Annual Technical Report

Indiana ILEARN Checkpoint Assessments

2024–2025

ACKNOWLEDGMENTS

This technical report was produced on behalf of the Indiana Department of Education (IDOE). Requests for additional information concerning this technical report or the associated appendices should be directed to IDOE at INassessments@doe.in.gov.

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Stephan Ahadi, Elizabeth Xiaoxin Wei, Maryam Pezeshki, Guanlan Xu, Kevin Clayton, Christina Sneed, Jessica Singh, Diego Heimlich, and Quentin Leifer. Major contributors from IDOE include Lynn Schemel, Mary Krivulka, Logan Miller, and Alyson Traficante.

CONTENTS

TABLES

FIGURES

EXHIBITS

APPENDICES

EXECUTIVE SUMMARY

This executive summary provides an overview of validity evidence for the Indiana's Learning Evaluation and Assessment Readiness Network (ILEARN) Through-Year Assessment (TYA) design to support a validity argument regarding the uses of and inferences for the ILEARN Checkpoint Assessments as well as a summary of the TYA program and its 2024–2025 school year pilot test administration. The program includes three Checkpoint assessments and the end-of-year summative ILEARN Assessment. This technical report addresses the ILEARN Checkpoint Assessments.

Overview of Validity Evidence

ILEARN Checkpoints are low-stakes assessments intended to inform instruction. These assessments are not incorporated into accountability models. Intended uses for Checkpoint test scores include the following:

- Determining (alongside other local data points) if individual students have obtained proficiency with specific sets of academic standards after opportunity to learn
- Identifying students who need additional support to achieve proficiency by the end of the year
- Supporting the development of intervention or remediation plans for students who need continued support for specific sets of standards (through data points and associated Performance-Level Descriptors [PLDs])
- Considering areas of strength and weakness for specific groups of students, including class, grade, and special populations
- Determining ways to support Tier I, Tier II, and Tier III instruction so that students can achieve mastery of academic standards prior to the end of the school year (including an early warning system for students who are less likely to achieve mastery without significant intervention)

Evidence for the validity of test score interpretations is central to substantiating claims that Checkpoint test scores can fulfill their intended purpose to support instructional decision-making and interventions aligned to the Indiana Academic Standards (IAS) throughout the school year prior to the summative assessment.

Sufficient evidence exists to support the principal claims for the Checkpoint test scores, including that test scores indicate the degree to which students have achieved specific

IAS measured by each Checkpoint at each grade level. The details of such validity evidence are presented fully in Chapter 2 and briefly summarized here.

- **Strong Alignment to IAS (Content Standards).** Alignment is achieved by a rigorous, highly-iterative item development process that incorporates Range Performance-Level Descriptors (RPLDs) and specific constructs as well as universal design to ensure access to all student populations. The Indiana Assessment Frameworks provide test blueprints that specify the academic standards that will be assessed on each Checkpoint, the priority given to each standard within the blueprint, and item specifications that outline exactly how that content will be measured. These frameworks are a strong link between IAS and content-based test score interpretations. (https://www.in.gov/doe/students/assessment/ilearn/#Indiana_Assessment_Frameworks)
- **Minimization of Construct-Irrelevant Factors.** ILEARN Checkpoints adopted the use of universal design to remove barriers to access for the widest range of students possible. Universal design principles are embedded in the item development process, considered in the development of the online test delivery platform, and a part of the design of universal features provided to all students during test administration.
- **Internal Structure of the Assessments.** 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 Checkpoints and 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 Checkpoint scores align with the underlying structure of the test and provide evidence for appropriateness of the selected item response theory (IRT) models. Section 2.4, Evidence Based on Internal Structure, summarizes results from confirmatory factor analysis, which provides evidence for the unidimensionality of the TYA design.

The interpretation of the Checkpoint 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 IAS. Checkpoint test scores are reported with respect to four proficiency levels, demarcating the degree to which Indiana students participating in the assessments have achieved the learning expectations defined by the subset of IAS measured by each respective Checkpoint. The standardized and rigorous procedures that Indiana educators, serving as standard-setting panelists, followed to recommend performance standards in the standard-setting process after the Spring 2019 ILEARN test administration provided central and strong evidence to support the validity of test score interpretations regarding performance standards. Checkpoint assessments use the same

performance standards and four proficiency levels as those from the summative ILEARN Assessment.

Summary of the Assessment Program

Checkpoint assessments measure student achievement and growth according to the IAS for grades 3–8 English/Language Arts and grades 3–8 Mathematics.

Students' achievement is measured at three Checkpoints following opportunity to learn. The measures support identification of students who need additional support and intervention to master specific academic standards before the end-of-year summative ILEARN Assessment. A student may have a second test opportunity for any given Checkpoint assessment to support teachers in monitoring progress through interventions or remediation. Since each Checkpoint assessment measures a subset of the summative blueprint, starting in the 2025–2026 school year, student performance on domains and subdomains will be combined throughout the year and reported after the summative assessment. The 2024–2025 school year data were collected to model and inform implementation of a combined domain and subdomain scoring starting in the 2025–2026 school year. This document provides Checkpoint-specific scores, validity arguments for the overall TYA design, results from research on prediction probability, and combined domain scoring. Since the summative ILEARN is the only assessment used for accountability purposes, summative-specific analyses are documented in the ILEARN technical report.

Checkpoint assessments were created using a variety of item types from several sources, including licensed item banks such as the Smarter Balanced Assessment Consortium (Smarter Balanced) item bank, Cambium Assessment, Inc.'s (CAI) ClearSight item bank, and custom Indiana item banks. 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.

Checkpoint 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* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 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 Checkpoint assessments.

Checkpoint scores are provided to corporations and schools through the Indiana Centralized Reporting System (CRS). CRS is designed to assist stakeholders in reviewing and downloading test results and in understanding and appropriately using 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 levels provide information about the strengths and improvement opportunities of students and can be used to improve teaching and student learning.

CAI's ClearFrame API for use with ILEARN TYAs will be launched in the 2025–2026 school year and available for all Indiana corporations.

ClearFrame supports the ILEARN program by allowing school corporations to easily create data connections with participating curriculum providers. These providers automatically receive high-quality scores from ILEARN Checkpoints during the school year and provide customized learning paths for each student in time to make a difference.

Corporation Test Coordinators (CTCs) and technology directors for all Indiana corporations will be provided access to the API Connection Manager feature in CRS, where they can choose the vendor applications to set up connections. Once the ILEARN Checkpoints are selected in the API Connection Manager and a new connection is created, corporations will work with participating vendors so that score data for Checkpoints 1, 2, and 3 will be automatically sent as these data become available in CRS. Finally, quality assurance (QA) procedures are enforced throughout all stages of Checkpoint test development, configuration, administration, 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 2024–2025 Checkpoint test administration, which provides summaries of the test-taking student population and their performance on the assessments. In addition, these sections describe test administration-specific evidence for the reliability of Checkpoint

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 Checkpoint assessments, including the IAS, which define the content domain to be assessed by Checkpoints; 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. Chapter 5 discusses test administration procedures, including eligibility for participation in Checkpoint 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 Checkpoint assessments for scoring and reporting. Chapter 7 provides a description of the score reporting system and the interpretation of test scores. Finally, Chapter 8 provides an overview of the 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 ILEARN Through-Year Assessment (TYA) consists of three Checkpoints and one summative across the course of the academic year. The primary purpose of the ILEARN TYA design is to improve the quality and usefulness of state assessment systems through innovation for instructional utility by

- providing actionable data throughout the year for educators and families;
- shortening overall testing time throughout the year, including summative assessment;
- connecting teachers directly to instructional support for students;
- providing redesigned reports for instructional purposes; and
- providing quality professional development for assessment, instruction, and data literacy.

The TYA theory of action is to teach, assess, and reteach.

- Teachers prepare using assessment and instructional frameworks. Students participate in learning activities for the set of standards.
- Teachers administer the Checkpoint assessments to see if students have mastered the required knowledge and skills.
- Teachers remediate, reinforce, or accelerate learning based on assessment results.
- Teachers administer an optional (Opportunity 2) assessment or other local progress-monitoring assessment to see if remediation or reinforcement of skills was successful.
- Students master the required content by the summative assessment.

Thus, by administering Checkpoint assessments multiple times during the year, educators receive more timely feedback to inform their instructional practices and address student learning needs, and students receive targeted remediations to reinforce their learning and prepare for the summative assessment.

The ILEARN Checkpoints are low-stakes assessments intended to inform instruction. These assessments are not incorporated into accountability models.

Additional intended uses include 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. The TYA design 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 Indiana Academic Standards (IAS). This design supports instruction and student learning by providing immediate feedback to educators and families 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.

The TYA design 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* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014).

1.2 Background of the Assessments

Indiana's TYA system comprises three Checkpoint assessments and a statewide end-of-year summative assessment. The Checkpoint assessments are designed to align academic content standards temporally through the school year so that each Checkpoint assesses students' knowledge and skills with respect to each standard more nearly at the point where instruction on the standards has been completed. The summative assessment is the only component used for the state's accountability model.

Each Checkpoint assessment was constructed to measure student achievement in English/Language Arts (ELA) and Mathematics relative to the IAS. Checkpoint assessments were first administered to students during the 2024–2025 school year as fixed forms. Starting in the 2025–2026 school year, Checkpoint assessments (Opportunity 1) will be administered adaptively. The 2024–2025 school year was a pilot year for the ILEARN TYA test administration. Schools and corporations opted in to participate in the Checkpoints. Seventy-five percent of schools opted in to participate in the Checkpoints.

1.2.1 Development of the IAS

The IAS were approved by the Indiana State Board of Education in April 2014 for ELA and Mathematics. The IAS were most recently updated in 2023 for all subjects (see the Spring 2023 Standards Implementation memo for more details regarding instructional and assessment timelines). The Checkpoints assess the 2023 IAS.

The IAS are intended to implement more rigorous standards that promote college and career readiness, with the goal of challenging and motivating Indiana students to acquire stronger critical thinking, problem-solving, and communication skills. For additional information on the development of the IAS, see https://www.in.gov/doe/students/indiana-academic-standards/.

1.2.2 Online Item Pool Construction

The 2024–2025 Checkpoint item pools each contain sufficient numbers of items per grade and content area to administer two fixed-form test opportunities, representing the breadth and depth of the content standards identified in the test specifications.

With new items being developed and field-tested in the Spring ILEARN test administration each year, the operational pool size for each assessment is constantly increasing. Starting in the 2025–2026 school year, Checkpoint assessments (Opportunity 1) will also be administered adaptively like the summative ILEARN.

2. The Validity of Test Score Interpretations

2.1 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 [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 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, 1999, 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, 1999, 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.

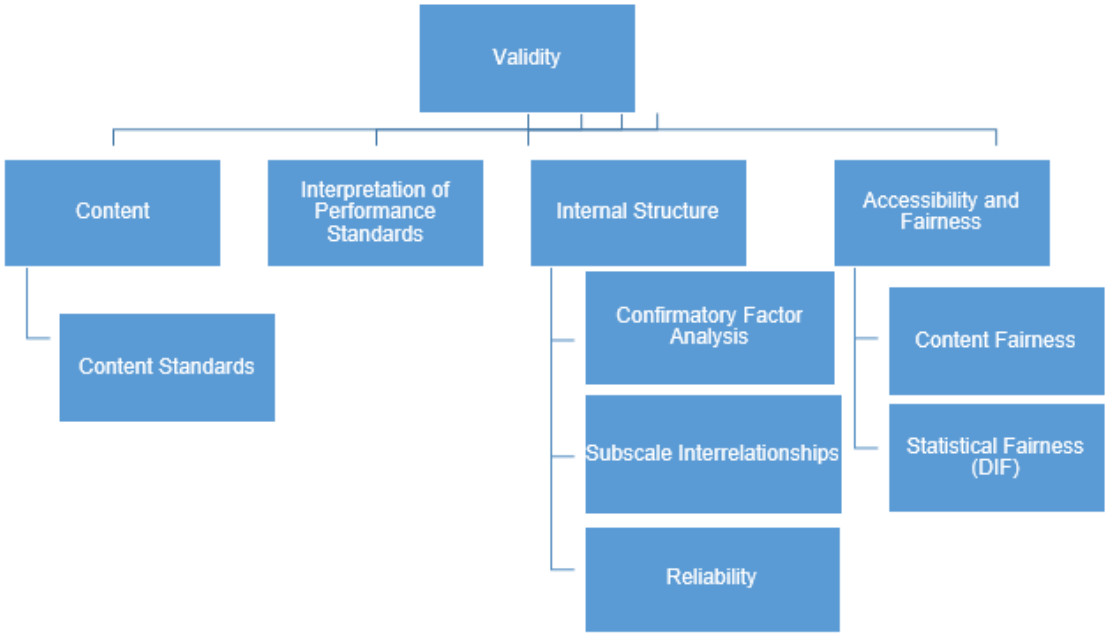
The kinds of evidence required to support the validity of test score interpretations depend on the claims made for how test scores may be interpreted. Moreover, the *Standards* make it explicit that validity is an attribute not of tests but rather of test score interpretations. Thus, the test itself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

There are several intended uses for Checkpoint test scores, including 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 the claims, including evidence showing that test items and administrations align with the Indiana Academic Standards (IAS). 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 Checkpoint and summative assessments, the definition of the measurement construct is provided by the IAS.

A summary of the types of validity evidence gathered is captured and illustrated in the following chart.

Figure 1. Types of Validity Evidence



2.2 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 the ILEARN Through-Year Assessment (TYA), the definition of the measurement construct is provided by the IAS, which specify what students should know and be able to do by the end of the year for each grade level. The ILEARN TYA is designed to measure student progress toward achievement of the IAS. Therefore, the validity of through-year test score interpretations critically depends on the degree to which test content aligns with expectations for student learning as specified in the IAS.

Several processes are in place to ensure that Checkpoints fully align 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 IAS with the Checkpoint item content. The Checkpoint 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 IAS and the Checkpoint content-based test score interpretations.

2.2.1 Content Standards

The IAS were most recently revised in 2023. The standards are available for review in the following URLs:

- https://www.in.gov/doe/students/indiana-academic-standards/englishlanguage-arts/
- https://www.in.gov/doe/students/indiana-academic-standards/mathematics/

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.

Tables 1 and 2 present the number of items in the 2024–2025 item pool that measured each reporting category by grade for English/Language Arts (ELA) and Mathematics.

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

Grade	Checkpoint	Reporting Category	Number of Items
3	CP1	Reading Foundations	19
		Reading and Understanding Fiction	24
	CP2	Reading Foundations	15
		Reading and Understanding Informational Text Media	25
	CP3	Reading Foundations	8
		Reading and Understanding Informational Text Media	10
		Writing	15
4	CP1	Reading and Understanding Fiction	29
		Understanding and Using Vocabulary	6
	CP2	Reading and Understanding Informational Text Media	29
		Understanding and Using Vocabulary	8
	CP3	Understanding and Using Vocabulary	10
		Writing	18
5	CP1	Reading and Understanding Fiction	30
		Understanding and Using Vocabulary	6
	CP2	Reading and Understanding Informational Text Media	33
		*NA**	--
	CP3	Understanding and Using Vocabulary	8
		Writing	17
6	CP1	Analyzing Literature	30
		Understanding and Using Vocabulary	9
	CP2	Analyzing Informational Text Media	30
		Understanding and Using Vocabulary	1
	CP3	Analyzing Informational Text Media	17
		Writing	15
7	CP1	Analyzing Literature	32
		Understanding and Using Vocabulary	10
	CP2	Analyzing Informational Text Media	36

Grade	Checkpoint	Reporting Category	Number of Items
		Understanding and Using Vocabulary	5
	CP3	Analyzing Informational Text Media	14
		Writing	22
8	CP1	Analyzing Literature	26
		Understanding and Using Vocabulary	16
	CP2	Analyzing Informational Text Media	25
		Understanding and Using Vocabulary	15
	CP3	Analyzing Informational Text Media	13
		Writing	22

*Communication and Collaboration items were added to the test but they were not considered a second reporting category.

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

Grade	Checkpoint	Reporting Category	Number of Items
3	CP1	Addition and Subtraction of Money	22
		Place Value	16
	CP2	Multiplication and Division	26
		Understanding Fractions	13
	CP3	Measurement and Data	16
		Understanding Fractions	22
4	CP1	Multiplication	19
		Place Value	14
	CP2	Measurement	17
		Understanding Fractions, Multiplication, and Division	23
	CP3	Calculation Fractions and Decimals	18
		Solving Problems	4
		Understanding Fractions and Decimals	18
5	CP1	Multiplication and Division of Whole Numbers	18
		Volume	20
	CP2	Computing Decimals and Fractional	17
		Understanding Percents, Decimals, and Fractions	21
	CP3	Computing Decimals and Fractions	19
		Measurement and Data Analysis	17
6	CP1	Integers and Rational Numbers	19
		Operations with Positive Numbers	23
	CP2	Equations and Inequalities	20
		Expressions and Data Analysis	10
	CP3	Solving Problems with Rates, Ratios, and Proportions	22
		Working with Rates and Ratios	16

Grade	Checkpoint	Reporting Category	Number of Items
Grade	Checkpoint	Reporting Category	Number of Items
7	CP1	Computing with Rational Numbers	28
		Exponents and Roots	12
	CP2	Proportional Relationships	27
		Slope	15
	CP3	Expressions, Equations, and Inequalities	28
		Geometry	10
8	CP1	Linear Equations and Inequalities	16
		Real Numbers	6
	CP2	Functions	15
		Linear Functions	24
	CP3	Functions	12
		Geometry	24

2.2.2 Review Process for Items Appearing in Checkpoint Operational Test Administration

This section describes the item review procedures used to ensure item accuracy and alignment with the IAS. Checkpoint items are field-tested in ILEARN. All items developed by Cambium Assessment, Inc. (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 One, performed by a Level 3–4 CAI content specialist
3. Edit Review One, in which a copyeditor 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 (i.e., it assesses something worthwhile in a reasonable way).
- The item aligns correctly to a Depth of Knowledge (DOK) level.
- 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 (SR) 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.

Based on their reviews of each item, 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 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 item pools are initially reviewed as follows:

- IDOE reviews to ensure that Mathematics and ELA items developed in the Indiana item bank are eligible for Content and Fairness Committee (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 Indiana passage and item specifications and existing items in the bank to determine whether the requested edits will be made. IDOE reviews all items to ensure the following:
    - Alignment to content standard

  - Adherence to all item specification requirements
  - Accessibility for all student populations
  - Appropriate linguistic complexity
  - Accurate scoring rules
  - Representation of content across the depth and breadth of the academic standard (comparing what is already in the item bank and ensuring that gaps are addressed)
- Indiana 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 Language Accessibility and Bias and Sensitivity (LABS) Committee.
- After all IDOE- and committee-recommended edits have been applied, experts apply accessibility markups (e.g., translations, 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.
- Each year, CAI content leads conduct gap analyses in which the operational pool is summarized by IAS and compared against blueprint target ranges. CAI used the results of this gap analysis to identify Indiana standards to potentially assess with licensed items. Licensed items available for Indiana's use were first tentatively aligned to the IAS via crosswalks provided by IDOE. These tentative alignments were then reviewed and updated, if necessary, by IDOE content specialists prior to being subjected to educator review. This item acceptance review (IAR) process for ensuring alignment of licensed items to the IAS was jointly developed by CAI and IDOE and has been used multiple times to bring items into the ILEARN item pool from several external item banks.
- In June 2023, an IAR meeting was held for ELA and Mathematics for items licensed from Smarter Balanced. Items that were accepted as aligned to the IAS were accepted for operational use in the Indiana Checkpoint item pool.

Indiana-owned 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. Items licensed from Smarter Balanced have already been field-tested prior to their acceptance into the Indiana checkpoint item pool and are deemed ready for operational use in the Indiana checkpoint item pool upon acceptance at the IAR.

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 (DIF analyses).

Items flagged for review based on 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, the response keys are correct, and there are no other obvious problems with the items.

IDOE then convenes the data review committee to evaluate flagged field-test items in the context of each item's statistical performance. Based on 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.

2.3 Evidence for Interpretation of Performance Standards

Alignment of test content to the IAS ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the IAS. However, the interpretation of the Checkpoint 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 IAS. Checkpoint test scores are reported with respect to four proficiency levels, demarcating the degree to which Checkpoint students have achieved the learning expectations defined by the IAS. 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 IAS and that they are prepared to benefit from instruction at the next grade level. Procedures used to adopt performance standards for the Checkpoint assessments are therefore central to the validity of test score interpretations.

Checkpoint assessments adopted the same proficiency levels as those of the summative ILEARN Assessment. Following the operational administration of the summative ILEARN Assessments in 2018–2019, a standard-setting workshop was conducted to recommend a set of performance standards to IDOE for reporting student performance of the IAS. This section summarizes 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 IAS and student achievement to map the Performance-Level Descriptors (PLDs) adopted by IDOE onto an ordered-item booklet based on operational test forms administered to students in Spring 2019. Chapter 7 of the ILEARN Technical Report, Performance Standards, explains the standard-setting procedures in more detail and the accompanying appendices provide additional information and documents about standard-setting meetings.

