

**Indiana Learning Evaluation  
Readiness Network  
(*ILEARN*)**

**2019–2020**

**Volume 1  
Annual Technical Report**

## **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](mailto:INassessments@doe.in.gov).

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Stephan Ahadi, Elizabeth Ayers-Wright, Kyra Bilenki, Kevin Clayton, Aleah Pepper, and Elizabeth Xiaoxin Wei. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and Program Leads.

## TABLE OF CONTENTS

1.	INTRODUCTION .....	1
1.1	COVID-19 Considerations .....	1
1.2	Background and Historical Context of Tests.....	1
1.3	Purpose and Intended Uses of the <i>ILEARN</i> Assessments .....	1
1.4	Participants in the Development and Analysis of <i>ILEARN</i> .....	2
1.5	Available Test Formats and Special Versions .....	3
1.6	Student Participation .....	3
2.	SUMMARY OF OPERATIONAL PROCEDURES .....	6
2.1	Administration Procedures .....	6
2.2	Universal Features, Designated Features, and Accommodations.....	6
3.	MAINTENANCE OF THE ITEM BANK.....	8
3.1	Overview of Item Development.....	8
3.2	Review of Operational Items .....	8
3.3	Field Testing.....	8
3.4	Operational Form Construction and Adaptive Simulations .....	9
4.	CLASSICAL ANALYSES OVERVIEW .....	10
4.1	Classical Item Analyses.....	10
4.1.1	<i>Item Discrimination</i> .....	10
4.1.2	<i>Distractor Analysis</i> .....	11
4.1.3	<i>Item Difficulty</i> .....	11
4.1.4	<i>Mean Total Score</i> .....	11
4.2	Differential Item Functioning Analysis.....	11
4.3	Item Analyses Results .....	15
5.	ITEM CALIBRATION .....	16
5.1	Item Response Theory Models.....	16
5.1.1	<i>ELA and Mathematics</i> .....	17
5.1.2	<i>Science</i> .....	17
5.1.3	<i>Social Studies</i> .....	17
5.2	IRT Analyses Results .....	18

5.2.1	IRT Summaries.....	18
5.2.2	2020 ILEARN Test Characteristic Curves .....	18
6.	SCORING AND REPORTING .....	19
6.1	Maximum Likelihood Estimation .....	19
6.1.1	Likelihood Function.....	19
6.1.2	Derivatives.....	19
6.1.3	Standard Errors of Estimates .....	20
6.1.4	Extreme Case Handling.....	21
6.1.5	Standard Errors of LOT/HOT Scores.....	22
6.2	Transforming Theta Scores to Reporting Scale Scores.....	22
6.3	Overall Performance Classification.....	23
6.4	Reporting Category Scores .....	24
6.4.1	MLE/MMLE Scoring.....	24
6.4.2	Strengths and Weaknesses.....	24
6.4.3	Standard Level Aggregate Scores.....	25
6.5	Lexile and Quantile Scores.....	26
6.6	Comparison of Scores to Previous Year.....	26
7.	QUALITY CONTROL PROCEDURES .....	28
7.1	Scoring Quality Check.....	28
8.	REFERENCES .....	29

## LIST OF APPENDICES

Appendix A: Operational Item Statistics

Appendix B: Test Characteristic Curves

Appendix C: Distribution of Scale Scores and Standard Deviations

Appendix D: Distribution of Reporting Category Scores

Appendix E: Operational Item Exposure and Blueprint Match

Appendix F: Simulation Report

## LIST OF TABLES

Table 1: Required Uses and Citations of ILEARN .....	2
Table 2: Number of Students Participating in 2019–2020, Biology.....	5
Table 3: Distribution of Demographic Characteristics of Tested Population, Biology ..	5
Table 4: 2019–2020 Testing Windows .....	6
Table 5: Thresholds for Flagging Items in Classical Item Analysis .....	10
Table 6: DIF Classification Rules.....	14
Table 7: Operational Item p-Value Five-Point Summary and Range, Biology .....	15
Table 8: Operational Item Parameter Five-Point Summary and Range, Biology .....	18
Table 9: ELA Theta and Scaled-Score Limits for Extreme Ability Estimates .....	21
Table 10: Mathematics Theta and Scaled-Score Limits for Extreme Ability Estimates.....	21
Table 11: Science Theta and Scaled-Score Limits for Extreme Ability Estimates .....	22
Table 12: Social Studies Theta and Scaled-Score Limits for Extreme Ability Estimates.....	22
Table 13: Scaling Constants on the Reporting Metric.....	23
Table 14: Proficiency Levels for ELA.....	23
Table 15: Proficiency Levels for Mathematics .....	23
Table 16: Proficiency Levels for Science .....	24
Table 17: Proficiency Levels for Social Studies Grade 5 .....	24
Table 18: Proficiency Levels for Social Studies U.S. Government .....	24
Table 19: Year-to-Year Biology Scale Score Comparisons – All Students .....	27
Table 20: Year-to-Year Biology Scale Score Comparisons – Matched Schools.....	27
Table 21: Year-to-Year Biology Performance Level Comparisons – Matched Schools.....	27

## **1. INTRODUCTION**

The *ILEARN* 2019–2020 technical report is provided to document and make transparent all methods used in item development, test construction, psychometrics, standard setting, score reporting, creating summaries of student assessment results, and supporting evidence for intended uses and interpretations of the test scores. The technical report is presented as five separate, self-contained volumes that cover the following topics:

- (1) *Annual Technical Report*. This annually updated volume provides a general overview of the tests administered to students each year.
- (2) *Test Development*. This volume details the procedures used to construct test forms and summarizes the item bank and its development process.
- (3) *Test Administration*. This volume describes the methods used to administer all available test forms, security protocols, and modifications or accommodations.
- (4) *Evidence of Reliability and Validity*. This volume provides an array of reliability and validity evidence that supports the intended uses and interpretations of the test scores.
- (5) *Score Interpretation Guide*. This volume describes the score types reported along with the appropriate inferences and intended uses of each score type.

IDOE communicates the quality of the *ILEARN* assessments to stakeholders and to the public by producing and providing these technical reports.

### **1.1 COVID-19 CONSIDERATIONS**

The Spring 2020 administration of *ILEARN* was cancelled due to the novel Coronavirus (COVID-19) pandemic. As a result, test summaries in the 2019-2020 technical reports are based on the Fall 2019 and Winter 2020 *ILEARN* Biology End-of-Course Assessments (ECA).

### **1.2 BACKGROUND AND HISTORICAL CONTEXT OF TESTS**

*ILEARN* was constructed to measure student achievement in English/Language Arts (ELA), Mathematics, Science, and Social Studies relative to the Indiana Academic Standards (IAS). *ILEARN* was first administered to students during the 2018-2019 academic year, replacing the Indiana Statewide Testing for Educational Progress-Plus (*ISTEP+*).

### **1.3 PURPOSE AND INTENDED USES OF THE *ILEARN* ASSESSMENTS**

*ILEARN* is Indiana's standards-referenced, summative accountability assessment measuring student achievement and growth. *ILEARN* is comprised of computer-adaptive and performance task test segments aligned to the IAS in English/Language Arts and Mathematics at grades 3 through 8, Science at grades 4 and 6, and Social Studies at grade 5. Additionally, Indiana develops two *ILEARN* ECAs to measure IAS for students completing high school biology and U.S. Government courses, respectively. *ILEARN* is developed with regular and frequent input from Indiana educators to help foster

transparency and ensure student-centeredness and appropriateness of content for Indiana students, using principles of evidence-centered design, and with accessibility for all student populations. *ILEARN* yields overall and reporting-category level test scores at the student level and at other levels of aggregation to reflect degrees of student performance and mastery of the IAS. *ILEARN* supports instruction and student learning by providing immediate feedback to educators and parents based on IAS which can be used to inform instructional strategies, remediate, or enrich curriculum. An array of reporting metrics allows achievement to be monitored at both student and aggregate levels and growth to be measured at both student and group levels over time. While *ILEARN* is designed as a school accountability assessment and *ILEARN* results inform the state's calculations for school accountability, the purpose of this report is to reflect and support validity expectations of *ILEARN* data and reporting.

The *ILEARN* assessments draw items from multiple item banks (see Volume 2) aligned with the IAS and other nationally recognized career and college readiness standards. *ILEARN* content standards are aligned with knowledge and skills that are essential for college and career readiness. CAI and IDOE collaborate to ensure that the items on the test forms constructed for all grades are technically sound and uniquely measure students' mastery of the IAS in ELA, Mathematics, Science, and Social Studies per the published test blueprints.

Table 1 outlines the required uses and citations of *ILEARN* based on the federal Every Student Succeeds Act (ESSA). *ILEARN* fulfills all the requirements described in Table 1.

*Table 1: Required Uses and Citations of ILEARN*

Required Use	Required Use Citation
Indicator of academic achievement and progress	IC 20-32-5.1-2

## 1.4 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF *ILEARN*

IDOE manages the Indiana state assessment program with the assistance of Indiana educators, the Indiana State Board of Education (ISBE), Technical Advisory Committee (TAC), and several vendors (listed below). IDOE fulfills the diverse requirements of implementing *ILEARN* while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014).

### Indiana Department of Education

The Office of Student Assessment oversees all aspects of the *ILEARN* program, including coordination with other IDOE offices, Indiana public schools, and vendors.

### Indiana Educators

Indiana educators participate in most aspects of the conceptualization and development of *ILEARN*. Educators participate in the development of the academic standards,

clarification of how these standards will be assessed, creation of blueprints and test design, and committee reviews of test items and passages.

### **Technical Advisory Committee**

ISBE convenes a panel three times a year to discuss psychometric, test development, administrative, and policy issues relevant to current and future Indiana assessments. This committee is composed of several nationally recognized assessment experts.

### **Cambium Assessment, Inc.**

Cambium Assessment, Inc. (CAI) is the current vendor for assessment testing and was selected through the state-mandated competitive procurement process. In the Winter of 2017, CAI became the primary party responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for *ILEARN* described in this report. Additionally, CAI is responsible for developing and maintaining the *ILEARN* bank, which is used for test construction.

### **Human Resources Research Organization**

For the 2019–2020 *ILEARN* assessment, the Human Resources Research Organization (HumRRO) conducted independent verifications of scoring activities.

## **1.5 AVAILABLE TEST FORMATS AND SPECIAL VERSIONS**

*ILEARN* is an online, adaptive assessment for ELA, Mathematics, and Science and an online, fixed-form assessment for Social Studies. During the 2019-2020 School Year only Fall and Winter Biology assessments were administered. All online adaptive assessments made use of technology-enhanced item types. Students unable to participate in the online administration had the option to use an online accommodated form or a paper-pencil form. Students participating in the computer-based *ILEARN* could use standard online testing features in the test delivery system (TDS), which included 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, there were accommodated forms for braille and Spanish. Students with disabilities could take *ILEARN*, with or without accommodations, or the alternate assessment I AM. Visually impaired students could take the braille version of *ILEARN* ELA, Mathematics, Science, and Social Studies. English Learners (ELs) could take the Spanish language version of *ILEARN* Mathematics, Science, and Social Studies. During test development, CAI ensured that scores obtained on the alternative modes of administrations were comparable to those received on the standard online test adhering to the same blueprints. Post administration checks were also performed, and no concerns were found. The test summary comparison between the standard online form and the alternative mode forms is provided in Volume 2.

## **1.6 STUDENT PARTICIPATION**

All Indiana public school students in ELA and Mathematics grades 3–8, Science grades 4, 6, and end-of-course Biology, Social Studies grade 5, and end-of-course U.S. Government can participate in the state assessments. Table 2 shows the number of



students tested and the number of students reported in the 2019-2020 *ILEARN* Biology ECA. Table 3 presents the distribution of students, in counts and percentages. The subgroup categories reported here are gender, ethnicity, students with special education (SPED), Section 504, and English Learners.

Table 2: Number of Students Participating in 2019–2020, Biology

Admin	Number Tested	Number Reported
Fall 2019	876	870
Winter 2020	1717	1713

Table 3: Distribution of Demographic Characteristics of Tested Population, Biology

Admin	Group	All Students	Male	Female	White	Black / African American	Asian	Hispanic	American Indian / Alaska Native	Native Hawaiian / Other Pacific Islander	Multiracial / Two or More Races	Special Education	Section 504	English Learner
Fall 2019	N	876	439	437	710	51	16	66	2	2	29	137	28	16
	%	100	50.11	49.89	81.05	5.82	1.83	7.53	0.23	0.23	3.31	15.64	3.20	1.83
Winter 2020	N	1717	904	813	1152	251	37	204	3	2	68	254	47	37
	%	100	52.65	47.35	67.09	14.62	2.15	11.88	0.17	0.12	3.96	14.79	2.74	2.15

## 2. SUMMARY OF OPERATIONAL PROCEDURES

### 2.1 ADMINISTRATION PROCEDURES

Table 4 shows the testing window schedule for the 2019–2020 *ILEARN* administrations by assessment.

*Table 4: 2019–2020 Testing Windows*

Assessment	Grade/Subject	Mode	Testing Window
<i>ILEARN</i>	Biology	Online Paper	December 2 – December 19, 2019 (Fall window) December 2 – December 12, 2019
		Online Paper	February 10 – February 27, 2020 (Winter window) February 10 – February 20, 2020

The key personnel involved with *ILEARN* administration included the Corporation Test Coordinators (CTCs), Co-Op role, Non-Public School Test Coordinators (NPSTCs), School Test Coordinators (SCs), Principal (PR), and Test Administrators (TAs) who proctored the test. Test administration manuals were provided so that personnel involved with statewide assessment administrations could maintain both standardized administration conditions and test security.

A secure browser developed by CAI was required to access the online *ILEARN* assessments. The online browser provided a secure environment for student testing by disabling the hot keys, copy, and screen-capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines). During the online assessment, students could pause a test, review previously answered questions, and modify their responses. Responses could only be modified if the test had not been paused for more than 20 minutes (pause rule). Note that the performance task did *not* have a pause rule.

### 2.2 UNIVERSAL FEATURES, DESIGNATED FEATURES, AND ACCOMMODATIONS

Accessibility supports discussed within this document include both embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features that are available to all students as they access instructional or assessment content, designated features that are available to students for whom a need has been identified by an informed educator or team of educators, and accommodations that are generally available for students for whom there is documentation on an Individualized Education Program (IEP), Section 504 Plan, or Individual Language Plan (ILP).

Scores achieved by students using designated features and accommodations are included for federal accountability purposes. All educators making decisions about uses for students are trained on the process and understand the range of designated features and accommodations available.

Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for

students who need them. Embedded accommodations (e.g., text-to-speech) are provided digitally through instructional or assessment technology, while non-embedded accommodations (e.g., scribe) are external to the test delivery system and may be digital or non-digital. Accommodations are available for students for whom there is a documented need on an IEP, Section 504 Plan, or ILP. State-approved accommodations do not compromise the learning expectations, constructs, or grade-level standards. Such accommodations help students with a documented need in an IEP, Section 504 Plan, or ILP generate valid outcomes of the assessments so that they can fully demonstrate what students know and are able to do. From the psychometric point of view, the purpose of providing accommodations is to “increase the validity of inferences about students with disabilities by offsetting specific disability-related, construct-irrelevant impediments to performance” (Koretz & Hamilton, 2006, p. 562).

The test administrators and school test coordinators in Indiana are responsible for ensuring that arrangements for accommodations are made before the test administration dates. The available accommodation options for eligible students include braille, American Sign Language, closed captioning, streamline, assistive technology (e.g., adaptive keyboards, touch screen, switches), calculation device, print-on-demand, multiplication table, and scribe. Detailed descriptions for each of these accommodations can be found in Appendix J of Volume 3.

### **3. MAINTENANCE OF THE ITEM BANK**

The results in this chapter are normally based on the current spring administration. Due to cancellation of the spring 2020 assessments, results that require empirical data are based only on the fall and winter Biology assessments that were administered. Overall test and form summaries that do not depend on empirical data are included as all item bank work was completed prior to the cancellation of the assessments.

#### **3.1 OVERVIEW OF ITEM DEVELOPMENT**

Operational items used on *ILEARN* test forms were drawn from a variety of sources including licensed items banks (Smarter Balanced (Smarter), Independent College and Career Ready (ICCR), and Hawaii EOC) and Indiana custom-developed items. Volume 2 is a separate, stand-alone report containing complete details on the *ILEARN* item banks.

New items are developed each year to be added to the operational item pool after being field tested. Several factors determine 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.

