Annual Technical Report

Indiana Learning Evaluation Assessment Readiness Network (ILEARN)

2023–2024

ACKNOWLEDGMENTS

1   CONTENTS

TABLES

FIGURES

EXHIBITS

APPENDICES

Appendix 7-B, 2024 ILEARN U.S. Government Standards-Confirmation Report
Appendix 7-C, 2024 ILEARN Science Standard-Setting Report

EXECUTIVE SUMMARY

This executive summary provides an overview of validity evidence of the Indiana Learning Evaluation Assessment Readiness Network (ILEARN) to support a validity argument regarding the uses of and inferences for the ILEARN assessments as well as a summary of the ILEARN program and its spring 2024 test administration.

Overview of Validity Evidence

Intended uses for ILEARN test scores include school accountability, feedback about student and class performance, measurement of student growth over time, evaluation of performance gaps between groups, and diagnosis of individual student strengths and opportunities for improvement. Evidence for the validity of test score interpretations is central to substantiating claims that ILEARN test scores can fulfill their intended purpose to evaluate the effectiveness with which Indiana corporations and schools teach students the Indiana Academic Standards (IAS) and individual student's performance on IAS by the end of each school year.

Sufficient evidence exists to support the principal claims for the ILEARN test scores, including that test scores indicate the degree to which students have achieved the Indiana Academic Standards at each grade level and that students scoring at the Proficient level or higher demonstrate levels of achievement consistent with national benchmarks that indicate they are on track for college readiness.

ILEARN test content validity is supported by the strong alignment of ILEARN to Indiana's Academic Standards (content standards). Alignment is achieved by the rigorous item development process, which begins with the content standard and considers those standards throughout the highly iterative development and review process. The ILEARN test blueprints specify the range and priority level with which each of the content strands and standards will be covered in each test administration and complete the link between the Indiana Academic Standards and the ILEARN content-based test score interpretations. Additionally, Smarter Balanced conducted cognitive lab studies to collect test takers' performance strategies or responses to particular items. Since the majority of ILEARN items pool come from the Smarter Balanced, results from these cognitive lab studies may be applied to ILEARN. ILEARN items are developed to measure specific constructs and intellectual processes; therefore, evidence described in this report that test takers have engaged in relevant performance strategies to answer the items correctly supports the validity of the test scores.

Additionally, to minimize the impact of construct-irrelevant factors in assessing student achievement, ILEARN adopted universal design which removes barriers to access for the widest range of students possible. Test development specialists and IDOE contracted item writers receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including tasks, items, and manipulatives. In the review process, adherence to the principles of universal design is verified.

The validity evidence regarding the internal structure of the assessments has also been provided in this report. Based on the analysis of the degree to which the underlying factor structure of a construct is congruent with the empirical investigations about the unidimensionality of that construct, the relationships among ILEARN test items and test components are representative of the proposed underlying construct for test score interpretations. The evidence showed that the methods for reporting ILEARN strand scores align with the underlying structure of the test and provide evidence for appropriateness of the selected Item Response Theory (IRT) models.

The interpretation of the ILEARN test scores rests fundamentally on how test scores relate to performance standards, which define the extent to which students have achieved the expectations defined in the Indiana Academic Standards. ILEARN test scores are reported with respect to four proficiency levels, demarcating the degree to which Indiana students participating in ILEARN have achieved the learning expectations defined by the Indiana Academic Standards. The standardized and rigorous procedures that Indiana educators, serving as standard setting panelists, followed to recommend performance standards in the standard setting process after spring 2019 test administration provided central and strong evidence to support the validity of test score interpretations regarding performance standards.

Summary of the Assessment Program

The ILEARN assessments are designed to measure college-and-career readiness as defined by the IAS and support the claim that students in grades 3 through 12 are demonstrating progress toward college-and-career readiness in ELA, mathematics, science, and social studies.

ILEARN assessments were created using a variety of item types from several sources, including licensed item banks Smarter Balanced Assessment Consortium (Smarter) and Independent College and Career Ready (ICCR), Memorandum of Understanding (MOU), and custom Indiana development. Item development efforts support the goal of high-quality items through rigorous development processes managed and tracked by a content development platform that ensures every item flows through the correct

sequence of reviews and captures every comment and change to the item. The blueprint design and test construction also follow rigorous procedures to support the validity of the claims that ILEARN assessments are designed to support.

ILEARN assessments, as assessment instruments, have established test administration procedures that support useful interpretations of score results, as specified in Standard 6.0 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014). Various sources of test-administration–related evidence for the validity of the assessment results are presented in this report, including testing procedures, accommodations, Test Administrator (TA) training and resources, and test security procedures implemented for ILEARN.

ILEARN scores are provided to corporations and schools through the Indiana Centralized Reporting System (CRS). The CRS is designed to assist stakeholders in reviewing and downloading the test results and in understanding and appropriately using the results of the state assessments. It provides information on student performance and aggregated summaries at several levels—state, corporation, school, and roster. Assessment results on student performance on the test can be used to help teachers or schools make decisions on how to support students' learning. Aggregate score reports on the teacher and school level provide information about the strengths and improvement opportunities of students and can be used to improve teaching and student learning.

Finally, quality assurance procedures are enforced throughout all stages of ILEARN test development, configuration, administration, and scoring and reporting. Those procedures ensure the accuracy and integrity of the test scores as well as strengthen the validity of the score interpretation.

Overview of the Chapters

This technical report begins with Chapter 1, an introduction and background of the assessment, to offer a brief but important overview of the purpose of the assessment and its background as well as recent changes to the test administration. Chapter 2 provides a review of validity evidence evaluated to date. Chapter 3 presents the results of the 2023–2024 ILEARN test administration, which provides summaries of the test-taking student population and their performance on the assessments. In addition, these sections describe administration-specific evidence for the reliability of ILEARN assessments, including internal consistency reliability, standard errors of measurement (SEMs), and the reliability of performance-level classifications. Chapter 4 describes the design and development of ILEARN assessments, including the Indiana Academic Standards, which define the content domain to be assessed by ILEARN; the

development of test specifications, including blueprints, that ensure the breadth and depth of the content domain is adequately sampled by the assessments; and test development procedures that ensure alignment of test forms with the blueprint specifications. This chapter also delineates Cambium Assessment, Inc.'s (CAI) adaptive algorithm that delivers the computerized ILEARN assessments to Indiana students. Chapter 5 discusses the test administration procedures, including eligibility for participation in ILEARN assessments; testing conditions, including accessibility tools and accommodations; systems security for assessments administered online; and test security procedures for all test administrations. Chapter 6 describes the procedures used to scale and equate ILEARN assessments for scoring and reporting. Chapter 7 outlines the procedures used to identify and adopt performance standards for the ILEARN assessments. Chapter 8 provides a description of the score reporting system and the interpretation of test scores. Finally, Chapter 9 provides an overview of the quality assurance (QA) processes CAI uses to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

1. INTRODUCTION AND BACKGROUND

1.1 PURPOSES OF THE ASSESSMENT

The Indiana Learning Evaluation Assessment Readiness Network (ILEARN) is Indiana's standards-referenced, summative accountability assessment measuring student achievement and growth.

The primary intended use of the ILEARN assessment system is for school accountability, to ensure that educators, schools, and districts are providing effective instruction of the Indiana Academic Standards (IAS). ILEARN yields overall and reporting-category-level test scores at the student level and at other levels of aggregation to reflect degrees of student performance and mastery of the IAS. ILEARN supports instruction and student learning by providing immediate feedback to educators and parents based on the IAS, which can be used to inform instructional strategies and to remediate or enrich curriculum. An array of reporting metrics allows achievement to be monitored at both the student and aggregate levels and growth to be measured at both levels over time.

ILEARN, as an assessment instrument, has established test administration procedures that support useful interpretations of score results, as specified in Standard 6.0 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

1.2 BACKGROUND OF THE ASSESSMENTS

ILEARN was constructed to measure student achievement in English/language arts (ELA), mathematics, science, and social studies relative to the Indiana Academic Standards. ILEARN was first administered to students during the 2018–2019 academic year, replacing the Indiana Statewide Testing for Educational Progress–Plus (ISTEP+). In 2023-2024 academic year, ILEARN adopted and started administering Next Generation Science Standards (NGSS) Science Assessment.

1.2.1 DEVELOPMENT OF INDIANA ACADEMIC STANDARDS

The IAS were approved by the Indiana State Board of Education in April 2014 for English/language arts (ELA) and mathematics, and in March 2015 for social studies. The IAS for science were revised in 2010, 2016, and 2020 to reflect changes in science content. The IAS were most recently updated in 2023 for all subject areas. The IAS are intended to implement more rigorous standards that promote college-and-career readiness, with the goal of challenging and motivating Indiana's students to acquire stronger critical thinking, problem solving, and communication skills.

1.2.2 ONLINE ITEM POOL CONSTRUCTION

The 2023–2024 *ILEARN* item pools each contain sufficient numbers of items per grade and content area to ensure that students would be administered items representing the breadth and depth of the content standards identified in the test specifications while also adapting item selection to maximize test information near each student's ability level. In ELA, since item selection is passage-dependent, it is more challenging to provide precise estimates of each student's true achievement level across the range of proficiency than in mathematics and science.

With new items being developed and field tested in the spring administration of each year, the operational pool size for each assessment has constantly increased since 2018. The simulations show that a larger operational pool improves the adaptive item selection in terms of blueprint match, content coverage, and precision of the student ability estimation, especially the ability estimation for students with more extreme test scores.

1.3 RECENT CHANGES TO THE TEST

1.3.1 SCIENCE NEXT GENERATION SCIENCE STANDARDS (NGSS) AND COMPUTER SCIENCE TEST

In September 2022, the Indiana State Board of Education (SBE) adopted the K-12 Indiana Academic Standards in Science and Computer Science built with a focus on increasing student engagement in Science learning. The standards build on science theories as practiced in real world applications, build logically for students from K-12, focus a strengthened understanding and application of content, and integrate practices, cross-cutting concepts, and core ideas. Each standard, or performance expectation, outlines three dimensions of science learning: a Science and Engineering Practice (SEP), a Disciplinary Core Idea (DCI), and a Crosscutting Concept (CCC). The alignment of each standard to these dimensions ensures that students engage in a scientific activity (SEP) related to the standard, understand concepts and gain knowledge in a discipline (DCI), and do so while using a scientific worldview that can be applied across disciplines (CCC). This three-dimensional nature of the NGSS ensures that students learn to apply their knowledge and relate concepts across science disciplines.

The new Computer Science standards are meant to reflect a future-focused vision for computer science education that build logically from kindergarten through grade eight, integrate core student practices, reflect the growing range of fields within computer science, and focus on computer science as a collaborative endeavor. The Core Practices and Core Concepts of each standard outlines the skills and knowledge students need to be able to communicate and think like a computer scientist. The Indiana Academic Standards for Grades K-8 Computer Science are designed to prepare students for high school courses in computer science and engage students to use knowledge and skills used in the real world.

CAI developed a shared science assessment item bank in collaboration with the states that were part of the Memorandum of Understanding (MOU) using a rigorous, structured process that engaged stakeholders at critical junctures. The items in the bank are linked to the Next Generation Science Standards, which participating states all use. The cluster-based science assessments utilizing this item bank were first administered in Indiana in spring 2024 for grades 4, 6, and End of Course (EOC) Biology.

1.4 OVERVIEW OF THE REPORT

This technical report documents the evidence that supports claims made for how the *ILEARN* assessment scores may be interpreted. While ILEARN is designed as a school accountability assessment and ILEARN results inform the state's calculations for school accountability, the primary and foremost purpose of this report is to reflect and support validity expectations of ILEARN data and reporting. Therefore, after Chapter 1 provides an overview of the purpose and intended uses of the assessment, Chapter 2 provides a review of validity evidence evaluated to date to support the intended uses and interpretations. Because evidence for the validity of test score interpretations will accrue over time, this chapter will be expanded as further evidence is collected.

Chapter 3 presents the results of the 2023–2024 ILEARN test administration. This chapter provides summaries of the test-taking student population and their performance on the assessments. In addition, these sections describe administration-specific evidence for the reliability of ILEARN assessments, including internal consistency reliability, standard errors of measurement (SEMs), and the reliability of performance-level classifications.

The remaining chapters are organized in a chronological order and document technical details of test development, administration, scoring, and reporting activities. Chapter 4 of this technical report describes the design and development of *ILEARN* assessments, including the Indiana Academic Standards, which define the content domain to be assessed by *ILEARN*; the development of test specifications, including blueprints, that ensure the breadth and depth of the content domain is adequately sampled by the assessments; and test development procedures that ensure alignment of test forms with the blueprint specifications. *ILEARN* is an online, adaptive assessment for English/language arts (ELA) for grades 3–8, mathematics for grades 3–8, science for grades 4, 6, and biology, and an online, fixed-form assessment for social studies for grades 5 and U.S. government. For the 2023–2024 school year, accommodated and paper-and-pencil versions of the assessments were available to students whose English Learner (EL) statuses or Section 504 Plans indicated that need. It describes the item development process and the sequence of reviews that each item must pass through before being eligible for *ILEARN* test administration. This chapter also delineates Cambium Assessment, Inc.'s (CAI) adaptive algorithm that delivers the computerized *ILEARN* assessments to Indiana students.

Chapter 5 discusses the test administration procedures, including eligibility for participation in *ILEARN* assessments; testing conditions, including accessibility tools and accommodations; systems security for assessments administered online; and test security procedures for all test administrations.

Chapter 6 describes the procedures used to scale and equate *ILEARN* assessments for scoring and reporting. Chapter 7 outlines the procedures used to identify and adopt performance standards for the *ILEARN* assessments. Chapter 8 provides a description of the score reporting system and the interpretation of test scores.

Finally, Chapter 9 provides an overview of the quality assurance (QA) processes CAI uses to ensure that all test development, administration, scoring, and reporting activities are conducted with fidelity to the developed procedures.

2. THE VALIDITY OF TEST SCORE INTERPRETATIONS

1.5 VALIDITY EVIDENCE

The term *validity* refers to the degree to which test score interpretations are supported by evidence, and it speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is the most fundamental component of test design and evaluation. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the standards describe the range of evidence that may be brought to support the validity of test score interpretations.

The first source of validity evidence is the relationship between the test content and the intended test construct. For test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct. Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance.

The second source of validity evidence is based on "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014). This evidence is collected by surveying test takers about their performance strategies or responses to particular items. Because items are developed to measure specific constructs and intellectual processes, evidence that test takers have engaged in relevant performance strategies to answer the items correctly supports the validity of the test scores.

The third source of validity evidence is based on the internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. Differential item functioning (DIF), which determines whether particular items may function differently for subgroups of test takers, is one method of analyzing the internal structure of tests. Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis.

A fourth source of validity evidence is the relationship of the test scores to external variables. The *Standards* (AERA, APA, & NCME, 2014) divide this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs; conversely, discriminant

evidence distinguishes the test from other measures intended to assess different constructs. A multi-trait multi-method matrix can be used to analyze both convergent and discriminant evidence. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends on the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

The fifth source of validity evidence is that the intended and unintended consequences of test use should be included in the test validation process. Determining the validity of the test should depend upon evidence directly related to the test; external factors should not influence this process. For example, if an employer administers a test to determine the hiring rates for different groups of people and the results indicate an unequal distribution of skills related to the measurement construct, that would not necessarily imply a lack of test validity. However, if the unequal distribution of scores is, in fact, due to an unintended, confounding aspect of the test, that would interfere with the test's validity. Test use should align with the test's intended purpose.

Supporting a validity argument requires multiple sources of validity evidence. This then allows for an evaluation of whether sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining test validity first requires an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

The kinds of evidence required to support the validity of test score interpretations depend on the claims made for how test scores may be interpreted. Moreover, the standards make it explicit that validity is an attribute not of tests but rather of test score interpretations. Thus, the test itself is not assessed for validity; instead, the intended interpretation and use of test scores are evaluated.

There are several intended uses for ILEARN test scores, including school accountability, feedback about student and class performance, measurement of student growth over time, evaluation of performance gaps between groups, and diagnosis of individual student strengths and weaknesses. Each of these intended uses requires claims to be made about the interpretation of test scores, and the strength of those claims rests on the validity evidence supporting them. Some validity evidence will be central to all of the claims, including evidence showing that test items and administrations align with Indiana Academic Standards. Other evidence may target more specific claims, such as evidence for measurement of student growth. Validity

evidence should therefore be evaluated with respect to the claim that it is purported to support.

Determining whether the test measures the intended construct is central to evaluating the validity of test score interpretations. Such an evaluation in turn requires a clear definition of the measurement construct. For Indiana's ILEARN assessments, the definition of the measurement construct is provided by the Indiana Academic Standards.

A summary of the types of validity evidence gathered is captured and illustrated in the following chart:
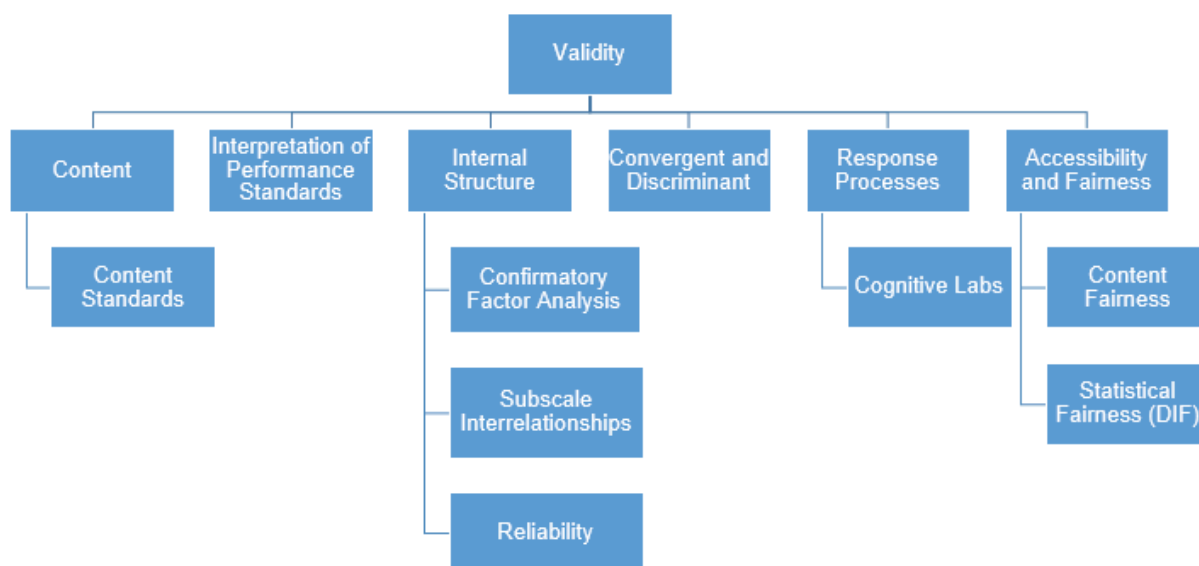
1.6 EVIDENCE BASED ON TEST CONTENT

Determining whether the test measures the intended construct is central to evaluating the validity of test score interpretations. Such an evaluation in turn requires a clear definition of the measurement construct. For Indiana's *ILEARN* assessments, the definition of the measurement construct is provided by the Indiana Academic Standards which specify what students should know and be able to do by the end of the year for each grade level in order for them to graduate prepared for post-secondary education or entry into the workforce. The *ILEARN* assessments are designed to measure student progress toward achievement of the Indiana Academic Standards. Therefore, the validity of *ILEARN* test score interpretations critically depends on the degree to which test content aligns with expectations for student learning as specified in the Indiana Academic Standards.

Several processes are in place to ensure ILEARN fully aligns to the content standards, including a rigorous item development process, adherence to test blueprints,

consideration of cognitive complexity, and standard setting based on content standards. These processes include the Indiana State Board of Education, test developers, and educator and stakeholder committees.

Ensuring the alignment of test items to their intended content standards establishes a critical link between the expectations for student achievement articulated in the Indiana Academic Standards with the *ILEARN* item content. The *ILEARN* test blueprints, in turn, specify the range and depth with which each of the content strands and standards will be covered in each test administration and complete the link between the Indiana Academic Standards and the *ILEARN* content-based test score interpretations.

The test blueprints drive item selection in the adaptive algorithm used to administer *ILEARN* assessments. The adaptive algorithm seeks to meet the following three objectives:

- To satisfy blueprint constraints
- To maximize overall test information near the student's ability estimate
- To maximize test information within each of the reporting strands.

Each item satisfies multiple blueprint elements. As the test progresses, the weight of item selections increases for blueprint elements that have not been met, while items measuring blueprint elements that have been satisfied are no longer considered. The adaptive algorithm is configured for each assessment to ensure that all critical blueprint elements are satisfied in each test administration.

Unlike fixed-form tests, in which the same test form is administered to all students statewide, the *ILEARN* assessments are administered adaptively to students within the same classrooms and schools administer different samples of items from the subject-area pool. While each student may be administered only one or two items per content standard, performance indicators at the classroom and school levels are based on a larger, more representative sample of the content domain than is possible with fixed-form assessments. This ensures that teachers and schools are held accountable for instruction across the full range of the academic content standards.

Because directly measuring student achievement against each benchmark in the Indiana Academic Standards would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the Indiana Academic Standards. To ensure that each student is assessed on the intended breadth and depth of the standards, item selection in the Test Delivery System (TDS) is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark. The test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the Indiana Academic Standards is evaluated, alignment of test blueprints with the content standards is critical.

1.6.1 CONTENT STANDARDS

The Indiana Academic Standards were approved by the Indiana State Board of Education in April 2014 for English/language arts (ELA) and mathematics and in March 2015 for social studies. The Indiana Academic Standards for science were revised in 2010, 2016, and 2020 to reflect changes in science content. The IAS for all subject areas were most recently revised in 2023. The standards are available for review at the following URLs:

- https://www.in.gov/doe/students/indiana-academic-standards/englishlanguage-arts/
- https://www.in.gov/doe/students/indiana-academic-standards/mathematics/
- https://www.in.gov/doe/students/indiana-academic-standards/science-and-computer-science/
- https://www.in.gov/doe/students/indiana-academic-standards/social-studies/

Blueprints were developed to ensure that the test and the items aligned to the prioritized standards they were intended to measure. A complete description of the blueprint and test construction process can be found in Chapter 4 of this report, Item Development and Test Construction.

Table 1 through Table 4 present the number of items in the 2023–2024 item pool that measured each reporting category by grade for ELA, mathematics, science, and social studies, respectively.

Table 1: Number of Items for Each Reporting Category, ELA

Grade	Reporting Category	Number of Items
3	Key Ideas and Textual Support/Vocabulary	144
	Structural Elements and Organization/ Connection of Ideas/Media Literacy	92
	Writing	124
	Reading Foundations	6
	Speaking and Listening	186
4	Key Ideas and Textual Support/Vocabulary	236
	Structural Elements and Organization/ Connection of Ideas/Media Literacy	80
	Writing	119
	Speaking and Listening	61
5	Key Ideas and Textual Support/Vocabulary	144
	Structural Elements and Organization/	118

Grade	Reporting Category	Number of Items
	Connection of Ideas/Media Literacy	
	Writing	167
	Speaking and Listening	72
6	Key Ideas and Textual Support/Vocabulary	150
	Structural Elements and Organization/ Connection of Ideas/Media Literacy	138
	Writing	62
	Speaking and Listening	91
7	Key Ideas and Textual Support/Vocabulary	222
	Structural Elements and Organization/ Connection of Ideas/Media Literacy	118
	Writing	151
	Speaking and Listening	30
8	Key Ideas and Textual Support/Vocabulary	166
	Structural Elements and Organization/ Connection of Ideas/Media Literacy	68
	Writing	125
	Speaking and Listening	128

Table 2: Number of Items for Each Reporting Category, Mathematics

Grade	Reporting Category	Number of Items
3	Algebraic Thinking and Data Analysis	168
	Computation	67
	Geometry and Measurement	134
	Number Sense	97
	Process Standards	33
4	Algebraic Thinking and Data Analysis	89
	Computation	124
	Geometry and Measurement	117
	Number Sense	141
	Process Standards	29

Grade	Reporting Category	Number of Items
5	Algebraic Thinking	95
	Computation	88
	Geometry and Measurement, Data Analysis, and Statistics	172
	Number Sense	85
	Process Standards	24
6	Algebra and Functions	215
	Computation	87
	Geometry and Measurement, Data Analysis, and Statistics	113
	Number Sense	119
	Process Standards	29
7	Algebra and Functions	219
	Data Analysis, Statistics, and Probability	72
	Geometry and Measurement	62
	Number Sense and Computation	190
	Process Standards	18
8	Algebra and Functions	149
	Data Analysis, Statistics, and Probability	40
	Geometry and Measurement	62
	Number Sense and Computation	79
	Process Standards	56

Table 3: Number of Items for Each Reporting Category, Science

Grade	Reporting Category	Number of Items
4	Physical Science	89
	Life Science	33
	Earth and Space Science	63
	Computer Science	19
6	Physical Science	27
	Life Science	73

Grade	Reporting Category	Number of Items
	Earth and Space Science	26
	Computer Science	13
	From Molecules to Organisms: Structure and Function	54
Biology	Ecosystems: Interactions, Energy, and Dynamics	55
	Heredity and Evolution	64

Table 4: Number of Items for Each Reporting Category, Social Studies

Grade	Reporting Category	Number of Items
	Civics and Government	27
5	Geography and Economics	21
	History	20
	Functions of Government	20
U.S. Government	Historical Foundations of American Government	14
	Institutions and Processes of Government	20

2.2.2 REVIEW PROCESS FOR ITEMS APPEARING IN ILEARN OPERATIONAL TEST ADMINISTRATION

This section describes the item review procedures used to ensure item accuracy and alignment with the Indiana Academic Standards. All items developed by CAI follow a standard item review process whereby item reviews proceed initially through a series of internal CAI reviews before items are deemed eligible for review by external content experts. Most of the CAI content staff members responsible for conducting internal reviews are former classroom teachers who hold degrees in education and/or their respective content areas. Each item passes through the following four internal review steps before it is designated as eligible for review by Indiana Department of Education (IDOE) content specialists:

1. Preliminary Review, conducted by a group of CAI content area experts
2. Content Review 1, performed by a Level 3–4 CAI content specialist
3. Edit Review 1, in which a copy editor checks the item for correct grammar and usage
4. Senior Content Review, conducted by a Level 4–5 lead content expert

At every stage of the item review process, beginning with the preliminary review, CAI's test developers analyze each item to ensure the following:

- The item is well aligned with the intended content standard.
- The item conforms to the item specifications for the target being assessed.
- The item is based on a quality idea or real-world phenomenon (science) (i.e., it assesses something worthwhile in a reasonable way).
- The item aligns correctly to a DOK level (except for NGSS science).
- The vocabulary used in the item is appropriate for the intended grade or age and subject matter, and it takes into consideration language accessibility, bias, and sensitivity.
- The item content is accurate and straightforward.
- Any accompanying graphic and stimulus materials are necessary to answer the question.
- The item stem is clear, concise, and succinct; it contains enough information to ensure that it will be understood; it is stated positively (and does not rely on negatives such as no, not, none, or never unless absolutely necessary); and it ends with a question.
- For selected-response items, the set of response options are succinct; parallel in structure, grammar, length, and content; sufficiently distinct from one another; and all plausible, but with only one correct option.
- There is no obvious or subtle cueing within the item.
- The score points for constructed-response items are clearly defined.
- For machine-scored constructed-response (MSCR) items, the item is scored as intended at each score point in the rubric.

On the basis of their reviews of each item, the test developers may accept the item and classification as written, revise the item, or reject the item outright.

Items passing through the internal review process are sent to IDOE for its review. At this stage, items may be further revised in accordance with any edits or changes requested by IDOE or rejected outright. Items at the IDOE review level pass through external reviews in which committees of Indiana educators and stakeholders assess each item's accuracy, alignment to the intended standard, and DOK level, as well as item fairness and language sensitivity. All items considered for inclusion in the *ILEARN* item pools are initially reviewed as follows:

- IDOE reviews to ensure that mathematics and ELA items developed in the ICCR item bank are eligible for CFC Review. At this stage, IDOE can request edits to wording, scoring, or alignment or DOK updates. A CAI director for mathematics or ELA reviews all IDOE-requested edits in light of the ICCR item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made.
- Indiana Content and Fairness Committee (CFC) Review ensures that each item is reviewed for content validity, grade-level appropriateness, alignment to the content standards, and accessibility and fairness. All custom and educator-authored Indiana development was taken to the CFC Review that combines the functions of CAI's Content Advisory Committee and the Language

Accessibility, Bias, and Sensitivity (LABS) Committee. Science items also undergo review by a cross-state Content Advisory Committee and a separate cross-state Fairness Advisory Committee, in which educators from other MOU states review the items as specified above.

- After all IDOE and committee recommended edits have been applied, experts apply accessibility markups (e.g., translations or text-to-speech), Accessibility markup is embedded into each item as part of the item development process rather than as a post-hoc process applied to completed test forms.
- Because ILEARN relies heavily on licensed banks, a process for ensuring alignment of those items to the IAS was developed by CAI and IDOE.
    - Prior to the spring 2019 administration, two item acceptance review meetings were held. Results of those meetings can be found in Volume 2 of the 2018–2019 Technical Report.
    - In November 2019, a third item acceptance review meeting was held for ELA and mathematics. Results of that meeting can be found in Volume 2 of the 2019–2020 Technical Report
    - Prior to the spring 2024 administration of the new Science tests, an item acceptance review meeting for acceptance of operational items was held in February 2023.

Items successfully passing through these committee review processes are then field-tested to ensure that they behave as intended when administered to students. Despite conscientious item development, some items perform differently than expected when administered to students. Using the item statistics gathered in field-testing to review item performance is an important step in constructing valid and equivalent operational test forms.

Classical item analyses ensure that items function as intended with respect to the underlying scales. Classical item statistics are designed not only to evaluate item difficulty and the relationship of each item to the overall scale (item discrimination) but also to identify items that may exhibit a bias across subgroups (differential item functioning [DIF] analyses).

Items flagged for review on the basis of their statistical performance must pass a three-stage review to be included in the final item pool from which operational forms are created. In the first stage of this review, a team of psychometricians reviews all flagged items to ensure that the data are accurate and properly analyzed, response keys are correct, and that there are no other obvious problems with the items.

The IDOE then convenes the data review committee to evaluate flagged field-test items in the context of each item's statistical performance. On the basis of their review of each item's performance, the data review committee may either recommend that a flagged item be rejected or deem the item eligible for inclusion in operational test administrations.

1.7 Evidence for Interpretation of Performance Standards

Alignment of test content to the Indiana Academic Standards ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the Indiana Academic Standards. However, the interpretation of the *ILEARN* test scores rests fundamentally on how test scores relate to performance standards, which define the extent to which students have achieved the expectations defined in the Indiana Academic Standards. *ILEARN* test scores are reported with respect to four proficiency levels, demarcating the degree to which *ILEARN* students have achieved the learning expectations defined by the Indiana Academic Standards. The cut score establishing the At Proficiency level of performance is the most critical, since it indicates that students are meeting grade-level expectations for achievement of the Indiana Academic Standards that they are prepared to benefit from instruction at the next grade level, and that they are on track for college and career readiness. Procedures used to adopt performance standards for the *ILEARN* assessments are therefore central to the validity of test score interpretations.

Following the operational administration of the *ILEARN* assessments in 2018–2019, a standard setting workshop was conducted to recommend a set of performance standards to the IDOE for reporting student performance of the Indiana Academic Standards. This section describes the standardized and rigorous procedures that Indiana educators, serving as standard setting panelists, followed to recommend performance standards. The workshops employed the *Bookmark* procedure, a widely used method in which standard setting panelists use their expert knowledge of the Indiana Academic Standards and student achievement to map the performance-level descriptors (PLDs) adopted by the IDOE onto an ordered-item book based on operational test forms administered to students in spring 2019. Chapter 7, Performance Standards, explains the standard setting procedures in more details.

Panelists were also provided with contextual information to help inform their primarily content-driven cut-score recommendations. The decision to provide panelists with contextual benchmark information was discussed during a meeting with the Indiana State Board of Education (SBOE) and Indiana's Technical Advisory Committee (TAC) and confirmed by the policy committee. Panelists recommending performance standards for the ELA and mathematics grades 3–8 assessments were provided with the approximate location of relevant National Assessment of Educational Progress (NAEP) and Smarter Balanced Assessment Consortium (Smarter) performance standards. Panelists recommending performance standards for the science grades 4, 6, and biology assessments were provided with the approximate location of relevant NAEP performance standards. Panelists recommending performance standards for the social studies grade 5 assessment were provided with the approximate location of relevant Smarter performance standards for grade 5 ELA. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists used benchmark information to locate performance

standards that converged across assessment systems, the validity of test score interpretations was bolstered.

In addition, panelists in ELA and mathematics were provided with feedback about the vertical articulation of their recommended performance standards so that they could view how the locations of their recommended cut scores for each grade-level assessment were placed in relation to the cut-score recommendations at the other grade levels. This approach allowed panelists to view their cut-score recommendations as a coherent system of performance standards, and further reinforced the interpretation of test scores as indicating both achievement of current grade-level standards, and preparedness to benefit from instruction in the subsequent grade level.

Following the recommendations of final performance standards and vertical moderation sessions to ensure articulation of recommended cut scores across grade levels, the recommended cut scores were presented to a stakeholder panel for review and comment.

Based on the recommended cut scores, Table 5 shows the estimated percentage of students meeting the *ILEARN* proficient standard for each assessment in spring 2019. Table 5 also shows the national percentages of students that meet the NAEP and Smarter proficient standards. Since NAEP is only delivered in Grades 4 and 8, the percentages in other grades were interpolated or extrapolated so estimated percentages were available in all grades. As Table 5 indicates, the performance standards recommended for *ILEARN* assessments are consistent with relevant NAEP and Smarter proficient benchmarks. Moreover, because the performance standards were vertically articulated in ELA and mathematics, the proficiency rates across grade levels are generally consistent.

Table 5: Estimated Percentage of Students Meeting ILEARN and Benchmark Proficient Standards in Spring 2019 (Year of Standard Setting)

Grade	ILEARN At Proficiency	NAEP Proficient	Smarter Proficient
ELA 3	46	41	45
ELA 4	45	41	47
ELA 5	47	41	50
ELA 6	47	41	48
ELA 7	49	41	50
ELA 8	50	41	50
Mathematics 3	58	51	47
Mathematics 4	53	48	43
Mathematics 5	47	46	36
Mathematics 6	46	43	38

Grade	ILEARN At Proficiency	NAEP Proficient	Smarter Proficient
Mathematics 7	41	41	38
Mathematics 8	37	38	37
Science 4	46	42	--
Science 6	47	39	--
Biology	39	35	--
Social Studies 5	45	--	50

Prior to the operational administration of the *ILEARN* US Government assessments in 2023-2024, a standard confirmation workshop was conducted to recommend a set of performance standards to the IDOE for reporting student performance of the Indiana Academic Standards using spring 2019 test data. Table 6 presents the percentage of U.S. Government students from spring 2019 meeting the proficient standard. Chapter 7, Performance Standards, explains the standard setting procedures in more details.

Table 6: Percentage of Students Meeting *ILEARN* Proficient Standard

Grade	ILEARN At Proficiency
U.S. Government	20

Following the operational administration of the ILEARN Science NGSS assessments in 2023-2024, a standard setting workshop was conducted to recommend a set of performance standards to the IDOE for reporting student performance of the Indiana Academic Standards. Table 7 presents the percentage of students meeting the proficient standard. Chapter 7, Performance Standards, explains the standard setting procedures in more details.

Table 7: Percentage of Students Meeting *ILEARN* and Benchmark Proficient Standards

Grade	ILEARN At or Above Proficiency	Indiana NAEP Proficient or Advanced
4	44	42
6	42	39
Biology	44	35

1.8 EVIDENCE BASED ON INTERNAL STRUCTURE

While the blueprints ensure that the full range of the intended measurement construct is represented in each test administration, tests may also inadvertently measure attributes that are not relevant to the construct of interest. For example, when a high level of English language proficiency is necessary to access content in mathematics and science items, language proficiency may unnecessarily limit the student's ability to demonstrate achievement in those subject areas. Although such tests may measure achievement of relevant mathematics and science content standards, they may also measure construct-irrelevant variation in language proficiency, limiting the universality of test score interpretations for some student populations.

Evidence based on internal structure is the degree to which the relationships among test items and test components are representative of the proposed underlying construct for test score interpretations. An analysis of the degree to which the underlying factor structure of a construct is congruent with the empirical investigations about the unidimensionality of that construct can provide evidence for the internal structure of the test.

One pathway to explore the internal structure of the test is via a second-order factor model, assuming a general construct (first factor) with reporting categories (second factor) and that the items load onto the reporting category they intend to measure. If the first-order factors are highly correlated and the model fits data well for the second-order model, this provides evidence of unidimensionality and reporting subscores.

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. This section presents results for subject area content models and correlations among relevant reporting categories. Disattenuated correlation among reporting category scores are presented in section 3.5, Subscale Intercorrelations.

Further, an analysis of the internal structure of a test can include evidence for test reliability by ensuring the test is measuring the construct consistently. Reliability analyses are discussed in detail in section 3.4 Reliability.

1.8.1 CONFIRMATORY FACTOR ANALYSIS

Indiana's *ILEARN* assessments represent a structural model of student achievement in grade-level and course-specific reporting categories. This section is based on a second-order confirmatory factor analysis (CFA), in which the first-order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto factors they are intended to measure. The underlying

structure of the *ILEARN* assessments was common across all grades, which is useful for comparing the results of our analyses across the grades.

Within each subject area (e.g., ELA), items are designed to measure a single reporting category (e.g., Key Ideas and Textual Support/Vocabulary, Structural Elements and Organization/Connection of Ideas/Media Literacy, and Writing). Reporting categories within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order confirmatory factor analyses is illustrated in Figure 1. As the figure illustrates, each item is an indicator of a reporting category. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each reporting category serves as an indicator of achievement in a subject area. As at the item level, the reporting categories include an error term indicating that they are not pure indicators of overall achievement in the subject area. The paths from the reporting categories to the items represent the first-order factor loadings, or the degree to which items are correlated with the underlying reporting category construct. Similarly, the paths from subject-area achievement to the reporting categories represent the second-order factor loading, indicating the degree to which academic reporting category constructs correlate with the underlying subject-area achievement construct.

Figure 1: Second-Order Structural Model for ILEARN Assessments
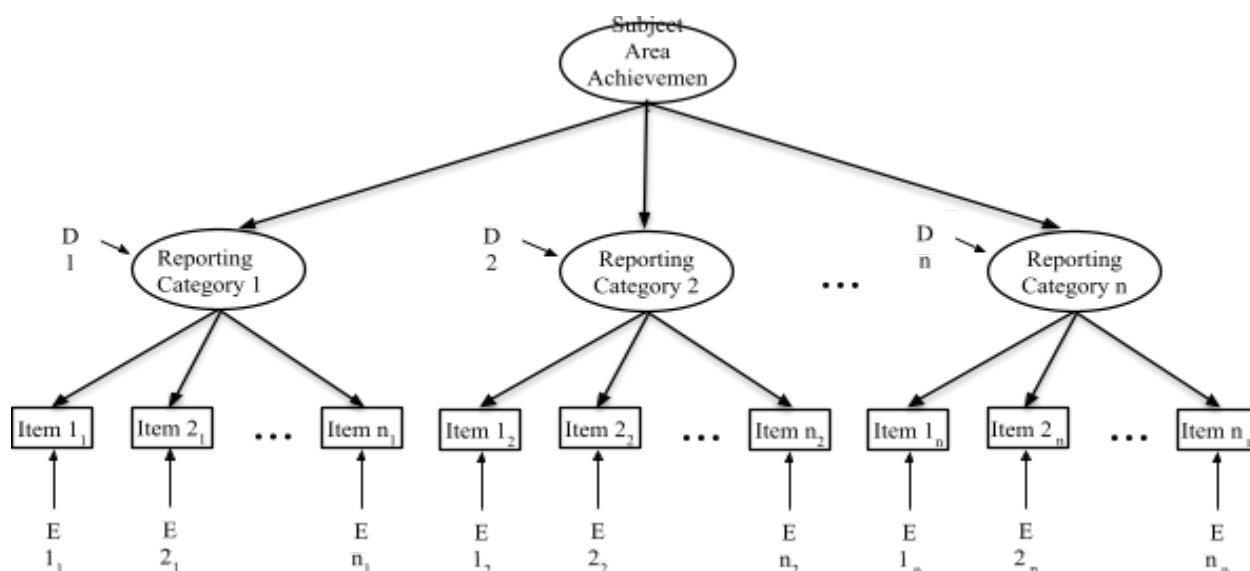
It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional Item Response Theory (IRT) model difficult. However, we could then easily justify reporting these separate scores. On the contrary, if the reporting categories were perfectly correlated, we could justify using a unidimensional model, but we could not justify reporting separate scores.

The *ILEARN* test items were designed to measure different standards and higher-level reporting categories. Test scores were reported as an overall performance measure. Additionally, scores on the various reporting categories were also provided as indices of strand-specific performance. The strand scores were reported in a fashion that aligned with the theoretical structure of the test derived from the test blueprint.

While the test consisted of items targeting different standards, all items within a grade and subject were calibrated concurrently using the various IRT models described in this technical report. This implies the pivotal IRT assumption of local independence (Lord, 1980). Formally stated, this assumption posits that the probability of the outcome on item *i* depends only on the student's ability and the characteristics of the item. Beyond that, the score of item *i* is independent of the outcome of all other items. From this assumption, the joint density (i.e., the likelihood) is viewed as the product of the individual densities. Thus, the maximum likelihood estimation of person and item parameters in traditional IRT is derived on the basis of this theory.

The results in this section were based on the data collected from the initial administration of the *ILEARN* assessments, which was the spring 2019 administration. The purpose is to provide validity evidence regarding the dimensionality of the assessments and to show that the methods for reporting *ILEARN* strand scores align with the underlying structure of the test and provide evidence for appropriateness of the selected IRT models. Given there is no major change in test design, this analysis does not need to be conducted in subsequent test administrations.

1.8.2 Factor Analytic Method

A series of Confirmatory Factor Analyses (CFAs) were conducted using the statistical program Mplus, version 7.31 (Muthén & Muthén, 2012) for each grade and subject assessment. The estimation method, weighted least squares means and variance adjusted (WLSMV), was employed because it is less sensitive to the size of the sample and the model and is also shown to perform well with categorical variables (Muthén, du Toit, & Spisic, 1997).

For each of the test forms, the goodness of fit between the structural model and the operational test data were examined. Goodness of fit is typically indexed by a $\chi 2$ statistic, with good model fit indicated by a non-significant $\chi 2$ statistic. However, the $\chi 2$ statistic is sensitive to sample size, so even well-fitting models will demonstrate highly significant $\chi 2$ statistics given a very large number of students. Therefore, fit indices, such as the Comparative Fit Index (CFI; Bentler, 1990), the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), and the Root Mean Square Error of Approximation (RMSEA) were also used to evaluate model fit. Table 8 provides a list of the goodness-of-fit statistics used to evaluate model fit, along with a guideline as to what constitutes a good fit.

Goodness-of-Fit Index	Indication of Good Fit
CFI	≥ .95
TLI	≥ .95
RMSEA	≤ .05

If the internal structure of the test was strictly unidimensional, then the overall person ability measure, theta (θ), would be the single common factor and the correlation matrix among test items would suggest no discernable pattern among factors. As such, there would be no empirical or logical basis to report scores for the separate performance categories. In factor analytic terms, a test structure that is strictly unidimensional implies a single-order factor model in which all test items load onto a single underlying factor. The following development expands the first-order model to a generalized second-order parameterization to show the relationship between the models.

The factor analysis models are based on the matrix S of tetrachoric and polychoric sample correlations among the item scores (Olsson, 1979), and the matrix W of asymptotic covariances among these sample correlations (Jöreskog, 1994) is employed as a weight matrix in a weighted least squares estimation approach (Browne, 1984; Muthén, 1984) to minimize the fit function:

$$F_{WLS} = vech(S - \hat{\Sigma})'W^{-1}vech(S - \hat{\Sigma}).$$

In this equation, $\hat{\Sigma}$ is the implied correlation matrix given the estimated factor model and the function vech vectorizes a symmetric matrix. That is, vech stacks each column of the matrix to form a vector. Note that the WLSMV approach (Muthén, du Toit, & Spisic, 1997) employs a weight matrix of asymptotic variances (i.e., the diagonal of the weight matrix) instead of the full asymptotic covariances.

We posit a first-order factor analysis where all test items load onto a single common factor as the base model. The first-order model can be mathematically represented as

$$\hat{\Sigma} = \Lambda\Phi\Lambda' + \Theta,$$

where Λ is the matrix of item factor loadings (with Λ' representing its transpose), and Θ is the uniqueness, or measurement error. The matrix Φ is the correlation among the separate factors. For the base model, items are thought only to load onto a single underlying factor. Hence Λ' is a *p x 1* vector, where *p* is the number of test items and Φ is a scalar equal to 1. Therefore, it is possible to drop the matrix Φ from the general notation. However, this notation is retained to more easily facilitate comparisons to the implied model, such that it can subsequently be viewed as a special case of the second-order factor analysis.

For the implied model, we posit a second-order factor analysis in which test items are coerced to load onto the reporting categories they are designed to target, and all reporting categories share a common underlying factor. The second-order factor analysis can be mathematically represented as

$$\hat{\Sigma} = \Lambda(\Gamma\Phi\Gamma' + \Psi)\Lambda' + \Theta,$$

where $\hat{\Sigma}$ is the implied correlation matrix among test items, Λ is the *p x k* matrix of first-order factor loadings relating item scores to first-order factors, Γ is the *k x 1* matrix of second-order factor loadings relating the first-order factors to the second-order factor with *k* denoting the number of factors, Φ is the correlation matrix of the second-order factors, and Ψ is the matrix of first-order factor residuals. All other notation is the same as the first-order model. Note that the second-order model expands the first-order model such that $\Phi\rightarrow\Gamma\Phi\Gamma' + \Psi$. As such, the first-order model is said to be nested within the second-order model.

There is a separate factor for each reporting category for ELA, mathematics, science, and social studies. Therefore, the number of rows in Γ (*k*) differed among subjects, but the general structure of the factor analysis was consistent.

1.8.3 ELA CONTENT MODEL

The goodness-of-fit statistics for the hypothesized *ILEARN* second-order models in ELA are shown in Table 9. All the statistics indicate that the second-order models posited by the *ILEARN* assessments fit the data well. This pattern was true across all grades. The CFI and TLI values are all equal to or greater than .98. The RMSEA values are all 0.01, well below the values used to indicate good fit.

Table 9: Goodness-of-Fit for the ILEARN ELA Second-Order Models

Grade	df	RMSEA	CFI	TLI	Convergence
Second-Order Models					
3	524	0.014	0.983	0.981	Yes
4	557	0.014	0.983	0.982	Yes
5	591	0.009	0.984	0.983	Yes
6	492	0.014	0.984	0.983	Yes
7	460	0.012	0.982	0.981	Yes
8	557	0.010	0.985	0.984	Yes

The estimated correlations between the reporting categories from the second-order factor model for ELA are shown in Table 10. Although the correlations are high, the

results provide empirical evidence that there is some detectable dimensionality among the reporting categories.

Table 10: Correlations Among ELA Factors

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10	0.997	1	
	Writing (Cat3)	9	0.792	0.790	1
4	Key Ideas and Textual Support/Vocabulary (Cat1)	13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11	0.975	1	
	Writing (Cat3)	9	0.714	0.732	1
5	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11	0.972	1	
	Writing (Cat3)	9	0.816	0.793	1
6	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10	0.985	1	
	Writing (Cat3)	9	0.780	0.792	1
7	Key Ideas and Textual Support/Vocabulary (Cat1)	10	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11	0.977	1	
	Writing (Cat3)	8	0.876	0.879	1
8	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10	0.924	1	
	Writing (Cat3)	9	0.807	0.746	1

1.8.4 MATHEMATICS CONTENT MODEL

The goodness-of-fit statistics for the strand-based second-order models in mathematics are shown in Table 11. The models generally show good fit although the CFI and TLI fit indices are less than the cutoff value of 0.95 for grades 6 and 8. Even for these grades, however, the RMSEA estimates are well below their respective 0.05 cutoff values. All of the statistics indicate the second-order models are a good fit for the data.

Table 11: Goodness-of-Fit for the ILEARN Mathematics Second-Order Models

Grade	df	RMSEA	CFI	TLI	Convergence
		Second-Order Models			
3	1076	0.017	0.983	0.982	Yes
4	1076	0.014	0.958	0.955	Yes
5	1076	0.015	0.977	0.976	Yes
6	1075	0.019	0.942	0.939	Yes
7	1075	0.013	0.983	0.982	Yes
8	1075	0.025	0.916	0.912	Yes

The estimated correlations between the reporting categories from the second-order factor model for mathematics can be seen in Table 12. Although the correlations are high, the results provide empirical evidence that there is some detectable dimensionality among the reporting categories.

Table 12: Correlations Among Mathematics Factors

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	9	1			
	Computation (Cat2)	13	0.989	1		
	Geometry and Measurement (Cat3)	10	0.969	0.959	1	
	Number Sense (Cat4)	11	0.908	0.898	0.880	1
4	Algebraic Thinking and Data Analysis (Cat1)	9	1			
	Computation (Cat2)	12	0.963	1		
	Geometry and Measurement (Cat3)	10	0.929	0.894	1	
	Number Sense (Cat4)	12	0.934	0.900	0.868	1
5	Algebraic Thinking (Cat1)	11	1			
	Computation (Cat2)	11	0.888	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.890	0.790	1	
	Number Sense (Cat4)	11	0.926	0.823	0.825	1
6	Algebra and Functions (Cat1)	11	1			
	Computation (Cat2)	11	0.820	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.763	0.645	1	
	Number Sense (Cat4)	11	0.973	0.823	0.766	1
7	Algebra and Functions (Cat1)	11	1			

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Data Analysis, Statistics, and Probability (Cat2)	10	0.865	1		
	Geometry and Measurement (Cat3)	10	0.891	0.859	1	
	Number Sense and Computation (Cat4)	11	0.912	0.880	0.906	1
8	Algebra and Functions (Cat1)	11	1			
	Data Analysis, Statistics, and Probability (Cat2)	10	0.748	1		
	Geometry and Measurement (Cat3)	12	0.821	0.712	1	
	Number Sense and Computation (Cat4)	10	0.815	0.707	0.775	1

1.8.5 SOCIAL STUDIES CONTENT MODEL

The goodness-of-fit statistics for the strand-based second-order models in social studies are shown in Table 13. All the statistics indicate that the second-order models posited by the *ILEARN* assessments fit the data well. This pattern was true across both grades. The CFI and TLI values are equal to or greater than .97. The RMSEA values are well below the values used to indicate good fit.

Table 13: Goodness-of-Fit for the ILEARN Social Studies Second-Order Models

Grade	df	RMSEA	CFI	TLI	Convergence
Second-Order Models					
5	699	0.020	0.977	0.975	Yes
U.S. Government	1322	0.015	0.986	0.986	Yes

The estimated correlations between the reporting categories from the second-order factor model for social studies can be seen in Table 14. Although the correlations are high, the results provide empirical evidence that there is some detectable dimensionality among the reporting categories.

Table 14: Correlations Among Social Studies Factors

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	16	1		
	Geography and Economics (Cat2)	11	0.982	1	
	History (Cat3)	12	0.947	0.950	1
U.S. Government	Functions of Government (Cat1)	19	1		
	Historical Foundations of American Government (Cat2)	14	0.962	1	

Institutions and Processes of Government (Cat3)	20	0.957	0.971	1

In all scenarios, the empirical results suggest that the implied model fits the data well. That is, these results indicate that reporting an overall score in addition to separate scores for the individual reporting categories is reasonable, as the intercorrelations among items suggest that there are detectable distinctions among reporting categories.

Clearly, the correlations among the separate factors are high, which is reasonable. This again provides support for the measurement model, given that the calibration of all items is performed concurrently. If the correlations among factors were very low, this could possibly suggest that a different IRT model would be needed (e.g., multidimensional IRT) or that the IRT calibration should be performed separately for items measuring different factors. The high correlations among the factors suggest that these alternative methods are unnecessary and that the current approach is in fact preferable.

Overall, these results provide empirical evidence and justification for the use of the chosen scoring and reporting methods. Additionally, the results provide justification for the current IRT model employed.

1.9 EVIDENCE OF SCIENCE NGSS BANK PARAMETER STABILITY

1.9.1 BACKGROUND

Science items of the shared science NGSS MOU bank are calibrated concurrently for all states that participated in NGSS MOU field testing using a multigroup Rasch testlet model. The testlets correspond to assertion-based items (e.g. stand alone or clusters) and the testlet variance accounts for local dependencies between the assertions pertaining to the same item. During calibration, overall differences across groups (state and grade combinations) are modeled by estimating group-specific mean and variance parameters for the ability distribution.

CAI NGSS technical team have compared the parameters of several state specific calibrations to the multigroup calibrations to assess parameter stability over time. For NGSS MOU joint multigroup calibration, field test data from 14 states and one US territory with an embedded field test design were used. Most field test items were administered in two or three states and the multigroup concurrent calibration of field items was conducted by anchoring on the bank values of the operational items. The result indicated correlations between bank and state-specific parameters were high, ranging from .976 to .998 across grades, states, and years.

To evaluate the stability of Indiana parameters, CAI conducted a single group calibration using Indiana spring 2024 field test data and compared the parameters from the state specific calibration with MOU bank parameters from multigroup calibration. The MOU field test items that were administered in Indiana were calibrated using concurrent

calibration method by anchoring on the MOU bank values of the operational items that were administered adaptively in Indiana.