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 grades 3–8 ELA and Mathematics assessments were provided with the approximate location of relevant National Assessment of Educational Progress (NAEP) and Smarter Balanced Assessment Consortium (Smarter) performance standards. 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 3 shows the estimated percentage of students meeting the ILEARN proficient standard for each assessment in Spring 2019. Table 3 also shows the national percentages of students who 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 3 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 3: 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

Grade	ILEARN At Proficiency	NAEP Proficient	Smarter Proficient
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

2.4 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 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 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. In case of the TYA design, the internal structure can be analyzed by assuming a general construct (second factor), which is the overall achievement, with time-specific Checkpoint and summative assessments (first factors) and the items loading onto the time-specific construct 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 for the TYA design. The internal structure of each Checkpoint assessment can also be analyzed using a second-order factor model in which the second order is the subject-specific construct and the first-order factors are the reporting categories with items loading onto the reporting categories they intend to measure.

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 among relevant reporting categories. Disattenuated correlations among reporting category scores are presented in Section 3.4, Subscale Intercorrelations.

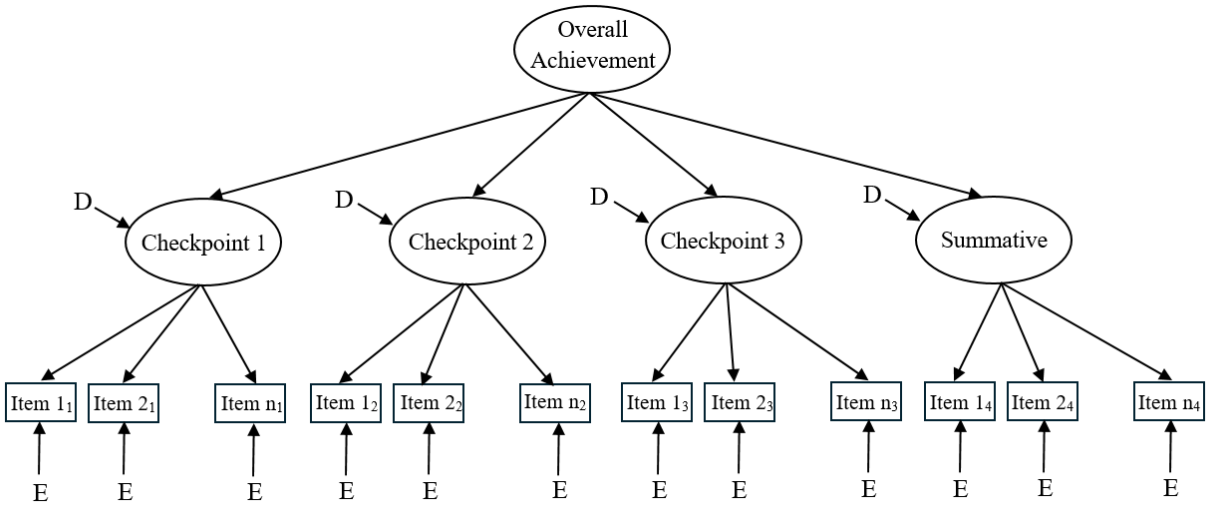
An analysis of the internal structure of a test can also include evidence for test reliability by ensuring that the test is measuring the construct consistently. Reliability analyses are discussed in detail in Section 3.3, Reliability.

2.4.1 Confirmatory Factor Analysis

Indiana's TYAs, comprising four measurement occasions (i.e., three Checkpoints and an end-of-year summative ILEARN), 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 four measurement occasions, and items load onto factors they are intended to measure. The underlying structure of the TYAs was common across all grades, which is useful for comparing the results of the analyses across the grades.

Within each subject area (e.g., ELA), items are designed to measure subject-specific knowledge. Measurement occasions within each subject area are, in turn, indicators of achievement in the subject area. The form of the second-order CFAs is illustrated in Figure 2. As the figure illustrates, each item is an indicator of a measurement occasion. Because items are never pure indicators of an underlying factor, each item also includes an error component. Similarly, each measurement occasion serves as an indicator of overall achievement in a subject area. As at the item level, the measurement occasions include an error term indicating that they are not pure indicators of overall achievement in the subject area. The paths from the measurement occasions to the items represent the first-order factor loadings, or the degree to which items are correlated with the underlying construct. Similarly, the paths from subject-area achievement to measurement occasions represent the second-order factor loading, indicating the degree to which they correlate with the underlying subject-area achievement construct.

Figure 2: Second-Order Structural Model for TYAs



It may not be reasonable to expect that the measurement occasion scores are completely orthogonal—this would suggest that there are no relationships among measurement occasion scores and would make justification of a unidimensional item response theory (IRT) model difficult. However, reporting these separate scores could be easily justified. On the contrary, if the measurement occasions were perfectly correlated, using a unidimensional model could be justified, but reporting separate scores could not be easily justified.

While the test consisted of items targeting different standards, all items within a grade and subject were calibrated concurrently using the 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 based on this theory.

The results in this section were based on the data collected from the initial administration of the TYAs, which was the 2024–2025 test administration. The purpose is to provide validity evidence regarding the dimensionality of the assessments and to show that the methods for reporting TYA 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.

2.4.2 Factor Analytic Method

A series of CFAs were conducted using the statistical program Mplus, version 8.10, (Muthén & Muthén, 2017) for each grade and subject assessment. The estimation method, weighted least squares means and variances (WLSMV) adjusted, 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 χ2 statistic, with good model fit indicated by a non-significant χ2 statistic. However, the χ2 statistic is sensitive to sample size, so even well-fitting models will demonstrate highly significant χ2 statistics given a large number of students. Therefore, fit indices, such as the comparative fit index (CFI; Bentler, 1990), Tucker–Lewis Index (TLI; Tucker & Lewis, 1973), and root mean square error of approximation (RMSEA) were also used to evaluate model fit. Table 4 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.

Table 4: Guidelines for Evaluating Goodness of 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 $\boldsymbol{S}$ of tetrachoric and polychoric sample correlations among the item scores (Olsson, 1979), and the matrix $\boldsymbol{W}$ 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:

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

In this equation, $\widehat{\boldsymbol{\Sigma}}$ 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

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

where $\boldsymbol{\Lambda}$ is the matrix of item factor loadings (with $\boldsymbol{\Lambda}'$ representing its transpose), and $\boldsymbol{\Theta}$ is the uniqueness, or measurement error. The matrix $\boldsymbol{\Phi}$ is the correlation among the separate factors. For the base model, items are thought only to load onto a single underlying factor. Hence $\boldsymbol{\Lambda}'$ is a $p \times 1$ vector, where p is the number of test items and $\boldsymbol{\Phi}$ is a scalar equal to 1. Therefore, it is possible to drop the matrix $\boldsymbol{\Phi}$ 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 for the TYA design in which test items are coerced to load onto the measurement occasions they are designed to target, and all measurement occasions share a common underlying factor. The second-order factor analysis can be mathematically represented as

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

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

2.4.3 ELA Content Model

The goodness-of-fit statistics for the hypothesized TYA second-order models in ELA are shown in Table 5. All the statistics indicate that the second-order models posited by the TYAs fit the data well. This pattern was true across all grades. The CFI and TLI values are all equal to or greater than .97. The RMSEA values are all around 0.01, well below the values used to indicate good fit.

Table 5: Goodness of Fit for the TYA ELA Second-Order Models

Grade	df	RMSEA	CFI	TLI	Convergence
Second-Order Models					

Grade	df	RMSEA	CFI	TLI	Convergence
3	4846	0.013	0.972	0.972	Yes
4	4273	0.009	0.986	0.985	Yes
5	4273	0.010	0.986	0.985	Yes
6	4090	0.010	0.988	0.987	Yes
7	4555	0.012	0.983	0.983	Yes
8	4555	0.010	0.984	0.984	Yes

2.4.4 Mathematics Content Model

The goodness-of-fit statistics for the TYA second-order models in Mathematics are shown in Table 6. The models generally show good fit, although the CFI and TLI fit indices are less than the cutoff value of 0.95 for grade 7. Even for this grade, however, the RMSEA estimate is well below its respective 0.05 cutoff value. All of the statistics indicate the second-order models are a good fit for the data.

Table 6: Goodness of Fit for the TYA Mathematics Second-Order Models

Grade	df	RMSEA	CFI	TLI	Convergence
Second-Order Models					
3	6436	0.016	0.968	0.967	Yes
4	6436	0.017	0.972	0.971	Yes
5	6323	0.020	0.951	0.950	Yes
6	6781	0.014	0.970	0.970	Yes
7	7016	0.023	0.940	0.939	Yes
8	7016	0.016	0.967	0.966	Yes

2.5 Evidence of Fairness and Accessibility

2.5.1 Fairness in Content

The universal design principles 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 universal design principles are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002). They include the following:

1. Inclusive assessment population

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

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

Coupled with the design of fair and accessible items, the TYAs include a full array of accessibility features and accommodations appropriate for each student. These options include appropriate universal features, designated features, and accommodations, when needed, based on the constructs being measured by the assessment.

IDOE implements systematic steps through item development and content presentation to ensure that accessibility is interwoven throughout all stages of assessment delivery and scoring outcomes. The validity of assessment results depends on the utilization of the full array of accessibility features and accommodations appropriately for each student.

2.5.2 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 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, or +C), signifying that the item favored the focal group (e.g., African American/Black, Hispanic, Female), or negatively (i.e., –A, –B, or –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 are 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
- Socioeconomic Status (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). The results of DIF analyses are presented in Appendix 4M of the ILEARN Technical Report.

2.6 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 Checkpoint test scores indicate the degree to which students have achieved the IAS at each grade level and that students scoring at the Proficient level or higher demonstrate levels of achievement consistent with national benchmarks. These claims are supported by evidence of a test development process that ensures alignment of test content to the IAS and evidence that the structural model described by the IAS and implemented in the TYAs is sound.

3. Summary of the Checkpoint Test Administration

Checkpoints are fixed-form assessments in the initial pilot for the 2024–2025 school year for English/Language Arts (ELA) and Mathematics. Students participating in the computer-based Checkpoint assessments 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, Checkpoints provide accommodated forms for braille and Spanish. Visually impaired students can take the braille version of Checkpoint ELA and Mathematics. English learners (ELs) can take the Spanish language version of Mathematics.

The following tests were available in the 2024–2025 administration:

- Grades 3–8 ELA
- Grades 3–8 Mathematics

3.1 Student Population and Participation

For the 2024–2025 pilot, 75% of Indiana schools participated in grades 3–8 ELA and Mathematics Checkpoint assessments. This section presents results for Opportunity 1 tests. Since only a small subset of the students took the Opportunity 2 tests, those results are reported in Appendix 3-G, Opportunity 2 Descriptive Statistics. Table 7 shows the number of students tested and the number of students reported for the 2024–2025 Checkpoint assessments. It is important to note that participation based on enrollment is high (i.e., 99%).

Table 7: Number of Students Participating in Checkpoints

		ELA					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
CP1	Number Tested	57,815	58,366	60,650	60,855	62,453	61,492
	Number Reported	57,674	58,255	60,448	60,633	62,032	61,172
CP2	Number Tested	57,441	58,096	60,393	60,434	61,423	60,584
	Number Reported	57,309	57,942	60,242	60,186	61,040	60,300
CP3	Number Tested	56,187	57,605	59,908	59,969	60,613	59,669
	Number Reported	56,103	57,538	59,826	59,746	60,346	59,447
		Mathematics					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
CP1	Number Tested	57,875	58,641	60,822	60,892	63,047	61,630
	Number Reported	57,754	58,554	60,533	60,645	62,766	61,015
CP2	Number Tested	57,288	58,246	60,491	60,442	62,142	60,703
	Number Reported	57,156	58,081	60,325	60,139	61,655	60,420
CP3	Number Tested	56,126	57,832	60,068	59,921	61,245	59,914
	Number Reported	55,974	57,723	59,917	59,690	60,616	59,609

Tables 8 and 9 present the distribution of student subgroups in percentages. The subgroup categories reported are gender, ethnicity, students with special education (SPED) status, students with Section 504 Plans, ELs, and socioeconomic status (SES).

Table 8: 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	CP1	57,815	51.16	48.84	61.38	13.43	3.02	15.87	0.14	0.12	6.05	17.74	2.16	14.99	54.84
	CP2	57,441	51.15	48.85	61.33	13.43	3.04	15.86	0.14	0.12	6.08	18.13	2.42	12.69	54.85
	CP3	56,187	51.09	48.91	61.94	12.95	3.10	15.68	0.15	0.11	6.08	18.60	2.81	12.54	55.03
G4	CP1	58,366	51.13	48.87	61.86	13.22	2.99	15.67	0.16	0.11	5.99	18.55	3.00	14.65	53.68
	CP2	58,096	51.07	48.93	61.79	13.28	3.01	15.67	0.16	0.10	5.98	18.65	3.20	11.86	53.43
	CP3	57,605	51.05	48.95	61.95	13.06	2.99	15.76	0.16	0.10	5.96	18.76	3.37	11.91	53.75
G5	CP1	60,650	51.15	48.85	62.12	13.09	3.04	15.69	0.12	0.12	5.83	18.17	3.39	13.93	52.60
	CP2	60,393	51.11	48.89	62.00	13.09	3.07	15.80	0.12	0.12	5.81	18.06	3.53	10.74	52.14
	CP3	59,908	51.11	48.89	62.24	12.81	3.07	15.83	0.11	0.12	5.83	18.09	3.72	10.78	52.32

Grade	Year	*N*	Male	Female	White	AfAm	Asian	Hisp	AmIndian	Pacific	Multi	SPED	S504	EL	SES
G6	CP1	60,855	51.10	48.90	62.60	13.03	3.20	15.40	0.14	0.10	5.52	17.42	3.69	15.06	51.37
	CP2	60,434	51.08	48.92	62.58	13.00	3.24	15.45	0.14	0.10	5.48	17.42	3.86	9.34	50.73
	CP3	59,969	51.03	48.97	62.64	12.85	3.27	15.48	0.15	0.10	5.50	17.36	3.97	9.40	50.93
G7	CP1	62,453	51.31	48.69	61.78	13.57	3.51	15.53	0.16	0.13	5.33	15.48	3.90	15.95	49.60
	CP2	61,423	51.32	48.68	61.87	13.49	3.45	15.57	0.16	0.13	5.34	15.51	4.01	9.80	49.17
	CP3	60,613	51.34	48.66	61.87	13.35	3.49	15.67	0.17	0.14	5.32	15.36	4.19	9.89	49.37
G8	CP1	61,492	51.13	48.87	62.16	13.41	3.19	15.57	0.17	0.10	5.40	15.29	4.07	15.36	48.43
	CP2	60,584	51.07	48.93	62.24	13.37	3.16	15.63	0.16	0.10	5.33	15.11	4.20	9.34	47.74
	CP3	59,669	51.04	48.96	62.17	13.26	3.21	15.75	0.16	0.11	5.34	15.04	4.37	9.45	47.87

Table 9: 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	CP1	57,875	51.24	48.76	61.45	13.44	3.03	15.77	0.14	0.12	6.04	17.75	2.17	15.02	54.76
	CP2	57,288	51.19	48.81	61.47	13.36	3.06	15.79	0.14	0.12	6.05	18.12	2.43	12.69	54.76
	CP3	56,126	51.19	48.81	62.02	12.90	3.10	15.64	0.14	0.11	6.09	18.58	2.82	12.54	55.00
G4	CP1	58,641	51.08	48.92	61.88	13.22	2.99	15.65	0.16	0.11	5.99	18.48	2.99	14.61	53.49
	CP2	58,246	51.06	48.94	61.82	13.22	3.01	15.72	0.16	0.10	5.98	18.62	3.19	11.85	53.40
	CP3	57,832	51.05	48.95	61.96	13.06	2.99	15.76	0.16	0.11	5.96	18.67	3.37	11.95	53.65
G5	CP1	60,822	51.16	48.84	62.21	13.06	3.04	15.65	0.12	0.12	5.82	18.18	3.38	14.17	52.57
	CP2	60,491	51.09	48.91	62.10	13.04	3.06	15.75	0.12	0.12	5.81	18.02	3.51	10.70	52.12
	CP3	60,068	51.11	48.89	62.30	12.80	3.06	15.77	0.12	0.12	5.83	18.07	3.75	10.72	52.35
G6	CP1	60,892	51.06	48.94	62.59	13.01	3.20	15.44	0.15	0.11	5.51	17.42	3.71	14.13	51.20
	CP2	60,442	51.07	48.93	62.53	13.03	3.24	15.49	0.15	0.10	5.46	17.44	3.78	9.36	50.73
	CP3	59,921	51.00	49.00	62.60	12.90	3.28	15.48	0.16	0.10	5.49	17.37	3.99	9.42	50.89
G7	CP1	63,047	51.37	48.63	61.84	13.52	3.49	15.51	0.16	0.13	5.34	15.56	3.97	15.24	49.67
	CP2	62,142	51.32	48.68	61.92	13.36	3.48	15.60	0.16	0.14	5.34	15.50	4.00	9.83	49.08
	CP3	61,245	51.33	48.67	61.91	13.27	3.51	15.68	0.17	0.13	5.33	15.39	4.24	9.89	49.33
G8	CP1	61,630	51.11	48.89	62.06	13.47	3.18	15.57	0.17	0.10	5.44	15.35	4.11	14.61	48.76
	CP2	60,703	51.06	48.94	62.15	13.36	3.18	15.68	0.16	0.10	5.38	15.16	4.20	9.42	47.81
	CP3	59,914	51.03	48.97	62.07	13.32	3.22	15.78	0.16	0.11	5.34	15.07	4.41	9.55	47.98

3.2 Summary of Overall Student Performance

The 2024–2025 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 10 and 11. Additionally, Figures 3 and 4 display the student percentages at each proficiency level. It should be noted that the Through-Year Assessment (TYA) is using a partial blueprint design (i.e., each Checkpoint assesses a subset of the standards). Although all Checkpoints measure the same underlying trait, care should be taken in comparing the scale scores across Checkpoints, as these scores are not on the same scale as the summative.