#### **3.2 REVIEW OF OPERATIONAL ITEMS**

During and after each assessment window operational item performance is reviewed based on their performance during the current administration. Flagging criteria are outlined in Table 5 on page 10 and further described in Section 4 below. Flagged items are reviewed by psychometricians and content experts.

#### **3.3 FIELD TESTING**

The *ILEARN* item pool grows each year through the field testing of new items. Any item used on an assessment is field tested before it is used as an operational item. The 2019-2020 *ILEARN* assessments contained newly developed field test items. The EFT slots are randomly positioned for the online adaptive ELA, Mathematics, and Biology assessments and are in fixed positions for the online fixed-form Social Studies assessments. To obtain high-quality responses to the EFT items, students were unaware of which items were operational and which were EFT. For all assessments, field test items were randomly distributed from the pool of available field test items.

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

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

This mimics the spiraling-by-student within a classroom model typically used with paper-pencil forms and ensures broad representation of the items across abilities and

demographic groups. To describe the distribution of forms, consider that  $J$  total forms are available for administration and a total of  $N$  students are participating in the field test. The probability that any one of the  $J$  forms can be assigned to one student is  $1/J$ . Thus, the distribution of forms would follow a uniform distribution with sample sizes per form equal to  $N/J$ .

Thus, field test item exposure rates depend on the number of field test slots and the number of field test items.

### **3.4 OPERATIONAL FORM CONSTRUCTION AND ADAPTIVE SIMULATIONS**

Prior to the operational testing window for adaptive tests, CAI psychometricians employ a simulation approach to configure the adaptive algorithm, seeking to maximize test score precision while meeting blueprint specifications based on the available pool of test items. The simulation report in Appendix F provides more details about the simulation approach and results.

Appendix E contains the operational item exposure rates, as well as the operational blueprint match results for the fall and winter end-of-course Biology assessments. Item exposure rates were calculated over all completed test cases. The location of the item on the form (e.g., first or last) does not matter; the calculation only considers if an operational item was administered on a given test. For the blueprint match analysis only students who completed all parts of the test were included. If a student did not finish the test, the algorithm did not have the opportunity to fully meet blueprint as not enough items were administered. In addition, reset cases were excluded because the algorithm will not administer items or passages that were previously administered, and in some cases a single item or passage was needed to meet blueprint. As can be seen in the appendix, 100% of students that completed tests were administered a set of operational items that met blueprint.

For all other non-adaptive assessments, CAI content and psychometric staff worked with IDOE to build fixed-forms. Volume 2 contains more detailed information about operational test form development.

## 4. CLASSICAL ANALYSES OVERVIEW

### 4.1 CLASSICAL ITEM ANALYSES

IDOE and the CAI psychometricians collectively monitored the behavior of items while test forms were administered in the live environment. This was accomplished using CAI's quality monitoring system, which yielded an item-analysis report on the performance of test items throughout the testing window. During administration of the 2019–2020 *ILEARN*, this system served 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. To examine the performance of test items, this report generated classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation. The report is configurable and was produced to flag only items with statistics falling outside a specified range or to generate reports based on all items in the pool. A minimum sample of 200 responses (Zwick, 2012) per item was applied for classical item analyses. The criteria for flagging and reviewing items is provided in Table 5, and a description of the statistics is provided below.

*Table 5: Thresholds for Flagging Items in Classical Item Analysis*

Analysis Type	Flagging Criteria
Item Discrimination	Adjusted biserial/polyserial correlation statistic is less than .25 for multiple-choice or constructed-response items. *
Distractor Analysis	Adjusted biserial correlation statistic is greater than .00 for multiple-choice item distractors. Proportion of students responding to a distractor exceeds the proportion responding to a keyed response for multiple-choice items.
Item Difficulty (MC items)	Proportion correct value is less than .25 or greater than .95 for multiple-choice items.
Item Difficulty (non-MC items)	Proportion of students receiving any single score point is greater than .95 for constructed-response items.
Inverted Mean Total Score	Mean total score for a lower score point exceeds the mean total score for a higher score point for multi-point constructed-response items.

\* IDOE reviewed any item with an adjusted biserial/polyserial correlation less than 0.10. CAI shared these items with IDOE to make final determinations.

#### 4.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those examinees who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for multiple-choice items was calculated as the correlation between the item score and the ability estimate for students. Biserial correlations for operational items can be found in Appendix A. Most of the operational items had a higher biserial correlation than the flagging criteria. Items with low biserial correlations were reviewed by CAI content experts, and all items behaved as expected.

### 4.1.2 Distractor Analysis

Distractor analysis for multiple-choice items was used to identify items that may have had marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracted high-scoring students. For MC items, the correct response should have been the most frequently selected option by high-scoring students. The discrimination value of the correct response should have been substantial and positive, and the discrimination values for distractors should have been lower and, generally, negative. Most of the operational items had a negative distractor. CAI content experts reviewed items with positive distractor correlations and did not find any issue.

### 4.1.3 Item Difficulty

Items that were either extremely difficult or extremely easy were flagged for review but were not necessarily removed if they were grade-level appropriate and aligned with the test specifications. For MC items, the proportion of students in the sample selecting the correct answer (the  $p$ -value) was computed in addition to the proportion of students selecting incorrect responses. For constructed-response items, item difficulty was calculated using the item's relative mean score and the average proportion correct (analogous to  $p$ -value and indicating the ratio of the item's mean score divided by the maximum possible score points). Conventional item  $p$ -values are summarized in Section 4.3. The  $p$ -values for operational items can be found in Appendix A. Most of the operational items had  $p$ -values within the expected range. Flagged items were verified by CAI content experts and psychometricians reported that all items behaved as expected.

### 4.1.4 Mean Total Score

For multi-point constructed-response items, mean total score was calculated using the item's relative mean score and the average proportion correct (analogous to  $p$ -value and indicating the ratio of the item's mean score divided by the maximum possible score points). Items were flagged when the proportion of students in any score point category was greater than 0.95. In addition, constructed-response items were flagged if the average ability estimate of students in a score-point category was lower than the average ability estimate of students in the next lower score-point category. For example, if students who received three points on a constructed-response item score lower, on average, on the total test than students who received only two points on the item, the item will be flagged for review. The  $p$ -values for operational items can be found in Appendix A. Flagged items were verified by CAI content experts and psychometricians reported that all of them behaved as expected.

## 4.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

Note that DIF summaries are only provided when field test analyses occur. No items were field tested during School Year 2019-2020 and thus no DIF summaries appear in this year's technical report.



The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014) provides a guideline for when sample sizes permitting subgroup differences in performance should be examined and appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors.

Differential item functioning (DIF) analysis was conducted for all items to detect potential item bias across major and special population groups, including gender and ethnicity. A minimum sample of 200 responses (Zwick, 2012) per item in each subgroup was applied for DIF analyses. Because of the limited number of students in some groups, DIF analyses were performed for the following groups:

- Male/Female;
- White/African-American;
- White/Hispanic;
- White/Asian;
- White/Native American;
- Text-to-Speech (TTS)/Not TTS;
- Student with Special Education (SPED)/Not SPED;
- Title 1/Not Title 1 (proxy for Free and Reduced-Price Lunch); and
- English Learners (ELs)/Not ELs.

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF was important because it provided a statistical indicator that an item may contain cultural or other bias. DIF-flagged items were further examined by content experts, who were asked to re-examine each flagged item to decide whether the item should have been excluded from the pool due to bias. Not all items that exhibit DIF are biased; characteristics of the education system may also lead to DIF. For example, if schools in certain areas were less likely to offer rigorous Mathematics classes, students at those schools might perform more poorly on Mathematics items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias, but the instruction. However, DIF can indicate bias, so all items were evaluated for DIF.

A generalized Mantel-Haenszel (MH) procedure was applied to calculate DIF. The generalizations include (1) adaptation to polytomous items and (2) 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) 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  $k = \{1, 2, \dots, K\}$  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 ( $\Delta_{MH}$ , Holland & Thayer, 1988) is then defined as

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

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

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

where  $\mathbf{a}_k$  is a  $(T - 1) \times 1$  vector of item response scores, corresponding to the  $T$  response categories of a polytomous item (excluding one response).  $E(\mathbf{a}_k)$  and  $var(\mathbf{a}_k)$ , 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

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

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  $k$ .

Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. DIF classification rules are illustrated in Table 6. Items were also indicated as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American, Hispanic, or female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., White or male). If the DIF statistics fell into the “C” category for any group, the item showed significant DIF and was reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts reviewed all items flagged based on DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance.

Table 6: DIF Classification Rules

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

### 4.3 ITEM ANALYSES RESULTS

This section presents a summary of results from the classical item analysis for the 2019-2020 *ILEARN* operational items. The summaries here are aggregates; item-specific details are found in Appendix A.

Table 7 provides summaries of the p-values by percentile and range by grade and subject for operational items. Note that the “Total OP Items” column shows the number of operational items that were used in the computation of the percentiles. The two-dimension scores for writing items are counted as two items in ELA. Indiana students’ performance indicates the desired variability across the scale in all grades and subjects. The variability informs us that the constructed operational forms had a good discrimination for Indiana students.

*Table 7: Operational Item p-Value Five-Point Summary and Range, Biology*

Grade	Total OP Items*	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Fall 2019	261	0.00	0.30	0.40	0.45	0.51	0.62	0.91
Winter 2020	263	0.00	0.28	0.41	0.47	0.52	0.62	0.90

\*While the item pool was identical for the two administrations, due to small sample sizes two items were not administered during the Fall 2019 administration.

## 5. ITEM CALIBRATION

Item response theory (IRT; van der Linden & Hambleton, 1997) is used to calibrate all items and derive scores for all *ILEARN* items and assessments. IRT is a general framework that models test responses resulting from an interaction between students and test items.

IRT encompasses many related measurement models that allow for varied assumptions about the nature of the data. Simple unidimensional models are the most common models used in K–12 operational testing programs. In some instances, item dependencies exist, and more complex models are employed.

### 5.1 ITEM RESPONSE THEORY MODELS

*ILEARN* employed IRT models for item calibration and student ability estimation across the subject area assessments. Each subject employed models consistent with the banks and item types from which the items originated. Depending on the assessment and IRT model, either maximum likelihood estimation (MLE) or marginal maximum likelihood estimation (MMLE) was used. The various IRT models used are described first and then the models used by each assessment are outlined.

#### **Two-Parameter Logistic Model**

In the case of the two-parameter logistic model (2PL), we have:

$$p_{ij}(z_{ij}|\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \begin{cases} \frac{\exp(1.7 * a_i(\theta_j - b_{i,1}))}{1 + \exp(1.7 * a_i(\theta_j - b_{i,1}))} = p_{ij}, & \text{if } z_{ij} = 1 \\ \frac{1}{1 + \exp(1.7 * a_i(\theta_j - b_{i,1}))} = 1 - p_{ij}, & \text{if } z_{ij} = 0 \end{cases}$$

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

#### **Generalized Partial Credit Model**

In the case of the generalized partial credit model (GPC or GPCM) for items with two or more points, we have:

$$p_{ij}(z_{ij}|\theta_j, a_i, b_{i,1}, \dots, b_{i,m_i}) = \begin{cases} \frac{\exp(\sum_{k=1}^{z_{ij}} 1.7 * a_i(\theta_j - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l 1.7 * a_i(\theta_j - b_{i,k}))}, & \text{if } z_{ij} > 0 \\ \frac{1}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l 1.7 * a_i(\theta_j - b_{i,k}))}, & \text{if } z_{ij} = 0 \end{cases}$$

where  $\mathbf{b}'_i = (b_{i,1}, \dots, b_{i,m_i})$  for the  $i$ th item's step parameters,  $m_i$  is the maximum possible score of this item,  $a_i$  is the discrimination parameter for item  $i$ ,  $z_{ij}$  is the observed item

score for the person  $j$ ,  $k$  indexes step of the item  $i$ , and  $b_{i,k}$  is the  $k^{\text{th}}$  step parameter for item  $i$  with  $m_i + 1$  total categories.

### **Rasch Model**

In the case of the Rasch model for one point items we have:

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

### **Rasch Testlet Model**

In the case of the Rasch testlet model for one point items we have:

$$p_{ij}(z_{ij}|\theta_j, b_{i,1}, \dots, b_{i,m_i}, u_g) = \left\{ \begin{array}{l} \frac{\exp((\theta_j + u_g - b_{i,1}))}{1 + \exp((\theta_j + u_g - b_{i,1}))} = p_{ij}, \text{ if } z_{ij} = 1 \\ \frac{1}{1 + \exp((\theta_j + u_g - b_{i,1}))} = 1 - p_{ij}, \text{ if } z_{ij} = 0 \end{array} \right\},$$

where  $u_g$  is the nuisance dimension parameter for cluster  $g$ .

## **5.1.1 ELA and Mathematics**

ELA and Mathematics adopted the Smarter IRT framework. For one point items the two-parameter logistic model was used and for multi-point items the generalized partial credit model was used.

## **5.1.2 Science**

Science item banks were newly established. For Science items, the conditional dependencies between the assertions of an item cluster were too strong to ignore. Science adopted the Rasch Testlet Model for performance tasks (PTs). Stand-alone Science items were analyzed with the Rasch model. More information about the performance tasks can be found in Volume 2.

## **5.1.3 Social Studies**

Social Studies item banks were newly established. Grade 5 adopted a process consistent with the ELA and Mathematics, and only used the 2PL and GPC models. U.S. Government returned low sample sizes, and in order to ensure reliable item parameter estimates the Rasch model was used.

## 5.2 IRT ANALYSES RESULTS

Following the Spring 2019 *ILEARN* assessments, IRT calibrations were completed that placed all items within a grade and subject on the same scale. More information about those calibrations can be found in the 2018-2019 Technical Reports. Beginning in 2019-2020 all assessments will be pre-equated, with item parameter stability checks following each administration.

### 5.2.1 IRT Summaries

The IRT statistical properties of the final operational test forms used for *ILEARN* are summarized in Table 8.

*Table 8: Operational Item Parameter Five-Point Summary and Range, Biology*

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Biology*	b	-2.86	-1.51	-0.53	0.09	0.66	2.43	4.15

\* The same item pool was used for the Fall and Winter Biology assessments

### 5.2.2 2020 *ILEARN* Test Characteristic Curves

Another way to view the technical properties of *ILEARN* test forms is via the test characteristic curves (TCCs). These plots are displayed in Appendix B.

## 6. SCORING AND REPORTING

### 6.1 MAXIMUM LIKELIHOOD ESTIMATION

Ability estimates were generated using pattern scoring, a method that scores students depending on how they answer individual items. Scoring details are provided below.

#### 6.1.1 Likelihood Function

The likelihood function for generating the maximum likelihood estimates (MLEs) is based on a mixture of item models and can therefore be expressed as

$$L(\theta) = L(\theta)^{2PL} L(\theta)^{CR},$$

where

$$L(\theta)^{2PL} = \prod_{i=1}^{N_{2PL}} P_i^{z_i} Q_i^{1-z_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N_{CR}} \frac{\exp \sum_{l=1}^{z_i} D a_i(\theta - b_{il})}{1 + \sum_{h=1}^{m_i} \exp \sum_{l=1}^h D a_i(\theta - b_{il})}$$

$$p_i = \frac{1}{1 + \exp [-D a_i(\theta - b_i)]}$$

$$q_i = 1 - p_i$$

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  $l$ th step for item  $i$  with  $m$  total categories, and  $D = 1.7$ .

A student's theta (i.e., MLE) is defined as  $\arg \max_{\theta} \log(L(\theta))$  given the set of items administered to the student.