1.9.2 RESULT

The correlations and median absolute differences between Indiana specific values and MOU bank values are presented in Table X. The result showed that for both cluster and standalone items, correlations are exceptionally high and the differences in parameter values are minimal between the Indiana specific parameters and MOU bank parameters across grades. Among three grades, all correlations are above 0.99 and all median absolute differences are below 0.1. The result is consistent with previous findings in other NGSS MOU states. Between cluster item and standalone items, standalone items demonstrated slightly higher correlations and slightly lower median absolute difference in parameter values. However, the pattern is highly consistent between cluster and standalone items.

Table 15. Summary of Comparison

Grade	Cluster			Standalone		
	N	Correlation	Median absolute difference	N	Correlation	Median absolute difference
4	20	0.997	0.060	23	0.997	0.051
6	22	0.994	0.067	23	0.999	0.041
Bio	37	0.995	0.063	35	0.997	0.041

Additionally, CAI identified assertions that showed absolute difference in parameter value greater than 0.3. As the Figure X-X showed, overall speaking, the flagged assertions are the more difficult ones in the pool due to the portion of high-performing students that can answer the assertion correctly is generally small and therefore the standard error of the data used in the state-specific calibration for those assertions could be relatively large. It was also observed that there are more assertions flagged in Biology than in lower grades because Biology pool is relatively more difficult than the other two lower grades. Those flagged assertions and associated items were reviewed by CAI technical and content team for potential flaws. None of assertions or items have been identified as problematic from psychometric and content perspectives.

Figure 2. Scatter Plot of Grade 4 Science MOU items
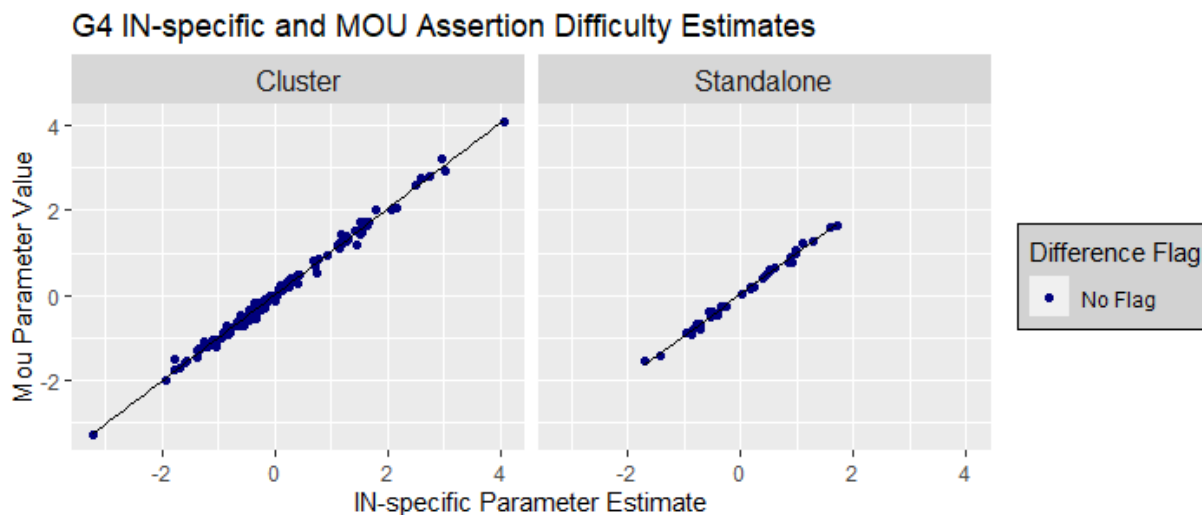
Figure 3. Scatter Plot of Grade 6 Science MOU items
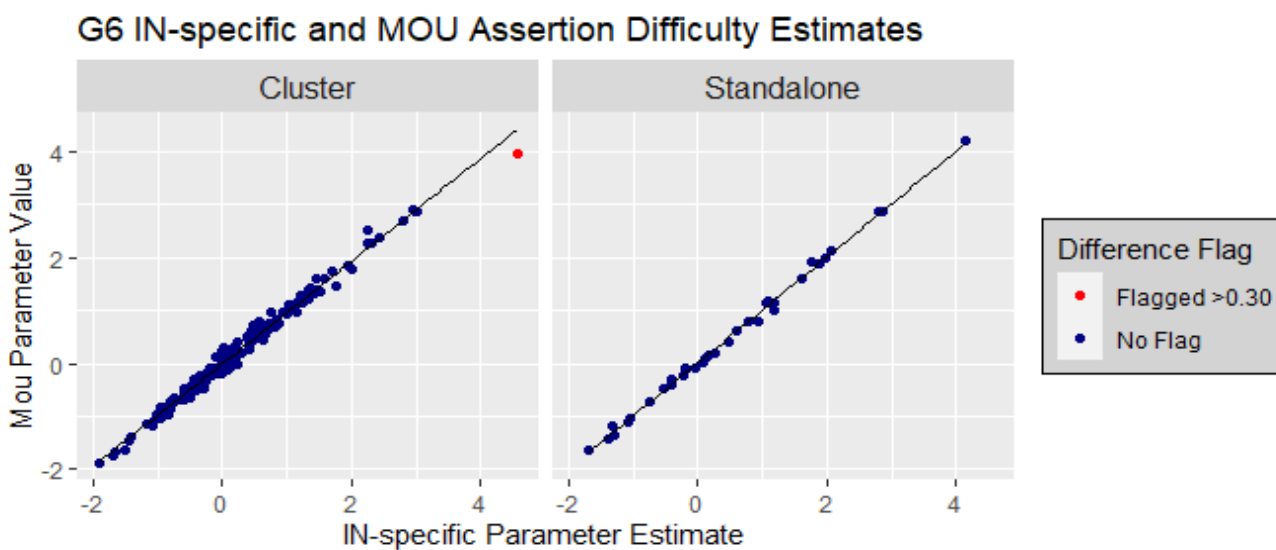
Figure 4. Scatter Plot of Biology MOU items

Bio IN-specific and MOU Assertion Difficulty Estimates
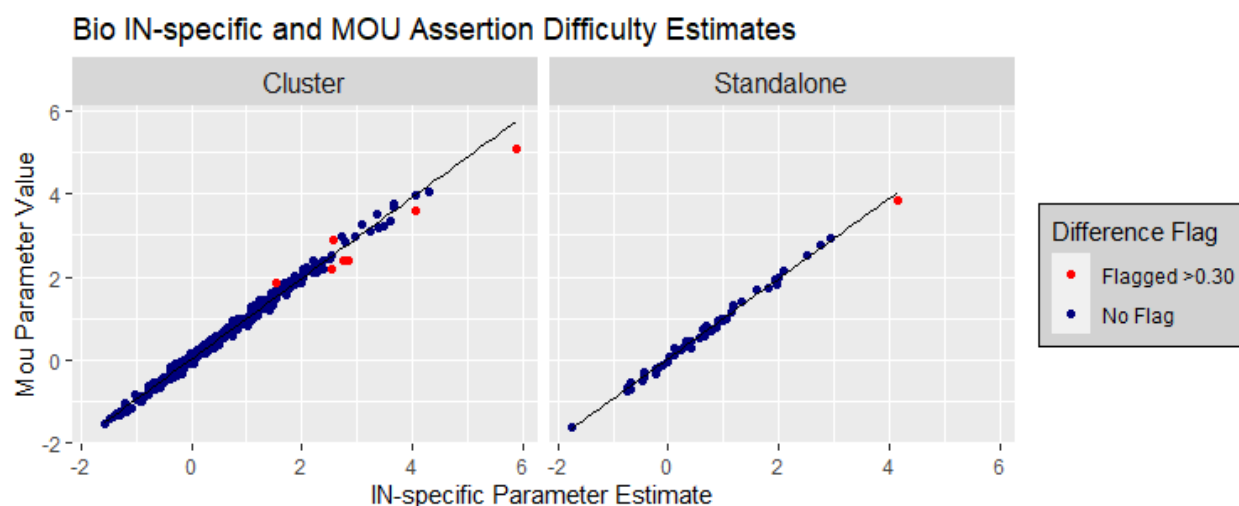
1.10 EVIDENCE OF CONVERGENT AND DISCRIMINANT VALIDITY

According to Standard 1.14 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), it is necessary to provide evidence of convergent and discriminant validity to support test score relationships with external variables. Convergent evidence supports the relationship between measures assessing the same construct while discriminant evidence distinguishes the test from other measures assessing different constructs. Since independent tests measuring the same constructs as ELA and mathematics were not available for Indiana only the correlations between subscores within and across tests were examined. The a priori expectation is that subscores within the same subject (e.g., ELA) will correlate more positively than subscore correlations across subjects (e.g., ELA and mathematics). These correlations are based on a small number of items, typically around 8 to 18; consequently, the observed score correlations are expected to be smaller in magnitude as a result of the very large measurement error at the subscore level. For this reason, both the observed score and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within subjects and across subjects for grades 3–8 ELA and mathematics using 2021–2022 spring administration data, except for grades 4 and 6 which used 2023-2024 spring administration data as science tests at grades 4 and 6 has become brand new NGSS tests starting 2023-2024. In grades 4 and 6, science was included and in grade 5, social studies was included. Table 16 through Table 27 show the observed and disattenuated score correlations among ELA, mathematics, science, and social studies subscores. In general, the pattern is consistent with the a priori expectation that subscores within a test correlate more highly than correlations between tests measuring different constructs.

Table 16: Grade 3 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1						
	Structural Elements and Organization/ Connection of Ideas/Media Literacy (Cat2)	0.69	1					
	Writing (Cat3)	0.63	0.59	1				
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.65	0.61	0.63	1			
	Computation (Cat2)	0.64	0.61	0.62	0.81	1		
	Geometry and Measurement (Cat3)	0.63	0.59	0.61	0.79	0.77	1	
	Number Sense (Cat4)	0.64	0.60	0.62	0.79	0.76	0.78	1

Table 17: Grade 3 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat 1	Cat 2	Cat 3	Cat1	Cat 2	Cat 3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1						
	Structural Elements and Organization/ Connection of Ideas/Media Literacy (Cat2)	0.98	1					
	Writing (Cat3)	0.89	0.86	1				
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.83	0.82	0.83	1			
	Computation (Cat2)	0.86	0.84	0.85	1.00*	1		
	Geometry and Measurement (Cat3)	0.82	0.80	0.81	0.97	0.97	1	
	Number Sense (Cat4)	0.82	0.80	0.81	0.95	0.95	0.95	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

Table 18: Grade 4 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.65	1									
	Writing (Cat3)	0.65	0.6	1								
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.62	0.59	0.63	1							
	Computation (Cat2)	0.6	0.57	0.62	0.75	1						
	Geometry and Measurement (Cat3)	0.6	0.57	0.61	0.72	0.73	1					
	Number Sense (Cat4)	0.6	0.57	0.62	0.75	0.76	0.74	1				
Science	Physical Science (Cat1)	0.64	0.62	0.61	0.64	0.63	0.63	0.64	1			
	Life Science (Cat2)	0.61	0.59	0.57	0.59	0.58	0.58	0.59	0.67	1		
	Earth and Space Science (Cat3)	0.61	0.59	0.58	0.61	0.6	0.61	0.61	0.68	0.64	1	
	Computer Science (Cat4)	0.56	0.53	0.55	0.57	0.56	0.56	0.56	0.59	0.55	0.56	1

Table 19: Grade 4 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.93	1									
	Writing (Cat3)	0.89	0.85	1								
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.84	0.82	0.85	1							
	Computation (Cat2)	0.79	0.77	0.81	0.96	1						
	Geometry and Measurement (Cat3)	0.80	0.78	0.81	0.94	0.92	1					
	Number Sense (Cat4)	0.79	0.78	0.80	0.97	0.95	0.94	1				
Science	Physical Science (Cat1)	0.87	0.88	0.82	0.85	0.81	0.83	0.83	1			
	Life Science (Cat2)	0.85	0.86	0.80	0.82	0.78	0.79	0.79	0.93	1		
	Earth and Space Science (Cat3)	0.87	0.88	0.82	0.86	0.82	0.85	0.84	0.96	0.94	1	
	Computer Science (Cat4)	0.79	0.78	0.77	0.79	0.75	0.76	0.76	0.82	0.8	0.82	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

Table 20: Grade 5 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Social Studies		
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1									
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.62	1								
	Writing (Cat3)	0.68	0.59	1							
Mathematics	Algebra and Functions (Cat1)	0.65	0.57	0.68	1						
	Computation (Cat2)	0.61	0.54	0.65	0.78	1					
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.61	0.53	0.64	0.76	0.76	1				
	Number Sense (Cat4)	0.61	0.54	0.63	0.76	0.75	0.73	1			
Social Studies	Civics and Government (Cat1)	0.63	0.58	0.61	0.61	0.58	0.59	0.59	1		
	Geography and Economics (Cat2)	0.58	0.53	0.56	0.59	0.56	0.56	0.57	0.67	1	
	History (Cat3)	0.62	0.56	0.59	0.60	0.57	0.58	0.58	0.71	0.65	1

<p align="center" style="color:blue">Table 21: Grade 5 Disattenuated Score Correlations</p>

Subject	Reporting Category	ELA			Mathematics				Social Studies		
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1									
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.94	1								
	Writing (Cat3)	0.91	0.86	1							
Mathematics	Algebra and Functions (Cat1)	0.84	0.81	0.85	1						
	Computation (Cat2)	0.80	0.78	0.83	0.98	1					
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.82	0.79	0.83	0.97	0.97	1				
	Number Sense (Cat4)	0.81	0.79	0.81	0.97	0.96	0.94	1			
Social Studies	Civics and Government (Cat1)	0.87	0.87	0.81	0.79	0.77	0.79	0.78	1		
	Geography and Economics (Cat2)	0.88	0.88	0.82	0.85	0.82	0.84	0.84	1.00*	1	
	History (Cat3)	0.88	0.88	0.81	0.80	0.78	0.80	0.79	1.00*	1.00*	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

Table 22: Grade 6 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.65	1									
	Writing (Cat3)	0.63	0.64	1								
Mathematics	Algebra and Functions (Cat1)	0.62	0.63	0.67	1							
	Computation (Cat2)	0.52	0.53	0.58	0.74	1						
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.58	0.58	0.62	0.77	0.67	1					
	Number Sense (Cat4)	0.61	0.61	0.65	0.81	0.7	0.74	1				
Science	Physical Science (Cat1)	0.54	0.55	0.54	0.61	0.52	0.57	0.59	1			
	Life Science (Cat2)	0.6	0.6	0.58	0.65	0.55	0.61	0.64	0.64	1		
	Earth and Space Science (Cat3)	0.49	0.5	0.48	0.58	0.5	0.55	0.57	0.55	0.6	1	
	Computer Science (Cat4)	0.56	0.56	0.57	0.6	0.5	0.57	0.6	0.53	0.58	0.48	1

Table 23: Grade 6 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.92	1									
	Writing (Cat3)	0.86	0.85	1								
Mathematics	Algebra and Functions (Cat1)	0.82	0.81	0.83	1							
	Computation (Cat2)	0.71	0.71	0.74	0.91	1						
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.79	0.78	0.8	0.95	0.85	1					
	Number Sense (Cat4)	0.82	0.80	0.82	0.98	0.88	0.94	1				
Science	Physical Science (Cat1)	0.85	0.85	0.8	0.86	0.76	0.84	0.86	1			
	Life Science (Cat2)	0.86	0.85	0.79	0.85	0.75	0.83	0.86	0.99	1		
	Earth and Space Science (Cat3)	0.80	0.80	0.75	0.86	0.76	0.84	0.86	0.98	0.97	1	
	Computer Science (Cat4)	0.83	0.81	0.79	0.81	0.69	0.80	0.82	0.85	0.85	0.80	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

Table 24: Grade 7 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.65	1					
	Writing (Cat3)	0.69	0.62	1				
Mathematics	Algebra and Functions (Cat1)	0.64	0.58	0.65	1			
	Data Analysis, Statistics, and Probability (Cat2)	0.64	0.58	0.64	0.74	1		
	Geometry and Measurement (Cat3)	0.53	0.49	0.54	0.66	0.64	1	
	Number Sense and Computation (Cat4)	0.64	0.58	0.64	0.78	0.75	0.67	1

Table 25: Grade 7 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.96	1					
	Writing (Cat3)	0.92	0.89	1				
Mathematics	Algebra and Functions (Cat1)	0.84	0.83	0.84	1			
	Data Analysis, Statistics, and Probability (Cat2)	0.85	0.84	0.83	0.95	1		
	Geometry and Measurement (Cat3)	0.75	0.74	0.74	0.88	0.87	1	
	Number Sense and Computation (Cat4)	0.82	0.81	0.81	0.98	0.95	0.89	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

Table 26: Grade 8 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.65	1					
	Writing (Cat3)	0.71	0.61	1				
Mathematics	Algebra and Functions (Cat1)	0.64	0.56	0.65	1			
	Data Analysis, Statistics, and Probability (Cat2)	0.62	0.55	0.63	0.75	1		
	Geometry and Measurement (Cat3)	0.58	0.51	0.60	0.73	0.72	1	
	Number Sense and Computation (Cat4)	0.54	0.48	0.56	0.69	0.67	0.68	1

Table 27: Grade 8 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat 1	Cat 2	Cat 3	Cat 1	Cat 2	Cat 3	Cat 4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.93	1					
	Writing (Cat3)	0.92	0.88	1				
Mathematics	Algebra and Functions (Cat1)	0.81	0.79	0.83	1			
	Data Analysis, Statistics, and Probability (Cat2)	0.80	0.78	0.82	0.95	1		
	Geometry and Measurement (Cat3)	0.76	0.74	0.78	0.93	0.93	1	
	Number Sense and Computation (Cat4)	0.73	0.72	0.76	0.91	0.89	0.91	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

1.11 Evidence Related to Cognitive Processes

1.11.1 ELA and Mathematics

Cognitive labs investigating claims about the cognitive processes students use to respond to test items, and other questions concerning interactions with test items, were conducted by Smarter Balanced and reported in their cognitive laboratories technical report (2013). Since the majority of ILEARN items come from the Smarter Balanced, results from these cognitive lab studies can be applied to ILEARN. Among the many research questions addressed in these studies, several were relevant to the DOK level elicited by items across item types.

For example, one study examined whether students who achieved full credit on multi-part selected-response (MPSR) items demonstrated, through their think-aloud sessions, greater understanding than those students who did not achieve full credit. In addition, this study examined whether students who received full credit on MPSR items demonstrated a depth of understanding similar to that of students receiving full credit on similarly challenging constructed-response (CR) items measuring the same target. With respect to the first hypothesis, students receiving full credit on the MPSR items demonstrated a greater understanding of the material than those who did not obtain full credit. With respect to the second hypothesis, results indicated that in most cases, the DOK demonstrated by the students receiving full credit on the MPSR items either equaled or exceeded the DOK demonstrated by students achieving full credit on the matched CR items.

The cognitive labs were also designed to assess whether different types of technology-enhanced (TE) items elicited DOK levels comparable to CR items matched for specific content claim/targets and DOK levels. Selected-response (SR) items were also included, where available, as a comparison item format.

With respect to ELA items, students demonstrated a higher DOK level for most of the TE item types rather than for the matched CR items, but with some exceptions. A similar pattern was observed for the matched SR items versus the CR items. Evidence for mathematics items was mixed, with some TE and SR item types showing evidence for greater DOK than matched CR items, while other CR items indicated greater DOK than the matched TE and SR items.

These cognitive lab studies also addressed questions concerning student use of online tools, such as the equation editor for mathematics items, indicating, for example, that some students across grade levels did have difficulty responding using the equation editor, but that grade 3 students, in particular, had greater difficulty than students in other grades. Studies also inquired whether accessibility tools improved student access to test content, finding, for example, that while text-to-speech (TTS) always improved access to ELA test content, especially for English language learners (ELLs) and

students with an Individualized Education Program (IEP), that in mathematics, access improved for students in grade 3 only.

1.11.2 SCIENCE

In 2017, when the development of item clusters for the MOU states began, cognitive lab studies were conducted to evaluate and refine the process of developing item clusters aligned to the NGSS. Results of the cognitive lab studies confirmed the feasibility of the approach used. Item clusters were completed within 12 minutes on average, and students reported being familiar with the format conventions and online tools used in the item clusters. They appeared to easily navigate the item clusters' interactive features and response formats. In general, students who received credit on a given item displayed a reasoning process that aligned with the skills that the item was intended to measure.

A second set of cognitive lab studies was conducted by CAI for Connecticut in 2018 and 2019 to determine if students using braille can understand the task demands of selected accommodated three-dimensional science standards-aligned item clusters and navigate the interactive features of these clusters in a manner that allows them to fully display their knowledge and skills relative to the constructs of interest. In general, both the students who relied entirely on braille and/or the Job Access with Speech (JAWS) screen-reading software and those who had some vision and were able to read the screen with magnification were able to find the information they needed to respond to the questions, navigate the various response formats, and finish within a reasonable amount of time. The item clusters were clearly different from (and more complex than) other tests with which the students were familiar, however, and the study recommended that students should be given adequate time to practice with at least one sample cluster before taking the summative test. The study also resulted in tool-specific recommendations for accessibility for visually impaired students. The reports of both sets of cognitive lab studies are presented in Appendix 2-A, Science Clusters Cognitive Lab Report, and Appendix 2-B, Braille Cognitive Lab Report.

1.12 EVIDENCE OF FAIRNESS AND ACCESSIBILITY

1.12.1 FAIRNESS IN CONTENT

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

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items

- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

Content experts receive extensive training on the principles of universal design and apply these principles in the development of all test materials. In the review process, adherence to the principles of universal design is verified.

1.12.2 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is insufficient for determining the fairness of a test. Rather, it must be accompanied by statistical processes. While a variety of item statistics were reviewed during form building to evaluate the quality of items, one notable statistic used was differential item functioning (DIF). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe evidence of DIF. Furthermore, items were categorized positively (i.e., +A, +B, +C), signifying that the item favored the focal group (e.g., African American/Black, Hispanic, Female), or negatively (i.e., –A, –B, –C), signifying that the item favored the reference group (e.g., White, Male). Items across all groups were flagged if their DIF statistics indicated the "C" category. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Items were reviewed by the Bias and Sensitivity Committee regardless of whether the DIF statistic favored the focal group or the reference group. The details about how these items were reviewed for bias is further described in Chapter 4, Item Development and Test Construction.

DIF analyses were conducted for all items to detect potential item bias from a statistical perspective across major ethnic and gender groups. These DIF analyses were performed for the following groups:

- Male/Female
- White/African American
- White/Hispanic
- White/Asian
- White/Native American
- Student with Special Education (SPED)/Not SPED
- SES/Not SES (proxy for Free and Reduced Lunch)
- English Learners (ELs)/Not ELs

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).

1.13 SUMMARY OF VALIDITY OF TEST SCORE INTERPRETATIONS

Evidence for the validity of test score interpretations is strengthened as evidence supporting test score interpretations accrues. In this sense, the process of seeking and evaluating evidence for the validity of test score interpretation is ongoing. Nevertheless, sufficient evidence exists to support the principal claims for the test scores, including that *ILEARN* test scores indicate the degree to which students have achieved the Indiana Academic Standards at each grade level and that students scoring at the Proficient level or higher demonstrate levels of achievement consistent with national benchmarks that indicate they are on track for college readiness. These claims are supported by evidence of a test development process that ensures alignment of test content to the Indiana Academic Standards and evidence that the structural model described by the Indiana Academic Standards and implemented in the *ILEARN* assessments is sound.

3. SUMMARY OF THE SUMMATIVE TEST ADMINISTRATION

ILEARN is an online, adaptive assessment for ELA, mathematics, and science and an online, fixed-form assessment for social studies. All online adaptive assessments make use of technology-enhanced item types. Students unable to participate in the online administrations have the option to use a paper-pencil form. Students participating in the computer-based *ILEARN* test can use standard online testing features in the Test Delivery System (TDS), which include a selection of font colors and sizes and the ability to zoom in and out and highlight text. In addition to the resources available to all students, *ILEARN* provides accommodated forms for braille and Spanish. Students with disabilities can take *ILEARN* with or without accommodations, or they can take the Indiana's Alternate Measure (*I AM*) assessment. Visually impaired students can take the braille version of *ILEARN* ELA, mathematics, science, and social studies. English learners (ELs) can take the Spanish language version of *ILEARN* mathematics, science, and social studies. During test development, CAI ensured that scores obtained on the alternative modes of administrations were comparable to those received on the standard online tests, which adhered to the same blueprints. Post-administration checks were also performed, and no concerns were found in the 2023–2024 administration.

The following tests were available in the 2023–2024 administration:

- ELA grades 3–8
- Mathematics grades 3–8
- Science grades 4, 6, and biology
- Social studies grade 5 and U.S. government

1.14 STUDENT POPULATION AND PARTICIPATION

All Indiana public and nonpublic school students in ELA and mathematics grades 3–8; science grades 4 and 6, and students taking the biology End-of-Course (EOC) assessment; and social studies grade 5, are required to participate in the state assessments. U.S. government is an optional EOC assessment. Table 28 shows the number of students tested and the number of students reported for the 2023–2024 *ILEARN* assessments. The numbers of students tested and reported for historical administrations (e.g., 2018–2019, 2020–2021, 2021–2022,etc.) are also provided to show the trend in student participation. As can be observed in this table, for all census assessments, the number of students participating in *ILEARN* decreased from 2018–2019 to 2020–2021, which is expected due to the pandemic. However, in the post-pandemic era, student participation increased in 2021–2022 and following years as expected. Across all grades, the student participations have been overall consistent between 2022-2023 to 2023-2024. For U.S. government, because it is an optional test, student participation decreased from the initial administration (i.e., 2018–2019) to subsequent administrations (i.e., 2020–2021, 2021–2022, etc.). It is important to note that participation based on enrollment is high (i.e., 97-99%) in the post pandemic years. Decrease in the number tested and reported is due to lower enrollment. Please note

that starting 2023-2024, ILEARN science assessment has become a brand new NGSS science assessment, therefore only spring 2024 statistics of science tests are presented in this chapter.

Table 28: Number of Students Participating in ILEARN

		G3	G4	G5	G6	G7	G8
ELA							
SP24	Number Tested	81792	82913	81266	82591	82191	83027
	Number Reported	81777	82896	81244	82558	82152	83001
SP23	Number Tested	82170	80480	81869	81640	82383	83617
	Number Reported	82145	80453	81840	81611	82341	83571
SP22	Number Tested	79953	81034	81136	82218	83391	85047
	Number Reported	79915	81003	81102	82180	83346	84990
SP21	Number Tested	79431	78998	80341	81683	83102	82717
	Number Reported	79389	78970	80286	81601	83030	82614
SP19	Number Tested	83096	84175	86407	85880	84669	83079
	Number Reported	83074	84147	86381	85833	84591	82991
Mathematics							
		G3	G4	G5	G6	G7	G8
SP24	Number Tested	81775	82883	81231	82572	82163	83011
	Number Reported	81738	82854	81205	82519	82064	82917
SP23	Number Tested	82175	80493	81860	81643	82424	83647
	Number Reported	82126	80452	81821	81571	82301	83524
SP22	Number Tested	79967	81028	81133	82230	83426	85073
	Number Reported	79940	80990	81080	82102	83262	84897
SP21	Number Tested	79359	78978	80311	81686	83065	82719
	Number Reported	79319	78907	80222	81547	82883	82546
SP19	Number Tested	83111	84183	86420	85895	84692	83066
	Number Reported	83080	84144	86369	85817	84580	82991
Science							
		G4	G6	Biology (Fall)	Biology (Winter)	Biology (Spring)	
SP24	Number Tested	82782	82426	--	1842	80847	
	Number Reported	82743	81849	--	1827	80154	
Social Studies							
		G5	U.S. Government				
SP24	Number Tested	81119	233				
	Number Reported	81101	231				
SP23	Number Tested	81721	323				

		Number Reported	81708	322					
SP22		Number Tested	80963	279					
		Number Reported	80939	278					
SP21		Number Tested	79870	645					
		Number Reported	79831	641					
SP19		Number Tested	86274	1245					
		Number Reported	86253	1230					

Table 29 through Table 32 present the distribution of students of subgroups in percentages. The subgroup categories reported are gender, ethnicity, students with special education (SPED) status, students with Section 504 Plans, English Learners (ELs), and Socioeconomic Status (SES). The percentage of participation by subgroup seems to be largely consistent from 2018–2019 to 2023–2024, with slightly increased representation of students from disadvantaged subgroups, such as SES.

Table 29: Distribution of Demographic Characteristics of Tested Population, ELA

Grade	Year	N	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	EL*	SES*
G3	SP24	81792	51.05	48.95	62.83	12.88	3.26	14.96	0.17	0.09	5.81	19.25	2.78	11.44	54.02
	SP23	82170	51.25	48.75	63.90	12.81	3.20	13.99	0.13	0.10	5.87	18.59	2.67	N/A	54.57
	SP22	79953	51.06	48.94	64.82	12.48	3.13	13.74	0.15	0.09	5.59	17.58	2.25	10.13	49.73
	SP21	79431	51.18	48.82	65.53	12.43	3.20	13.10	0.17	0.10	5.46	16.88	1.97	10.14	50.40
	SP19	83096	51.28	48.72	65.90	12.62	2.76	13.06	0.15	0.09	5.42	16.56	2.24	9.47	N/A
G4	SP24	82913	51.18	48.82	63.05	12.88	3.31	14.77	0.13	0.10	5.76	18.80	3.27	11.4	53.32
	SP23	80480	51.08	48.92	64.39	12.45	3.22	14.12	0.15	0.10	5.57	18.54	3.12	N/A	53.92
	SP22	81034	51.14	48.86	64.91	12.56	3.26	13.55	0.16	0.11	5.44	17.13	2.50	10.25	49.17
	SP21	78998	50.89	49.11	65.58	12.24	2.94	13.46	0.17	0.10	5.51	17.07	2.48	9.65	49.87
	SP19	84175	50.84	49.16	66.11	12.48	2.62	13.34	0.16	0.07	5.21	16.32	2.62	8.93	N/A
G5	SP24	81266	51.09	48.91	63.53	12.45	3.34	14.94	0.15	0.09	5.5	18.24	3.64	9.72	52.68
	SP23	81869	51.20	48.80	64.41	12.71	3.32	13.87	0.18	0.12	5.40	17.29	3.33	N/A	53.11
	SP22	81136	50.99	49.01	64.92	12.45	2.98	13.89	0.17	0.10	5.49	16.68	3.05	8.48	48.82
	SP21	80341	51.20	48.80	65.40	12.43	2.91	13.72	0.14	0.10	5.30	16.83	2.57	7.88	50.04
	SP19	86407	50.86	49.14	66.29	12.56	2.47	13.31	0.16	0.08	5.14	16.06	2.70	6.70	N/A

Grade	Year	N	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	EL*	SES*
G6	SP24	82591	51.22	48.78	63.67	12.68	3.43	14.64	0.17	0.12	5.29	16.52	3.84	8.64	51.89
	SP23	81640	51.03	48.97	64.40	12.57	3.03	14.27	0.17	0.10	5.46	16.61	3.67	N/A	52.65
	SP22	82218	51.19	48.81	64.79	12.58	2.98	14.09	0.14	0.10	5.32	16.19	3.07	7.07	48.77
	SP21	81683	50.96	49.04	65.73	12.19	2.73	13.98	0.15	0.08	5.15	15.83	2.93	6.65	49.07
	SP19	85880	50.95	49.05	66.91	12.25	2.29	13.34	0.17	0.08	4.97	15.14	2.94	4.33	N/A
G7	SP24	82191	50.93	49.07	63.77	12.55	3.06	14.94	0.17	0.10	5.41	15.98	4.00	8.50	51.00
	SP23	82383	51.16	48.84	64.36	12.56	3.07	14.45	0.15	0.11	5.31	15.92	3.78	N/A	52.19
	SP22	83391	50.97	49.03	65.05	12.47	2.76	14.36	0.16	0.09	5.10	15.15	3.41	6.92	47.77
	SP21	83102	50.94	49.06	66.07	12.21	2.61	13.89	0.15	0.08	5.00	15.34	2.88	6.19	48.03
	SP19	84669	51.17	48.83	67.50	12.10	2.46	12.80	0.19	0.08	4.86	14.70	2.65	3.53	N/A
G8	SP24	83027	51.12	48.88	63.66	12.59	3.13	15.10	0.14	0.11	5.26	15.45	4.11	8.46	50.41
	SP23	83617	50.93	49.07	64.55	12.53	2.83	14.74	0.17	0.09	5.09	15.16	3.91	N/A	51.27
	SP22	85047	50.89	49.11	65.39	12.34	2.65	14.32	0.15	0.10	5.05	14.91	3.50	6.44	46.68
	SP21	82717	50.94	49.06	66.71	12.05	2.42	13.75	0.16	0.08	4.83	14.77	3.05	4.84	46.31
	SP19	83079	51.10	48.90	68.58	11.77	2.29	12.38	0.19	0.09	4.69	14.55	2.70	3.37	N/A

* EL is not available in the spring 2023 data.
*SES was not available in the spring 2019 data.

Table 30: Distribution of Demographic Characteristics of Tested Population, Mathematics

Grade	Year	N	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	EL*	SES*
G3	SP24	81775	51.04	48.96	62.84	12.87	3.27	14.94	0.17	0.09	5.82	19.28	2.81	11.44	54.16
	SP23	82175	51.25	48.75	63.88	12.80	3.20	14.02	0.13	0.10	5.87	18.60	2.69	N/A	54.59
	SP22	79967	51.07	48.93	64.82	12.47	3.13	13.75	0.16	0.09	5.59	17.60	10.15	2.35	49.74
	SP21	79359	51.19	48.81	65.53	12.43	3.21	13.09	0.17	0.10	5.46	16.85	10.15	1.97	50.37
	SP19	83111	51.27	48.73	65.89	12.62	2.76	13.07	0.15	0.09	5.42	16.57	2.24	9.48	N/A
G4	SP24	82883	51.18	48.82	63.06	12.88	3.31	14.76	0.13	0.1	5.76	18.8	3.31	11.41	53.43
	SP23	80493	51.08	48.92	64.38	12.45	3.22	14.13	0.15	0.10	5.57	18.54	3.14	N/A	53.96
	SP22	81028	51.13	48.87	64.91	12.56	3.26	13.55	0.16	0.11	5.44	17.14	10.24	2.62	49.18
	SP21	78978	50.91	49.09	65.58	12.24	2.94	13.47	0.17	0.10	5.50	17.07	9.67	2.48	49.92
	SP19	84183	50.83	49.17	66.11	12.47	2.62	13.35	0.16	0.07	5.21	16.35	2.61	8.95	N/A

Grade	Year	*N*	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	EL*	SES*
G5	SP24	81231	51.09	48.91	63.53	12.45	3.34	14.94	0.15	0.09	5.5	18.25	3.68	9.72	52.79
	SP23	81860	51.20	48.80	64.41	12.70	3.32	13.88	0.18	0.12	5.39	17.30	3.34	N/A	53.14
	SP22	81133	50.99	49.01	64.91	12.45	2.98	13.90	0.17	0.10	5.50	16.69	8.48	3.11	48.82
	SP21	80311	51.19	48.81	65.42	12.42	2.91	13.71	0.14	0.10	5.30	16.83	7.89	2.58	50.02
	SP19	86420	50.86	49.14	66.27	12.56	2.47	13.32	0.16	0.08	5.14	16.07	2.70	6.72	N/A
G6	SP24	82572	51.22	48.78	63.66	12.68	3.43	14.64	0.17	0.12	5.29	16.53	3.87	8.66	51.96
	SP23	81643	51.04	48.96	64.39	12.57	3.03	14.28	0.17	0.10	5.46	16.61	3.67	N/A	52.67
	SP22	82230	51.20	48.80	64.79	12.59	2.98	14.09	0.14	0.10	5.32	16.21	7.07	3.14	48.80
	SP21	81686	50.98	49.02	65.72	12.19	2.73	13.98	0.15	0.08	5.15	15.80	6.64	2.93	49.10
	SP19	85895	50.95	49.05	66.90	12.24	2.29	13.36	0.17	0.08	4.97	15.18	2.94	4.35	N/A
G7	SP24	82163	50.93	49.07	63.78	12.53	3.06	14.94	0.17	0.1	5.41	15.99	4.02	8.51	51.08
	SP23	82424	51.18	48.82	64.33	12.55	3.07	14.49	0.15	0.10	5.31	15.92	3.78	N/A	52.22
	SP22	83426	50.98	49.02	65.03	12.48	2.76	14.36	0.16	0.09	5.11	15.17	6.93	3.54	47.79
	SP21	83065	50.94	49.06	66.04	12.21	2.61	13.90	0.15	0.08	5.01	15.30	6.20	2.87	48.09
	SP19	84692	51.17	48.83	67.50	12.08	2.46	12.82	0.19	0.08	4.87	14.71	2.64	3.55	N/A
G8	SP24	83011	51.13	48.87	63.66	12.59	3.13	15.11	0.14	0.11	5.25	15.46	4.14	8.47	50.49
	SP23	83647	50.95	49.05	64.54	12.53	2.83	14.76	0.17	0.09	5.08	15.16	3.90	N/A	51.30
	SP22	85073	50.89	49.11	65.39	12.35	2.65	14.32	0.15	0.10	5.05	14.90	6.45	3.59	46.71
	SP21	82719	50.93	49.07	66.69	12.05	2.42	13.77	0.16	0.08	4.82	14.75	4.83	3.05	46.33
	SP19	83066	51.11	48.89	68.58	11.74	2.29	12.41	0.20	0.09	4.70	14.52	2.70	3.39	N/A

* EL is not available in the spring 2023 data.
*SES was not available in the spring 2019 data.

Table 31: Distribution of Demographic Characteristics of Tested Population, Science

Grade	Year	*N*	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	ELL*	SES*
G4	SP24	82782	51.17	48.83	63.1	12.85	3.31	14.77	0.13	0.1	5.75	18.82	3.33	11.39	53.47
G6	SP24	82426	51.2	48.8	63.69	12.65	3.44	14.64	0.17	0.12	5.29	16.52	3.87	8.66	52.02

Grade	Year	*N*	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	ELL*	SES*
Biology (Fall)	SP24	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Biology (Winter)	SP24	1842	51.3	48.7	68.02	14.39	1.47	11.4	0.11	0.05	4.56	14.06	3.15	5.7	43.65
Biology (Spring)	SP24	80847	50.72	49.28	63.72	12.28	3.16	15.68	0.15	0.1	4.91	13.57	4.28	8.38	47.86

* EL is not available in the spring 2023 data.
*SES was not available in the spring 2019 data.

Table 32: Distribution of Demographic Characteristics of Tested Population, Social Studies

Grade	Year	*N*	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	ELL*	SES*
G5	SP24	81119	51.08	48.92	63.55	12.43	3.34	14.93	0.15	0.09	5.5	18.25	3.68	9.72	52.81
	SP23	81721	51.19	48.81	64.43	12.67	3.32	13.88	0.18	0.12	5.40	17.30	3.36	N/A	53.16
	SP22	80963	50.96	49.04	64.96	12.42	2.98	13.89	0.17	0.10	5.49	16.68	8.48	3.17	48.83
	SP21	79870	51.20	48.80	65.51	12.33	2.91	13.72	0.14	0.10	5.29	16.82	7.90	2.59	49.99
	SP19	86274	50.84	49.16	66.33	12.51	2.47	13.31	0.16	0.08	5.14	16.09	2.71	6.71	N/A
U.S. Government	SP24	233	55.79	44.21	66.95	20.6	1.29	6.01	0	0	5.15	18.45	2.15	1.29	36.91
	SP23	323	47.06	52.94	60.37	28.79	2.17	4.02	0.00	0.00	4.64	21.05	3.41	N/A	55.42
	SP22	279	49.82	50.18	65.23	21.15	1.43	9.32	0.36	0.00	2.51	20.43	3.23	5.73	43.01
	SP21	645	53.33	46.67	84.65	4.19	1.24	5.58	0.16	0.16	4.03	11.78	1.24	1.55	29.77
	SP19	1245	55.42	44.58	68.11	14.38	1.04	12.85	0.24	0.00	3.37	12.93	1.20	3.29	N/A

* EL is not available in the spring 2023 data.
*SES was not available in the spring 2019 data.

1.15 SUMMARY OF OVERALL STUDENT PERFORMANCE

The 2023–2024 state summary results for the average scale scores and the percentage of students in each proficiency level by grade and content area are presented in Tables 32–35. In terms of both average scale scores and percentages at or above proficiency, there is a drastic decline in student performance from 2018–2019 to 2020–2021, which was expected due to the impact of the pandemic, followed by an increasing trend beginning 2021–2022 as a result of the post-pandemic recovery. In 2022–2023, the increasing trend seems to continue for math. For ELA, however, there seems to be a mix, with slight increases for Grades 6 and 8 and slight decreases for Grades 3, 4, 5, and 7. In 2023–2024, overall speaking, student performance seems to be comparable with 2022-2023. For ELA, however, there seems to be a mix, with slight decreases for Grade 3 and 8 and slight increases for the rest of the grades. For Math and Social Studies, all grades showed slight decreases except for Grade 7 and Grade 8 of Math. For Science, the 2023-2024 assessment is a completely new and different assessment, therefore the student performance should not be compared directly with previous administrations and only spring 2024 statistics of science assessment is presented in this chapter. It is also important to note that changes in performance are likely confounded by shifts in state demographics as shown in Table 33 through Table 36.

Table 33: Percentage of Students in Proficiency Levels, ELA

Grade	Admin	Number Tested	Scale Score Mean	Scale Score SD	% Below Proficiency	% Approaching Proficiency	% At Proficiency	% Above Proficiency	% At or Above Proficiency
G3	SP24	81777	5435.39	75.22	39.80	21.60	23.44	15.16	38.60
	SP23	82145	5436.96	76.47	39.36	20.97	23.40	16.27	39.67
	SP22	79915	5439.07	74.77	38.50	20.83	24.12	16.54	40.66
	SP21	79389	5435.89	74.19	40.13	21.14	23.41	15.32	38.73
	SP19	83074	5449.74	69.13	31.05	23.16	27.91	17.88	45.79
G4	SP24	82896	5472.02	82.79	36.04	22.18	22.80	18.98	41.78
	SP23	80453	5470.31	82.31	37.48	22.19	21.94	18.39	40.33
	SP22	81003	5473.04	83.07	36.89	21.97	21.57	19.56	41.13
	SP21	78970	5469.02	81.58	37.99	22.46	21.81	17.75	39.56
	SP19	84147	5481.24	75.49	30.53	24.14	25.62	19.70	45.32
G5	SP24	81244	5498.24	86.08	38.01	21.68	26.98	13.34	40.31
	SP23	81840	5498.34	85.78	38.19	21.62	26.72	13.46	40.18
	SP22	81102	5499.60	84.22	36.59	22.44	27.79	13.17	40.96
	SP21	80286	5497.66	81.55	37.14	23.34	27.74	11.77	39.51
	SP19	86381	5513.26	79.85	29.00	23.96	31.83	15.21	47.04
G6	SP24	82558	5520.9	83.61	35.91	22.96	24.45	16.68	41.12
	SP23	81611	5520.87	81.43	35.79	23.39	24.64	16.17	40.81
	SP22	82180	5517.12	80.85	37.04	23.91	24.51	14.53	39.04
	SP21	81601	5520.43	77.80	34.67	25.46	25.62	14.24	39.86
	SP19	85833	5534.31	73.36	27.05	25.64	29.81	17.51	47.32
G7	SP24	82152	5546	85.62	32.13	26.07	24.82	16.97	41.79

Grade	Admin	Number Tested	Scale Score Mean	Scale Score SD	% Below Proficiency	% Approaching Proficiency	% At Proficiency	% Above Proficiency	% At or Above Proficiency
	SP23	82341	5541.42	84.65	33.47	27.04	24.18	15.32	39.50
	SP22	83346	5546.30	85.85	31.24	26.33	25.38	17.04	42.42
	SP21	83030	5543.52	84.33	32.27	26.66	25.53	15.55	41.08
	SP19	84591	5559.97	82.16	24.69	26.24	28.83	20.24	49.07
G8	SP24	83001	5553.19	92.97	32.58	24.80	23.30	19.32	42.62
	SP23	83571	5558.08	85.79	29.07	27.09	25.54	18.31	43.85
	SP22	84990	5557.31	85.86	29.59	27.31	25.03	18.06	43.09
	SP21	82614	5559.57	84.70	28.69	27.46	24.97	18.88	43.85
	SP19	82991	5572.88	79.23	21.21	28.68	28.64	21.47	50.11

Table 34: Percentage of Students in Proficiency Levels, Mathematics

Grade	Admin	Number Tested	Scale Score Mean	Scale Score SD	% Below Proficiency	% Approaching Proficiency	% At Proficiency	% Above Proficiency	% At or Above Proficiency
	SP24	81738	6426.44	83.72	29.29	18.30	28.42	23.98	52.41
	SP23	82126	6427.64	84.31	28.80	18.12	28.59	24.48	53.07
G3	SP22	79940	6425.07	82.87	29.55	18.57	28.92	22.97	51.89
	SP21	79319	6419.01	83.35	32.15	19.15	27.81	20.89	48.70
	SP19	83080	6437.16	75.70	23.18	18.74	32.63	25.45	58.08
	SP24	82854	6464.33	86.13	32.72	19.32	28.97	18.99	47.96
	SP23	80452	6466.29	84.57	31.75	19.50	29.37	19.38	48.75
G4	SP22	80990	6464.18	83.35	32.41	20.05	29.23	18.31	47.54
	SP21	78907	6456.50	83.41	35.96	20.30	27.57	16.17	43.74
	SP19	84144	6476.73	77.78	25.80	20.75	32.84	20.62	53.46
	SP24	81205	6485.52	87.40	35.53	23.66	21.95	18.85	40.80
	SP23	81821	6486.07	87.68	35.15	23.74	21.88	19.22	41.10
G5	SP22	81080	6483.05	89.16	35.25	23.92	22.78	18.04	40.82
	SP21	80222	6479.32	86.78	36.79	24.61	22.49	16.11	38.60
	SP19	86369	6501.15	84.83	27.32	25.34	25.26	22.08	47.34
	SP24	82519	6508.19	101.98	39.35	22.56	21.22	16.87	38.09
	SP23	81571	6508.78	101.74	39.65	22.06	20.92	17.38	38.30
G6	SP22	82102	6503.45	99.33	41.42	23.06	20.41	15.11	35.52
	SP21	81547	6499.54	95.63	43.04	23.95	19.71	13.30	33.01
	SP19	85817	6527.18	93.34	30.29	23.93	25.59	20.19	45.78
	SP24	82064	6514.76	104.7	40.74	25.41	18.33	15.52	33.85
	SP23	82301	6513.30	102.90	41.39	25.75	18.27	14.58	32.85
G7	SP22	83262	6512.11	100.42	41.47	26.62	18.17	13.74	31.91
	SP21	82883	6512.78	95.92	40.63	28.89	17.96	12.52	30.48
	SP19	84580	6535.57	97.61	31.94	26.68	22.94	18.44	41.38
	SP24	82917	6529.86	114.09	44.54	24.08	15.93	15.45	31.38
	SP23	83524	6529.75	113.59	44.59	24.03	15.79	15.59	31.38
G8	SP22	84897	6526.30	111.82	45.22	25.01	15.36	14.40	29.76
	SP21	82546	6523.36	107.31	46.89	25.35	14.80	12.96	27.76
	SP19	82991	6550.37	108.11	34.77	27.83	19.11	18.30	37.41

Table 35: Percentage of Students in Proficiency Levels, Science

Grade	Admin	Number Tested	Scale Score Mean	Scale Score SD	% Below Proficiency	% Approaching Proficiency	% At Proficiency	% Above Proficiency	% At or Above Proficiency
G4	SP24	82743	160.03	15.99	18.88	37.19	32.12	11.82	43.93
G6	SP24	81849	360	16.18	21.07	36.82	19.08	23.03	42.11
Biology (Fall)	SP24	--	--	--	--	--	--	--	--
Biology (Winter)	SP24	1827	559.33	15.7	18.88	39.41	22.06	19.65	41.71
Biology (Spring)	SP24	80154	559.88	16.16	19.10	37.23	22.65	21.02	43.66

* Science tests of Spring 2024 are new NGSS tests

Table 36: Percentage of Students in Proficiency Levels, Social Studies

Grade	Admin	Number Tested	Scale Score Mean	Scale Score SD	% Below Proficiency	% Approaching Proficiency	% At Proficiency	% Above Proficiency	% At or Above Proficiency
G5	SP24	81101	8490.36	52.51	44.38	17.54	21.67	16.42	38.08
G5	SP23	81708	8491.59	53.44	43.44	17.42	21.88	17.27	39.15
G5	SP22	80939	8490.24	53.28	44.37	17.26	21.68	16.69	38.37
G5	SP21	79831	8490.41	53.31	43.42	17.97	21.97	16.64	38.61
G5	SP19	86253	8500.82	54.94	36.45	17.96	24.11	21.49	45.60
U.S. Government	SP24	231	8446.73	52.31	84.42	--	15.58	--	--
U.S. Government	SP23	322	8450.99	50.29	80.75	--	19.25	--	19.25
U.S. Government	SP22	278	8447.94	53.33	80.58	--	19.42	--	19.42
U.S. Government	SP21	641	8470.65	53.55	68.02	--	31.98	--	31.98
U.S. Government	SP19	1230	8449.44	51.55	79.92	--	20.08	--	20.08

1.16 STUDENT PERFORMANCE BY SUBGROUP

The 2023–2024 state summary results for the average scale scores and the percentage of students in each proficiency level by grade and by content area were calculated for several subcategories—including female, male, White, African American, Asian, Hispanic/Latino, American Indian/Alaskan, Native Hawaiian/Pacific Islander, Multi-Racial, special education (SPED), section 504 plan, EL, and SES.

Distribution of scale scores by subgroups along with historical statistics are presented in Appendix 3-A, Distribution of Scale Scores and Standard Deviations. Percentage of students in performance levels for overall and by subgroup along with historical statistics are presented in Appendix 3-B, Percentage of Students in Performance Levels for Overall and by Subgroup. In addition, the summary of scale scores by subgroup for each reporting category along with historical statistics are provided in Appendix 3-C, Distribution of Reporting Category Scores by Subgroup.

Figure 5 through Figure 8 display the average scale scores, overall and by subgroup, for the 2023–2024 administration as well as for historical administrations. As shown in the figures, average scale scores decreased drastically in 2020–2021 comparing to the pre-pandemic 2018–2019 test administration, both statewide and across all subgroups. Starting from the 2021–2022 administration, by and large, average scale scores seem to be increasing for all subgroups as a result of the post-pandemic recovery. Please note that subgroups with the size smaller than 200 were suppressed from the graphs.

Figure 5: ELA Average Scale Score by Subgroup
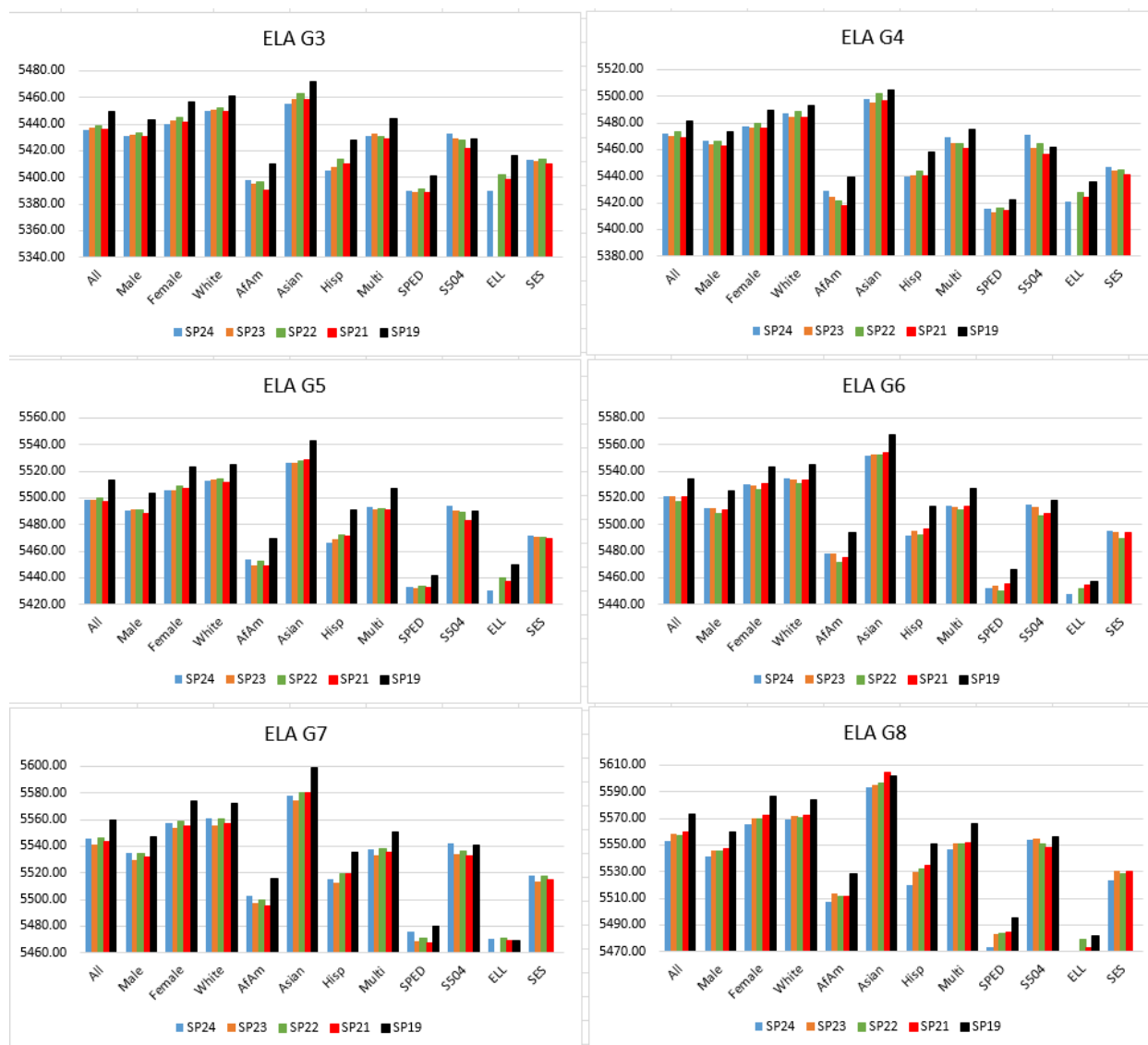
* EL is not available in the spring 2023 data.