Table 10: 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	CP1	57,674	5410.35	78.23	57.09	19.39	14.27	9.25	23.52
	CP2	57,309	5431.25	87.82	46.45	21.71	16.86	14.98	31.83
	CP3	56,103	5439.06	90.82	41.12	20.41	20.45	18.01	38.46
G4	CP1	58,255	5456.50	91.15	45.70	21.45	18.70	14.15	32.85
	CP2	57,942	5462.86	87.89	45.36	21.25	17.08	16.32	33.40
	CP3	57,538	5472.27	93.92	37.80	23.38	20.28	18.53	38.81
G5	CP1	60,448	5487.63	92.95	44.62	23.36	21.39	10.62	32.01
	CP2	60,242	5493.64	91.70	43.93	21.81	21.12	13.15	34.26
	CP3	59,826	5491.92	102.67	45.60	22.05	18.96	13.39	32.35
G6	CP1	60,633	5510.89	97.58	44.47	23.19	19.25	13.09	32.34
	CP2	60,186	5509.04	89.88	44.30	22.00	19.45	14.25	33.70
	CP3	59,746	5518.93	95.93	38.96	23.17	21.51	16.36	37.87
G7	CP1	62,032	5528.45	92.61	41.02	26.55	19.72	12.71	32.43
	CP2	61,040	5532.64	89.50	41.52	25.34	19.32	13.82	33.14
	CP3	60,346	5543.31	99.71	36.80	25.46	19.98	17.76	37.74
G8	CP1	61,172	5542.32	100.76	38.87	28.41	18.53	14.19	32.72
	CP2	60,300	5544.88	96.94	39.51	27.44	17.67	15.38	33.05
	CP3	59,447	5557.61	103.07	33.43	25.35	20.66	20.55	41.21

Table 11: 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
G3	CP1	57,754	6403.96	95.44	41.70	18.93	22.75	16.62	39.37
	CP2	57,156	6415.96	94.03	35.16	21.19	25.16	18.49	43.65
	CP3	55,974	6424.26	95.56	34.69	18.68	24.79	21.83	46.63
G4	CP1	58,554	6455.87	106.94	40.30	17.51	23.91	18.28	42.19
	CP2	58,081	6454.79	98.16	40.10	20.29	23.49	16.12	39.61
	CP3	57,723	6486.35	110.16	27.58	20.28	28.00	24.14	52.14
G5	CP1	60,533	6488.51	111.40	38.72	22.01	19.20	20.07	39.27
	CP2	60,325	6490.07	105.25	38.42	21.91	18.83	20.84	39.67
	CP3	59,917	6499.43	103.83	32.80	22.43	20.83	23.94	44.77
G6	CP1	60,645	6489.62	107.44	50.42	21.65	15.49	12.44	27.93

Grade	Admin	Number Tested	Scale Score Mean	Scale Score SD	% Below Proficiency	% Approaching Proficiency	% At Proficiency	% Above Proficiency	% At or Above Proficiency
	CP2	60,139	6506.51	109.80	44.39	21.64	17.52	16.45	33.97
	CP3	59,690	6513.26	122.87	41.81	20.95	18.32	18.92	37.25
G7	CP1	62,766	6525.15	126.62	44.19	22.05	15.26	18.50	33.77
	CP2	61,655	6526.11	121.86	39.31	25.73	17.78	17.18	34.96
	CP3	60,616	6545.93	125.86	35.59	20.51	17.23	26.67	43.90
G8	CP1	61,015	6533.98	129.41	45.31	23.59	13.77	17.33	31.10
	CP2	60,420	6526.58	126.65	47.41	25.27	12.47	14.85	27.32
	CP3	59,609	6567.02	146.25	35.97	23.36	14.68	26.00	40.68

Figure 3. Percentage of Students in Proficiency Levels, ELA

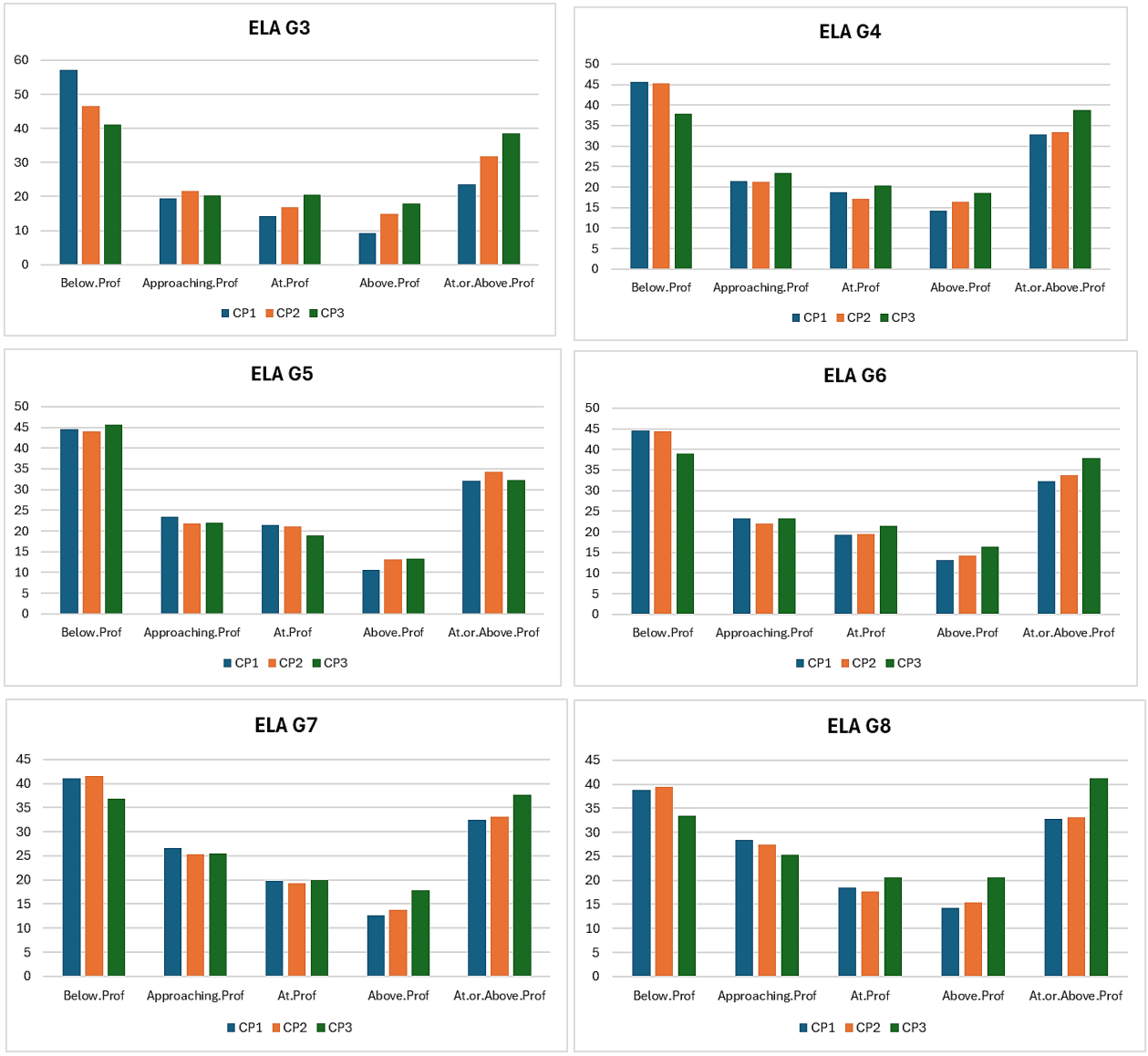
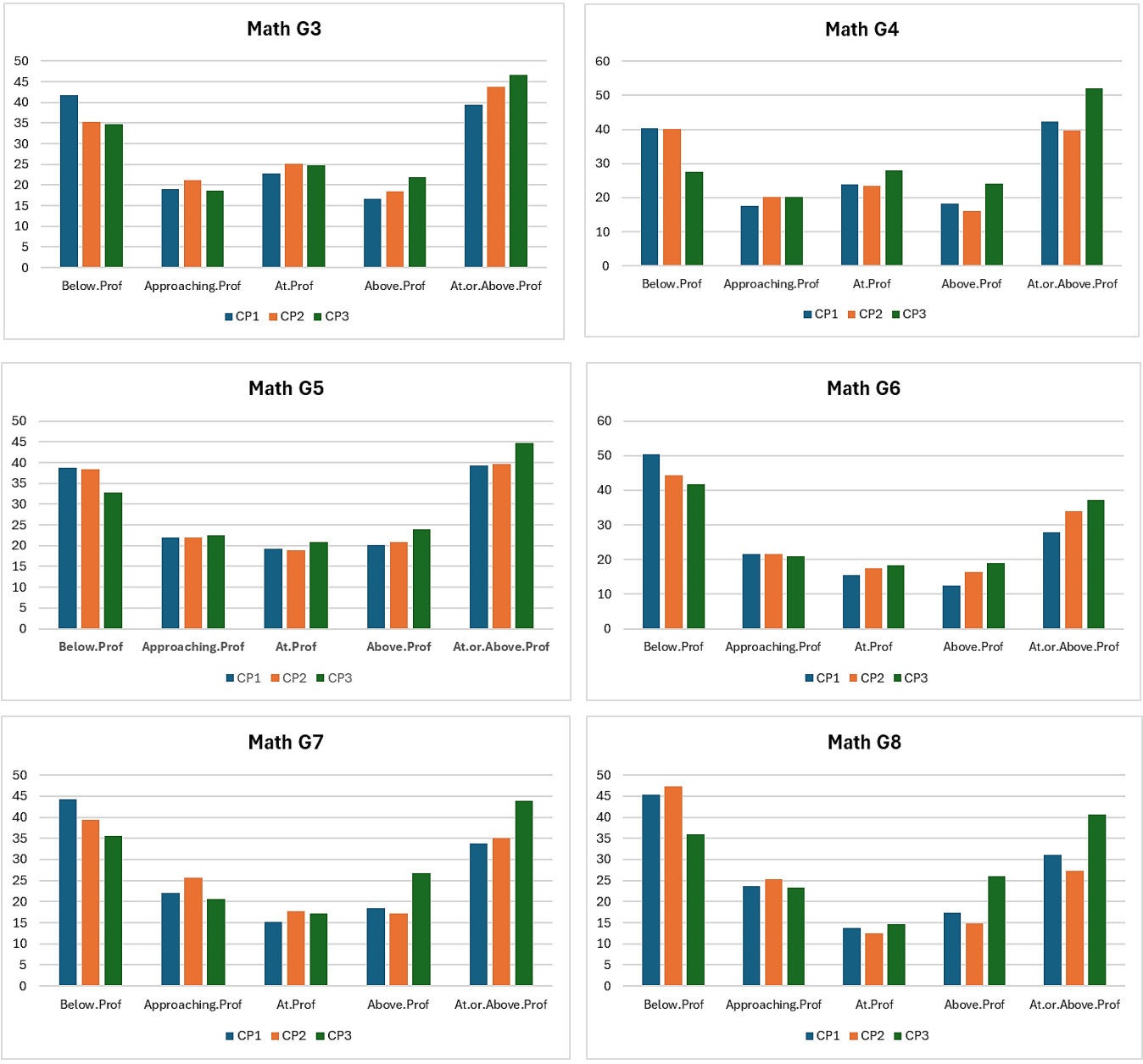


Figure 4. Percentage of Students in Proficiency Levels, Mathematics



3.3 Student Performance by Subgroup

The 2024–2025 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,

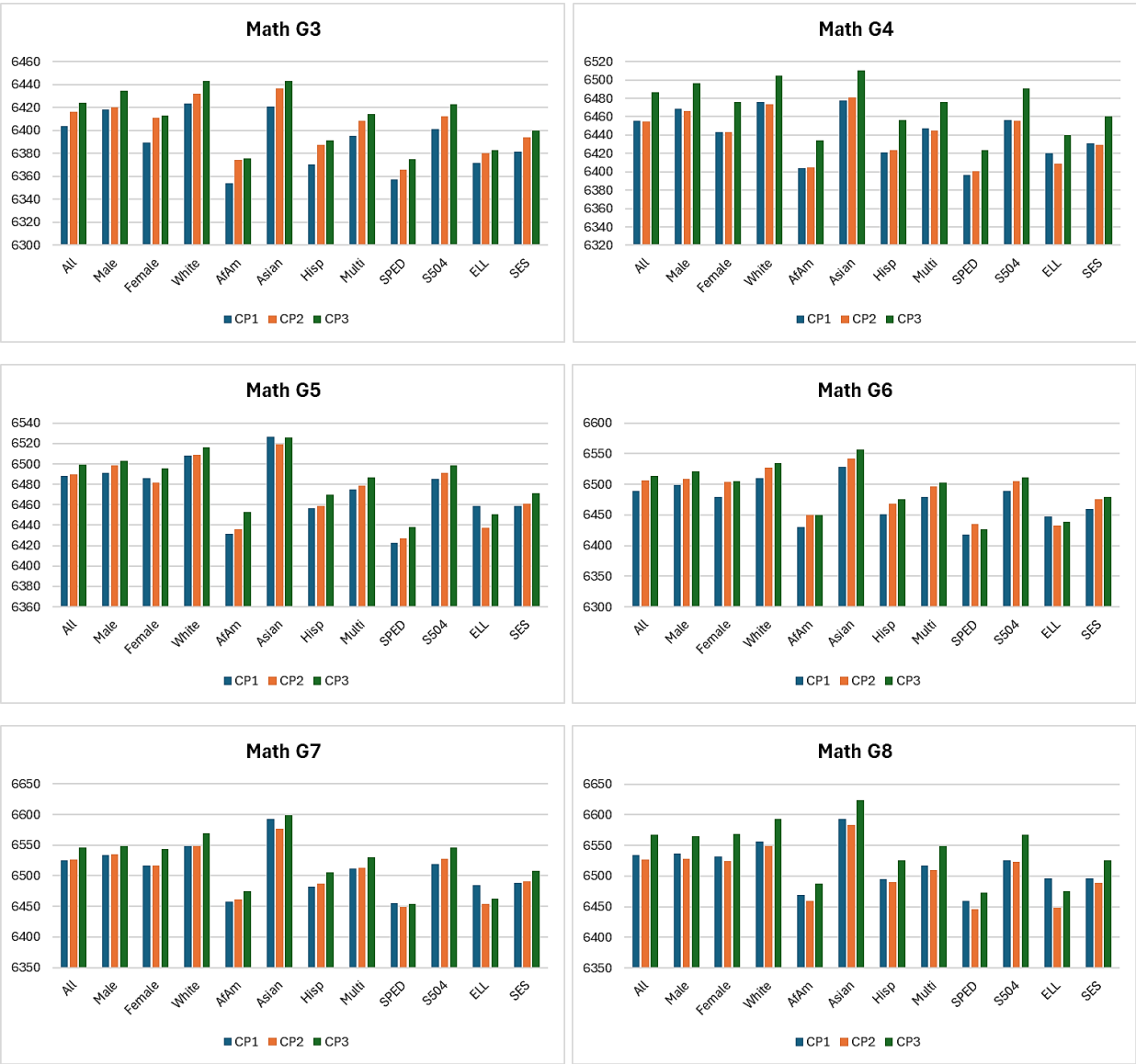
Hispanic/Latino, American Indian/Alaskan, Native Hawaiian/Pacific Islander, Multi-Racial, SPED, Section 504 Plan, EL, and SES.

Distribution of scale scores by student population groups are presented in Appendix 3-A, Distribution of Scale Scores and Standard Deviations. Percentage of students in performance levels for overall and by subgroup 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 are provided in Appendix 3-C, Distribution of Reporting Category Scores by Subgroup. It should be noted that the TYA is using a partial blueprint design (i.e., each Checkpoint assesses a subset of the standards). Therefore, scores across Checkpoints cannot be readily compared.

Figures 5 and 6 display the average scale scores, overall and by subgroup, for the 2024–2025 test administration. Please note that subgroups with the size smaller than 200 were suppressed from the graphs.

Figure 5: ELA Average Scale Score by Subgroup

Figure 6: Mathematics Average Scale Score by Subgroup



3.4 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. 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.

3.4.1 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.

Tables 12 and 13 present the marginal reliability coefficients and the average standard error of measurements (SEMs) for the total scale scores for the 2024–2025 test administration. Results show that marginal reliability is in the high .70s and .80s for all subjects and grades. Within a subject and grade, reliability seems to be consistent across test administrations.

Table 12: Marginal Reliability for ELA

Grade	Admin	Marginal Reliability	*N*	Mean	SD	SEM
3	CP1	0.810	57,674	5410.35	78.23	31.65
	CP2	0.801	57,309	5431.25	87.82	35.51
	CP3	0.802	56,103	5439.06	90.82	36.80
4	CP1	0.800	58,255	5456.50	91.15	35.54
	CP2	0.795	57,942	5462.86	87.89	37.96

Grade	Admin	Marginal Reliability	*N*	Mean	SD	SEM
	CP3	0.770	57,538	5472.27	93.92	41.55
5	CP1	0.780	60,448	5487.63	92.95	38.50
	CP2	0.795	60,242	5493.64	91.70	38.78
	CP3	0.768	59,826	5491.92	102.67	44.64
6	CP1	0.774	60,633	5510.89	97.58	38.53
	CP2	0.787	60,186	5509.04	89.88	39.83
	CP3	0.796	59,746	5518.93	95.93	40.26
7	CP1	0.806	62,032	5528.45	92.61	37.34
	CP2	0.825	61,040	5532.64	89.50	35.52
	CP3	0.796	60,346	5543.31	99.71	42.32
8	CP1	0.788	61,172	5542.32	100.76	38.93
	CP2	0.798	60,300	5544.88	96.94	39.32
	CP3	0.798	59,447	5557.61	103.07	44.46

Table 13: Marginal Reliability for Mathematics

Grade	Admin	Marginal Reliability	*N*	Mean	SD	SEM
3	CP1	0.839	57,754	6403.96	95.44	34.08
	CP2	0.794	57,156	6415.96	94.03	35.75
	CP3	0.809	55,974	6424.26	95.56	34.49
4	CP1	0.817	58,554	6455.87	106.94	36.57
	CP2	0.828	58,081	6454.79	98.16	34.62
	CP3	0.760	57,723	6486.35	110.16	40.38
5	CP1	0.810	60,533	6488.51	111.40	37.97
	CP2	0.819	60,325	6490.07	105.25	36.54
	CP3	0.805	59,917	6499.43	103.83	44.35
6	CP1	0.840	60,645	6489.62	107.44	38.31
	CP2	0.841	60,139	6506.51	109.80	41.50
	CP3	0.820	59,690	6513.26	122.87	42.25
7	CP1	0.852	62,766	6525.15	126.62	44.46
	CP2	0.830	61,655	6526.11	121.86	44.57
	CP3	0.867	60,616	6545.93	125.86	40.15
8	CP1	0.841	61,015	6533.98	129.41	48.36
	CP2	0.833	60,420	6526.58	126.65	48.78
	CP3	0.855	59,609	6567.02	146.25	50.72

3.4.2 Standard Error of Measurement

Within the 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 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.

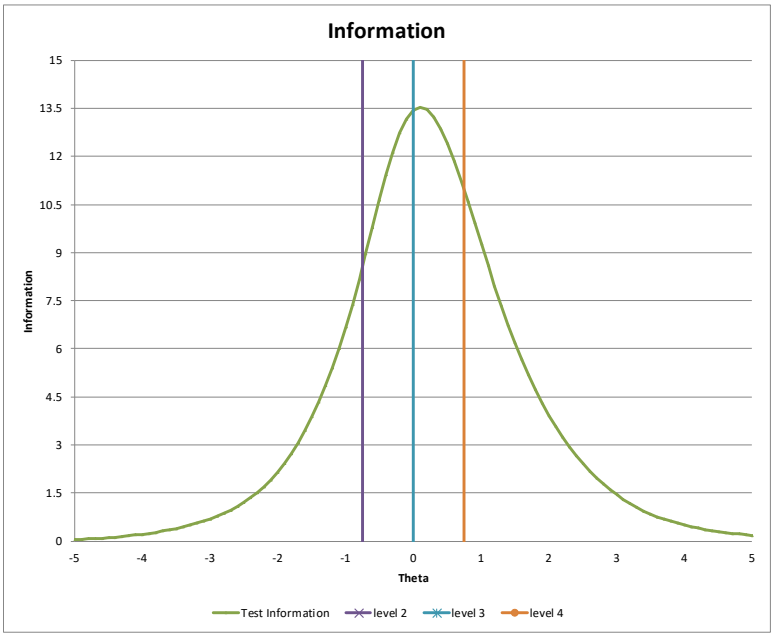
Figure 7 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. Where the curve is lower at the tails indicates that 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 Checkpoint assessments is calculated as:

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

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

Figure 7: Sample TIF



The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

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

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.

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, 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. 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.

Tables 14 and 15 provide 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 both 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.

Table 14: Average SEM by Performance Level, ELA

Grade	Admin	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency	Overall
G3	CP1	27.052	28.889	35.690	59.575	31.649
	CP2	26.846	31.800	38.257	64.687	35.513
	CP3	28.130	31.247	38.774	60.632	36.797
G4	CP1	27.383	28.836	37.759	69.148	35.545
	CP2	34.154	32.725	37.611	55.730	37.961
	CP3	31.323	36.058	44.043	66.624	41.552
G5	CP1	31.530	31.780	40.299	78.962	38.503
	CP2	34.896	31.924	38.225	64.004	38.778
	CP3	38.264	37.200	42.561	81.531	44.636
G6	CP1	29.522	29.874	38.797	84.048	38.527
	CP2	38.502	34.145	38.604	54.416	39.832
	CP3	42.755	32.335	34.699	52.843	40.258
G7	CP1	31.705	31.214	39.743	64.587	37.338
	CP2	31.720	30.721	35.846	55.316	35.524
	CP3	35.679	36.037	42.415	64.968	42.318
G8	CP1	27.237	30.985	42.014	82.828	38.930
	CP2	32.498	32.660	39.363	68.652	39.316
	CP3	41.520	38.343	41.790	59.472	44.460

Table 15: Average SEM by Performance Level, Mathematics

Grade	Admin	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency	Overall
G3	CP1	31.506	26.992	30.215	53.880	34.077
	CP2	29.214	26.939	31.387	64.195	35.746
	CP3	34.601	24.175	24.599	54.365	34.489
G4	CP1	32.560	24.855	29.113	66.410	36.575
	CP2	35.283	24.545	26.501	57.459	34.618
	CP3	30.003	26.446	29.539	76.495	40.376
G5	CP1	31.402	25.442	30.457	71.554	37.968
	CP2	35.237	25.592	27.298	58.816	36.542
	CP3	52.638	40.319	35.196	44.727	44.348
G6	CP1	37.823	30.007	32.425	62.050	38.309
	CP2	43.772	34.012	33.852	53.356	41.498
	CP3	35.315	29.615	34.453	79.123	42.253
G7	CP1	40.406	36.245	38.768	68.638	44.463
	CP2	51.417	31.088	30.761	63.400	44.573
	CP3	47.257	29.572	28.217	46.523	40.154
G8	CP1	54.480	36.525	37.851	56.832	48.362
	CP2	51.039	38.699	39.736	66.304	48.777
	CP3	47.705	37.703	41.406	71.859	50.724

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.

3.4.3 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], 1999, 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 Checkpoint test 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 based on an alternate form, that is, the percentages of students who are consistently classified in the same proficiency levels on two equivalent test administrations.