#### 6.1.2 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:

$$\theta_{t+1} = \theta_t - \frac{\partial \ln L(\theta_t)}{\partial \theta_t} / \frac{\partial^2 \ln L(\theta_t)}{\partial^2 \theta_t},$$

where

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



$$\begin{aligned}\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} &= \frac{\partial^2 \ln L(\theta)^{2PL}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} \\ \frac{\partial \ln L(\theta)^{2PL}}{\partial \theta} &= \sum_{i=1}^{N_{2PL}} D a_i \frac{(z_i - p_i)(p_i)}{p_i} \\ \frac{\partial^2 \ln L(\theta)^{2PL}}{\partial^2 \theta} &= - \sum_{i=1}^{N_{2PL}} D^2 a_i^2 \frac{p_i q_i}{1} \left( 1 - \frac{z_i}{p_i^2} \right) \\ \frac{\partial \ln L(\theta)^{CR}}{\partial \theta} &= \sum_{i=1}^{N_{CR}} D a_i \left( z_i - \frac{\sum_{h=1}^{m_i} h \exp(\sum_{l=1}^j D a_i (\theta - b_{il}))}{1 + \sum_{h=1}^{m_i} \exp(\sum_{l=1}^h D a_i (\theta - b_{il}))} \right) \\ \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} &= \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left( \left( \frac{\sum_{h=1}^{m_i} h \exp(\sum_{l=1}^h D a_i (\theta - b_{il}))}{1 + \sum_{h=1}^{m_i} \exp(\sum_{l=1}^h D a_i (\theta - b_{il}))} \right)^2 \right. \\ &\quad \left. - \frac{\sum_{h=1}^{m_i} h^2 \exp(\sum_{l=1}^h D a_i (\theta - b_{il}))}{1 + \sum_{h=1}^{m_i} \exp(\sum_{l=1}^h D a_i (\theta - b_{il}))} \right),\end{aligned}$$

and where  $\theta_t$  denotes the estimated  $\theta$  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 two-parameter logistic (2PL) model.

### 6.1.3 Standard Errors of Estimates

When the MLE or MMLE is available and within the LOT and HOT, the standard error (SE) is estimated based on the test information function and is estimated by

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

where

$$\begin{aligned}I(\theta_j) &= \sum_{i=1}^I D^2 a_i^2 \left( \frac{\sum_{l=1}^{m_i} l^2 \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))} \right. \\ &\quad \left. - \left( \frac{\sum_{l=1}^{m_i} l \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l D a_i (\theta_j - b_{ik}))} \right)^2 \right),\end{aligned}$$

where  $m_i$  is the maximum possible score point (starting from 0) for the  $i$ th item,  $D$  is the scale factor, 1.7.

### 6.1.4 Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded and an MLE or MMLE cannot be generated. For all incorrect tests, score by adding 0.5 to an item score with smallest a-parameter among the administered operational items for a test. For all correct tests, score by subtracting 0.5 from an item score with smallest a-parameter among the administered operational items for a student. Adding 0.5 to an incorrect item score with smallest a-parameter adds less benefit than selecting any other items, e.g., selecting the hardest item. Subtracting 0.5 from a correct item score with smallest a-parameter penalizes less than selecting any other items, e.g., selecting the easiest item.

Extreme unreliable student ability estimates are truncated to the lowest observable scores (LOT/LOSS) or the highest observable scores (HOT/HOSS). Note that LOT = lowest observable theta score, LOSS = lowest observable scale score, HOT = highest observable theta score, and HOSS = highest observable scale score. Estimated theta values lower than the LOT or higher than the HOT will be truncated to the LOT and HOT values and will be assigned the LOSS and HOSS associated with the LOT and HOT.

Table 9 through Table 12 give the LOT/LOSS and HOT/HOSS for the *ILEARN* assessments.

*Table 9: ELA Theta and Scaled-Score Limits for Extreme Ability Estimates*

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
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

*Table 10: Mathematics Theta and Scaled-Score Limits for Extreme Ability Estimates*

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
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

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
8	-5.0667	6.0000	6120	6950

Table 11: Science Theta and Scaled-Score Limits for Extreme Ability Estimates

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
4	-3	3	7350	7650
6	-3	3	7350	7650
Biology	-3	3	7350	7650

Table 12: Social Studies Theta and Scaled-Score Limits for Extreme Ability Estimates

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
5	-3	3	8350	8650
U.S. Government	-3	3	8350	8650

### 6.1.5 Standard Errors of LOT/HOT Scores

For standard error of LOT/HOT scores, theta in the formula in Section 6.1.3 is replaced with the LOT/HOT values. The upper bound of the SE was set to 2.5 for all grades and subjects.

## 6.2 TRANSFORMING THETA SCORES TO REPORTING SCALE SCORES

For 2019-2020, scale scores were reported for each student who took the *ILEARN* assessments. The scale scores were based on the operational items presented to the student and did not include any field-test items. The scale score is a linear transformation of the IRT ability estimate,  $\theta$ :

$$SS = a * \theta + b,$$

where  $a$  is the slope and  $b$  is the intercept. Table 13 lists the scaling constants  $a$  and  $b$  for the *ILEARN* assessments.

When administered, ELA and Mathematics are reported on a vertical scale. The IRT vertical scale was established by Smarter 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 2013–2014 Technical Report (Smarter Balanced, 2016).

The slope and intercept used to transform the IRT ability estimate to a scale score are unique to Indiana and the *ILEARN* assessments.

Each Science and Social Studies assessment was reported on a separate within-test scale.

The summary of *ILEARN* scale scores for each test is provided in Appendix C, and the summary of scale scores for each reporting category is provided in Appendix D.

*Table 13: Scaling Constants on the Reporting Metric*

Subject	Grade	Slope (a)	Intercept (b)
ELA	3–8	75	5500
Mathematics	3–8	75	6500
Science	4, 6, Biology	50	7500
Social Studies	5, U.S. Government	50	8500

### 6.3 OVERALL PERFORMANCE CLASSIFICATION

Each student who tested during the 2019-2020 school year was assigned an overall performance category in accordance with his or her overall scale score. Table 14 through Table 18 provide the scale score range for performance standards for *ILEARN*. The lower bound of the Level 3, At Proficiency, marks the minimum cut score for proficiency.

*Table 14: Proficiency Levels for ELA*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
3	5060–5415	5416–5459	5460–5514	5515–5760
4	5090–5443	5444–5492	5493–5546	5547–5810
5	5110–5471	5472–5523	5524–5594	5595–5850
6	5130–5491	5492–5543	5544–5603	5604–5870
7	5130–5506	5507–5567	5568–5628	5629–5890
8	5150–5510	5511–5576	5577–5637	5638–5920

*Table 15: Proficiency Levels for Mathematics*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
3	6080–6381	6382–6424	6425–6487	6488–6730
4	6100–6428	6429–6473	6474–6540	6541–6800
5	6110–6452	6453–6509	6510–6565	6566–6850

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
6	6110–6487	6488–6544	6545–6604	6605–6870
7	6120–6492	6493–6561	6562–6624	6625–6920
8	6120–6508	6509–6589	6590–6650	6651–6950

*Table 16: Proficiency Levels for Science*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
4	7350–7481	7482–7505	7506–7534	7535–7650
6	7350–7465	7466–7503	7504–7544	7545–7650
Biology	7350–7477	7478–7508	7509–7546	7547–7650

*Table 17: Proficiency Levels for Social Studies Grade 5*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
5	8350–8476	8477–8501	8502–8542	8543–8650

*Table 18: Proficiency Levels for Social Studies U.S. Government*

Grade	Level 1 Below Proficiency	Level 2 At Proficiency
U.S. Government	8350–8496	8497–8650

## 6.4 REPORTING CATEGORY SCORES

### 6.4.1 MLE/MMLE Scoring

Reporting category theta scores were calculated using either MLE or MMLE, depending on the assessment, 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.

### 6.4.2 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 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  $\text{round}(SS_{rc} + 1.5 * SE(SS_{rc}), 0) < SS_p$ ;
- At/Near (Code = 2): if  $\text{round}(SS_{rc} + 1.5 * SE(SS_{rc}), 0) \geq SS_p$  and  $\text{round}(SS_{rc} - 1.5 * SE(SS_{rc}), 0) < SS_p$ , a strength or weakness is indeterminable; and
- Above (Code = 3): if  $\text{round}(SS_{rc} - 1.5 * SE(SS_{rc}), 0) \geq SS_p$ ,

where  $SS_{rc}$  is the student's scale score on a reporting category;  $SS_p$  is the proficiency scale score cut (Level 3 cut); and  $SE(SS_{rc})$  is the standard error of the student's scale score on the reporting category.

### 6.4.3 Standard Level Aggregate Scores

Standard level information was reported relative to the proficiency standard for tests that were adaptively administered. In Spring 2020 standard level information would have been reported for the ELA, Mathematics, and Science assessments.

Start by defining  $p_{ij} = p(z_{ij} = 1)$ , representing the probability that student  $j$  responds correctly to item  $i$  ( $z_{ij}$  represents the  $j^{\text{th}}$  student's score on the  $i^{\text{th}}$  item). For items with one score point we use the 2PL IRT model to calculate the expected score on item  $i$  for student  $j$  with  $\theta_{\text{Level 3 cut}}$  as:

$$E(z_{ij}) = \frac{\exp(1.7 * a_i(\theta_{\text{Level 3 cut}} - b_i))}{1 + \exp(1.7 * a_i(\theta_{\text{Level 3 cut}} - b_i))}.$$

For items with two or more score points, using the generalized partial credit model, the expected score for student  $j$  with a Level 3 cut on an item  $i$  with a maximum possible score of  $m_i$  was calculated as:

$$E(z_{ij}) = \sum_{l=1}^{m_i} \frac{\exp(\sum_{k=1}^l 1.7 * a_i(\theta_{\text{Level 3 cut}} - b_{i,k}))}{1 + \sum_{l=1}^{m_i} \exp(\sum_{k=1}^l 1.7 * a_i(\theta_{\text{Level 3 cut}} - b_{i,k}))}.$$

For each item  $i$ , the residual between observed and expected score for each student was defined as:

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

Residuals are summed for items within a standard. The sum of residuals was divided by the total number of points possible for items within the standard,  $S$ :

$$\delta_{jS} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a standard score was computed by averaging individual student standard scores for the standard, across students of different abilities receiving different items measuring the same standard at different levels of difficulty,

$$\bar{\delta}_{Sg} = \frac{1}{n_g} \sum_{j \in g} \delta_{js},$$

and

$$se(\bar{\delta}_{Sg}) = \sqrt{\frac{1}{n_g(n_g - 1)} \sum_{j \in g} (\delta_{js} - \bar{\delta}_{Sg})^2},$$

where  $n_g$  is the number of students who responded to any of the items that belong to the standard  $S$  for an aggregate unit  $g$ . If a student did not see any items on a particular standard, the student was NOT included in the  $n_g$  count for the aggregate.

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

The statistic  $\bar{\delta}_{Tg}$  was not directly reported; instead, the aggregate was reported to show if a group of students performed better, worse, or as expected on this standard. In some cases, insufficient information was available and that was indicated as well.

For standard level strengths/weaknesses, the following were reported:

- If  $\bar{\delta}_{Sg} \geq +1.5 * se(\bar{\delta}_{Sg})$ , then performance is *above* the Proficiency Standard.
- If  $\bar{\delta}_{Sg} \leq -1.5 * se(\bar{\delta}_{Sg})$ , then performance is *below* the Proficiency Standard.
- Otherwise, performance is *near* the Proficiency Standard.
- If  $se(\bar{\delta}_{Sg}) > 0.2$ , data are insufficient.

## 6.5 LEXILE AND QUANTILE SCORES

ILEARN reports Lexile and Quantile measures with ELA and Mathematics test scores. MetaMetrics provided conversion tables between ELA scale scores and Lexile measures and between Mathematics scale scores and Quantile measures for each grade and subject.

## 6.6 COMPARISON OF SCORES TO PREVIOUS YEAR

The ILEARN Biology assessments administered in fall and winter of 2019-2020 were also administered during the 2018-2019 school year, and thus year-to-year comparisons are possible. Table 19 provides a summary of the Biology test administrations between the 2018-2019 and 2019-2020 test administrations. As Table 19 indicates, the number of students participating in the Biology assessment decreased sharply, especially with respect to the Winter test administration.

*Table 19: Year-to-Year Biology Scale Score Comparisons – All Students*

Administration Window and Year	Total Number of Schools Tested	Number of Students Included	Mean Scale Score	Standard Deviation of Scale Scores
Fall 2018	18	1215	7496	50
Fall 2019	41	870	7479	49
Winter 2019	123	6605	7504	52
Winter 2020	26	1713	7489	48

To provide a more meaningful comparison of student Biology achievement between school years, Table 20 shows the means and standard deviations between school years for only those schools that participated in both the 2018-2019 and 2019-2020 test administrations. Although the schools represented in Table 20 participated in the Biology test administrations during both school years, the number of participating students dropped substantially. The pattern of decreased Biology achievement between school years persists even when restricting the analysis to those schools participating in both school years.

*Table 20: Year-to-Year Biology Scale Score Comparisons – Matched Schools*

Administration Window and Year	Number of Schools Included in Comparison	Number of Students Included	Mean Scale Score	Standard Deviation of Scale Scores
Fall 2018	9	729	7498	48
Fall 2019	9	607	7487	43
Winter 2019	24	2629	7508	52
Winter 2020	24	1678	7489	48

For schools that participated in both the 2018-2019 and 2019-2020 test administrations, Table 21 provides the percent of students in each of the four performance levels. While the number of students in Level 3 (At Proficiency) is steady between the administrations, the number of students in Level 4 (Above Proficiency) varies.

*Table 21: Year-to-Year Biology Performance Level Comparisons – Matched Schools*

Administration and Year	Number of Students Included	Percent of Students in Level 1	Percent of Students in Level 2	Percent of Students in Level 3	Percent of Students in Level 4
Fall 2018	729	35	26	23	16
Fall 2019	607	39	31	22	8
Winter 2019	2629	30	24	23	23
Winter 2020	1678	41	26	21	12



## **7. QUALITY CONTROL PROCEDURES**

CAI's quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Scoring procedures that cannot be reasonably automated are replicated by two independent analysts at CAI.

### **7.1 SCORING QUALITY CHECK**

All student test scores were produced using CAI's scoring engine. Prior to releasing any scores, a second score verification system was used to verify that all test scores match with 100% agreement in all tested grades. This second system is independently constructed and maintained from the main scoring engine and separately estimates marginal maximum likelihood estimations using the procedures described within this report.

Additionally, HumRRO provided replication of the psychometric scoring process for *ILEARN*. Scores were approved and published by the IDOE only when all three independent systems matched. For the Fall Biology administration HumRRO was initially unable to produce code to fully replicate results and an abbreviated process was used. IDOE signed-off on the abbreviated process and scores were released in January 2020. CAI provided support to HumRRO regarding the estimation method of the IRT models. CAI also provided documentation from another state indicating their replication and validation of CAI's scoring when the same models were used. HumRRO was able to produce fully functioning code prior to the release of the Winter Biology scores.

## 8. REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.
- Bock R.D., Zimowski M.F. (1997) Multiple Group IRT. In: van der Linden W.J., Hambleton R.K. (eds) *Handbook of Modern Item Response Theory*. Springer, New York, NY
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.
- Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach (ETS Research Report No. 91–47). Princeton, NJ: Educational Testing Service.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.
- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176.
- Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40:106–108.
- Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- van der Linden, W. J. & Hambleton, R. K. (Eds.) (1997) *Handbook of modern item response theory*. New York: Springer-Verlag.
- Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement* (ETS Research Report No. 12–08). Princeton, NJ: Educational Testing Service.



**Indiana's Learning Evaluation  
and Readiness Network  
(*ILEARN*)**

**2019–2020**

**Volume 2  
Test Development**

## **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](mailto:INassessments@doe.in.gov).

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Stephan Ahadi, Elizabeth Ayers-Wright, Elizabeth Xiaoxin Wei, Kevin Clayton, Aleah Pepper, Kyra Bilenski, Christopher Johnston, and Gabriel Martinez. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and Program Leads.