*SES was not available in the spring 2019 data.

Figure 6: Mathematics Average Scale Score by Subgroup
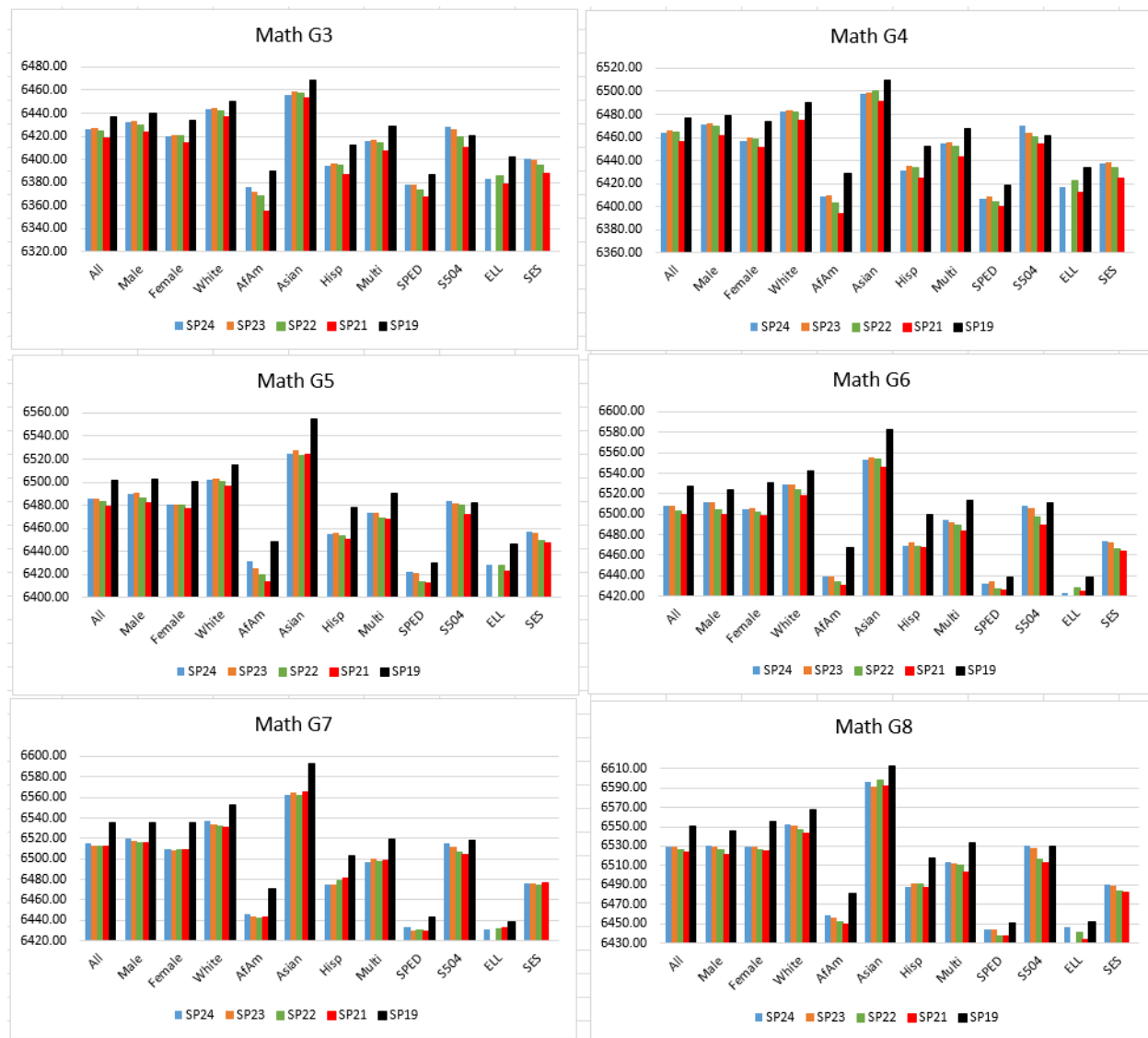
* EL is not available in the spring 2023 data.
*SES was not available in the spring 2019 data.

Figure 7: Science Average Scale Score by Subgroup

Science G4

Science G6

Bio
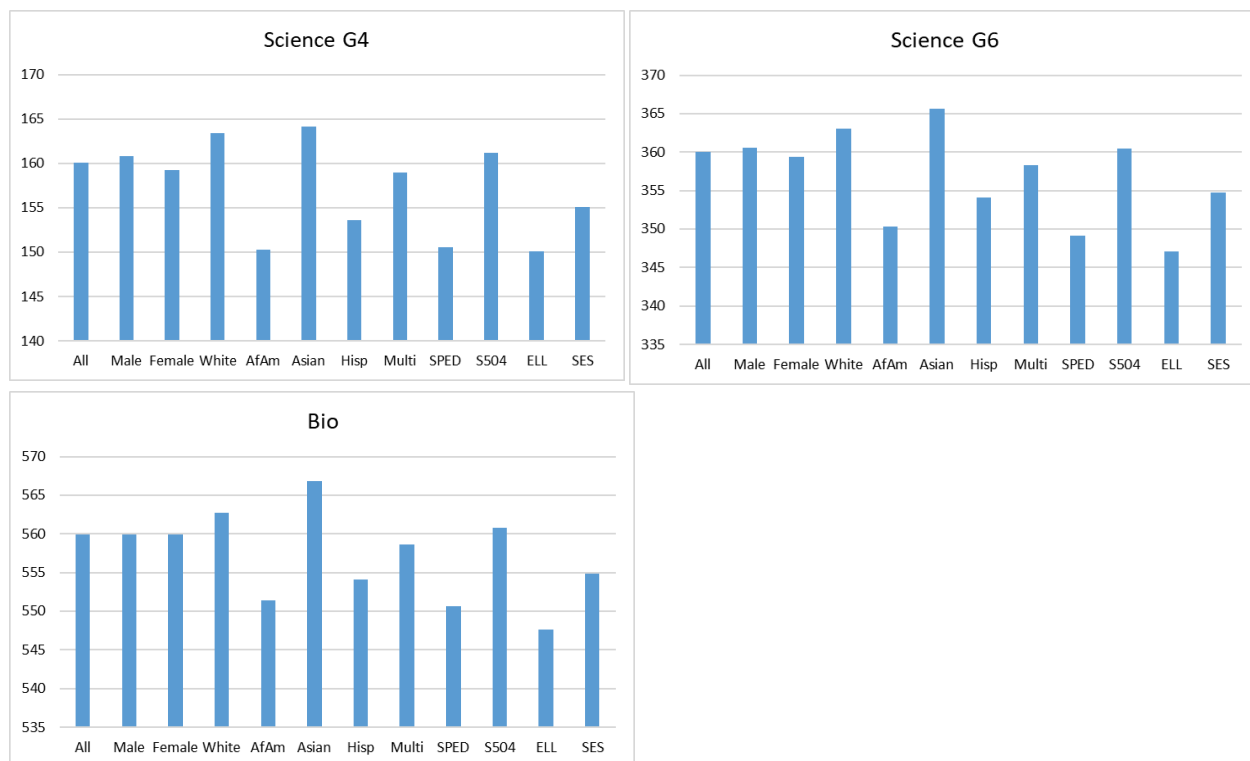
Figure 8: Social Studies Average Scale Score by Subgroup

Social Studies G5
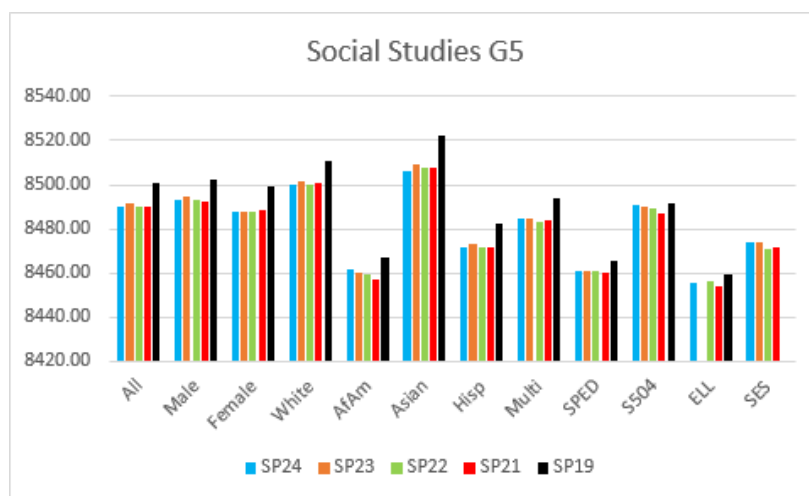
* EL is not available in the spring 2023 data.
*SES was not available in the spring 2019 data.

1.17 RELIABILITY

Test score reliability is traditionally estimated using both classical and item response theory (IRT) approaches. Classical estimates of test reliability, such as Cronbach's alpha, provide an index of the internal consistency reliability of the test or the likelihood that a student would achieve the same score in an equivalently constructed test form. While classical indicators provide a single estimate of the reliability of test forms, the precision of test scores varies with respect to the information value of the test at each location. For example, most fixed-form assessments target test information near important cut scores or near the population mean so that test scores are most precise in targeted locations. Because adaptive tests target test information near each student's ability level, the precision of test scores may increase, especially for lower- and higher-ability students. The precision of individual test scores is critically important to valid test score interpretation and is provided along with test scores as part of all student-level reporting.

1.17.1 MARGINAL RELIABILITY

While measurement error is conditional on test information, it is nevertheless desirable to provide a single index of a test's internal consistency reliability. Such an index is provided by the marginal reliability coefficient, which considers 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, which are estimated at different points on the ability scale for all students. The marginal reliability coefficients are nearly identical or close to the coefficient *alpha*.

The marginal reliability $(\bar{p})$ 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 standard error of measurement of the scaled score for student i, and σ^2 is the variance of the scaled score. The higher the reliability coefficient, the greater the precision of the test.

Table 37 through * Science tests of Spring 2024 are new NGSS tests

Table 40 present the marginal reliability coefficients and the average standard error of measurements for the total scale scores for the 2023–2024 administration as well as for historical administrations. Results show that marginal reliabilities are in high .80s and .90s for all subjects and grades. Within a subject and grade, reliabilities seem to be consistent across administrations. However, for science, the 2023-2024 NGSS science tests differ from the science tests administered in prior years, in terms of item types and

score scale. Therefore, the reliabilities of the 2023-2024 science tests are slightly lower than previous administrations due to the clustering nature of the items and shorter test length.

Table 37: Marginal Reliability for ELA

Grade	Admin	Marginal Reliability	N	Mean	SD	SEM
3	SP24	0.905	81777	5435.39	75.22	22.90
	SP23	0.906	82145	5436.97	76.47	23.20
	SP22	0.895	79915	5439.07	74.77	23.80
	SP21	0.892	79389	5435.89	74.19	23.95
	SP19	0.872	83074	5449.74	69.13	24.14
4	SP24	0.895	82896	5472.02	82.79	26.50
	SP23	0.894	80453	5470.31	82.31	26.45
	SP22	0.899	81003	5473.04	83.07	26.14
	SP21	0.898	78970	5469.02	81.58	25.71
	SP19	0.880	84147	5481.24	75.49	25.75
5	SP24	0.902	81244	5498.24	86.08	26.69
	SP23	0.902	81840	5498.34	85.78	26.67
	SP22	0.896	81102	5499.60	84.22	26.92
	SP21	0.890	80286	5497.66	81.55	26.78
	SP19	0.878	86381	5513.26	79.85	27.51
6	SP24	0.898	82558	5520.90	83.61	26.37
	SP23	0.895	81611	5520.87	81.43	26.04
	SP22	0.889	82180	5517.12	80.85	26.52
	SP21	0.887	81601	5520.43	77.80	25.69
	SP19	0.881	85833	5534.31	73.36	24.62
7	SP24	0.895	82152	5546.01	85.62	27.30
	SP23	0.894	82341	5541.42	84.65	27.09
	SP22	0.894	83346	5546.30	85.85	27.56
	SP21	0.889	83030	5543.52	84.33	27.65
	SP19	0.880	84591	5559.97	82.16	27.92
8	SP24	0.906	83001	5553.20	92.97	28.14
	SP23	0.900	83571	5558.08	85.79	26.77
	SP22	0.902	84990	5557.31	85.86	26.37
	SP21	0.901	82614	5559.57	84.70	26.20
	SP19	0.879	82991	5572.88	79.23	27.11

Table 38: Marginal Reliability for Mathematics

Grade	Admin	Marginal Reliability	N	Mean	SD	SEM
3	SP24	0.958	81738	6426.44	83.72	16.87
	SP23	0.956	82126	6427.64	84.31	17.32
	SP22	0.960	79940	6425.07	82.88	16.24
	SP21	0.961	79319	6419.01	83.35	16.24
	SP19	0.943	83080	6437.16	75.70	17.62
4	SP24	0.956	82854	6464.33	86.13	17.71
	SP23	0.955	80452	6466.29	84.57	17.62
	SP22	0.955	80990	6464.18	83.35	17.34
	SP21	0.956	78907	6456.50	83.41	17.24

	SP19	0.944	84144	6476.73	77.78	18.10
5	SP24	0.951	81205	6485.52	87.40	19.10
	SP23	0.951	81821	6486.07	87.68	19.10
	SP22	0.952	81080	6483.05	89.16	18.97
	SP21	0.950	80222	6479.32	86.78	18.86
	SP19	0.938	86369	6501.15	84.83	20.40
6	SP24	0.957	82519	6508.19	101.98	20.69
	SP23	0.953	81571	6508.78	101.74	21.39
	SP22	0.951	82102	6503.46	99.33	21.19
	SP21	0.948	81547	6499.54	95.63	21.14
	SP19	0.947	85817	6527.18	93.34	20.93
7	SP24	0.952	82064	6514.76	104.70	22.11
	SP23	0.950	82301	6513.30	102.90	22.12
	SP22	0.945	83262	6512.11	100.42	22.66
	SP21	0.944	82883	6512.78	95.92	21.97
	SP19	0.934	84580	6535.57	97.61	23.53
8	SP24	0.945	82917	6529.86	114.09	26.18
	SP23	0.944	83524	6529.75	113.59	26.42
	SP22	0.944	84897	6526.30	111.82	25.86
	SP21	0.942	82546	6523.36	107.31	25.37
	SP19	0.940	82991	6550.37	108.11	25.76

<div align="center">

Table 39: Marginal Reliability for Science

</div>

Grade	Admin	Marginal Reliability	*N*	Mean	SD	SEM
4	SP24	0.894	82743	160.03	15.99	5.17
6	SP24	0.866	81849	360.00	16.18	5.91
Biology (Fall)	SP24	-	-	-	-	-
Biology (Winter)	SP24	0.852	1827	559.33	15.70	6.00
Biology (Spring)	SP24	0.861	80154	559.88	16.16	5.99

* Science tests of Spring 2024 are new NGSS tests

<div align="center">

Table 40: Marginal Reliability for Social Studies

</div>

Grade	Admin	Marginal Reliability	*N*	Mean	SD	SEM
5	SP24	0.875	81101	8490.36	52.51	18.24
	SP23	0.883	81708	8491.59	53.44	17.85
	SP22	0.871	80939	8490.24	53.28	18.75
	SP21	0.864	79831	8490.41	53.31	19.34
	SP19	0.874	86253	8500.82	54.94	19.00
US Government	SP24	0.830	231	8446.73	52.31	21.26
	SP23	0.875	322	8450.99	50.29	17.53
	SP22	0.885	278	8447.94	53.33	17.85
	SP21	0.899	641	8470.65	53.55	16.82
	SP19	0.880	1230	8449.44	51.55	17.58

1.17.2 STANDARD ERROR OF MEASUREMENT

Within the Item Response Theory (IRT) framework, measurement error varies across the range of abilities. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the test information function (TIF) represents the Standard Error of Measurement (SEM). The SEM is equal to the inverse square root of information. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, unlike students with more extreme scores. Conversely, measurement error is minimal for the part of the underlying scale at the middle of the test distribution and greater on scaled values farther away from the middle.

Within the IRT framework, measurement error varies across the range of abilities as a result of the test, providing varied information across the range of abilities as displayed by the TIF. The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. For instance, if the measurement error is large, then less information is being provided by the assessment at the specific ability level.

Figure 9 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. Where the curve is lower at the tails indicates that the test provides less information about test takers at the tails relative to the center.

Computing these TIFs is useful for evaluating where the test is maximally informative. In IRT, the TIF is based on the estimates of the item parameters in the test, and the formula used for the *ILEARN* assessment is calculated as:

$$TIF\left(\theta_s\right) = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left(\frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} Da_i\left(\theta_s - b_{il}\right)\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i\left(\theta_s - b_{il}\right)\right)} - \left(\frac{\sum_{h=1}^{m_i} h\,exp\left(\sum_{l=1}^{h} Da_i\left(\theta_s - b_{il}\right)\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i\left(\theta_s - b_{il}\right)\right)} \right)^2 \right) + \sum_{i=1}^{N_{2PL}} D^2 a_i^2 \left(\frac{q_i}{p_i}\left[p_i\right]^2 \right),$$

where N_{GPCM} is the number of items that are scored using Generalized Partial Credit Model (GPC) items, N_{2PL} is the number of items scored using the two-parameter logistic (2PL) model, *i* indicates item *i* ($i \in \{1, 2, . . . , N\}$), m_i is the maximum possible score of the item, *s* indicates student *s*, and θ_s is the ability of student *s*.

Figure 9: Sample Test Information Function
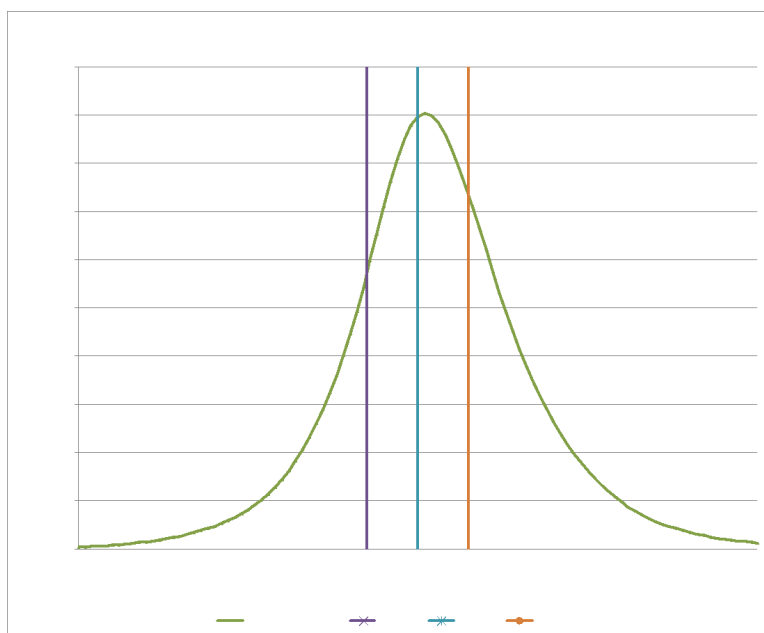
The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

$$se\left(\theta_s\right) = \frac{1}{\sqrt{TIF\left(\theta_s\right)}}.$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the SEMs are more useful for score interpretation. The magnitude of the conditional standard errors can be evaluated at the cut scores. For tests administered adaptively, we can evaluate whether the algorithm selected items appropriately to match a student's ability given the current item pool and identify the areas with a shortage of items.

Theoretically, with an infinitely large item bank comprising sufficient items to assess the range of achievement within all benchmarks and a perfect match-to-ability for each item presented, standard error of measurement (SEM) curves would be flat along the score range—an indication that all students are measured with the same precision. However, this is not practical because the real-world item pools are limited in size, especially in the early years of the computer-adaptive test (CAT) administrations. Thus, the SEM will be larger at locations characterized by relatively few items, typically at either end of the distribution where comprehensive sets of easy or difficult items are lacking. To improve measurement precision for adaptive assessments, items that measure the range of blueprint elements across the range of abilities are desirable. Nevertheless, because items targeting information near the population mean will be most frequently administered, it remains important to ensure sufficient items of normative difficulty to avoid overexposing items.

Table 41 through * Science tests of Spring 2024 are new NGSS tests

Table 44 provides the results of the average standard errors for each performance level. Generally, the average standard error is largest in the Below Proficiency and Above Proficiency performance level for all subjects, which can be expected given a shortage of very easy and very difficult items in the item pools to better measure low-performing and high-performing students. Within a subject and grade, average standard errors seem to be consistent across administrations both overall and for each performance level.

Table 41: Average Standard Error of Measurement by Performance Level, ELA

Grade	Admin	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency	Overall
G3	SP24	23.463	21.017	22.184	25.177	22.895
	SP23	23.708	21.260	22.451	25.566	23.203
	SP22	26.699	21.337	21.257	23.881	23.803
	SP21	26.982	21.360	21.210	23.755	23.948
	SP19	28.699	22.133	21.244	23.319	24.136
G4	SP24	27.849	24.598	25.06	27.9	26.502
	SP23	27.794	24.531	24.937	27.828	26.450
	SP22	27.282	24.104	24.520	28.052	26.139
	SP21	27.151	23.179	23.919	28.039	25.712
	SP19	28.271	23.407	23.774	27.279	25.749
G5	SP24	27.444	24.317	25.621	30.565	26.691
	SP23	27.441	24.378	25.622	30.214	26.666
	SP22	28.278	25.034	25.716	28.865	26.915
	SP21	28.385	25.036	25.404	28.421	26.781
	SP19	30.198	25.662	25.628	29.229	27.509
G6	SP24	28.082	24.280	24.894	27.738	26.372
	SP23	27.933	23.981	24.545	27.076	26.035
	SP22	29.282	24.218	24.446	26.734	26.515
	SP21	28.591	23.192	23.889	26.310	25.687
	SP19	28.698	22.248	22.564	25.313	24.623
G7	SP24	29.432	24.855	25.446	29.704	27.296
	SP23	29.399	24.591	25.087	29.583	27.085
	SP22	30.512	24.955	25.357	29.442	27.558
	SP21	30.997	24.952	25.306	29.205	27.654
	SP19	31.583	25.382	25.646	29.993	27.922
G8	SP24	29.821	25.588	26.642	30.385	28.139
	SP23	27.893	24.515	25.813	29.636	26.766
	SP22	28.541	23.666	24.952	28.865	26.370
	SP21	28.189	23.512	24.901	28.817	26.202
	SP19	29.808	24.982	25.704	29.172	27.112

Table 42: Average Standard Error of Measurement by Performance Level, Mathematics

Grade	Admin	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency	Overall
G3	SP24	18.384	14.923	15.122	18.573	16.869
	SP23	19.051	15.135	15.361	19.175	17.317

	SP22	17.328	14.602	14.856	17.904	16.239
	SP21	17.429	14.593	14.756	17.899	16.241
	SP19	19.198	15.686	15.701	20.058	17.618
G4	SP24	19.335	16.569	16.240	18.319	17.711
	SP23	19.124	16.712	16.310	18.037	17.617
	SP22	18.986	16.621	16.070	17.265	17.344
	SP21	19.238	16.223	15.534	16.994	17.242
	SP19	20.471	16.996	16.415	18.905	18.095
G5	SP24	21.783	17.924	16.972	17.984	19.098
	SP23	22.056	17.949	16.882	17.624	19.097
	SP22	22.884	17.095	16.318	17.144	18.968
	SP21	22.497	17.028	16.233	17.043	18.864
	SP19	25.014	18.081	17.251	20.960	20.401
G6	SP24	23.339	18.921	18.038	20.220	20.691
	SP23	24.496	19.257	18.384	20.616	21.387
	SP22	24.278	19.127	18.268	19.818	21.190
	SP21	23.971	19.197	18.301	19.660	21.137
	SP19	24.893	19.638	18.434	19.655	20.925
G7	SP24	26.941	19.209	18.553	18.388	22.112
	SP23	26.677	19.346	18.720	18.348	22.121
	SP22	27.475	20.262	18.745	17.954	22.660
	SP21	26.181	19.882	18.650	17.889	21.970
	SP19	31.086	20.923	18.907	19.944	23.526
G8	SP24	30.280	23.864	21.536	22.733	26.176
	SP23	30.736	23.883	21.796	22.687	26.423
	SP22	29.943	23.473	21.578	21.758	25.861
	SP21	29.099	23.428	21.208	20.460	25.374
	SP19	31.421	23.757	21.621	22.394	25.764

Table 43: Average Standard Error of Measurement by Performance Level, Science

Grade	Admin	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency	Overall
G4	SP24	5.401	4.975	5.053	5.742	5.171
G6	SP24	6.098	5.837	5.801	5.937	5.908
Biology (Fall)	SP24	--	--	--	--	--
Biology	SP24	6.771	6.007	5.658	5.643	6.003
Biology (Spring)	SP24	6.754	6.001	5.644	5.669	5.994

* Science tests of Spring 2024 are new NGSS tests

Table 44: Average Standard Error of Measurement by Performance Level, Social Studies

Grade	Admin	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency	Overall
G5	SP24	18.069	16.000	16.793	23.012	18.241
	SP23	17.793	15.008	16.017	23.203	17.853
	SP22	19.030	16.000	16.657	23.537	18.745
	SP21	19.259	16.994	17.700	24.257	19.341
	SP19	17.541	15.995	17.750	25.377	18.997
US Government	SP24	21.533	--	19.750	-	21.255
	SP23	18.012	--	15.484	--	17.525

SP22	18.321	--	15.870	--	17.845
SP21	17.360	--	15.673	--	16.821
SP19	18.050	--	15.692	--	17.576

Appendix 3-D, Standard Error of Measurement Curves by Subgroup, shows SEM curves for overall students and by subgroup. Appendix 3-E, Standard Error of Measurement Curves by Reporting Category, shows SEM curves by reporting category.

1.17.3 STUDENT CLASSIFICATION RELIABILITY

When student performance is reported in terms of performance categories, a reliability index 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* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014)). This index considers the consistency of classifications for the percentage of test takers who would, hypothetically, be classified in the same category on a second *ILEARN* administration, using either the same form or an alternate, equivalent form.

Students can be misclassified in one of two ways. Students who are truly below a proficiency cut point but are classified based on the assessment as being above the cut point are considered to be *false positives*. Similarly, students who are truly above a proficiency cut point but are classified as being below the cut point are considered to be *false negatives*.

Decision accuracy refers to the agreement between the classifications based on the form taken and the classifications that would be made based on the test taker's true scores. *Decision consistency* refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate form, that is, the percentages of students who are consistently classified in the same proficiency levels on two equivalent administrations of the test.

For a fixed-form test, the consistency of classifications is estimated on single-form test scores from a single test administration based on the true-score distribution that is estimated by fitting a bivariate beta-binomial model or a four-parameter beta model (Huynh, 1976; Livingston & Lewis, 1995). For the spring 2019 administration and all future CAT administrations, the consistency classification is based on all sets of items administered across students because the item selection algorithm constructs a test form unique to each student.

The classification index can be examined for decision accuracy and decision consistency. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of the test takers' true scores, if their true scores could somehow be known. 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 on the basis of an alternate, equivalently constructed test form or test administration (e.g., 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 administrations.

The true score is an expected value of the test score with measurement error. For a student with estimated ability $\hat{\theta}$ and associated standard error $se\left(\hat{\theta}\right)$, we can assume that $\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $se\left(\hat{\theta}\right)$, that is, $\hat{\theta} \sim N\left(\theta, se\left(\hat{\theta}\right)^2\right)$. The probability of the true score at or above the cut score θ_c is estimated as

$$P\left(\theta \geq \theta_c\right) = P\left(\frac{\theta - \hat{\theta}}{se\left(\hat{\theta}\right)} \geq \frac{\theta_c - \hat{\theta}}{se\left(\hat{\theta}\right)}\right) = P\left(\frac{\hat{\theta} - \theta}{se\left(\hat{\theta}\right)} < \frac{\hat{\theta} - \theta_c}{se\left(\hat{\theta}\right)}\right) = \Phi\left(\frac{\hat{\theta} - \theta_c}{se\left(\hat{\theta}\right)}\right),$$

where $\Phi(\cdot)$ is the cumulative function of standard normal distribution. Similarly, the probability of the true score being below the cut score is estimated as

$$P\left(\theta < \theta_c\right) = 1 - \Phi\left(\frac{\hat{\theta} - \theta_c}{se\left(\hat{\theta}\right)}\right).$$

1.17.4 CLASSIFICATION ACCURACY

Instead of assuming a normal distribution, we can directly estimate the probability of consistent classification using the likelihood function. The likelihood function of the achievement attribute, designated θ, given a student's item scores, represents the likelihood of the student's ability at that theta value. Integrating the likelihood values over the range of theta at and above the cut score (with proper normalization) represents the probability of the student's latent ability or the true score being at or above that cut point.

If a student's 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 a student with true ability θ being classified at or above the cut score θ_c, given the student's item scores $x = \left(x_1, \cdots, x_N\right)$, can be estimated as

$$P\left(\theta \geq \theta_c | x\right) = \frac{\int_{\theta_c}^{+\infty} L(\theta|x)d\theta}{\int_{-\infty}^{+\infty} L(\theta|x)d\theta},$$

where the likelihood function is

$$L(\theta|x) = \prod_{i=1}^{N} P(\theta),$$

and $P(\theta)$ is calculated from the Rasch model or partial credit model based on the estimated item parameters.

Similarly, we can estimate the probability of below the cut score as:

$$P\left(\theta < \theta_c|x\right) = \frac{\int_{-\infty}^{\theta_c} L(\theta|x)d\theta}{\int_{-\infty}^{+\infty} L(\theta|x)d\theta}$$

Mathematically, we have

$$N_{11} = \sum_{i \in N_1} P\left(\theta_i \geq \theta_c|x\right),$$

$$N_{01} = \sum_{i \in N_1} P\left(\theta_i < \theta_c|x\right),$$

$$N_{10} = \sum_{i \in N_0} P\left(\theta_i \geq \theta_c|x\right), \text{ and}$$

$$N_{00} = \sum_{i \in N_0} P\left(\theta_i < \theta_c|x\right),$$

where N_1 consists of the students with estimated $\hat{\theta}_i$ being at and above the cut score, and N_0 contains the students with estimated $\hat{\theta}_i$ being below the cut score. The accuracy index is then computed as:

$$\frac{N_{11}+N_{00}}{N_1+N_0} \, .$$

In Exhibit A, accurate classifications occur when the decision made based on the true score agrees with the decision made based on the form taken. Misclassifications, false positives, and false negatives occur when students' true-score classifications differ from their observed-score classifications (e.g., a student whose true score results in a Proficient level classification but is classified incorrectly as Approaching Proficient). N_{11} represents the expected numbers of students who are truly above the cut score; N_{01} represents the expected number of students falsely above the cut score; N_{00} represents

the expected number of students truly below the cut score; and N_{10} represents the number of students falsely below the cut score.

<p style="text-align:center">Exhibit A: Classification Accuracy</p>

		Classification on a Form Actually Taken	
		At or Above the Cut Score	**Below the Cut Score**
Classification on True Score	**At or Above the Cut Score**	N_{11} (Truly above the cut score)	N_{10} (False negative)
	Below the Cut Score	N_{01} (False positive)	N_{00} (Truly below the cut)

1.17.5 CLASSIFICATION CONSISTENCY

To estimate the consistency, we assume students are tested twice independently; hence, the probability of the student being classified as at or above the cut score θ_c in both tests can be estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c) = P(\theta_1 \geq \theta_c)P(\theta_2 \geq \theta_c) = \left(\frac{\int_{\theta_c}^{+\infty} L(x)d\theta}{\int_{-\infty}^{+\infty} L(x)d\theta} \right)^2 .$$

Similarly, the probability of consistency for at or above the cut score is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 \geq \theta_c | x) = \left(\frac{\int_{\theta_c}^{+\infty} L(x)d\theta}{\int_{-\infty}^{+\infty} L(x)d\theta} \right)^2 .$$

The probability of consistency for below the cut score is estimated as

$$P(\theta_1 < \theta_c, \theta_2 < \theta_c | x) = \left(\frac{\int_{-\infty}^{\theta_c} L(x)d\theta}{\int_{-\infty}^{+\infty} L(x)d\theta} \right)^2 .$$

The probability of inconsistency is estimated as

$$P(\theta_1 \geq \theta_c, \theta_2 < \theta_c | x) = \frac{\int_{\theta_c}^{+\infty} L(x)d\theta \int_{-\infty}^{\theta_c} L(x)d\theta}{\left[\int_{-\infty}^{+\infty} L(x)d\theta \right]^2}, \; and$$

$$P(\theta_1 < \theta_c, \theta_2 \geq \theta_c | x) = \frac{\int\limits_{-\infty}^{\theta_c} L(x)d\theta \int\limits_{\theta_c}^{+\infty} L(x)d\theta}{\left[\int\limits_{-\infty}^{+\infty} L(x)d\theta\right]^2}.$$

The consistent index is computed as

$$\frac{N_{11} + N_{00}}{N},$$

where

$$N_{11} = \sum_{i \in N} P\left(\theta_{i,1} \geq \theta_c, \theta_{i,2} \geq \theta_c | x\right),$$

$$N_{01} = \sum_{i \in N} P\left(\theta_i < \theta_c, \theta_{i,2} \geq \theta_c | x\right),$$

$$N_{10} = \sum_{i \in N} P\left(\theta_i \geq \theta_c, \theta_{i,2} < \theta_c | x\right),$$

$$N_{00} = \sum_{i \in N} P\left(\theta_i < \theta_c, \theta_{i,2} < \theta_c | x\right), \text{ and}$$

$$N = N_{11} + N_{10} + N_{01} + N_{00}.$$

As shown in Exhibit B, consistent classification occurs when two forms agree on the classification of a student as either at or above or below the performance standard, whereas inconsistent classification occurs when the decisions made by the forms differ.

Exhibit B: Classification Consistency

		Classification on the Second Form Taken	
		Above the Cut Score	**Below the Cut Score**
Classification on the First Form Taken	**At or Above the Cut Score**	N_{11} (Consistently above the cut)	N_{10} (Inconsistent)
	Below the Cut Score	N_{01} (Inconsistent)	N_{00} (Consistently below the cut)

1.17.6 CLASSIFICATION ACCURACY AND CONSISTENCY ESTIMATES

The analysis of the classification index is performed for test scores in the 2022–2023 administration. Table 45 through * Science tests of Spring 2024 are new NGSS tests which are different from the previous tests

Table 48 present the decision accuracy and consistency indices. Accuracy classifications are slightly higher than the consistency classifications in all performance standards. The consistency classification rate can be somewhat lower than the accuracy rate because consistency assumes two test scores, both of which include measurement error, while the accuracy rate assumes a single test score and the true score, which does not include measurement error. The classification index ranged from 89% to 96% for accuracy, and from 86% to 95% for consistency across all grades and subjects. The accuracy and consistency rates for each performance standard are greater for the performance standards associated with smaller standard errors. The better the test is targeted to the student's ability, the higher the classification index. Within a subject and grade, classification indices seem to be consistent across administrations.

Table 45: Decision Accuracy and Consistency Indices for Performance Standards, ELA

Grade	Admin	Accuracy			Consistency		
		Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4	Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4
3	SP24	0.923	0.918	0.940	0.892	0.885	0.916
	SP23	0.925	0.919	0.938	0.894	0.886	0.913
	SP22	0.921	0.889	0.917	0.884	0.937	0.911
	SP21	0.918	0.919	0.941	0.885	0.886	0.916
	SP19	0.912	0.906	0.931	0.872	0.863	0.863
4	SP24	0.920	0.911	0.930	0.887	0.875	0.902
	SP23	0.917	0.913	0.934	0.883	0.878	0.907
	SP22	0.920	0.887	0.914	0.878	0.931	0.903
	SP21	0.922	0.916	0.932	0.922	0.881	0.907
	SP19	0.918	0.904	0.925	0.882	0.861	0.861
5	SP24	0.924	0.916	0.940	0.893	0.883	0.916
	SP23	0.924	0.918	0.940	0.892	0.884	0.916
	SP22	0.922	0.889	0.910	0.874	0.937	0.911
	SP21	0.918	0.908	0.942	0.885	0.871	0.918
	SP19	0.917	0.901	0.931	0.881	0.856	0.856
6	SP24	0.921	0.914	0.934	0.888	0.879	0.907
	SP23	0.919	0.914	0.935	0.886	0.879	0.908
	SP22	0.916	0.882	0.910	0.873	0.934	0.908
	SP21	0.918	0.908	0.934	9.884	0.871	0.908
	SP19	0.922	0.908	0.926	0.890	0.865	0.865
7	SP24	0.923	0.910	0.931	0.892	0.876	0.905
	SP23	0.922	0.913	0.938	0.890	0.878	0.912
	SP22	0.925	0.894	0.906	0.868	0.930	0.904
	SP21	0.922	0.906	0.933	0.890	0.867	0.906
	SP19	0.928	0.902	0.918	0.898	0.858	0.858
8	SP24	0.931	0.918	0.929	0.903	0.885	0.901
	SP23	0.931	0.913	0.927	0.902	0.878	0.898
	SP22	0.931	0.903	0.914	0.878	0.927	0.898
	SP21	0.932	0.916	0.925	0.903	0.881	0.895
	SP19	0.931	0.903	0.917	0.904	0.858	0.858

Table 46: Decision Accuracy and Consistency Indices for Performance Standards, Mathematics

Grade	Admin	Accuracy			Consistency		
		Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4	Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4
3	SP24	0.952	0.946	0.950	0.933	0.923	0.929
	SP23	0.953	0.945	0.949	0.934	0.922	0.928
	SP22	0.952	0.933	0.946	0.922	0.951	0.931
	SP21	0.950	0.945	0.954	0.932	0.923	0.935
	SP19	0.951	0.938	0.940	0.930	0.912	0.913
4	SP24	0.948	0.939	0.954	0.927	0.914	0.935
	SP23	0.947	0.939	0.953	0.926	0.914	0.934
	SP22	0.945	0.922	0.939	0.913	0.955	0.937
	SP21	0.945	0.940	0.960	0.923	0.915	0.944
	SP19	0.946	0.932	0.948	0.923	0.902	0.926
5	SP24	0.944	0.940	0.957	0.921	0.915	0.938
	SP23	0.944	0.939	0.957	0.921	0.915	0.939
	SP22	0.945	0.922	0.942	0.918	0.958	0.940
	SP21	0.943	0.941	0.959	0.919	0.917	0.942
	SP19	0.942	0.933	0.948	0.917	0.905	0.927
6	SP24	0.943	0.940	0.960	0.919	0.916	0.943
	SP23	0.943	0.942	0.957	0.919	0.918	0.940
	SP22	0.942	0.919	0.941	0.917	0.960	0.943
	SP21	0.937	0.941	0.962	0.913	0.917	0.947
	SP19	0.943	0.930	0.948	0.918	0.899	0.927
7	SP24	0.946	0.949	0.963	0.923	0.928	0.948
	SP23	0.944	0.948	0.964	0.920	0.927	0.950
	SP22	0.936	0.910	0.948	0.930	0.967	0.956
	SP21	0.932	0.948	0.969	0.905	0.928	0.957
	SP19	0.937	0.937	0.955	0.910	0.911	0.938
8	SP24	0.940	0.947	0.964	0.915	0.925	0.949
	SP23	0.940	0.947	0.851	0.920	0.925	0.949
	SP22	0.938	0.912	0.950	0.934	0.970	0.962
	SP21	0.933	0.951	0.972	0.906	0.933	0.964
	SP19	0.938	0.938	0.938	0.909	0.911	0.937

Table 47: Decision Accuracy and Consistency Indices for Performance Standards, Science

Grade	Admin	Accuracy			Consistency		
		Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4	Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4
4	SP24	0.929	0.913	0.945	0.905	0.883	0.925
6	SP24	0.905	0.898	0.919	0.882	0.869	0.896

Biology (Fall)	SP24	-	-	-	-	-	-
Biology (Winter)	SP24	0.893	0.901	0.942	0.866	0.871	0.920
Biology (Spring)	SP24	0.897	0.906	0.934	0.896	0.876	0.913

* Science tests of Spring 2024 are new NGSS tests which are different from the previous tests

Table 48: Decision Accuracy and Consistency Indices for Performance Standards, Social Studies

Grade	Admin	Accuracy			Consistency		
		Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4	Cut 1–Cut 2	Cut 2–Cut 3	Cut 3–Cut 4
5	SP24	0.905	0.917	0.942	0.861	0.880	0.917
	SP23	0.911	0.921	0.944	0.870	0.885	0.919
	SP22	0.904	0.865	0.918	0.884	0.944	0.923
	SP21	0.902	0.911	0.941	0.861	0.874	0.917
	SP19	0.907	0.908	0.930	0.866	0.866	0.898
US Government	SP24	0.956	--	--	0.936	--	--
	SP23	0.944	--	--	0.919	--	--
	SP22	0.956	--	--	0.936	--	--
	SP21	0.933	--	--	0.903	--	--
	SP19	0.954	--	--	0.932	--	--

1.17.7 RELIABILITY FOR SUBGROUPS IN THE POPULATION

The 2023–2024 marginal reliability results for each of the identified subgroups (gender, ethnicity [White, African American, Asian, Hispanic, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, and Multi-Racial], special education students, section 504 students, and SES students) were calculated. The marginal reliability coefficients for subgroups along with historical statistics are provided in Appendix 3-F, Marginal Reliability Coefficients for Overall and by Subgroup. As the appendix indicates, reliabilities are consistent across subgroups, indicating that the ILEARN assessments measure a common underlying achievement dimension across all subgroups. Where reliability estimates are attenuated, there is an associated decrease in variance within the subgroup population, indicating that the decrease in reliability is likely due to a restriction in range.

1.17.8 REPORTING CATEGORY RELIABILITY

The marginal reliability coefficients and the measurement errors are computed for the reporting categories. Table 49 through Table 52 present the marginal reliability coefficients for reporting categories.

Table 49: Marginal Reliability Coefficients for ELA Reporting Categories

Grade	Reporting Categories	BP min	BP max	Mean	SD	Min	Max	Marginal Reliability
3	Key Ideas and Textual Support/Vocabulary	12	15	5442.26	96.93	5060	5760	0.76
	Structural Elements and Organization/Connection of Ideas/Media Literacy	10	12	5446.11	95.21	5060	5760	0.72
	Writing	6	8	5397.1	116.36	5060	5760	0.70
4	Key Ideas and Textual Support/Vocabulary	11	14	5479.62	98.19	5090	5810	0.73
	Structural Elements and Organization/Connection of Ideas/Media Literacy	11	14	5475.55	107.65	5090	5810	0.68
	Writing	6	8	5453.31	109.64	5090	5810	0.74
5	Key Ideas and Textual Support/Vocabulary	11	14	5503.7	101.41	5110	5850	0.74
	Structural Elements and Organization/Connection of Ideas/Media Literacy	11	14	5501.16	96.96	5110	5850	0.69
	Writing	6	8	5489.23	115.62	5110	5850	0.79
6	Key Ideas and Textual Support/Vocabulary	10	13	5521.65	106.07	5130	5870	0.69
	Structural Elements and Organization/Connection of Ideas/Media Literacy	10	13	5528.29	101.11	5130	5870	0.72
	Writing	6	8	5509.4	105.33	5130	5870	0.78
7	Key Ideas and Textual Support/Vocabulary	10	13	5543.56	106.03	5130	5890	0.74
	Structural Elements and Organization/Connection of Ideas/Media Literacy	10	13	5539.92	113.86	5130	5890	0.67
	Writing	6	8	5548.88	108.55	5130	5890	0.77
8	Key Ideas and Textual Support/Vocabulary	10	12	5558.68	114.97	5150	5920	0.77
	Structural Elements and Organization/Connection of Ideas/Media Literacy	10	12	5552.48	127.49	5150	5920	0.62
	Writing	6	8	5541.88	113.18	5150	5920	0.79

Table 50: Marginal Reliability Coefficients for Mathematics Reporting Categories

Grade	Reporting Categories	BP min	BP max	Mean	SD	Min	Max	Marginal Reliability
3	Algebraic Thinking and Data Analysis	9	11	6421.23	111.92	6080	6730	0.76
	Computation	11	13	6427.46	96.44	6080	6730	0.78
	Geometry and Measurement	9	11	6423.78	102.03	6080	6730	0.80

Grade	Reporting Categories	BP min	BP max	Mean	SD	Min	Max	Marginal Reliability
	Number Sense	11	13	6433.99	99.93	6080	6730	0.83
4	Algebraic Thinking and Data Analysis	9	11	6468.03	109.16	6100	6800	0.75
	Computation	11	13	6472.31	107.48	6100	6800	0.80
	Geometry and Measurement	9	11	6455.09	109.78	6100	6800	0.78
	Number Sense	10	13	6467.37	106.22	6100	6800	0.79
5	Algebraic Thinking	10	12	6489.39	98.81	6110	6850	0.78
	Computation	11	13	6488.43	108.15	6110	6850	0.80
	Geometry and Measurement, Data Analysis, and Statistics	9	11	6481.56	123.08	6110	6850	0.76
	Number Sense	11	13	6488.25	105.6	6110	6850	0.77
6	Algebra and Functions	11	13	6508.84	120.89	6110	6870	0.84
	Computation	10	12	6511.65	134.94	6110	6870	0.78
	Geometry and Measurement, Data Analysis, and Statistics	9	11	6505.12	131.5	6110	6870	0.78
	Number Sense	10	12	6507.75	121.22	6110	6870	0.80
7	Algebra and Functions	11	13	6509.42	126.37	6120	6920	0.82
	Data Analysis, Statistics, and Probability	9	11	6511.06	134.8	6120	6920	0.78
	Geometry and Measurement	9	11	6485.88	156.69	6120	6920	0.73
	Number Sense and Computation	11	13	6516.82	119.98	6120	6920	0.83
8	Algebra and Functions	11	13	6514.51	132.57	6120	6950	0.82
	Data Analysis, Statistics, and Probability	10	12	6532.07	145.58	6120	6950	0.78
	Geometry and Measurement	10	12	6521.87	142.75	6120	6950	0.78
	Number Sense and Computation	9	11	6542.03	158.28	6120	6950	0.74

Table 51: Marginal Reliability Coefficients for Science Reporting Categories

Grade	Reporting Categories	BP min	BP max	Mean	SD	Min	Max	Marginal Reliability
	Physical Science	6	6	159.50	19.22	100	220	0.70
4	Life Science	4	4	159.72	19.41	100	220	0.67
	Earth and Space Science	5	5	162.88	26.71	100	220	0.69

Grade	Reporting Categories	BP min	BP max	Mean	SD	Min	Max	Marginal Reliability
	Computer Science	8	8	358.74	18.99	300	420	0.59
6	Physical Science	4	4	358.17	19.14	300	420	0.69
	Life Science	6	6	362.22	18.65	300	420	0.54
	Earth and Space Science	4	4	364.09	26.88	300	420	0.66
	Computer Science	8	8	159.50	19.22	100	220	0.70
Biology (Fall)	From Molecules to Organisms: Structures and Processes	6	6	--	--	--	--	--
	Ecosystems: Interactions, Energy and Dynamics	6	6	--	--	--	--	--
	Heredity and Evolution	6	6	--	--	--	--	--
Biology (Winter)	From Molecules to Organisms: Structures and Processes	6	6	560.03	18.75	500	620	0.61
	Ecosystems: Interactions, Energy and Dynamics	6	6	557.56	18.36	500	620	0.66
	Heredity and Evolution	6	6	559.15	18.56	503	620	0.70
Biology (Spring)	From Molecules to Organisms: Structures and Processes	6	6	560.79	19.33	500	620	0.63
	Ecosystems: Interactions, Energy and Dynamics	6	6	558.03	19.02	500	620	0.68
	Heredity and Evolution	6	6	559.70	18.66	500	620	0.70

Table 52: Marginal Reliability Coefficients for Social Studies Reporting Categories

Grade	Reporting Categories	BP min	BP max	Mean	SD	Min	Max	Marginal Reliability
5	Civics and Government	15	17	8494.96	60.84	8350	8650	0.74
	Geography and Economics	11	13	8489.06	63.27	8350	8650	0.61
	History	11	13	8485.56	60.99	8350	8650	0.69
U.S. Government	Functions of Government	19	21	8446.13	60.57	8350	8650	0.66
	Historical Foundations of American Government	13	15	8448.61	59.62	8350	8618	0.34
	Institutions and Processes of Government	19	21	8442.84	61.62	8350	8647	0.61

1.17.9 RELIABILITY FOR ACCOMMODATED TESTERS

Internal consistency reliabilities are also calculated for accommodated test administrations, including Spanish and Braille forms and the paper form, and online accommodations provided to eligible students. Given the small number of students for accommodated test administrations, Spanish and Braille forms and the paper form are collapsed into a single category for the reliability analysis.

Table 53 shows the marginal reliabilities for accommodated forms, online accommodations, and non-accommodated test administrations. Note that even when collapsing across forms, some assessments had no accommodated test administrations, and for others, the number of accommodated testers was very small, limiting the generalizability of the results. Nevertheless, the internal consistency reliabilities of online accommodated test administrations were comparable to those of non-accommodated test administrations, indicating that, like the non-accommodated assessments, accommodated test administrations result in test scores of similar precision.

Table 53: Marginal Reliability Coefficients for Accommodated vs Non-Accommodated Students

Grade	Accommodated Forms		Online Accommodations		Non-Accommodated	
	N	Reliability	*N*	Reliability	*N*	Reliability
ELA						
3	114	0.873	24937	0.884	56726	0.894
4	107	0.888	24917	0.872	57872	0.880
5	94	0.818	23055	0.874	58095	0.885
6	75	0.875	21531	0.862	60952	0.880
7	51	0.823	21211	0.858	60890	0.878
8	49	0.882	20836	0.873	62116	0.889
Mathematics						
3	1137	0.922	23968	0.955	56633	0.952
4	1232	0.907	23871	0.951	57751	0.949
5	1308	0.884	21920	0.936	57977	0.946
6	1355	0.880	20333	0.943	60831	0.952
7	1359	0.804	19967	0.923	60738	0.952
8	1403	0.809	19540	0.911	61974	0.945
Science						
4	1217	0.774	23832	0.879	57694	0.882
6	1350	0.669	20135	0.828	60364	0.853
Biology (Spring)	1236	0.584	17722	0.788	61196	0.862
Social Studies						
5	1305	0.657	21875	0.839	57921	0.869
U.S. Government	0	--	41	0.736	190	0.842

1.18 SUBSCALE INTERCORRELATIONS

Table 54 through

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
4	Physical Science (Cat1)	6	1			
	Life Science (Cat2)	4	0.67	1		
	Earth and Space Science (Cat3)	5	0.68	0.64	1	
	Computer Science (Cat4)	8	0.59	0.55	0.56	1
6	Physical Science (Cat1)	4	1			
	Life Science (Cat2)	6	0.64	1		
	Earth and Space Science (Cat3)	4	0.55	0.60	1	
	Computer Science (Cat4)	8	0.53	0.58	0.48	1
Biology (Fall)	From Molecules to Organisms: Structures and Processes (Cat1)	6	--	--	--	--
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	--	--	--	--
	Heredity and Evolution (Cat3)	6	--	--	--	--
Biology (Winter)	From Molecules to Organisms: Structures and Processes (Cat1)	6	1			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	0.62	1		
	Heredity and Evolution (Cat3)	6	0.65	0.69	1	
Biology (Spring)	From Molecules to Organisms: Structures and Processes (Cat1)	6	1			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	0.65	1		
	Heredity and Evolution (Cat3)	6	0.66	0.69	1	

Table 57 present the observed correlation matrix of the reporting category scores for each subject area. In ELA, the correlations among the reporting categories ranged from 0.55 to 0.71. In mathematics, the correlations were between 0.64 and 0.81. In science, the correlations among reporting categories ranged from 0.48 to 0.69. In social studies, the correlations ranged from 0.54 to 0.72.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the strand level, given the limited number of items from which the scores were derived. Consequently, overinterpretation of these correlations, as either high or low, should be avoided cautiously.