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, 1979; Livingston & Lewis, 1995).

The classification index can be examined for decision accuracy and consistency. Decision accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made based on the test takers' true scores, if their true scores could somehow be known. Decision consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made based on an alternate, equivalently constructed test form or test administration, 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 $\mathrm{se}(\hat{\theta})$, we can assume that

$\hat{\theta}$ follows a normal distribution with mean of true ability θ and standard deviation of $\mathrm{se}(\hat{\theta})$, that is, $\hat{\theta} \sim N\left(\theta, \mathrm{se}(\hat{\theta})^2\right)$. The probability of the true score at or above the cut score θ_c is estimated as

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

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

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

3.4.4 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 as 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 $\boldsymbol{x} = (x_1, \cdots, x_N)$, can be estimated as

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

where the likelihood function is

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

and $P(x_i|\theta)$ 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:

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

Mathematically, we have

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

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

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

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

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 number 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.

Exhibit A: Classification Accuracy

		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)

3.4.5 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

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

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

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

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

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

The probability of inconsistency is estimated as

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

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

The consistent index is computed as

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

where

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

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

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

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

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

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. The summary of all classification statistics results is presented in the next section.

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)

3.4.6 Classification Accuracy and Consistency Estimates

The analysis of the classification index is performed for test scores in the 2024–2025 test administration. Tables 16 and 17 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 87% to 94% for accuracy, and from 82% to 92% 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.

Table 16: 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	CP1	89.332	91.149	94.288	85.019	87.612	92.054
	CP2	88.270	89.091	92.436	83.715	84.767	89.445
	CP3	90.047	88.544	90.594	85.941	84.112	86.998
4	CP1	91.382	89.760	91.678	87.768	85.818	88.613
	CP2	88.597	89.697	92.138	83.933	85.502	89.039
	CP3	88.645	86.728	89.438	84.087	81.704	85.397
5	CP1	89.581	89.252	92.918	85.338	85.043	90.213
	CP2	89.377	90.029	92.865	84.994	86.001	90.106
	CP3	87.703	88.964	93.235	82.821	84.560	90.487
6	CP1	90.442	89.762	92.195	86.494	85.764	89.406

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
	CP2	88.785	89.183	92.073	84.160	84.868	88.971
	CP3	89.123	89.116	92.307	84.649	84.750	89.213
7	CP1	90.644	89.264	92.023	86.792	85.103	88.986
	CP2	89.734	90.412	92.719	85.523	86.505	89.903
	CP3	89.654	88.376	91.061	85.384	83.836	87.541
8	CP1	91.155	89.379	91.756	87.514	85.300	88.695
	CP2	89.686	89.901	92.641	85.437	85.896	89.709
	CP3	89.530	88.465	90.618	85.254	83.880	86.898

Table 17: 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	CP1	91.129	90.662	93.115	87.456	86.891	90.349
	CP2	90.561	89.551	91.963	86.718	85.337	88.796
	CP3	91.583	91.877	93.377	88.072	88.518	90.714
4	CP1	92.834	92.015	92.775	89.852	88.734	90.020
	CP2	91.424	91.706	94.002	87.896	88.303	91.622
	CP3	91.920	90.032	91.727	88.627	86.041	88.453
5	CP1	92.758	91.891	92.610	89.732	88.586	89.752
	CP2	91.688	92.565	93.241	88.275	89.452	90.561
	CP3	87.835	88.629	91.461	83.151	83.984	87.938
6	CP1	90.697	91.992	94.817	86.899	88.751	92.684
	CP2	89.569	90.914	93.747	85.372	87.177	91.175
	CP3	92.147	91.439	92.837	88.913	87.970	90.039
7	CP1	90.216	91.677	93.882	86.211	88.260	91.376
	CP2	90.728	91.932	94.465	86.935	88.609	92.195
	CP3	92.434	93.507	94.222	89.332	90.821	91.816
8	CP1	90.482	92.597	94.209	86.545	89.527	91.843
	CP2	89.132	92.303	94.628	84.740	89.084	92.406
	CP3	91.783	91.644	92.669	88.419	88.241	89.699

3.4.7 Reliability for Subgroups in the Population

The 2024–2025 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], SPED students, Section 504 students, EL, and SES students) were calculated. The marginal reliability coefficients for subgroups are provided in Appendix 3-F, Marginal Reliability Coefficients for Overall and by Subgroup. As the appendix indicates, reliability is consistent across subgroups, indicating that the Checkpoint 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.

3.4.8 Reporting Category Reliability

The marginal reliability coefficients and the measurement errors are computed for the reporting categories. Tables 18–23 present the marginal reliability coefficients for reporting categories. Some of these reporting categories are measured with very few items, therefore their reliabilities are relatively lower and they are not reported at individual student level. Data from this pilot year have been used to conceptualize the combined subdomain scoring for reporting more reliable reporting category scores. The model is elaborated on in Section 6.3, Combined Domain and Subdomain Scoring.

Table 18: Marginal Reliability Coefficients for ELA Reporting Categories – CP1

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
3	Reading Foundations	10	10	5429.04	121.43	0.627
	Reading and Understanding Fiction	14	14	5408.40	87.70	0.704
4	Reading and Understanding Fiction	17	17	5454.72	97.31	0.778
	Understanding and Using Vocabulary*	3	3	5583.22	240.33	0.633
5	Reading and Understanding Fiction	17	18	5487.85	97.52	0.754
	Understanding and Using Vocabulary*	4	4	5516.33	216.37	0.589
6	Analyzing Literature	15	15	5512.17	110.39	0.744
	Understanding and Using Vocabulary*	5	5	5575.43	203.97	0.641
7	Analyzing Literature	16	16	5527.35	101.13	0.762
	Understanding and Using Vocabulary*	5	5	5565.16	184.51	0.658
8	Analyzing Literature	13	13	5542.75	121.96	0.732
	Understanding and Using Vocabulary	10	10	5572.16	149.77	0.671

* Reporting category scores are not reported for categories with fewer than nine items.

Table 19: Marginal Reliability Coefficients for ELA Reporting Categories – CP2

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
3	Reading Foundations	10	10	5458.75	131.48	0.584
	Reading and Understanding Informational Text and Media	13	13	5427.79	95.53	0.689
4	Reading and Understanding Informational Text and Media	17	17	5462.49	91.33	0.742
	Understanding and Using Vocabulary*	4	4	5496.06	199.28	0.642
5	Reading and Understanding Informational Text and Media	18	18	5493.20	93.99	0.765

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
6	Analyzing Informational Text and Media	19	19	5509.43	90.72	0.782
	Understanding and Using Vocabulary*	1	1	5518.11	369.56	0.741
7	Analyzing Informational Text and Media	18	18	5531.59	94.93	0.781
	Understanding and Using Vocabulary*	5	5	5567.23	159.56	0.490
8	Analyzing Informational Text and Media	13	13	5540.67	108.55	0.696
	Understanding and Using Vocabulary	10	10	5569.44	146.93	0.679

* Reporting category scores are not reported for categories with fewer than nine items.

Table 20: Marginal Reliability Coefficients for ELA Reporting Categories – CP3

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
3	Reading Foundations*	5	5	5477.10	160.07	0.512
	Reading and Understanding Informational Text and Media*	6	6	5460.22	147.02	0.596
	Writing	10	10	5425.43	117.18	0.661
4	Understanding and Using Vocabulary	10	10	5510.43	146.64	0.502
	Writing	10	11	5461.17	117.41	0.648
5	Understanding and Using Vocabulary*	8	8	5515.87	145.55	0.581
	Writing	12	12	5482.77	122.44	0.669
6	Analyzing Informational Text and Media	10	10	5504.44	130.92	0.667
	Writing	10	10	5525.29	123.44	0.679
7	Analyzing Informational Text and Media	10	10	5552.46	127.28	0.658
	Writing	11	11	5541.89	119.53	0.646
8	Analyzing Informational Text and Media	10	10	5565.99	136.65	0.692
	Writing	11	11	5553.96	118.10	0.620

* Reporting category scores are not reported for categories with fewer than nine items.

Table 21: Marginal Reliability Coefficients for Mathematics Reporting Categories – CP1

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
3	Place Value	10	10	6410.00	121.26	0.683
	Addition, Subtraction, and Money	12	12	6403.80	112.62	0.703
4	Place Value	11	11	6478.42	168.84	0.709
	Multiplication	11	11	6445.83	119.61	0.709

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
5	Multiplication and Division of Whole Numbers	11	12	6486.36	139.04	0.737
	Volume	10	11	6507.62	159.01	0.726
6	Integers and Rational Numbers	10	10	6506.27	148.22	0.713
	Operations with Positive Numbers	13	13	6481.55	130.37	0.739
7	Computing with Rational Numbers	16	16	6516.91	131.84	0.787
	Exponents and Roots*	7	7	6539.39	202.18	0.705
8	Real Numbers	14	14	6539.97	146.54	0.745
	Linear Equations and Inequalities	10	10	6513.69	176.77	0.710

* Reporting category scores are not reported for categories with fewer than nine items.

Table 22: Marginal Reliability Coefficients for Mathematics Reporting Categories – CP2

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
3	Understanding Fractions	10	10	6453.95	150.25	0.609
	Multiplication and Division	13	14	6413.80	102.01	0.700
4	Understanding Fractions Multiplication and Division	13	13	6455.96	115.46	0.722
	Measurement	10	10	6455.94	129.25	0.683
5	Understanding Percents, Decimals, and Fractions	11	12	6503.62	135.77	0.722
	Computing with Percents, Decimals and Fractions	10	11	6482.51	145.58	0.727
6	Expressions and Data Analysis	15	15	6496.74	126.33	0.734
	Equations and Inequalities	10	10	6512.11	141.49	0.699
7	Proportional Relationships	14	14	6521.37	135.08	0.756
	Slope	10	10	6516.21	182.58	0.639
8	Functions	9	10	6508.05	148.78	0.653
	Linear Functions	13	15	6535.50	143.85	0.749

* Reporting category scores are not reported for categories with fewer than nine items.

Table 23: Marginal Reliability Coefficients for Mathematics Reporting Categories – CP3

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
3	Understanding Fractions	11	12	6426.94	122.65	0.671
	Measurement and Data	10	11	6424.22	119.24	0.653
4	Understanding Fractions and Decimals	10	11	6499.41	135.83	0.661
	Calculating with Fractions and Decimals	10	11	6502.76	149.42	0.657
	Solving Problems*	1	2	6534.36	274.42	0.660

Grade	Reporting Categories	BP Min	BP Max	Mean	SD	Marginal Reliability
5	Computing with Percents, Decimals, and Fractions	12	12	6500.75	115.58	0.675
	Measurement and Data Analysis	11	11	6486.45	143.45	0.663
6	Working with Rates and Ratios	10	10	6548.33	165.32	0.679
	Solving Problems with Rates, Ratios, and Proportions	12	12	6499.78	146.13	0.752
7	Expressions Equations and Inequalities	15	15	6533.76	131.66	0.778
	Geometry	10	10	6541.79	198.03	0.753
8	Geometry	14	14	6568.34	168.97	0.770
	Functions	10	10	6568.48	177.32	0.701

* Reporting category scores are not reported for categories with fewer than nine items.

3.5 Subscale Intercorrelations

Tables 24 and 25 present the observed correlation matrix of the reporting category scores for each subject area. In ELA, the correlations among the reporting categories ranged from 0.17 to 0.68. In Mathematics, the correlations were between 0.44 and 0.71.

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 reporting category 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. Further, since some of these reporting categories are measured with very few items and their reliability is low, they are not reported at the individual student level.

Table 24: Observed Correlations Among Reporting Category Scores for ELA, Grades 3–8

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
3	CP1	Reading Foundations	10	1						
		Reading and Understanding Fiction	14	0.54	1					
	CP2	Reading Foundations	10	0.60	0.51	1				
		Reading and Understanding Informational Text and Media	13	0.56	0.63	0.56	1			
	CP3	Reading Foundations*	5	0.54	0.44	0.54	0.47	1		

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Reading and Understanding Informational Text and Media*	6	0.47	0.51	0.47	0.53	0.43	1	
		Writing	10	0.57	0.56	0.58	0.58	0.53	0.53	1
4	CP1	Reading and Understanding Fiction	17	1						
		Understanding and Using Vocabulary*	3	0.58	1					
	CP2	Reading and Understanding Informational Text and Media	17	0.67	0.50	1				
		Understanding and Using Vocabulary*	4	0.54	0.45	0.56	1			
	CP3	Understanding and Using Vocabulary	10	0.55	0.45	0.54	0.47	1		
		Writing	10-11	0.58	0.46	0.58	0.49	0.52	1	
5	CP1	Reading and Understanding Fiction	17-18	1						
		Understanding and Using Vocabulary*	4	0.49	1					
	CP2	Reading and Understanding Informational Text and Media	18	0.68	0.47	1				
	CP3	Understanding and Using Vocabulary	8	0.58	0.42	0.62		1		
		Writing	12	0.59	0.41	0.62		0.56	1	
6	CP1	Analyzing Literature	15	1						
		Understanding and Using Vocabulary*	5	0.48	1					
	CP2	Analyzing Informational Text and Media	19	0.65	0.51	1				
		Understanding and Using Vocabulary*	1	0.20	0.17	0.23	1			
	CP3	Writing	10	0.54	0.43	0.60	0.19	1		
		Analyzing Informational Text and Media	10	0.58	0.45	0.64	0.19	0.55	1	
7	CP1	Analyzing Literature	16	1						

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Understanding and Using Vocabulary	5	0.58	1					
	CP2	Analyzing Informational Text and Media	18	0.67	0.56	1				
		Understanding and Using Vocabulary*	5	0.47	0.42	0.50	1			
	CP3	Writing	10	0.58	0.48	0.59	0.42	1		
		Analyzing Informational Text and Media	11	0.60	0.50	0.63	0.45	0.56	1	
8	CP1	Analyzing Literature	13	1						
		Understanding and Using Vocabulary	10	0.57	1					
	CP2	Analyzing Informational Text and Media	13	0.61	0.54	1				
		Understanding and Using Vocabulary	10	0.57	0.52	0.63	1			
	CP3	Writing	10	0.57	0.52	0.58	0.56	1		
		Analyzing Informational Text and Media	11	0.56	0.51	0.59	0.57	0.57	1	

* Reporting category scores are not reported for categories with fewer than nine items.

Table 25: Observed Correlations Among Reporting Category Scores for Mathematics, Grades 3–8

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
3	CP1	Place Value	10	1						
		Addition Subtraction Money	12	0.66	1					
	CP2	Understanding Fractions	10	0.51	0.53	1				
		Multiplication and Division	13-14	0.62	0.65	0.55	1			
	CP3	Understanding Fractions	11-12	0.62	0.62	0.54	0.64	1		
		Measurement and Data	10-11	0.59	0.62	0.53	0.64	0.65	1	
4	CP1	Place Value	11	1						
		Multiplication	11	0.62	1					
	CP2	Understanding Fractions, Multiplication, and Division	13	0.62	0.65	1				
		Measurement	10	0.61	0.63	0.68	1			
	CP3	Understanding Fractions and Decimals	10-11	0.59	0.62	0.66	0.63	1		
		Calculating with Fraction and Decimals	10-11	0.57	0.58	0.62	0.59	0.64	1	
		Solving Problems*	1-2	0.45	0.44	0.47	0.46	0.47	0.45	1
Grade	**Test**									

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
5	CP1	Multiplication and Division of Whole Numbers	11-12	1						
		Volume	10-11	0.60	1					
	CP2	Understanding Percents, Decimals, and Fractions	11-12	0.62	0.60	1				
		Computing with Percents, Decimals, and Fractions	10-11	0.61	0.58	0.65	1			
	CP3	Computing with Percents, Decimals, and Fractions	12	0.58	0.57	0.64	0.61	1		
		Measurement and Data Analysis	11	0.54	0.54	0.57	0.57	0.56	1	
6	CP1	Measurement and Data Analysis	10	1						
		Integers and Rational Numbers	13	0.61	1					
	CP2	Expressions and Data Analysis	15	0.59	0.65	1				
		Equations and Inequalities	10	0.60	0.63	0.65	1			
	CP3	Working with Rates and Ratios	10	0.59	0.64	0.64	0.62	1		

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Solving Problems with Rates, Ratios, and Proportions	12	0.62	0.66	0.66	0.66	0.71	1	
7	CP1	Computing with Rational Numbers	16	1						
		Exponents and Roots*	7	0.62	1					
	CP2	Proportional Relationships	14	0.68	0.56	1				
		Slope	10	0.57	0.49	0.56	1			
	CP3	Expressions, Equations, and Inequalities	15	0.71	0.58	0.70	0.55	1		
		Geometry	10	0.60	0.54	0.60	0.52	0.64	1	
8	CP1	Real Numbers	14	1						
		Linear Equations and Inequalities	10	0.62	1					
	CP2	Functions	9-10	0.60	0.59	1				
		Linear Functions	13-15	0.65	0.65	0.69	1			
	CP3	Geometry	14	0.61	0.60	0.60	0.67	1		
		Functions	10	0.61	0.59	0.62	0.67	0.65	1	

* Reporting category scores are not reported for categories with fewer than nine items.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.

Tables 26 and 27 present disattenuated correlations. Disattenuated correlation is capped if the correlation is greater than 1. These values suggest the validity evidence of the internal structure is supported.

Table 26: Disattenuated Correlations Among Reporting Category Scores for ELA

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
3	CP1	Reading Foundations	10	1						
		Reading and Understanding Fiction	14	0.82	1					
	CP2	Reading Foundations	10	1	0.80	1				
		Reading and Understanding Informational Text and Media	13	0.85	0.91	0.88	1			
	CP3	Reading Foundations*	5	0.96	0.73	1	0.79	1		
		Reading and Understanding Informational Text and Media*	6	0.78	0.78	0.81	0.83	0.78	1	
		Writing	10	0.90	0.82	0.94	0.86	0.91	0.85	1
4	CP1	Reading and Understanding Fiction	17	1						
		Understanding and Using Vocabulary*	3	0.83	1					
	CP2	Reading and Understanding Informational Text and Media	17	0.88	0.74	1				
		Understanding and Using Vocabulary*	4	0.77	0.71	0.82	1			

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
	CP3	Understanding and Using Vocabulary	10	0.89	0.81	0.88	0.83	1		
		Writing	10-11	0.82	0.72	0.84	0.76	0.92	1	
5	CP1	Reading and Understanding Fiction	17-18	1						
		Understanding and Using Vocabulary*	4	0.74	1					
	CP2	Reading and Understanding Informational Text and Media	18	0.90	0.70	1				
	CP3	Understanding and Using Vocabulary	8	0.88	0.72	0.93		1		
		Writing	12	0.84	0.65	0.87		0.91	1	
6	CP1	Analyzing Literature	15	1						
		Understanding and Using Vocabulary*	5	0.70	1					
	CP2	Analyzing Informational Text and Media	19	0.86	0.72	1				
		Understanding and Using Vocabulary*	1	0.27	0.24	0.30	1			
	CP3	Writing	10	0.77	0.67	0.82	0.27	1		

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Analyzing Informational Text and Media	10	0.81	0.68	0.88	0.27	0.82	1	
7	CP1	Analyzing Literature	16	1						
		Understanding and Using Vocabulary	5	0.81	1					
	CP2	Analyzing Informational Text and Media	18	0.87	0.78	1				
		Understanding and Using Vocabulary*	5	0.78	0.75	0.80	1			
	CP3	Writing	10	0.83	0.73	0.82	0.75	1		
		Analyzing Informational Text and Media	11	0.86	0.77	0.89	0.80	0.86	1	
8	CP1	Analyzing Literature	13	1						
		Understanding and Using Vocabulary	10	0.82	1					
	CP2	Analyzing Informational Text and Media	13	0.85	0.79	1				
		Understanding and Using Vocabulary	10	0.81	0.78	0.91	1			
	CP3	Writing	10	0.80	0.77	0.84	0.81	1		

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Analyzing Informational Text and Media	11	0.84	0.79	0.9	0.87	0.87	1	

* Reporting category scores are not reported for categories with fewer than nine items.