## TABLE OF CONTENTS

1.	INTRODUCTION.....	1
1.1	Claim Structure .....	2
1.2	Underlying Principles Guiding Development.....	3
1.3	Organization of this Volume .....	4
2.	<i>ILEARN</i> ITEM BANK SUMMARY.....	5
2.1	Item Banks .....	5
2.2	Item Acceptance Meetings.....	7
2.3	Item Bank Composition .....	7
3.	ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS ....	10
3.1	Overview .....	10
3.2	Passage and Item Specifications .....	11
3.2.1	<i>Passage Specifications</i> .....	12
3.2.2	<i>Item Specifications</i> .....	13
3.3	Selection and Training of Item Writers .....	16
3.4	Internal Review .....	16
3.4.1	<i>Preliminary Review</i> .....	16
3.4.2	<i>Content Review 1</i> .....	17
3.4.3	<i>Edit Review 1</i> .....	18
3.4.4	<i>Senior Content Review</i> .....	18
3.5	Review by State Personnel and Stakeholder Committees .....	19
3.5.1	<i>State (Client) Review</i> .....	19
3.5.2	<i>Content/Fairness Committee Review</i> .....	19
3.5.3	<i>Markup for Translation and Accessibility Features</i> .....	20
3.5.4	<i>Indiana Educator Review of Licensed Item Banks</i> .....	20
3.6	Field Testing.....	20
3.7	Post-Field-Test Review .....	20
3.7.1	<i>Key Verification</i> .....	21
3.7.2	<i>Rubric Validation</i> .....	21
3.7.3	<i>Rangefinding</i> .....	22
3.7.4	<i>Data Review</i> .....	22
4.	<i>ILEARN</i> BLUEPRINTS AND STATE ASSESSMENT TEST CONSTRUCTION...	23
4.1	Test Blueprints .....	23
4.1.1	<i>Blueprint Construction Meeting</i> .....	23
4.1.2	<i>ILEARN Test Specifications</i> .....	23
4.1.3	<i>ELA Blueprints</i> .....	27
4.1.4	<i>Mathematics Blueprints</i> .....	27
4.1.5	<i>Science Blueprints</i> .....	28
4.1.6	<i>Social Studies Blueprints</i> .....	28
4.2	Test Form Construction.....	28
4.3	Test Form Assembly .....	29

4.4	Roles and Responsibilities .....	30
4.4.1	Role of the CAI Content Team .....	30
4.4.2	Role of the CAI Technical Team .....	31
4.4.3	Role of IDOE .....	31
4.5	Target Guidelines .....	31
4.6	Accommodated Form Construction .....	31
4.6.1	Test Characteristic Curve .....	33
4.6.2	Test Characteristic Curve Difference .....	34
4.6.3	Conditional Standard Error of Measurement Curve .....	34
5.	PERFORMANCE LEVEL DESCRIPTORS .....	36
6.	REFERENCES .....	37

## LIST OF TABLES

Table 1:	Sources of Items for the ILEARN 2019–2020 Assessments .....	1
Table 2:	ELA Claims .....	2
Table 3:	Mathematics Categories .....	3
Table 4:	Operational Item Counts by Source .....	5
Table 5:	Operational Performance Task Counts by Source .....	6
Table 6:	ILEARN Item Types and Descriptions .....	7
Table 7:	ELA Operational Items by Item Type and Grade .....	8
Table 8:	Mathematics Operational Items by Item Type and Grade .....	8
Table 9:	Science Operational Items by Item Type and Grade .....	9
Table 10:	Social Studies Operational Items by Item Type and Grade .....	9
Table 11:	How Each Step of Development Supports the Validity of Claims .....	10
Table 12:	ILEARN Item Specifications .....	11
Table 13:	Sample ELA Item Specification for Grade 4 .....	14
Table 14:	Number of Hand-Scored Items by Form .....	24
Table 15:	Number of Embedded Field-Test Items by Form .....	24
Table 16:	Blueprint Percentage of Test Items Assessing Each Reporting Category in ELA .....	25
Table 17:	Blueprint Percentage of Test Items Assessing Each Reporting Category in Mathematics .....	25
Table 18:	Blueprint Percentage of Test Items Assessing Each Reporting Category in Science .....	26
Table 19:	Blueprint Percentage of Test Items Assessing Each Reporting Category in Social Studies .....	26
Table 20:	Statistical Test Summary Comparison for Grade 5 Social Studies Online and Paper Forms .....	32

## **LIST OF FIGURES**

Figure 1: Features of the REVISE Software .....	22
Figure 2: TCC Comparisons of Grade 5 Social Studies Online and Paper Forms .....	33
Figure 3: TCC Differences of Grade 4 Science Online and Accommodated Forms .....	34
Figure 4: CSEM Comparisons of Grade 4 Science Online and Accommodated Forms	35

## **LIST OF APPENDICES**

Appendix A: English/Language Arts Blueprints
Appendix B: Mathematics Blueprints
Appendix C: Science Blueprints
Appendix D: Social Studies Blueprints
Appendix E: <i>ILEARN</i> Passage Specifications
Appendix F: Example Item Types
Appendix G: Item Review Checklist
Appendix H: Item Writer Training Materials

# 1. INTRODUCTION

As discussed in Volume 1, due to the COVID-19 pandemic, all Spring 2020 ELA, Mathematics, Science, and Social Studies assessments were cancelled. However, the Fall and Winter Biology test administrations were completed and students that tested during these windows received scores. Since test development for the Spring 2020 assessments was completed prior to administration cancellation, this volume elaborates on the test development process for all 2020 forms.

*ILEARN* assessments were designed to align with the Indiana Academic Standards (IAS) and encompass a variety of item types from several sources.

The IAS were approved by the Indiana State Board of Education in April 2014 for English/Language Arts (ELA) and Mathematics, and in March 2015 for Social Studies. The IAS for Science were originally revised in 2010 but were updated in 2016 to reflect changes in Science content. The IAS are intended to implement more rigorous standards that promote college-and-career readiness, with the goal of challenging and motivating Indiana’s students to acquire stronger critical thinking, problem solving, and communications skills.

Table 1 denotes the sources of the items used in 2019-2020, including licensed item banks (Smarter Balanced Assessment Consortium [Smarter], Independent College and Career Ready [ICCR], and Hawaii End-of-Course [EOC]), legacy Indiana Statewide Testing for Educational Progress-Plus (*ISTEP+*) items, and custom Indiana development. Each item source is outlined in more detail in Section 2.

The Smarter and ICCR ELA, Mathematics, and Science item banks were developed to measure college-and-career readiness standards as embodied in the Common Core State Standards (CCSS). The item banks are designed to measure the full breadth and depth of the standards and cover a range of difficulty that matches the distribution of student performance in each grade and subject. The item banks are designed primarily for accountability assessments. However, not all CCSS map directly to the IAS, so items from other sources (e.g., legacy *ISTEP+* and custom development) were needed to fill those gaps.

*Table 1: Sources of Items for the ILEARN 2019–2020 Assessments*

Subject and Grade(s)	Licensed Bank(s)	Legacy <i>ISTEP+</i> Items	Custom Development	Notes
ELA 3–8	Smarter  ICCR	Yes	Yes	ICCR items were used to augment the pool where the Smarter item pool could not provide items or provided items only to a limited extent. <i>ISTEP+</i> items were used only when required to ensure blueprint was met.



Subject and Grade(s)	Licensed Bank(s)	Legacy ISTEP+ Items	Custom Development	Notes
Mathematics 3–8	Smarter  ICCR	Yes	Yes	ICCR items were used to augment the pool where the Smarter item pool could not provide items or provided items to a limited extent only. ISTEP+ items were used only when required to ensure blueprint was met.
Science 4 and 6	ICCR	Yes	Yes	Very few ICCR items were used operationally in 2019–2020.
Science Biology	Hawaii EOC  ICCR	Yes	Yes	Very few ICCR items were used operationally in 2019–2020.
Social Studies 5	No	Yes	Yes	
U.S. Government	No	No	Yes	

## 1.1 CLAIM STRUCTURE

The *ILEARN* assessments are designed to measure college-and-career readiness and support the assessments claim that students in grades 3–8 demonstrate progress toward college-and-career readiness in ELA, Mathematics, Science, and Social Studies.

Within ELA, items are designed to support the following claims about proficient students, shown in Table 2.

Table 2: ELA Claims

ELA Claims
Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
Students can write well-structured, focused texts for a variety of purposes, analytically integrating information from multiple sources.
Students know and can apply the rules of standard, written English.

In Mathematics, assessments support claims such as the following: *Proficient students in grade 7 can use procedures involving rational numbers to solve problems, model real-world phenomena, and reason mathematically.* The specific claims vary by grade level and are summarized for Mathematics in Table 3.

**Table 3: Mathematics Categories**

<b>Grade</b>	<b>Reporting Categories</b>				
<b>Grade 3</b>	Algebraic Thinking and Data Analysis	Computation	Geometry and Measurement	Number Sense	Process Standards
<b>Grade 4</b>	Algebraic Thinking and Data Analysis	Computation	Geometry and Measurement	Number Sense	Process Standards
<b>Grade 5</b>	Algebraic Thinking	Computation	Geometry and Measurement, Data Analysis, and Statistics	Number Sense	Process Standards
<b>Grade 6</b>	Algebra and Functions	Computation	Geometry and Measurement, Data Analysis, and Statistics	Number Sense	Process Standards
<b>Grade 7</b>	Algebra and Functions	Data Analysis, Statistics, and Probability	Geometry and Measurement	Number Sense and Computation	Process Standards
<b>Grade 8</b>	Algebra and Functions	Data Analysis, Statistics, and Probability	Geometry and Measurement	Number Sense and Computation	Process Standards

## 1.2 UNDERLYING PRINCIPLES GUIDING DEVELOPMENT

The Smarter and ICCR item banks were established using a highly structured, evidence-centered design. The process for their development, as well as for the custom development and legacy *ISTEP+* banks, began with detailed item specifications. The specifications, discussed in a later section, described the interaction types that could be used, provided guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and offered sample items.

Items were written with the goal that virtually every item would be accessible to all students, either by itself or in conjunction with accessibility tools, such as text-to-speech, translations, or assistive technologies. This goal is supported by the delivery of the items on CAI's test delivery platform, which has received an internationally recognized accessibility standard known as Web Content Accessibility Guidelines (WCAG) 2.0 AA certification, offers a wide array of accessibility tools, and is compatible with most assistive technologies.

Item development efforts support the goal of high-quality items through rigorous development processes managed and tracked by a content development platform that

ensures that every item flows through the correct sequence of reviews and captures every comment and change to the item.

IDOE sought to ensure that the items were measuring the standards in a fair and meaningful way by engaging educators and other stakeholders at each step of the development process. Educators evaluated the alignment of items to the standards and offered guidance and suggestions for improvement. They participated in the review of items for fairness and sensitivity. Following the field testing of items, educators engaged in *rubric validation*, a process that refines rule-based rubrics upon review of student responses, as well as data review.

For the licensed Smarter and ICCR items, in coordinating among states, educators in multiple states frequently reviewed the same items using the same criteria. In general, one state was assigned rights to modify the items, and other states were offered the modified items on an accept-reject basis.

Combined, these principles and the processes that support them have led to an item bank that measures the IAS with fidelity and does so in a way that minimizes construct-irrelevant variance and barriers to access. The details of these processes follow.

### **1.3 ORGANIZATION OF THIS VOLUME**

This volume is organized in three sections:

- An overview of the item pool, the types of assessments the pool is designed to support, and methods for refreshing the pool;
- An overview of the item development process that supports the validity of the claims that *ILEARN* assessments are designed to support; and
- A description of test construction for the *ILEARN* assessments for ELA, Mathematics, Science, and Social Studies, including the blueprint design and test construction.

## 2. ILEARN ITEM BANK SUMMARY

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

### 2.1 ITEM BANKS

Table 4 provides the count of items, by source, used on the 2019–2020 *ILEARN* assessments.

The *ILEARN* ELA and Mathematics operational item banks draw primarily from the Smarter item bank, which includes more than 30,000 items across grades and subjects. However, not all IAS are covered by Smarter items. Items from CAI's ICCR item bank and legacy *ISTEP+* items were used when needed to fill existing gaps in IAS coverage across grades. In addition, in a few small instances, new, custom Indiana item development was needed to complete the item bank and ensure complete coverage of the IAS.

For Science grades 4 and 6, the item banks consisted mostly of previous *ISTEP+* items, augmented by custom development. In Biology, the Hawaii EOC Biology item pool was used primarily and was augmented by ICCR, previous *ISTEP+*, and custom Indiana development items as needed to fill gaps in coverage to the IAS.

The Social Studies grade 5 item pool contains custom Indiana development and previous *ISTEP+* items. The U.S. Government item pool is comprised of completely custom Indiana development items.

*Table 4: Operational Item Counts by Source*

Subject and Grade	# of Smarter Items	# of ICCR Items	# of <i>ISTEP+</i> Legacy Items	# of Custom Items	# of Hawaii EOC items
ELA 3	351	25	22	12	
ELA 4	260	29	23	21	
ELA 5	235	17	22	20	
ELA 6	181	37	11	14	
ELA 7	265	39	21	21	
ELA 8	348	13	23	13	
Mathematics 3	387	47	25	30	
Mathematics 4	441	17	19	24	
Mathematics 5	340	50	21	29	

Subject and Grade	# of Smarter Items	# of ICCR Items	# of ISTEP+ Legacy Items	# of Custom Items	# of Hawaii EOC items
Mathematics 6	488	19	15	14	
Mathematics 7	446	32	17	20	
Mathematics 8	325	23	14	17	
Science 4		20	48	70	
Science 6		13	64	77	
Biology		17	38	15	183
Social Studies 5			46	16	
U.S. Government				54	

Additionally, all assessments other than Social Studies included one performance task per grade. Table 5 lists the counts of performance tasks in the 2019–2020 item pool.

*Table 5: Operational Performance Task Counts by Source*

Subject and Grade	# of Smarter Performance Tasks	# of Custom Indiana Performance Tasks
ELA 3	2	-
ELA 4	3	-
ELA 5	3	-
ELA 6	2	-
ELA 7	3	-
ELA 8	3	-
Mathematics 3	2	-
Mathematics 4	3	-
Mathematics 5	2	-
Mathematics 6	2	-
Mathematics 7	2	-
Mathematics 8	5	-
Science 4	-	1
Science 6	-	1
Biology	-	2

## 2.2 ITEM ACCEPTANCE MEETINGS

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

Prior to the Spring 2019 administration two item acceptance review meetings were held. Results of those meetings can be found in Volume 2 of the 2018-2019 Technical Reports.

In November 2019 a third item acceptance review meeting was held for ELA and Mathematics. A description of the meeting, as well as the results, can be found in Appendix I.

## 2.3 ITEM BANK COMPOSITION

Table 6 lists the ELA, Mathematics, Science, and Social Studies item types and provides a brief description of each. Examples of various item types can be found in Appendix F, Example Item Types. Table 7 through Table 10 list the number of items by type for each grade and subject.

*Table 6: ILEARN Item Types and Descriptions*

Response Type	Description
Edit Task with Choice (ETC)*	Student chooses a word or phrase from several options in order to complete a sentence.
Equation Response (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. Responses can include numbers, fractions, expressions, inequalities, functions, and equations.
Evidence-Based, Selected-Response (EBSR)	Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A.
Extended Response (ER)	Student is directed to provide a longer, written response in the form of an essay.
Graphic Response (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hot Text (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference.
Multiple-Choice (MC)	Student selects one correct answer from four options.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Performance Task (PT)	Student works through a group of items measuring multiple standards and using various item types to demonstrate the ability to integrate knowledge and skills.

Response Type	Description
Simulation (SIM)	Student selects inputs to “run” trials. Data is presented in a table after trials are run.
Table Input (TI)	Student types numeric values into a given table.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Text Entry (TE)	Student is directed to type their response in a text box.

*\*Note: Four legacy ISTEP+ ETC items were approved for inclusion in the pool by IDOE content specialists; however, CAI did not develop any custom ETC items for ELA.*

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

**Table 7: ELA Operational Items by Item Type and Grade**

Item Type	3	4	5	6	7	8
TE	22	25	28	17	34	43
ETC	1	1	1			1
EBSR	62	35	34	41	24	41
HT	40	42	38	24	53	50
MI	23	13	7	11	4	7
MC	188	166	125	106	155	169
MS	72	49	58	42	74	84
ER	2	2	3	2	2	2

**Table 8: Mathematics Operational Items by Item Type and Grade**

Item Type	3	4	5	6	7	8
TE	6	7	5	6	3	10
EQ	261	280	244	274	289	111
GI	52	22	17	26	19	31
MI	32	70	72	53	37	66
MC	125	95	81	78	75	96
MS	11	11	15	83	90	60
HT		1		1		
TI	2	15	6	15	2	5

*Table 9: Science Operational Items by Item Type and Grade*

Item Type	4	6	Biology**
TE	15	5	4
ETC	10	9	6
EBSR			1
EQ	2	3	2
GI		2	36
HT	1	3	
MI	2	7	4
MC	90	99	183
MS	12	19	9
PT*	1	1	2
SIM			1
TI	2	3	3
IC & MC**		1	
IC & MS**	2	1	1
EQ & MC**			1
MS & MC**		1	
TI & MC**	1		

\*A PT has multiple interactions of various item types that sometimes include a simulation.

\*\*Eight items required two response types.

*Table 10: Social Studies Operational Items by Item Type and Grade*

Item Type	5	U.S. Government
TE	4	
EBSR	1	19
MC	54	10
MI	2	1
MS	1	24



### 3. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

#### 3.1 OVERVIEW

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

The process began with the definition of passage and item specifications, and continued with the following steps:

- 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 had a role in ensuring that the items could support the claims on which they were based. Table 11 describes how each step contributed to these goals. Each step in the process is discussed in more detail below.