Table 54: Observed Correlations Among Reporting Category Scores for ELA, Grades 3–8

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Key Ideas and Textual Support/Vocabulary (Cat1)	12–15	1			
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.69	1		
	Writing (Cat3)	6–8	0.58	0.55	1	
4	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1			
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.65	1		
	Writing (Cat3)	5-6	0.65	0.6	1	
5	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1			
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.69	1		
	Writing (Cat3)	4-6	0.69	0.64	1	
6	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1			
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.65	1		
	Writing (Cat3)	5-6	0.63	0.64	1	
7	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1			
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.65	1		
	Writing (Cat3)	5-6	0.68	0.61	1	
8	Key Ideas and Textual Support/Vocabulary (Cat1)	10–12	1			
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.62	1		
	Writing (Cat3)	4-6	0.71	0.6	1	

Table 55: Observed Correlations Among Reporting Category Scores for Mathematics, Grades 3–8

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	8-10	1			
	Computation (Cat2)	9-11	0.77	1		
	Geometry and Measurement (Cat3)	9–11	0.75	0.76	1	
	Number Sense (Cat4)	11–13	0.75	0.77	0.78	1
4	Algebraic Thinking and Data Analysis (Cat1)	9–10	1			
	Computation (Cat2)	11–13	0.75	1		
	Geometry and Measurement (Cat3)	8-9	0.72	0.73	1	
	Number Sense (Cat4)	11–13	0.75	0.76	0.74	1
5	Algebraic Thinking (Cat1)	10	1			
	Computation (Cat2)	11–13	0.76	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9–10	0.72	0.73	1	
	Number Sense (Cat4)	11–13	0.73	0.74	0.70	1
6	Algebra and Functions (Cat1)	11–12	1			
	Computation (Cat2)	10–12	0.74	1		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9–11	0.77	0.67	1	
	Number Sense (Cat4)	9-11	0.81	0.70	0.74	1
7	Algebra and Functions (Cat1)	11–13	1			
	Data Analysis, Statistics, and Probability (Cat2)	9–11	0.74	1		
	Geometry and Measurement (Cat3)	8-10	0.70	0.65	1	
	Number Sense and Computation (Cat4)	11-12	0.79	0.75	0.69	1
8	Algebra and Functions (Cat1)	10-12	1			
	Data Analysis, Statistics, and Probability (Cat2)	10–12	0.76	1		
	Geometry and Measurement (Cat3)	10–12	0.73	0.72	1	
	Number Sense and Computation (Cat4)	9–11	0.67	0.64	0.64	1

Table 56: Observed Correlations Among Reporting Category Scores for Science

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
4	Physical Science (Cat1)	6	1			
	Life Science (Cat2)	4	0.67	1		
	Earth and Space Science (Cat3)	5	0.68	0.64	1	
	Computer Science (Cat4)	8	0.59	0.55	0.56	1
6	Physical Science (Cat1)	4	1			
	Life Science (Cat2)	6	0.64	1		
	Earth and Space Science (Cat3)	4	0.55	0.60	1	
	Computer Science (Cat4)	8	0.53	0.58	0.48	1
Biology (Fall)	From Molecules to Organisms: Structures and Processes (Cat1)	6	--	--	--	--
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	--	--	--	--
	Heredity and Evolution (Cat3)	6	--	--	--	--
Biology (Winter)	From Molecules to Organisms: Structures and Processes (Cat1)	6	1			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	0.62	1		
	Heredity and Evolution (Cat3)	6	0.65	0.69	1	
Biology (Spring)	From Molecules to Organisms: Structures and Processes (Cat1)	6	1			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	0.65	1		
	Heredity and Evolution (Cat3)	6	0.66	0.69	1	

Table 57: Observed Correlations Among Reporting Category Scores for Social Studies

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	0-20	1		
	Geography and Economics (Cat2)	0-20	0.66	1	
	History (Cat3)	0-20	0.72	0.66	1
U.S. Government	Functions of Government (Cat1)	13	1		
	Historical Foundations of American Government (Cat2)	9	0.54	1	
	Institutions and Processes of Government (Cat3)	14	0.72	0.57	1

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

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

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 58 through Table 61 present disattenuated correlations. Disattenuated correlation is capped if the correlation is greater than 1. These values suggest that validity evidence of the internal structure is supported.

Table 58: Disattenuated Correlations Among Reporting Category Scores for ELA

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	12–15	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.94	1	
	Writing (Cat3)	6–8	0.80	0.78	1
4	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.92	1	
	Writing (Cat3)	7–8	0.89	0.85	1
5	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.96	1	
	Writing (Cat3)	6–8	0.91	0.87	1

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
6	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.92	1	
	Writing (Cat3)	7–8	0.85	0.85	1
7	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.92	1	
	Writing (Cat3)	7–8	0.90	0.85	1
8	Key Ideas and Textual Support/Vocabulary (Cat1)	10–12	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.90	1	
	Writing (Cat3)	6–8	0.90	0.86	1

Table 59: Disattenuated Correlations Among Reporting Category Scores for Mathematics

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	9–11	1			
	Computation (Cat2)	11–13	0.99	1		
	Geometry and Measurement (Cat3)	9–11	0.96	0.97	1	
	Number Sense (Cat4)	11–13	0.94	0.95	0.96	1
4	Algebraic Thinking and Data Analysis (Cat1)	9–11	1			
	Computation (Cat2)	11–13	0.96	1		
	Geometry and Measurement (Cat3)	9–11	0.94	0.92	1	
	Number Sense (Cat4)	11–13	0.97	0.95	0.94	1
5	Algebraic Thinking (Cat1)	10–12	1			
	Computation (Cat2)	11–13	0.97	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9–11	0.94	0.94	1	
	Number Sense (Cat4)	11–13	0.94	0.95	0.92	1
6	Algebra and Functions (Cat1)	11–13	1			
	Computation (Cat2)	10–12	0.91	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9–11	0.95	0.85	1	
	Number Sense (Cat4)	10–12	0.98	0.88	0.94	1
7	Algebra and Functions (Cat1)	11–12	1			
	Data Analysis, Statistics, and Probability (Cat2)	9–11	0.93	1		
	Geometry and Measurement (Cat3)	9–11	0.90	0.87	1	
	Number Sense and Computation (Cat4)	12–13	0.96	0.93	0.89	1
8	Algebra and Functions (Cat1)	11–13	1			
	Data Analysis, Statistics, and Probability (Cat2)	10–12	0.95	1		
	Geometry and Measurement (Cat3)	10–12	0.91	0.92	1	

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Number Sense and Computation (Cat4)	9–11	0.86	0.84	0.84	1

Table 60: Disattenuated Correlations Among Reporting Category Scores for Science

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
4	Physical Science (Cat1)	6	1			
	Life Science (Cat2)	4	0.93	1		
	Earth and Space Science (Cat3)	5	0.96	0.94	1	
	Computer Science (Cat4)	8	0.82	0.79	0.82	1
6	Physical Science (Cat1)	4	1			
	Life Science (Cat2)	6	0.99	1		
	Earth and Space Science (Cat3)	4	0.97	0.97	1	
	Computer Science (Cat4)	8	0.85	0.85	0.79	1
Biology (Fall)	From Molecules to Organisms: Structures and Processes (Cat1)	6	--			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	--	--		
	Heredity and Evolution (Cat3)	6	--	--	--	
Biology (Winter)	From Molecules to Organisms: Structures and Processes (Cat1)	6	1			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	0.97	1		
	Heredity and Evolution (Cat3)	6	0.99	1	1	
Biology (Spring)	From Molecules to Organisms: Structures and Processes (Cat1)	6	1			
	Ecosystems: Interactions, Energy and Dynamics (Cat2)	6	0.98	1		
	Heredity and Evolution (Cat3)	6	0.98	0.99	1	

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

Table 61: Disattenuated Correlations Among Reporting Category Scores for Social Studies

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	17	1		
	Geography and Economics (Cat2)	11	0.98	1	
	History (Cat3)	12	1	1	1
U.S. Government	Functions of Government (Cat1)	20	1		
	Historical Foundations of American Government (Cat2)	14		1	1
	Institutions and Processes of Government (Cat3)	20	1	1	1

Note: Disattenuated values greater than 1.00 are reported as 1.00*.

1.19 HANDSCORED ITEMS INTER-RATER RELIABILITY

The basic method to compute inter-rater reliability (IRR) is percentage agreement. All English/Language Arts (ELA) writing prompts were handscored by a human with a 10% second read. As shown in Table 62, the percentage of exact agreement (when two raters gave the same score), the percentage of adjacent ratings (when the difference between two raters was 1), and the percentage of non-adjacent ratings (when the difference was greater than 1) were all computed. In this example, the percentage of exact agreement was 2/4, or 50%, and the adjacent and non-adjacent percentages were 25% each.

Table 62: Percentage Agreement Example

Response	Rater 1	Rater 2	Agreement
1	2	3	1
2	1	1	0
3	2	2	0
4	2	0	2

Likewise, IRR monitors how often scorers are in exact agreement with each other and ensures that an acceptable agreement rate is maintained. The calculations for the IRR in this report are as follows:

- *Percentage Exact* is the total number of responses by the scorer in which scores are equal, divided by the number of responses that were scored twice.

- *Percentage Adjacent* is the total number of responses by the scorer in which scores are one score point apart, divided by the number of responses that were scored twice.

- *Percentage Non-Adjacent* is the total number of responses by the scorer where scores are more than one score point apart, divided by the number of responses that were scored twice.

Table 63 displays the rater-agreement percentages for hand-scored writing items. The percentage of exact agreement between two raters ranged from 87% to 98%. The percentage of adjacent rating was between 2% and 13%. The non-adjacent percentages fell close to 0%. The total number of processed responses does not necessarily correspond to the number of student responses selected to be second read by a human reader. These numbers could potentially be higher, as some students might request rescoring and have their responses rescored as requested.

Table 63: Inter-Rater Reliability of Hand-Scored Writing Items

Grade	Dimension	Percentage Exact	Percentage Adjacent	Percentage Not Adjacent	Total Number of Processed Responses
3	Purpose, Focus, & Organization	98	2	0	39739
	Evidence & Elaboration	98	2	0	
	Conventions	98	2	0	
4	Purpose, Focus, & Organization	90	9	0	20821
	Evidence & Elaboration	90	9	0	
	Conventions	92	8	0	
5	Purpose, Focus, & Organization	91	9	0	17516
	Evidence & Elaboration	91	9	0	
	Conventions	92	8	0	
6	Purpose, Focus, & Organization	89	11	0	18329
	Evidence & Elaboration	89	11	0	
	Conventions	91	9	0	
7	Purpose, Focus, & Organization	87	13	0	14238
	Evidence & Elaboration	87	13	0	
	Conventions	88	12	0	
8	Purpose, Focus, & Organization	91	9	0	18084
	Evidence & Elaboration	91	9	0	
	Conventions	92	8	0	

Table 64 displays the rater-agreement percentages for hand-scored non-writing items. The percentage of exact agreement between two raters ranged from 82% to 97%. The percentage of adjacent rating was between 3% and 17%. The non-adjacent percentages fell between 0% and 1%. The total number of processed responses does not necessarily correspond to the number of student responses selected to be second read by a human reader. These numbers could potentially be higher, as some students might request rescoring and have their responses rescored as requested.

Table 64: Inter-Rater Reliability of Hand-Scored Non-Writing Items

Grade	Percentage Exact	Percentage Adjacent	Percentage Not Adjacent	Total Number of Processed Responses
ELA				
3	92	8	0	38552
4	87	12	0	33581

5	85	15	0	31621
6	85	14	0	32408
7	86	14	0	31745
8	85	15	0	33266
Mathematics				
3	96	4	0	34639
4	96	4	0	27960
5	91	9	0	31513
6	95	5	0	27871
7	97	3	0	31288
8	93	7	0	27902
Social Studies				
5	82	17	1	20145

Cohen's kappa (Cohen, 1968) is an index of inter-rater agreement after accounting for the agreement that could be expected due to chance. This statistic can be computed as

$$K = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the proportion of observed agreement, and P_c indicates the proportion of agreement by chance. Cohen's kappa treats all disagreement values with equal weights. Weighted kappa coefficients (Cohen, 1968), however, allow unequal weights, which can be used as a measure of validity. Weighted kappa coefficients were calculated using the following formula:

$$K_w = \frac{P'_o - P'_c}{1 - P'_c},$$

where

$$P'_o = \frac{\Sigma w_{ij} p_{oij}}{w_{max}},$$

$$P'_c = \frac{\Sigma w_{ij} p_{cij}}{w_{max}},$$

where p_{oij} is the proportion of the judgments observed in the *ij*th cell, p_{cij} is the proportion in the *ij*th cell expected by chance, and w_{ij} is the disagreement weight.

Weighted kappa coefficients for operational hand-scored writing prompts by dimension are presented in Table 65.

Table 65: Weighted Kappa Coefficients for Hand-Scored Writing Items

Grade	*N*	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
3	10712	0.862	0.862	0.865
4	6951	0.820	0.820	0.627
5	8811	0.850	0.850	0.746
6	9761	0.811	0.810	0.720
7	7783	0.831	0.830	0.602
8	9183	0.836	0.836	0.714

Table 66 presents weighted kappa coefficients for operational hand-scored non-writing items.

Table 66: Weighted Kappa Coefficients for Hand-Scored Non-Writing Items

Grade	*N*	Purpose, Focus, & Organization
ELA		
3	6388	0.439
4	9227	0.507
5	11928	0.526
6	10779	0.480
7	11042	0.511
8	13090	0.549
Mathematics		
3	9823	0.829
4	9458	0.862
5	15150	0.762
6	5196	0.696
7	7327	0.815
8	7831	0.752
Social Studies		
5	15150	0.762

1.20 ACCESSIBILITY RESOURCES ASSIGNMENT AND USAGE

The purpose of the analysis on accessibility resources was to monitor the assignment and usage of various accommodation tools and usage consistency throughout the test. The tools investigated were masking, print-on-demand, audio transcript, speech-to-text (STT), text-to-speech for all items including reading comprehension (TTS-All), and text-to-speech (TTS) for ELA, and masking, print-on-demand, TTS, and STT for mathematics. The text to speech tool is distinguished between TTS and TTS-All: TTS is for students for whom text to speech is not allowed for reading comprehension passages, whereas students for whom text to speech is assigned as an accommodation may access it on all items and passages in ELA (TTS-All).

Table 67 and Table 68 provide the number and percentage of students assigned each accessibility resource in ELA and mathematics. The number in parentheses shows the percentage assigned, which is calculated using the following formula:

$$\text{Percentage assigned} = \frac{\text{The number of students assigned the tool}}{\text{Total number of students}} * 100$$

Table 67: Number and Percentage Assigned Accessibility Resources (ELA)

Grade	Admin	Total *N*	Masking*	Print-On-Demand	Audio-Transcript	STT	TTS-All	TTS
3	SP24	81567	32(.04)	21(.03)	29(.04)	812(1.00)	3704(4.54)	12510(15.34)
	SP23	81934	55(.07)	11(.01)	22(.03)	443(.54)	3636(4.44)	12216(14.91)
	SP22	79668	32(.04)	17(.02)	21(.03)	141(.18)	3209(4.03)	11526(14.47)
4	SP24	82673	12(.01)	26(.03)	39(.05)	1126(1.4)	4570(5.53)	12939(15.65)
	SP23	80256	69(.09)	17(.02)	25(.03)	505(.63)	3969(4.95)	12278(15.30)
	SP22	80778	38(.05)	23(.03)	25(.03)	171(.21)	3512(4.35)	12277(15.20)
5	SP24	81052	32(.04)	18(.02)	34(.04)	1124(1.4)	4488(5.54)	11771(14.52)
	SP23	81621	47(.06)	20(.02)	19(.02)	482(.59)	3910(4.79)	11452(14.03)
	SP22	80806	50(.06)	21(.03)	28(.03)	230(.28)	3233(4.00)	11731(14.52)
6	SP24	82355	101(.12)	18(.02)	19(.02)	973(1.2)	4243(5.15)	10984(13.34)
	SP23	81403	200(.25)	22(.03)	32(.04)	428(.53)	3719(4.57)	10751(13.21)
	SP22	81958	149(.18)	28(.03)	26(.03)	169(.21)	3648(4.45)	10723(13.08)
7	SP24	81972	116(.14)	12(.01)	35(.04)	718(.88)	3979(4.85)	10877(13.27)
	SP23	82167	204(.25)	18(.02)	30(.04)	302(.37)	3707(4.51)	10873(13.23)
	SP22	83103	151(.18)	22(.03)	33(.04)	95(.11)	3454(4.16)	10576(12.73)
8	SP24	82808	111(.13)	10(.01)	27(.03)	568(.69)	3928(4.74)	10926(13.19)
	SP23	83350	214(.26)	18(.02)	29(.04)	201(.24)	3372(4.05)	10805(12.96)
	SP22	84737	175(.21)	38(.04)	21(.02)	59(.07)	3486(4.11)	10516(12.41)

*The first number is the number of students and the number in parentheses is the percentage assigned.

Table 68: Number and Percentage Assigned Accessibility Resources (Mathematics)

Grade	Admin	Total *N*	Masking*	Print-On-Demand	TTS	STT
3	SP24	81608	25(.03)	21(.03)	16175(19.82)	794(.97)
	SP23	81972	40(.05)	12(.01)	15825(19.31)	419(.51)
	SP22	79731	26(.03)	22(.03)	14760(18.51)	130(.16)
4	SP24	82749	14(.02)	29(.04)	17434(21.07)	1068(1.3)
	SP23	80312	49(.06)	17(.02)	16235(20.21)	487(.61)
	SP22	80828	27(.03)	26(.03)	15783(19.53)	169(.21)
5	SP24	81105	24(.03)	15(.02)	16231(20.01)	1087(1.3)
	SP23	81722	32(.04)	22(.03)	15369(18.81)	457(.56)
	SP22	80861	48(.06)	23(.03)	14952(18.49)	230(.28)
6	SP24	82429	101(.12)	19(.02)	14957(18.15)	914(1.1)
	SP23	81444	114(.14)	18(.02)	14452(17.74)	402(.49)
	SP22	81983	153(.19)	24(.03)	14366(17.52)	169(.21)
7	SP24	82045	113(.14)	11(.01)	14536(17.72)	679(.83)
	SP23	82254	126(.15)	16(.02)	13898(16.90)	285(.35)
	SP22	83230	148(.18)	16(.02)	14028(16.85)	93(.11)

8	SP24	82887	110(.13)	12(.01)	14370(17.34)	548(.66)
	SP23	83465	142(.17)	13(.02)	13467(16.13)	192(.23)
	SP22	84841	171(.20)	29(.03)	13916(16.40)	57(.07)

*The first number is the number of students and the number in parentheses is the percentage assigned.

TTS and TTS-All seem to be the most frequently assigned tools. For mathematics, assignment rates ranged from 16% to 20%. For ELA, assignment rates ranged from 12% to 16% for TTS and were approximately from 4% to 5% for TTS-All. By contrast, all the other accessibility resources seem to be rather infrequently assigned, with assignment rates less than 1%.

Table 69 and Table 70 represent the number and percentage of students who used the assigned tools in ELA and mathematics. The number in parentheses shows percentage use, which is calculated using the following formula:

$$\text{Percentage use} = \frac{\text{The number of students who used the tool}}{\text{The number of students who were assigned the tool}} * 100$$

Table 69: Number and Percentage Usage of Accessibility Resources (ELA)

Grade	Admin	Masking*	Print-On-Demand	Audio-Transcript	STT	TTS-All	TTS
3	SP24	11(34.38)	7(33.33)	6(20.69)	699(86.08)	3446(93.03)	9427(75.36)
	SP23	11(20.00)	5(45.45)	1(4.55)	382(86.23)	3415(93.92)	9739(79.72)
	SP22	8(25)	7(41.18)	0	134(95.04)	3087(96.20)	9323(80.89)
4	SP24	1(8.33)	5(19.23)	5(12.82)	917(81.44)	4287(93.81)	9792(75.68)
	SP23	30(43.48)	7(41.18)	0	418(82.77)	3730(93.98)	9757(79.47)
	SP22	15(39.47)	7(30.43)	3(12.00)	144(84.21)	3357(95.59)	9505(77.42)
5	SP24	4(12.50)	5(27.78)	2(5.88)	811(72.15)	4171(92.94)	8872(75.37)
	SP23	12(25.53)	13(65.00)	1(5.26)	356(73.86)	3635(92.97)	8882(77.56)
	SP22	22(44.00)	4(19.05)	2(7.14)	185(80.43)	3004(92.92)	8848(75.42)
6	SP24	38(37.62)	9(50.00)	1(5.26)	582(59.82)	3745(88.26)	7481(68.11)
	SP23	56(28.00)	11(50.00)	1(3.12)	301(70.33)	3236(87.01)	7728(71.88)
	SP22	49(32.89)	8(28.57)	1(3.85)	117(69.23)	3260(89.36)	7690(71.72)
7	SP24	35(30.17)	2(16.67)	3(8.57)	322(44.85)	3223(81.00)	6167(56.70)
	SP23	65(31.86)	2(11.11)	2(6.67)	159(52.65)	3013(81.28)	6814(62.67)
	SP22	32(21.19)	2(9.09)	3(9.09)	60(63.16)	2759(79.88)	6585(62.26)
8	SP24	24(21.62)	3(30.00)	4(14.81)	235(41.37)	2851(72.58)	5671(51.90)
	SP23	42(19.63)	5(27.78)	3(10.34)	110(54.73)	2542(75.39)	6106(56.51)
	SP22	41(23.43)	13(34.21)	1(4.76)	34(57.63)	2626(75.33)	5886(55.97)

*The first number is the number of assigned students using each tool and the number in parentheses is the percentage usage.

Table 70: Number and Percentage Usage of Accessibility Resources (Mathematics)

Grade	Admin	Masking*	Print-On-Demand	TTS	STT
3	SP24	5(20.00)	4(19.05)	11550(17.41)	552(69.52)
	SP23	7 (17.50)	2(16.67)	11710(74.00)	310(73.99)
	SP22	5(19.23)	5(22.73)	11293(76.51)	108(83.08)

4	SP24	1(7.14)	2(6.90)	11031(63.27)	651(60.96)
	SP23	3(6.12)	2(11.76)	10559(65.04)	322(66.12)
	SP22	2(7.41)	4(15.38)	10236(64.85)	117(69.23)
5	SP24	0	3(20.00)	9640(59.39)	603(55.47)
	SP23	8(25.00)	5(22.73)	9196(59.83)	277(60.61)
	SP22	6(12.50)	3(13.04)	8814(58.95)	150(65.22)
6	SP24	20(19.80)	4(21.05)	6822(45.61)	347(37.96)
	SP23	33(28.95)	4(22.22)	6857(47.45)	173(43.03)
	SP22	36(23.53)	7(29.17)	6880(47.89)	74(43.79)
7	SP24	31(27.43)	0	4386(30.17)	171(25.18)
	SP23	48(38.10)	1(6.25)	4322(31.10)	83(29.12)
	SP22	38(25.68)	1(6.25)	4601(32.80)	38(40.86)
8	SP24	20(18.18)	1(8.33)	3876(26.97)	108(19.71)
	SP23	46(32.39)	3(23.08)	3626(26.93)	45(23.44)
	SP22	38(22.22)	3(10.34)	3981(28.61)	22(38.60)

*The first number is the number of assigned students using each tool and the number in parentheses is the percentage usage.

STT, TTS, and TTS-All (ELA only) seem to show high usage rates, which seem to decrease as the grade level increases. For all other accessibility resources, the usage rates seem to range from moderately low to low.

To assess the degree to which students interact with accessibility resources in a consistent manner within a test, a test was divided into three equal portions. Tests were divided by the number of items, regardless of any built-in test segmentation.

Then, usage frequency was calculated to determine the number of items on which a given student used an accessibility resource out of the number of items on which the student could use that resource. Finally, the students' individual usage frequencies were averaged:

$$Usage\ Frequency = Avg\left(\frac{N\ items\ on\ which\ resource\ was\ used}{N\ items\ on\ which\ resource\ could\ be\ used}\right) * 100$$

For example, to calculate the usage frequency for TTS, students who were assigned TTS were identified. Each student's individual usage frequency was calculated. If a student used a resource on 3 out of 10 items, the usage frequency for that student would be 30 percent; then, that student's usage frequency was averaged with every other student assigned TTS.

Table 71 and Table 72 provide results for usage consistency in each portion of the test. For example, the usage frequency of TTS in Portion 1 of the grade 3 ELA test in spring 2022 was 7.16%. In other words, TTS was used, on average, on 7.16% of the items in Portion 1 of the grade 3 ELA test.

While STT and TTS seem to be used more on the third portion of the test, the other tools are consistently used throughout the test.

Table 71. Tool Usage Frequency in Each Portion of the Test (**ELA**)

Grade	Admin	Masking 1	2	3	Print-on-Demand 1	2	3	Audio Transcript 1	2	3	STT 1	2	3	TTS-All 1	2	3	TTS 1	2	3
G3	SP24	1.25	0.42	72.00	0.63	0.95	0.37	0	0.69	0	0.87	3.54	15.17	50.63	45.68	50.03	6.09	3.20	17.75
	SP23	0.97	0	0.77	2.42	1.21	2.98	0	0.30	0	0.89	3.67	17.24	47.70	42.40	47.94	5.23	2.58	18.30
	SP22	0.42	0.62	0.57	3.14	1.96	1.61	N/A	N/A	N/A	4.49	1.28	20.21	51.82	41.76	49.67	7.10	3.18	18.42
G4	SP24	0	0	0.60	0.77	0.77	0.30	0	0.51	0.39	0.73	3.23	12.73	48.73	41.82	46.71	4.44	1.83	14.09
	SP23	2.61	0.87	0.94	0.39	1.18	1.26	N/A	N/A	N/A	2.14	2.23	14.56	48.61	41.40	47.47	4.62	2.14	15.20
	SP22	3.33	0.88	1.32	0.29	0.29	0.62	0	0	1.71	3.47	1.25	18.06	53.14	40.04	48.99	5.63	1.93	12.05
G5	SP24	0.21	0	1.20	0.74	0.74	0.85	0	0.00	0	0.56	2.67	11.79	44.90	38.14	43.68	4.05	1.57	13.50
	SP23	0.85	0.28	0.88	3	4	7.83	0	0	0.40	1.70	1.67	12.85	42.69	38.16	44.26	3.94	2.46	15.23
	SP22	2.53	1.73	1.22	1.90	0.63	1.43	0	0	0.60	2.99	1.36	16.55	46.00	35.98	42.07	5.60	1.98	11.16
G6	SP24	1.72	1.06	2.36	1.85	2.22	3.42	0	0.00	0	1.53	0.76	8.78	34.86	27.09	32.48	3.07	0.99	9.51
	SP23	0.97	0.33	1.05	3.03	3.03	3.65	0	0	0	2.87	0.33	11.23	32.49	25.47	30.14	3.28	1.42	9.78
	SP22	1.52	0.31	1.99	0	0.71	1.59	0	0	0	2.56	0.87	14.23	31.64	23.53	30.78	4.25	1.22	9.11
G7	SP24	0.75	0.29	0.93	0	0	0.64	0	0.19	0	1.23	0.25	6.44	29.01	22.66	27.04	2.18	0.61	6.83
	SP23	0.98	0.49	1.62	0	0	0	0	0.22	0.30	1.90	0	8.49	26.41	18.83	24.69	2.21	0.65	7.56
	SP22	0.97	0.53	1.13	0	0	0	0	0	0	2.81	0.14	11	23.79	16.30	24.45	2.58	0.83	6.78
G8	SP24	0.51	0.45	0.83	0	0	0.77	0	1.32	0	1.24	0.34	5.12	23.93	19.06	21.27	1.84	0.58	5.43
	SP23	1.03	0.63	0.75	0	0	0.97	0	0	0.63	2.10	0	7.56	21.61	16.63	20.48	1.83	0.50	5.59
	SP22	0.61	0.20	0.83	1.13	0.94	1.40	0	0	0.79	2.06	0.48	10.03	19.89	14.19	18.92	2.00	0.64	4.84

Table 72. Tool Usage Frequency in Each Portion of the Test (Mathematics)

Grade	Admin	Masking 1	2	3	Print-on-Demand 1	2	3	TTS 1	2	3	STT 1	2	3
G3	SP24	0.67	0.22	0.47	0.53	0.53	0.28	28.29	22.14	23.61	0.08	0.09	9.83
	SP23	0.56	0.69	0.29	2.31	0.93	2.94	29.69	23.25	24.53	0.29	0.07	10.79
	SP22	0.21	0.64	0.68	1.26	0.51	0.80	26.64	23.92	23.76	0	0	12.58
G4	SP24	0.40	0	0	0	0.19	0.22	20.23	14.99	19.87	0.06	0.02	6.58
	SP23	0.23	0	0.13	0.33	0	0.37	20.54	15.08	19.54	0.08	0.02	7.38
	SP22	0	0.21	0.23	0.43	0.21	0.24	18.82	13.73	18.98	0.10	0.03	8.17

G5	SP24	0	0	0	1.11	1.48	0	19.23	12.52	14.58	2.06	0.03	6.13
	SP23	1.04	0.52	1.30	1.26	1.26	2.46	19.50	12.95	14.93	2.37	0.02	6.99
	SP22	0.46	0.12	0.37	0.24	0.24	0	16.60	10.35	13.36	2.46	0	6.16
G6	SP24	0.44	0.61	0.61	0.29	0	1.64	10.33	7.78	11.22	0.05	0.02	5.05
	SP23	0.54	0.54	1.33	0	0.62	1.67	10.34	8.08	11.02	0	0.03	5.81
	SP22	0.44	0.36	1.00	0	0.93	0.98	8.65	7.35	10.18	0	0.03	6.47
G7	SP24	1.08	0.44	0.78	0	0	0	6.70	5.46	5.42	1.33	0.61	0.38
	SP23	1.19	0.88	1.26	0	0.35	0	6.46	5.01	4.99	1.46	0.72	0.41
	SP22	0.71	0.60	0.79	0	0.35	0	6.60	5.36	5.15	2.93	0.24	0.63
G8	SP24	0.56	0.45	0.37	0	0	0.49	3.88	3.62	4.69	0.01	0	1.80
	SP23	1.13	0.63	0.62	0.43	0.43	0.45	3.88	3.12	4.25	0	0	2.25
	SP22	0.65	0.42	0.55	0.96	0	0.41	4.00	3.18	4.13	0	0.10	3.30

4. ITEM DEVELOPMENT AND TEST CONSTRUCTION

1.21 TEST DESIGN AND TEST SPECIFICATIONS

The IDOE sought the participation of Indiana educators in the development of ILEARN test specifications (test blueprints). The ILEARN assessments are designed to measure student achievement of the IAS. The IAS were designed and adopted to ensure that Indiana students graduate from high school ready to succeed in their college and career endeavors. To ensure that the ILEARN assessments provide a valid assessment of college-and-career-readiness, the test blueprints were constructed to ensure that the assessments represent the range of content defined in the IAS and result in accurate classification of student achievement as college-and-career-ready.

Indiana assessment forms were constructed using the ILEARN blueprints and item pools. The construction of test forms is a process that requires both judgement from content experts and psychometric criteria to ensure that certain technical characteristics of the test forms meet industry expected standards. The processes used for blueprint development and test form construction are described to support the claim that they are technically sound and consistent with expectations of current professional standards.

ILEARN is designed to support the claims described at the outset of this chapter.

1.21.1 ELA AND MATHEMATICS ITEM SPECIFICATIONS

CAI developed the IN ELA and mathematics item bank using a rigorous, structured process that engages stakeholders at critical junctures. This process is managed by CAI's Item Tracking System (ITS), which is an auditable content-development tool that enforces workflow and captures every change to, and comment about, each item. Reviewers, including internal CAI reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, with all accessibility features and tools.

The process begins with the definition of passage and item specifications, and continues with

- selection and training of item writers;
- writing and internal review of items;
- review by state personnel and stakeholder committees;
- markup for translation and accessibility features;
- field-testing; and
- post field-test reviews.

Each of these steps has a role in ensuring that the items can support the claims that will be based on them. Exhibit C describes how the steps contribute to these goals, and later sections of this report include detailed discussions of every step in the process.

Exhibit C: Summary of How Each Step of Development Supports the Validity of Claims

Development steps	Supports alignment to the standards	Reduces construct-irrelevant variance through universal design	Expands access through linguistic and other supports
Passage and item specifications	Specifies item types, content limits, and guidelines for meeting Depth of Knowledge (DOK) requirements and adjusting difficulty	Avoids the use of any item types with accessibility constraints and provides language guidelines; allows for multiple response modes to accommodate different styles	
Selection and training of item writers	Ensures that item writers have the background to understand the standards and specifications; teaches item writers about selection of item types for measurement and accessibility	Training in language accessibility, bias, and sensitivity, helping item writers to avoid unnecessary barriers	
Writing and internal review of items	Checks content and DOK alignment and evaluates and improves overall quality	Eliminates editorial issues and flags and removes bias and accessibility issues	
Markup for translation and accessibility features		Adds universal features, such as text-to-speech for mathematics, that reduce barriers	Adds text-to-speech, braille, American Sign Language (ASL), translations, and glossaries
Review by state personnel and stakeholder committees	Checks content and DOK alignment and evaluates and improves overall quality	Flags sensitivity issues	
Field testing	Provides statistical check on quality and flags issues	Flags items that appear to function differently for subsequent review for issues	May reveal usability or implementation issues with markup
Post field-test reviews	Provides final, more focused check on flagged items; rubric validation and rangefinding ensure that scoring reflects standards and expectations	Final, focused review on items flagged for differential item functioning	

1.21.1.1 Passage and Item Specifications

The Indiana Department of Education leveraged quality content from third-party item banks for use on *ILEARN* assessments. These item banks were accompanied by item

specifications which were utilized when alignment was confirmed by Indiana educators. The available specifications are described in Table 72 below.

Table 72: ILEARN Item Specifications

Specification	Developer	Content Areas Included
Indiana Item Specifications	Developed by Indiana for Indiana standards and define custom item development	Mathematics, English/language arts, science, social studies
Independent College and Career Ready (ICCR) Item Specifications*	Developed by Cambium Assessment, Inc (CAI) for their Independent College-and-Career-Ready item bank.	Mathematics, English/language arts, science
Smarter Balanced Item Specifications*	Developed by Smarter Balanced for their Smarter Balanced item bank.	Mathematics, English/language arts

Some third-party item specifications include content beyond the scope of the associated Indiana Academic Standards. For these specifications, only those portions which align to the Indiana Academic Standards are used for ILEARN assessments. Indiana educators approved alignment of items to each Indiana Academic Standard.

Smarter item and passage specifications were informed by best practices described in the Common Core State Standards (CCSS), the Smarter Content Specifications for ELA, and the practices prevalent in Smarter states' guidelines.

Independent College and Career Ready (ICCR) items and passage specifications were developed through a collaboration between content experts in one of CAI's partner states and CAI content experts. The specifications align to nationally recognized standards. Over time, the specifications have been expanded to reflect continuous improvement and the availability of new interaction types.

ILEARN item specifications (used for custom Indiana development) were developed by Indiana educators at a workshop in February 2018. They were further reviewed both by CAI test developers and IDOE content specialists, which resulted in minor updates and clarifications being made in 2020 and 2022.

In all cases, item and passage specifications ensure that items are written to the highest caliber and align to the standards being assessed.

1.21.1.2 Passage Specifications

ELA development begins with passage specifications. Detailed passage specifications ensure that all passages align to the correct grade level and provide sufficient complexity for close analytical reading. These specifications augment, rather than replace, quantitative syntactic measures such as Lexiles. The qualities called out in the specifications are derived from the ELA standards and accompanying material. The specifications help test developers create or select passages that will support a range of difficulty, furthering the goal of measuring the full range of performance found in the

population, but remaining on grade level. Appendix 4-A, ILEARN Passage Specifications, contains sample ILEARN passage specifications.

4.1.1.3 Item Specifications

Item specifications guided the item development process for Smarter, ICCR, and custom Indiana development.

Depending upon the source of the item, specifications in ELA may include any or all of the following.

- *Content Standard.* This identifies the standard being assessed.
- *Content Limits.* This section delineates the specific content that the standard measures and the parameters in which items must be developed to assess the standard accurately, including the lower and upper complexity limits of items.
- *Acceptable Response Mechanisms.* This section identifies the various ways in which students may respond to an item or prompt. Here, we note whether evidence-based selected-response (two-part items), extended response, hot text, multiple-choice, multiple select, and/or short answer (to be scored automatically with our *proposition scorer*) items may be used, and if so, how.
- *DOK Demands.* This section is broken into three subsections—DOK, task demand, and response mechanism. The task demands explain the skills the students may be required to demonstrate and connect these skills to the DOK. The task demands show how the DOK level requires higher-order thinking. Finally, the DOK and task demand are connected to appropriate response mechanisms used to assess these skills. All *ILEARN* item specifications have a standard-level DOK value.
- *Sample Items.* In this section, sample items present a range of response mechanisms and their corresponding expected difficulties (easy, medium, and hard). Notes delineating the cognitive demands of the item and an explanation of its difficulty level are detailed for each sample item.
- *Accessibility and Accommodation Considerations.* This section includes Allowable Tools (e.g., calculator), Literacy Considerations (e.g. glossary words), Visual and Auditory Considerations (including American Sign Language), and Linguistic Complexity.
- *Construct relevant vocabulary.* This section denotes the terms related to the skills and concepts of the standard that students are expected to understand and recognize with the items.

Table 73 is a sample of the item specifications that content experts, in collaboration with Indiana educators, developed for a grade 4 Reading: Vocabulary standard. It outlines the limits of the item content to fully address the standard. The acceptable response mechanisms that are recommended to assess this standard are noted. The DOK sections explain the demands for the DOK level and provide the acceptable response mechanisms. This level of detail provides the item writer with guidance when developing

items, ensuring that the items address the standard and are correctly aligned at the DOK and difficulty levels.

Additionally, accessibility and linguistic complexity considerations are provided for item writers. Item writers consider how each item will be rendered or adapted to reach the largest number of students possible without violating the construct. Specifically, this section of the item specifications includes Literacy Considerations (e.g., glossary words), Visual and Auditory Considerations (including American Sign Language), and Linguistic Complexity.

Table 73: Sample ELA Item Specification for Grade 4

Content Standard	**4.RV.2.2:** Identify relationships among words, including more complex homographs, homonyms, synonyms, antonyms, and multiple meanings.
Content Limits	Items should ask students not to define the type of word that is being used but rather to demonstrate its meaning between the words. Items may refer only to synonym and antonym in the stimuli. All words should be provided with sufficient context for support.
Construct-Relevant Vocabulary	antonyms, meaning, opposite, phrase, relationship, replace, similar/same as, synonyms,
Recommended Response Mechanisms (Item Types)	Drag and Drop Evidence-Based Selected Response Hot Text Multiple Choice Multi-Select
DOK	2

Evidence Statements
Students replace a given word with synonyms, antonyms, homographs, homonyms, and multiple-meaning words.
Students use context to determine or support meaning.
Students identify a word, sentence, or phrase that uses a given word in the same way.
(NOTE: Level of difficulty will depend on subtlety/amount of text and/or complexity of interpretation required.)

Sample Item
Why is "[word X]" a better word to use from paragraph 4 than "[word Y]"? A.   [Word X] suggests [something more formal] B.   [Word X] suggests [something more precise] C.   [Word X] suggests [something more aligned to the tone] D.   [Word X] suggests [something more audience appropriate]

Literacy Considerations	Word List: Content can select construct-irrelevant words for glossing, which gives students access to the definition and an audio clip of those words. Considerations will include the question/task, standard, and construct-relevant words necessary for the item.

Visual and Auditory Considerations (NOTE: These considerations generally refer to the passage/media source rather than the item.)	American Sign Language: Allows a student to see a video of an ASL interpreter. This option will be included only if the media contains audio. Audio Transcriptions: Written transcripts of audio for students of varying auditory and visual abilities can be provided as needed. The same transcripts will be used for ASL videos. Closed Captioning: Captions media so that audio is available for students who are hearing impaired. Can be used for both audio-only and video media. Graphics: Graphics will be provided in formats that are accessible to students with varying abilities, including students who are blind or visually impaired. Graphics should contain only content that will help students understand or process information; those that do not contribute to the student's understanding should not be included. Graphics should be braillable whenever possible; those that cannot be brailled will be provided to blind/visually impaired students through a verbal or written description.
Linguistic Complexity	Rating to be completed after all final edits have been applied and approved by IDOE.

Similar to ELA, mathematics, science, and social studies item specifications may include any or all of the following information.

- *Content Limits*. This section delineates the specific content measured by the standard and the extent to which the content is different across grade levels. In mathematics, for example, content limits can include acceptable denominators, number of place values for rounding or computation, acceptable shapes for geometry standards, etc.
- *Acceptable Response Mechanisms*. This section identifies the various ways in which students may respond to a prompt, such as multiple-choice, graphic response, proposition response, equation response, and multi-select items. The identified acceptable response mechanisms were identified with accessibility concerns taken into consideration. For example, a graphic response item should only be used when the standard or task demand requires a graphic representation (e.g., graphing a system of equations). Other items, such as multiple-choice, can still be used with static images that can be used for all student populations.
- *Depth of Knowledge (DOK)*. The task demands of each standard can be classified as DOK 1, DOK 2, or DOK 3.
- *Task Demands*. In this section, the standards are broken down into specific task demands aligned to each standard. Task demands denote the specific ways in which students will provide evidence of their understanding of the concept or skill. In addition, each task demand is assigned appropriate response mechanisms, DOK, and PCs specifically relevant to that particular task demand.
- *Examples and Sample Items*. In this section, sample items are delineated along with their corresponding expected difficulties (easy, medium, and difficult). Notes

for modifying the difficulty of each task demand are detailed with suggestions for the item writer. The suggestions for adapting the difficulty based on the task demands are research based and have been reviewed by both content experts and a cognitive psychologist.

1.21.2 SCIENCE CLUSTERS

The cluster-based science assessments were first administered in grades 4, 6, and High School in spring 2024.

CAI developed a shared science assessment item bank in collaboration with the states that were part of the Memorandum of Understanding (MOU) using a rigorous, structured process that engaged stakeholders at critical junctures. The items in the bank are linked to the Next Generation Science Standards, which participating states all use.

A performance expectation is a point in a three-dimensional space formed by three dimensions of science learning: crosscutting concepts (CCCs), science and engineering practices (SEPs), and disciplinary core ideas (DCIs). That is, a performance expectation (PE) is characterized by a specific CCC, SEP, and DCI. When the MOU states first convened, many sessions were spent discussing how to assess these new three-dimensional standards. These group sessions are where the idea of an item cluster was conceived. An item cluster consists of a stimulus (scientific phenomenon) associated with multiple parts. Each of these parts contains questions that allow the student to explore the phenomenon. Each of the parts assesses at least two dimensions, and the entire item or cluster assesses a student on all three dimensions for a specific PE. Exhibit D is a visual representation of the structure of a three-dimensional cluster.

Exhibit D: Structure of Three-Dimensional Item Clusters
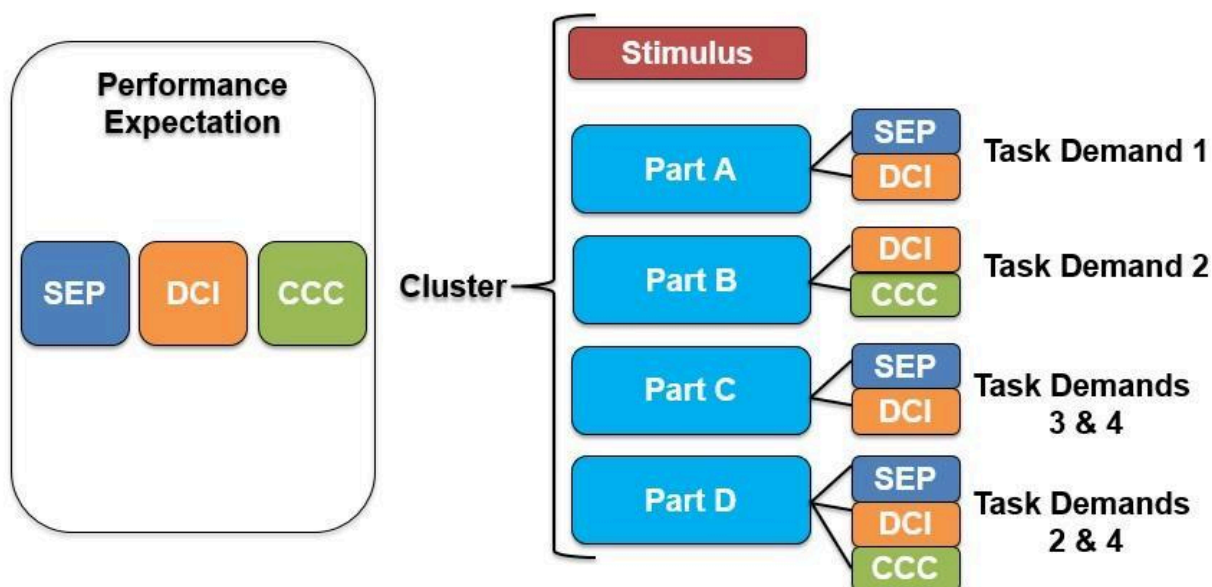
Each part of an item cluster contains questions that require the student to interact with the item cluster. There are many different interactions that can be included in a cluster. Appendix 4-N provides an overview of the different interaction types. The interactions used in an item cluster are chosen intentionally to best assess different aspects of the three-dimensional construct.

Exhibit E provides an example of an item cluster that has a phenomenon, five parts, and eight interactions; each part of an item cluster assesses multiple dimensions.

Exhibit E: Example of an NGSS Item Cluster

A student rings a doorbell. When the person inside the house is on the main floor, he can easily hear the doorbell. When he is upstairs, though, he cannot so easily hear the doorbell.

Figure 1 shows the circuit of a simple doorbell when it is on (pressed) and off (not pressed).

Figure 1. Simple Doorbell Circuit
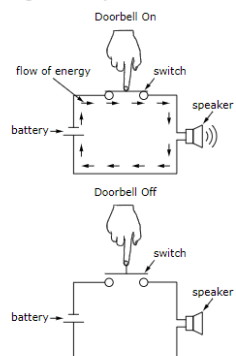
Table 1 shows the types of doorbell speakers available and their cost, in dollars ($).

Table 1. Types of Speakers and Cost

Speaker	Cost ($)
Bell	11
Buzzer	17
Chimes	25

Table 2 shows the types of batteries available based on their voltage (V), the amount of power each produces, and their cost.

Table 2. Types of Batteries and Cost

Battery (V)	Amount of Power	Cost ($)
12	A lot	27
9	Average	3
1.5	A little	1

Table 3 shows the types of switches and their cost.

Table 3. Types of Switches and Cost

Switches	Cost ($)
Rectangular	4
Circular	5
Lighted	11

Your Task

In the questions that follow, you will design a main-floor doorbell that can be heard from upstairs in a house.

Part A

Click on each blank box and select a phrase to describe what is happening to the energy at each part of the circuit when the doorbell is turned on.

Parts	Energy Pathway when Doorbell Is on
Battery	Energy is stored.
Wires	Energy is transferred.
Speaker	Electrical energy is converted to sound energy.

Part B

Use the simulation to select the materials necessary to conduct fair experiments and create a doorbell that can be heard from upstairs and costs less than $40. The student can only hear a doorbell from upstairs if it is loud or very loud.

- Select the speaker, battery, and switch to determine the overall cost and loudness of the doorbell.
- Then click Run Trial.
- The cost of wire has already been included in the total cost.
- You must complete **two** trials.
- You may run up to **five** trials.
- Click the trash can icon if you want to delete a trial and generate new data.
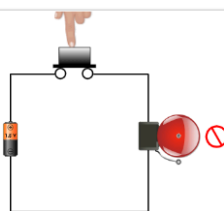
Speaker [Bell]
Battery [1.5]
Switch [Rectangular]

[Run Trial]

Trial	Speaker	Battery (V)	Switch	Loudness	Cost ($)	
1	Bell	9.0	Rectangular	Loud	18	🗑
2	Bell	12.0	Rectangular	Very Loud	42	🗑
3	Bell	1.5	Rectangular	No Sound	16	🗑
4	Chimes	9.0	Lighted	Quiet	39	🗑
5	Bell	9.0	Lighted	Loud	25	🗑

Part C

Select **all** of the trials that meet the criteria for being heard upstairs and cost less than $40.

- ☑ Trial 1
- ☐ Trial 2
- ☐ Trial 3
- ☐ Trial 4
- ☑ Trial 5
- ☐ None

Part D

Click on the blank boxes and select words or phrases to predict what will happen to the loudness of the doorbell when the battery power increases.

The loudness of the doorbell will [increase] because [more energy is stored in the battery.]

Part E

Select **two** trials that support the relationship between the loudness of the doorbell and the power of the battery.

- ☐ Trial 1
- ☑ Trial 2
- ☑ Trial 3
- ☐ Trial 4
- ☐ Trial 5
- ☐ Cannot be determined

This item cluster is aligned to the NGSS PE of 4-PS3-4: Apply scientific ideas to design, test, and refine a device that converts energy from one form to another. The PE uses the following three elements of the three-dimensional standards: (1) Constructing Explanations and Designing Solutions (i.e., SEP), (2) Conservation of Energy and Energy Transfer (i.e., DCI), and (3) Energy and Matter (i.e., CCC).

Part A requires students to demonstrate their knowledge of how energy is stored, transferred, or used within the system. In this item cluster, they must know how a battery, wires, and a speaker work within the circuit. This aligns with the DCI and the CCC.

Part B requires students to design and test designs that use electricity to produce a sound. This aligns with the DCI (how changes in current influence the production of sound) and the SEP (designing and testing solutions to a design problem).

Part C requires students to compare their designs with some criteria and constraints. This aligns with the SEP (designing and testing solutions) and the CCC (energy can be transferred in various ways and between objects). The answer for Part C is directly determined by how the student completes Part B. If all the trials the student runs in Part B meet the given criteria, then all of those must be selected to be considered as correct in Part C. Therefore, there are multiple different ways to get this item correct.

Part D requires students to make a prediction from the evidence that they generated in Part B. This part is aligned to all three dimensions. The student has used their designs and information (representing SEP) from Part B to show how energy is transferred between objects (representing CCC) and specifically how increasing the current changes the volume (representing DCI).

Like Part C, Part E is dependent on Part B. The students are determining which trials support the prediction that they made in Part D. This part, combined with Part D and Part B, address all three dimensions of the PE.

The next big challenge for the MOU states was to properly score these item clusters so that all evidence of understanding the PEs and three dimensions could be collected. It was determined that scoring assertions would be the best way to capture and score student responses on item clusters. Scoring assertions are evidence statements that relate specific features from the student response to skills and knowledge being tested (of which they provide evidence). The use of these assertions in scoring creates a direct linkage between what the student does and the inferences about the skills and knowledge that the student's response supports. This approach provides a physical embodiment of evidence-centered design, Mislevy and Haertel's well-regarded approach to cognitive measurement (Mislevy & Haertel, 2006). This also provides a structure for ensuring and reviewing alignment during test development and a clear explanation of what was measured, how it was measured, and why it was measured when tests are scored and reported.

By inspecting the student response for every meaningful piece of student input, more information about student skills and knowledge can be harvested than in a single interaction. In fact, evidence for some scoring assertions may derive from two or more interactions within an item cluster. This may happen if one interaction is dependent on another interaction, allowing for multiple solution paths. This is one of the primary reasons that scoring assertions within item clusters can show deeper cognitive understanding and higher-order thinking that is required of the three-dimensional science standards.

Each of the parts in an item cluster likely has one or more scoring assertions where student skills and knowledge are being collected. The scoring mechanism has the capability to focus on one interaction, one part, or across multiple interactions and parts as determined by the item writers, subject-matter expert (SME) reviewers, and performance expectations. All permutations and combinations of measurable moments can be captured with scoring assertions.

The example item cluster from this section has seven assertions. Each scoring assertion is described in detail in Exhibit F.

Exhibit F: Example of NGSS Scoring Assertions

Your response earned **7** points of a possible **7**

Score Rationale	
When asked to describe what is happening to the energy for the battery when the doorbell is turned on, the student selected "energy is stored" or "energy is transferred." This provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and /or motion.	✔
When asked to describe what is happening to the energy of the wires when the doorbell is turned on, the student selected "energy is transferred." This provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and /or motion.	✔
When asked to describe what is happening to the energy of the speaker when the doorbell is turned on, the student selected "electrical energy is converted to sound energy." This provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and /or motion.	✔
The student ran at least two trials and ran at least one trial in which they selected components of a doorbell that produced "Loud" or "Very Loud" sound and that included components that cost less than $40. This provides some evidence of an ability to select characteristics to be manipulated while gathering information to determine the loudest, cost-effective doorbell.	✔
When asked to select the trial that met the criteria for being heard upstairs and cost less than $40, the student selected all trials from their simulation that produced "Loud" or "Very Loud" sound and cost less than $40. This provides some evidence of an ability to use given information to design and test a device that converts energy from one form to another.	✔
When asked to predict what will happen to the sound of the doorbell if the battery power increases, the student selected "The loudness of the doorbell will increase because more energy is stored in the battery." This provides some evidence of an ability to use an explanation to predict how the sound of an object changes, given a change in the conversion of stored energy.	✔
When asked to select the trials that support the relationship between the loudness of the doorbell and the power of the battery, the student selected two trials from the simulation in which the loudness was higher for the trial with a battery with more power. This provides some evidence of an ability to use evidence to support an inference.	✔

Assertion texts like the one shown in Exhibit F are written for every assertion in every item. They describe the correct response and what evidence should be provided by the student's response.

In the example item cluster, Part A has three assertions. Each one "provides some evidence of an ability to complete a causal chain explaining how energy can be transferred via electric current to produce light, sound, heat, and/or motion." The student must know something about electrical energy (DCI) and how it is transferred or used

(DCI and CCC) to correctly respond. One assertion corresponds to each row in the table (i.e., one for Battery, one for Wires, and one for Speaker).

Part B has two assertions. The first "provides some evidence of an ability to select characteristics to be manipulated while gathering information to determine the loudest, most cost-effective doorbell." The second assertion "provides some evidence of an ability to use given information to design and test a device that converts energy from one form to another." The student must use their knowledge of how electrical energy is used and transferred (DCI) and how to design and test a design of a device using electricity (SEP) to correctly interact with Part B.