Table 27: Disattenuated Correlations Among Reporting Category Scores for Mathematics

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
3	CP1	Place Value	10	1						
		Addition Subtraction Money	12	0.95	1					
	CP2	Understanding Fractions	10	0.80	0.82	1				
		Multiplication and Division	13-14	0.90	0.93	0.85	1			
	CP3	Understanding Fractions	11-12	0.92	0.91	0.85	0.93	1		
		Measurement and Data	10-11	0.89	0.92	0.84	0.94	0.98	1	
4	CP1	Place Value	11	1						
		Multiplication	11	0.87	1					
	CP2	Understanding Fractions,	13	0.87	0.91	1				

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Multiplication, and Division								
		Measurement	10	0.87	0.90	0.96	1			
	CP3	Understanding Fractions and Decimals	10-11	0.87	0.90	0.95	0.93	1		
		Calculating with Fraction and Decimals	10-11	0.84	0.86	0.9	0.89	0.97	1	
		Solving Problems*	1-2	0.67	0.65	0.68	0.69	0.72	0.69	1
5	CP1	Multiplication and Division of Whole Numbers	11-12	1						
		Volume	10-11	0.83	1					
	CP2	Understanding Percents, Decimals, and Fractions	11-12	0.85	0.83	1				
		Computing with Percents, Decimals, and Fractions	10-11	0.83	0.80	0.89	1			
	CP3	Computing with Percents, Decimals, and Fractions	12	0.82	0.82	0.92	0.86	1		
		Measurement Data Analysis	11	0.77	0.78	0.82	0.82	0.83	1	
6	CP1	Integers and Rational Numbers	10	1						

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Operations with Positive Numbers	13	0.84	1					
	CP2	Expressions and Data Analysis	15	0.82	0.88	1				
		Equations and Inequalities	10	0.85	0.88	0.90	1			
	CP3	Working with Rates and Ratios	10	0.85	0.91	0.90	0.91	1		
		Solving Problems with Rates, Ratios, and Proportions	12	0.85	0.88	0.89	0.91	1	1	
7	CP1	Computing with Rational Numbers	16	1						
		Exponents and Roots*	7	0.84	1					
	CP2	Proportional Relationships	14	0.88	0.77	1				
		Slope	10	0.81	0.73	0.80	1			
	CP3	Expressions, Equations, and Inequalities	15	0.91	0.79	0.91	0.78	1		
		Geometry	10	0.78	0.73	0.80	0.75	0.84	1	
8	CP1	Real Numbers	14	1						
		Linear Equations and Inequalities	10	0.85	1					
	CP2	Functions	9-10	0.86	0.87	1				

Grade	Test	Reporting Category	Number of Items	CP1		CP2		CP3		
				Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7
		Linear Functions	13-15	0.87	0.88	0.98	1			
	CP3	Geometry	14	0.81	0.81	0.84	0.88	1		
		Functions	10	0.84	0.84	0.92	0.92	0.89	1	

* Reporting category scores are not reported for categories with fewer than nine items.

4. Item Development and Test Construction

4.1 Test Design and Test Specifications

The Indiana Department of Education (IDOE) sought the participation of Indiana educators in the development of ILEARN through-year blueprint and item specifications. The Through-Year Assessments (TYAs) are designed to measure student achievement of the Indiana Academic Standards (IAS) throughout the course of the school year. 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 TYAs 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 Checkpoint assessment forms were constructed using 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.

4.1.1 English/Language Arts and Mathematics Item Specifications

Cambium Assessment, Inc. (CAI) developed the English/Language Arts (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 students, 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 Development Step 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 to ensure that scoring reflects standards and expectations	Final, focused review on items flagged for differential item functioning	

4.1.1.1 Passage and Item Specifications

IDOE leveraged quality content from third-party item banks for use on ILEARN Checkpoint Assessments. These item banks were accompanied by item specifications

that were used when alignment was confirmed by Indiana educators. The available specifications are described in Table 28.

Table 28: Through-Year Item Specifications

Specification	Developer	Content Areas Included
Indiana Item Specifications	Developed by Indiana for Indiana standards and define custom item development	Mathematics, ELA, Science, Social Studies
Smarter Balanced Item Specifications*	Developed by Smarter Balanced for its Smarter Balanced item bank	Mathematics, ELA

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

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

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, 2022, 2023 and 2024.

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

4.1.1.2 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 Balanced and custom Indiana development.

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

- **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 (MC), multiselect, 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 ASL), 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 29 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 ASL), and Linguistic Complexity.

Table 29: 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 MC Multiselect
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.)	ASL: 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 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. For example, content limits can include acceptable denominators, number of place values for rounding or computation, and acceptable shapes for Geometry standards.
- **Acceptable Response Mechanisms.** This section identifies the various ways in which students may respond to a prompt, such as MC, graphic-response, proposition-response, equation-response, and multiselect 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 MC, can still be used with static images that can be used for all student populations.
- **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.

4.1.2 Target Blueprints

4.1.2.1 Checkpoint 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 test (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 designated as Checkpoint in the public-facing blueprint 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 8 points for each reporting category.

Tables 30 and 31 present the Checkpoint test blueprint requirements specified in the Test Delivery System (TDS) for the 2024–2025 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.

Table 30: Minimum/Maximum Percentages of Test Items by Score Reporting Category for Checkpoint ELA

Reporting Category	Min	Max
Grade 3 ELA Checkpoint 1 (20–25 scored items)		
Reading Foundations	42%	52%
Reading and Understanding Fiction	48%	58%
Grade 3 ELA Checkpoint 2 (20–25 scored items)		
Reading Foundations	42%	52%
Reading and Understanding Informational Text and Media	48%	58%
Grade 3 ELA Checkpoint 3 (20–22 scored items)		
Reading Foundations (Scale Score Only)	24%	29%
Reading and Understanding Informational Text and Media (Scale Score Only)	24%	29%
Writing	45%	50%
Grade 3 ELA Combined Subdomains (60–72 scored items)		
Reading Foundations	35%	47%
Reading and Understanding Fiction	14%	23%
Reading and Understanding Informational Text and Media	21%	33%
Writing	14%	17%
Grade 4 ELA Checkpoint 1 (20–23 scored items)		
Reading and Understanding Fiction	75%	85%
Understanding and Using Vocabulary (Scale Score Only)	14%	25%
Grade 4 ELA Checkpoint 2 (20–23 scored items)		
Reading and Understanding Informational Text and Media	75%	85%
Understanding and Using Vocabulary (Scale Score Only)	15%	25%
Grade 4 ELA Checkpoint 3 (21–25 scored items)		
Understanding and Using Vocabulary	45%	50%

Reporting Category	Min	Max
Writing	50%	55%
Grade 4 ELA Combined Subdomains (61–71 scored items)		
Reading and Understanding Fiction	21%	30%
Understanding and Using Vocabulary	23%	34%
Reading and Understanding Informational Text and Media	21%	30%
Writing	15%	20%
Grade 5 ELA Checkpoint 1 (20–23 scored items)		
Reading and Understanding Fiction	75%	86%
Understanding and Using Vocabulary (Scale Score Only)	14%	25%
Grade 5 ELA Checkpoint 2 (20–23 scored items)		
Reading and Understanding Informational Text and Media	75%	86%
5.CC.7 (Scale Score Only)	14%	25%
Grade 5 ELA Checkpoint 3 (20–25 scored items)		
Understanding and Using Vocabulary	45%	57%
Writing	43%	55%
Grade 5 ELA Combined Subdomains (60–71 scored items)		
Reading and Understanding Fiction	21%	30%
Understanding and Using Vocabulary	18%	30%
Reading and Understanding Informational Text and Media	21%	30%
Writing	14%	20%
Grade 6 ELA Checkpoint 1 (20–24 scored items)		
Analyzing Literature	70%	82%
Understanding and Using Vocabulary (Scale Score Only)	18%	30%
Grade 6 ELA Checkpoint 2 (20–25 scored items)		
Analyzing Informational Text and Media	75%	87%
Understanding and Using Vocabulary (Scale Score Only)	13%	25%
Grade 6 ELA Checkpoint 3 (20–23 scored items)		
Writing	48%	55%
Analyzing Informational Text and Media	45%	52%
Grade 6 ELA Combined Subdomains (60–72 scored items)		
Analyzing Literature	19%	30%
Understanding and Using Vocabulary	10%	18%
Analyzing Informational Text and Media	35%	52%
Writing	14%	20%
Grade 7 ELA Checkpoint 1 (20–24 scored items)		
Analyzing Literature	70%	78%
Understanding and Using Vocabulary (Scale Score Only)	22%	30%
Grade 7 ELA Checkpoint 2 (20–24 scored items)		
Analyzing Informational Text and Media	71%	78%
Understanding and Using Vocabulary (Scale Score Only)	22%	29%
Grade 7 ELA Checkpoint 3 (20–24 scored items)		

Reporting Category	Min	Max
Writing	45%	55%
Analyzing Informational Text and Media	45%	55%
Grade 7 ELA Combined Subdomains (60–72 scored items)		
Analyzing Literature	19%	30%
Understanding and Using Vocabulary	14%	20%
Analyzing Informational Text and Media	35%	50%
Writing	14%	20%
Grade 8 ELA Checkpoint 1 (20–24 scored items)		
Analyzing Literature	48%	58%
Understanding and Using Vocabulary	42%	52%
Grade 8 ELA Checkpoint 2 (21–24 scored items)		
Analyzing Informational Text and Media	52%	58%
Understanding and Using Vocabulary	42%	48%
Grade 8 ELA Checkpoint 3 (20–24 scored items)		
Writing	45%	55%
Analyzing Informational Text and Media	45%	55%
Grade 8 ELA Combined Subdomains (61–72 scored items)		
Analyzing Literature	15%	23%
Understanding and Using Vocabulary	28%	36%
Analyzing Informational Text and Media	29%	43%
Writing	14%	20%

Table 31: Minimum/Maximum Percentages of Test Items by Score Reporting Category for Checkpoint Mathematics

Reporting Category	Min	Max
Grade 3 Mathematics Checkpoint 1 (21–25 scored items)		
Place Value	42%	50%
Addition, Subtraction, and Money	50%	58%
Grade 3 Mathematics Checkpoint 2 (22–25 scored items)		
Understanding Fractions	42%	48%
Multiplication and Division	52%	58%
Grade 3 Mathematics Checkpoint 3 (20–25 scored items)		
Understanding Fractions	45%	57%
Measurement and Data	43%	55%
Grade 3 Mathematics Combined Subdomains (63–75 scored items)		
Place Value	13%	17%
Addition, Subtraction, and Money	15%	22%
Understanding Fractions	27%	38%

Reporting Category	Min	Max
Multiplication and Division	16%	22%
Measurement and Data	13%	19%
Grade 4 Mathematics Checkpoint 1 (21–24 scored items)		
Place Value	43%	50%
Multiplication	50%	57%
Grade 4 Mathematics Checkpoint 2 (22–27 scored items)		
Understanding Fractions, Multiplication, and Division	50%	60%
Measurement	40%	50%
Grade 4 Mathematics Checkpoint 3 (22–27 scored items)		
Understanding Fractions and Decimals	40%	50%
Calculating with Fractions and Decimals	40%	50%
Solving Problems (Scale Score Only)	8%	13%
Grade 4 Mathematics Combined Subdomains (65–78 scored items)		
Place Value	13%	17%
Multiplication	14%	20%
Understanding Fractions, Multiplication, and Division	15%	23%
Measurement	13%	18%
Understanding Fractions and Decimals	13%	18%
Calculating with Fractions and Decimals	13%	18%
Solving Problems	3%	5%
Grade 5 Mathematics Checkpoint 1 (20–25 scored items)		
Multiplication and Division of Whole Numbers	45%	57%
Volume	43%	55%
Grade 5 Mathematics Checkpoint 2 (21–25 scored items)		
Understanding Percents, Decimals, and Fractions	48%	57%
Computing with Percents, Decimals, and Fractions	43%	52%
Grade 5 Mathematics Checkpoint 3 (22–27 scored items)		
Computing with Percents, Decimals, and Fractions	46%	56%
Measurement and Data Analysis	44%	54%
Grade 5 Mathematics Combined Subdomains (63–77 scored items)		
Multiplication and Division of Whole Numbers	13%	21%
Volume	13%	19%
Understanding Percents, Decimals, and Fractions	14%	21%
Computing with Percents, Decimals, and Fractions	27%	41%
Measurement and Data Analysis	14%	21%
Grade 6 Mathematics Checkpoint 1 (22–27 scored items)		
Integers and Rational Numbers	40%	50%
Operations with Positive Numbers	50%	60%
Grade 6 Mathematics Checkpoint 2 (25–25 scored items)		
Expressions and Data Analysis–Non-Calculator	20%	20%
Expressions and Data Analysis–Calculator	40%	40%

Reporting Category	Min	Max
Equations and Inequalities	40%	40%
Grade 6 Mathematics Checkpoint 3 (21–24 scored items)		
Working with Rates and Ratios	43%	50%
Solving Problems with Rates, Ratios, and Proportions	50%	57%
Grade 6 Mathematics Combined Subdomains (68–76 scored items)		
Integers and Rational Numbers	13%	18%
Operations with Positive Numbers	16%	22%
Expressions and Data Analysis–Non-Calculator	6%	7%
Expressions and Data Analysis–Calculator	13%	15%
Equations and Inequalities	13%	15%
Working with Rates and Ratios	13%	16%
Solving Problems with Rates, Ratios, and Proportions	14%	19%
Grade 7 Mathematics Checkpoint 1 (21–25 scored items)		
Computing with Rational Numbers	61%	70%
Exponents and Roots (Scale Score Only)	30%	39%
Grade 7 Mathematics Checkpoint 2 (21–25 scored items)		
Proportional Relationships	52%	60%
Slope	40%	48%
Grade 7 Mathematics Checkpoint 3 (23–25 scored items)		
Expressions, Equations, and Inequalities–Non-Calculator	33%	38%
Expressions, Equations, and Inequalities–Calculator	21%	25%
Geometry	40%	43%
Grade 7 Mathematics Combined Subdomains (65–75 scored items)		
Computing with Rational Numbers	19%	25%
Exponents and Roots	9%	14%
Proportional Relationships	15%	23%
Slope	13%	15%
Expressions, Equations, and Inequalities–Non-Calculator	11%	14%
Expressions, Equations, and Inequalities–Calculator	7%	9%
Geometry	13%	15%
Grade 8 Mathematics Checkpoint 1 (23–26 scored items)		
Real Numbers–Non-Calculator	36%	42%
Real Numbers–Calculator	17%	24%
Linear Equations and Inequalities	38%	43%
Grade 8 Mathematics Checkpoint 2 (22–25 scored items)		
Functions	40%	45%
Linear Functions	55%	60%
Grade 8 Mathematics Checkpoint 3 (22–25 scored items)		
Geometry	55%	60%
Functions	40%	45%
Grade 8 Mathematics Combined Subdomains (67–76 scored items)		

Reporting Category	Min	Max
Real Numbers–Non-Calculator	12%	15%
Real Numbers–Calculator	5%	9%
Linear Equations and Inequalities	13%	15%
Functions	26%	30%
Linear Functions	16%	22%
Geometry	16%	22%

4.1.2.2 ELA Checkpoint Score Reporting Categories

The ELA Checkpoint assessments measure students' understanding of the standards throughout the year in grades 3–8. These assessments measure students' proficiency in ELA knowledge and skills. ILEARN Checkpoint individual student reports describe "at or near" proficient ELA performance in the following reporting categories:

Grade 3

- Reading Foundations
- Reading and Understanding Fiction
- Reading and Understanding Informational Text and Media
- Writing

Grade 4

- Reading and Understanding Fiction
- Understanding and Using Vocabulary
- Reading and Understanding Informational Text and Media
- Writing

Grade 5

- Reading and Understanding Fiction
- Understanding and Using Vocabulary
- Reading and Understanding Informational Text and Media
- Writing

Grade 6

- Analyzing Literature
- Understanding and Using Vocabulary
- Analyzing Informational Text and Media
- Writing

Grade 7

- Analyzing Literature
- Understanding and Using Vocabulary

- Analyzing Informational Text and Media
- Writing

Grade 8

- Analyzing Literature
- Understanding and Using Vocabulary
- Analyzing Informational Text and Media
- Writing

4.1.2.3 Mathematics Checkpoint Score Reporting Categories

The Mathematics Checkpoint assessments measure students' understanding of the standards throughout the year in grades 3–8. These assessments measure students' proficiency in mathematical knowledge and skills. ILEARN Checkpoint individual student reports describe "at or near" proficient mathematical performance in the following reporting categories:

Grade 3

- Place Value
- Addition, Subtraction, and Money
- Understanding Fractions
- Multiplication and Division
- Measurement and Data

Grade 4

- Place Value
- Multiplication
- Understanding Fractions, Multiplication, and Division
- Measurement
- Understanding Fractions and Decimals
- Calculating with Fractions and Decimals
- Solving Problems

Grade 5

- Multiplication and Division of Whole Numbers
- Volume
- Understanding Percents, Decimals, and Fractions
- Computing with Percents, Decimals, and Fractions
- Measurement and Data Analysis

Grade 6

- Integers and Rational Numbers
- Operations with Positive Numbers

- Expressions and Data Analysis–Non-Calculator
- Expressions and Data Analysis–Calculator
- Equations and Inequalities
- Working with Rates and Ratios
- Solving Problems with Rates, Ratios, and Proportions

Grade 7

- Computing with Rational Numbers
- Exponents and Roots
- Proportional Relationships
- Slope
- Expressions, Equations, and Inequalities–Non-Calculator
- Expressions, Equations, and Inequalities–Calculator

Grade 8

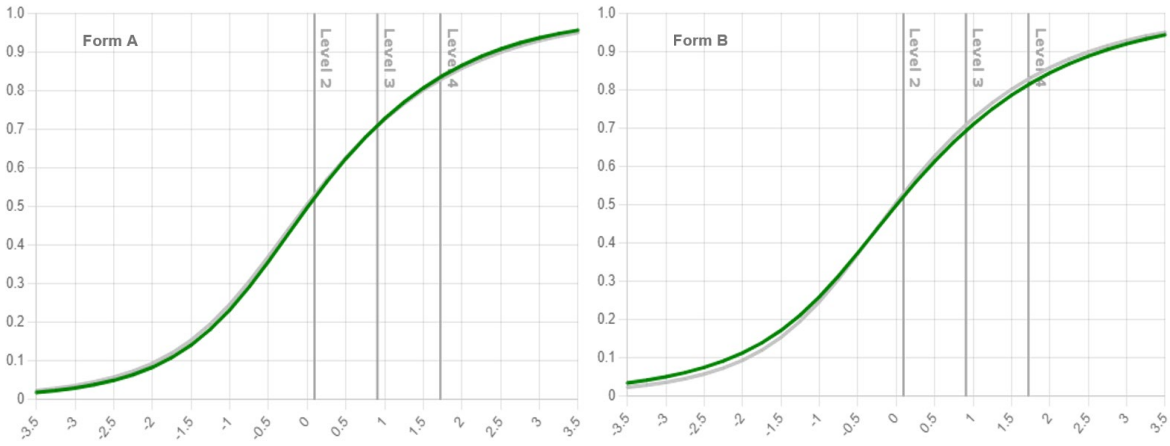
- Real Numbers–Non-Calculator
- Real Numbers–Calculator
- Linear Equations and Inequalities
- Functions
- Linear Functions
- Geometry

4.1.2.4 Test Characteristic Curves

An item characteristic curve (ICC) shows the probability of a correct response as a function of ability, given an item's parameters. Test characteristic curves (TCCs) can be constructed as the sum of ICCs for the items included in 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.

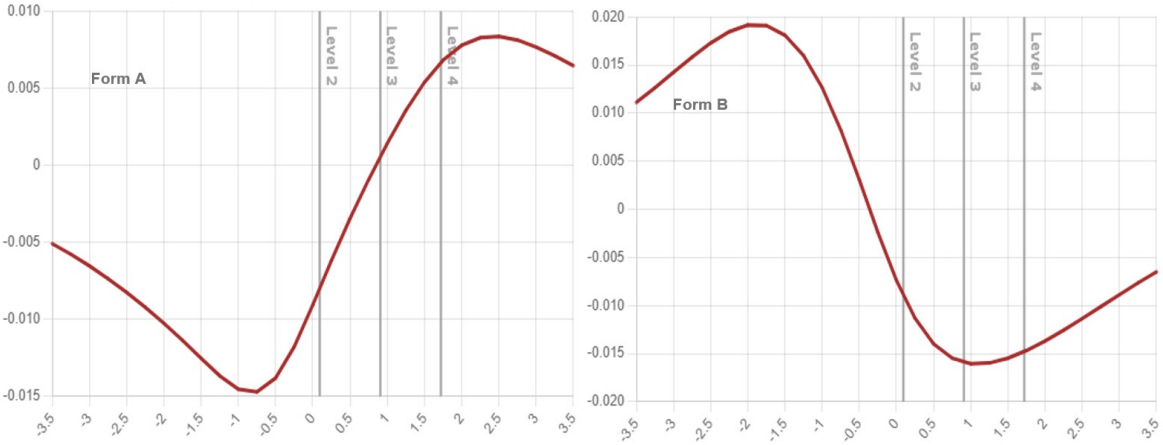
Items were selected for two fixed forms so that the form TCCs matched as closely as possible with a reference form. Figure 8 compares the TCCs for the two ELA grade 7 forms with a reference form. Appendix 4-B provides the TCCs for all grades in ELA and Mathematics.

Figure 8: TCC Comparisons of Grade 7 ELA Form A and Form B



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 9 illustrates a sample TCC difference, which allows us to evaluate the degree to which the parallelism is achieved between each form and the reference form.