*Table 11: How Each Step of Development Supports the Validity of Claims*

	<b>Supports alignment to the standards</b>	<b>Reduces construct-irrelevant variance through universal design</b>	<b>Expands access through linguistic and other supports</b>
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	Training in language accessibility, bias, and sensitivity to help item writers avoid unnecessary barriers.	

	<b>Supports alignment to the standards</b>	<b>Reduces construct-irrelevant variance through universal design</b>	<b>Expands access through linguistic and other supports</b>
	item writers about selection of item types for measurement and accessibility.		
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; 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	Final, more focused check on flagged items. Rubric validation and rangefinding ensure that scoring reflects standards and expectations.	Final, focused review on items flagged for differential item function.	

## 3.2 PASSAGE AND ITEM SPECIFICATIONS

Per the recommendations of the 2016 *ISTEP+* Panel, the Indiana Department of Education is leveraging quality content from third-party item banks for use on *ILEARN* assessments. These item banks are accompanied by item specifications which will be utilized where alignment was confirmed by Indiana educators. The specifications available are described in Table 12 below.

*Table 12: ILEARN Item Specifications*

<b>Specification</b>	<b>Developer</b>	<b>Content Areas Included</b>
Indiana Item Specifications	Developed by Indiana for Indiana standards and define custom item development	Mathematics, English/Language Arts, Science, Social Studies

Specification	Developer	Content Areas Included
ICCR Item Specifications*	Developed by Cambium Assessment, Inc (CAI) for their Independent College-and-Career-Ready item bank.	Mathematics, English/Language Arts, Science
Smarter Balanced Item Specifications*	Developed by Smarter Balanced for their Smarter Balanced item bank.	Mathematics, English/Language Arts

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

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

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

ILEARN item specifications (used for custom Indiana development) were developed by Indiana educators at a workshop in February 2018. They were further reviewed both by CAI test developers and IDOE content specialists.

Item specifications for the Hawaii Biology EOC items were created by CAI assessment specialists in conjunction with the Hawaii Department of Education's Office of Curriculum, Instruction, and Student Support. The specifications use content specialist understanding of the CCSS, as well as information about the Biology course design, to detail information for development of items to the standards.

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

### 3.2.1 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 E, *ILEARN* Passage Specifications, contains sample *ILEARN* passage specifications.

### 3.2.2 Item Specifications

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

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

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

Table 13 is a sample of the item specifications that content experts, in collaboration with Indiana educators, developed for a grade 4 Reading: Vocabulary standard. It outlines the limits of the item content to fully address the standard. The acceptable response mechanisms that are recommended to assess this standard are noted. The DOK sections explain the demands for the DOK level and provide the acceptable response mechanisms. This level of detail provides the item writer with guidance when developing

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

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

*Table 13: Sample ELA Item Specification for Grade 4*

Content Standard	<b>4.RV.2.2:</b> Identify relationships among words, including more complex homographs, homonyms, synonyms, antonyms, and multiple meanings.
Content Limits	Items should ask students not to define the type of word that is being used but rather to demonstrate its meaning between the words.  Items may refer only to synonym and antonym in the stimuli.  All words should be provided with sufficient context for support.
Construct-Relevant Vocabulary	antonyms, meaning, opposite, phrase, relationship, replace, similar/same as, synonyms,
Recommended Response Mechanisms (Item Types)	Drag and Drop Evidence-Based Selected Response Hot Text Multiple Choice Multi-Select
DOK	2
<b>Evidence Statements</b>	
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.)	
<b>Sample Item</b>	
<p>Why is “[word X]” a better word to use from paragraph 4 than “[word Y]”?</p> <p>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]</p>	
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.)	<p>American Sign Language: Allows a student to see a video of an ASL interpreter. This option will be included only if the media contains audio.</p> <p>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.</p>

	<p>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.</p> <p>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 brailleable whenever possible; those that cannot be brailled will be provided to blind/visually impaired students through a verbal or written description.</p>
Linguistic Complexity	Rating to be completed after all final edits have been applied and approved by IDOE.

Similar to ELA, Mathematics, Science, and Social Studies item specifications may include any or all of the following information.

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

### **3.3 SELECTION AND TRAINING OF ITEM WRITERS**

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

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

Key materials are included in Appendix H, Item Writer Training Materials. These include:

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

### **3.4 INTERNAL REVIEW**

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

All Smarter, ICCR, Hawaii Biology, 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 G, Item Review Checklist. The CAI internal review cycle includes the following phases:

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

#### **3.4.1 Preliminary Review**

Preliminary review is conducted by team leads or senior content staff. Sometimes, preliminary review is conducted in a group setting, led by a senior test developer. During

the preliminary review process, test developers, either individually or as a group, analyze items to ensure the following is true for all items.

- The item aligns with the academic standard.
- The item matches the item specification for the skill being assessed.
- The item is based on a quality idea (i.e., it assesses something worthwhile in a reasonable way).
- The item is properly aligned to a DOK level.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility, bias, and sensitivity.
- The content is accurate and straightforward.
- The graphic and stimulus materials are necessary to answer the question.
- The stimulus is clear, concise, and succinct (i.e., it contains enough information to know what is being asked, it is stated positively, and it does not rely on negatives—such as *no*, *not*, *none*, *never*—unless absolutely necessary).

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

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

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

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

### 3.4.2 Content Review 1

Content Review 1 is conducted by a senior content specialist who was not part of the Preliminary Review. This reviewer carefully examines each item based on all the criteria identified for Preliminary Review. Note that the criteria used for these internal reviews matches the same criteria used by committee members during



Content/Fairness Committee Reviews, as documented in Appendix G. 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.

### **3.4.3 Edit Review 1**

During Edit Review 1, editors have four primary tasks.

First, editors perform basic line editing for correct spelling, punctuation, grammar, and mathematical and scientific notation, ensuring consistency of style across the items.

Second, editors ensure that all items are accurate in content. Editors compare reading passages against the original publications to make sure that all information is internally consistent across stimulus materials and items, including names, facts, or cited lines of text that appear in the item. Editors ensure that the answer keys are correct and that all information in the item is correct. For mathematics items, editors perform all calculations to ensure accuracy.

Third, editors review all material for fairness and language accessibility issues, using CAI's Language Accessibility, Bias, and Sensitivity (LABS) Guidelines.

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

### **3.4.4 Senior Content Review**

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

### **3.5 REVIEW BY STATE PERSONNEL AND STAKEHOLDER COMMITTEES**

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

#### **3.5.1 State (Client) Review**

After items have been developed in the ICCR item bank, state content experts review any eligible items prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, or alignment or DOK updates. A CAI director for Mathematics or ELA reviews all client-requested edits in light of the ICCR item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients have the option to present these items to committee (based on the edits made) or withhold them from committee review.

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

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.

#### **3.5.2 Content/Fairness Committee Review**

During the Content/Fairness Committee Reviews, items are reviewed for content validity, grade-level appropriateness, and alignment to the content standards. Content Advisory Committee Review members are typically grade-level and subject-matter experts, but may also be mathematics coaches (who can speak to standards across grades) or literacy specialists. During this review, educators also ensure that the rubrics for machine-scored constructed-response items reflect the anticipated correct responses (see more information Section 3.7.2, Rubric Validation).

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

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 representatives 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.

### **3.5.3 Markup for Translation and Accessibility Features**

After all approved state and committee recommended edits have been applied, the items are considered “locked” and ready for all accessibility tagging. 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.

Accessibility markup, such as translations or for text-to-speech, follows similar processes. One trained expert enters the markup. A second expert reviews the work and recommends changes if necessary. If there is disagreement, a third expert is engaged to resolve the conflict.

### **3.5.4 Indiana Educator Review of Licensed Item Banks**

Because *ILEARN* relies heavily on licensed banks, a process for ensuring alignment of those items to the Indiana Academic Standards was developed by CAI and IDOE. Prior to the Spring 2019 administration two item acceptance review meetings were held. Results of those meetings can be found in Volume 2 of the 2018-2019 Technical Reports.

In November 2019 a third item acceptance review meeting was held for ELA and Mathematics. A description of the meeting, as well as the results, can be found in Appendix I.

## **3.6 FIELD TESTING**

All Smarter and ICCR items were field tested embedded in operational, summative, accountability assessments in participating states. Previously operational *ISTEP+* legacy items were field tested in Indiana prior to Spring 2019. Custom Indiana development was field tested (either as embedded field-test items or operational field-test items) in Spring 2019.

Due to the cancellation of the spring 2020 test administrations, no field test data was collected.

## **3.7 POST-FIELD-TEST REVIEW**

Following field testing, items were subject to additional reviews. These included:

- Key verification, for items that are key-scored,
- Rubric validation, for machine-scored items that are rule-based or heuristic based,
- Rangefinding, for essays and other hand-scored items, and

- Data review, for items that failed standard flagging criteria.

Each process is discussed below.

### **3.7.1 Key Verification**

Key verification is a simple process by which a table of response frequencies and the scores they received is created. These are reviewed by qualified CAI content staff to ensure that all correct responses, and only correct responses, receive a score.

### **3.7.2 Rubric Validation**

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

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

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

Figure 1 shows some features from REVISE.

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

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

Figure 1: Features of the REVISE Software

The screenshot displays the REVISE software interface, which is used for Rubric Evaluation and Verification for Items Scored Electronically. The interface includes a top navigation bar with tabs for Item List, Samples, Rubric, Summary, and Responses. The main content area is divided into several sections:

- Sample Details:** This section provides information about the sample, including the Sample Name (RV Sample), Sample Details, and Sample Create Date (5/25/2017 3:12:05 PM). It also includes a table of Rule Short Names, Rule Descriptions, and Number of Responses.
- Responses:** This section lists responses in the sample, including Mark as Reviewed, Original Score, Processed Score, Current Score, Processed Score, Response ID, Sample Score, and Sample Score.
- Test Item:** This section displays the actual test item, including the item number (17185), the item description, and the item content. The item content includes a table showing the relationship between time and distance for a plane traveling at a constant speed.
- Student Response:** This section displays the actual student response, including the response text (570d / 1r) and the response score (0).

Annotations highlight key features of the software:

- Users can automatically draw samples according to a variety of sample designs. Revisions to the rubric can be checked against the original sample and independent samples.
- Responses in the sample are listed here.
- The committee records its comments and consensus score here.
- Users can see the actual test item here.
- Users can see the actual student response here.

### 3.7.3 Rangefinding

Items requiring hand-scoring undergo a committee process called *rangefinding*, which engages educators and content experts in interpreting the rubric and selecting exemplars that will be used to train and validate hand-scoring. Volume 4 addresses rangefinding in more detail; it is referenced here as part of the natural sequence of item development.

### 3.7.4 Data Review

Volume 1 of this technical report describes in detail the statistical flags that send items to data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members were taught to interpret these flags and given guidelines for examining the items for content or fairness issues.

## 4. ILEARN BLUEPRINTS AND STATE ASSESSMENT TEST CONSTRUCTION

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

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

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

### 4.1 TEST BLUEPRINTS

#### 4.1.1 Blueprint Construction Meeting

In February 2018, IDOE and CAI worked closely with Indiana educators to create blueprints that guided the item development process for all subjects and grades. More details can be found in Volume 2 of the 2018-2019 *ILEARN* Technical reports.

#### 4.1.2 ILEARN Test Specifications

Test blueprints provided the following guidelines:

- Length of the assessment;
- Content areas to be covered and the acceptable number of items across standards within each content area or reporting category;
- Number of hand-scored items; and
- Approximate number of field-test items.

Table 14 presents the number of operational or operational field-test hand-scored items per form. Note that in ELA and Mathematics, all PTs included one or more hand-scored items. In Science, most of the PTs included one hand-scored interaction. Additionally, Indiana educators were invited to participate in the hand-scoring of these items in a partnership with Measurement Incorporated (MI).

*Table 14: Number of Hand-Scored Items by Form*

Subject	# of Operational Writing Prompts	# of Additional Operational or Operational Field-Test Hand-Scored Items	Comments
ELA	1	3	There were no embedded field-test hand-scored items.
Mathematics	n/a	3	Each form included up to two embedded field-test hand-scored items.
Science	n/a	2	Each form included up to two embedded field-test hand-scored items.
Social Studies	n/a	2	Each form included up to two embedded field-test hand-scored items.
U.S. Government	n/a	n/a	There were no field-test hand-scored items.

In addition to operational and non-operational field-test items, each form included embedded field-test (EFT) items. It is important to note that DOK ranges were not included in the blueprints because each IAS includes a target DOK. Other than U.S. Government, all IAS target DOK values were determined during the *ISTEP+* administrations. Table 15 denotes the number of EFT items per form.

*Table 15: Number of Embedded Field-Test Items by Form*

Subject	Grade or Course	# of EFT Items per form
ELA	All	8
Mathematics	All	5
Science	Grades 4 and 6	10
Science	Biology	5
Social Studies	Grade 5 and U.S. Government	5

Note that ELA EFT items were divided between the non-text-to-speech (non-TTS) (Reporting Categories 1 and 2) and TTS (Reporting Category 3, Speaking and Listening and Reading Foundations, grade 3). Similarly, in Mathematics grades 6 through 8, EFT items were divided between the non-calculator and calculator segments.

The Spring 2020 online *ILEARN* ELA and Mathematics assessment forms included slots for embedded field testing as well as linking items to establish the link between MetaMetrics Lexile and Quantile scales. Lexile and Quantile anchor items were stand-alone items and were randomly distributed in field-test slots along with the true field-test items.

Table 16 through Table 19 provide the percentage of operational items required in the blueprints by reporting category, for each grade level or course. The percentages below represent an acceptable range of item counts.

*Table 16: Blueprint Percentage of Test Items Assessing Each Reporting Category in ELA*

Grade	Key Ideas and Textual Support/ Vocabulary	Structural Elements and Organization/Connection of Ideas/ Media Literacy	Writing	Speaking and Listening	Reading Foundations
3	33—44%	28—35%	33—41%	6—9%	0—6%
4	31—41%	31—41%	33—41%	6—9%	n/a
5	31—41%	31—41%	33—41%	6—9%	n/a
6	29—39%	29—39%	34—42%	6—9%	n/a
7	29—39%	29—39%	34—42%	6—9%	n/a
8	29—36%	29—36%	34—42%	6—9%	n/a

*Table 17: Blueprint Percentage of Test Items Assessing Each Reporting Category in Mathematics*

Grade	Reporting Category				
	Algebraic Thinking and Data Analysis	Computation	Geometry and Measurement	Number Sense	Process Standards
3	19—24%	23—28%	19—24%	23—28%	8—13%
4	19—24%	23—28%	19—24%	23—28%	8—13%
	Algebraic Thinking	Computation	Geometry and Measurement, Data Analysis, and Statistics	Number Sense	Process Standards
5	20—26%	22—28%	18—23%	22—28%	8—13%
	Algebra and Functions	Computation	Geometry and Measurement, Data Analysis, and Statistics	Number Sense	Process Standards



Grade	Reporting Category				
6	23—28%	21—26%	19—24%	21—26%	8—13%
	<b>Algebra and Functions</b>	<b>Data Analysis, Statistics, and Probability</b>	<b>Geometry and Measurement</b>	<b>Number Sense and Computation</b>	<b>Process Standards</b>
7	23—28%	19—24%	19—24%	23—28%	8—13%
8	23—28%	21—26%	21—26%	19—24%	8—13%

*Table 18: Blueprint Percentage of Test Items Assessing Each Reporting Category in Science*

Grade	Reporting Categories				
	<b>Questioning and Modeling</b>	<b>Investigating</b>	<b>Analyzing, Interpreting, and Computational Thinking</b>	<b>Explaining Solutions, Reasoning, and Communicating</b>	
4	25—29%	25—29%	21—25%	21—25%	
6	21—25%	21—25%	25—29%	25—29%	
	<b>Developing and Using Models to Describe Structure and Function</b>	<b>Developing and Using Models to Explain Processes</b>	<b>Analyzing Data and Mathematical Thinking</b>	<b>Constructing and Communicating an Explanation</b>	<b>Evaluating Claims with Evidence</b>
Biology	18—22%	18—22%	18—22%	18—22%	18—22%

*Table 19: Blueprint Percentage of Test Items Assessing Each Reporting Category in Social Studies*

Grade	Reporting Categories		
	<b>Civics and Government</b>	<b>Geography and Economics</b>	<b>History</b>
5	38—43%	28—33%	28—33%
	<b>Functions of Government</b>	<b>Historical Foundations of American Government</b>	<b>Institutions and Processes of Government</b>
U.S. Government	35—39%	24—28%	35—39%

### **4.1.3 ELA Blueprints**

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

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

- In grades 3-5: Two nonfiction reading passages with associated items and two literary reading passages with associated items;
- In grades 6-8: Three nonfiction reading passages with associated items and one literary reading passage with associated items;
- Two to three speaking and listening items;
- Stand-alone writing and/or research items; and
- One PT which includes two “precursor” items leading up to a text-based writing task.

