ	2.08789	0.79	0.49
23665	F-LE.2	EQ	1.80991	0.82	0.53
23669	A-REI.1	EQ	-0.70209	0.76	0.71
23672	A-SSE.3	EQ	3.19454	1.02	1.20
23674	A-APR.4	EQ	0.33913	0.78	0.67
23680	F-LE.5	EQ	1.38375	0.93	0.79
23630	A-SSE.1	ETC	0.79701	1.03	0.98
23673	F-IF.7	ETC	0.39112	0.85	0.82

Note: EQ = equation response item; ETC = editing task choice item

Appendix D: Field-Test Items: Differential Item Functioning Classifications

Table D-1. Field-Test Items: Differential Item Functioning Classifications for EOC Algebra 1

Item Number	Benchmark	Type	ELL/No ELL	SPED/No SPED	Lunch/No Lunch	Female/ Male	Hawaiian/ White	Filipino/ White	Japanese/ White	Hawaiian/ Filipino	Hawaiian/ Japanese	Filipino/ Japanese
23625	A-CED.1	EQ	-A	-A	-A	+A	-A	-A	+A	-A	-A	-A
23626	A-CED.4	EQ	-A	+A	-A	+A	-A	-A	-A	+A	+A	-A
23627	A-REI.12	EQ	+A	-A	-A	-A	-A	+A	-A	-A	-A	+A
23628	A-REI.4	EQ	-A	-A	-A	-A	-A	-A	+A	-A	-A	-A
23633	A-SSE.3	EQ	-A	+A	-A	+A	+A	+A	+A	+A	-A	-A
23635	F-IF.6	EQ	+A	-A	+A	+A	+A	+A	+A	-A	-A	-A
23636	F-IF.7	EQ	-A	-A	-A	-A	-A	-A	+A	-A	-A	-A
23638	F-IF.9	EQ	-A	-A	+A	-A	-A	+A	-A	-A	-A	+A
23639	F-IF.9	EQ	-A	-A	-A	-A	-A	-A	-A	-A	-A	-A
23641	A-REI.1	EQ	-A	+A	+A	+A	+A	+A	+B	+A	-A	-A
23644	F-BF.1	EQ	+A	-B	-A	+A	-A	+A	+A	-A	-A	-A
23646	F-IF.6	EQ	-A	+A	-A	-A	-A	+A	+A	-A	-A	-A
23651	F-IF.4	EQ	-A	-A	-A	+A	-A	-A	-A	-A	-A	-A
23652	A-SSE.3	EQ	-A	+A	-A	-A	+A	+A	+A	+A	+A	-A
23660	F-IF.2	EQ	+A	-A	-A	+A	-A	-A	+A	-A	-A	-A
23661	F-IF.5	EQ	-A	-A	-A	-A	-A	-A	+A	-A	-C	-C
23662	F-IF.5	EQ	-A	-A	-A	-A	+A	-A	+A	+A	-A	-A
23668	F-LE.5	EQ	-A	-A	-A	-A	-A	-A	-A	+A	-A	+A
23670	A-CED.2	EQ	-A	-C	-A	+A	-A	+A	-A	-A	-A	+A
23671	F-IF.4	EQ	+A	+A	-A	-A	+A	+A	-A	-A	+A	+A
23676	F-LE.5	EQ	-A	-A	-A	-A	+A	+A	+A	+A	+A	+A
23677	S-ID.6	EQ	-A	-A	-A	-A	-A	+A	-A	-A	+A	+A
23678	A-REI.11	EQ	-A	+A	+A	+A	+A	+A	+C	-A	-C	-A
23679	A-SSE.3	EQ	+A	-A	-A	-A	-A	+A	+C	-A	-A	-A
23643	N-Q.2	ETC	+A	+A	-A	-A	-A	-A	+A	-A	-A	-A
23647	S-ID.6	ETC	-A	+A	+A	+A	-A	+A	-A	-B	-A	+A
23648	A-SSE.1	ETC	-B	-A	-A	-A	-A	+A	-A	-A	-A	+A
23649	A-SSE.1	ETC	-A	-C	-A	+A	-B	-C	-A	+A	-A	-A
23655	F-LE.5	ETC	-A	-B	-A	+A	-B	-A	-A	-A	-A	+A

Item Number	Benchmark	Type	ELL/No ELL	SPED/No SPED	Lunch/No Lunch	Female/ Male	Hawaiian/ White	Filipino/ White	Japanese/ White	Hawaiian/ Filipino	Hawaiian/ Japanese	Filipino/ Japanese
23624	A-APR.3	GI	-A	-A	-A	+A	-A	+A	-A	-A	-A	+A
23642	A-REI.12	GI	+A	-A	-A	+A	+A	+A	+A	-A	+A	+A
2001162	A-REI.12	GI	+A	-A	+A	-A	+A	+A	+A	-A	+A	+A

Note: EQ = equation response item; ETC = editing task choice item; GI = grid item

Table D-2. Field-Test Items: Differential Item Functioning Classifications for EOC Algebra 2

Item Number	Benchmark	Type	ELL/No ELL	SPED/No SPED	Lunch/No Lunch	Female/ Male	Hawaiian/ White	Filipino/ White	Japanese/ White	Hawaiian/ Filipino	Hawaiian/ Japanese	Filipino/ Japanese
23629	A-REI.4	EQ	-A	-A	-A	-A	-A	-A	-A	+A	+A	+A
23631	A-SSE.2	EQ	-A	-A	-A	+A	-B	-A	-A	-A	-A	-A
23632	A-SSE.3	EQ	-A	-A	+A	-A	+A	+A	+A	-A	+A	+A
23634	A-SSE.3	EQ	-A	-A	-A	-B	+A	-A	-A	+A	+A	-A
23640	A-CED.4	EQ	-A	-A	-A	-A	-A	-A	-A	+A	-A	-A
23645	F-BF.5	EQ	-A	-A	+A	+A	+A	+B	+A	-A	+A	+B
23653	F-BF.1	EQ	-A	+A	-A	+A	-A	-A	+A	-A	-B	-A
23654	A-REI.2	EQ	-A	-A	-A	-A	-A	+A	+A	-A	-A	-A
23658	A-SSE.2	EQ	+A	-A	-A	+A	-A	+A	+A	-A	-A	-A
23663	F-IF.7	EQ	-A	-A	+A	+A	-A	+A	-A	-A	+A	+A
23664	F-IF.8	EQ	+A	-A	+A	+A	-A	-A	-A	-A	-A	-A
23665	F-LE.2	EQ	-A	+A	-A	-A	-A	-A	-A	-A	-A	+A
23669	A-REI.1	ETC	-A	-A	-A	+A	+A	-A	+B	+A	-C	-C
23672	A-SSE.3	ETC	+A	+A	-A	-B	-A	-A	-A	-A	-A	+A
23674	A-APR.4	EQ	-A	+A	+A	+A	+A	+A	+B	+A	-A	-A
23680	F-LE.5	EQ	+A	-A	-A	-A	-A	-A	+A	-A	-A	-A
23630	A-SSE.1	ETC	+A	-A	-A	+A	-C	-B	-A	-A	-A	-A
23673	F-IF.7	ETC	+A	-A	+A	-A	-A	+A	-A	-A	+A	+A

Note: EQ = equation response item; ETC = editing task choice item