Figure 9: TCC Differences of Grade 7 ELA Form A and Form B



4.1.3 Blueprint Match

The blueprints developed for the Checkpoint assessments are provided in Appendices 4-E and 4-F. 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.

4.1.3.1 ELA Blueprints

The ELA blueprint results in an assessment design that delivers some passage-based items and some items not associated with passages.

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. Within a reporting category, the blueprint shows the standards assessed on each Checkpoint assessment.

4.1.3.2 Mathematics Blueprints

Within the blueprints developed for Mathematics, 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 shows the standards assessed on each Checkpoint assessment.

4.2 Item Development Process

Both Smarter Balanced, CAI, and ClearSight developed the ELA and Mathematics item banks using a rigorous, structured process that engaged stakeholders at critical junctures. ClearSight items were not included in the 2024–2025 Checkpoint assessments, but they were included in the 2024–2025 summative field-test administration to be used for future Checkpoint assessments. Similarly, all custom Indiana development followed a very similar review process. This process was managed by CAI's 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 students, with all accessibility features and tools.

4.2.1 Summary of Item Sources

TYAs were designed to measure proficiency on the IAS, meet federal requirements for school accountability testing, and provide information to schools, teachers, families, and students to support teaching and learning.

The IAS were approved by the Indiana State Board of Education in April 2014 for ELA and Mathematics. Minor text updates and clarifications were made to the ELA and Mathematics IAS in 2020. The IAS are intended to implement more rigorous standards

that promote college and career readiness, with the goal of challenging and motivating Indiana students to acquire stronger critical thinking, problem-solving, and communications skills.

TYAs were created using a variety of item types from several sources. Table 32 denotes the sources of the items used in 2024–2025, including licensed item banks (Smarter Balanced Assessment Consortium [Smarter Balanced], ClearSight, and custom Indiana development. Each item source is outlined in more detail in Section 2.

The Smarter Balanced and ClearSight ELA and Mathematics item banks were developed to measure college- and career-readiness standards as embodied in the 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 32: Sources of Items for the ILEARN 2024–2025 Assessments

Subject and Grade(s)	Licensed Bank(s)	Indiana-Owned Items
ELA 3–8	Smarter Balanced ClearSight (not included in Checkpoint test administration)	Yes
Mathematics 3–8	Smarter Balanced ClearSight (not included in Checkpoint test administration)	Yes

4.2.2 Development of New Items

Operational items used on Checkpoint test forms were drawn from a variety of sources, including licensed item banks (Smarter Balanced, 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 ClearSight items (for use on future Checkpoint test administrations) or Indiana custom-developed items have at least a bachelor's degree, and many bring teaching experience. All item writers are trained in

- universal design principles;

- appropriate use of item types;
- ClearSight specifications; and
- Indiana passage and item specifications.

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

- CAI's language accessibility and bias and sensitivity (LABS) guidelines, which include a focus on linguistic complexity
- Indiana item specifications
- A training presentation (using Microsoft PowerPoint) for the appropriate use of item types

4.3 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 8, Quality Assurance Procedures.

4.3.1 Item Review Processes

CAI's test development structure uses 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 Balanced, ClearSight, 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-E, Item Review Checklist. The CAI internal review cycle includes the following phases:

- Preliminary Review
- Content Review One
- Edit Review One
- Senior Content Review

4.3.1.1 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 all items to ensure that the following is true:

- 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 response type.

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

4.3.1.2 Content Review One

Content Review One 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 and Fairness Committee (CFC) 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.

4.3.1.3 Edit Review One

During Edit Review One, 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 ensure 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 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 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 MC 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.

4.3.1.4 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, ensuring that all 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 response type.

4.3.2 Committee Review of Item Pool

All Smarter Balanced, ClearSight, and custom Indiana items have been through an exhaustive external review process. Items in the Smarter and ClearSight item banks were reviewed by content experts in several states, as well as reviewed and approved by multiple stakeholder committees, to evaluate both content and bias/sensitivity. Custom Indiana items were reviewed only by Indiana educators. After items have been developed in the ClearSight 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 ClearSight 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.

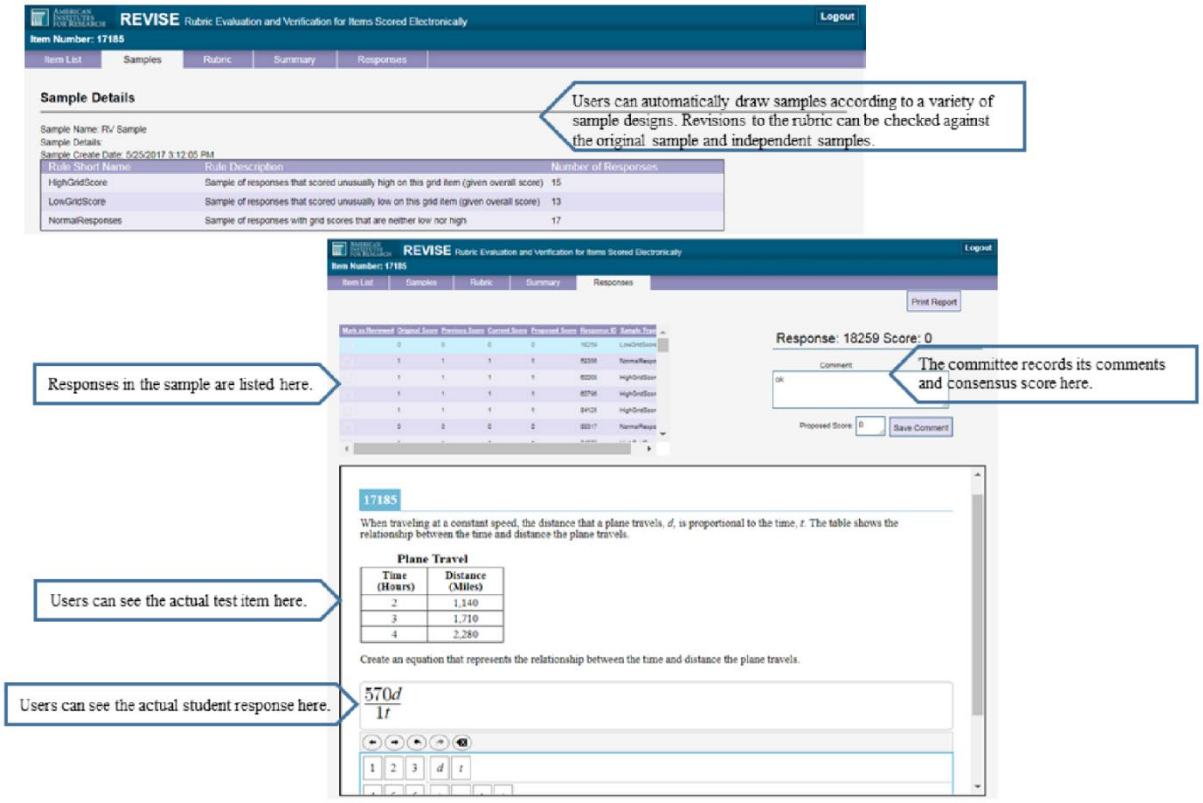
For items that have already been field-tested in other states, wording and scoring edits are not eligible to be made, as such edits risk altering the function of calibrated items. Clients can simply select items from the available item bank to present to the committee.

During the CFC 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 in Chapter 4 of the ILEARN technical report.).

Note that all custom and educator-authored Indiana development was taken to the CFC review. This committee combines the functions of the Content Advisory Committee and the 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 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.



4.4 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 and Mathematics. 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 (differential item functioning [DIF] analyses).

4.4.1 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. In this chapter, classical analyses are reported for operational items.

4.4.1.1 ELA and Mathematics

4.4.1.1.1 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 can 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 MC items and polyserial correlations for constructed-response items). Items are flagged for review if biserial/polyserial values are less than 0.25.

4.4.1.1.2 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 MC 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).

MC items are flagged for review if the *p*-value is less than 0.25 or greater than 0.95. 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 3 points on a constructed-response item score, on average, lower on the total test than students who receive only 2 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 MC or constructed-response items.
- Adjusted biserial correlations for MC item distractors is greater than 0.00.
- Proportion correct value is less than 0.25 or greater than 0.95 for MC 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.

4.4.2 IRT Statistics

4.4.2.1 ELA and Mathematics IRT 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 Checkpoint items are calibrated together with ILEARN items using the 2PL IRT model for MC items and the generalized partial credit model (GPCM) for constructed-response items, scored polytomously.

For MC models, the 2PL model takes the form

$$p_{ij}\left(\theta_j, a_{i,} b_{i,1}, \ldots b_{i,m_i}\right) = \left\{ \begin{array}{l} \dfrac{exp\left(1.7 * a_i\left(\theta_j - b_{i,1}\right)\right)}{1 + exp\left(1.7 * a_i\left(\theta_j - b_{i,1}\right)\right)} = p_{ij}, \;\; if\; z_{ij} \\ = 1\; \dfrac{1}{1 + exp\left(1.7 * a_i\left(\theta_j - b_{i,1}\right)\right)} = 1 - p_{ij}, \;\; if\; z_{ij} = 0 \end{array} \right\},$$

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

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

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

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

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:

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

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 Checkpoint items field-tested in ILEARN are presented in Appendix 4-C of the ILEARN Technical Report, Field-Test Item Parameters.

4.4.3 DIF Analysis

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

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:

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

where $k = \{1, 2, \ldots K\}$ 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

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

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

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

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

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

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

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

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

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

where $\boldsymbol{a}_k$ is a $(T-1)\ X\ 1$ vector of item response scores, corresponding to the T response categories of a polytomous item (excluding one response). $E(\boldsymbol{a}_k)$ and $var(\boldsymbol{a}_k)$, 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 and Mathematics. 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
- Socioeconomic Status (SES)/Not SES (proxy for Free and Reduced-Price Lunch)
- English Learners (ELs)/Not ELs

Table 33 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). The result indicated that across all grades in ELA, there was one field-test item classified as B category and no field-test item classified as C category. Across all grades in Mathematics, there was no field-test item classified as B or C category. Appendix 4-M of the ILEARN Technical Report summarizes and presents more details of the DIF flagging results of the Spring 2025 field-test items.

Table 33: DIF Classification Rules

DIF Category	Flag Criteria
Dichotomous Items	

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

4.5 Item Banks

The Checkpoint item bank is increasing in robustness. It contains licensed items that have been constructed explicitly to support multiple statewide assessment programs. As described previously, all items used on TYAs are aligned to the IAS. The summaries of current Checkpoint item inventories are provided in this section.

The ILEARN ELA and Mathematics operational item banks draw partially from the Smarter Balanced 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 ClearSight item bank and custom Indiana-developed items were also used to ensure complete coverage of the IAS and support a more robust item pool.

Table 34 provides the count of items, by source, used on the 2024–2025 Checkpoint assessments.

Table 34: Operational Item Counts by Source

Subject and Grade	Smarter Balanced Items	Indiana-Owned Items	Total Items
ELA 3	9	107	116
ELA 4	10	91	101
ELA 5	11	86	97
ELA 6	18	84	102
ELA 7	4	115	119
ELA 8	2	115	117
Mathematics 3	42	73	115
Mathematics 4	64	49	113
Mathematics 5	34	80	114

Subject and Grade	Smarter Balanced Items	Indiana-Owned Items	Total Items
Mathematics 6	35	88	123
Mathematics 7	44	73	117
Mathematics 8	38	79	117

4.5.1 Establishing the Banks

Since the TYA bank 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 test administration, two item acceptance review meetings were held. Results of those meetings can be found in Chapter 2 of the 2018–2019 ILEARN 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 Chapter 2 of the 2019–2020 ILEARN Technical Report.

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.

4.5.1.1 Item Bank Composition

Table 35 lists the ELA and Mathematics item types and provides a brief description of each. Examples of various item types can be found in Appendix 4-C, Example Item Types. Tables 36 and 37 list the number of items by type for each grade and subject.

Table 35: TYA Item Types and Descriptions

Response Type*	Description
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.
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.

Response Type*	Description
Multiple Choice (MC)	Student selects one correct answer from four options.
Multiselect (MS)	Student selects all correct answers from several options.
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.

*Response types 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 36: ELA Checkpoint Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
EBSR	0	6	2	11	5	5
HT	3	2	3	4	0	0
MI	4	0	0	0	0	0
MC	93	82	80	73	84	97
MS	16	11	12	14	30	15

Table 37: Mathematics Checkpoint Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
EQ	35	52	43	38	43	26
MI	2	9	5	6	1	3
MC	70	45	60	67	64	72
MS	7	5	6	9	9	16
TI	1	2	0	3	0	0

4.5.2 Bank Maintenance

4.5.2.1 ELA and Mathematics

To maintain the Indiana item banks, new Checkpoint items are developed and field-tested in the Spring administration of the summative ILEARN Assessment of each year, using CAI's field-test engine, and then calibrated and analyzed following the procedures described in Section 4.4.2, IRT 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 EFT 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 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. Tables 38 and 39 present the number of field-test items administered in 2024–2025 by subject, grade, and ownership.

Table 38: Number of Field-Test Items in Summative 2024–2025 for ELA Test Administration

Grade	ClearSight	Smarter Balanced	Indiana-Owned
3	0	34	359
4	0	19	178
5	6	29	209
6	0	22	213
7	0	44	205
8	0	43	270

Table 39: Number of Field-Test Items in Summative 2024–2025 for Mathematics Test Administration

Grade	ClearSight	Smarter Balanced	Indiana-Owned
3	4	0	231
4	4	2	218
5	5	2	201
6	2	3	200
7	3	0	207
8	1	3	218

4.5.3 Braille Item Pools

Across all grades and subject areas, braille forms were provided for the Checkpoint online test forms, as Checkpoints are administered online only. For ELA and Mathematics, enough items from the general education CAT pools were appropriate for the braille pool.

4.5.4 Spanish Item Pools

Spanish tests were provided in the online mode only and were provided only for the Mathematics Checkpoints. Enough items from the general education CAT pools were appropriate for the Spanish pool. Students taking the online Spanish form in Mathematics were exposed to the same item pools used for braille.

5. Test Administration

The State of Indiana implemented a pilot online assessment for interim use beginning with the 2024–2025 school year. This assessment program, referred to as the ILEARN Checkpoint Assessments, comprises English/Language Arts (ELA) and Mathematics assessments for grades 3–8 students. During the 2024–2025 ILEARN Checkpoint test administrations, ELA and Mathematics were offered as fixed-form online assessments. The ELA and Mathematics assessments consist of two opportunities each. It was required for students to complete the first opportunity within the designated window to have results included in the pilot study.

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* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014). This chapter of the ILEARN Checkpoint 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 Checkpoints
- A description of the training and documentation provided to TAs necessary for them to follow the standardized test 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
- A description of the Quality Monitor (QM) subsystem and test irregularity investigation process to detect cheating, monitor item quality in real time, and evaluate test integrity used by Cambium Assessment, Inc. (CAI)

5.1 Testing Options

Administering the 2024–2025 ILEARN Checkpoint Assessments required coordination, detailed specifications, and proper training. In addition, several individuals in each corporation and school were involved in the test administration process, from those setting up secure testing environments to those administering the tests. The Indiana Department of Education (IDOE) worked with CAI to develop and provide the training and documentation necessary for the administration of ILEARN Checkpoints under standardized conditions within all testing environments.

All students were required to take a practice test at their school prior to taking the 2024–2025 ILEARN Checkpoint Assessments. These practice tests contained sample test items similar to the test items that students would encounter on the ILEARN Checkpoint

Assessments to help students become familiar with the item types that would be presented on the online assessments. Indiana students also had the opportunity to interact with released, nonsecure 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 July 2024. A quick guide for the RIR is available to the public (Appendix 5-A).

The ILEARN Checkpoints were administered as a single segment that could be administered over multiple days.

The ILEARN Assessments were untimed, but timing estimates were provided by IDOE to ensure that schools had resources available to create local testing schedules as testing time data was collected from the pilot year. The ILEARN Checkpoint testing windows were as follows:

- Checkpoint 1 Opportunity 1: September 16 – November 15, 2024
- Checkpoint 1 Opportunity 2: September 17, 2024 – April 11, 2025
- Checkpoint 2 Opportunity 1: November 18, 2024 – February 7, 2025
- Checkpoint 2 Opportunity 2: November 19, 2024 – April 11, 2025
- Checkpoint 3 Opportunity 1: February 10 – April 11, 2025
- Checkpoint 3 Opportunity 2: February 11 – April 11, 2025

Corporations and schools who volunteered for the pilot program had students enrolled in tested grade levels participate in the 2024–2025 ILEARN Checkpoint test administration with or without accommodations, as appropriate. Students must take the tests appropriate for the grade level in which they are enrolled. Off-grade testing is not available for ILEARN Checkpoints.

Public and Nonpublic School Students. Students enrolled in accredited Indiana public (including charter schools) and nonpublic schools (including Choice schools) were eligible to participate in the pilot ILEARN Checkpoint Assessment(s).

English Learners (ELs). All ELs enrolled in tested grade levels were eligible to participate in the pilot ILEARN Checkpoint Assessments, including ELA, regardless of how long these students had been enrolled in a U.S. school. Mathematics 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 five 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. In Indiana, a student with an Individualized Education Program (IEP) can participate in ILEARN Checkpoints with the appropriate testing supports and accommodations

prescribed by the IEP. 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.

Remote Proctoring. Students enrolled in virtual schools have the option to be administered the ILEARN Checkpoint pilot remotely. Virtual schools are approved by IDOE prior to being provided with access for remote proctoring, and all students must have documentation signed by a parent or guardian. Both administrators and students have cameras enabled for the entirety of the test administration. Remote proctoring is allowed for ILEARN Checkpoints and their corresponding practice tests only.

5.1.1 Administrative Roles

Corporation Test Coordinators (CTCs), Corporation Information Technology Coordinators (CITCs), Non-Public School Test Coordinators (NPSTCs), School Test Coordinator (STCs), and TAs each had specific roles and responsibilities in the online testing systems. See the *Online 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 systems.

CITCs

CITCs were responsible for ensuring that testing devices were properly configured to support testing and for coordinating participation in a 2024–2025 ILEARN Practice Test. All schools were required to complete a practice test to prepare for online testing and ensure that student testing devices and local school networks are 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 systems.

STCs

Before each test administration, STCs and CTCs were required to verify that student eligibility was correct in the Test Information Distribution Engine (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 ensured that computers and devices were prepared for testing and technical issues were resolved to ensure a smooth testing experience for the students. During the testing 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 TA Certification Course (Appendix 5-G). TAs administered the ILEARN Checkpoint Assessments 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 Checkpoint 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 Checkpoint test administrations.

5.1.2 Online Test Administration

The *Online 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-N) provided information about hardware, software, and network configurations to run CAI's various testing applications.

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

5.1.2.1 Scheduling 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 30 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.

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, the student 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 a local education agency (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 grade 6 Mathematics 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.

5.1.3 Accommodated Test Administration

The ILEARN Checkpoint 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 Checkpoint Assessments. These features allow students to 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 Test Administrator's Manuals (TAMs) (Appendix 5-B), 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 *Test Information Distribution Engine (TIDE) User Guide* (Appendix 5-C).

5.1.4 Allowable Resources for Online Testing

Table 40 provides a list of the designated features and accommodations that were offered in the 2024–2025 test 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 40: Designated Features and Accommodations Available in 2024–2025 for ILEARN

Designated Features	Accommodations
Embedded	
Color contrast (Onscreen) Glossaries (Language) Spanish Masking Mouse Pointer Print Size	American Sign Language (ASL) Audio Transcriptions Calculator Closed Captioning Co:Writer Permissive Mode Print-on-Demand Speech-to-Text Streamline Text-to-Speech (TTS) Except Reading Comprehension

Designated Features	Accommodations
	TTS Including Reading Comprehension Refreshable Braille
Non-Embedded	
Assistive technology to Magnify/Enlarge Access to Sound Amplification Program Special Furniture or Equipment for Viewing Test Special Lighting Conditions Time of Day for Testing Altered Color Acetate Film for Paper Assessments	Braille Transcript for Audio Items Read-Aloud to Self Scribe Speech-to-Text Tested Individual Interpreter for Sign Language Multiplication Table Hundreds Chart Additional Breaks Bilingual Word-to-Word Dictionary Calculator Multiplication Table

**See Appendix 5-E for a complete list of the Read-Aloud Scripts available to students during the 2024–2025 ILEARN Assessments.*

The TA and the 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 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., TTS) were provided the opportunity to participate in practice activities for the statewide assessments with appropriate allowable accommodations. TAs 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.

Grades 3–8 students who participated in ILEARN Checkpoint for ELA could be assigned to either of two TTS modalities:

1. TTS **except** for items and passages measuring reading comprehension
2. 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 *2024–2025 Accessibility and Accommodations Guidance Manual* (Appendix 5-M), as well as in periodic communications with the field.