Part C has one assertion, as the student's selections are not independent of each other. The assertion "provides some evidence of an ability to use given information to design and test a device that converts energy from one form to another." The student must be able use generated evidence to support a design decision (SEP) about the transfer of energy (CCC). This assertion is pulling responses from both Parts B and C. This is precisely how item clusters and assertions can assess multiple dimensions and higher levels of complexity, as students are running their own experiments and analyzing the outcomes, no matter what those outcomes are.

Part D has one assertion. The assertion "provides some evidence of an ability to use an explanation to predict how the sounds of an object changes, given a change in the conversion of stored energy." This shows how the student must use elements from all three dimensions to respond correctly to this assertion. The student uses data from their generated designs and makes a prediction using that data to support their knowledge of energy and energy transformations.

Part E also has one assertion. The assertion "provides some evidence of an ability to use evidence to support an inference." In this case, it is an inference about the relationship between the available battery power and the loudness of the bell. Again, this scoring assertion is pulling information from three different parts (Parts B, D, and E).

While each part of the item, each interaction within the item, or each assertion may not be three-dimensional, the item cluster as a whole represents all three dimensions. It also provides an organized flow of cognition from scaffolding (Part A), through the engineering process (Parts B and C), to a conclusion and evidentiary support of the conclusion (Parts D and E).

The assertion text explains how a student responded to a given task and what that task shows evidence of. This allows us to ensure that items allow each student an opportunity to show what they know and what their knowledge, skills, and abilities show about their understanding of science and engineering.

Once the item cluster, along with interactions and scoring assertions, came to fruition, CAI and the group of states were able to begin item and test development in earnest.

The item development process was managed by CAI's Item Tracking System (ITS), which is an auditable content-development tool that enforces rigorous workflow and

captures all changes made to and comments associated with each item. Reviewers, including internal CAI reviewers or stakeholders in committee meetings, can review items in ITS as they will appear to the student, with all accessibility features and tools.

The process begins with the definition of item specifications and continues with

- selection and training of item writers;
- writing and internal review of items;
- review by state personnel and stakeholder committees;
- markup for translation and accessibility features;
- field testing; and
- post-field-test reviews.

Each of these steps has a role in ensuring that the items can support the claims on which they will be based. Exhibit G describes how each step contributes to these goals. Each step in the process is discussed in more detail below.

Exhibit G: Summary of How Each Step of Development Supports the Validity of Claims

	Supports Alignment to the Standards	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Item Specifications	Specifies item interactions, content limits, and guidelines for meeting task demands and levels of cognitive engagement requirements and adjusting difficulty.	Avoids the use of any item interactions with accessibility constraints and provides language guidelines. Allows for multiple response modes to accommodate different styles.	
Selection and Training of Item Writers	Ensures that teachers who are writing items have the background to understand the standards and specifications. Teaches item writers about selection of item interactions for measurement and accessibility.	Training in language accessibility, bias, and sensitivity helps item writers avoid unnecessary barriers.	
Writing and Internal Review of Items	Checks content alignment and evaluates and improves overall quality.	Eliminates editorial issues and flags and removes bias and accessibility issues.	
Markup that Prepares Items for Translation and Accessibility Features		Adds universal features, such as text-to-speech (TTS), for science that reduce barriers.	Adds TTS, braille, ASL, translations, and glossaries.
Review by State Personnel and Stakeholder Committees	Checks content and cognitive complexity alignment; evaluates and improves overall quality.	Flags sensitivity issues.	
Field Testing	Provides statistical checks on quality and flags issues.	Flags items that appear to function differently for subsequent review for issues.	May reveal usability or implementation issues with markup.

	Supports Alignment to the Standards	Reduces Construct-Irrelevant Variance Through Universal Design	Expands Access Through Linguistic and Other Supports
Post-Field-Test Reviews	Final, more focused check on flagged items. Rubric validation ensures that scoring reflects standards.	Final, focused review on items flagged for differential item functioning (DIF).	

1.21.2.1.1 *Science Cluster Item Specifications*

CAI worked with a group of states, psychometricians, and science experts, including the authors of the Next Generation Science Standards (NGSS), to develop powerful innovative solutions to the challenges of measuring three-dimensional science standards based on the National Research Council's A Framework for K–12 Science Education (2012). Participating states included Connecticut, Hawaii, Idaho, Montana, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. New Hampshire, North Dakota, and South Dakota participated in some activities. This collaboration yielded item specifications for performance expectations (PEs), sample item clusters for some specifications, and hundreds of science item clusters and stand-alone items in various stages of development. Under this collaboration, utilizing guidelines for item specifications proposed by WestEd in collaboration with the Council of Chief State School Officers (CSSO), state members, and content experts (CCSSO, 2015), states developed item specifications jointly. These item specifications were also reviewed and approved by state educators to ensure adherence to NGSS.

Item specifications are documents designed to guide item writers as they craft test questions and stakeholders as they review those items. These specifications are intended to serve writers as a roadmap to facilitate the creation of items that are properly aligned to the three dimensions comprising each science standard and that together form coherent item clusters. Exhibit H provides a sample of the item specifications developed by content experts for a middle school standard. Item specifications in science include the following:

- **Standard.** This identifies the NGSS Performance Expectation being assessed.
- **Dimensions.** This identifies the Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs) that the standard assesses.
- **Clarifications and Content Limits.** This delineates the specific content that the standard measures and the parameters in which items must be developed to assess the standard accurately, including the lower and upper complexity limits of items. Specifically, content limits refine the intent of the standard and provide limits of what may be asked of test takers. For example, content limits may identify the specific formulae that students are expected to know or not know.
- **Science Vocabulary.** This section identifies the relevant technical words that students are expected to know, and related words that they are explicitly not

expected to know. These categories should not be considered exhaustive, as the boundaries of relevance are ambiguous, and the list is limited by the imagination of the writers.

- **Content/Phenomena.** This section provides examples of the types of phenomena that would support the effective items related to the standard in question. In general, these are guideposts, and item writers seek comparable phenomena, rather than drawing on those within the documents.
- **Task Demands.** In this section, the standard and associated evidence statements are broken down into specific task demands aligned to each standard. Task demands denote the specific ways in which students will provide evidence of their understanding of the concept or skill. Specifically, the task demands identify the types of interactions and activities that item writers should employ. Each item should be clearly linked to one or more of the task demands, and the verbs guide the types of interactions writers might employ to elicit the student response.

Exhibit H: Sample Science Item Cluster Specifications for a Middle School Standard

Standard	6.1.2 **Develop and use a model** to describe the role of gravity and inertia in orbital motions of objects in our solar system.		
Dimensions	**Developing and Using Models** ● Develop and use a model to describe phenomena.	**ESS1.A: The Universe and Its Stars** ● The Earth and its solar system are part of the Milky Way galaxy, which is one of many galaxies in the universe. **ESS1.B: Earth and the Solar System** ● The solar system consists of the Sun and a collection of objects, including planets, their moons, and asteroids that are held in orbit around the Sun by its gravitational pull on them. ● The solar system appears to have formed from a disk of dust and gas, drawn together by gravity.	**Systems and System Models** ● Models can be used to represent systems and the interactions in a system.
Clarifications and Content Limits	**Assessment Clarifications** ● Emphasis is on understanding that inertia and gravity work together to keep the objects of the Solar System (the planets, the moons, the space station, and satellites) in orbit. The emphasis is on conceptual understanding that inertia is a property that works with gravity to keep objects in orbit. The concept of, and the term *balance,* is included in this definition. ● Understanding that gravity is a force and is a function of mass and distance. ● Emphasis is on knowing the mass of an object and not the concept of weight, which is a force. At this grade level, those terms can be used interchangeably. **Assessment Content Limits** ● <u>Students do not need to know</u>: The mathematical formula for calculating force, inertia, gravity, or Kepler's law, or how to calculate trajectories or perform any computational analysis.		

Terms That Do Not Need Definition	inertia, gravity, force, mass, orbit, Earth, moon, names of planets
Terms That MUST Be Defined	perihelion, aphelion, names of specific moons, names of space shuttles, moment of inertia, Kepler's laws of planetary motion, black hole, specific facts on any planets or moons, computational analysis on any relative motions
Phenomena	
Context/ Phenomena	Example phenomena for 6.1.2: • Satellites orbit Earth but can fall out of orbit (Skylab, UART satellite). • Halley's Comet can be seen as it travels past Earth every 75–76 years. • Rings are present around some planets. • Mars has two moons at different distances from the planet, which orbit the planet at different speeds. • Objects that are very distant can still be held in orbit around the Sun. • A belt of rocks and gases circles the Sun between Mars and Jupiter.
Task Demands	
1.	Identify from a collection, including distractors, the components of a model that include depictions of celestial bodies and/or man-made objects and the forces among them.
2.	Assemble or complete, from a collection of potential model components, an illustration, diagram, or description that is capable of representing forces and their influences on the motion of celestial bodies and/or man-made objects in orbit. This <u>does not</u> include the simple labeling of an existing diagram.
3.	Make predictions about the effects of changes in mass/distance/how fast an object travels in a given model on other objects in the system. Predictions can be based on manipulating model components, completing illustrations, or selecting from a list including distractors.
4.	Summarize data or evidence to highlight trends, patterns, or correlations.
5.	Describe, select, or identify the relationships among components of a model that describe the role of gravity and/or inertia in orbital motions, or explains how gravity and/or inertia affect the orbital motion of objects in our solar system.

The specifications help test developers create item clusters that will support a range of difficulties, furthering the goal of measuring the full range of performance found in the population, but remaining at grade level.

1.21.3 TARGET BLUEPRINTS

1.21.3.1 Summative Target Blueprints

Blueprints specify a range of items to be administered in each reporting category (or strand). 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. Allowing a range in the number of required items allows the computer-adaptive testing (CAT) algorithm the flexibility to select items that balance matching items to the ability of the student while matching the blueprints.

To ensure that the CATs accurately reflect the content of the curriculum standards, best practice requires that at least 50% of the standards for each reporting category be assessed on each test. In the aggregate, however, all the standards are assessed. Providing the student performance on all standards at an aggregate level is very

beneficial for instructional purposes. The blueprints require a minimum of eight points for each reporting category.

Table 73 through Table 76 present the summative test blueprint requirements specified in the Test Delivery System (TDS) for the 2023–2024 school year. Each test must include items within the range of the minimum and maximum number of items for the total test and for the score-reporting categories.

<div align="center">

Table 73: Minimum/Maximum Percentages of Test Items by Score-Reporting Category for Summative ELA

</div>

Reporting Category	Min	Max
Grade 3 ELA (34–36 scored items)		
Key Ideas and Textual Support/ Vocabulary	33%	44%
Structural Elements and Organization/Connection of Ideas/ Media Literacy	28%	35%
Writing*	33%	41%
Speaking and Listening	6%	9%
Reading Foundations	0%	6%
Grade 4 ELA (34–36 scored items)		
Key Ideas and Textual Support/ Vocabulary	31%	41%
Structural Elements and Organization/Connection of Ideas/ Media Literacy	31%	41%
Writing*	31%	41%
Speaking and Listening	6%	9%
Grade 5 ELA (34–36 scored items)		
Key Ideas and Textual Support/ Vocabulary	31%	41%
Structural Elements and Organization/Connection of Ideas/ Media Literacy	31%	41%
Writing*	31%	41%
Speaking and Listening	6%	9%
Grade 6 ELA (33–35 scored items)		
Key Ideas and Textual Support/ Vocabulary	29%	39%
Structural Elements and Organization/Connection of Ideas/ Media Literacy	29%	39%
Writing*	34%	42%
Speaking and Listening	6%	9%
Grade 7 ELA (33–35 scored items)		
Key Ideas and Textual Support/ Vocabulary	29%	39%
Structural Elements and Organization/Connection of Ideas/ Media Literacy	29%	39%
Writing*	34%	42%
Speaking and Listening	6%	9%
Grade 8 ELA (33–35 scored items)		

Reporting Category	Min	Max
Key Ideas and Textual Support/ Vocabulary	29%	36%
Structural Elements and Organization/Connection of Ideas/ Media Literacy	29%	36%
Writing*	34%	42%
Speaking and Listening	6%	9%

* Each student receives one writing prompt (argumentative, informative, or narrative) at every grade

Table 74: Minimum/Maximum Percentages of Test Items by Score-Reporting Category for Summative Mathematics

Reporting Category	Min	Max
Grade 3 Mathematics (46–48 scored items)		
Algebraic Thinking and Data Analysis	19%	24%
Computation	23%	28%
Geometry and Measurement	19%	24%
Number Sense	23%	28%
Process Standards	8%	13%
Grade 4 Mathematics (46–48 scored items)		
Algebraic Thinking and Data Analysis	19%	24%
Computation	23%	28%
Geometry and Measurement	19%	24%
Number Sense	23%	28%
Process Standards	8%	13%
Grade 5 Mathematics (47–49 scored items)		
Algebraic Thinking	20%	26%
Computation	22%	28%
Geometry and Measurement, Data Analysis, and Statistics	18%	23%
Number Sense	22%	28%
Process Standards	8%	13%
Grade 6 Mathematics (46–48 scored items)		
Algebra and Functions	23%	28%
Computation	21%	26%
Geometry and Measurement, Data Analysis, and Statistics	19%	24%
Number Sense	21%	26%
Process Standards	8%	13%
Grade 7 Mathematics (46–48 scored items)		
Algebra and Functions	23%	28%
Data Analysis, Statistics, and Probability	21%	26%
Geometry and Measurement	21%	26%

Reporting Category	Min	Max
Number Sense and Computation	23%	28%
Process Standards	8%	13%
Grade 8 Mathematics (46–48 scored items)		
Algebra and Functions	23%	28%
Data Analysis, Statistics, and Probability	21%	26%
Geometry and Measurement	21%	26%
Number Sense and Computation	19%	24%
Process Standards	8%	13%

Table 75: Number of Test Items by Score-Reporting Category for Summative Science

Reporting Category	Clusters	Stand-Alone Items
Grade 4 Science (23 scored items)		
Physical Science	2	24
Life Science	2	2
Earth and Space Science	2	3
Computer Science	0	8
Grade 6 Science (22 scored items)		
Physical Science	2	2
Life Science	2	4
Earth and Space Science	2	2
Computer Science	0	8
Biology (18 scored items)		
From Molecules to Organisms: Structure and Function	2	4
Ecosystems: Interactions, Energy, and Dynamics	2	4
Heredity and Evolution	2	4

Table 76: Minimum/Maximum Percentages of Test Items by Score-Reporting Category for Summative Social Studies

Reporting Category	Min	Max
Grade 5 Social Studies (40 scored items)		
Civics and Government	38%	43%
Geography and Economics	28%	33%
History	28%	33%
U.S. Government (54 scored items)		
Functions of Government	35%	39%
Historical Foundations of American Government	24%	28%
Institutions and Processes of Government	35%	39%

1.21.3.2English Language Arts Score-Reporting Categories

The ILEARN ELA assessments measure students' understanding of the standards at the end of grades 3, 4, 5, 6, 7, and 8. These assessments measure students' proficiency in ELA knowledge and skills. ILEARN individual student reports describe "at or near" proficient ELA performance in the following reporting categories:

Grade 3

- ***Key Ideas and Textual Support/Vocabulary.*** Your student can often answer questions independently about literary and nonfiction texts. He or she can find the theme/main idea of a text and use key details to support it, describe characters and relationships, and find the meanings of words/phrases using textual clues.
- ***Structural Elements and Organization/Connection of Ideas/Media Literacy.*** Your student can often independently plan writing to tell a story, express opinions, or convey information; can organize facts/information into categories; can use correct grammar in simple and complex sentences; and can correctly spell high-frequency and studied words.
- ***Writing.*** Your student can often independently plan writing to tell a story, express opinions, or convey information; can organize facts/information into categories; can use correct grammar in simple and complex sentences; and can correctly spell high-frequency and studied words.

Grade 4

- ***Key Ideas and Textual Support/Vocabulary.*** Your student can often independently interact with literary or nonfiction texts. He or she can determine the main idea and how it is supported by key details; describe how characters/setting affect plot; paraphrase key events; and determine the meaning of unknown words.
- ***Structural Elements and Organization/Connection of Ideas/Media Literacy.*** Your student can often independently recognize claims in media sources; compare points of view, themes, and different accounts; distinguish between fact and opinion; understand text structures and features; and combine information from two texts to demonstrate understanding.
- ***Writing.*** Your student can often independently organize and develop writing for persuasive, informative, and narrative purposes, introducing a topic to an audience, and using facts and examples to support ideas. He or she often uses appropriate word choice and punctuation.

Grade 5

- ***Key Ideas and Textual Support/Vocabulary.*** Your student can often independently interact with literary or nonfiction texts. He or she quotes evidence to support inferences, determines main ideas and key events, describes multiple

characters and settings, and determines how simple figurative language adds meaning.

- ***Structural Elements and Organization/Connection of Ideas/Media Literacy.*** Your student can often independently explain reasoning used to support claims in different media, describe various viewpoints and how they influence information, describe a text's overall structure, and compare stories in the same genre on their approaches to similar themes.
- ***Writing.*** Your student can often independently organize and develop writing for persuasive, informative, and narrative purposes; introduce a topic; and use facts and examples to support ideas. He or she often uses appropriate word choice, sentence structure, and punctuation.

Grade 6

- ***Key Ideas and Textual Support/Vocabulary.*** Your student can often independently interact with literary, informational, historical, and scientific texts. He or she makes inferences, explains central ideas and how plots unfold and characters change, cites details, and determines the meaning and impact of words.
- ***Structural Elements and Organization/Connection of Ideas/Media Literacy.*** Your student can often independently explain how authors structure information, develop points of view, and support ideas with details. He or she can compare how literary and nonfiction texts from different sources, genres, or media approach similar themes and topics.
- ***Writing.*** Your student can often independently organize and develop writing for argumentative, informative, and narrative purposes, using evidence or details to support ideas. He or she often uses appropriate word choice, sentence structure, and punctuation.

Grade 7

- ***Key Ideas and Textual Support/Vocabulary.*** Your student can often independently interact with literary, informational, historical, and scientific texts. He or she explains multiple central ideas and how plot/characters/setting interact, cites several details to support inferences, and analyzes word impact on meaning.
- ***Structural Elements and Organization/Connection of Ideas/Media Literacy.*** Your student can often independently describe a text's structure, compare various points of view, and trace an argument and its support. He or she can compare and contrast the information presented by different authors or in different media formats.
- ***Writing.*** Your student can often independently organize and develop writing for argumentative, informative, and narrative purposes; introduce claims and acknowledge opposing views; choose evidence to support ideas; and use appropriate word choice, sentence structure, and punctuation.

Grade 8

- ***Key Ideas and Textual Support/Vocabulary.*** Your student can often independently interact with literary, informational, historical, and scientific texts to explain how central ideas develop, describe how dialogue affects plot and characters, cite strong and relevant evidence, and interpret figures of speech.
- ***Structural Elements and Organization/Connection of Ideas/Media Literacy.*** Your student can often independently compare structures in related texts, describe points of view/cultural experiences, and distinguish authors' perspectives, purposes, and positions. He or she can identify and describe persuasive techniques used by different media.
- ***Writing.*** Your student can often independently organize and develop writing for argumentative, informative, and narrative purposes; clearly distinguish a topic/claim; support ideas with relevant details; use transitions to clarify ideas; establish style; and use correct punctuation.

1.21.3.3 Mathematics Score-Reporting Categories

The ILEARN mathematics assessments measure students' understanding of the standards at the end of grades 3, 4, 5, 6, 7, and 8. These assessments measure students' proficiency in mathematical knowledge and skills and whether they are adept in demonstrating the process standards. ILEARN individual student reports describe "at or near" proficient mathematics performance in the following reporting categories:

Grade 3

- ***Algebraic Thinking and Data Analysis.*** Your student can often independently represent and interpret data, interpret a multiplication equation as equal groups, and solve one- and two-step real-world problems that involve whole numbers with all four operations.
- ***Computation.*** Your student can often independently and fluently add and subtract whole numbers up to 1,000, fluently multiply and divide numbers up to 100 or use strategies, and use models to represent the concepts of multiplication and division.
- ***Geometry and Measurement.*** Your student can often independently understand time intervals, volume and mass measurements, concepts of money, and concepts of area and perimeter. He or she can often identify and describe two- and three-dimensional shapes.
- ***Number Sense.*** Your student can often independently and fluently use the concept of place values to round numbers to the nearest 10 or 100, compare two fractions, understand a fraction a/b as a pieces of a whole that is divided into b equal parts, and represent fractions on a number line.

Grade 4

- ***Algebraic Thinking and Data Analysis.*** Your student can often independently represent and interpret data; formulate questions addressed by data; recognize the relationships between the four operations; solve real-world problems with the four operations on multi-digit whole numbers; and generate a number pattern.
- ***Computation.*** Your student can often independently add and subtract multi-digit whole numbers fluently; multiply fluently within 100 or using strategies; find quotients and remainders using strategies; and decompose a fraction.
- ***Geometry and Measurement.*** Your student can often independently measure lengths; understand degrees; solve problems with composed angles; know relative measurement sizes; solve real-world problems involving measurements or the perimeter/area of rectangles; and describe two-dimensional figures.
- ***Number Sense.*** Your student can often independently name and write mixed numbers; explain why two fractions are equivalent using models; write tenths and hundredths using decimals and fractions; and use place value to round multi-digit whole numbers.

Grade 5

- ***Algebraic Thinking.*** Your student can often independently solve real-world problems involving whole numbers, fractions, and decimals using the four operations; graph ordered pairs on a coordinate plane; and write linear expressions that arise from real-world problems.
- ***Computation.*** Your student can often independently multiply and divide multi-digit whole numbers; use the four operations with fractions and decimals to the hundredths; and evaluate expressions with parentheses or brackets.
- ***Geometry and Measurement, Data Analysis, and Statistics.*** Your student can often independently identify, describe, and draw triangles and circles; identify and classify polygons; collect and represent data; find measures of center; convert among measurement units; and apply area, perimeter, and volume formulas.
- ***Number Sense.*** Your student can often independently use a number line to compare fractions; explain different fraction interpretations; continue patterns when multiplying/dividing by powers of 10; round decimals; interpret percents; and understand place value in a multi-digit number.

Grade 6

- ***Algebra and Functions.*** Your student can often independently substitute values into expressions with variables and exponents; create equivalent linear expressions; write and solve one-step equations; graph inequalities; graph points; and find vertical and horizontal distances between points.
- ***Computation.*** Your student can often independently divide multi-digit whole numbers; solve problems with fractions and decimals; divide with two fractions; and use order of operations to evaluate expressions (including those with exponents), justifying each step.

- ***Geometry and Measurement, Data Analysis, and Statistics.*** Your student can often independently convert measurement units; know interior-angle sum formulas; find and graph sides of a polygon; solve problems involving volume, surface area, and the area of composite shapes; and summarize and graphically represent data.
- ***Number Sense.*** Your student can often independently understand the integer number system; compare rational numbers; connect fractions to percents; identify prime numbers; find the greatest common factor or least-common multiple of two numbers; and understand ratios and unit rates.

Grade 7

- ***Algebra and Functions.*** Your student can often independently create and solve linear equations and inequalities; find and explain slopes for linear graphs in context; identify unit rates from graphs and contexts; and determine if graphs or contexts represent proportional relationships.
- ***Data Analysis, Statistics, and Probability.*** Your student can often independently determine if a sample is appropriate for a population; find and use means, medians, and ranges to draw conclusions; find trends and outliers in graphs and tables; and find probabilities and all possible outcomes of a probability model.
- ***Geometry and Measurement.*** Your student can often independently use scale factors to find sides of similar polygons and lengths in scale drawings; identify angle relationships; find the area and circumference of a circle; and find volumes, surface areas, and nets of cylinders and rectangular prisms.
- ***Number Sense and Computation.*** Your student can often independently use number lines to add, subtract, multiply, and divide whole numbers; find unit rates; use ratios and percents; write expressions from contexts; find prime factorizations and square roots of perfect squares; and plot irrational numbers.

Grade 8

- ***Algebra and Functions.*** Your student can often independently solve linear equation and inequality problems; identify key features of linear tables and graphs, such as slopes and y-intercepts; find solutions of two linear equations; and decide if a graph is a function.
- ***Data Analysis, Statistics, and Probability.*** Your student can often independently construct and interpret a scatterplot; create and use a line of best fit to solve real-world problems; understand independent and compound events; and find the sample space of compound events and calculate their probabilities.
- ***Geometry and Measurement.*** Your student can often independently find volumes and surface areas of 3-D figures; describe rotations, reflections, translations, and dilations for congruent and similar figures; and use the Pythagorean Theorem to find missing sides and distances.

- ***Number Sense and Computation.*** Your student can often independently identify a number written in scientific notation or find its decimal expansion; find the approximate value of an irrational number; apply properties of exponents; and solve an equation in the form $x2 = p$ if p is a perfect square.

1.21.3.4 Science Score-Reporting Categories

The ILEARN science assessments measure students' understanding of the standards at the end of grades 4 and 6, and High School biology. These assessments measure students' proficiency in science knowledge and skills. ILEARN individual student reports describe "at or near" proficient science performance in the following reporting categories:

Grade 4

- ***Physical Science.*** Proficiency in this category requires students to explain how an object's speed and energy are related, explain energy transfer between objects, design a device that converts energy, develop a model of waves or a path of light, and use patterns to transfer information. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- ***Life Science.*** Proficiency in this category requires students to construct an argument about how an organism's structures help it survive and/or reproduce and to develop a model to describe how animals receive, process, and respond to sensory information. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- ***Earth and Space Science.*** Proficiency in this category requires students to explain changes in a landscape over time, use maps to describe patterns in Earth's features, investigate the effects of natural Earth processes, and describe relationships between humans and the environment. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- ***Computer Science.*** Proficiency in this category requires students to use visuals or variables to represent data, describe strategies to solve computing problems, design and debug programs, predict outcomes from data sets, and describe relationships among technology, humans, and society. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.

Grade 6

- ***Physical Science.*** Proficiency in this category requires students to support a claim that digital signals are more reliable for transmitting information than analog signals and to develop or use wave models to show the relationship between amplitude and energy and how materials affect waves. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.

- ***Life Science.*** Proficiency in this category requires students to explain how changes to an environment or resource affect a population of organisms and explain or model how photosynthesis or living and non-living factors affect the cycling of matter and flow of energy in an ecosystem. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- ***Earth and Space Science.*** Proficiency in this category requires students to develop models of the solar system that describe patterns of lunar phases, eclipses, seasons, and the role of gravity in the motion of celestial objects or analyze data to demonstrate scale properties of those objects. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- ***Computer Science.*** Proficiency in this category requires students to identify examples of binary code, represent complex problems as algorithms, explain how cybersecurity can protect private information, and recommend changes to computing devices to improve accessibility and user experience. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.

Biology

- ***From Molecules to Organisms: Structures and Processes.*** Proficiency in this category requires students to explain how the structure of DNA affects proteins, investigate how feedback mechanisms maintain homeostasis in organisms, explain how sugar molecules combine with other elements to form large molecules, and use models to describe systems in multicellular organisms, cellular division, photosynthesis, and energy change in cellular respiration. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- **Ecosystems: Interactions, Energy, and Dynamics.** Proficiency in this category requires students to use math to explain carrying capacity, biodiversity, and matter and energy in an ecosystem in aerobic and anaerobic conditions, describe carbon cycling through Earth's systems, explain the stability of ecosystems, find solutions to reduce the impacts of human activities on the environment, and explain the role of group behavior in species' survival and reproduction. Your student inconsistently demonstrates the knowledge and skills expected of a proficient student.
- **Heredity: Inheritance and Variation of Traits, and Biological Evolution: Unity and Diversity.** Proficiency in this category requires students to ask questions about the role of DNA and chromosomes in heredity and explain the causes of genetic variation, the increase in advantageous traits in a population over time, the common ancestry among species, the effects of environmental and biological factors on a species' evolution, and how natural selection drives adaptation of populations. Your student inconsistently demonstrates the

knowledge and skills expected of a proficient student.Social Studies Score-Reporting Categories

The ILEARN social studies assessment measures students' understanding of the standards at the end of grade 5. The assessment measures students' proficiency in social studies knowledge and skills. ILEARN individual student reports describe "at or near" proficient social studies performance in the following reporting categories:

- ***Civics and Government.*** Your student can often independently explain key ideas and concepts relating to the founding of the United States, the U.S. Constitution, elections, and the branches of government. Your student can often identify ways that citizens can bring about political change.
- ***Geography and Economics.*** Your student can often independently use maps to locate places and regions and to identify physical and human systems from both today and the past. Your student can define market economies and can often describe factors that make them work.
- ***History.*** Your student can often identify early cultures and settlements in North America and major leaders who influenced the American Revolution. Your student often thinks chronologically and can use sources to examine historical events.

1.21.4 ITEM SELECTION ALGORITHM

CAI's adaptive algorithm takes as input two sources of information: an item pool and a test blueprint. The adaptive algorithm is then configured to execute maximally adaptive test administrations under the constraint of blueprint match. Configuration of the adaptive algorithm is critical because the composition of the item pool, which changes from administration to administration, interacts with the blueprint to influence the performance of the adaptive algorithm. A more detailed overview of the adaptive algorithm is presented in Appendix 4-B.

1.21.4.1 Test Blueprint

Test blueprints may contain specifications from the content hierarchy (strand, benchmark, standard, etc.) and other constraints, such as DOK, item type, or any other test item attribute that may be stored.

CAI's adaptive engine supports blueprints that meet the following conditions (which have been advocated by the Consortium for Citizens with Disabilities, an umbrella group encompassing most national advocacy groups for students with disabilities and other exceptional students):

1. Every student is tested on the full range of grade-level content, with no discernible differences in the content assessed.
2. Every student is tested on items measuring the same mix of cognitively complex skills, with no discernible difference—regardless of student proficiency.

3. Every student is tested on items reflecting the full range of other aspects (e.g. prescribed number of open-ended items) of the grade-level curriculum as may be appropriate for the grade and subject.
4. Students are tested on items that provide the best measurement possible within these constraints.

These four principles ensure that every student can accurately demonstrate his or her academic skills and knowledge across the entire grade-level curriculum. CAI's adaptive algorithm supports blueprints that align with these principles.

1.21.4.2 Item Selection

The adaptive algorithm, built on our partnerships with client states over the years, ensures that each student will receive a test that (1) matches the blueprint and (2) contains the items that best match their performance level, as defined by the blueprint.

To accomplish this goal, the algorithm implements a highly parameterized multiple-objective utility function that includes:

- a measure of the content match to the blueprint,
- a measure of overall test information, and
- measures of test information for each reporting category on the test.

We define an objective function that measures an item's contribution to each of these objectives, weighting them to achieve the desired balance among them. The equation below sketches this objective function for a single item.

$$f_{ijt} = w_2 \left(\frac{\sum_{r=1}^{R} s_{rit} p_r d_{rj}}{\sum_{r=1}^{R} d_{rj}} \right) + w_1 \sum_{k=1}^{K} q_k h_{1k}(v_{kijt}, V_{kit}, t_k) + w_0 h_0(u_{ijt}, U_{it}, t_0)$$

Where the *w* terms represent user-supplied weights that assign relative importance to meeting each of the objectives, d_{rj} indicates whether item *j* has the blueprint-specified feature *r*, and p_r is the user-supplied priority weight for feature *r*. The term s_{rit} is an adaptive control parameter that is described below. In general, s_{rit} increases for features that have not met their designated minimum as the end of the test approaches.

The remainder of the terms represent an item's contribution to measurement precision:

v_{kjit} is the value of item *j* toward reducing the measurement error for reporting category *k* for test taker *i* at time of selection *t*; and

u_{ijt} is the value of item *j* in terms of reducing the overall measurement error for test taker *i* at time of selection *t*.

The terms U_{it} and V_{kit} represent the total information overall and on reporting category *k*, respectively.

The term q_k is a user-supplied priority weight associated with the precision of the score estimate for reporting category k. The t terms represent precision targets for the overall score (t_0) and each score reporting category score. The functions h(.) are given by:

$$h_0\left(u_{ijt}, U_{it}, t_0\right) = \{au_{ijt} \; if \; U_{it} < t_0 \; bu_{ijt} \; otherwise$$

$$h_{1k}\left(v_{kijt}, V_{kit}, t_k\right) = \{c_k v_{kijt} \; if \; V_{kit} < t_k \; d_k v_{kijt} \; otherwise$$

Items can be selected to maximize the value of this function. This objective function can be manipulated to produce a pure, standards-free adaptive algorithm by setting w_2 to zero or to produce a completely blueprint-driven test by setting $w_1 = w_0 = 0$. Adjusting the weights to optimize performance for a given item pool will enable users to maximize information subject to the constraint that the blueprint is virtually always met.

We note that the computations of the content values and information values generate values on very different scales and that the scale of the content value varies as the test progresses. Therefore, we normalize both the information and content values before computing the value of Equation 1.

This normalization is given by $x = \{1 \; if \; min = max \; \frac{v-min}{max-min} \; otherwise$, where min and max represent the minimum and maximum, respectively, of the metric computed over the current set of items or item groups.

Figure 10 summarizes the item selection process. If the item position has been designated for a field-test item, then that item is administered. Otherwise, the adaptive algorithm is triggered.

Figure 10: Summary of Item Selection Process
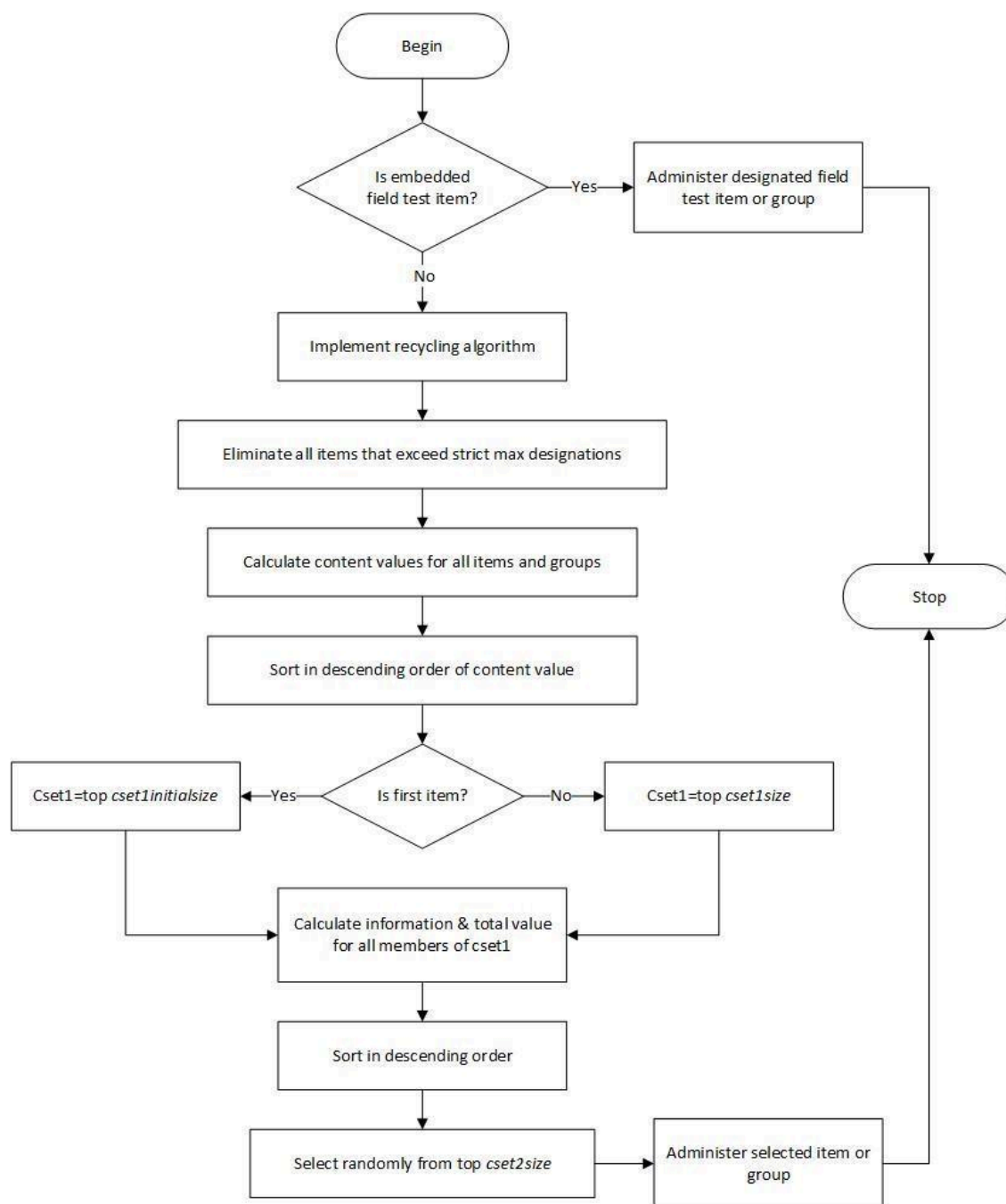
Items (or groups of items in the case of ELA tests) are sorted by their "content value," their value toward meeting the content constraints in the blueprint. Information measures are added to the content measures, and the items are sorted based on their overall value for the objective function. The final item selection is made based on a random selection from among the small subset of items that have the highest combined content and information value.

We further note that at startup for each test administration, the item pool is customized based on the student's access needs. Any items indicated as access-limited for characteristics associated with the student are removed from the item pool at the initiation of the test; therefore, all item selection computations are based only on items to which the student has access. For example, this applies to items that have been brailled and can be delivered to students who require the accommodation of braille. Further, any items that do not have any audio files associated to them, or audio files that have an associated ASL video file, would be administered to students with the ASL accommodation.

1.21.4.3 Accommodated Paper Form Construction

For all grades and subjects, a fixed form was created for use as paper form when a student's Individualized Education Program (IEP) called for such an accommodation. This form was transcribed to Spanish (except for ELA) and braille.

During test development, forms across all modes were required to adhere to the same test blueprints, content-level, and psychometric considerations. The online and accommodated forms were then reviewed for their comparability of item counts, both at the overall test level and at the reporting category levels. ELA assessments in both administration modes were additionally compared for the distribution of passages by length. The forms were then submitted for psychometric reviews, during which the following statistics were computed and compared between the online and paper-and-pencil accommodated forms where possible, given the various item sources and differing scales of the item pools:

    IRT *b*-parameter (difficulty) mean and standard deviation;

    IRT *b*-parameter minimum and maximum;

    IRT *a*-parameter mean and standard deviation;

    IRT *a*-parameter minimum and maximum;

    Item *p*-value mean and standard deviation;

    Item *p*-value minimum and maximum; and

    Lowest bi/polyserial.

A sample output with summary statistics for grade 5 social studies is presented in Table 77. As the table shows, the IRT *b*-parameter (difficulty) mean and the item *p*-value mean are similar between the forms.

As mentioned, parallelism among test forms was further evaluated by comparing Test Characteristics Curves (TCCs), test information curves, and Conditional Standards Errors of Measurement (CSEMs) between the online and paper-and-pencil forms.

Table 77: Statistical Test Summary Comparison for Grade 5 Social Studies
Online and Paper Forms

Type	Statistics	Online Form	Paper Form
Overall	Number of Items	40	40
	Possible Score	42	42
	Difficulty Mean	0.18	0.13
	Difficulty StDev	1.02	0.89
	Difficulty Minimum	−1.21	−2.21
	Difficulty Maximum	4.04	2.06
	Parameter-A Mean	0.56	0.53
	Parameter-A StDev	0.24	0.21
	Parameter-A Minimum	0.19	0.19
	Parameter-A Maximum	1.19	0.97
	P-Value Mean	0.50	0.50
	P-Value StDev	0.14	0.13
	P-Value Minimum	0.09	0.28
	P-Value Maximum	0.75	0.86
	Lowest Bi/Poly-Serial	0.22	0.25

Test Characteristic Curve

An Item Characteristic Curve (ICC) shows the probability of a correct response as a function of ability, given an item's parameters. TCCs can be constructed as the sum of ICCs for the items included on any given assessment. The TCC can be used to determine test taker raw scores or percentage-correct scores that are expected at a given ability level. When two tests are developed to measure the same ability, their scores can be equated using TCCs.

Items were selected for the paper form so that the form TCC matched the regular online form TCC as closely as possible. Paper forms are accommodations and are therefore only administered to a small subset of students who share specific characteristics. Figure 11 compares the TCCs for both online and paper forms of grade 5 social studies. Appendix 4-C provides the TCCs for online fixed form tests.

Figure 11: TCC Comparisons of Grade 5 Social Studies Online and Paper Forms
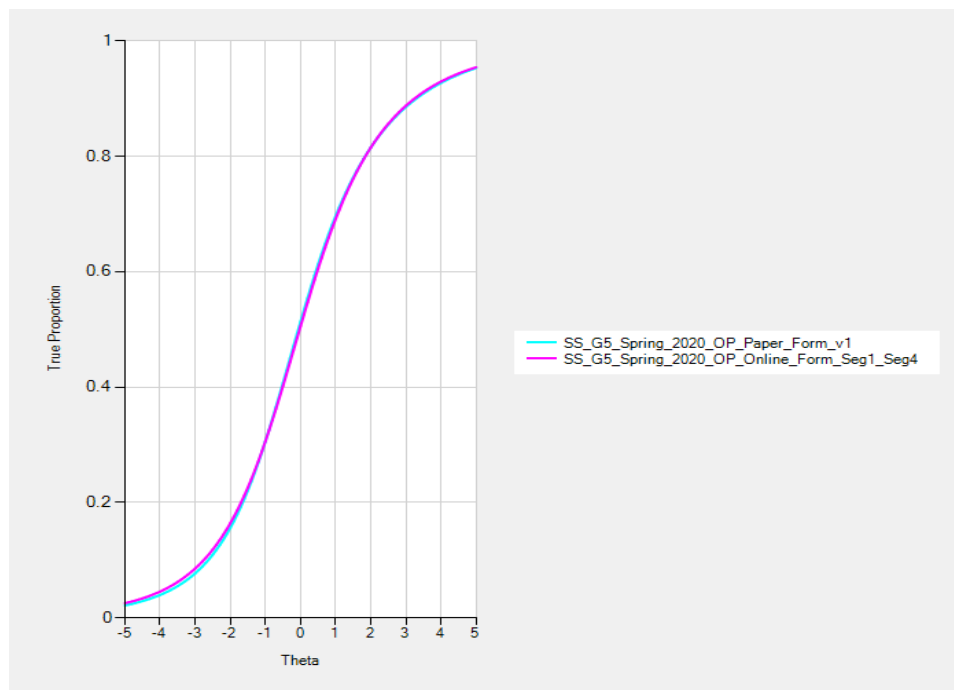
Assembly of parallel forms is a critical step in the test development process when there is a need for developing more than one form. For the test scores to be comparable across forms, such forms must meet both statistical and content requirements. Figure 12 illustrates a sample TCC difference, which allows us to evaluate the degree to which the parallelism is achieved between the forms.

Figure 12: TCC Differences of Grade 5 Social Studies Online and Paper Forms
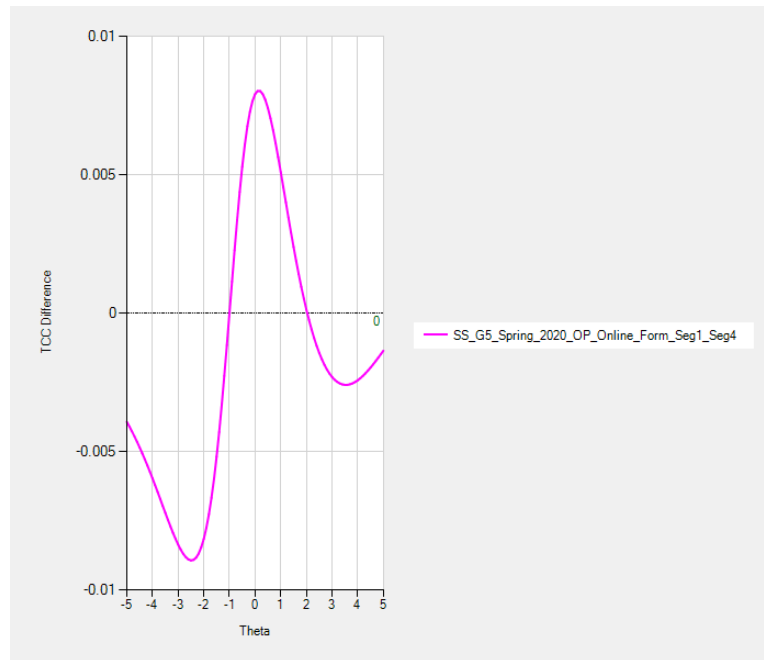
Conditional Standard Error of Measurement Curve

The CSEM curve shows the level of error of measurement expected across the range of student ability, and the Form Analyzer tool allows test developers to compare the statistical comparability of multiple forms simultaneously. The example in Figure 13 superimposes two CSEM curves onto one plot so that test developers can view the degree to which the two test forms are statistically parallel, and this is provided as an example of how test developers use the CSEM curves when building forms.

Figure 13: CSEM Comparisons of Grade 5 Social Studies Online and Paper Forms
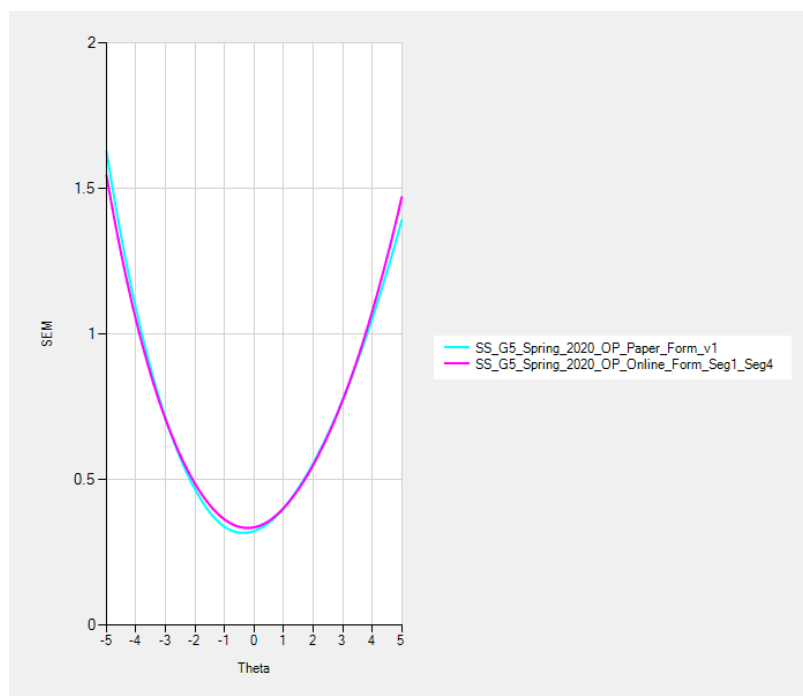
1.21.5 BLUEPRINT MATCH

The item selection algorithm delivers a test covering more benchmarks and with better precision compared with a fixed-form test. Across all grades and subjects, almost all tests met the blueprint specifications with a 100% match. The spring 2023 Simulation Summary Report is presented as Appendix 4-D, Simulation Summary Report.

The blueprints developed for ELA are provided in Appendix 4-E, English/Language Arts Blueprints. The blueprints are organized by strand and specify the number of items required for each reporting category, ensuring that the form contains enough items in that category to elicit enough information from the student to justify strand-level scores. Appendix 4-E also shows the reporting categories and required number of items in the proposed ELA blueprints.

1.21.5.1 ELA Blueprints

The ELA blueprint results in an assessment design that delivers the following to each student:

- In grades 3–5: Two nonfiction reading passages with associated items and two literary reading passages with associated items;
- In grades 6–8: Three nonfiction reading passages with associated items and one literary reading passage with associated items;
- Two to three speaking and listening items and up to four Media Literacy items;
- Stand-alone writing and/or research items; and

- One PT which includes two "precursor" items leading up to a text-based writing task.

The blueprint defines the reading standards within each strand. The standards have assigned item ranges to ensure that the material is represented on a test form with the proper emphasis relative to other standards in that reporting category. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction or the adaptive assessment. Writing is measured by an extended text-based writing task representing the writing dimensions of Organization/Purpose, Evidence/Elaboration, and Conventions.

1.21.5.2 Mathematics Blueprints

The blueprints developed for mathematics are shown in Appendix 4-F, Mathematics Blueprints. Reporting categories at a specific grade consist of a single content domain or, when necessary and appropriate, a combination of content domains. For each reporting category, the blueprints specify a minimum and maximum number of items on each form that should contribute to that category. This ensures that the form contains enough items in each category to elicit enough information from the student to generate an ability estimate.

Within a reporting category, the blueprint lists the associated standards and the assigned item ranges. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction or the adaptive assessment.

1.21.5.3 Science Blueprints

The blueprints developed for science are shown in Appendix 4-G, Science Blueprints. Reporting categories at a specific grade consist of a single content domain or, when necessary and appropriate, a combination of content domains. For each reporting category, the blueprints specify a minimum and maximum number of clusters and items on each form that should contribute to that category. This ensures that the form contains enough clusters and items in each category to elicit enough information from the student to generate an ability estimate.

Within a reporting category, the blueprint lists the associated performance expectations and the assigned item ranges. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction or the adaptive assessment.

1.21.5.4 Social Studies Blueprints

The blueprints developed for social studies are shown in Appendix 4-H, Social Studies Blueprints. Reporting categories at a specific grade consist of a single content domain or, when necessary and appropriate, a combination of content domains. For each reporting category, the blueprints specify a minimum and maximum number of items on

each form that should contribute to that category. This ensures that the form contains enough items in each category to elicit enough information from the student to generate an ability estimate.

Within a reporting category, the blueprint lists the associated standards and the assigned item ranges. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction or the adaptive assessment.

1.22 ITEM DEVELOPMENT PROCESS

Both Smarter and CAI ICCR developed the ELA and mathematics item banks using a rigorous, structured process that engaged stakeholders at critical junctures. Similarly, all custom Indiana development followed a very similar review process. This process was managed by CAI's Item Tracking System (ITS), which is an auditable content-development tool that enforces rigorous workflow and captures every change to, and comment about, each item. Reviewers, including internal CAI reviewers and stakeholders in committee meetings, reviewed items in ITS as they would appear to the student, with all accessibility features and tools.

1.22.1 SUMMARY OF ITEM SOURCES

ILEARN assessments were designed to measure proficiency on the Indiana Academic Standards (IAS), meet federal requirements for school accountability testing, and provide information to schools, teachers, parents, and students to support teaching and learning.

The IAS were approved by the Indiana State Board of Education in April 2014 for English/language arts (ELA) and mathematics, and in March 2015 for social studies. Minor text updates and clarifications were made to the ELA, Math, and social studies IAS in 2020. The IAS for science and computer science were approved by the State Board in June 2023. The IAS are intended to implement more rigorous standards that promote college-and-career readiness, with the goal of challenging and motivating Indiana's students to acquire stronger critical thinking, problem solving, and communications skills.

ILEARN assessments were created using a variety of item types from several sources. Table 78 denotes the sources of the items used in 2023–2024, including licensed item banks (Smarter Balanced Assessment Consortium [Smarter], Independent College and Career Ready [ICCR], and the Memorandum of Understanding [MOU]), and custom Indiana development. Each item source is outlined in more detail in Section 2.

The Smarter and ICCR ELA, mathematics, and science item banks were developed to measure college-and-career readiness standards as embodied in the Common Core State Standards (CCSS). The item banks are designed to measure the full breadth and depth of the standards and cover a range of difficulty that matches the distribution of

student performance in each grade and subject. The item banks are designed primarily for accountability assessments. However, not all CCSS map directly to the IAS, so Indiana custom developed items were needed to fill those gaps.

Table 78: Sources of Items for the ILEARN 2023–2024 Assessments

Subject and Grade(s)	Licensed Bank(s)	Indiana-Owned Items
ELA 3–8	Smarter ICCR	Yes
Mathematics 3–8	Smarter ICCR	Yes
Science 4 and 6	ICCR, MOU	Yes
Science Biology	ICCR, MOU	Yes
Social Studies 5	No	Yes
U.S. Government	No	Yes

1.22.2 DEVELOPMENT OF NEW ITEMS

Operational items used on ILEARN test forms were drawn from a variety of sources, including licensed items banks (Smarter Balanced [Smarter], Independent College and Career Readiness [ICCR], and the Memorandum of Understanding [MOU]), Indiana-owned items from external sources, and Indiana custom-developed items.

New items are developed each year to be added to the operational item pool after field testing. Several factors play into the development of new items; the item development team conducts a gap analysis for distributions of items across multiple dimensions, such as item counts, item types, item difficulty, and numbers in each strand or benchmark.

All CAI item writers who developed ICCR items have at least a bachelor's degree, and many bring teaching experience. All item writers are trained in:

- the principles of universal design,
- the appropriate use of item types, and
- the ICCR specifications.

Key materials are included in Appendix 4-I, Item Writer Training Materials. These include:

- CAI's Language Accessibility, Bias, and Sensitivity (LABS) Guidelines, which include a focus on Linguistic Complexity;
- the Indiana item specifications; and
- a training presentation (using Microsoft PowerPoint) for the appropriate use of item types.

1.23 ITEM REVIEW

During and after each operational test administration, a series of quality assurance reports is generated and used to evaluate whether operational items are performing as intended. These reports serve 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. Flagged items are reviewed by psychometricians and content experts. Details can be found in Chapter 9, Quality Assurance Procedures.

1.23.1 ITEM REVIEW PROCESSES

CAI's test development structure utilizes highly effective units organized around each content area. Unit directors oversee team leaders who work with team members to ensure item quality and adherence to best practices. All team members, including item writers, are content-area experts. Teams include senior content specialists, who review items prior to client review and provide training and feedback for all content-area team members.