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

### **4.1.4 Mathematics Blueprints**

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

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

### **4.1.5 Science Blueprints**

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

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

### **4.1.6 Social Studies Blueprints**

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

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

## **4.2 TEST FORM CONSTRUCTION**

During Fall 2019, CAI psychometricians and content experts worked with IDOE to build forms for the Spring 2020 administration. *ILEARN* assessment test form construction utilized test construction guidelines, explicit blueprints, and collaborative participation from all parties. The Spring 2020 *ILEARN* test forms were built by CAI test developers to match exactly the detailed test blueprint and target distributions of item difficulty and assessment information when information was available and to the extent possible.

Item parameters based on separate, item bank-specific calibrations are on different item response theory (IRT) scales and are not directly comparable. Thus, when items from separate pools combine on a single form, some typical test construction summaries must be modified or are not applicable. In ELA and Mathematics, the existing Smarter IRT item parameters and vertical scales were used. For Science and Social Studies, new scales were established.

For the online ELA, Mathematics, and Science computer-adaptive test (CAT), item pools of available items were used, and there was no single test form constructed. For online Social Studies and all paper assessments, a single fixed form was constructed. The operational items were selected to represent the blueprint for that grade and subject. The subsequent sections outline the roles and responsibilities of the participants, test construction process, materials used, and sample statistical and graphical summaries used during the review process.

While blueprints describe the content to be covered and other content-relevant aspects of the assessment, other considerations exist. The psychometric considerations, ensuring that students will receive scores of similar precision, include the following:

- A reasonable range of item difficulties was present;
- $p$ -values for items were reasonable and within specified bounds ( $> 5\%$  and  $< 95\%$ );
- Biserial correlations were reasonable and within specified bounds;
- For all items, IRT  $a$ -parameters were reasonable; and
- For all items, IRT  $b$ -parameters were reasonable, with the range dependent on the scale.

More information about  $p$ -values, biserial correlations, and IRT parameters can be found in Volume 1 of this technical report. The details on calibration, equating, and scoring of the *ILEARN* can also be found in Volume 1.

Using Fixed-Form Builder, a test form-building tool, CAI test developers selected items appropriately aligned to the IAS from the *ILEARN* item bank that met the various test blueprint requirements and statistical targets. Once the form was created to meet the blueprint and statistical criteria, the items were rearranged to reflect the order in which they would be presented on the assessment, following the procedures described in Section 4.3, Test Form Assembly.

### 4.3 TEST FORM ASSEMBLY

Test form assembly integrates the skills of psychometricians and content experts. Each form must measure the same construct with similar precision. For fixed-form tests, the statistical criteria try to ensure that the construct is measured with items of similar difficulty and discrimination across years. This review will ensure that new forms match the information curve and test characteristic curves from the Spring 2019 first-year form.

The *ILEARN* forms were created using CAI's standard process. Content specialists work with a tool that:

- guides them in selecting items needed to meet the test blueprint, and
- graphically presents statistical information, helping them form tests that meet the statistical criteria in the first draft.

Draft forms are reviewed by senior test developers for adherence to blueprints, possible cueing issues, and balance in terms of item types.

Upon passing the internal content reviews, the forms are passed to psychometricians, where experts review more detailed technical output from Form Analyzer. This software provides a detailed statistical summary of the forms. The Form Analyzer tool is a web-based component of the test construction suite that provides real-time information about test forms as they are constructed by content development teams. As test developers input items to satisfy a specific blueprint, Form Analyzer provides psychometric teams with psychometric characteristics of the form and compares those statistical characteristics to a previously developed form to ensure that new forms are statistically parallel to prior forms. Specifically, Form Analyzer provides the following information when constructing test forms:

- Test characteristics curves for the new form overlaid with a prior reference form;
- Standard error of measurement curves for the new form overlaid with a prior reference form;
- Test characteristics curve differences between current and reference form;
- Statistical summary of current and reference form, including:
  - Classical item statistics (e.g.,  $p$ -value, biserials),
  - IRT-based statistics,
  - Individual item-level statistics; and
- Real-time blueprint satisfaction reports updated as items are added to the forms.

In year 1, the first three bullets were not reviewed as no reference form existed. Statistical summaries under bullet 4 were calculated and compared only to guideline specifications as no reference form existed. For example,  $p$ -values were reviewed so that no items with extreme values (e.g., less than 0.05) were used, but there was no comparison for overall item  $p$ -values to reference forms.

## 4.4 ROLES AND RESPONSIBILITIES

### 4.4.1 Role of the CAI Content Team

CAI content teams were responsible for the initial form construction and subsequent revisions. They performed the following tasks:

- Selection of the operational items;
- Revision of the operational item sets according to feedback from senior CAI content staff;
- Revision of the operational item sets according to feedback from the CAI technical team;

- Revision of the operational item sets according to feedback from IDOE;
- Assistance in the generation of materials for IDOE review; and
- Revision of the forms to incorporate feedback from IDOE.

#### **4.4.2 Role of the CAI Technical Team**

The CAI technical team, which includes psychometricians and statistical support associates, prepares the item bank by updating ITS with current item statistics and provides test construction training to the internal content team. The technical team performs the following tasks:

- Preparation of item bank statistics and updating of CAI's ITS;
- Creation of the master data sheets (MDS) for each grade and subject;
- Providing feedback on the statistical properties of initial item selections;
- Providing feedback on the statistical properties of each subsequent item selection; and
- Assisting in the generation of materials for IDOE review.

#### **4.4.3 Role of IDOE**

The IDOE team, which includes the Assessment Director, Assistant Assessment Director, and content specialists, previews proposed test forms and provides feedback. IDOE performs the following tasks:

- Review of proposed test forms; and
- Final approval of all test forms.

### **4.5 TARGET GUIDELINES**

During test construction of Spring 2020 operational forms, the Spring 2019 operational forms were used as the reference curve and statistical targets. In addition, the statistical targets for the forms were set by choosing items that met general guidelines (e.g., no extreme  $p$ -values).

### **4.6 ACCOMMODATED FORM CONSTRUCTION**

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

During test development, forms across all modes were required to adhere to the same test blueprints, content-level, and psychometric considerations. The online and accommodated forms were then reviewed for their comparability of item counts, both at

the overall test level and at the reporting category levels. ELA assessments in both administration modes were additionally compared for the distribution of passages by length. The forms were then submitted for psychometric reviews, during which the following statistics were computed and compared between the online and paper-and-pencil accommodated forms where possible given the various item sources and differing scales of the item pools:

- IRT  $b$ -parameter (difficulty) mean and standard deviation;
- IRT  $b$ -parameter minimum and maximum;
- IRT  $a$ -parameter mean and standard deviation;
- IRT  $a$ -parameter minimum and maximum;
- Item  $p$ -value mean and standard deviation;
- Item  $p$ -value minimum and maximum; and
- Lowest bi/polyserial.

A sample output with summary statistics for grade 5 Social Studies is presented in Table 20. As the table shows, the IRT  $b$ -parameter (difficulty) mean and the item  $p$ -value mean are similar between the forms.

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

*Table 20: Statistical Test Summary Comparison for Grade 5 Social Studies Online and Paper Forms*

Type	Statistics	Online Form	Paper Form
Overall	Number of Items	40	40
	Possible Score	42	42
	Difficulty Mean	0.18	0.13
	Difficulty StDev	1.02	0.89
	Difficulty Minimum	-1.21	-2.21
	Difficulty Maximum	4.04	2.06
	Parameter-A Mean	0.56	0.53
	Parameter-A StDev	0.24	0.21
	Parameter-A Minimum	0.19	0.19
	Parameter-A Maximum	1.19	0.97
	P-Value Mean	0.50	0.50

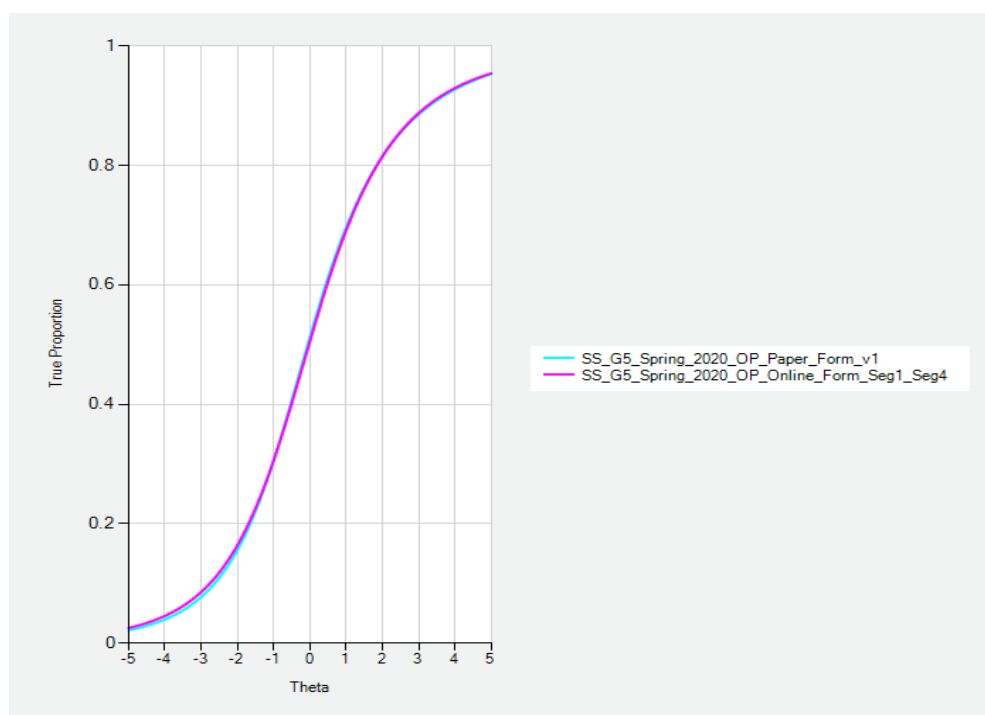
Type	Statistics	Online Form	Paper Form
	P-Value StDev	0.14	0.13
	P-Value Minimum	0.09	0.28
	P-Value Maximum	0.75	0.86
	Lowest Bi/Poly-Serial	0.22	0.25

### 4.6.1 Test Characteristic Curve

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

Items were selected for the paper form so that the form TCC matched the regular online form TCC as closely as possible. Figure 2 compares the TCCs for both online and paper forms of grade 5 Social Studies. Appendix C of Volume 1 provides the TCC for all administered assessments.

*Figure 2: TCC Comparisons of Grade 5 Social Studies Online and Paper Forms*





## 4.6.2 Test Characteristic Curve Difference

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 3 illustrates a sample TCC difference, which allows us to evaluate the degree to which the parallelism is achieved between the forms.

## 4.6.3 Conditional Standard Error of Measurement Curve

The CSEM curve shows the level of error of measurement expected across the range of student ability, and the Form Analyzer tool allows test developers to compare the statistical comparability of multiple forms simultaneously. The example in Figure 4 superimposes two CSEM curves onto one plot so that test developers can view the degree to which the two test forms are statistically parallel, and this is provided as an example of how test developers use the CSEM curves when building forms.

*Figure 3: TCC Differences of Grade 4 Science Online and Accommodated Forms*

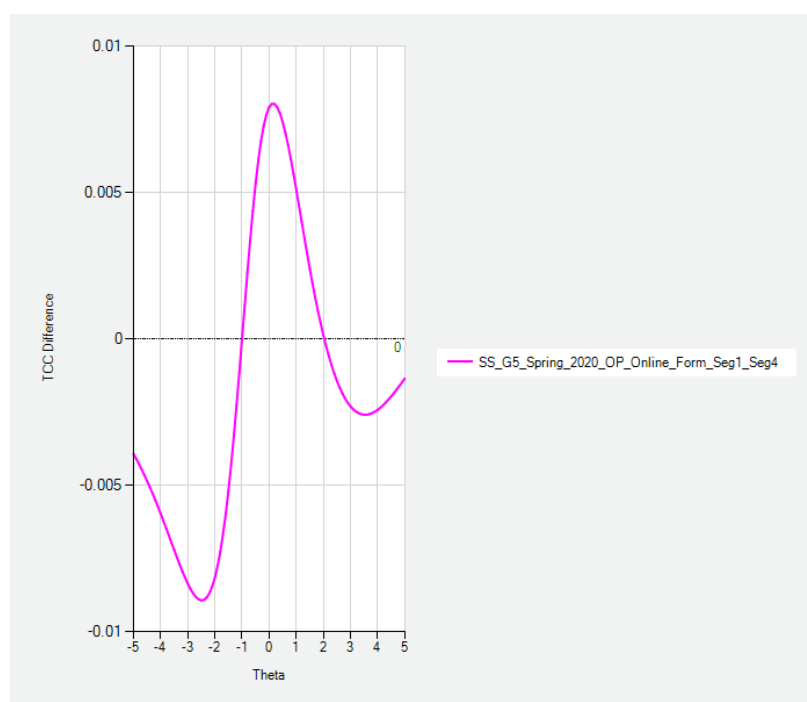
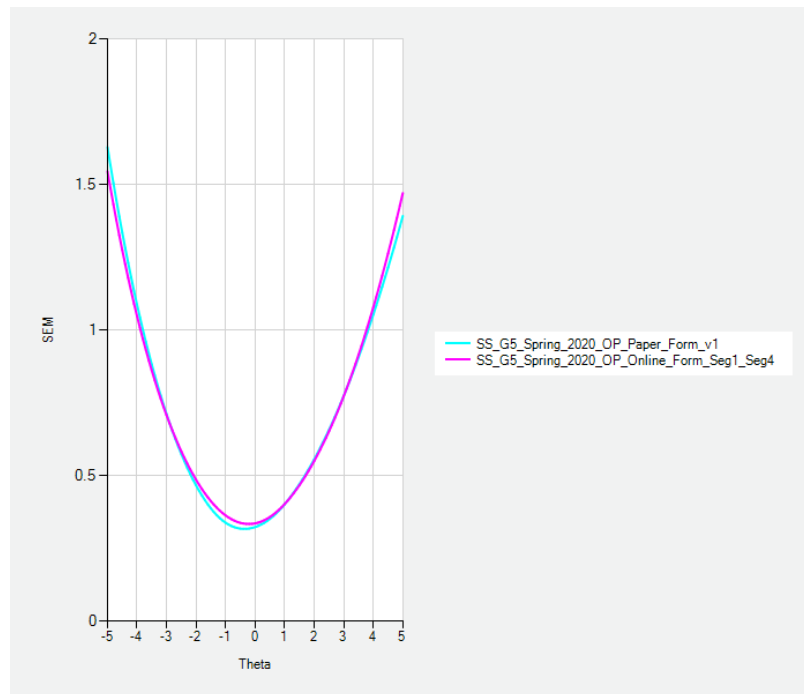


Figure 4: CSEM Comparisons of Grade 4 Science Online and Accommodated Forms



## **5. PERFORMANCE LEVEL DESCRIPTORS**

The Indiana Department of Education (IDOE) held a meeting with Indiana educators in June 2018 to develop performance level descriptors (PLDs). The main purpose of the meeting was for educators to develop Policy and Range PLDs for each grade and content area and recommend proficiency level names to be used for reporting following their review of the policy PLDs.

PLDs describe levels of achievement or categories of performance on a large-scale assessment. PLDs are used to inform the evidence required for item development, inform items selected during the form construction process, and support standard setting panelist recommendations during the standard setting process. They are then ultimately used to inform stakeholder interpretation of student scores once standards are set. The focus of the June 2018 meetings was on Policy and Range PLDs.

After the June 2018 educator workshop, CAI and IDOE revised the PLDs based on feedback from the policy review panel. CAI worked with IDOE to edit the Range PLDs for consistency of format, language, and grammar, prior to finalizing the documents for presentation to the Indiana State Board of Education (SBOE). The Range PLDs approved by this body were then posted to the IDOE website.

More information about the PLD meeting can be found in Volume 2 of the 2018-2019 *ILEARN* Technical Report.