If an EL or a student with an IEP or Section 504 Plan used any accommodations during the test administration, this information was recorded by the TA in the required test administration information and was captured by CAI in the Database of Record (DOR). CAI included these 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 who were not enrolled in public schools or public school programs were permitted access to accommodations if evidence was provided that the student had been found eligible as a student with a disability as defined by IDEA. Documentation that the requested accommodations had been regularly used for instruction was provided.

The contracted Unified English Braille (UEB) and Nemeth Code for Mathematics accommodations were available for eligible students with IEPs or Section 504 Plans participating in paper-based assessments.

IDOE monitors test administration in corporations and schools to ensure that appropriate assessments, with or without accommodations, are administered to all students with disabilities and ELs and are consistent with Indiana's policies.

5.2 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 Checkpoint Assessments, including the process for giving students access to accommodations. Key personnel involved with ILEARN Checkpoint test administration included CTCs, NPSTCs, 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 *Online Test Delivery System (TDS) User Guide* (Appendix 5-D) was designed to familiarize TAs with TDS and contained tips and screen captures throughout the text. The user guide described the following:

- 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
- Information on CAI's Secure Browser features and keyboard shortcuts

The User Support sections of both the *Online 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.

5.2.1 Manuals and User Guides

The list of webinars and training resources available to corporations and schools for the 2024–2025 ILEARN test administration is provided on the following page. 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.

- **TA Certification Course:** All educators who administered the ILEARN Assessment were required to complete the online TA Certification Course (Appendix 5-G).
- **What is a Computer-Adaptive Test (CAT) Webinar:** This online module described computer-adaptive testing and the student test experience (Appendix 5-H).
- **Why It Is Important to Assess Webinar Module:** This online module illustrated the importance of statewide testing (Appendix 5-I).
- **Remote Proctoring Videos:** These online videos, available in English and Spanish, provided TAs and families with step-by-step instructions on important parts of the remote testing process.

Table 41 presents the list of available user guides and manuals related to ILEARN test 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.

Table 41: User Guides and Manuals

Resource	Description
Online 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, RIR 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-J)	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 TIDE for ILEARN Assessments.
Assistive Technology Manual (Appendix 5-K)	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 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-L)	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 and Accommodations Guidance Manual (Appendix 5-M)	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, ELs, ELs with disabilities, and students without an identified disability or EL 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.
ILEARN Checkpoint Administrator's Manual (CAM)	The ILEARN Checkpoint Administrator's Manual (CAM) provides an overview of the specific roles and responsibilities required before, during, and after testing for ILEARN Checkpoints.
Remote Proctoring User Guide	This user guide provides important information for TAs and assessment personnel on administering assessments remotely.

5.3 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 Checkpoint 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 Checkpoints included the following:

- Procedures to ensure security of test materials
- Procedures to investigate test irregularities
- Guidelines to determine if test invalidation was appropriate or necessary

5.3.1 Student-Level Testing Confidentiality

To support these policies and procedures, IDOE leveraged security measures within CAI systems. For example, students taking the ILEARN Checkpoint 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 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
- A requirement for the corporation or school to administer a breach form

5.3.2 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 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, 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 test 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 test 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 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 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
- Cause achievement of schools to be inaccurately measured or reported

A secure browser was required to access the online ILEARN Checkpoint Assessments. 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, 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 could not print from the CAI Secure Browser unless testing with the print-on-demand accommodation. Print-on-demand allows students to participate in CATs 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 the *Online Test Delivery System (TDS) User Guide* in Appendix 5-D for further details.

5.3.3 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 CAI systems encrypt data at rest and in transit. ILEARN Checkpoint data resides on servers at Amazon Web Services (AWS), CAI's online hosting provider. AWS maintains 24-hour surveillance of both the interior and exterior of its facilities. Staff at both CAI and AWS 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 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 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.

5.3.3.1 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 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, TDS, and the Centralized 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, AWS, 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. AWS 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 that 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.

5.3.3.2 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 CAI systems encrypt data at rest and in transit.

Physical Security

Indiana data reside on servers at AWS, CAI's hosting provider. AWS 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 AWS receive formal training in security procedures to ensure that they know the procedures and implement them properly. CAI and AWS 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 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 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.

5.4 Tracking and Resolving Test Irregularities

Throughout the testing 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., reopen test, reopen 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 Testing 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 testing 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 2025–2026 school year. Table 42 presents examples of test irregularities and test security violations.

Table 42: 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.

Description
Student(s) accessing or using unauthorized electronic equipment (e.g., cell phones, smart watches, iPods, 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 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, 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.

5.5 ClearFrame API

IDOE and CAI will introduce ClearFrame API for use with the ILEARN Through-Year Assessment system beginning in the 2025–2026 school year. This initiative is designed to enhance how assessment data are used across Indiana school corporations by making it more accessible, timely, and actionable for educators and instructional systems.

ClearFrame serves as a secure and streamlined data integration tool that connects ILEARN Assessment results with approved instructional and curriculum platforms. Specifically, it enables school corporations to establish automated data-sharing connections between the state assessment system and participating educational technology vendors. This integration supports the broader goals of the ILEARN program by ensuring that assessment results are not only collected, but also immediately leveraged to support teaching and learning.

Through ClearFrame, student performance data from ILEARN Checkpoints, administered periodically throughout the school year, can be transmitted directly to connected curriculum providers. Once shared through ClearFrame, participating vendors can use these data to generate personalized learning pathways, targeted instructional resources, and timely interventions tailored to each student's needs. Because data are delivered quickly after each Checkpoint, educators can act on it during the same school year, improving instructional responsiveness and student outcomes.

Access to this system will be managed through the Centralized Reporting System (CRS). Within CRS, CTCs and technology directors from all Indiana school corporations will be granted access to a feature called the API Connection Manager. This tool provides a user-friendly interface for managing data-sharing relationships with approved vendor applications.

Once a connection is established, corporations will collaborate with their selected vendors to complete the setup and ensure proper data flow. After configuration, student score data from each Checkpoint assessment will be automatically transmitted to the connected systems as soon as it becomes available in CRS. This eliminates the need for manual data uploads or exports, reducing administrative burden and minimizing delays.

By facilitating these seamless integrations, ClearFrame helps ensure that ILEARN Assessment data are not only informative but also immediately useful in guiding instruction and supporting student success throughout the school year.

6. Scaling and Equating

6.1 Item Response Theory Procedures

6.1.1 Calibration of Checkpoint Item Banks

Checkpoint items are field-tested in and calibrated with the summative ILEARN Assessment. 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 are 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. English/Language Arts (ELA) and Mathematics field-test items are calibrated using IRTPRO software, version 6.0. In each calibration, the parameters of the operational items are fixed to their bank values, and the item parameters of the field-test items, as well as the mean and variance of each group, are estimated.

6.1.2 Estimating Student Ability Using Maximum Likelihood Estimation

6.1.2.1 ELA and Mathematics — Maximum Likelihood Estimation

The Checkpoint 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 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 $log\big(L(\theta)\big)$ 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

$$\frac{\partial lnL(\theta)}{\partial \theta} = \frac{\partial lnL(\theta)^{2PL}}{\partial \theta} + \frac{\partial lnL(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 lnL(\theta)}{\partial^2 \theta} = \frac{\partial^2 lnL(\theta)^{2PL}}{\partial^2 \theta} + \frac{\partial^2 lnL(\theta)^{CR}}{\partial^2 \theta}$$

$$\frac{\partial lnL(\theta)^{2PL}}{\partial \theta} = \sum_{i=1}^{N_{2PL}} Da_i \frac{(z_i - p_i)(p_i)}{p_i}$$

$$\frac{\partial^2 lnL(\theta)^{2PL}}{\partial^2 \theta} = -\sum_{i=1}^{N_{2PL}} D^2 a_i^2 \frac{p_i q_i}{1} \left(1 - \frac{z_i}{p_i^2}\right)$$

$$\frac{\partial lnL(\theta)^{CR}}{\partial \theta} = \sum_{i=1}^{N_{CR}} Da_i \left(z_i - \frac{\sum_{h=1}^{m_i} hexp\left(\sum_{l=1}^{j} Da_i(\theta - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta - b_{il})\right)} \right)$$

$$\frac{\partial^2 lnL(\theta)^{CR}}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{h=1}^{m_i} hexp\left(\sum_{l=1}^{h} Da_i(\theta - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta - b_{il})\right)} \right)^2 - \frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} Da_i(\theta - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} Da_i(\theta - b_{il})\right)} \right),$$

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.

6.1.2.2 Standard Errors of Measurement

The standard error of measurement (SEM) of the marginal maximum likelihood estimation (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 $I(\theta^2) = -\frac{d^2 l(\theta)}{d\theta^2}$ where $\frac{d^2 l(\theta)}{d\theta^2}$ 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.

6.1.3 Calibrating Field-Test Items onto the ILEARN Scale

Checkpoint items are field-tested in the summative ILEARN Assessment. Following the Spring 2019 ILEARN Assessments, item response theory (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 43 presents the number of students used in field-test calibrations for ELA and Mathematics.

Table 43: Number of Students Used in Field-Test Calibrations

ELA		Mathematics	
Grade	Calibration *N* Count	Grade	Calibration *N* Count
3	84055	3	82598
4	81969	4	80338
5	83258	5	81602
6	81672	6	79959
7	82705	7	80934
8	82313	8	80499

6.2 Checkpoint Reporting Scale (Scale Scores)

6.2.1 Overall Performance

For 2024–2025, scale scores were reported for each student who took the Checkpoint 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 44 lists the scaling constants a and b for the Checkpoint 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 44 and are unique to Indiana and the Checkpoint and ILEARN Assessments.

Table 44: Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3–8	75	5500
Mathematics	3–8	75	6500

6.2.1.1 LEXILE® Scores

Checkpoints 1 and 2 report Lexile® range measures with ELA test scores. MetaMetrics provided conversion tables between ELA scale scores and Lexile® measures for each grade and subject.

6.2.2 Reporting Category Performance

In addition to a total scale score, performance on each reporting category is reported. Reporting category theta scores were calculated using MLE, 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.

6.2.2.1 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 the standard error 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 overall 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.

6.2.3 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 theta (LOT) score and lowest observable scale score (LOSS), or the highest observable theta (HOT) score and highest observable scale score (HOSS). 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 45 gives the LOT, LOSS, HOT, and HOSS for the Checkpoint assessments.

Table 45: 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

6.2.4 Rules for Scoring and Reporting of Incomplete Test Administrations

Reporting for each of the subject-area test administrations (ELA and Mathematics) 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.

Checkpoints 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 zero items, the student did not attempt the test. Test scores for these records are not computed or reported.

Undetermined (Attempt = Y). If a student has not attempted enough items to receive an overall score or reporting category scores, their score will be identified as Undetermined.

Attempted (Attempt = Y). If the student responds to at least one item in the test, the test is considered attempted.

6.2.4.1 Online Tests

For Checkpoint tests that are attempted but not complete, both overall scores and subdomain scores will be suppressed and no scores will be reported. Students must complete an entire Checkpoint test to receive a score report.

6.3 Combined Domain and Subdomain Scoring

The Checkpoint Assessments test a subset of the standards in each Checkpoint through domains and subdomains. All these standards are assessed in the end-of-year (EOY) summative ILEARN. However, since the redesigned summative ILEARN Assessment will be shortened to reduce testing time in the 2025–2026 school year, this can affect the reliability of domain score reporting for individual students. Starting in the 2025–2026 school year, the ILEARN Through-Year Assessment system will provide a reliable domain score to show student mastery at the end of the year, and student assessment results across the Checkpoint and summative assessments will be aggregated via a statistically and psychometrically sound approach.

In the 2024–2025 school year, Cambium Assessment, Inc. (CAI) conducted research studies comparing several IRT models for combining scores across Checkpoints and the summative ILEARN. Unidimensional IRT models were compared with longitudinal models (i.e., correlated factors, second-order model, and latent growth factor model) using both simulated and empirical data from a statewide assessment program. Results showed an outperformance of the longitudinal models in producing accurate and reliable aggregate scores.

Another consideration made in selecting the best approach to combine scores was the type of blueprint designed for Checkpoint assessments. A comparison was made between using a full EOY blueprint and a partial blueprint targeting specific content standards (as used in Indiana). Results suggested different construct stability by subject. Although the confirmatory factor analysis results showed both ELA and Mathematics measure the same underlying construct in each test occasion, the nature of the subject matter (being continuous and cumulative versus discrete and topic specific) may interact with the blueprint design to affect construct stability. For more detailed explanations and result comparisons, please refer to CAI's publication website (https://www.cambiumassessment.com/publications.html). Considering Indiana's use of partial blueprints in each Checkpoint, the time-adjusted IRT approach was selected, as it provides the most reliable combined domain and subdomain scores for students.

Starting in the 2025–2026 school year, to provide reliable domain or subdomain-level scores, a student's combined ability on a specific domain (or subdomain) is calculated by applying MLE to their responses to all items within that domain or subdomain from both the Checkpoints and the final summative ILEARN Assessment. This process aggregates a student's performance on a particular set of skills throughout the entire year, using the

time-adjusted item parameters to produce a comprehensive and reliable measurement score for that specific domain or subdomain. Each Checkpoint is statistically linked to the final summative assessment scale by applying a set of predetermined linking constants. These linking constants are calculated and determined by using state-level average ability and standard deviations of Checkpoints and summative assessments. The linking constants ensure that scores from any Checkpoint are time adjusted and comparable to the scores on the measurement scale of the EOY summative assessment.

6.4 Predictive Measure

A second line of research was conducted using the 2024–2025 school year data to compare different models for predicting EOY proficiency using scores from each Checkpoint assessment. Using the pilot year (2024–2025 school year) data, CAI conducted regression analyses, comparing linear regression with the time-adjusted model, to show that prediction accuracy can be enhanced by using cumulative data from Checkpoints. Since prediction accuracy was essentially similar between the two methods, and for system consistency, the time-adjusted model was selected to be used for predicting EOY proficiency starting in the 2025–2026 school year.

Starting in the 2025–2026 school year, to provide a concrete, predictive measure of student achievement, CAI will calculate and report the probability that each student will meet or exceed the EOY proficiency standard. This information will be updated every time a student completes a Checkpoint test opportunity through the school year and will display the prediction based on all Checkpoints completed, regardless of the order of completion. This probabilistic prediction leverages the projected EOY ability scores from the time-adjusted IRT models.

A student's predicted EOY ability score will be derived by applying MLE to their responses to items from all available Checkpoints, using the time-adjusted item parameters. This process synthesizes a student's performance throughout the year to produce a single, forward-looking projection of their likely EOY ability.

7. Reporting and Interpreting Checkpoint Scores

The online Centralized Reporting System (CRS) generates a set of online score reports that includes information describing student performance for students, families, 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. The ILEARN Checkpoint items are all machine-scored, so reports are immediately available upon test completion. 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, CRS also produces aggregate score reports by corporations, 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 score types reported in CRS and a description of the ways to interpret and use these scores in detail.

7.1 Confidentiality of Student Data

There are two reporting systems available for the ILEARN Checkpoints for the 2025–2026 school year. CRS is available for educators, while the Family Portal is designed with the intention of helping parents and guardians understand more about how their child performed on an assessment. Both reporting systems have been designed to help answer how well students performed on the assessments through an online tool that provides educators and stakeholders with timely, relevant score reports.

Both reporting systems for the Checkpoint assessments are 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 systems are also designed to present student performance in a uniform format. For example, the same or 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 systems, 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 corporation, 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, CRS provides two categories of online score reports: (1) aggregate score reports across the state, corporation, and school, and (2) student individual score reports. Table 46 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 *CRS User Guide*, located in a help button on CRS and posted in the Resources section of the ILEARN Portal.

The Family Portal provides parents and guardians access to student test results, both current and historical, starting from the 2025–2026 school year. Content in the Family Portal is designed to be accessible to parents and guardians and is separated by both assessment program and subject area to aid in navigation. For ILEARN Checkpoints, the Family Portal provides (1) broad scoring information for each subject-area Checkpoint and (2) detailed information on category results, item-level results, and suggested resources for parents and guardians to review when considering next steps based on their child's score. Users can also download their child's Detailed Report directly from the Family Portal.

Table 46: Types of Online Score Reports by Aggregation Level

Type of Report	Description
Corporation 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 Overall performance level Average scale scores for student's school and corporation Performance category at the reporting category level

7.2 Family Portal for Parents and Guardians

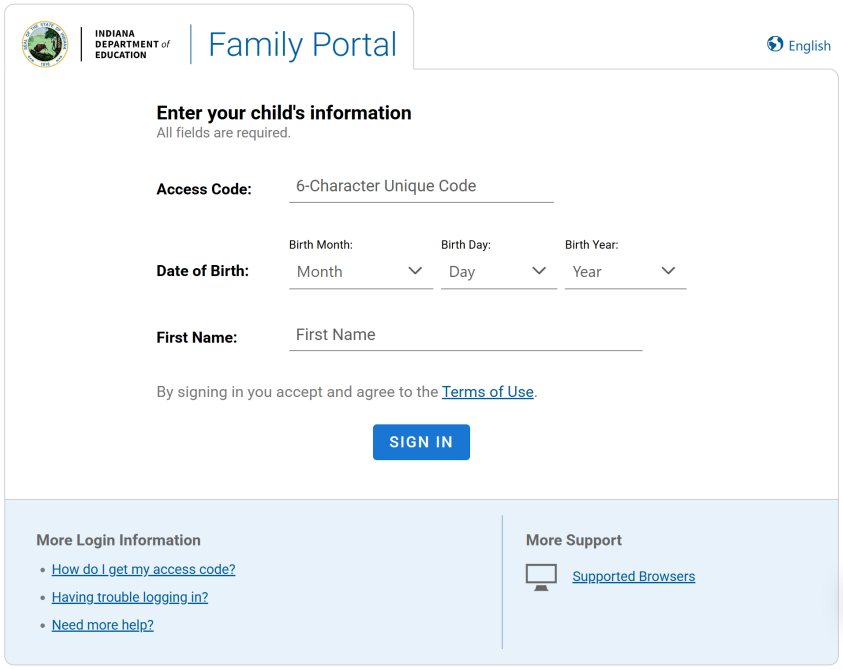
7.2.1 Logging in and the Dashboard

The Family Portal's user base is parents and guardians rather than teachers and assessment personnel. When logging in to the Family Portal, parents and guardians must have three pieces of information:

1. **Their child's unique Family Portal access code**. This code is available in the Test Information Distribution Engine (TIDE) and must be provided to the student's parent or guardian by the school or corporation. This code is considered personally identifiable information (PII) and must be communicated securely.
2. **Their child's date of birth**.
3. **Their child's first name, as entered into TIDE**. The student's first name must match the legal first name on record for the student in TIDE. Nicknames, abbreviations, or alternate spellings will not be accepted by the system.

Figure 10 displays the login page.

Figure 10: Family Portal Login Page



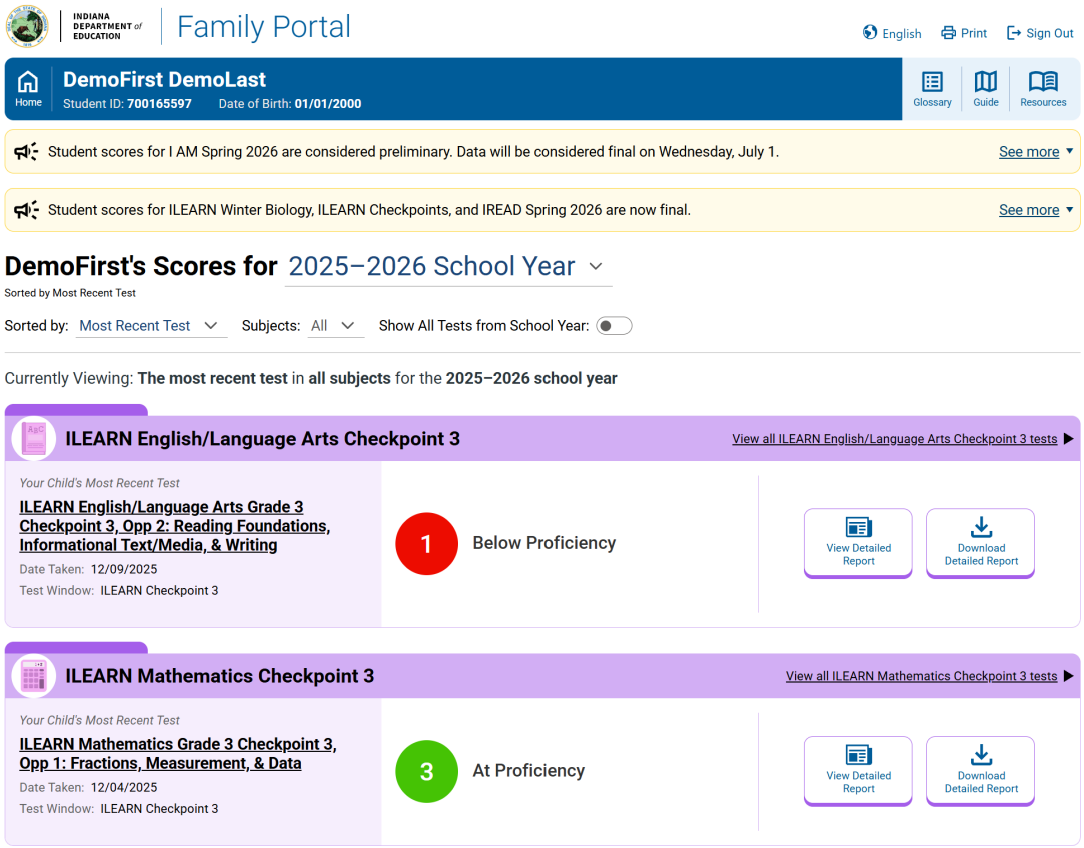
When users log on to the Family Portal, the dashboard page shows overall test results for all tests that the students have taken grouped by test family and subject (e.g., Checkpoint 1 English/Language Arts [ELA]). The dashboard summarizes students' performance by test family for ELA and Mathematics across all grades, including information on (1) the specific Checkpoint (1, 2, or 3) and Opportunity (1 or 2) that was taken, (2) the grades of the students who have tested, (3) the date the test was last taken,

and (4) the proficiency level the student received. Users can also download the student's detailed report directly from the dashboard.