All Smarter, ICCR, and custom Indiana items go through a rigorous, multiple-level internal review process before they are sent to external review. Staff members are trained to review items for both content and accessibility throughout the entire process. A sample item review checklist that our test developers use is included in Appendix 4-J, Item Review Checklist. The CAI internal review cycle includes the following phases:

- Preliminary Review;
- Content Review 1;
- Edit Review 1; and
- Senior Content Review.

1.23.1.1 Preliminary Review

A preliminary review is conducted by team leads or senior content staff. Sometimes the preliminary review is conducted in a group setting, led by a senior test developer. During the preliminary review process, test developers, either individually or as a group, analyze items to ensure the following is true for all items.

- The item aligns with the academic standard.
- The item matches the item specification for the skill being assessed.
- The item is based on a quality idea (i.e., it assesses something worthwhile in a reasonable way).
- The item is properly aligned to a DOK level.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question.

- The stimulus is clear, concise, and succinct (i.e., it contains enough information to know what is being asked, it is stated positively, and it does not rely on negatives—such as *no, not, none, never*—unless absolutely necessary).

For selected-response items, test developers also check to ensure that the set of response options are

- as succinct and short as possible (without repeating text);
- parallel in structure, grammar, length, and content;
- sufficiently distinct from one another;
- all plausible (but with a clear and single correct option); and
- free of obvious or subtle cuing.

For machine-scored constructed-response items, item developers also check that the items score as intended at each score point in the rubric and that scoring assertions address the skill that the student is demonstrating with each type of response.

At the conclusion of the Preliminary Review, items that were accepted as written or revised during this review moved on to Content Review 1. Items that were rejected during this review did not advance.

1.23.1.2 Content Review 1

Content Review 1 is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on all the criteria identified for Preliminary Review. Note that the criteria used for these internal reviews matches the same criteria used by committee members during Content/Fairness Committee Reviews, as documented in Appendix 4-J. The specialist also ensures that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item from the perspective of potential clients as well as from the specialist's own experience in test development.

1.23.1.3 Edit Review 1

During Edit Review 1, editors have four primary tasks.

First, editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items. Second, editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to make sure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. Editors ensure that the answer keys and that all information in the item is correct. For mathematics items, editors perform all calculations to ensure accuracy. Third, editors review all material for fairness and language accessibility issues, using CAI's Language Accessibility, Bias, and Sensitivity (LABS) Guidelines.

Finally, editors confirm that the items reflect the accepted guidelines for good item construction. In all items, they look for language that is simple, direct, and free of

ambiguity with minimal verbal difficulty. Editors confirm that a problem or task and its stem are clearly defined and concisely worded with no unnecessary information. For multiple-choice items, editors check that options are parallel in structure and fit logically and grammatically with the stem and that the key accurately and correctly answers the question as it is posed, is not inappropriately obvious, and is the only correct answer to an item among the distractors. For constructed-response items, editors review the rubrics for appropriate style and grammar.

1.23.1.4 Senior Content Review

By the time an item arrives at Senior Content Review, it has been thoroughly vetted by both content reviewers and editors. Senior reviewers (in particular, Senior Content Specialists) look back at the item's entire review history, making sure that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy and alignment to the standard. For machine-scored constructed-response items, senior reviewers carefully check the rubric and scoring logic by responding to the task just as the student would in the testing environment. They check full-credit, partial-credit, and zero-credit responses to verify that the scoring is working as intended and the scoring assertions adequately address the evidence the student provides with each type of response.

1.23.2 COMMITTEE REVIEW OF ITEM POOL

All Smarter, ICCR, and custom Indiana items have been through an exhaustive external review process. Items in the Smarter and ICCR item banks were reviewed by content experts in several states, as well as reviewed and approved by multiple stakeholder committees, in order to evaluate both content and bias/sensitivity. Custom Indiana items were reviewed only by Indiana educators. After items have been developed in the ICCR item bank, state content experts review any eligible items prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, or alignment or DOK updates. A CAI director for mathematics or ELA reviews all client-requested edits in light of the ICCR item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients have the option to present these items to committee (based on the edits made) or withhold them from committee review.

For items that have already been field-tested in other states, wording and scoring edits are not eligible to be made as such edits risk altering the function of calibrated items. Clients can simply select items from the available item bank to present to the committee.

During the Content/Fairness Committee Reviews for custom Indiana content, passages and items are reviewed for content validity, grade-level appropriateness, and alignment to the content standards. Content Advisory Committee Review members are typically grade-level and subject-matter experts but may also be mathematics coaches (who can

speak to standards across grades) or literacy specialists. During this review, educators also ensure that the rubrics for machine-scored constructed-response items reflect the anticipated correct responses (see more information Section 4.3.4, Rubric Validation).

Note that all custom and educator-authored Indiana development was taken to the Content and Fairness Committee Review. This committee combines the functions of the Content Advisory Committee and the Language Accessibility, Bias, and Sensitivity (LABS) Committee.

Additionally, each committee contains two members who are specifically charged with reviewing for accessibility and fairness. These stakeholders review items to check for issues that might unfairly impact students based on their background. For example, these members can include representatives from the special education, low vision, hearing impaired, and other student populations, including English Learners. Further, diverse members of this committee represent students of various ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns.

Once items have been accepted by IDOE and are ready for Content and Fairness Committee (CFC), Linguistic complexity ratings are applied in ITS. For CAI-authored items, content staff trained on IDOE's Linguistic Complexity rubric assigned ratings. IDOE staff assigned Linguistic Complexity ratings for educator-authored items.

1.23.3 FIELD-TESTING

The ILEARN item pool grows each year through the field-testing of new items. Any item used on an assessment is field tested before it is used as an operational item. The 2022–2023 ILEARN assessments contained newly developed field-test items. The embedded field-test (EFT) slots are randomly positioned for the online adaptive English/language arts (ELA), mathematics, and science assessments and are in fixed positions for the online fixed-form social studies assessments. To render high-quality responses to the EFT items, students were unaware of which were operational items and which were EFT items. For all assessments, field-test items were randomly distributed from the pool of available field-test items.

CAI's field-test item distribution algorithm minimizes design effects by using an algorithm that randomly draws an item from the pool for each student, ensuring that:

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

This design mimics the "spiraling-by-student within a classroom" model typically used with paper-pencil forms and ensures broad representation of the items across abilities and demographic groups. To describe the distribution of forms, consider that J total forms are available for administration and a total of N students are participating in the field test. The probability that any one of the J forms can be assigned to one student is 1/J. Thus, the distribution of forms would follow a uniform distribution with sample sizes

per form equal to N/J. Therefore, field-test item exposure rates depend on the number of field-test slots and the number of field-test items.

After items are field tested, psychometric analysis of classical item statistics (see next section) is used to flag items that do not perform as expected. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Data Review committee members were taught to interpret these flags and were given guidelines for examining the items for content or fairness issues.

1.23.4 RUBRIC VALIDATION

More complex selected-response items, as well as machine-scored constructed-response items, undergo rubric validation, which occurs in two phases. During the first phase, CAI content experts draw one or more samples to identify anomalous or unforeseen responses and ensure they are scored correctly. At this point, the rubrics may be adjusted and the responses rescored.

The second phase of rubric validation involves state content experts. During this phase, a fresh sample of responses is drawn from three strata in equal numbers: low-scoring responses from otherwise high-scoring students, high-scoring responses from otherwise low-scoring students, and a random sample from the remainder.

During these reviews, experts review responses and scores in a CAI system called REVISE. Items are reviewed as the students saw them, along with the student's response. The experts' comments are captured, and rubrics are accepted or updated as consensus is reached. Often, these discussions adjust tolerances. For example, in drawing a best-fitting line, the experts may choose to be more or less lenient in accepting a line as "close enough." In this regard, the process is similar to rangefinding, which is discussed in Section 3.7.3, Rangefinding. Figure 14 shows some features from REVISE.

The ITS archives critical information regarding the scoring certification completed during the rubric validation process. This includes any rubric changes made during the scoring decision meetings and the sign-off completed by the CAI senior content expert once the rubric has been changed, rescoring has been completed, and it has been verified that the scoring using the final rubric functioned as intended.

Following rubric validation, all items are subject to statistical checks, and flagged items are presented in data review committees.

Figure 14: Features of the REVISE Software

1.24 Item Statistics

The item analyses included classical item statistics and item calibrations using the two-parameter logistic (2PL) and generalized partial credit (GPC) item response theory (IRT) models for ELA, mathematics, and social studies (grade 5), Rasch for social studies (U.S. government), and Rasch testlet model for science. Classical item statistics are designed to evaluate the item difficulty and the relationship of each item to the overall scale (item discrimination) and to identify items that may exhibit a bias across subgroups (DIF analyses).

1.24.1 Classical Statistics

Classical item statistics are sample dependent, which means item difficulty and item discrimination indices are dependent on the sample of students selected to answer the items. If the same items are given to a different sample, they may vary substantially depending on the nature of the sample. This property is particularly important for *ILEARN* assessments because ELA, mathematics, and science assessments are administered via adaptive algorithms, while social studies assessments are fixed forms. For fixed-form tests, forms are randomly assigned to students, ensuring that each item is seen by a representative sample of participating students. By contrast, in an adaptive setting, items are selected to maximize test information near the student's ability

estimate, which causes the resulting data to include students with a restricted range of ability levels. That is, only high-performing students are administered the most difficult items, and vice versa. This characteristic of adaptive testing data has implications on the meaning and interpretation of the resulting classical test statistics. Specifically, the item difficulty index tends to migrate toward 0.5, regardless of how difficult an item is, and the item discrimination index is likely to be attenuated (or weakened) due to the restricted ability range in the adaptive data. As such, classical test statistics do not provide the same meaning or interpretation for items administered via adaptive algorithms. It is a standard practice in the field of psychometrics that operational items from an adaptive test do not use their operational adaptive test data to derive classical test statistics for item evaluation or item banking purposes. Therefore, classical item analyses were not conducted for operational items for ELA, mathematics, and science. In this chapter, classical analyses are reported only for operational items for social studies and field-test items for all assessments.

1.24.1.1 ELA, Mathematics, and Social Studies Classical Statistics

1.24.1.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those test takers who possess the skills being measured and those who do not. In general, the higher the value, the better the item 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). Items are flagged for review if biserial/polyserial values are less than 0.25.

1.24.1.1.2 Item Difficulty

Extremely difficult or extremely easy items are flagged for review but are not necessarily rejected if the item discrimination index is not flagged. For multiple-choice items, the proportion of test takers in the sample selecting the correct answer (*p*-values) and those selecting each of the incorrect responses, is computed. For constructed-response items, 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).

Multiple-choice items are flagged for review if the *p*-value is less than 0.25 or greater than .95. Constructed-response items are flagged if the proportion of students in any score-point category is greater than 0.95. 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. Constructed-response items 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, then the item is flagged. This situation may indicate that the scoring rubric is flawed.

The criteria used for flagging based on the classical statistics are as follows:

- Adjusted biserial/polyserial correlation statistic is less than 0.25 for multiple-choice or constructed-response items.
- Adjusted biserial correlations for multiple-choice item distractors is greater than 0.00.
- Proportion correct value is less than 0.25 or greater than 0.95 for multiple-choice and constructed-response items; proportion of students receiving any single score point is greater than 0.95 for constructed-response items.
- The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items.
- Mean total score for a lower score point exceeds the mean total score for a higher score point for constructed-response items.

1.24.1.2 Science Cluster Classical Statistics

1.24.1.2.1   Item Discrimination

The item discrimination index indicates the extent to which each item differentiated between those test takers who possess the skills being measured and those who do not. Generally, the higher the value, the better the item was able to differentiate between high- and low-achieving students. For each assertion within an item, the discrimination index was calculated as the biserial correlation between the assertion score and the ability estimate for students. The average biserial correlation was then calculated across the assertions within an item. Items are flagged for review if the average biserial correlations are less than 0.25, or one or more assertions have biserial correlations less than 0.0.

1.24.1.2.2   Item Difficulty

Both the percentage correct (often referred to as a p-value) for individual assertions and the average p-value across all assertions of a cluster item were calculated by grade for items field tested in science assessments. The average p-value across the assertions within an item cluster is defined as the item difficulty of an item cluster. Items are flagged for review if the average p-values are less than 0.30 or greater than 0.85.

1.24.1.2.3   Response Time

Because these items require students to perform multiple interactions, they may require more time for students to complete. To ensure a good balance between the amount of

information an item provides, and the time students spend on the item, item response time were recorded and analyzed. Specifically, the statistic "percentile 80" was computed for each item. A percentile 80 of x minutes means that 80% of the students spend x minutes or fewer on the item. An item is flagged for review when the percentile 80 is greater than 15 minutes, or the assertions per (percentile 80) minute is less than 0.5. The classical item statistics for the field-test items are presented in Appendix 4-K, Field Test Item Classical Statistics.

1.24.2 ITEM RESPONSE THEORY STATISTICS

1.24.2.1 ELA and Mathematics Item Response Theory Statistics

Traditional item response models assume a single underlying trait, and they assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This basic simplifying assumption allows the likelihood function for these models to take the relatively simple form of a product over items for a single student:

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

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

The ILEARN items are calibrated using the 2PL item response theory (IRT) model for multiple-choice items and the generalized partial credit model (GPCM) for constructed-response items, scored polytomously.

For multiple-choice models, the two-parameter logistic (2PL) model takes the form

$$p_{ij}\left(\theta_j, a_i, b_{i,1}, ...b_{i,m_i}\right) = \left\{ \frac{exp\left(1.7*a_i\left(\theta_j - b_{i,1}\right)\right)}{1+exp\left(1.7*a_i\left(\theta_j - b_{i,1}\right)\right)} = p_{ij}, \quad if \; z_{ij} = 1 \; \frac{1}{1+exp\left(1.7*a_i\left(\theta_j - b_{i,1}\right)\right)} = 1 - p_{ij}, \quad if \; z_{ij} = 0 \right\},$$

where $b_{i,1}$ is the difficulty parameter for item i, a_i is the discrimination parameter for item i, and z_{ij} is the observed item score for person j.

For items that have multiple, ordered response categories (i.e., partial credit items), we again have the choice of a simple Rasch family model (Masters' 1982 partial credit model) or a more general variant such as Muraki's (1992) generalization of Samejima's (1972) graded response model. For smaller-sample tests, such as state-specific alternate assessments, we recommend the Rasch-family variants because they can be reliably estimated with fewer cases. Under Masters' model, the probability of a response in category i for an item with m_j categories can be written as

$$P\left(x_j = i|\theta_k, b_{j0} \cdots b_{jm_j-1}\right) = \frac{e^{\sum_{v=0}^{i} 1.7(\theta_k - b_{jv})}}{\sum_{g=0}^{m_{j-1}} e^{\sum_{v=0}^{g} 1.7(\theta_k - b_{jv})}}.$$

Muraki's generalization adds an item-dependent discrimination parameter as follows (again, Masters' formulation does not usually include the arbitrary constant 1.7):

$$P\left(x_j = i|\theta_k, b_{j0} \cdots b_{jm_j-1}\right) = \frac{e^{\sum_{v=0}^{i} 1.7a_j(\theta_k - b_{jv})}}{\sum_{g=0}^{m_{j-1}} e^{\sum_{v=0}^{g} 1.7a_j(\theta_k - b_{jv})}}.$$

Returning to the likelihood equation, the contribution of each item to the overall likelihood function remains independent of all other items, given *θ*. This is convenient for two reasons: mixing models within an analysis (e.g., one-parameter and partial credit items on the same scale) becomes no more complicated, and the likelihood of the response pattern may be calculated as the product of the likelihood of responses to individual items.

In the case of the Rasch model for 1-point items, we have:

$$p_{ij}\left(\theta_j, b_{i,1}, \ldots b_{i,m_i}\right) = \left\{ \frac{exp(\theta_j - b_{i,1})}{1+exp(\theta_j - b_{i,1})} = p_{ij}, \quad if \; z_{ij} = 1 \; \frac{1}{1+exp(\theta_j - b_{i,1})} = 1 - p_{ij}, \quad if \; z_{ij} = 0 \right\}.$$

The field test item calibration is conducted using IRTPRO 6.0. IRTPRO implements the method of Maximum Likelihood (ML) for item parameter estimation. The item parameter estimates of the field-test items are presented in Appendix 4-L, Field Test Item Parameters.

1.24.2.2 Science Item Response Theory Statistics

In discussing item response theory (IRT) models for the Indiana science assessments, we distinguish between the underlying latent structure of a model and the parameterization of the item response function conditional on that assumed latent structure. Subsequently, we discuss how group effects are considered.

1.24.2.2.1 Latent Structure

Most operational assessment programs rely on a unidimensional IRT model for item calibration and computing scores for students. These models assume a single underlying trait, and they assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This assumption of conditional independence implies that the conditional probability of a pattern of *I* item responses takes the relatively simple form of a product over items for a single student as shown below:

$$P\left(\theta_j\right) = \prod_{i=1}^{I} P\left(z_{ij}|\theta_j\right) \tag{1}$$

where z_{ij} represents the scored response of student j ($j = 1, ..., N$) to item i ($I = 1, ..., I$), z_j represents the pattern of scored item responses for student j, and θ_j represents student j's proficiency. Unidimensional IRT models differ with respect to the functional relation between the proficiency θ_j and the probability of obtaining a score z_{ij} on item i.

Some items in the Indiana science assessments are more complex than traditional item types. A single item may contain multiple parts, and each part may contain multiple student interactions. For example, a student may be asked to select a term from a set of terms at several places in a single item. Instead of receiving a single score for each item, multiple inferences are made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses to the item. These scoring units are called assertions and are the basic unit of analysis in our IRT analysis. That is, they fulfill the role of items in traditional assessments; however, for the Indiana assessment items, multiple assertions are typically developed around a single item so that assertions are clustered within items.

One approach is to apply one of the traditional IRT models to the scored assertions; however, a substantial complexity that arises from the use of this new item types is that local dependencies exist between assertions pertaining to the same stimulus (i.e., item or item cluster). The local dependencies between the assertions pertaining to the same stimulus constitute a violation of the assumption that a single latent trait can explain all dependencies between assertions. Fitting a unidimensional model in the presence of local dependencies may result in biased item parameters and standard errors of measurement. In particular, it is well documented that ignoring local item dependencies leads to an overestimation of the amount of information conveyed by a set of responses and an underestimation of the SEM (e.g., Sireci, Wainer, & Thissen, 1991; Yen, 1993).

The effects of groups of assertions developed around a common stimulus can be accounted for by including additional dimensions corresponding to those groupings in the IRT model. These dimensions are considered nuisance dimensions. Whereas traditional unidimensional IRT models assume that all assertions (the basic units of analysis) are independent given a single underlying trait θ, we now assume the conditional independence of assertions, given the underlying latent trait θ and all nuisance dimensions:

$$P\left(\theta_j, u_j\right) = \prod_{i \in SA} P\left(z_{ij}|\theta_j\right) \prod_{g=1}^{G} \prod_{i \in g} P\left(z_{ij}|\theta_j, u_{jg}\right) \tag{2}$$

where SA indicates stand-alone assertions, u_g indicates the nuisance dimension for assertion group g (with the position of student j on that dimension denoted as u_{jg}), and

u is the vector of all G nuisance dimensions. It can be seen that the conditional probability $P\left(z_{ij}|\theta_j, u_{jg}\right)$ becomes a function of two latent variables: the latent trait θ, representing a student's proficiency in science (the underlying trait of interest), and the nuisance dimension u_g, accounting for the conditional dependencies between assertions of the same group. Furthermore, we assume that the nuisance dimensions are all uncorrelated with one another and with the general dimension. It is important to point out that even though every group of assertions introduces an additional dimension, models with this latent structure do not suffer from the complications of dimensionality like other multidimensional IRT models because one can take advantage of this special structure during model calibration (Gibbons & Hedeker, 1992). In this regard, Rijmen (2010) showed that it is unnecessary to assume all nuisance dimensions are uncorrelated; rather, it is sufficient that they are independent, given the general dimension θ.

The model structure of the IRT model for science is illustrated in Figure 15. Note that stand-alone items can be scored with more than one assertion. The assertions of stand-alone items with more than one assertion but fewer than four assertions are also modeled as stand-alone assertions. Even though these assertions are likely to exhibit conditional dependencies, the variance of the nuisance dimension cannot be reliably estimated if it is based on a very small number of assertions. The few stand-alone items with four or more assertions are treated as item clusters to take into account the conditional dependencies.

Figure 15: Directed Graph of the Science IRT Model
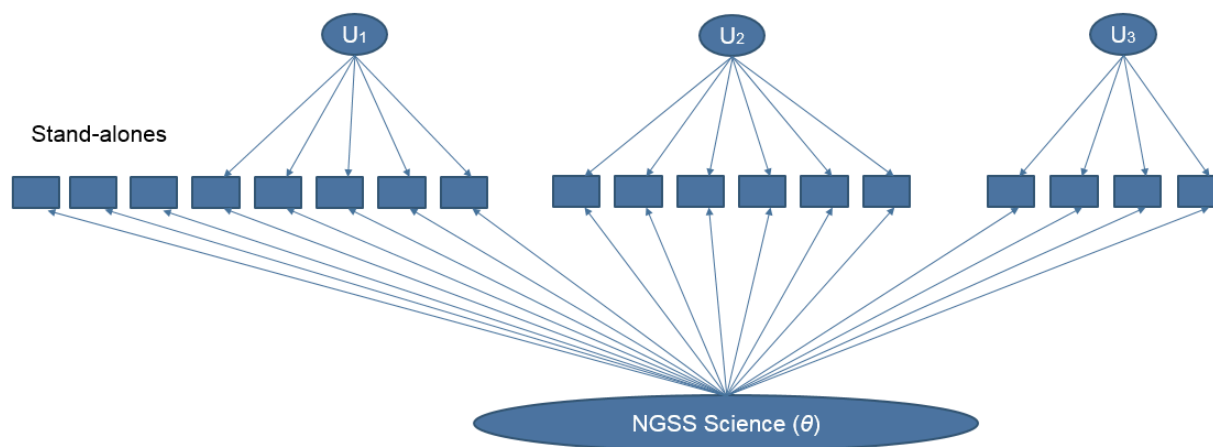
1.24.2.2.2   *Item Response Function*

For the grouped assertions, like in unidimensional models, different parametric forms can be assumed for the conditional probability of obtaining a score of z_{ij}. The Rasch

testlet model is adopted as the IRT model for the Indiana science assessments (Wang & Wilson, 2005). For binary data, the Rasch testlet model is defined as:

$$P\left(z_{ij}|\theta_j, u_{jg}; b_i\right) = \frac{exp\left(\theta_j + u_{jg} - b_i\right)}{1 + exp\left(\theta_j + u_{jg} - b_i\right)} \tag{3}$$

The item response function of the Rasch testlet model models the probability of a correct answer (i.e., a true assertion), as a function of the overall proficiency θ, the nuisance dimension u_g, and the item (i.e., assertion) difficulty b_i. The Rasch testlet model does not include item discrimination parameters. Furthermore, only models for binary data are considered. Assertions are always binary because they are either true or false. Nevertheless, the model could easily accommodate polytomous responses by using the same response function that is incorporated in unidimensional models for polytomous data.

1.24.2.2.3 Multigroup Model

The science item bank is calibrated concurrently using all the items administered in any of the states that collaborate with CAI on their new science assessments. In the calibration, each state is treated as a population of students or group. Overall group differences are taken into account by allowing a group-specific distribution of the overall proficiency variable θ. Specifically, for every student j belonging to group k, $k = 1, ..., K$, a normal distribution is assumed,

$$\theta_j \sim N\left(\mu_k, \sigma_k^2\right),$$

where μ_k and σ_k^2 are the mean and variance of a normal distribution. The mean of the reference distribution ($k = 1$) is set to 0 to identify the model (for free item calibrations, where there are no anchor items with their location parameters set to specific values). For each of the nuisance variables u_g, a common variance parameter across groups is assumed, and the means are set to 0 in order to identify the model,

$$u_{jg} \sim N\left(0, \sigma_{u_g}^2\right).$$

In 2018 and 2019, the IRT models were calibrated using the Bayesian networks with logistic regression (BNL) suite of Matlab functions (Rijmen, 2006) and flexMIRT (Cai, 2017). The resulting parameters from BNL were used as starting values for flexMIRT, to reduce the estimation time for flexMIRT. Starting from 2021, CAIRT (Cambium Assessment IRT) was used for calibration, which was specifically developed by CAI to calibrate advanced IRT models on very large data sets. It relies on the same estimation methods as BNL. CAI has cross-validated parameter estimates from CAIRT with BNL and flexMIRT under a variety of scenarios (Rijmen, Liao, & Lin, 2021). In 2024, Indiana

computer science field test items were calibrated using CAIRT by anchoring on the parameters of operational NGSS MOU items on the test.

1.24.3 ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999, 2014) provide a guideline for when sample sizes permitting subgroup differences in performance should be examined and appropriate actions taken to ensure that differences in performance are not attributable to construct-irrelevant factors.

Differential Item Functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important because it provides a statistical indicator that an item may contain cultural or other bias. DIF flagged items are further examined by content experts who are asked to re-examine each flagged item to decide whether the item should be excluded from the pool due to bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF.

CAI uses a generalized Mantel-Haenszel (MH) procedure to calculate DIF. The generalizations include adaptation to polytomous items; and improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's raw score on the operational items on a given test is used as the ability-matching variable. That score is divided into 10 intervals to compute the $MH\chi^2$ DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD [Dorans & Schmitt, 1991]) are computed for polytomous items.

The MH chi-square statistic (Holland & Thayer, 1988) is calculated as:

$$MH\chi^2 = \frac{\left(\left|\sum_k n_{R1k} - \sum_k E\left(n_{R1k}\right)\right| - 0.5\right)^2}{\sum_k var\left(n_{R1k}\right)},$$

where $k = \{1, 2, ...K\}$ for the strata, n_{R1k} is the number of correct responses for the reference group in stratum k, and 0.5 is a continuity correction. The expected value is calculated as

$$E\left(n_{R1k}\right) = \frac{n_{+1k} n_{R+k}}{n_{++k}},$$

where n_{+1k} is the total number of correct responses, n_{R+k} is the number of students in the reference group, and n_{++k} is the number of students in stratum k, and the variance is calculated as

$$var\left(n_{R1k}\right) = \frac{n_{R+k}n_{F+k}n_{+1k}n_{+0k}}{n_{++k}^2\left(n_{++k}-1\right)},$$

where n_{F+k} is the number of students in the focal group, n_{+1k} is the number of students with correct responses, and n_{+0k} is the number of students with incorrect responses in stratum k.

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k \frac{n_{R1k}n_{F0k}}{n_{++k}}}{\sum_k \frac{n_{R0k}n_{F1k}}{n_{++k}}}.$$

The MH-delta (Δ_{MH} [Holland & Thayer, 1988]) is then defined as

$$\Delta_{MH} = -2.35 ln\left(\alpha_{MH}\right).$$

The generalized MH statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = \left(\sum_k a_k - \sum_k E(a_k)\right)'\left(\sum_k var(a_k)\right)^{-1}\left(\sum_k a_k - \sum_k E(a_k)\right)$$

where a_k is a $(T-1) X 1$ vector of item response scores, corresponding to the T response categories of a polytomous item (excluding one response). $E(a_k)$ and $var(a_k)$, a $(T-1) \times (T-1)$ variance matrix are calculated analogously to the corresponding elements in $MH\chi^2$ in stratum k.

The SMD (Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{FK}m_{FK} - \sum_k p_{FK}m_{RK}$$

where

$$p_{FK} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group students in stratum k,

$$m_{FK} = \frac{1}{n_{F+k}}\left(\sum_t a_t n_{Ftk}\right)$$

is the mean item score for the focal group in stratum k, and

$$m_{RK} = \frac{1}{n_{R+k}}\left(\sum_t a_t n_{Rtk}\right)$$

is the mean item score for the reference group in stratum k.

DIF analysis was conducted for all field-test items with at least 200 responses per item in each subgroup (Zwick, 2012) to detect potential item bias for major demographic groups. DIF statistics were calculated at the item level for ELA, mathematics, and social studies and at the assertion level for science. DIF analyses were performed for the following groups:

- Male/Female
- White/African American
- White/Hispanic
- White/Asian
- White/Native American
- Student with Special Education (SPED)/Not SPED
- SES/Not SES (proxy for Free and Reduced-Price Lunch)
- ELs/Not ELs

Table 79 details the DIF classification rules. Similar to how the general MH statistic is used to classify items on traditional tests, assertions were classified into three categories (i.e., A, B, or C) for DIF, ranging from "no evidence of DIF" to "severe DIF." Furthermore, assertions were categorized positively (i.e., +A, +B, or +C), signifying that an item favors the focal group (e.g., African American/Black, Hispanic, or female), or negatively (i.e., –A, –B, or–C), signifying that an item favors the reference group (e.g., white or male). For science, an item cluster is flagged for data review if two or more assertions show "C" DIF in the same direction. Appendix 4-M summarizes the DIF flagging results of the spring 2024 field-test items.

<div align="center">

Table 79: DIF Classification Rules

</div>

DIF Category	Flag Criteria
Dichotomous Items	
C	MH_{χ^2} is significant, and $\left|\hat{\Delta}_{MH}\right| \geq 1.5$.
B	MH_{χ^2} is significant, and $1 \leq \left|\hat{\Delta}_{MH}\right| < 1.5$.
A	MH_{χ^2} is not significant, or $\left|\hat{\Delta}_{MH}\right| < 1$.
Polytomous Items and Assertions	
C	MH_{χ^2} is significant, and $|SMD|/|SD| > .25$.
B	MH_{χ^2} is significant, and $.17 < |SMD|/|SD| \leq .25$.

A	MH_{X^2} is not significant, or $|SMD|/|SD| \leq .17.$

1.25 ITEM BANKS

The ILEARN item bank is quite robust, containing licensed items which have been constructed explicitly to support multiple statewide assessment programs. As described above, all items used on ILEARN assessments are aligned to the IAS. The ILEARN item banks support an adaptive assessment for ELA, mathematics, and science, and a fixed-form assessment in social studies grade 5 and U.S. government. Summaries of current item inventories are provided in this section.

The *ILEARN* ELA and mathematics operational item banks draw primarily from the Smarter item bank, which includes more than 30,000 items across grades and subjects. However, not all IAS are covered by Smarter items. Items from CAI's ICCR item bank and custom Indiana-developed items were also used to ensure complete coverage of the IAS and support a more robust item pool for the computer-adaptive assessment.

For grades 4 and 6 science and biology, the item banks consisted mostly of items licensed from the MOU and ICCR, and some Indiana-owned items. The grade 5 social studies item pool and the U.S. government item pool contain solely custom Indiana items. Table 80 provides the count of items, by source, used on the 2023–2024 *ILEARN* assessments.

Table 80: Operational Item Counts by Source

Subject and Grade	Smarter Items	ICCR Items	Indiana-Owned Items	MOU Items	Total Items
ELA 3	349	40	53		442
ELA 4	265	44	46		355
ELA 5	262	58	50		370
ELA 6	195	67	35		297
ELA 7	249	53	50		352
ELA 8	305	27	43		375
Mathematics 3	369	58	72		499
Mathematics 4	419	25	61		505
Mathematics 5	333	71	65		469
Mathematics 6	483	35	46		564
Mathematics 7	467	33	61		561
Mathematics 8	307	30	49		386
Science 4		43	19	142	204

Subject and Grade	Smarter Items	ICCR Items	Indiana-Owned Items	MOU Items	Total Items
Science 6		42	13	84	139
Biology		56		117	173
Social Studies 5			61		61
U.S. Government			36		36

Additionally, ELA and mathematics assessments included one performance task per grade. Table 81 lists the counts of performance tasks in the 2023–2024 item pool.

Table 81: Operational Performance Task Counts by Source

Subject	Grade	Number of Smarter Performance Tasks
ELA	3	2
ELA	4	3
ELA	5	6
ELA	6	4
ELA	7	3
ELA	8	1
Mathematics	3	2
Mathematics	4	2
Mathematics	5	2
Mathematics	6	2
Mathematic	7	2
Mathematics	8	5

1.25.1 ESTABLISHING THE BANKS

1.25.1.1 ELA and Mathematics

Since *ILEARN* relies heavily on licensed item banks, a process for ensuring alignment of those items to the IAS was developed. CAI and IDOE worked to determine a crosswalk between the IAS and the standards for the licensed banks. During item acceptance review meetings, educators reviewed the IAS and then worked through items in small batches to rate their levels of agreement about the alignment of the standard to the given item.

Prior to the spring 2019 administration, two item acceptance review meetings were held. Results of those meetings can be found in Chapter 2 of the 2018–2019 Technical Reports.

In November 2019, a third item acceptance review meeting was held for ELA and mathematics. Results of that meeting can be found in Chapter 2 of the 2019–2020 Technical Reports.

Subsequent item acceptance reviews were convened in November 2021 and September 2022 during which alignment was considered for Smarter performance tasks and field test items that were approved for use on *ILEARN*.

1.25.1.2 Science

Starting from 2018, science items were field-tested in MOU states, as well as the states that mainly use ICCR items. Starting from 2024, Indiana joined MOU and science items were field-tested in Indiana. All items administered in Indiana were aligned to the Indiana Academic Standards (IAS) for science.

There was a target of a minimum sample size of 1,500 students per item for any given state. Most items were administered in two or more states so that the item pools for all individual states were linked through common items. The common item design was used to calibrate all the items on a common science scale for each grade band. The calibration model is explained in detail in Section 4.4.2.2, Science Item Response Theory Statistics.

Before being eligible for administration, Indiana-owned science field-test items went through all review processes described in previous sections. MOU-owned items went through a two-stage Item Acceptance Review process that whereby content specialists and educators affirmed alignment to the IAS and content fairness for Indiana students. Following the close of the test administration window, classical statistics were performed on all administered field-test items using the data of the students testing in the state that owned the item. DIF statistics were computed based on combined states' data whenever possible (i.e., for states with an independent field test or an operational test for which the relevant demographic variable was available), following the recommendations of several Technical Advisory Committees (TACs). During the item data review meetings, items were reviewed by either the owner state committee, or the MOU cross-state data review committee; items were either accepted or rejected. All items accepted from the Indiana-specific data review will be incorporated into the operational item bank. MOU items accepted at the cross-state data review or other states' data reviews will be incorporated into Indiana's operational item bank if they align with Indiana standards and are accepted by Item Acceptance Review committees.

1.25.1.3 Item Bank Composition

Table 82 lists the ELA, mathematics, science, and social studies item types and provides a brief description of each. Examples of various item types can be found in

Appendix 4-N, Example Item Types. Table 83 through Table 86 list the number of items by type for each grade and subject.

<p align="center">Table 82: ILEARN Item Types and Descriptions</p>

Response Type*	Description
Cluster	Student works through a group of interactions measuring multiple dimensions of a science performance expectation. Clusters use assertion-based scoring to assess students' knowledge and skills.
Edit Task with Choice (ETC)**	Student chooses a word or phrase from several options in order to complete a sentence.
Equation Response (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. Responses can include numbers, fractions, expressions, inequalities, functions, and equations.
Evidence-Based, Selected-Response (EBSR)	Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A.
Extended Response (ER)	Student is directed to provide a longer, written response in the form of an essay.
External Copy (ECI)	Student selects text to complete a table or steps in a cause-and-effect chain.
Graph Item	Student creates a graph by clicking and dragging dynamic graph elements (e.g., a bar representing categorical data) to the desired location on a static graph.
Graphic Response (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hot Text (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference.
Multiple-Choice (MC)	Student selects one correct answer from four options.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Performance Task (PT)	Student works through a group of items measuring multiple standards and using various item types to demonstrate the ability to integrate knowledge and skills.
Simulation (SIM)	Student selects inputs to "run" trials. Data is presented in a table after trials are run.
Table Input (TI)	Student types numeric values into a given table.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Text Entry (TE)	Student is directed to type their response in a text box.

*Response Types ETC, EQ, MC, MS, and TI are sometimes presented together as Part A and Part B of one item.

**Four Indiana-developed items were approved for inclusion in the pool by IDOE content specialists; however, CAI did not develop any custom ETC items for ELA.

Table 83: ELA Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
TE	25	25	33	21	31	32
ETC	1		1		1	1
EBSR	66	39	43	52	28	38
HT	40	44	41	25	48	36
MI	25	12	19	16	4	7
MC	214	188	172	138	168	191
MS	71	47	61	45	72	70
ER	2	3	6	4	3	2

Table 84: Mathematics Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
TE	6	6	8	6	3	10
EQ	260	277	252	269	309	114
GI	50	23	14	23	17	18
MI	28	74	78	57	40	70
MC	134	92	94	90	90	106
MS	17	16	18	102	100	63
HT	2	2		1	1	
TI	2	15	5	16	1	5

Table 85: Science Operational Items by Item Type and Grade

Item Type	4	6	Biology**
ETC	45	31	45
ECI	2	1	1
EQ		5	2
GI	2	3	
HT	2		
MI	7	7	8
MC	28	22	13
MS	16	5	14
TI			1
MC & MS	2		2

ETC & MC		4	2
ETC & MS	1	1	2
EQ & MC			2
MI & MC	2		
Cluster	94	57	70
Graph Item	1		
ETC & EQ			7
GI, ETC & EQ			1
GI & EQ			1
TI & ETC			1
EQ & MS			1
MI & ETC	2	2	
GI & ETC		1	

Table 86: Social Studies Operational Items by Item Type and Grade

Item Type	5	U.S. Government
TE	4	
EBSR	1	13
MC	50	9
MI	3	1
MS	3	13

1.25.2 BANK MAINTENANCE

1.25.2.1 ELA, Mathematics, and Science

To maintain the Indiana item banks, new items are developed and field-tested in the spring administration of each year, using CAI's field-test engine, and then calibrated and analyzed following the procedures described in Section 4.4.2, Item Response Theory Statistics.

The field-test engine that CAI employs for embedding field-test items randomly samples field-test items for each individual test administration, essentially creating thousands of unique embedded field-test (EFT) forms. This sampling approach to embedding field-test items results in several important outcomes:

- Reduction in the number of embedded field-test items that each student must respond to and more efficient "spiraling" of items, which reduces clustering of item responses, resulting in more precise parameter estimates

- More generalizable item statistics because they are not based on items appearing in a single position
- A truly representative sample of respondents for each item

The embedded field-testing algorithm actually consists of two different algorithms—one for identifying which field-test items will be administered to which student (the *distribution algorithm*), and one for selecting the position on the test for each item administered to the student (the *positioning algorithm*).

When a student starts a test, the system randomly selects a predetermined number of item groups, stopping when it has selected item groups containing at least the minimum number of field-test items designated for administration to each student. We refer to item groups rather than items because field-test items, like items in the operational tests, can either be stand-alone items or appear together as a group, such as when items are bound with a reading passage or some other common stimulus. We use the term *item groups* to refer to both cases, with stand-alone items representing item groups of one. This randomization ensures that (1) each item is seen by a representative sample of participating students, and (2) every item is as likely as every other item to appear in a class or school, minimizing the clustering effects.

Construction of item groups for reading passages or other stimulus-based item sets similarly reduces clustering. With static embedded field test (EFT) blocks, reading passages and other stimuli are typically field tested with two or more sets of fixed items, so that each administration of a passage or stimulus is associated with a fixed set of items in a fixed order. The distribution algorithm, however, randomly selects a group of items from within the stimulus or passage set for administration, so that all items within a stimulus or passage set are administered with all other items from within the set, which reduces clustering by distributing items across all students rather than within a limited number of forms, and results in a more representative sample of students responding to each item.

A second, *positioning algorithm*, determines where an item appears on a given student's test, with the result that the position of each item is randomized among the positions designated as available for field-test items. This way, the field-test items can be interspersed with operational items (making them more difficult to detect) and each item is seen across all available positions. This approach helps "average out" position effects on item functioning, yielding more robust and generalizable estimates of their statistical properties. Our algorithm accomplishes what paper test "balanced block" designs seek to approximate. For item groups, averaging out position effects also means that any effects of item cueing are removed from item parameter estimates.

The procedures for item review are discussed in Section 4.3, Item Review. Table 87 through Table 90 present the number of field-test items administered in 2023–2024 by subject, grade, and ownership.

Table 87: Number of Field-Test Items in 2023–2024 for ELA

Grade	Smarter	Indiana-Owned
3		226
4	20	149
5	16	142
6	10	147
7		204
8		209

Table 88: Number of Field-Test Items in 2023–2024 for Mathematics

Grade	Smarter	Indiana-Owned
3	1	88
4	3	71
5	5	103
6	15	130
7	17	87
8	10	87

Table 89: Number of Field-Test Items in 2023–2024 for Science

Grade	ICCR	MOU	Indiana-Owned
4		16	27
6		21	24
Biology	2	40	31

Table 90: Number of Field-Test Items in 2023–2024 for Social Studies

Grade	Indiana-Owned
5	17
U.S Government	5

1.25.3 BRAILLE ITEM POOLS

Across all grades and subject areas, braille forms were provided in both online and paper testing modes. Hardcopy braille forms were transcribed from the regular-print paper forms and were accompanied by braille notes that indicated to Test

Administrators where modifications to the content was necessary for students taking the braille test. Online braille tests were constructed differently by subject area. For ELA and mathematics, enough items from the general education CAT pools were appropriate for the braille pool to support fully adaptive, online refreshable braille pools at each grade. For science and social studies, fixed forms were delivered for online refreshable braille.

1.25.4 SPANISH ITEM POOLS

Spanish tests were provided in the online mode only and were provided for mathematics, science, and social studies. Online Spanish tests were constructed differently by subject area. For mathematics and science, enough items from the general education CAT pools were appropriate for the Spanish pool to support fully adaptive, online refreshable Spanish pools at each grade. Students taking the online Spanish form in mathematics were exposed to the same pools of items used for braille. For social studies, fixed forms were delivered for Spanish.

5. TEST ADMINISTRATION

The State of Indiana implemented an online assessment for operational use beginning with the 2018–2019 school year. This assessment program, referred to as the ILEARN assessments, replaced Indiana Statewide Testing for Educational Progress-Plus (ISTEP+). ILEARN comprises English/language arts (ELA), mathematics, science, and social studies assessments for students ranging from grade 3 through the end of high school. ELA and mathematics assessments are administered in grades 3–8. Science is administered in grades 4 and 6, and biology is administered as an end-of-course assessment, typically in high school. Social studies is administered in grade 5, and U.S. government is administered in high school as an end-of-course assessment. The U.S. government assessment is optional. During the 2023–2024 ILEARN administrations, ELA, mathematics, science, and biology assessments were offered as computer-adaptive tests (CATs), while the social studies and U.S. government tests were offered as fixed-form online assessments. The ELA and mathematics assessments consist of a CAT segment and a performance task segment, while science and biology assessments consist of a CAT segment only. Students needed to complete the CAT segment of the test to receive a final overall scale score and both the CAT segment and the performance task segment for ELA and mathematics to receive an overall scale score and reporting category level scores.

Assessment instruments have established test administration procedures that support useful interpretations of score results, as specified in Standard 6.0 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). This chapter of the ILEARN technical report provides details on the testing procedures, accommodations, Test Administrator (TA) training and resources, and test security procedures implemented for ILEARN. Specifically, it provides the following test-administration–related evidence for the validity of the assessment results:

- A description of the student population that takes ILEARN;
- A description of the training and documentation provided to TAs necessary for them to follow the standardized administration procedures;
- A description of offered test accommodations intended to remove barriers that otherwise would interfere with a student's ability to take a test;
- A description of the test security process implemented to mitigate loss, theft, and test content reproduction of any kind; and
- A description of the quality monitoring (QM) system and test irregularity investigation process to detect cheating, monitor item quality in real-time, and evaluate test integrity used by Cambium Assessment, Inc. (CAI).

1.26 TESTING OPTIONS

Administering the 2023–2024 ILEARN assessments required coordination, detailed specifications, and proper training. In addition, several individuals in each corporation and school were involved in the administration process, from those setting up secure

testing environments to those administering the tests. IDOE worked with CAI to develop and provide the training and documentation necessary for the administration of ILEARN under standardized conditions within all testing environments, both online and on paper-and-pencil tests.

All students were required to take a practice test at their school prior to taking the 2023–2024 ILEARN assessments. These practice tests contained sample test items similar to the test items that students would encounter on the ILEARN assessments to help students become familiar with the item types that would be presented on the online or paper-and-pencil assessments. Indiana students also had the opportunity to interact with released, non-secure items on public-facing Released Items Repository (RIR) assessments available on the ILEARN portal. A completely updated ILEARN RIR was deployed for all tests in late January 2023. A quick guide for the RIR is available to the public (Appendix 5-A).

The ILEARN assessments were administered in multiple segments over multiple days. The test segments administered were as follows:

- ELA: CAT and a performance task segment.
- Mathematics: CAT and a performance task segment.
- Science: CAT.
- Social studies: fixed-form segment.

The ILEARN assessments were untimed, but timing estimates were included in the ILEARN Test Administrator's Manuals (TAM) (Appendix 5-B, 5-P,5-R) to ensure that schools had resources available to create local testing schedules. The fall and winter biology tests were not administered in the 2023–2024 school year to allow for the development of new NGSS assessments. The spring ILEARN test window for grades 3–8 was held from April 15 through May 10, 2024. The spring biology and U.S. government tests were available from April 15 through May 17, 2024.

All students enrolled in tested grade levels and courses participated in the spring 2024 ILEARN administration with or without accommodations, with the exception of students with significant cognitive disabilities (approximately 1% of the student population) who participated in the alternate assessment (I AM). I AM has a distinct administration that is described in a separate technical report. Students took the fall, winter, or spring biology ECA upon completion of the respective high school course to coincide with one of the three test windows. Section 1111(b)(2)(A) of the Elementary and Secondary Education Act of 1965 (as amended by the Every Student Succeeds Act [ESSA]) requires the implementation of high-quality student academic assessments in mathematics, reading or language arts, and science. Section 1111(b)(2)(B)(i)(II) requires that these assessments be administered to all elementary and secondary school students. In addition, Section 1111(c)(4)(E) requires participation rates in statewide assessments of at least 95% for all students and each subgroup of students and factors this percentage into the state's federal accountability system. Students' failure to take Indiana's assessments may result in a lower federal accountability rating. Students must take the

tests appropriate for the grade level and subject in which they are enrolled. Off-grade testing is not available for ILEARN.

Public and Nonpublic School Students. Students enrolled in accredited Indiana public (including charter schools) and nonpublic schools (including Choice schools) were required to participate in course-level appropriate ILEARN assessment(s).

English Learners (ELs). All ELs enrolled in tested grade levels were expected to participate in all ILEARN assessments, including English/Language Arts, regardless of how long these students had been enrolled in a U.S. school. Mathematics, science, and social studies assessments are available in Spanish in the online Test Delivery System (TDS). Spanish is available via the Spanish Toggle Designated Feature. Spanish Toggle provides students with two presentations of an item (English and Spanish). Students can toggle between these presentations using the global icon on TDS.

Translated glossaries are also available as a support for the top 5 student home languages in Indiana: Arabic, Burmese, Mandarin, Vietnamese, and Spanish.

Students with Disabilities. Indiana established procedures to ensure the inclusion in statewide testing of all public elementary and secondary school students with disabilities. Federal and state laws require that all students participate in the state testing system. In Indiana, a student with an Individualized Education Program (IEP) will participate in ILEARN with the appropriate testing supports and accommodations prescribed by the IEP. If required by the student's IEP, the student will participate in Indiana's Alternate Measure (I AM). Per the Individuals with Disabilities Education Improvement Act (IDEA) and Title 5 Article 7-Special Education, published December 2014 by the Indiana State Board of Education, decisions regarding the appropriate assessment for a student with disabilities are made annually by the student's IEP team. These decisions are based on the student's curriculum, present levels of academic achievement, functional performance, and learning characteristics. Decisions cannot be based on program setting, category of disability, percentage of time in a particular placement or classroom, or any considerations regarding a school's Adequate Yearly Progress (AYP) designation.

Indiana does not have an opt-out policy for statewide assessments. IDOE advised schools to maintain documentation locally in the event a student is unable to participate for any reason in one or more ILEARN assessments. IDOE recommended schools document relevant information (e.g., test(s) not completed, reason for nonparticipation, efforts to communicate with parents) and include any supporting documentation (e.g., physician's note) for use with grade promotion, IEP goals, accountability appeals, graduation criteria, etc.

1.26.1 ADMINISTRATIVE ROLES

CTCs, NPSTCs, STCs, and TAs each had specific roles and responsibilities in the online testing systems. See the *Test Delivery System (TDS) User Guide* (Appendix 5-D) for their specific responsibilities before, during, and after testing.

CTCs

CTCs were responsible for coordinating testing at the corporation level, ensuring that the STCs in each school were appropriately trained and aware of policies and procedures, and ensuring that they were trained to use CAI's systems.

CITCs

CITCs were responsible for ensuring that testing devices were properly configured to support testing and for coordinating participation in the 2023–2024 systems readiness test (SRT). All schools were required to complete the SRT to prepare for online testing. The SRT was a simulation of online testing at the state level that ensured student testing devices and local school networks were correctly configured to support online testing.

NPSTCs

NPSTCs were responsible for coordinating testing at the school level for non-public schools, ensuring that the STCs within the school were appropriately trained and aware of policies and procedures, and that the STCs were trained to use CAI's systems.

STCs

Before each administration, STCs and CTCs were required to verify that student eligibility was correct in TIDE, and that any accommodations or test settings were correct. To participate in a computer-based online test, students had to be listed as eligible for that test in TIDE. See the *Test Information Distribution Engine (TIDE) User Guide* (Appendix 5-C) for more information.

STCs were responsible for ensuring that testing at their schools was conducted in accordance with the test security measures and other policies and procedures established by IDOE. STCs were primarily responsible for identifying and training TAs. STCs who worked with technology coordinators to ensure that computers and devices were prepared for testing and technical issues were resolved to ensure a smooth testing experience for the students. During the test window, STCs monitored testing progress, ensured that all students participated as appropriate, and handled testing issues as necessary by contacting the CAI Help Desk.

TAs

To be certified as a TA, educators need to complete an online Test Administrator Certification Course (Appendix 5-G). TAs administered the ILEARN assessment to students as well as a practice test session prior to the assessment.

TAs were responsible for reviewing necessary user manuals and user guides to prepare the testing environment and ensure that students did not have unauthorized books, notes, scratch paper, or electronic devices. They were required to administer the ILEARN assessment according to the directions found in the guide. TAs were required to report to the STC any deviation in test administration, at which time the STC was required to report it to the CTC. Then, if necessary, the CTC was to report it to IDOE.

TAs also ensured that the only available resources accessible to students were those allowed for specific ILEARN test administrations.

1.26.2 ONLINE ADMINISTRATION

The *Test Delivery System (TDS) User Guide* (Appendix 5-D) provided instructions for creating test sessions; monitoring sessions; verifying student information; assigning test accommodations; and starting, pausing, and submitting tests. The *Technology Guide* (Appendix 5-U) provided information about hardware, software, and network configurations to run CAI's various testing applications.

Personnel involved with statewide assessment administration played an important role in ensuring the validity of the assessment by maintaining both standardized administration conditions and test security.

1.26.2.1 Test Participation

There are circumstances in which a student did not participate in an expected assessment or participated in an assessment but in a non-standard way. In such instances, participation codes control and document how the test record is handled for reporting aggregates and accountability calculations. Available participation codes and descriptions are presented in Table 91. For more information on test participation, please refer to the *Test Information Distribution Engine (TIDE) User Guide*, presented as Appendix 5-C.

Table 91: Participation Codes and Their Descriptions

Participation Code	State	Federal	Description
101: Did Not Test	Countable for Participation only	Countable for Participation only	Student was enrolled at the school and eligible to test (with or without reasonable accommodations) but did not test.
103: ELL First Year in U.S. April 15 or Later	Not Countable	Not Countable	The student is an English language learner (ELL) and first enrolled in the U.S. on or after April 15 of current school year. Student is not required to test, but testing is made available.
104: ELL First Year in U.S. Before April 15	Counted for Participation only	Counted for Participation only	The student is ELL and first enrolled in the U.S. before April 15 of current school year. Student must take ELA, mathematics, and science.
205: ELL in Second Year of Enrollment	Counted in Participation and Growth	Counted in Participation and Growth	Student is ELL and first enrolled in the U.S. during the 2019–2020 school year. Student must take ELA, mathematics, and science.

Participation Code	State	Federal	Description
106: Student Refused to Test	Countable	Countable	Student refuses to start the assessment or refuses to complete at least six items of the assessment.
107: Excused for Health Emergency	Not Countable	Not Countable	Student is unable to test during the testing window due to an unanticipated health circumstance.
108: Course Instruction Not Complete	Not Countable	Not Countable	Student will not complete the relevant course instruction during the current academic year.
109: Course Not Provided	Not Countable	Not Countable	Student did not take a course associated with the assessment (e.g., student is assigned a test for a course they did not take at any time during the current school year).
110: Test Has Already Been Taken	Not Countable	Not Countable	Student has already taken the same assessment during a previous administration year.
111: IDOE Excused – Approval Needed	Not Countable	Not Countable	Requires IDOE authorization. Used in rare circumstances to capture irregular test circumstances.
112: Student Transferred Before Testing Window	Not Countable	Not Countable	Student transferred out of school before the LEA had a reasonable opportunity to administer the assessment.
200: Standard Participation	Countable	Countable	Student took the assessment under normal circumstances.
201: Accommodated	Countable	Countable	Student took the assessment with allowed accommodation(s).
202: Modified	Counted for Participation only	Counted for Participation Only	Student took the assessment with non-allowed modifications which interfere with the validity/reliability of the test.
203: Invalidated	Not Countable	Not Countable	LEA determines that the test was spoiled or invalid (e.g., student cheated; TA broke protocol).
204: Parental Exclusion*	Not Countable	Countable	A parent or guardian has requested in writing that the student be exempt from the assessment.
208: Test System Irregularity	Not Countable	Not Countable	The test event was interrupted by a system error without reasonable opportunity to reset or re-open the test. IDOE approval required.
209: Incorrect Course Code Assigned	Countable	Countable	An incorrect course code or grade was assigned, triggering an incorrect test. LEA correction of the course code is required.