## 6. REFERENCES

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*.
- Calisir, F., & Gurel, Z. (2003). Influence of text structure and prior knowledge of the learner on reading comprehension, browsing and perceived control. *Computers in Human Behavior*, 19(2), 135–145.
- Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance-level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York: Routledge.
- Fisher, D., Frey, N., & Lapp, D. (2012). *Text complexity: Raising rigor in reading*. Newark, DE.: International Reading Association.
- Freebody, P., & Anderson, R. C. (1983). Effects on Text Comprehension of Differing Proportions and Locations of Difficult Vocabulary. *Journal of Reading Behavior*, 15(3), 19–39.
- Gillioz, C., Gygax, P., & Tapiero, I. (2012). Individual differences and emotional inferences during reading comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66(4), 239–250.
- Kucer, S. B. (2010). Going beyond the author: What retellings tell us about comprehending narrative and expository texts. *Literacy*, 45(2), 62–69.
- Long, D. L., & De Ley, L. (2000). Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language*, 42(4), 545–570.
- McConaughy, S. (1985). Good and Poor Readers' Comprehension of Story Structure Across Different Input and Output Modalities. *Reading Research Quarterly*, 20(2), 219–232. doi:10.2307/747757.
- Rapp, D. N., & Mensink, M. C. (2011). Focusing effects from online and offline reading tasks. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 141–164). Charlotte, NC, US: IAP Information Age Publishing.
- Rich, S. S., & Taylor, H. A. (2000). Not all narrative shifts function equally. *Memory & Cognition*, 28(7), 1257–1266.
- Riding, R. J., & Taylor, E. M. (1976). Imagery performance and prose comprehension in seven-year-old children. *Educational Studies*, 2(1), 21–2.

- Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, 25(5), 762–776.
- Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). A causal model of sentence recall: Effects of familiarity, concreteness, comprehensibility, and interestingness. *Journal of Reading Behavior*, 25(1), 5–16.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal*, 95(1), 26–43.
- Sparks, J. R., & Rapp, D. N. (2011). Readers reliance on source credibility in the service of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1), 230–247.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 15, 2012, from <http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis44.html>.
- Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.



**Indiana Learning Evaluation  
and Readiness Network  
(ILEARN)**

**2019-2020**

**Volume 3  
Test Administration**

## **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](mailto:inassessments@doe.in.gov).

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Elizabeth Ayers-Wright, Xiaoxin Wei, Tracie Morris, Suzanne Huston, Kevin Clayton, and Kyra Bilenki. Major contributors from the Indiana Department of Education include the Assessment Director, Assistant Assessment Director, and Program Leads.

.

## TABLE OF CONTENTS

1. INTRODUCTION.....	1
2. TESTING PROCEDURES AND TESTING WINDOWS.....	2
2.1 Eligible Students .....	3
2.2 Testing Accommodations.....	4
2.3 Available Accommodations.....	6
3. ADMINISTRATOR TRAINING.....	7
3.1 Online Administration.....	7
3.2 Test Administration Resources .....	9
4. TEST SECURITY PROCEDURES .....	13
4.1 Security of Test Materials .....	13
4.2 Identifying Test Irregularities or Potential Test Security Concerns.....	15
4.3 Tracking and Resolving Test Irregularities.....	15
4.4 CAI's System Security .....	17
REFERENCES .....	18

## LIST OF TABLES

Table 1: Designated Features and Accommodations Available in Spring 2019.....	4
Table 2: User Guides and Manuals .....	10
Table 3: Examples of Test Irregularities and Test Security Violations .....	16



## LIST OF APPENDICES

- Appendix A: *Online Test Delivery System (TDS) User Guide*
- Appendix B: *Technology Setup for Online Testing Quick Guide*
- Appendix C: *2018–2019 Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux*
- Appendix D: *Indiana Online Practice Test User Guide*
- Appendix E: *Test Information Distribution Engine User Guide*
- Appendix F: *Braille Requirements Manual for Online Testing*
- Appendix G: *Online Reporting System User Guide*
- Appendix H: *Indiana Accessibility and Accommodations Guidance Manual*
- Appendix I: *ILEARN ISR Interpretive Guide*
- Appendix J: *Accessibility and Accommodations Implementation and Setup Module*
- Appendix K: *Indiana Assessments Policy Manual*
- Appendix L: *ILEARN Biology Test Administrator’s Manual*
- Appendix M: *ILEARN Test Coordinator’s Manual*
- Appendix N: *Educator Brochure and Graphics*
- Appendix O: *Understanding Indiana’s New Assessment System Webinar Module*
- Appendix P: *Released Items Repository Quick Guide*
- Appendix Q: *Computer-Adaptive Tests Webinar Module*
- Appendix R: *Why It Is Important to Assess Webinar Module*
- Appendix S: *Test Administrator Training Webinar Module*
- Appendix T: *Request an Item Rescore Webinar Module*
- Appendix U: *Parent Brochure*
- Appendix V: *Test Administration Overview Webinar Module*
- Appendix W: *Test Information Distribution Engine (TIDE) Webinar Module*
- Appendix X: *Test Delivery System (TDS) Webinar Module*
- Appendix Y: *Online Reporting System (ORS) Webinar Module*
- Appendix Z: *Technology Requirements for Online Testing Webinar Module*
- Appendix AA: *How the Scoring Process Works Webinar Module*
- Appendix AB: *Test Administrator Certification Course Storyboard*

## 1. INTRODUCTION

The State of Indiana implemented a new online assessment for operational use beginning with the 2018–2019 school year. This new assessment program, referred to as the ILEARN assessments, replaced Indiana Statewide Testing for Educational Progress-Plus (ISTEP+). ILEARN comprises English/Language Arts (ELA) and Mathematics assessments in grades 3–8. Science is administered in grades 4 and 6, and Biology is administered in high school. Social Studies is administered in grade 5, and U.S. Government is administered in high school. The U.S. Government assessment is optional. The ELA, Mathematics, and Science assessments are computer-adaptive tests (CATs), and the Social Studies and U.S. Government tests are fixed-form online assessments. The ELA, Mathematics, and Science assessments consist of a non-performance task segment and a performance task segment. Students needed to complete the non-performance task segment of the test to receive their final overall scale score and both the non-performance task segment and the performance task segment to receive an overall scale score and reporting category level scores.

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

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

### 1.1 COVID-19 CONSIDERATIONS

The spring 2020 administration of *ILEARN* was cancelled due to the novel Coronavirus (COVID-19) pandemic. As a result, test summaries in the 2019-2020 technical reports are based on the fall 2019 and winter 2020 end-of-course Biology administrations which were successfully administered prior to COVID-19.

## 2. TESTING PROCEDURES AND TESTING WINDOWS

Administering the 2019-2020 ILEARN Biology assessments required coordination, detailed specifications, and proper training. In addition, several individuals in each corporation and school were involved in the administration process, from those setting up secure testing environments to those administering the tests. Without the proper training and coordination of these individuals, the standardization of the test administration could have been compromised. IDOE worked with CAI to develop and provide the training and documentation necessary for the administration of ILEARN under standardized conditions within all testing environments, both online and on paper-and-pencil tests.

All students were required to take a practice test at their school prior to taking the 2019-2020 ILEARN Biology assessments. These practice tests contained sample test items similar to the test items that students would encounter on the ILEARN assessments to help students become familiar with the item types that would be presented to the students on the online or paper-and-pencil assessments. Indiana students also had the opportunity to interact with released, non-secure items on a public-facing [Released Items Repository](#) (RIR) assessments available on the [ILEARN portal](#). The ILEARN RIR was deployed in October 2019, which allowed students to have online access to the items for two months prior to the opening of the first test administration.

The ILEARN assessments for the fall and winter Biology test windows were administered in multiple segments over multiple days. The test segments administered were as follows:

- Science: non-performance task CAT segment and a performance task segment.

The ILEARN assessments were untimed, but timing estimates were included in the ILEARN Biology Test Administrator's Manual (TAM) (Appendix L in this volume) to ensure that schools had resources available to create local testing schedules. The ILEARN testing window for grades 3–8 was scheduled to be available from April 20 through May 15, 2020 and was cancelled on April 7, 2020, as a result of COVID-19. The fall Biology test was available from December 2 through December 19, 2019, and the winter Biology test was available February 10 through February 27, 2020. The spring Biology and U.S. Government tests were scheduled to be available from April 20 through May 22, 2020 but were also canceled as a result of COVID-19.

## 2.1 ELIGIBLE STUDENTS

All students enrolled in Biology were required to participate in the 2019-2020 ILEARN Biology administrations with or without accommodations. The spring administrations were canceled due to COVID-19 and a waiver was granted by the United States Department of Education that exempted Indiana students from having to take state assessments in Spring 2021. Students took the fall or winter Biology ECA where completion of the respective high school course coincided with one of those test windows. Section 1111(b)(2)(A) of the Elementary and Secondary Education Act of 1965 (as amended by the Every Student Succeeds Act [ESSA]) requires the implementation of high-quality student academic assessments in Mathematics, Reading or Language Arts, and Science.

- **Public and Private School Students.** Students enrolled in Indiana public, charter, accredited nonpublic, and Choice schools were required to participate in course-level appropriate ILEARN assessment(s).
- **English Learners (ELs).** All ELs enrolled in tested courses were expected to participate in all ILEARN assessments, including English/Language Arts, regardless of how long these students had been enrolled in a U.S. school. Mathematics, Science, and Social Studies assessments are available in stacked Spanish in the online Test Delivery System (TDS). Stacked Spanish is represented on the screen with the stimuli, passage, and item all appearing in both Spanish and English for students whose test setting language is Spanish.
- **Students with Disabilities.** Indiana has established procedures to ensure the inclusion in statewide testing of all public elementary and secondary school students with disabilities. Federal and state laws require that all students participate in the state testing system. In Indiana, a student on an Individualized Education Program (IEP) participates under one of these three general options:
  1. ILEARN without accommodations;
  2. ILEARN with approved accommodations; or
  3. Indiana Alternate Measure (I AM) Alternate Assessment.

Per the Individuals with Disabilities Education Improvement Act (IDEA) and Title 511 Article 7-Special Education, published December 2014 by the Indiana State Board of Education, decisions regarding which assessment option a student will participate in are made annually by the student's IEP team and 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.

If a student requested an extraordinary exemption option due to a medical complexity, he or she may have been exempt from participating in statewide, standardized assessments

pursuant to the provisions of School Accountability, a letter requesting the exemption is required.

## 2.2 TESTING ACCOMMODATIONS

Students participating in the online ILEARN assessment were able to use the designated standard online testing features in the TDS. These features included the ability to select an alternate background and font color, mouse pointer size and color, and font size before testing. During the tests, students could 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.

All Indiana state assessments have appropriate accommodations available to make these options accessible to students with disabilities and ELs, including ELs with disabilities. Accommodations were provided to students with disabilities enrolled in public schools with current IEPs or Section 504 Plans, as well as to students identified as ELs.

The accommodations available for eligible students participating in the ILEARN assessments are described in the test administrator's manual (TAM) (Appendix L of this report volume), which were accessible to schools before and during testing in the [Resources](#) section of the [ILEARN Portal](#).

The ILEARN assessments provide two categories of assessment features to students. These include designated features and accommodations, both embedded and non-embedded in the TDS. Table 1 provides a list of designated features and accommodations that were offered in the 2019-2020 administration. Designated features for the ILEARN are those supports that 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). The *Online Test Delivery System (TDS) User Guide* at the ILEARN portal (Appendix A of this report volume) provides instructions on how to access and use these features.

*Table 1: Designated Features and Accommodations Available in 2019-2020 for ILEARN Biology*

Designated Features	Accommodations
<b><i>Embedded</i></b>	
Color contrast (Onscreen)	Permissive Mode
Glossaries (Language)	Print on Demand
Spanish	Streamline
Masking	Text-to-Speech
Mouse pointer	Refreshable Braille
Print size	
Translation Stacked Spanish	

**Non-Embedded**

Assistive technology to Magnify/Enlarge	Paper Booklet
Access to Sound Amplification Program	Large Print Booklet
Special Furniture or Equipment for Viewing	Read-Aloud to Self
Test	Read-Aloud Script for Paper Booklet
Special Lighting Conditions	Scribe
Time of Day for Testing Altered	Speech-to-Text
Color Acetate Film for Paper Assessments	Tested Individual
	Interpreter for Sign Language
	Braille Booklet
	Additional Breaks
	Bilingual Word-to-Word Dictionary
	Spanish Booklet
	Calculator
	Multiplication Table

Non-standard accommodation requests were recorded under a Special Requests section in the Test Information Distribution Engine (TIDE). These special requests required IDOE approval.

Students who required online accommodations (e.g., text-to-speech) were provided the opportunity to participate in practice activities for the statewide assessments with appropriate allowable accommodations. Test administrators had to identify 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 the test.

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 Test Administrator (TA) in his or her required administration information and captured by CAI in the database of record (DoR). CAI included this data in the state output student data score files (SDFs) provided to IDOE at the end of each test administration.

Guidelines recommended for making accommodation decisions included the following:

- Accommodations should facilitate an accurate demonstration of what the student knows or can do.
- Accommodations should not provide the student with an unfair advantage or negate the validity of a test; accommodations must not change the underlying skills that are being measured by the test.
- Accommodations must be the same or nearly the same as those needed and used by the student in completing daily classroom instruction and routine assessment activities.
- Accommodations must be necessary for enabling the student to demonstrate knowledge, ability, skill, or mastery.

Students with disabilities not enrolled in public schools or receiving services through public school programs who required accommodations to participate in a test

administration were permitted access to accommodations if the following information was provided:

- Evidence that the student had been found eligible as a student with a disability as defined by Individuals with Disabilities Education Improvement Act (IDEA).
- Documentation that the requested accommodations had been regularly used for instruction.

## **2.3 AVAILABLE ACCOMMODATIONS**

The TA and the School Test Coordinator (STC) were responsible for ensuring that arrangements for accommodations had been made before the test administration dates. As a supplement to the TAMs, IDOE provided a separate accessibility manual, the *Indiana Assessments Policy Manual* (Appendix K of this report volume) for individuals involved in administering tests to students who required accommodations.

The following accommodations were available for eligible students with IEPs or Section 504 Plans participating in paper-based assessments:

- Contracted UEB braille and UEB Nemeth for Math.

For eligible students with IEPs, Section 504 Plans, or Individual Learning Plans participating in online assessments, a comprehensive list of accommodations is given in the *Test Information Distribution Engine (TIDE) User Guide* (Appendix E of this report volume).

The accommodation guidelines provide information about the tools, supports, and accommodations that are available to students taking the Science assessments. For further information, please refer to the *Indiana Assessments Policy Manual* (Appendix K of this report volume).

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

### 3. ADMINISTRATOR TRAINING

IDOE established and communicated a clear, standardized procedure to educators and key personnel involved with administration of ILEARN assessments, including the process for giving students access to accommodations. Key personnel involved with ILEARN administration included Corporation Test Coordinators (CTCs), Non-Public School Test Coordinators (NPSTCs), Corporation Information Technology Coordinators (CITCs), STCs, and TAs. The roles and responsibilities of staff involved in testing are further detailed in the next section.

TAs were required to complete the online CAI TA Certification Course before administering the test. 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 the appendices to this volume of the technical report.

Test administrator manuals and guides were available online for school and corporation staff. The *Online Test Delivery System (TDS) User Guide* (Appendix A of this report volume) was designed to familiarize TAs with the TDS and contained tips and screenshots throughout the text. The user guide described:

- Steps to take prior to accessing the system and logging in;
- Navigation instructions for the TA Interface application;
- Details about the Student Interface, used by students for online testing;
- Instructions for using the training sites available for TAs and students; and
- Information on secure browser features and keyboard shortcuts.

The User Support sections of both the *Online Test Delivery System (TDS) User Guide* (Appendix A of this report volume) and the *Test Information Distribution Engine (TIDE) User Guide* (Appendix E of this report volume) 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.

#### 3.1 ONLINE ADMINISTRATION

The *Online Test Delivery System (TDS) User Guide* (Appendix A of this report volume) provided instructions for creating test sessions; monitoring sessions; verifying student information; assigning test accommodations; and starting, pausing, and submitting tests. The *Technology Setup for Online Testing Quick Guide* (Appendix B of this report volume) provided information about hardware, software, and network configurations to run CAI's various testing applications.

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



## **Roles and Responsibilities in the Online Testing Systems**

CTCs, NPSTCs, STCs, and TAs each had specific roles and responsibilities in the online testing systems. See the *Online Test Delivery System (TDS) User Guide* (Appendix A of this report volume) for their specific responsibilities before, during, and after testing.