From the dashboard, users have two main options for accessing additional data: they can select the long test name from the left side of the test's information block to view the Detailed Report, or they can select *View all [Test Name] Tests* from the right side of the header to view the Test Overview Page.

Figure 11 presents an example dashboard page.

Figure 11: Dashboard



7.2.2 Detailed Report

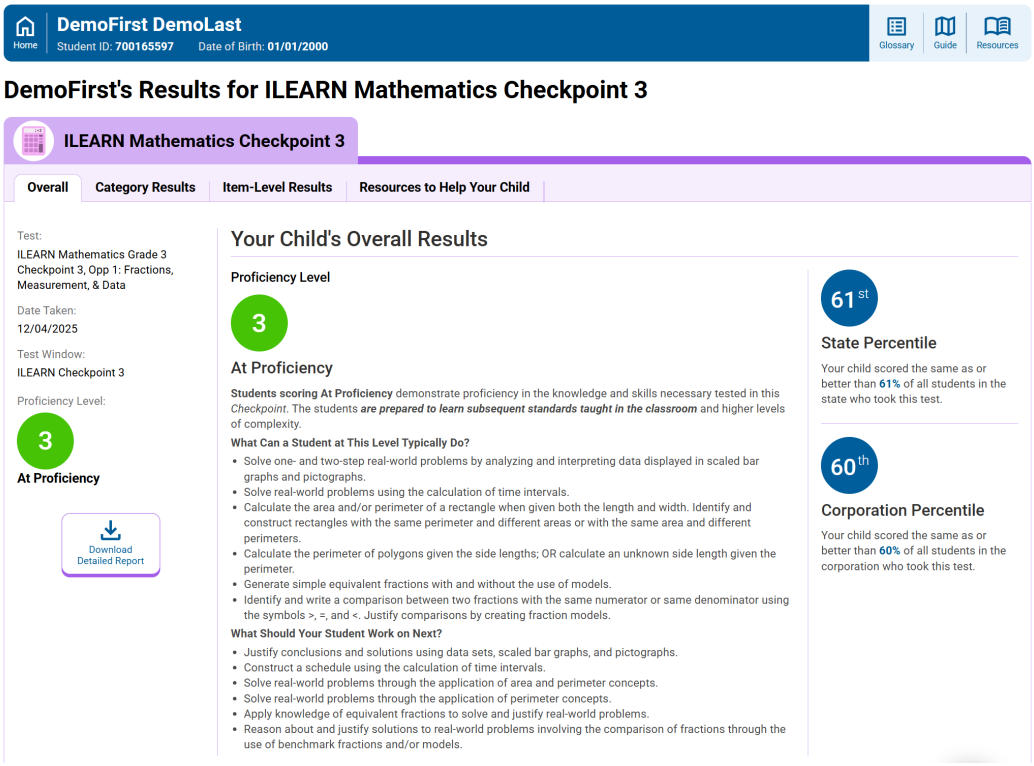
Selecting the test name on the left side of the panel brings users to the detailed report for the selected test. The detailed report loads into the Overall tab, which displays the same information for the test that is visible on the dashboard in addition to (1) the full Performance-Level Descriptor (PLD) for the Proficiency Level achieved by the student, (2) the State Percentile value, and (3) the Corporation Percentile value. Using the tabs at the top of the page, users can view additional information about the student's performance, including (1) how the student performed in each reporting category, (2) how

the student performed on each individual item, and information about each item's reporting category, academic standard, and PLD, and (3) additional resources for parents and guardians to help them understand their child's score and what the child should focus on going forward.

For ILEARN Checkpoints, the information on this page pertains only to the most recent opportunity of a given Checkpoint. For example, if a student has taken both Opportunity 1 and Opportunity 2 for Checkpoint 1 ELA, the detailed report will only display data on Opportunity 2. Data for Opportunity 1 can be found on the Test Overview Page, detailed in the following section.

Figure 12 presents an example detailed report.

Figure 12: Detailed Report



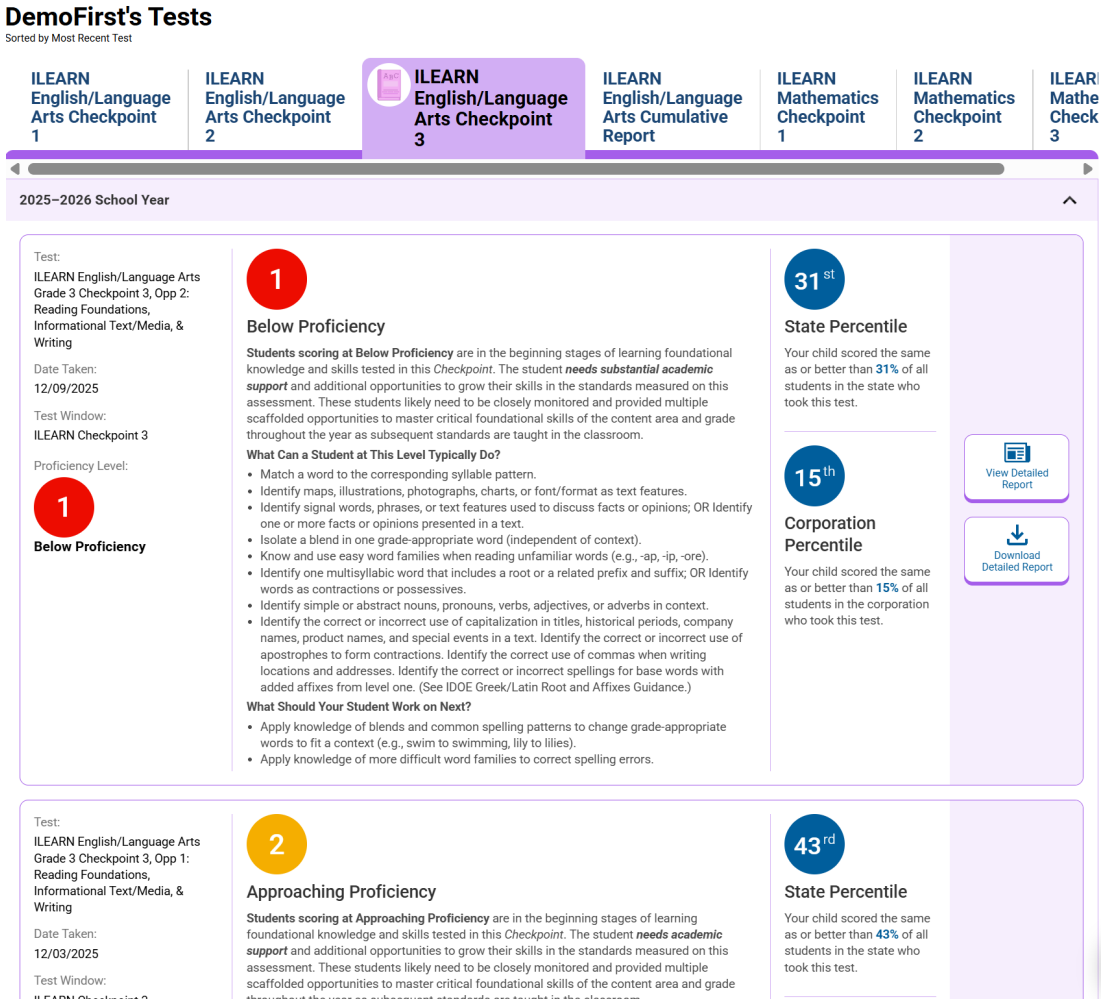
7.2.3 Test Overview Page

Selecting *View all [Test Name] Tests* will take the user to the test overview page. This page loads by default onto the test selected from the dashboard, but users can quickly switch between all of a student's tests using the tabs at the top of the page. The test overview page displays the same information provided on the Overall tab of the detailed

report, and by scrolling down users can view this information for all opportunities of the selected test. To find more information about any tests, users can select *View Detailed Report* from the right side of the page to go to the detailed report for that test.

Figure 13 presents a sample test overview page.

Figure 13: Test Overview Page



7.2.4 Accessibility and Additional Resources

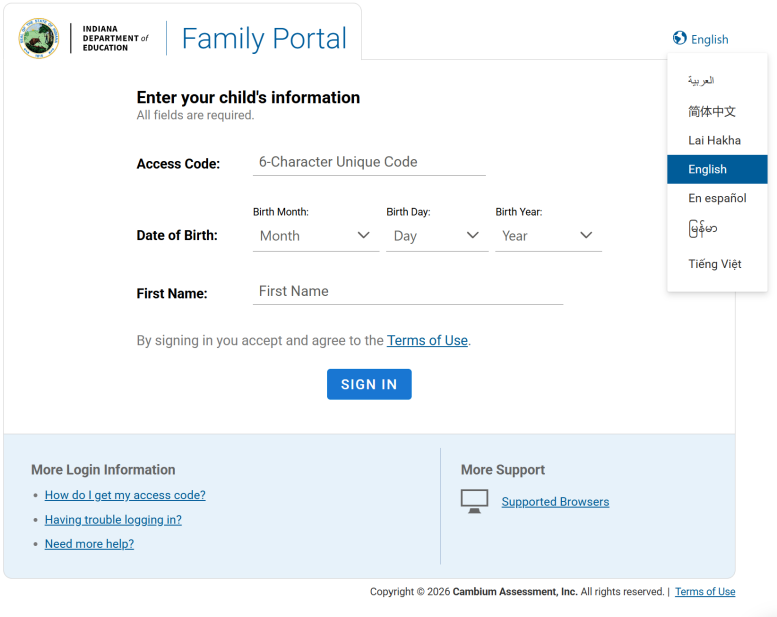
The Family Portal is available in six languages: Arabic, Burmese, Chinese, English, Spanish, and Vietnamese. Users can select the globe icon from the upper-right corner of the sign-in page to change the language, or they can select the globe at any point after signing in to adjust the language settings. The default language on the Family Portal is English.

Most but not all of the content in the Family Portal is available to be translated into the selected language. Some content, such as newer PLDs, is in the process of being

translated to ensure full cross-language accessibility. A limited amount of content, such as the name of testing programs (e.g., ILEARN), is never translated.

Figure 14 presents a sample sign-in page with the language drop-down engaged.

Figure 14: Family Portal Sign-In Page



The Family Portal provides three embedded guides to assist users in navigating the site and understand the data provided: a Glossary, a Guide, and a list of Resources. All three of these guides are pop-up windows that are accessible from icons in the upper-right corner of any page on the Family Portal site.

- **Glossary**: An alphabetized list of terms that may be unfamiliar to parents or guardians while reviewing the Family Portal.
- **Guide**: An example of the page a parent or guardian is currently viewing with numbered icons placed on each page element. Each numbered icon has its own page in the guide, which explains that element and what data it provides. Users can navigate through the guide using the arrows at the bottom of the page or by clicking each numbered icon to go to that element's page directly.
- **Resources**: A hyperlinked list of websites and materials that parents or guardians can review to learn more about assessments, their child's score, and the Family Portal.

7.3 Interpretation of Reported Scores

A student's performance on a Checkpoint test is reported as a scale score and a performance level for the overall test. Scale scores are only available within the CRS user

interface and are not available on the printable individual student report (ISR). There is also a separate performance level for each subdomain. Students' scores and performance levels are summarized at the aggregate levels. This section describes how to interpret these scores.

7.3.1 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.

7.3.2 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). PLDs are a description of content-area knowledge and skills that test takers at each performance level are expected to possess.

7.3.3 Aggregated Score

Students' scale scores are aggregated at the roster, school, and corporation 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.

7.4 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 can 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 in different areas and can be used to improve teaching and student learning. By narrowing the student performance result by subgroup in CRS, 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 broken down by English learner (EL) and observe students in the subgroup category that are struggling with Mathematics. 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 corporations 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.

7.5 ILEARN Checkpoint Reporting Focus Group Synthesis

In November 2025, four virtual focus group sessions were conducted with a total of 41 Indiana educators, including teachers, instructional coaches, principals, and corporation-level administrators. The participant profile included a diverse range of roles, including grades 5–8 ELA and Mathematics teachers, elementary literacy coaches, Mathematics instructional coaches, curriculum specialists, and school administrators. These sessions focused on gathering feedback to streamline data and information displays for educators, ensuring that the data are actionable and organized in a way to support instructional utility. The specific objective of the focus groups was to understand educator needs around data from the ILEARN Checkpoints and evaluate early mock-ups for ILEARN Checkpoint Assessment reports.

Executive Summary

The proposed ILEARN Checkpoint reporting mock-ups received an overwhelmingly positive response from participants, who characterized the new designs as “dreamy” “clear,” and “much more actionable” than the current system. The most critical takeaway is the reports’ potential to serve as a “great time saver,” effectively replacing the manual work teachers currently perform to group students. Educators reacted positively to the shift toward standard-level results that allows them to use assessment data as a “formative piece” to drive daily instruction.

Summary of Key Findings

- Participants characterized the graphically rich displays as “user-friendly,” “organized,” and “easy to understand,” with several participants noting that these reports are exactly what teachers need.
- Participants unanimously emphasized that the new reports would save a “huge amount of time during their instructional planning” by identifying “who needs what, without teachers having to go in and form these groups themselves.” Teachers noted that the new reports show teachers exactly which students need more support and what to work on with them.
- The **Standard-Level Report** was noted for its ease of use, with educators stating, “this is what I’m doing [manually] in my intervention time or in my skill groups daily.”
- The **Student Grouping Report** was highlighted as a critical tool for “reteaching, enrichment, [and] remediation,” with participants noting that it would make the turnaround between testing opportunities “shorter because [teachers] won’t spend so many days looking at the data, before they can actually figure out what to do for the students.”
- Participants also unanimously preferred the new **Navigation** tab structure over the current CRS navigation, stating the layout is “very quick to get the information you’re looking for.”

Overall, participants felt the new reports to be highly actionable tools to help drive small group instruction and planning for intervention groups.

8. Quality Assurance Procedures

Quality assurance (QA) procedures are enforced throughout all stages of Checkpoint test development, administration, 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, CAI notes 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.

8.1 QA in Item Development and Test Construction

Chapter 4 details the item development and test configuration processes. The blueprint describes the content to be covered, the Depth of Knowledge (DOK) with which it will be covered, the item types that will measure the constructs, and every other content-relevant aspect of the test.

The 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 to ensure consistency of test characteristics.

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 test 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). 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.

8.2 QA IN COMPUTER-DELIVERED TEST PRODUCTION

8.2.1 CONTENT PRODUCTION

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 students.
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 CAI ensures 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 the Indiana Department of Education (IDOE) for user acceptance testing (UAT) and final review.

8.2.2 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 students. 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 students, transforming the item representation to the exact representation that will be rendered to students. 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 students.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as students 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. In the following paragraphs, CAI discusses 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 timeline, reducing the work that must be performed during the busy period just before tests go live.

8.2.3 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.

8.2.4 UAT and Final Review

Each release of every CAI system goes through a complete testing cycle, including regression testing. With each release, and every time CAI publishes a test, the system goes through UAT. During UAT, CAI provides clients with login information to an identical (though smaller scale) testing environment to which the system has been deployed. CAI provides 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 test 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 shake-out 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 CAI's 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 QA 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.

8.2.4.1 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 Amazon Web Services (AWS), CAI maintains 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 AWS.

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 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.

8.2.5 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, CAI documents QA 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 CAI's system. They include the following:

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, CAI's servers are monitored at the hardware, operating system, and software platform levels with monitoring software that alerts CAI's 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 CAI 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 are 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 CAI to automatically identify schools or corporations experiencing unusual slowdowns, often before they even notice.

8.3 QA 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. CAI's QA 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.

Once data are prepared for psychometric analyses, the analysis team notifies the psychometric team about the final data being available in the DOR. The psychometric team then extracts the data directly from the DOR and conducts both classical and item response theory (IRT) analyses. The proprietary software, Stats Workshop, is used to automate the analyses, and the resulting outputs are reviewed by both the psychometric and content teams. Final results are uploaded to CAI's ITS.

In the data delivery phase, data are extracted from the DOR and provided to two independent Statistical Analysis System (SAS) programmers. These two programmers are provided with the client-assigned business rules, and they independently prepare data files suitable for delivery. 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 until they are resolved.

CAI's TDS has a real-time quality-monitoring component built in. As students test, data flow through CAI's Quality Monitor (QM) subsystem. The QM subsystem 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 subsystem 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 subsystem also aggregates data to detect problems that become apparent only in the aggregate. For example, the system 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 checks performed for data review, but they are conducted (1) on operational data, and (2) in real time to allow CAI's psychometricians to catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM subsystem to the DOR, which serves as the repository for all test information, and from which all test information for reporting is retrieved. The Data Extract Generator is the tool that is used to retrieve data from the DOR for delivery to IDOE and its QA contractor. CAI psychometricians ensure that data in the extract files match the DOR prior to delivery to IDOE.

8.4 QA in Item Analyses and Equating

Prior to operational work, CAI produces simulated data sets for testing software and analysis procedures. The QA 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 workflow among participants (If necessary, the psychometricians will reconcile differences and correct production or verification programs.)

Following the completion of these activities and the resolution of questions that arise, analysis specifications are finalized.

8.5 QA 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 TDS is tested thoroughly before, during, and after the testing window.

8.5.1 QA 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 Checkpoint 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 Centralized Reporting System (CRS), CAI merges item response data with the demographic information taken either from previous-year assessment data, or if current-year enrollment data are available by the time simulated data files are created, CAI 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 corporations are being handled properly in CRS.

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 testing 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 testing window, a series of QA reports can be generated at any time during the online testing 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 testing 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 QA 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 47 presents an overview of the QA reports.

Table 47: Overview of QA Reports

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology 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

8.5.1.1 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 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.

8.5.2 QA in Reporting

Scores for the Checkpoint 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 subsystem, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the “official” record is stored. Only after scores have passed the QM subsystem checks and are uploaded to the DOR are they passed to CRS, which is responsible for presenting individual-level results and calculating and presenting aggregate results. Absolutely no score is reported in CRS until it passes all validation checks.

Data accuracy and integrity is critical to the success of assessment validity. Additional QA 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 are thoroughly reviewed and accurate.

CAI psychometrics perform item pool simulations 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 performed 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.

IDOE established a production test deck verification process used annually to ensure that 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 CAI’s CRS. This check ensures that student scores are populating through the system accurately, using end-to-end testing.

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.

9. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* https://www.aera.net/Portals/38/1999%20Standards_revised.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

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. RR-91-47). Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1991.tb01414.x

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

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). Lawrence Erlbaum Associates. https://files.eric.ed.gov/fulltext/ED272577.pdf

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics*, *4*(3), 231–246. https://doi.org/10.2307/1164672

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381–389. https://doi.org/10.1007/BF02296131

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21. https://doi.org/10.1177/0146621605275414

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. https://doi.org/10.1002/j.2333-8504.1993.tb01559.x

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. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://hdl.handle.net/11299/115645

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://www.statmodel.com/download/Article_011.pdf

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]. https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. https://doi.org/10.1007/BF02296207

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison* (Research Report # RR-09-37). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02194.x

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. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1972). *A general model for free-response data* [Psychometric Monograph No. 18]. *Psychometrika*, *37*(1, Suppl. 1), 1–68. https://www.psychometricsociety.org/sites/main/files/file-attachments/mn18.pdf?1576606867

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. https://www.academia.edu/17360936/On_the_Reliability_of_Testlet_Based_Tests

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*(2),106–108. https://doi.org/10.1080/00031305.1986.10475369

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/onlinepubs/synth44.pdf

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. https://doi.org/10.1177/0146621604271053

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

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. https://doi.org/10.1007/BF02294531

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1109842.pdf