1.26.2.2Scheduling Make-Up Testing and Test Completion Sessions

Test completion sessions could include students working on different tests.

Unexpected circumstances (e.g., fire drills, power failures) could interrupt testing. Test completion sessions could be scheduled when normal conditions were restored. Interruptions could not reduce the total amount of time students were given to complete tests.

After a test had been paused for 20 minutes, the student could no longer view or modify responses from that testing session. Students could not view or change prior answers during a make-up session. A make-up or completion session was only to finish the remaining portions of the test. This limit did not apply to the ELA writing test, which could be modified up to the point of submission.

5.1.2.3 Test Irregularities

On rare occasions, a non-standard situation arose during test administration. Three ways to account for irregularities were provided. Steps for dealing with test irregularities are outlined in more detail in the sections on Appeals or Appeal Requests in the *TIDE User Guide*.

- **Reset a Test.** Resetting a test eliminates all responses for a student. When that student logged in to the test again, the test would start over. Resetting could only be implemented in situations where the test could not be appropriately completed as is (e.g., two students accidentally log in to each other's test, a student requiring braille was not given the accommodation). A test could never be reset to give a student a second opportunity.
- **Grace Period Extension.** Extending a test's grace period gives a student access to his or her previous responses. This extension could be granted if a test session was interrupted unexpectedly (e.g., fire drill, lockdown). The grace period extension could not be applied if the test session ended normally or if the student was given time to review his or her answers before logging out of a test.
- **Invalidate a Test.** Tests could be invalidated when a student's performance was not an accurate measure of his or her ability (e.g., the student cheated, used inappropriate materials). If a test was invalidated, the student was not given another opportunity to take the test. Invalidating a test required the approval of an LEA-level user.
- **Reopen a Test.** Reopening a test changed the test's status from completed or reported to paused. This capability was useful if a student accidentally submitted a test before reviewing it. After the test was reopened, a student could resume testing. A test was not reopened once a student saw a score.
- **Reopen a Test Segment.** Reopening a test segment allowed a student to return to a prior segment in cases where the student moved to the next segment in error. This could occur on both summative and interim mathematics grade 6 tests

or summative writing tests. After the test segment was reopened, a student could return to the prior segment and complete his or her work.

1.26.3 ACCOMMODATED TEST ADMINISTRATION

The ILEARN assessments make available to students three categories of assessment tools and supports, which may be embedded or non-embedded in TDS: universal features, designated features and accommodations.

Universal features are available in TDS to all students taking ILEARN assessments. These features include. During the tests, students can zoom in and zoom out to increase or decrease the size of text and images, highlight items and passages (or sections of items and passages), cross out response options by using the strikethrough function, use a notepad to make notes, and mark a question for review using the flag function.

Designated features, such as the ability to select an alternate background and font color, mouse pointer size and color, and font size before testing, as well as glossaries that provide definitions for approved words in a second language, are available for use by any student for whom the need has been indicated by an educator, or team of educators with parent/guardian and student.

Accommodations are supports provided to students with disabilities enrolled in public schools with current IEPs or Section 504 Plans, as well as to students identified as ELs. All Indiana state assessments have appropriate accommodations available to make test content accessible to students with disabilities and ELs, including ELs with disabilities. The accommodations available for eligible students participating in the ILEARN assessments are described in the ILEARN TAMs (Appendix 5-B,5-P,5-R), which were accessible to schools before and during testing in the Resources section of the ILEARN Portal. A comprehensive list of accommodations available for eligible students with IEPs, Section 504 Plans, or Individual Learning Plans participating in online assessments is given in the in the *Test Information Distribution Engine* (*TIDE*) *User Guide* (Appendix 5-C).

1.26.4 ALLOWABLE RESOURCES FOR ONLINE TESTING

Table 92 provides a list of the designated features and accommodations and that were offered in the 2023–2024 administration. The *Online Test Delivery System (TDS) User Guide* can be found on the ILEARN portal (Appendix 5-D) and provides instructions on how to access and use these features.

Table 92: Designated Features and Accommodations Available in 2023–2024 for ILEARN

Designated Features	Accommodations
Embedded	

Color contrast (Onscreen)	American Sign Language (ASL)
Glossaries (Language)	Audio Transcriptions
Spanish	Calculator
Masking	Closed Captioning
Mouse Pointer	Permissive Mode
Print Size	Print-on-Demand
Translation	Streamline
	Text-to-Speech Except Reading Comprehension
	Text-to-Speech Including Reading Comprehension
	Refreshable Braille

<div align="center">

Non-Embedded

</div>

Assistive technology to Magnify/Enlarge	Braille Transcript for Audio Items
Access to Sound Amplification Program	Paper Booklet
Special Furniture or Equipment for Viewing Test	Large Print Booklet
Special Lighting Conditions	Read-Aloud to Self
Time of Day for Testing Altered	Read-Aloud Script for Paper Booklet*
Color Acetate Film for Paper Assessments	Scribe
	Speech-to-Text
	Tested Individual
	Interpreter for Sign Language
	Braille Booklet
	Multiplication Table
	Hundreds Chart
	Additional Breaks
	Bilingual Word-to-Word Dictionary
	Spanish Booklet
	Calculator
	Multiplication Table

See Appendix 5-E for a complete list of the Read-Aloud Scripts available to students during the 2023–2024 ILEARN assessments.

The TA and the School Test Coordinator (STC) were responsible for ensuring that arrangements for appropriate accommodations were made before the test administration dates. Requests for any non-standard accommodations were recorded under a Special Requests section in the Test Information Distribution Engine (TIDE) and required IDOE approval. IDOE provided a separate, supplemental accessibility manual—the *Indiana Assessments Policy Manual* (Appendix 5-F)—for individuals involved in administering tests to students who required accommodations. Students who required online accommodations (e.g., text-to-speech) were provided the opportunity to participate in practice activities for the statewide assessments with appropriate allowable accommodations. Test Administrators identified test settings and accommodations in TIDE before students could start an online test session. Some settings and accommodations could not be changed once a student started a test. IDOE approved updates to incorrectly assigned accommodations before any updates were applied to subsequent student testing. IDOE also determined which testing attempts to invalidate prior to score reporting.

Starting in the 2020–2021 school year, TTS was expanded and split into two separate accommodations for ELA. The 2023–2024 tests continued this pattern of accommodations wherein one accommodation read aloud only content that was not designed to assess reading compression. The second accommodation read aloud all test content, including those items and passages designed to assess reading compression. As a result, students who participated in ILEARN ELA in grades 3 through 8 could be assigned to either of two TTS modalities:

- TTS **except** for items and passages measuring reading comprehension; or
- TTS **including** items and passages measuring reading comprehension.

Case conference committees determined which of these accommodation modalities was appropriate for their students requiring TTS. Guidance to schools and case conference committees on assigning TTS for all items including reading comprehension was provided in the *2023–2024 Accessibility and Accommodations Guidance* manual (Appendix 5-N), as well as in periodic communications with the field.

If an ELL or a student with an IEP or Section 504 Plan used any accommodations during the test administration, this information was recorded by the Test Administrator (TA) in the required administration information and was captured by CAI in the database of record (DoR). CAI included this data in the state output student data score files (SDFs) provided to IDOE at the end of each test administration. Guidelines recommended for making accommodation decisions included the following:

- Accommodations should facilitate an accurate demonstration of what the student knows or can do.
- Accommodations should not provide the student with an unfair advantage or negate the validity of a test; accommodations must not change the underlying skills that are being measured by the test.
- Accommodations must be the same or nearly the same as those needed and used by the student in completing daily classroom instruction and routine assessment activities.
- Accommodations must be necessary for enabling the student to demonstrate knowledge, ability, skill, or mastery.

Students with disabilities not enrolled in public schools or receiving services through public school programs who required accommodations to participate in a test administration were permitted access to accommodations if the following information was provided:

- Evidence that the student had been found eligible as a student with a disability as defined by Individuals with Disabilities Education Improvement Act (IDEA).

Documentation that the requested accommodations had been regularly used for instruction. The following accommodations were available for eligible students with IEPs or Section 504 Plans participating in paper-based assessments:

- Contracted Unified English Braille (UEB) and Nemeth Code for Mathematics.

- Uncontracted braille and Nemeth Code for Mathematics.

The IDOE monitors test administration in corporations and schools to ensure that appropriate assessments, online or paper-based, with or without accommodations, are administered to all students with disabilities and ELs and are consistent with Indiana's policies.

1.27 TRAINING AND INFORMATION FOR TEST COORDINATORS AND ADMINISTRATORS

IDOE established and communicated a clear, standardized procedure to educators and key personnel involved with the administration of ILEARN assessments, including the process for giving students access to accommodations. Key personnel involved with ILEARN administration included Corporation Test Coordinators (CTCs), Non-Public School Test Coordinators (NPSTCs), Corporation Information Technology Coordinators (CITCs), STCs, and TAs. The roles and responsibilities of staff involved in testing are further detailed in the next section.

TAs were required to complete CAI's online TA Certification Course before administering any tests. There were also several training modules developed by CAI in collaboration with IDOE to facilitate test administration. These modules included topics on CAI systems, test administration, and accessibility and accommodations. These modules are included in this chapter's appendices.

TAMs and user guides were available online for school and corporation staff. The *Test Delivery System (TDS) User Guide* (Appendix 5-D) was designed to familiarize TAs with TDS and contained tips and screenshots throughout the text. The user guide described:

- Steps to take prior to accessing the system and logging in;
- Navigation instructions for the TA Interface application;
- Details about the Student Interface, used by students for online testing;
- Instructions for using the training sites available for TAs and students; and
- Information on secure browser features and keyboard shortcuts.

The User Support sections of both the *Test Delivery System (TDS) User Guide* (Appendix 5-D) and the *Test Information Distribution Engine (TIDE) User Guide* (Appendix 5-C) provided instructions that addressed technology challenges that could occur during test administration. The CAI Help Desk collaborated with IDOE to provide support to Indiana schools as they administered the state assessment.

1.27.1 MANUALS AND USER GUIDES

The list of webinars and training resources available to corporations and schools for the 2023–2024 ILEARN administration is provided below. All training materials were available online at the ILEARN Portal. PDFs of these resources have also been included as appendices in this technical report. Test administration resources comprising various tutorials and documents (e.g., user guides, manuals, quick guides) also were available through the ILEARN Portal.

- **Test Administrator Certification Course:** All educators who administered the ILEARN assessment were required to complete the online TA Certification Course (Appendix 5-G).
- **What is a Standards-Based Computer-Adaptive Test (CAT) Webinar:** This online module described computer-adaptive-testing and the student test experience (Appendix 5-H).
- **Why It's Important to Assess Webinar Module:** This online module illustrated the importance of statewide testing (Appendix 5-I).
- **Centralized Reporting System (CRS) Webinar Module:** This module provided a general overview of ORS where student scores, including individual scores and aggregate scores, are displayed after students complete the ILEARN assessments (Appendix 5-J).

Table 93 presents the list of available user guides and manuals related to ILEARN administration. The table also includes a short description of each resource and its intended use. PDFs of these eight publications have also been included in this technical report as appendices.

<div align="center">Table 93: User Guides and Manuals</div>

Resource	Description
Test Delivery System (TDS) User Guide (Appendix 5-D)	This user guide supports TAs who manage testing for students participating in the ILEARN practice tests, released item repository tests, and operational tests.
Technology Guide (Appendix 5-U)	This HTML guide on the Indiana Assessment Portal provides information about hardware, software, and network configurations for running various testing applications.
Online Practice Test User Guide (Appendix 5-K)	This user guide provides an overview of the ILEARN Practice Test.
Released Items Repository (RIR) Quick Guide (Appendix 5-A)	The Released Items Repository (RIR) Quick Guide provides instructions for taking and administering demo tests in the RIR.
Test Information Distribution Engine (TIDE) User Guide (Appendix 5-C)	This user guide describes the tasks performed in the Test Information Distribution Engine (TIDE) for ILEARN assessments.
Assistive Technology Manual (Appendix 5-L)	This manual provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with special accessibility needs complete online tests in the Test Delivery System (TDS). It includes lists of supported devices and applications for each type of assistive technology that students may need, as well as setup instructions for the assistive technologies that require additional configuration to work with TDS.
Centralized Reporting System (CRS) User Guide (Appendix 5-M)	This user guide provides an overview of the different features available to educators to support viewing student scores and downloadable score data files for the ILEARN assessment.

| *Accessibility & Accommodations Information for Statewide Assessments (Appendix 5-N)* | The accessibility manual establishes the guidelines for the selection, administration, and evaluation of accessibility supports for instruction and assessment of all students, including students with disabilities, English learners (ELs), ELs with disabilities, and students without an identified disability or ELL status. |
| *ILEARN Test Administrator's Manual (TAM) (Appendix 5-B)* | The ILEARN Test Administrator's Manual (TAM) provides an overview of the specific roles and responsibilities required before, during, and after testing for ILEARN 3-8, ILEARN Biology and ILEARN U.S. Government. |

1.28 TEST SECURITY

Test security involves maintaining the confidentiality of test questions and answers and is critical in ensuring the integrity of a test and the validity of test results. Indiana has developed an appropriate set of policies and procedures to prevent test irregularities and ensure test result integrity. These include maintaining the security of test materials, assuring adequate trainings for everyone involved in test administration, outlining appropriate incident-reporting procedures, detecting test irregularities, and planning for investigation and handling of test security violations.

All personnel who administered ILEARN assessments were required to complete the online TA Certification Course accessible through the ILEARN portal. TDS was configured so that personnel could not administer tests without first completing the TA Certification Course. Access to the course was limited to the following roles: CTC, Co-Op, CITC, NPSTC, STC, and TA.

The test security procedures for ILEARN included the following:

- Procedures to ensure security of test materials;
- Procedures to investigate test irregularities; and
- Guidelines to determine if test invalidation was appropriate/necessary.

1.28.1 STUDENT-LEVEL TESTING CONFIDENTIALITY

To support these policies and procedures, IDOE leveraged security measures within CAI systems. For example, students taking the ILEARN assessments were required to acknowledge a security statement confirming their identity and acknowledging that they would not share or discuss test information with others. Additionally, students taking the online assessments were logged out of a test within the CAI Secure Browser after 20 minutes of inactivity.

In developing the *ILEARN Test Coordinator's Manual* (Appendix 5-T) and the ILEARN TAMs (Appendix 5-B), IDOE and CAI ensured that all test security procedures were available to everyone involved in test administration. Each manual included protocols for reporting any deviations in test administration.

If IDOE determined that an irregularity in test administration or security occurred, it acted based upon approved procedures including, but not limited to, the following:

- Invalidation of student scores; and
- A requirement for the corporation or school to administer a breach form

1.28.2 Maintaining Test Security

Before test materials were finalized, test items and performance tasks went through multiple reviews, including review by various committees. Maintaining security of all test content was of high priority before, during, and after committee meetings. Printed copies of items and performance task content were not provided to educator participants. Any secure materials created or distributed during the meetings were collected and destroyed following the meetings.

All test items and performance tasks, test materials, and student-level testing information were deemed secure and were required to be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of assessment questions, prompts, and student results. Any deviation in test administration was required to be reported to protect the validity of the assessment results.

Secure handling of all test materials was required before, during, and after test administration. After any administration, initial or make-up test session, secure materials (e.g., scratch paper) were required to be returned immediately to the STC and placed in locked storage. Secure materials were never to be left unsecured and were not permitted to remain in classrooms or be removed from the school's campus overnight. Secure materials that did not need to be returned to the print vendor for scanning and scoring were to be destroyed securely following outlined security guidelines but were not allowed to be discarded in the trash. In addition, any monitoring software that might have allowed test content on student workstations to be viewed or recorded on another computer or device during testing had to be disabled.

It was considered a testing security violation for authorized corporation or school personnel to fail to follow security procedures set forth by the IDOE, and no individual was permitted to do the following:

- Read, copy, share or view the passages, test items, or performance tasks before, during, or after testing;
- Explain the passages, test items, or performance tasks to students;
- Change or otherwise interfere with student responses to test items or performance tasks;
- Copy or read student responses; and
- Cause achievement of schools to be inaccurately measured or reported.

All accommodated assessment books (regular print, large print, braille, and Spanish) were treated as secure documents, and processes were in place to protect them from loss, theft, and reproduction of any kind.

A secure browser was required to access the online ILEARN tests. The CAI Secure Browser provided a secure environment for student testing by disabling hot keys, copy,

and screen capture capabilities and preventing access to the desktop (e.g., Internet, email, and other files or programs installed on school machines). Users could not access other applications from within the CAI Secure Browser, even if they knew the keystroke sequences.

Students were not able to print from the CAI Secure Browser unless testing with the Print-on-Demand accommodation. Print-on-Demand allows students to participate in computer-adaptive assessments while using paper to read and respond to items when necessary. This accommodation requires a one-on-one testing environment in a secure location and additional test security management. Printed content is securely destroyed at the local level once testing is complete, in accordance with established protocols.

The CAI Secure Browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. Review Appendix 5-D of the *Online Test Delivery System (TDS) User Guide* for further details.

1.28.3 ONLINE MANAGEMENT SYSTEM

CAI has built-in security controls in all its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of CAI's systems encrypt data at rest and in transit. ILEARN data resides on servers at Rackspace, CAI's online hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. Staff at both CAI and Rackspace receive formal training in security procedures to ensure that they know the procedures and implement them properly.

Hardware firewalls and intrusion detection systems protect CAI networks from intrusion. CAI's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. All CAI's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality consistent with the Family Educational Rights and Privacy Act (FERPA).

CAI's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. CAI maintains logs of key activities and indicators, including data backup, server response time, user accounts, system events and security, and load test results.

1.28.3.1 Secure System Design

CAI has developed a custom single sign-on application that is made available in Indiana's secure portal. This application is used to support access to CAI's systems in accordance with Indiana's user ID and password policy. Authorized users can log in to Indiana's single sign-on using their current user IDs and passwords and can be redirected to CAI's portal, where they have access to CAI's secure applications such as TIDE, the TDS, and the Reporting System. Nightly backups protect the data. 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 they will need to rerun the backup. The system can withstand failure of almost any component with little or no interruption of service.

CAI's hosting provider, Rackspace, has redundant power generators that can continue to operate for up to 60 hours without refueling. With multiple refueling contracts in place, these generators can operate indefinitely. Rackspace partners with nine different network providers, providing multiple, redundant data routes. Every installation is served by multiple servers, any one of which can take over for an individual test upon failure of another.

CAI's architecture ensures data are recoverable at all times. Each disk array is internally redundant, with multiple disks containing each data element. Immediate recovery from failure of any individual disk is performed by accessing the redundant data on another disk. CAI maintains support and maintenance agreements through our hosting provider for all hardware used by our systems.

1.28.3.2 System Security Components

CAI has built-in security controls in all its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of CAI's systems encrypt data at rest and in transit.

Physical Security

IN data reside on servers at Rackspace, CAI's hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. All access is keycard controlled, and sensitive areas require biometric scanning.

Secure data are processed at CAI facilities and are accessed from CAI machines. CAI's servers are in a secure, climate-controlled location with access codes required for entry. Access to our servers is limited to our network engineers, all of whom, like all CAI employees, have undergone rigorous background checks.

Staff at both CAI and Rackspace receive formal training in security procedures to ensure that they know the procedures and implement them properly. CAI and Rackspace protect data from accidental loss through redundant storage, backup procedures, and secure off-site storage.

Network Security

Hardware firewalls and intrusion detection systems protect our networks from intrusion. They are installed and configured to prevent access for services other than hypertext transfer protocol secure (HTTPS) for our secure sites.

CAI's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts.

Software Security

All of CAI's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with Indiana's privacy laws, FERPA, and other federal laws.

CAI's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. Different states interpret FERPA differently, and our system is designed to support these interpretations flexibly. CAI has worked with IDOE to maintain data security according to its specifications.

CAI maintains logs of key activities and indicators, including data backup, server response time, user accounts, system events and security, and load test results. In addition, CAI runs automated functional tests of our TDS every morning, and logs from these runs are available for at least one week from the time of the run.

CAI psychometricians monitor the quality and performance of test administrations statewide through a series of quality assurance (QA) reports. The QA reports provide information on item behavior, blueprint match rates, and item exposure rates, and also provide cheating analysis reports.

1.29 DATA FORENSICS PROGRAM

CAI's quality monitoring (QM) system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QM system, and any anomalies (such as tests not meeting blueprint, unexpected test lengths, or other unlikely issues) are flagged. CAI psychometricians run quality assurance reports and alert the program team of any issues. The forensic analysis report from the QM system flags unlikely patterns of behavior in testing administrations aggregated at the following levels: test administration, TA, and school.

Item statistics and blueprint reports were run and reviewed weekly during the 2023–2024 ILEARN test windows. In addition, response change analyses for multiple-choice and multiple-select items were conducted. The last and next to last (if it existed) responses were compared and students or aggregates were flagged if the number or average number of wrong to right response changes was above the flagging criteria.

CAI psychometricians monitored testing anomalies throughout the test window. A variety of evidence was collected for the evaluation. These evidences include blueprint match, unusual or much longer test times as compared to the state average, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be set by IDOE. While analyses used to detect the testing anomalies could be run anytime within the test window, analyses relying on state averages are typically held until the close of the test window to ensure final data is being used.

The lead psychometrician will alert the program team leads if any unexpected results are identified in order to immediately resolve any issues.

CAI also contracts with a third-party vendor, Caveon, to detect security breaches. IDOE, CAI, and Caveon conducted routine check-in calls prior to, during, and after each assessment window.

CAI and IDOE utilize an escalation protocol to address system and security concerns. A Teams hotline was established for each assessment window as an immediate means to address concerns or work through irregularities. At the beginning of each window, afternoon check-ins were established to review test administration and frequently asked questions and to address specific test irregularities.

1.30 TRACKING AND RESOLVING TEST IRREGULARITIES

Throughout the test window, TAs were instructed to report breaches of protocol and testing irregularities to the appropriate STC. Test irregularity requests were submitted, as appropriate, through the IDOE Testing Irregularity Report. IDOE instructed schools to submit a specific action request in TIDE, if appropriate.

TIDE allowed CTCs, NPSTCs, and STCs to request action to a test (e.g., Re-open test, re-open test segment) in response to a test irregularity that occurred in the testing environment. In many cases, schools were required by IDOE to provide formal documentation of test irregularities before creating an Irregularity Request in TIDE.

CTCs, NPSTCs, STCs, and TAs had to discuss the details of a test irregularity to determine whether test invalidation was appropriate. CTCs, NPSTCs, and STCs were required to submit to IDOE a T*esting Concerns and Security Violations Report* when invalidating any student test in response to a test security breach or interaction that compromised the integrity of the student's test administration.

During the test window, TAs were also required to immediately report any test incidents (e.g., disruptive students, loss of Internet connectivity, student improprieties) to the STC. A test incident could include testing that was interrupted for an extended period due to a local technical malfunction or severe weather. STCs notified CTCs or NPSTCs of any test irregularities that were reported. CTCs or NPSTCs were responsible for completing test invalidations via TIDE. Schools managed the invalidation process based on local decisions or guidance from IDOE regarding test irregularities or test security concerns. This information was stored in TIDE for the school year and remained available until TIDE was updated for the 2023–2024 school year. Table 94 presents examples of test irregularities and test security violations.

Table 94: Examples of Test Irregularities and Test Security Violations

Description
Student(s) making distracting gestures/sounds or talking during the test session that creates a disruption in the test session for other students.

Student(s) leaving the test room without authorization.
TA or Test Coordinator leaving related instructional materials on the walls in the testing room.
Student(s) cheating or providing answers to each other, including passing notes, giving help to other students during testing, or using handheld electronic devices to exchange information.
Student(s) accessing or using unauthorized electronic equipment (e.g., cell phones, smart watches, iPods, or electronic translators) during testing.
Disruptions to a test session such as a fire drill, school-wide power outage, earthquake, or other acts.
TA or Test Coordinator failing to ensure administration and supervision of the assessments by qualified, trained personnel.
TA giving incorrect instructions.
TA or Test Coordinator giving out his or her username/password (via email or otherwise), including to other authorized users.
TA allowing students to continue testing beyond the close of the test window.
TA or teacher coaching or providing any other type of assistance to students that may affect their responses. This includes both verbal cues (e.g., interpreting, explaining, or paraphrasing the test items or prompts) and nonverbal cues (e.g., voice inflection, pointing, or nodding head) to the correct answer. This also includes leading students through instructional strategies such as think-aloud, asking students to point to the correct answer or otherwise identify the source of their answer, requiring students to show their work to the TA, or reminding students of a recent lesson on a topic.
TA providing students with unallowable materials or devices during test administration or allowing inappropriate designated features and/or accommodations during test administration.
TA providing a student access to another student's work/responses.
TA or Test Coordinator modifying student responses or records at any time.
TA providing students with access to a calculator during a portion of the assessment that does not allow the use of a calculator.
TA uses another staff member's username and/or password to access vendor systems or administer tests.
TA uses a student's login information to access practice tests or operational tests.

6.	**S**CALING AND **E**QUATING

1.31 ITEM RESPONSE THEORY PROCEDURES

1.31.1 CALIBRATION OF ILEARN ITEM BANKS

The embedded field-test design, in conjunction with the adaptive administration of operational tests, produces item response data in a sparse data matrix. The items in the sparse data matrix were concurrently calibrated by grade and content area, with parameter estimates for operational items fixed to their bank values and field-test items calibrated under that constraint. From spring 2019 to spring 2023, the field-test items were calibrated using the IRTPRO software, version 4.2. Starting from spring 2024, the field-test items of ELA, mathematics, and social studies are calibrated using the IRTPRO software, version 6.0. The operational field-items of science were calibrated using CAI's proprietary software CAIRT. In each calibration, the parameters of the operational items were fixed to their bank values, and the item parameters of the field-test items, as well as the mean and variance of each group, were estimated.

1.31.2 ESTIMATING STUDENT ABILITY USING MAXIMUM LIKELIHOOD ESTIMATION

1.31.2.1 ELA and Mathematics — Maximum Likelihood Estimation

The *ILEARN* assessments are scored using maximum likelihood estimation (MLE). MLEs are useful since an estimate of a person's ability can be obtained after one item has been answered correctly and one item has been answered incorrectly. With number-correct scoring, the test must be completed before an assessment of ability can be computed. This "early" estimate of ability is what allows tests to be adaptive.

However, when all the items administered at a specific point in the test have been answered correctly or incorrectly, the estimate of ability goes to positive or negative infinity, respectively, or the highest or lowest score. This has implications for determining what constitutes a completed test. Theoretically, with maximum likelihood scoring, the student could answer the first item correctly, quit the test, and receive the maximum score. To avoid this, the definition for a complete test needs to be based on something in addition to a minimum number of items attempted, as is often the case with number-correct scored tests.

Ability estimates were generated using pattern scoring, a method that scores students depending on how they answer individual items.

The likelihood function for generating maximum likelihood estimates (MLEs) is based on a mixture of item models and can therefore be expressed as

$$L(\theta) = L(\theta)^{2PL} L(\theta)^{CR},$$

where

$$L(\theta)^{2PL} = \prod_{i=1}^{N_{2PL}} P_i^{z_i} Q_i^{1-z_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N_{CR}} \frac{exp \sum_{l=1}^{z_i} Da_i(\theta-b_{il})}{1+ \sum_{h=1}^{m_i} exp \sum_{l=1}^{h} Da_i(\theta-b_{il})}$$

$$p_i = \frac{1}{1+exp[-Da_i(\theta-b_i)]},$$

$$q_i = 1 - p_i$$

and where a_i is the slope of the item response curve (i.e., the discrimination parameter), b_i is the location parameter, z_i is the observed response to the item, i indexes item, h indexes step of the item, m_i is the maximum possible score point, b_{il} is the ith step for item i with m total categories, and $D = 1.7$.

A student's theta (i.e., MLE) is defined as $log(L(\theta))$ given the set of items administered to the student.

Derivatives

Finding the maximum of the likelihood requires an iterative method, such as Newton-Raphson iterations. The estimated MLE is found via the following maximization routine:

$$\theta_{t+1} = \theta_t - \frac{\frac{\partial lnL(\theta_t)}{\partial \theta_t}}{\frac{\partial^2 lnL(\theta_t)}{\partial^2 \theta_t}},$$

where

$$\frac{\partial lnL(\theta)}{\partial \theta} = \frac{\partial lnL(\theta)^{2PL}}{\partial \theta} + \frac{\partial lnL(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 lnL(\theta)}{\partial^2 \theta} = \frac{\partial^2 lnL(\theta)^{2PL}}{\partial^2 \theta} + \frac{\partial^2 lnL(\theta)^{CR}}{\partial^2 \theta}$$

$$\frac{\partial lnL(\theta)^{2PL}}{\partial \theta} = \sum_{i=1}^{N_{2PL}} Da_i \frac{(z_i-p_i)(p_i)}{p_i}$$

$$\frac{\partial^2 lnL(\theta)^{2PL}}{\partial^2 \theta} =- \sum_{i=1}^{N_{2PL}} D^2 a_i^2 \frac{p_i q_i}{1}\left(1 - \frac{z_i}{p_i^2}\right)$$

$$\frac{\partial lnL(\theta)}{\partial \theta}^{CR} = \sum_{i=1}^{N_{CR}} Da_i \left(z_i - \frac{\sum_{h=1}^{m_i} h\,exp\left(\sum_{l=1}^{j} Da_i(\theta-b_{il})\right)}{1+\sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta-b_{il})\right)} \right)$$

$$\frac{\partial^2 lnL(\theta)}{\partial^2 \theta}^{CR} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{h=1}^{m_i} h\,exp\left(\sum_{l=1}^{h} Da_i(\theta-b_{il})\right)}{1+\sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta-b_{il'})\right)} \right)^2 - \frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} Da_i(\theta-b_{il'})\right)}{1+\sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta-b_{il})\right)} \right),$$

and where θ_t denotes the estimated θ at iteration t. N_{CR} is the number of items that are scored using the Generalized Partial Credit Model (GPCM), and N_{2PL} is the number of items scored using the two-parameter logistic (2PL) model.

1.31.2.2 Science MLE

Student scores are obtained by marginalizing out the nuisance dimensions $\mathbf{u}_j$ from the likelihood of the observed response pattern $\mathbf{z}_j$ for student j,

$$\ell_i(\theta_j) = \log \int_{\mathbf{u}_j} P\left(\mathbf{\Sigma}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0},\ \right) d_j ,$$

and maximizing this marginalized likelihood function for θ_j. The Marginal Maximum Likelihood Estimation (MMLE) is a hybrid of the expected a posteriori (EAP) estimator (by marginalizing out the nuisance dimensions) and the MLE estimator (by maximizing the resulting marginal likelihood for θ). The marginal likelihood is maximized with respect to θ using the Newton Raphson method.

The proposed model reduces to the unidimensional Rasch model when the nuisance variances are zero for all g. Likewise, the proposed MMLE is equivalent to the MLE of the unidimensional Rasch model when all the nuisance variances are zero. This can be shown by using the variable transformation $v = {}^{-\frac{1}{2}} u$. Then we have

$$\int_{\mathbf{u}_j} P\left(\mathbf{\Sigma}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0},\ \right) d_j = \int_{\mathbf{v}_j} P\left(\mathbf{\Sigma}_j \middle| \theta_j,\ {}^{\frac{1}{2}}\mathbf{v}_j\right) N\left(\mathbf{v}_j \middle| \mathbf{v},\ \right) d_j .$$

If $\sigma_{u_g}^2 = 0$ for all g, then

$$\int_{\mathbf{u}_j} P\left(\mathbf{\Sigma}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0},\ \right) d_j = P\left(\ _j \middle| \theta_j\right) ,$$

which is the likelihood under the unidimensional Rasch model.

Derivatives

The marginal log likelihood function based on the IRT model with one overall dimension and one nuisance dimension for each grouping of assertions can be written as

$$l(\theta) = \sum_{i \in SA} log\left(P\left(z_i|\theta\right)\right) + \sum_{g=1}^{G} log\left\{\int Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right]N\left(0, \sigma_{u_g}^2\right)du_g\right\}$$

The first derivative of the marginal log likelihood function with respect to θ is

$$\frac{dl(\theta)}{d\theta} = \sum_{i \in SA} \frac{\frac{dP\left(z_i|\theta\right)}{d\theta}}{P\left(z_i|\theta\right)} + \sum_{g=1}^{G} \frac{\int\left\{Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right]\left(\sum_{i \in g} \frac{\frac{dP\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)}\right)N\left(0, \sigma_{u_g}^2\right)\right\}du_g}{\int\left\{Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right]N\left(0, \sigma_{u_g}^2\right)\right\}du_g},$$

and the second derivative of the marginal log likelihood function with respect to θ is

$$\frac{d^2l(\theta)}{d\theta^2} = \sum_{i \in SA}\left[\frac{\frac{P\left(z_i|\theta\right)}{d\theta^2}}{P\left(z_i|\theta\right)} - \left(\frac{\frac{P\left(z_i|\theta\right)}{d\theta}}{P\left(z_i|\theta\right)}\right)^2\right] + \sum_{g=1}^{G} \frac{\int Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right]\left(\sum_{i \in g} \frac{\frac{P\left(z_{ig}|\theta, u_g\right)}{d\theta}}{P\left(z_{ig}|\theta, u_g\right)}\right)^2 N\left(0, \sigma_{u_g}^2\right)du_g}{\int\left\{Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right]N\left(0, \sigma_{u_g}^2\right)\right\}du_g} + \sum_{g=1}^{G} \frac{\int Exp\left[\sum_{i \in g} log\left(P\left(z_{ig}|\theta, u_g\right)\right)\right]\left(\sum_{i \in g} \right.}{\int\left\{Exp\left[\sum_{i \in g} log\left(P\right.\right.}$$

Based on the above equations, we need to define only the ratios of the first and second derivatives of the item response probabilities with respect to θ to the response probabilities. For the Rasch testlet model, these are obtained as

$$p_i = \left(z_i = 1|\theta\right) = \frac{\left(\theta - b_i\right)}{1 + \left(\theta - b_i\right)} \ , \ q_i = \left(z_i = 0|\theta\right) = 1 - p_i,$$

and

$$p_{ig} = \left(z_{ig} = 1|\theta, u_g\right) = \frac{\left(\theta + u_g - b_i\right)}{1 + \left(\theta + u_g - b_i\right)} \ , \ q_{ig} = \left(z_{ig} = 0|\theta, u_g\right) = 1 - p_{ig}.$$

Therefore, we have,

$$\frac{\frac{dp_i}{d\theta}}{p_i} = q_i \ , \ \frac{\frac{dq_i}{d\theta}}{q_i} = -p_i,$$

$$\frac{\frac{dp_{ig}}{d\theta}}{p_{ig}} = q_{ig} \ , \ \frac{\frac{dq_{ig}}{d\theta}}{q_{ig}} = -p_{ig},$$

$$\frac{\frac{p_i}{d\theta^2}}{p_i} - \left(\frac{\frac{p_i}{d\theta}}{p_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{q_i}{d\theta^2}}{q_i} - \left(\frac{\frac{q_i}{d\theta}}{q_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{p_{ig}}{d\theta^2}}{p_{ig}} - \left(\frac{\frac{p_{ig}}{d\theta}}{p_{ig}}\right)^2 = - p_{ig}q_{ig}, \ and$$

$$\frac{\frac{q_{ig}}{d\theta^2}}{q_{ig}} - \left(\frac{\frac{q_{ig}}{d\theta}}{q_{ig}}\right)^2 = - p_{ig}q_{ig}.$$

1.31.2.3 Standard Errors of Measurement

The SEM of the MMLE score estimate is:

$$SEM(\hat{\theta}_{MMLE}) = \frac{1}{\sqrt{I(\hat{\theta}_{MMLE})}}$$

where $I(\hat{\theta}_{MMLE})$ is the observed information evaluated at $\hat{\theta}_{MMLE}$. The observed information is calculated as $I(\theta^2) = - \frac{d^2l(\theta)}{d\theta^2}$ where $\frac{d^2l(\theta)}{d\theta^2}$ is defined in the previous section on derivatives. Note that the calculation of the standard error of estimate depends on the unique set of items that each student answers and their estimate of θ. Different students have different SEMs, even if they have the same raw score and/or theta estimate.

1.31.3 CALIBRATING FIELD-TEST ITEMS ONTO THE ILEARN SCALE

Following the spring 2019 *ILEARN* assessments, IRT calibrations were completed that placed all items within a grade and subject on the same scale. More information about these calibrations can be found in the *ILEARN 2018–2019 Technical Report*. As of 2023–2024, all assessments are pre-equated.

For field-test item calibrations, all operational items were anchored to their bank values and field-test item parameters were estimated. Table 95 presents the number of students used in field-test calibrations for ELA, mathematics, and social studies, as well as number of students used in operational field-test calibrations of computer science items for science.

Table 95: Number of Students Used in Field-Test (ELA, Mathematics, Social Studies) and Operational-Field Test (Science) Calibrations

ELA		Mathematics		Science		Social Studies	
Grade	Calibration *N* Count	Grade	Calibration *N* Count	Grade	Calibration *N* Count	Grade	Calibration *N* Count
3	-	3	-				
4	82,623	4	-	4	82,643		
5	81,018	5	-			5	79,790

6	-	6	-	6	82,284		
7	-	7	-				
8	-	8	-				
						U.S. Government	231

1.32 *ILEARN* REPORTING SCALE (SCALE SCORES)

1.32.1 OVERALL PERFORMANCE

For 2023–2024, scale scores were reported for each student who took the *ILEARN* assessments. The scale scores were based on the operational items presented to the student and did not include any field-test items. The scale score is a linear transformation of the IRT ability estimate, θ:

$$SS = a*\theta + b,$$

where *a* is the slope and *b* is the intercept. Table 97 lists the scaling constants *a* and *b* for the *ILEARN* assessments.

ELA and mathematics were reported on a vertical scale. The IRT vertical scale was established by Smarter Balanced and formed by linking across grades using common items in adjacent grades. Grade 6 was used as the baseline, and each grade was successively linked onto the scale. More details about the vertical scaling methods can be found in Chapter 9 of the Smarter Balanced 2013–2014 Technical Report (Smarter Balanced, 2016). The slope and intercept used to transform the IRT ability estimate to a scale score are presented in Table 96 and are unique to Indiana and the *ILEARN* assessments.

Each science and social studies assessment was reported on a separate within-test scale.

Table 96: Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3–8	75	5500
Mathematics	3–8	75	6500
Science	4	15.583	166.711
Science	6	17.516	367.140
Science	Biology	17.909	563.622
Social Studies	5, U.S. Government	50	8500

1.32.1.1 LEXILE® and QUANTILE® Scores

ILEARN reports Lexile® and Quantile® measures with ELA and mathematics test scores. MetaMetrics provided conversion tables between ELA scale scores and Lexile®

measures and between mathematics scale scores and Quantile® measures for each grade and subject.

1.32.2 REPORTING CATEGORY PERFORMANCE

In addition to a total scaled score, performance on each reporting category is reported. Reporting category theta scores were calculated using either MLE or MMLE, depending on the assessment and based on the items contained in a particular reporting category. The same rules for scoring all correct and all incorrect cases were applied to reporting category scores.

1.32.2.1 Strengths and Weaknesses

For reporting categories, relative strengths and weaknesses were reported for each student at the reporting-category level. The difference between the proficiency cut score and the reporting category score plus or minus 1.5 times SE of the reporting category was used to determine the relative strengths and weaknesses.

The specific rules for mastery are as follows:

Below (Code = 1): if $round(SS_{rc} + 1.5 * SE(SS_{rc}), 0) < SS_p$

At/Near (Code = 2): if $round(SS_{rc} + 1.5 * SE(SS_{rc}), 0) \geq SS_p$ and $round(SS_{rc} - 1.5 * SE(SS_{rc}), 0) < SS_p$, a strength or weakness is indeterminable

Above (Code = 3): if $round(SS_{rc} - 1.5 * SE(SS_{rc}), 0) \geq SS_p$

SS_{rc} is the student's scale score on a reporting category; SS_p is the proficiency scale cut score (Level 3 cut scored); and $SE(SS_{rc})$ is the standard error of the student's scale score on the reporting category.

1.32.2.2 Standard-Level Aggregate Scores

Standard-level information was reported relative to the proficiency standard for tests that were adaptively administered. In spring 2024, standard-level information have been reported for the ELA, mathematics, and science assessments.

First, $p_{ij} = p(z_{ij} = 1)$ was defined, representing the probability that student j responded correctly to item i (z_{ij} represents the jth student's score on the ith item). For items with one score point, the 2PL IRT model was used to calculate the expected score on item i for student j with $\theta_{Level\,3\,cut}$ as:

$$E(z_{ij}) = \frac{exp(1.7*a_i(\theta_{Level\,3\,cut} - b_i))}{1 + exp(1.7*a_i(\theta_{Level\,3\,cut} - b_i))}.$$

For items with two or more score points, using the GPCM, the expected score for student *j* with a Level 3 cut score on item *i* with a maximum possible score of m_i was calculated as:

$$E\left(z_{ij}\right) = \sum_{l=1}^{m_i} \frac{l \exp\left(\sum_{k=1}^{l} 1.7*a_i\left(\theta_{Level\,3\,cut} - b_{i,k}\right)\right)}{1 + \sum_{l=1}^{m_i} \exp\left(\sum_{k=1}^{l} 1.7*a_i\left(\theta_{Level\,3\,cut} - b_{i,k}\right)\right)}.$$

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

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

Residuals are summed for items within a standard. The sum of residuals was divided by the total number of points possible for items within the standard, *S*:

$$\delta_{jS} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a standard score was computed by averaging individual student standard scores for the standard, across students of different abilities receiving different items measuring the same standard at different levels of difficulty,

$$\overline{\delta}_{Sg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jS},$$

and

$$se\left(\overline{\delta}_{Sg}\right) = \sqrt{\frac{1}{n_g(n_g-1)} \sum_{j \in g} \left(\delta_{jS} - \overline{\delta}_{Sg}\right)^2},$$

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

A statistically significant difference from zero in these aggregates was evidence that a class, teacher, school, or corporation was more effective (positive $\overline{\delta}_{Tg}$) or less effective (negative $\overline{\delta}_{Tg}$) in teaching a given standard.

The statistic $\overline{\delta}_{Tg}$ was not directly reported; instead, the aggregate was reported to show if a group of students performed better, worse, or as expected on this standard. In some cases, insufficient information was available, and that was indicated, as well.

For standard-level strengths/weaknesses, the following were reported:

If $\delta_{-Sg} \geq + 1.5 * se\left(\delta_{-Sg}\right)$, then performance is above the Proficiency Standard.

If $\delta_{-Sg} \leq - 1.5 * se\left(\delta_{-Sg}\right)$, then performance is below the Proficiency Standard.

Otherwise, performance is near the Proficiency Standard.

If $se\left(\delta_{-Sg}\right) > 0.2$, data are insufficient.

1.32.3 RULES FOR ZERO AND PERFECT SCORES

In IRT maximum likelihood ability estimation methods, zero and perfect scores are assigned the ability of minus and plus infinity. For all the tests, the extreme student ability estimates are truncated to the lowest observable scores (LOT/LOSS), or the highest observable scores (HOT/HOSS). Note that LOSS = lowest observable scale score and HOSS = highest observable scale score. Estimated theta values lower than the LOT or higher than the HOT will be truncated to the LOT and HOT values and will be assigned the LOSS and HOSS associated with the LOT and HOT. Table 97 gives the LOT, LOSS, HOT, and HOSS for the ILEARN assessments.

Table 97: Theta and Scaled Score Limits for Extreme Ability Estimates

Grade	Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable Scale Score (HOSS)
ELA				
3	−5.8667	3.4667	5060	5760
4	−5.4667	4.1333	5090	5810
5	−5.2000	4.6667	5110	5850
6	−4.9333	4.9333	5130	5870
7	−4.9333	5.2000	5130	5890
8	−4.6667	5.6000	5150	5920
Mathematics				
3	−5.6000	3.0667	6080	6730
4	−5.3333	4.0000	6100	6800
5	−5.2000	4.6667	6110	6850
6	−5.2000	4.9333	6110	6870
7	−5.0667	5.6000	6120	6920
8	−5.0667	6.0000	6120	6950
Science				
4	-4.28	3.41	100	220
6	-3.83	3.01	300	420

Grade	Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable Scale Score (HOSS)
Biology	-3.55	3.14	500	620
Social Studies				
5	−3	3	8350	8650
U.S. Government	−3	3	8350	8650

1.32.4 RULES FOR SCORING AND REPORTING OF INCOMPLETE TEST ADMINISTRATIONS

Reporting for each of the subject area test administrations (ELA, mathematics, science, and social studies) is based both on an attemptedness criterion and on whether the test administration is completed.

All operational items are included in the evaluation of test records for attemptedness, or whether students attempted or completed a test. Field-test items are excluded.

ILEARN implemented the following rules for participation and attemptedness as well as when to report overall and reporting category scores.

> **Not Attempted (Attempt = N).** If a student responds to four or fewer than four (<4) items, the student did not attempt the test. Test scores for these records are not computed or reported.

> **Partial Attempt (Attempt = Y).** If a student responds to at least five (≥5) items in the test.

> **Attempted (Attempt = Y).** Tests are considered "complete" if a student responds to all the items in each operational segment.

1.32.4.1 Online Tests

For tests that are attempted but not complete, if students have responded to 32 or more items for ELA, mathematics, and social studies and 16 or more items for science but have not completed the entire test, overall scores will be calculated and reported but reporting category scores will be suppressed. If fewer than 32 items for ELA, mathematics, and social studies or fewer than 16 items for science have been responded to, no scores will be reported. Since students cannot skip items in the TDS online testing environment, any administered item will have a response and score.

1.32.4.2 Paper Tests

For ELA, mathematics, and social studies, in each segment, the last item responded to will be identified. Any item prior to this item will be considered administered and blank items will be treated as incorrect. Any item after this item will be considered not administered and treated as not answered and not used for scoring. If more than 4 but less than 32 items are considered responded to, only total subject area scores will be

reported. If 32 or more items are considered responded to, total subject area and reporting category scores will be reported. For science paper tests that are attempted but not complete, if students have responded to 16 or more items but have not completed the entire test, total subject scores will be calculated and reported but reporting category (i.e. domain) scores will be suppressed in the reporting system. If fewer than 16 items have been attempted, no scores will be reported for the student in the reporting system.

For paper tests all items are considered administered, and blank items are scored as incorrect. This differs from the online assessment where test administration can be tracked at the item level rather than the test level. For this reason, overall and reporting category scores are reported for all paper test attempts, per the Technical Advisory Committee.

1.32.5 COMPARISON OF SCORES TO PREVIOUS YEAR

As a quality assurance check for aberrant test administrations in the context of the COVID-19 pandemic, CAI conducted a study to confirm the integrity of the test administration prior to the final release of spring 2024 test scores. In this study, a weighted linear regression model was run for each assessment to identify expected levels of achievement for corporations in spring 2024, given their observed achievement levels in spring 2023. Corporations with large deviations from expected levels of achievement were identified. IDOE investigated flagged schools prior to final score release. Details of this test administration study can be found in Appendix 6-A.

7. PERFORMANCE STANDARDS

In the summer of 2019, following the close of the first testing window, Cambium Assessment, Inc. (CAI), convened a panel of Indiana educators to recommend proficiency standards on each of the Indiana Learning Evaluation Readiness Network (*ILEARN*) assessments. In February 2024, prior to the opening of the spring 2024 testing window, CAI convened a panel of Indiana educators to recommend a proficiency standard for the revised U.S. Government assessment. In the summer of 2024, following the close of the first testing window for Indiana's science assessments in grades 4, 6, and biology, CAI again convened panels of Indiana educators to recommend proficiency standards on each of the new *ILEARN* science assessments.

This chapter briefly describes the procedures used by educators to recommend standards and resulting proficiency standards. Details of the panels, procedures, and outcomes are documented in the 2019 and 2024 *ILEARN* standard-setting technical reports, which are presented in Appendix 7-A, 2019 ILEARN Standard-Setting Report, Appendix 7-B, 2024 *ILEARN* U.S. Government Standards-Confirmation Report, and Appendix 7-C, 2024 *ILEARN* Science Standard-Setting Report.

1.33 STANDARD-SETTING PROCEDURES

Student achievement on *ILEARN* is classified into four performance levels: Below Proficiency, Approaching Proficiency, At Proficiency, and Above Proficiency. Interpretation of the *ILEARN* test scores rests fundamentally on how test scores relate to proficiency standards that define the extent to which students have achieved the expectations defined in the Indiana Academic Standards. The cut score establishing the Proficient level of performance is the most critical because it indicates that students are meeting grade-level expectations for achievement of the Indiana Academic Standards, that they are prepared to benefit from instruction at the next grade level, and that they are on track to pursue post-secondary education or enter the workforce. Procedures used to adopt proficiency standards for the *ILEARN* assessments are therefore central to the validity of test score interpretations.

1.33.1 ILEARN PROCEDURES IN 2019

Following the first operational administration of the *ILEARN* assessments in spring 2019, a standard-setting workshop was conducted to recommend to the State Board of Education (SBOE) a set of proficiency standards for reporting student achievement of the Indiana Academic Standards. The workshop consisted of a series of standardized and rigorous procedures that the Indiana educators serving as standard-setting panelists followed to recommend proficiency standards. The workshops employed the Bookmark procedure, a widely used method where standard-setting panelists used their expert knowledge of the Indiana Academic Standards and student achievement to map the PLDs adopted by the SBOE to an ordered-item booklet (OIB) based on the first

operational test form administered. The IDOE previously used the Bookmark method to recommend performance standards for the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) assessments. The Bookmark method was implemented in three rounds, providing panelists with feedback and benchmark information prior to Round 2, and panelist feedback, benchmark, and impact data prior to Round 3.

Following discussion of panelist feedback, panelists were presented with benchmark data, performance standards comparable to other important assessment systems, including national and international benchmarks such as NAEP and Smarter. To facilitate comparisons of Indiana performance standards with other national and international benchmarks, panelists were provided with the locations of performance standards from these other assessment systems in their OIBs. In particular, performance standard locations for the following assessments were provided as part of panelists' OIB review:

- Smarter ELA and mathematics performance standards in grades 3–8; social studies grade 5 used the performance standard cut from ELA grade 5
- NAEP performance standards in ELA and mathematics in grades 3–8 and science in grades 4, 6, and biology

When panelists can use benchmark information to locate proficiency standards that converge across assessment systems, the validity of test score interpretations is bolstered.

Panelists were also provided with feedback about the vertical articulation of their recommended proficiency standards so that they could view how the locations of their recommended cut scores for each grade-level assessment related to the cut-score recommendations at the other grade levels. This approach allowed panelists to view their cut-score recommendations as a coherent system of proficiency standards, and further reinforced the interpretation of test scores as indicating not only achievement of current grade-level standards, but also preparedness to benefit from instruction in the subsequent grade level.

1.33.1.1ILEARN Performance-Level Descriptors

Performance-Level Descriptors (PLDs) define the content-area knowledge and skills that students at each performance level are expected to demonstrate. The standard-setting panelists based their judgments about the location of the performance standards on the PLDs, as well as the Indiana Academic Standards.

Indiana's policy group is made up of a member from the SBOE, a member from higher education, administrators at the high school and grade 3–8 levels, special education administrators and leaders, and the IDOE leadership. The policy group created the Policy PLDs in May 2018. The Range PLDs were drafted by educators in a meeting held June 18–21, 2018. Policy PLDs define, at a broad policy level, what it means to be proficient across the performance levels. Policy PLDs must convey an appropriate

sense of rigor, clearly setting Indiana's expectations for a progression toward college and career readiness. Prior to the Range PLD meeting in June 2018, AIR and IDOE drafted Policy PLDs for educator review. The Policy PLDs were informed by Department leadership for educators to consider in light of the new assessment. During the first part of May 2018, IDOE sent a survey to educators to inform the labels for performance levels. On May 15, 2018, IDOE convened a stakeholder panel to make recommendations for ILEARN Policy PLDs.

IDOE provided panelists with background, the purpose and role of PLDs within the ILEARN assessment system. IDOE shared the educator survey information with the panelists and asked for their input on proficiency level names. Panelists agreed with the educators' top choice for the following proficiency level names:

- LEVEL 1: Below Proficiency

Indiana students below proficiency have not met current grade level standards. Students may require significant support to develop the knowledge, application, and analytical skills needed to be on track for college and career readiness.

- LEVEL 2: Approaching Proficiency

Indiana students approaching proficiency have nearly met current grade level standards by demonstrating some basic knowledge, application, and limited analytical skills. Students may require support to be on track for college and career readiness.

- LEVEL 3: At Proficiency

Indiana students at proficiency have met current grade level standards by demonstrating

essential knowledge, application, and analytical skills to be on track for college and career

readiness.

- LEVEL 4: Above Proficiency

Indiana students above proficiency have mastered current grade level standards by

demonstrating more complex knowledge, application, and analytical skills to be on track for college and career readiness.

Panelists used the PLDs to develop a representation of students who are "just barely" described by each of the PLDs. During this training task, panelists learned that while PLDs are written to characterize typical members of each performance level, their bookmark placements would be directed toward characterizing and identifying the most minimally qualified members of each performance level. Characterizing a student as "just barely" meeting the performance standard is not an intuitive judgment, and panelists worked to identify the minimum characteristics of student achievement for entry into each performance level. Each panel produced a "just barely" PLD to help

guide their discussions and bookmark placements. To develop a common understanding among panelists, each panel was asked to

- review and parse PLDs;
- discuss characteristics of students classified near thresholds of performance standards;
- identify the characteristics that distinguish students "just above" the performance standard from those "just below";
- determine what evidence was necessary to conclude that a student possessed the minimum knowledge and skills needed to meet the performance standard; and
- summarize knowledge and skills of students who "just barely" meet each performance standard, or are "just barely" described by each PLD.

These discussions yielded common descriptions of students "just barely" characterized by each PLD within each room.

1.33.2 ILEARN U.S. Government Procedures in 2024

Due to changes in state policy, IDOE updated the ILEARN U.S. Government ECA assessment by spring 2024 to measure new streamlined academic standards. Prior to the spring 2024 operational administration, a standards-confirmation workshop was conducted to verify to the State Board of Education (SBOE) that the established At Proficiency standard for reporting student achievement of the Indiana Academic Standards was still valid. The workshop consisted of a series of standardized and rigorous procedures that the Indiana educators serving as standards-confirmation panelists followed to confirm the established U.S. Government At Proficiency standard. The workshop employed the Bookmark procedure, a widely used method where standard-setting panelists used their expert knowledge of the Indiana Academic Standards and student achievement to map the PLDs adopted by the SBOE to an ordered-item booklet (OIB) based on the spring 2024 operational test form. The IDOE previously used the Bookmark method to recommend performance standards for the *ILEARN* assessments in 2019. The Bookmark method was implemented in two rounds, providing panelists with panelist feedback prior to Round 2.

1.33.2.1 ILEARN U.S. Government Performance-Level Descriptors

Determining the nature of the categories in which students are classified is a prerequisite to standard setting. These categories, or performance levels, are associated with PLDs that define the content-area knowledge, skills, and processes that students at each performance level can demonstrate. Indiana uses two performance levels to describe student performance for U.S. Government:

- LEVEL 1: Below Proficiency
- LEVEL 2: At Proficiency

The panel reviewed a set of just barely PLDs produced in the February 2019 workshop to help guide their discussions and evaluation of the existing At Proficiency performance standard location. These PLDs were sent to the educators one week in advance of the meeting to provide them a chance to become more familiar with the information in advance of the meeting.

1.33.3 ILEARN Science Procedures in 2024

Following the first operational administration of the new *ILEARN* science assessments in spring 2024, a standard-setting workshop was conducted to recommend to the State Board of Education (SBOE) a set of proficiency standards for reporting student achievement of the Indiana Academic Standards. The workshop consisted of a series of standardized and rigorous procedures that the Indiana educators serving as standard-setting panelists followed to recommend proficiency standards.

A new method for standard setting is necessary for tests based on the Next Generation Science Standards (NGSS) due to the structure of the content standards and, subsequently, the structure of test items assessing the standard. The workshops employed the test-centered Assertion Mapping Procedure (AMP), an adaptation of the widely used Item-Descriptor (ID) Matching method where standard-setting panelists used their expert knowledge of the Indiana Academic Standards and student achievement to map the PLDs adopted by the SBOE to an ordered set of score assertions derived from student interactions within a representative set of item clusters. These scoring assertions are not test items but rather inferences that are (or are not) supported by students' responses in one or more interactions within an item cluster. Because item clusters represent multiple, interdependent interactions through which students engage in scientific phenomena, scoring assertions cannot be meaningfully evaluated independently of the cluster from which they are derived. Thus, panelist review ordered scoring assertions for each cluster separately rather than for the overall test.

Panelists were also provided with contextual information to help inform their primarily content-driven cut-score recommendations. Panelists were provided with information about the approximate percentage of students scoring in each performance level on the 2015 NAEP assessments, where grade 6 and biology were interpolated and extrapolated from grades 4 and 8 NAEP. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists can use benchmark information to locate proficiency standards that converge across assessments systems, the validity of test score interpretations are bolstered.

1.33.3.1 ILEARN Science Performance-Level Descriptors

With the adoption of new standards in science and the development of new statewide assessments to assess achievement in those standards, IDOE adopted a similar

system of proficiency standards to determine whether students had met the learning goals defined by the new standards in science.

Determining the nature of the categories in which students are classified is a prerequisite to standard setting. These categories, or performance levels, are associated with PLDs that define the content-area knowledge, skills, and processes that students at each performance level can demonstrate. Indiana uses four performance levels to describe student performance:

- LEVEL 1: Below Proficiency
- LEVEL 2: Approaching Proficiency
- LEVEL 3: At Proficiency
- LEVEL 4: Above Proficiency

PLDs were reviewed and revised in a separate workshop before the standard-setting workshop.

1.34 RECOMMENDED PROFICIENCY STANDARDS

1.34.1 ILEARN STANDARDS IN 2019

Panelists were tasked with recommending three proficiency standards (Approaching Proficient, Proficient, and Highly Proficient) that resulted in four performance levels (Below Proficiency, Approaching Proficiency, At Proficiency, and Above Proficiency). As panelists discussed the reasons for their bookmark placements in the context of feedback from other panelists and impact data, variability often decreased across rounds. In general, there was considerable consistency in the placement of performance standards across rounds.

The final recommended performance standards for each assessment, grade, and performance standard are presented in Table 98 along with the projected impact each performance standard would have on Indiana public school students tested in 2019. The final recommended OIB page numbers are the median bookmarks of each panel following Round 3 bookmark placement, and subsequent moderation.

Following the standard-setting workshop, panelist recommendations were submitted to IDOE; IDOE formally adopted the standards in July 2019.

Table 98: Final Recommended Performance Standards

Grade	Performance Level	OIB Page	RP67	Estimated Percentage of Students At or Above Performance Standard
ELA 3	Approaching Proficiency	9	−1.12	69%
	At Proficiency	25	−0.54	46%
	Above Proficiency	43	0.20	18%
ELA 4	Approaching Proficiency	8	−0.75	69%

Grade	Performance Level	OIB Page	RP67	Estimated Percentage of Students At or Above Performance Standard
	At Proficiency	24	−0.10	45%
	Above Proficiency	45	0.63	19%
ELA 5	Approaching Proficiency	9	−0.37	71%
	At Proficiency	26	0.32	47%
	Above Proficiency	44	1.26	15%
ELA 6	Approaching Proficiency	7	−0.11	73%
	At Proficiency	21	0.59	47%
	Above Proficiency	41	1.38	17%
ELA 7	Approaching Proficiency	5	0.09	75%
	At Proficiency	24	0.90	49%
	Above Proficiency	43	1.72	20%
ELA 8	Approaching Proficiency	6	0.15	79%
	At Proficiency	21	1.03	50%
	Above Proficiency	44	1.85	21%
Mathematics 3	Approaching Proficiency	7	−1.57	76%
	At Proficiency	17	−0.99	58%
	Above Proficiency	47	−0.16	25%
Mathematics 4	Approaching Proficiency	9	−0.95	74%
	At Proficiency	22	−0.35	53%
	Above Proficiency	49	0.54	21%
Mathematics 5	Approaching Proficiency	7	−0.62	72%
	At Proficiency	23	0.14	47%
	Above Proficiency	47	0.88	22%
Mathematics 6	Approaching Proficiency	8	−0.16	70%
	At Proficiency	23	0.59	46%
	Above Proficiency	47	1.39	20%
Mathematics 7	Approaching Proficiency	10	−0.10	68%
	At Proficiency	28	0.83	41%
	Above Proficiency	43	1.67	18%
Mathematics 8	Approaching Proficiency	12	0.13	65%
	At Proficiency	29	1.20	37%
	Above Proficiency	48	2.01	18%
Science 4	Approaching Proficiency	12	−0.36	65%
	At Proficiency	24	0.12	46%
	Above Proficiency	40	0.69	24%
Science 6	Approaching Proficiency	12	−0.68	73%
	At Proficiency	26	0.08	47%
	Above Proficiency	46	0.89	19%
Biology	Approaching Proficiency	12	−0.43	63%

Grade	Performance Level	OIB Page	RP67	Estimated Percentage of Students At or Above Performance Standard
	At Proficiency	28	0.18	39%
	Above Proficiency	47	0.93	17%
	Approaching Proficiency	8	−0.46	63%
Social Studies 5	At Proficiency	18	0.04	45%
	Above Proficiency	42	0.87	21%

Table 99 shows the estimated percentage of student classified at each performance level based on final panelist-recommended standards for the overall student population across grade levels and courses.

Table 99: Percentage of Students at Each Performance Level Based on Final Recommended Performance Standards

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
ELA 3	31	23	28	18
ELA 4	31	24	26	19
ELA 5	29	24	31	15
ELA 6	27	26	29	17
ELA 7	25	26	29	20
ELA 8	21	29	29	21
Mathematics 3	24	19	32	25
Mathematics 4	26	21	33	21
Mathematics 5	28	25	25	22
Mathematics 6	30	24	26	20
Mathematics 7	32	27	23	18
Mathematics 8	35	28	19	18
Science 4	35	19	22	24
Science 6	27	25	28	19
Biology	37	24	22	17
Social Studies 5	37	18	24	21

Table 100 shows the estimated percentage of students meeting the *ILEARN* proficient standard for each assessment in spring 2019. It also shows the national percentages of students that meet the NAEP and Smarter proficient standards. Since NAEP is delivered in grades 4 and 8 only, the percentages in other grades were interpolated or extrapolated so estimated percentages were available in all grades. As Table 5 indicates, the performance standards recommended for *ILEARN* assessments are consistent with relevant NAEP and Smarter proficient benchmarks. Moreover, because the performance standards were vertically articulated in ELA and mathematics, the proficiency rates across grade levels are generally consistent.

Table 100: Estimated Percentage of Students Meeting ILEARN and Benchmark Proficient Standards

Grade	ILEARN At Proficiency	NAEP Proficient	Smarter Proficient
ELA 3	46	41	45
ELA 4	45	41	47
ELA 5	47	41	50
ELA 6	47	41	48
ELA 7	49	41	50
ELA 8	50	41	50
Mathematics 3	58	51	47
Mathematics 4	53	48	43
Mathematics 5	47	46	36
Mathematics 6	46	43	38
Mathematics 7	41	41	38
Mathematics 8	37	38	37
Science 4	46	42	--
Science 6	47	39	--
Biology	39	35	--
Social Studies 5	45	--	50

The IDOE reported ELA and mathematics student performance on the vertically linked scale established by Smarter. The IRT vertical scale was formed by linking across grades using common items in adjacent grades. Grade 6 was used as the baseline, and each grade was successively linked onto the scale. Each science and social studies assessment was reported on a separate within-test scale. Applying the ILEARN scale score transformations to the performance standards recommended by the workshop panels results in the system of scale score ranges for each of the *ILEARN* performance-level classifications identified in Table 101.

Table 101: ILEARN Scale Score Ranges Based on Final Performance Standards

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
ELA 3	5060–5415	5416–5459	5460–5514	5515–5760
ELA 4	5090–5443	5444–5492	5493–5546	5547–5810
ELA 5	5110–5471	5472–5523	5524–5594	5595–5850
ELA 6	5130–5491	5492–5543	5544–5603	5604–5870
ELA 7	5130–5506	5507–5567	5568–5628	5629–5890
ELA 8	5150–5510	5511–5576	5577–5637	5638–5920
Mathematics 3	6080–6381	6382–6424	6425–6487	6488–6730
Mathematics 4	6100–6428	6429–6473	6474–6540	6541–6800
Mathematics 5	6110–6452	6453–6509	6510–6565	6566–6850

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
Mathematics 6	6110–6487	6488–6544	6545–6604	6605–6870
Mathematics 7	6120–6492	6493–6561	6562–6624	6625–6920
Mathematics 8	6120–6508	6509–6589	6590–6650	6651–6950
Science 4*	7350–7481	7482–7505	7506–7534	7535–7650
Science 6*	7350–7465	7466–7503	7504–7544	7545–7650
Biology*	7350–7477	7478–7508	7509–7546	7547–7650
Social Studies 5	8350–8476	8477–8501	8502–8542	8543–8650
U.S. Government	8350–8496	-	8497–8650	-

*Science SP24 is a new test, and the standards and new cuts scores for that test are included in section 7.2.3.

1.34.2 ILEARN U.S. Government Standards in 2024

Panelists were tasked with confirming one proficiency standards (At Proficiency) that resulted in two performance levels (Below Proficiency and At Proficiency). Table 102 presents the proficiency standard associated with the percentage of students classified as meeting or exceeding the standard using spring 2019 administration data. Following the standards-confirmation workshop, panelist recommendations were submitted to IDOE; the Board formally adopted the standard in February 2024.

Table 102: Final Recommended Performance Standards for *ILEARN* U.S. Government using Spring 2019 Data

Grade	At Proficiency	
	Scale Score	% At or Above
U.S. Government	8497	20

Table 103 shows the percentage of students classified at each performance level in spring 2024, the initial year of the new U.S. Government administration, based on final panelist-recommended standards for the student population.

Table 103: Percentage of Spring 2024 Students at Each Performance Level Based on Final Recommended Proficiency Standards

Grade	Below Proficiency	At Proficiency
U.S. Government	84	16

1.34.3 ILEARN Science Standards in 2024

Panelists were tasked with recommending three proficiency standards (Approaching Proficiency, At Proficiency, and Above Proficiency) that resulted in four performance levels (Below Proficiency, Approaching Proficiency, At Proficiency, and Above

Proficiency). Table 104 presents the proficiency standard associated with the percentage of students classified as meeting or exceeding each standard. Following the standard-setting workshop, panelist recommendations were submitted to IDOE; the Board formally adopted the standards in June 2023.

Table 104: Final Recommended Performance Standards for *ILEARN* Science

Grade	Approaching Proficiency		At Proficiency		Above Proficiency	
	Scale Score	% At or Above	Scale Score	% At or Above	Scale Score	% At or Above
4	145	81	163	44	180	12
6	347	79	363	42	372	23
Biology	546	81	561	44	573	21

Table 105 shows the percentage of students classified at each performance level in 2024, the initial year of the new science administration, based on final panelist-recommended standards for the student population.

Table 105: Percentage of Students at Each Performance Level Based on Final Recommended Proficiency Standards

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
4	19	37	32	12
6	21	37	19	23
Biology	19	37	23	21

Table 106 shows the percentage of students meeting the *ILEARN* proficient standard for each assessment based on the spring 2024 operational test administration, and the approximate percentage of Indiana students meeting the NAEP science proficient standards. As the table indicates, the proficiency standards recommended are quite consistent with NAEP proficient benchmarks.

Table 106: Percentage of Students Meeting *ILEARN* and Benchmark Proficient Standards

Grade	ILEARN At or Above Proficiency	Indiana NAEP Proficient or Advanced
4	44	42
6	42	39
Biology	44	35

8. REPORTING AND INTERPRETING ILEARN SCORES

The online centralized Reporting System generates a set of online score reports that includes information describing student performance for students, parents, educators, and other stakeholders. The online score reports are generally produced immediately after students complete tests with machine scored items and by 12 business days for tests that contain human handscored items. Because the performance score report is updated each time a student completes a test, authorized users (e.g., school principals, teachers) can access available information on students' performance scores quickly and use it to immediately mediate to improve student learning. In addition to individual student's score reports, the Reporting System also produces aggregate score reports by districts, schools, and teacher rosters. The timely accessibility of aggregate score reports can help users monitor students' performance in each subject by grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year.

This section contains a description of the types of scores reported in the Reporting System and a description of the ways to interpret and use these scores in detail.

1.35 CONFIDENTIALITY OF STUDENT DATA

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

Once authorized users log in to the Reporting System, the online score reports are presented hierarchically. The system starts by presenting summaries on student performance on all assessments by subject and grade at a selected aggregate level. To view student performance for a specific aggregate unit, users can select the specific aggregate unit from a drop-down list of aggregate units (e.g., schools within a district, or rosters within a school). For more detailed student assessment results for a school, a teacher, or a roster, users can select the subject and grade on the online score reports.

Generally, the Reporting System provides two categories of online score reports: (1) aggregate score reports across the state, district, and school, and (2) student individual

score reports. Table 107 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 Reporting System User Guide, located in a help button on the Reporting System and posted in the Resources section of the ILEARN assessment portal.

Table 107: Types of Online Score Reports by Aggregation Level

Type of Report	Description
District School Teacher Roster	Number of students (for overall students and by subgroup)
	Average scale score (for overall students and by subgroup)
	Percentage and count of students at each performance level on the overall test (for overall students and by subgroup)
	Percentage and count of students at each performance category on the reporting category level (for overall students and by subgroup)
	Standard performance relative to proficiency (for overall students and by subgroup)
	Standard performance relative to test as a whole (for overall students and by subgroup)
	On-demand student roster report
Student	Overall scale score and standard error of measurement
	Overall performance level
	Average scale scores for student's school and district
	Performance category at the reporting category level

1.36 REPORTING SYSTEM FOR STUDENTS AND EDUCATORS

1.36.1 DASHBOARD

When users log on to the Reporting System, the dashboard page shows overall test results for all tests that the students have taken grouped by test family (e.g., Summative ELA). The dashboard summarizes students' performance by test family for ELA, mathematics, and science across all grades, including (1) the grades of the students who have tested, (2) the number of tests taken, (3) the test date last taken, and (4) the percentage and counts of students at each achievement level. District personnel see district summaries, school personnel see school summaries, and teachers see summaries of their students, assigned to them via a roster.

Figure 16 presents an example dashboard page at the district level.

Figure 16: Dashboard: District Level
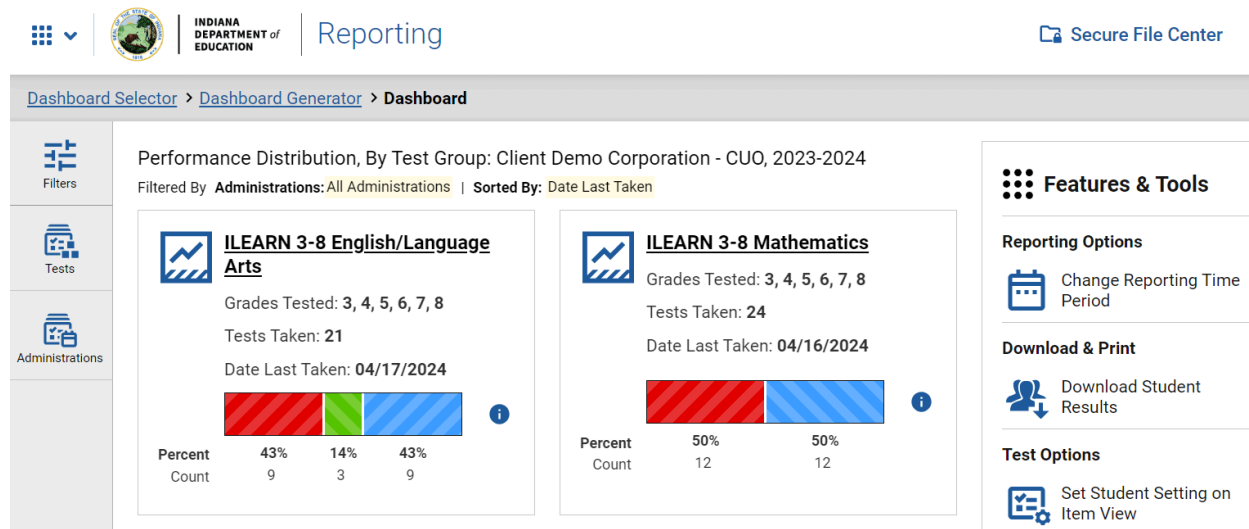
Once the user clicks on the test family that he or she wants to explore further, the system will take the user to the detailed dashboard, where the results will be displayed by test (e.g., Grade 3 ELA). The detailed dashboard page will appear by test in each grade. The detailed dashboard summarizes students' performance by test in each grade, including (1) student count, (2) average scale score and standard error of the average scale score, (3) the percentage and counts of students at each achievement level, and (4) test date last taken.

Figure 17 presents an example dashboard page at the district level.

Figure 17: Detailed Dashboard: District Level
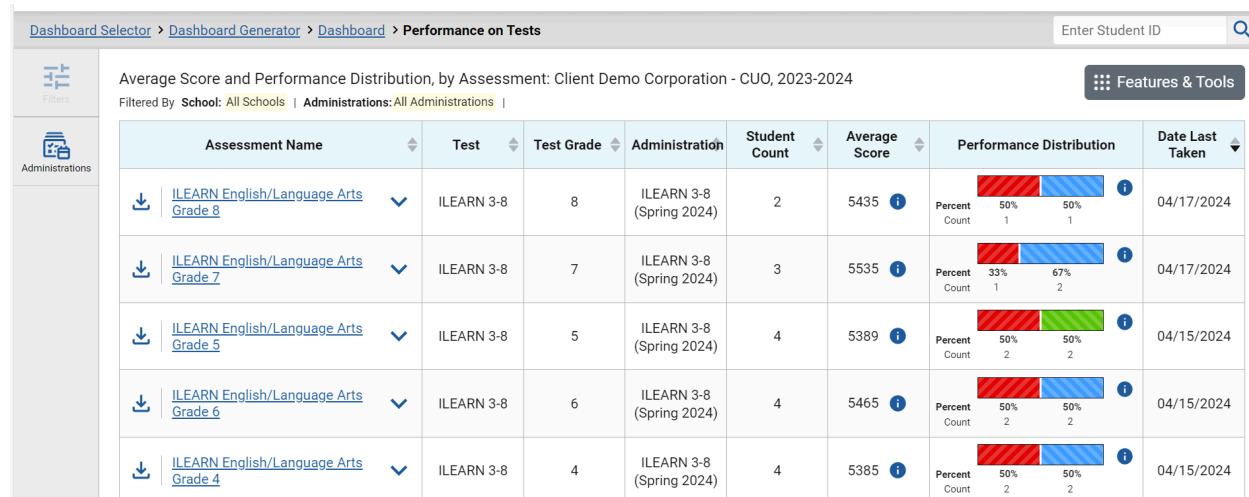
1.36.2 AGGREGATE-LEVEL SUBJECT DETAIL PAGE

More detailed summaries of student performance in each grade in a subject area for a selected aggregate level are presented when users select an assessment on the dashboard page. On each aggregate report, the summary report presents the summary results for the selected aggregate unit and the summary results for all aggregate units above the selected aggregate. For example, at the roster level, summaries appear for the teacher, school, and district aggregate. The roster performance can be compared with the above aggregate levels.

The subject detail page provides the aggregate summaries on a specific subject area including: (1) number of students, (2) average scale score, (3) percentage proficient, and (4) percentage of students in each performance level. The summaries are also presented for overall students and by subgroup.

Figure 18 presents an example of subject detail pages for mathematics at the district level.

Figure 18: Subject Detail Page for ELA: District View
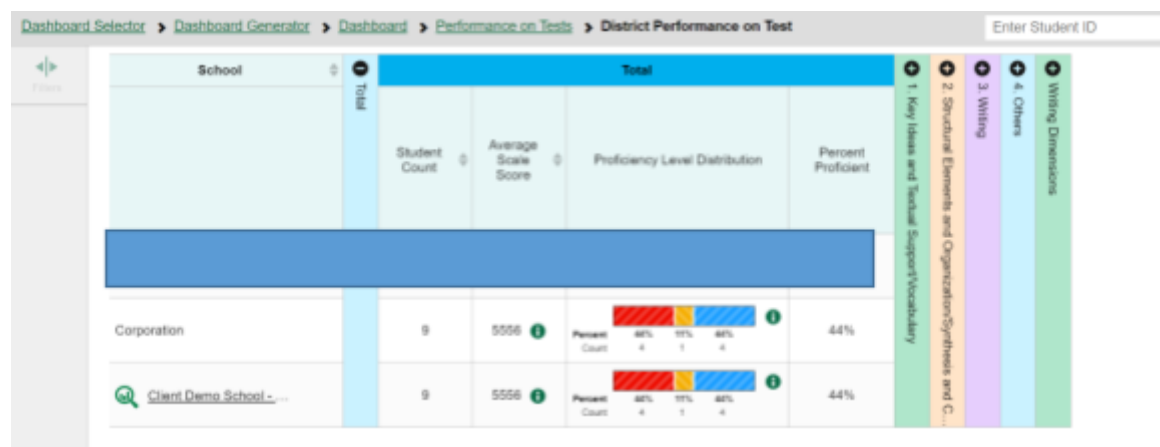
1.36.3 AGGREGATE-LEVEL REPORTING CATEGORY AND STANDARD REPORT

The Aggregate-Level Reporting Category Report provides the aggregate summaries on student performance in each reporting category for a particular grade and subject. The summaries on the Aggregate-Level Reporting Category Report include: (1) percentage of students in each achievement category for each reporting category, (2) performance relative to proficiency for each standard, and (3) performance on each standard relative to test as a whole.

For *Areas Where Performance Indicates Proficiency*, a performance indicator produces information on how a group of students in a roster, school, or district performed on the standard compared to the proficiency cuts. The performance indicator shows whether performance on this standard for this group was above, at/near, or below what is expected of students at the proficient level. *Areas of Strongest and Weakest*

Performance works in a similar manner but reports on specific areas of performance (via standards) relative to the group's overall performance instead of proficiency. It shows whether performance on this standard was above, at/near, or below what is expected of students in this group given the students' overall test performance. These indicators show strengths and weaknesses for a group of students and are provided at an aggregate level only because they are technically unreliable at the individual level.

Similar to the Aggregate-Level Subject Report, this report presents the summary results for the selected aggregate unit as well as the summary results for the aggregate units above the selected aggregate.

Figure 19 presents examples of the District Aggregate-Level Reporting Category and Standard Detail for mathematics.

Figure 19: Reporting Category and Standard Detail Page for ELA: District Level
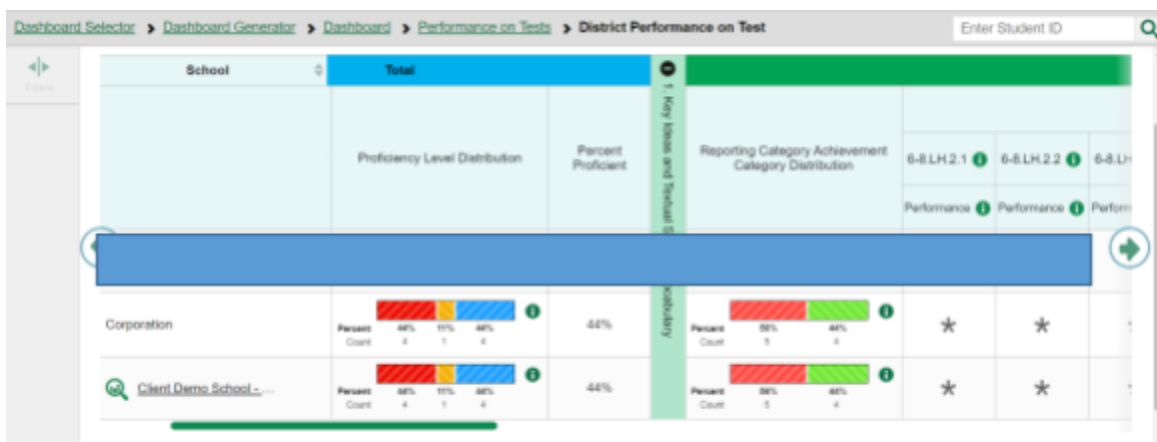
1.36.4 STUDENT PERFORMANCE ON TEST REPORT: PERFORMANCE BY ROSTER

The Student Roster Subject Report lists all students who belong to the selected aggregate level, such as a school, and reports the following measures for each student: (1) scale score, and (2) overall subject performance level.

Figure 20 contains examples of the Student Roster Subject Report for mathematics.

Figure 20: Student Performance on Test Report: Performance by Roster
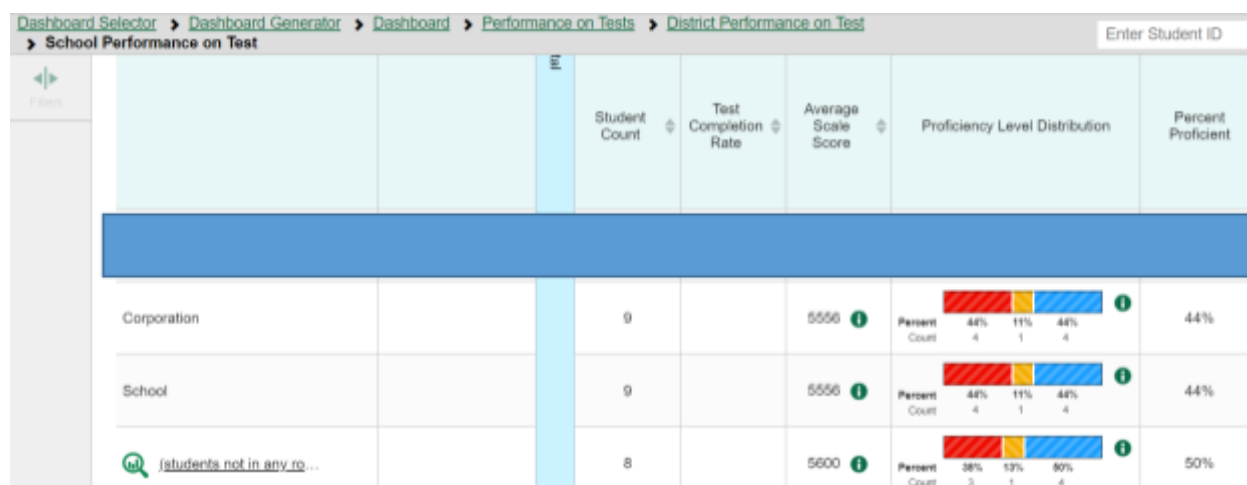
1.36.5 STUDENT PERFORMANCE ON TEST REPORT: PERFORMANCE BY ROSTER WITH EXPANDED REPORTING CATEGORY SECTION

The Student Roster Reporting Category Report records the reporting category achievement measures for individual students. Figure 21 presents an example of the Student Roster Reporting Category Report for mathematics.

Figure 21: Student Performance on Test Report: Performance by Roster with Expanded Reporting Category Section
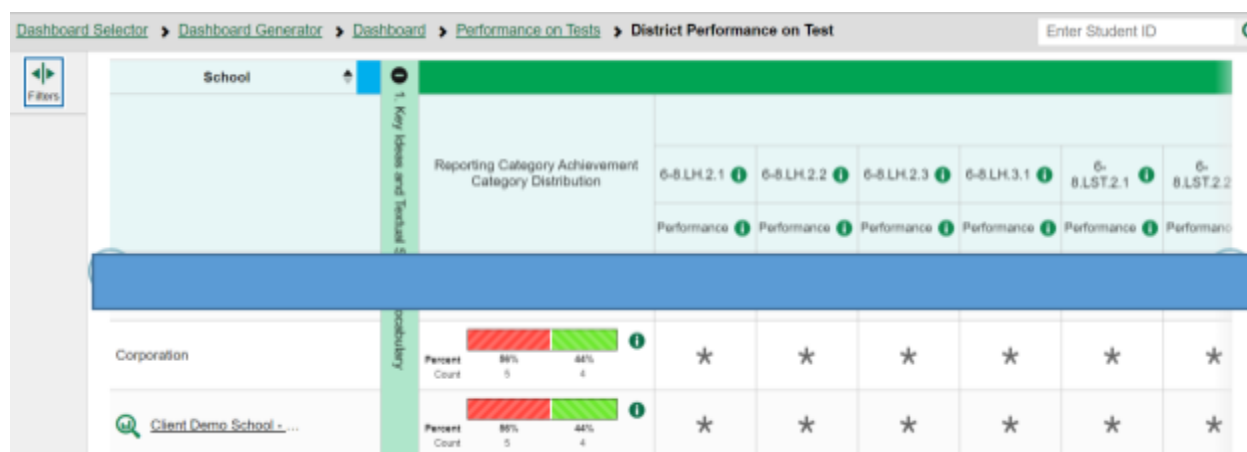
1.36.6 STUDENT INDIVIDUAL SCORE REPORT PAGE

When a student completes a test, an online score report appears in the student detail page in the Reporting System. The student detail page provides information about individual student performance on the test. It also provides (1) average scale score (2) performance level for the overall test, and (3) average scale scores for the student's state, district and school in each subject area.

On the top of the page, the student's name, scale score, and performance level are presented. On the left side section, the student's performance is described in detail using a horizontal bar chart. The student scale score is presented in the horizontal bar chart. On the right side, average scale scores for the student's state, district, and school are displayed so that the student achievement can be compared with the above aggregate levels. Student's performance on each reporting category are shown under the overall performance where the performance is shown graphically followed by a description of the performance. The following section of this technical report shows the longitudinal graph and table that shows historical performance over time for the subject. Figure 22 presents an example of the student detail pages for ELA.
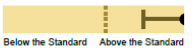
Figure 22: Student Individual Score Report for ELA
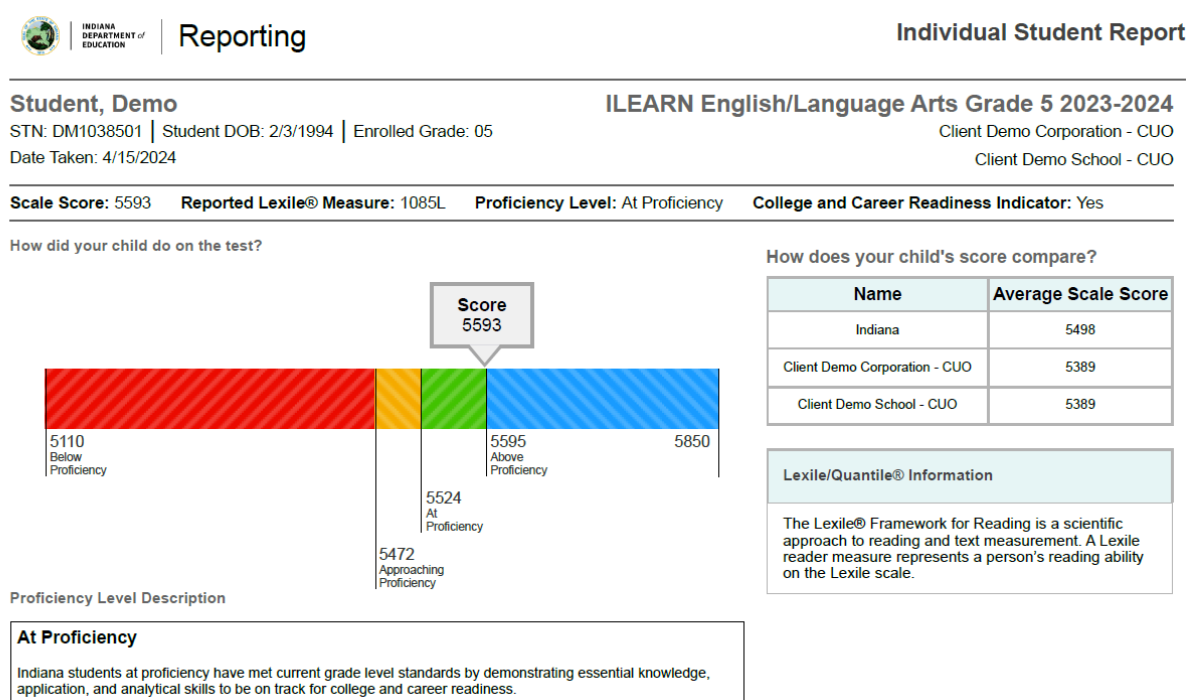
How Did Your Child Perform on Different Areas of the Test?

The table and the graph below indicate student performance on individual reporting categories. The black dot indicates the student's score on each reporting category. The lines to the left and right of the dot show the range of likely scores your student would receive if he or she took the test multiple times.

⚠ Below   ▨ At/Near   ✓ Above

Category	Reporting Category Achievement Category	Performance Level	Reporting Category Achievement Category Description
1. Key Ideas and Textual Support/Vocabulary	Below the Standard — Above the Standard	✓	Your student can almost always independently interact with literary and nonfiction texts. He or she quotes accurately to draw complex inferences, explain main ideas, describe how characters/events impact plot, and determine the meanings of complex words and phrases.
2. Structural Elements and Organization/Connection of Ideas/Media Literacy	Below the Standard — Above the Standard	▨	Your student can often independently explain reasoning used to support claims in different media, describe various viewpoints and how they influence information, describe a text's overall structure, and compare stories in the same genre on their approaches to similar themes.
3. Writing	Below the Standard — Above the Standard	▨	Your student can often independently organize and develop writing for persuasive, informative, and narrative purposes; introduce a topic; and use facts and examples to support ideas. He or she often uses appropriate word choice, sentence structure, and punctuation.

How Did Your Child Perform on the Essay?

Essay	Raw Score	Conventions	Evidence/Elaboration	Organization/Purpose
Narrative	Not available	The response was not able to be scored for the following reason: In a language other than English.	The response was not able to be scored for the following reason: In a language other than English.	The response was not able to be scored for the following reason: In a language other than English.

1.37 INTERPRETATION OF REPORTED SCORES

A student's performance on a test is reported as a scale score and a performance level for the overall test. There is also a separate performance level for each reporting category. Students' scores and performance levels are summarized at the aggregate levels. This section describes how to interpret these scores.

1.37.1 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 the student's knowledge and skills measured based on their performance on the assessment. 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. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels.

1.37.2 PERFORMANCE LEVELS

Performance levels are proficiency categories on a test that students are categorized into based on their scale scores. For summative assessments, scale scores are mapped into four performance levels (i.e., Below Proficiency, Approaching Proficiency, At Proficiency, and Above Proficiency.) using three performance standards (i.e., cut scores). Performance-Level Descriptors (PLDs) are a description of content area knowledge and skills that test takers at each performance level are expected to possess.

1.37.3 AGGREGATED SCORE

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

1.37.4 RELATIVE STRENGTHS AND WEAKNESSES

For standard performance, relative strengths and weaknesses at each standard are reported for aggregate levels only (e.g., classroom, school, district). Because an individual student responds to too few items within a given standard to generate reliable, technical data, the standard performance is produced by aggregating all items within a standard across students at an aggregate level. Standard reports include data on both Performance Relative to the Test as a Whole and Performance Relative to Proficiency for each standard. The difference between these two data reports is similar to the difference between norm-referenced data and standards-based data.

The Performance Relative to the Test as a Whole data for a standard show how a group of students performed in each standard relative to their performance on the total test. This is a norm-referenced report, with group performance in each standard being compared to the same group's overall test performance. Unlike performance levels provided for the total test, these data are not an indication of students' achievement in the standard.

The Performance Relative to Proficiency data for a standard show how a group of students performed in each standard relative to the expected performance for proficiency. For summative tests, this is the expected level of performance necessary to achieve Level 3 or Proficient performance. This is a standards-based report with the group performance in each standard being compared to the performance standard for that standard. Similar to the performance levels provided for the total test, these data indicate students' achievement in the standard with respect to the standards.

The Performance Relative to the Test as a Whole data for each standard are computed within a group; therefore, it is not appropriate to compare these data between groups. However, because the Performance Relative to Proficiency data for each standard are comparable to the standards-based expectations, performance across groups can be compared.

1.38 APPROPRIATE USES FOR SCORES AND REPORTS

Assessment results can be used to provide information on individual students' achievement on the test. Overall, assessment results show what students know and are able to do in certain subject areas. Further, they give information on whether students are on track to demonstrate the knowledge and skills necessary for college and their careers.

Assessment results on student achievement on the test can be used to help teachers or schools make decisions on how to support students' learning. Aggregate score reports for teacher and school levels provide information regarding the strengths and weaknesses of their students and can be utilized to improve teaching and student learning. By narrowing the student performance result by subgroup in the Reporting System, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from disadvantaged subgroups. For example, teachers can review student assessment results by IEP code and observe that students in the subgroup category that are struggling with ELA. Teachers can then provide additional instructions for these students to enhance their achievement in a specific subject.

In addition, assessment results can be used to compare students' performance among different students and among different groups. Teachers can evaluate how their students perform compared with other students in schools and districts overall.

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

9. QUALITY ASSURANCE PROCEDURES

Quality assurance (QA) procedures are enforced throughout all stages of ILEARN test development, administration, and scoring and reporting. This chapter describes QA procedures associated with the following activities:

- Test construction
- Test production
- Data preparation
- Equating and scaling
- Scoring and reporting

Because QA procedures pervade all aspects of test development, we note that discussion of QA procedures is not limited to this chapter but is also included in chapters describing all phases of test development and implementation.

1.39 QUALITY ASSURANCE IN ITEM DEVELOPMENT AND TEST CONSTRUCTION

Chapter 4 details the item development and test configuration processes. Each test administration is generated by the adaptive algorithm to exactly match the detailed test blueprint while targeting test information to student ability. The blueprint describes the content to be covered, the Depth of Knowledge (DOK) with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test.

The adaptive test configuration process is managed through Cambium Assessment, Inc.'s (CAI) test simulator. Upon completion of a simulation, the test simulator immediately generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, the test simulator produces a statistical summary of form characteristics in order to ensure consistency of test characteristics across simulated test forms.

Prior to its implementation in the operational test administration, the CAI scoring engine and the accuracy of data files are checked using a simulated student response data file. The simulated data are used to check whether the student responses entered in the Test Delivery System (TDS) were captured accurately, and the scoring specifications were applied accurately. The simulated data file is scored independently by two programmers, following the scoring rules.

In addition to checking the scoring accuracy, the test configuration file is checked thoroughly. For the operational administration, a test configuration file is the key file that contains all specifications for the item selection algorithm, and eventually for the scoring algorithm, such as the test blueprint specification, slopes and intercepts for theta-to-scale score transformation, and the item information (e.g., cut scores, answer keys, item attributes, item parameters, passage information, etc.). The accuracy of the

information in the configuration file is checked and confirmed numerous times independently by multiple staff members before the testing window opens.

1.40 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION

1.40.1 PRODUCTION OF CONTENT

While the online workflow requires some additional steps, it removes a substantial amount of work from the time-critical path, reducing the likelihood of errors. Like a test book, an online system can deliver a sequence of items; however, the online system makes the layout of that sequence algorithmic. The appearance of the item screen can be known with certainty before the final test is configured.

The production of computer-based tests includes four key steps:

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.
2. The complete test configuration is approved, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
3. Tests are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the state for testing purposes.
4. The final system is deployed to a staging environment accessible to IDOE for user acceptance testing (UAT) and final review.

1.40.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT

The Item Tracking System (ITS) integrates directly with the TDS display module and displays each item exactly as it will appear to the student. This process is called Web Preview and is tied to specific item review levels. Upon approval at those levels, the system locks content as it will be displayed to the student, transforming the item representation to the exact representation that will be rendered to the student. No change to the display content can occur without a subsequent Web Preview. This process freezes the display code that will present the item to the student.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as the student will view it. Layout changes can be made after this process in two ways:

1. Content can be revised and re-approved for web display.
2. Online style sheets can be changed to revise the layout of all items on the test.

Both processes are subject to strict change-control protocols to ensure that accidental changes are not introduced. Below, we discuss automated quality control processes during content publication that raise warnings if item content has changed after the most recent web-approved content was generated. The web approval process offers the

benefit of allowing final layout review much earlier in the process, reducing the work that must be performed during the very busy period just before tests go live.

1.40.3 PLATFORM REVIEW

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 significantly different platforms.

Platform review is conducted by a team. The team leader projects the item in its web-approved ITS format, and team members, each behind a different platform, look at the same item to gauge whether it renders as expected.

1.40.4 USER ACCEPTANCE TESTING AND FINAL REVIEW

Each release of every one of our systems goes through a complete testing cycle, including regression testing. With each release, and every time we publish a test, the system goes through UAT. During UAT, we provide our client with login information to an identical (though smaller scale) testing environment to which the system has been deployed. We provide recommended testing scenarios and constant support during the UAT period. Identified issues will be resolved before the opening of the test administration or noted for future review and resolution if a current resolution is not feasible within the timeline. IDOE signs off on the administration go-live date at the conclusion of UAT activities.

Deployments to the production environment follow specific, approved deployment plans. Teams working together execute the deployment plan. Each step in the deployment plan is executed by one team member and verified by a second. Each deployment undergoes shakeout testing following the deployment. This careful adherence to deployment procedures ensures that the operational system is identical to the system evaluated on the testing and staging servers. Upon completion of each deployment project, management approves the deployment log.

During the year, some changes may be required to the production system. Outside of routine maintenance, no change is made to the production system without approval of the Production Control Board (PCB). The PCB includes the director of CAI's Assessment Program or the chief operating officer, the director of our Computer and Statistical Sciences Center, and the project director. Any request for a change to the production system requires the signature of the system's lead engineer. The PCB reviews risks, test plans, and test results. In addition, if any proposed change will affect client functionality or pose risk to operation of a client system, the PCB ensures that the client is informed and in agreement with the decision.

The PCB approves a maintenance plan that includes every scheduled change to the system.

Deviations from the maintenance plan must be approved by the PCB, including server or driver patches that differ from those approved in the maintenance plan.

Every bug fix, enhancement, data correction, or new feature must be presented with the results of a quality assurance plan and approved by the PCB.

An emergency procedure is in place that allows rapid response in the event of a time-critical change needed to avert compromise of the system. Under those circumstances, any member of the PCB can authorize the senior engineer to make a change, with the PCB reviewing the change retroactively.

Typically, deployments happen during a maintenance window, and deployments are scheduled at a time that can accommodate full regression testing on the production machines. Any changes to the database or procedures that in any way might affect performance are typically subject to a load test at this time.

1.40.4.1 Cutover and Parallel Processing

CAI maintains multiple environments to ensure smooth cutover and parallel processing. With a centralized hosting site in Washington, DC, multiple development environments and a test environment can be maintained. At Rackspace, we maintain a staging environment and the production environment.

The production environment runs independently of the other environments and is changed only with the approval of the PCB. When developing enhancements, they are developed and tested initially on the development and test environments in Washington, DC, before being deployed to the staging environment in Rackspace.

The staging environment is a scaled-down version of the production environment. It is in this environment that UAT takes place. Only when UAT is complete and the PCB signs off is the production environment updated. In this way, the system continues to function uninterrupted as testing takes place in parallel until a clean cutover takes place.

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to UAT. UAT of the TDS serves both a software evaluation and content approval role. The UAT period provides IDOE with an opportunity to interact with the exact test with which the students will interact.

1.40.5 FUNCTIONALITY AND CONFIGURATION

The items, both individually and as configured onto the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document quality assurance procedures for delivering the online assessments.

One area of quality unique to online delivery is the quality of the delivery system. Three activities provide for the predictable, reliable, quality performance of our system. They include:

1. Testing on the system itself to ensure function, performance, and capacity

2. Capacity planning
3. Continuous monitoring

CAI statisticians examine the delivery demands, including the number of tests to be delivered, the length of the testing window, and the historic state-specific behaviors to model the likely peak loads. Using data from the load tests, these calculations indicate the number of each type of server necessary to provide continuous, responsive service, and CAI contracts for service in excess of this amount. Once deployed, our servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts our engineers at the first signs that trouble may be ahead. 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.

In addition, latency data is captured for each assessed student—data about how long it takes to load, view, or respond to an item. All this information is logged, as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

1.41 QUALITY ASSURANCE IN DATA PREPARATION

When a student responds to test questions online, his or her response to each item is immediately captured and stored in the Database of Record (DOR) at CAI, a repository for all data relevant to a student's testing experience. Our quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at CAI.

When data are prepared for psychometric analyses, they undergo two phases: a data preparation phase and a psychometric phase. In the former phase, data are extracted from the DOR and provided to two independent SAS programmers. These two programmers are provided with the client-assigned business rules, and they independently prepare data files suitable for subsequent psychometric analysis. The data files prepared by the different programmers are formally compared for congruency. Any discrepancies identified are resolved through code review meetings with the lead programmer and the lead psychometrician.

When the two data files match exactly, they are then passed over to two independent psychometricians, who each perform classical and IRT analyses. Any discrepancies are identified and resolved. When all results match from the independent analysts, the final results are uploaded to CAI's ITS.

CAI's Test Delivery System (TDS) has a real-time quality-monitoring component built in. As students test, data flow through our Quality Monitor (QM) system. The QM conducts a series of data integrity checks, ensuring, for example, that the record for each test

contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. In addition, the QM scores the test, recalculates performance-level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

The QM also aggregates data to detect problems that become apparent only in the aggregate. For example, the QM monitors item statistics and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem. This automated process is similar to the sorts of checks performed for data review, but they are conducted (a) on operational data, and (b) in real time to allow our psychometricians to catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM System to the DOR, which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator is the tool that is used to pull data from the DOR for delivery to IDOE and their QA contractor. CAI psychometricians ensure that data in the extract files match the DOR prior to delivery to the IDOE.

1.42 Quality Assurance in Item Analyses and Equating

Prior to operational work, CAI produces simulated datasets for testing software and analysis procedures. The quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are independently replicated by two CAI psychometricians. Two psychometricians complete a dry run calibration and linking activities and compare results. The practice runs serve two functions:

1. To verify accuracy of program code and procedures
2. To evaluate the communication and work flow among participants. If necessary, the team will reconcile differences and correct production or verification programs.
3. Following the completion of these activities and the resolution of questions that arise, analysis specifications are finalized.

1.43 Quality Assurance in Scoring and Reporting

CAI implements a series of quality control steps to ensure error-free production of score reports in an online format. The quality of the information produced in the TDS is tested thoroughly before, during, and after the testing window.

1.43.1 HANDSCORING

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the QM System, where the integrated record is passed for scoring. Once the integrated scores are sent to the QM System, the records are rescored in the test-scoring system that applies the ILEARN scoring rules and assigns scores from the calibrated items, including calculating performance-level indicators, subscale scores, and other features, which then pass automatically to the Reporting System and the DOR. The scoring system is tested extensively prior to deployment, including checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

1.43.2 QUALITY ASSURANCE IN TEST SCORING

CAI verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each simulated student is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they also provide a check of the full range of item responses and test scores in fixed-form tests. Additionally, these simulations ensure that students at all performance levels are exposed to the full range of test item content as dictated by the ILEARN test blueprints. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the Reporting System, we merge item response data with the demographic information taken either from previous year assessment data, or if current year enrollment data is available by the time simulated data files are created, we can verify online reporting using current year testing information. By populating the simulated data files with real school information, it is possible to verify that special school types and special districts are being handled properly in the Reporting System.

Specifications for generating simulated data files are included in the analysis specifications document submitted to IDOE each year. Review of all simulated data is scheduled to be completed before the opening of the test administration window, so that the integrity of item administration, data capture, and item and test scoring and reporting can be verified before the system goes live.

To monitor the performance of the assessment system during the test administration window, a series of quality assurance reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational test window. In the context of adaptive test administrations,

other reports such as blueprint match and item exposure reports allow psychometricians to verify that test administrations conform to specifications.

The quality assurance reports are generated on a regular 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.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Table 108 presents an overview of the quality assurance (QA) reports.

<div align="center">

Table 108: Overview of Quality Assurance Reports

</div>

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 items)
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
Blueprint Match	To monitor match to test blueprint	Early detection of blueprint violation

1.43.2.1 Item Analysis Report

The item 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 the 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. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as item response theory (IRT)–based item fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or generating reports based on all items in the pool.

Item* p-*Value. For multiple-choice items, the proportion of students selecting each response option is computed; for constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for selected-response items and polyserial correlations for polytomous constructed response, performance, and technology items are computed. CAI psychometric staff evaluates all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item fit index is produced for each item. For each student, a residual between the observed and expected scores given the student's ability is computed for each item. The residuals are averaged across all students, and the average residual is used to flag an item.

1.43.3 QUALITY ASSURANCE IN REPORTING

Scores for the ILEARN 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 IRT parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to psychometricians.

After passing through the series of validation checks in the QM System, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring there is only one place where the "official" record is stored. Only after scores have passed the QM checks and are uploaded to the DOR are they passed to the Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in the Reporting System until it passes all   validation checks.

Data accuracy and integrity is critical to the success of assessment validity. Additional quality assurance steps and procedures outside of the online systems are performed by staff on the assessment data to ensure accuracy before score reporting is considered final. These additional verifications ensure that the final data is thoroughly reviewed and accurate.

CAI psychometrics perform item pool simulations testing the adaptive algorithm on the ILEARN assessments to ensure that the system coding is performing as intended and students are receiving the appropriate test items based on their performance. These verifications are done by CAI psychometric staff with the resulting deliverable of a simulation report to IDOE containing information on test items and item distribution to the student population. The simulation report is also provided to a third party vendor to replicate simulation results independently from CAI ahead of the testing window.

IDOE established a production test deck verification process used annually to ensure the scores are being reported as expected based on specific results entered into the online test administration systems and reporting system. Test deck cases are entered

for demo students following certain patterns that are then verified in the reporting system prior to the initial score release in the CAI reporting system. This check ensures that student scores are populating through the system accurately, using end to end testing. Both online and paper-pencil test deck cases are entered and processed to ensure both modes provide accurate score reporting results which are verified in the CAI reporting system.

The student data files contain the final student test score results, and these files are delivered to IDOE annually at the conclusion of each test administration. A third party vendor will replicate sample, initial and final student data files to confirm accuracy of the scoring as part of quality control measures. The third party vendor is provided a layout, configuration files and scoring specifications from CAI to support this verification. CAI, IDOE and the third party vendor will meet as needed during this process to ensure that questions are answered, and replication is completed on time.

10. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

———. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach (ETS Research Report No. 91–47). Princeton, NJ: Educational Testing Service.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436. https://doi.org/10.1007/BF02295430

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

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics*, *4*, 231–246.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*(3), 381–389.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*, 7th Ed.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/BF02296207

Rijmen, F. (2010). Formal relations and an empirical comparison among the Bi-Factor, the Testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Samejima, F. (1972). A general model for free-response data. *Psychometrika* Monograph Supplement, 37(1, Pt. 2), 68.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*,106–108.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. 12–08). Princeton, NJ: Educational Testing Service.