### **CTCs**

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

### **CITCs**

CITCs were responsible for ensuring that testing devices were properly configured to support testing and for coordinating participation in the 2019-2020 systems readiness test (SRT). All schools were required to complete the SRT to prepare for online testing. The SRT was a simulation of online testing at the state level that ensured student testing devices and local school networks were correctly configured to support online testing.

### **NPSTCs**

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

### **STCs**

Before each administration, STCs and CTCs were required to verify that student eligibility was correct in TIDE, and that any accommodations or test settings were correct. To participate in a computer-based online test, students had to be listed as eligible for that test in TIDE. See the *Test Information Distribution Engine (TIDE) User Guide* (Appendix E of this report volume) 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 worked with technology coordinators to ensure that computers and devices were prepared for testing and technical issues were resolved to ensure a smooth testing experience for the students. During the 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**

TAs administered the ILEARN assessment to students as well as a practice test session prior to the assessment.

TAs were responsible for reviewing necessary user manuals and user guides to prepare the testing environment and ensure that students did not have unauthorized books, notes, scratch paper, or electronic devices. They were required to administer the ILEARN

assessment according to the directions found in the guide. TAs were required to report to the STC any deviation in test administration, at which time the STC was required to report it to the CTC. Then, if necessary, the CTC was to report it to IDOE. TAs also ensured that the only available resources accessible to students were those allowed for specific ILEARN test administrations.

## 3.2 TEST ADMINISTRATION RESOURCES

The list of webinars and training resources available to corporations and schools for the 2019-2020 ILEARN administration is provided below. All training materials were available online at the [ILEARN Portal](#). (PDFs of these resources have also been included in this technical report as Appendices J, O, Q–T, and V–AB, respectively.) Test administration resources comprising various tutorials and documents (user guides, manuals, quick guides, etc.) were available through the [ILEARN Portal](#).

- **Test Administrator Certification Course:** All educators who administered the ILEARN assessment were required to complete an online TA Certification Course.
- **Accessibility and Accommodations Implementation and Setup Module:** This online module provided information on accessibility and accommodations in Indiana for the ILEARN tests.
- **Understanding Indiana's New Assessment System Webinar Module:** This online module provided an overview of the new ILEARN assessment to prepare parents, educators, and administrators for what to expect from the assessments.
- **Computer-Adaptive Tests Webinar Module:** This online module described computer-adaptive-testing and the student test experience.
- **Why It Is Important to Assess Webinar Module:** This online module illustrated the importance of statewide testing.
- **Test Administrator Training Webinar Module:** This online module provided information and a step-by-step guide through the TA Interface in the TDS.
- **Request an Item Rescore Webinar Module:** This online module provided additional information regarding Indiana legislation that allows a principal or parent/guardian to request an item rescore for handscored items on the ILEARN tests.
- **Test Administration Overview Webinar Module:** This module provided a general overview of the TA role in the test administration process, including key responsibilities before, during, and after the testing window.
- **Test Information Distribution Engine (TIDE) Webinar Module:** This module provided a general overview of TIDE and the features applicable to educators and administrators before, during, and after testing.
- **Test Delivery System (TDS) Webinar Module:** This module provided a general overview of CAI's TDS and the features available in both the TA Interface and the Student Interface within TDS.

- **Online Reporting System (ORS) Webinar Module:** This module provided a general overview of the ORS where student scores, including individual scores and aggregate scores, are displayed after students complete the ILEARN assessments.
- **Technology Requirements for Online Testing Webinar Module:** This module provided technology requirements for corporation and school technology coordinators to ensure that their testing devices are set up properly before testing.
- **How the Scoring Process Works Webinar Module:** This module provided information for educators to better understand the scoring process tests go through prior to reporting.

Table 2 presents the list of available user guides and manuals related to ILEARN administration. The table also includes a short description of each resource and its intended use. (PDFs of these eight publications have also been included in this technical report as Appendices [A–H], respectively.)

*Table 2: User Guides and Manuals*

Resource	Description
<i>Online Test Delivery System (TDS) User Guide</i>	This user guide supports TAs who manage testing for students participating in the ILEARN practice tests, released item repository tests, and operational tests.
<i>Technology Setup for Online Testing Quick Guide</i>	This document explains in four steps how to set up technology in Indiana corporations and schools.
<i>2019-2020 Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux</i>	This manual provides information about hardware, software, and network configurations for running various testing applications provided by CAI.
<i>Indiana Online Practice Test User Guide</i>	This user guide provides an overview of the ILEARN Practice Test.
<i>Test Information Distribution Engine (TIDE) User Guide</i>	This user guide describes the tasks performed in the Test Information Distribution Engine (TIDE) for ILEARN assessments.
<i>Assistive Technology Manual</i>	This manual provides an overview of the embedded and non-embedded assistive technology tools that can be used to help students with special accessibility needs complete online tests in the Test Delivery System (TDS). It includes lists of supported devices and applications for each type of assistive technology that students may need, as well as setup instructions for the assistive technologies that require additional configuration in order to work with TDS.
<i>Online Reporting System (ORS) User Guide</i>	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.
<i>Indiana Accessibility and Accommodations Guidance</i>	The accessibility manual establishes the guidelines for the selection, administration, and evaluation of accessibility supports for instruction and assessment of all students, including students with disabilities, English learners (ELs), ELs with disabilities, and students without an identified disability or EL status.

## Department Resources and Support

In addition to the resources listed in Table 2, IDOE provided the following resources for corporations:

- Weekly newsletter distributed via email from the IDOE Office of Assessment to all officially designated CTCs in IDOE's database. The newsletter was titled "ILEARN Assessment Update" and included information on new announcements relevant to the ILEARN assessment, reminders of upcoming milestones, and a planning ahead section with important dates in the ILEARN program. The IDOE Office of Assessment contact information was also available at the end of each weekly newsletter so that corporations and schools could contact the IDOE directly if there were any questions.
- Communications via email memos took place on an "as needed" basis. These messages generally addressed specific issues that needed to be transmitted quickly to administrators and teachers in the field or important information that the IDOE wanted to ensure was clearly outlined due to its importance to the ILEARN program. An example of this was a memo the IDOE sent in fall 2019 that contained extensive information about ILEARN scheduling and timing guidance, which was intended to help schools develop their ILEARN testing schedules. The distribution was to superintendents, principals, and school leaders.
- General information about the assessments was posted on the IDOE Office of Assessment website (<https://www.doe.in.gov/assessment>), such as dates of testing windows for all state-administered assessments. The Accessibility and Accommodations Guidance in the ILEARN Policy and Guidance section of IDOE's website was often referenced to address questions pertaining to accommodations and overall accessibility.

## ILEARN Released Items Repository

The ILEARN Released Item Repository (RIR) is a collection of non-secure items and performance tasks that were available to the public via the ILEARN Portal and were intended to allow students, parents, and educators access to content similar to what the student encountered when taking the ILEARN assessment. The ILEARN RIR was deployed on October 7, 2019 and remained available throughout the testing window. A scoring guide accompanied the RIR, which provided educators the opportunity to see how their students performed on the assessment and where to focus efforts to improve student performance prior to the administration of the ILEARN assessment.

## ILEARN Practice Tests

The purpose of the practice tests was to familiarize students with the TDS functionality and item types that students would experience on the ILEARN tests. The practice tests did not contain performance tasks and were not computer-adaptive. The items provided a grade-specific testing experience, including a variety of question types. The practice

tests were not intended to guide classroom instruction. Users could also use the tutorials on each item to familiarize themselves with the different features and response instructions for each item type.

The ILEARN practice tests were deployed on October 7, 2019 and remained available throughout the testing window. The ILEARN practice tests were designed for use with the CAI Secure Browser and a supported web browser. The portal provided a list of supported web browsers on which to administer the practice tests. CAI's TDS delivered the practice tests in secure mode and used the same test delivery engine as the operational test to ensure that the student testing experience on the practice test matches the student experience for the operational test. IDOE required all students to take the practice test before taking the operational ILEARN test.

Students taking the ILEARN assessment on paper were also required to take a paper-and-pencil practice test prior to taking the operational ILEARN assessment. The practice test items were delivered to students at the beginning of the paper-and-pencil test booklets. The TA script provided specific instructions to ensure that the students completed the paper-and-pencil practice test items prior to starting the operational ILEARN assessment. A practice test answer key was included within the TA script and provided educators the opportunity to ensure that their students understood how to respond to the different question types represented on the ILEARN assessment.

## 4. TEST SECURITY PROCEDURES

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 that administered ILEARN assessments were required to complete the online TA Certification Course accessible through the [ILEARN portal](#). TDS was configured so that personnel could not administer tests without completing the TA Certification Course. Access to the course was limited to the following roles: CTC, Co-Op, CITC, NPSTC, STC, and TA.

The test security procedures for ILEARN included the following:

- Procedures to ensure security of test materials;
- Procedures to investigate test irregularities; and
- Guidelines to determine if test invalidation was appropriate/necessary.

To support these policies and procedures, IDOE leveraged security measures within CAI systems. For example, students taking the ILEARN assessments were required to acknowledge a security statement confirming their identity and acknowledging that they would not share or discuss test information with others. Additionally, students taking the online assessments were logged out of a test within the CAI Secure Browser after 20 minutes of inactivity.

In developing the *ILEARN Test Coordinator's Manual* (Appendix M of this report volume) and the ILEARN Biology TAM (Appendix L of this report volume), IDOE and CAI ensured that all test security procedures were available to everyone involved in test administration. Each manual included protocols for reporting any deviations in test administration.

If IDOE determined that an irregularity in test administration or security occurred, it acted based upon approved procedures including but not limited to the following:

- Invalidation of student scores; and
- A requirement for the corporation or school to administer a breach form.

### 4.1 SECURITY OF TEST MATERIALS

Before test materials were finalized, test items and performance tasks went through multiple reviews, including review by various committees. It was critical to maintain the security of test items and performance tasks during these committee meetings. Items were accessed directly from CAI's secure Item Tracking System (ITS) for online committee meetings. Printed copies of items and performance task content were not

provided to educators participating in the committee meetings. Any secure materials created at the meetings or distributed during the meetings were collected and destroyed following the meetings. Secure content was printed on light green paper with each page marked as secure in the header and/or footer. No materials were viewed by participants until after they signed the CAI and IDOE non-disclosure forms. CAI staff reviewed the security procedures with the committee members prior to obtaining their written acknowledgement.

All test items and performance tasks, test materials, and student-level testing information were deemed secure and were required to be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of assessment questions, prompts, and student results. Any deviation in test administration was required to be reported to protect the validity of the assessment results.

The security of all test materials was required before, during, and after test administration. After any administration, initial or make-up test session, secure materials (e.g., scratch paper) were required to be returned immediately to the STC and placed in locked storage. Secure materials were never to be left unsecured and were not permitted to remain in classrooms or be removed from the school's campus overnight. Secure materials that did not need to be returned to the print vendor for scanning and scoring were to be destroyed securely following outlined security guidelines but were not allowed to be discarded in the trash. In addition, any monitoring software that might have allowed test content on student workstations to be viewed or recorded on another computer or device during testing had to be disabled.

It was considered a testing security violation for authorized corporation or school personnel to fail to follow security procedures set forth by the IDOE, and no individual was permitted to do the following:

- Read, copy, share or view the passages, test items, or performance tasks before, during, or after testing;
- Explain the passages, test items, or performance tasks to students;
- Change or otherwise interfere with student responses to test items or performance tasks;
- Copy or read student responses; and
- Cause achievement of schools to be inaccurately measured or reported.

All accommodated assessment books (regular print, large print, braille, and Spanish) were treated as secure documents, and processes were in place to protect them from loss, theft, and reproduction of any kind.

To access the online ILEARN tests, a secure browser was required. The CAI Secure Browser provided a secure environment for student testing by disabling hot keys, copy, and screen capture capabilities and preventing access to the desktop (e.g., Internet, email, and other files or programs installed on school machines). Users could not access other applications from within the CAI Secure Browser, even if they knew the keystroke sequences. Students were not able to print from the CAI Secure Browser. During testing,

the desktop was locked down. The CAI Secure Browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. See the *Online Test Delivery System (TDS) User Guide* in Appendix A for further details.

## **4.2 IDENTIFYING TEST IRREGULARITIES OR POTENTIAL TEST SECURITY CONCERNS**

CAI's quality monitoring (QM) system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QM system, and any anomalies (such as tests not meeting blueprint, unexpected test lengths, or other unlikely issues) are flagged. CAI psychometricians run quality assurance reports and alert the program team of any issues. The forensic analysis report from the QM system flags unlikely patterns of behavior in testing administrations aggregated at the following levels: test administration, TA, and school.

Item statistics and blueprint reports were run and reviewed weekly during the 2019-2020 ILEARN testing windows. In addition, response change analyses for multiple-choice and multi-select items were conducted. The last and next to last (if it existed) responses were compared and students or aggregates were flagged if the number or average number of wrong to right response changes was above the flagging criteria.

CAI psychometricians monitored testing anomalies throughout the testing window. A variety of evidence was collected for the evaluation. These evidences include blueprint match, unusual or much longer test times as compared to the state average, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and can be set by IDOE. While analyses used to detect the testing anomalies could be run anytime within the testing window, analyses relying on state averages are typically held until the close of the testing window to ensure final data is being used.

The lead psychometrician will alert the program team leads if any unexpected results are identified in order to immediately resolve any issues.

## **4.3 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 Irregularities module under Administering Tests in TIDE.

TIDE allowed CTCs, NPSTCs, and STCs to report test irregularities (i.e., re-open test, re-open test segment) that occurred in the testing environment. In many cases, formal documentation prescribed by IDOE was required in addition to the submission of 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 had 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 2019-2020 school year.

Table 3 presents examples of test irregularities and test security violations.

*Table 3: Examples of Test Irregularities and Test Security Violations*

Description
Student(s) making distracting gestures/sounds or talking during the test session that creates a disruption in the test session for other students.
Student(s) leaving the test room without authorization.
TA or Test Coordinator leaving related instructional materials on the walls in the testing room.
Student(s) cheating or providing answers to each other, including passing notes, giving help to other students during testing, or using handheld electronic devices to exchange information.
Student(s) accessing or using unauthorized electronic equipment (e.g., cell phones, smart watches, iPods, or electronic translators) during testing.
Disruptions to a test session such as a fire drill, school-wide power outage, earthquake, or other acts.
TA or Test Coordinator failing to ensure administration and supervision of the assessments by qualified, trained personnel.
TA giving incorrect instructions.
TA or Test Coordinator giving out his or her username/password (via email or otherwise), including to other authorized users.
TA allowing students to continue testing beyond the close of the testing window.
TA or teacher coaching or providing any other type of assistance to students that may affect their responses. This includes both verbal cues (e.g., interpreting, explaining, or paraphrasing the test items or prompts) and nonverbal cues (e.g., voice inflection, pointing, or nodding head) to the correct answer. This also includes leading students through instructional strategies such as think-aloud, asking students to point to the correct answer or otherwise identify the source of their answer, requiring students to show their work to the TA, or reminding students of a recent lesson on a topic.
TA providing students with unallowable materials or devices during test administration or allowing inappropriate designated features and/or accommodations during test administration.
TA providing a student access to another student's work/responses.
TA or Test Coordinator modifying student responses or records at any time.
TA providing students with access to a calculator during a portion of the assessment that does not allow the use of a calculator.
TA uses another staff member's username and/or password to access vendor systems or administer tests.
TA uses a student's login information to access practice tests or operational tests.

## **4.4 CAI's SYSTEM SECURITY**

CAI has built-in security controls in all its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of CAI's systems encrypt data at rest and in transit. ILEARN data resides on servers at Rackspace, CAI's online hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. Staff at both CAI and Rackspace receive formal training in security procedures to ensure that they know the procedures and implement them properly.

Hardware firewalls and intrusion detection systems protect CAI networks from intrusion. CAI's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. All of CAI's secure websites and software systems enforce role-based security models that protect individual privacy and confidentiality in a manner consistent with the Family Educational Rights and Privacy Act (FERPA).

CAI's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. CAI maintains logs of key activities and indicators, including data backup, server response time, user accounts, system events and security, and load test results.

## REFERENCES

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



**Indiana Learning Evaluation and  
Readiness Network  
(*ILEARN*)**

**2019–2020**

**Volume 4  
Evidence of Reliability and  
Validity**

## **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 the IDOE at [INassessments@doe.in.gov](mailto:INassessments@doe.in.gov).

Major contributors to this technical report include the following staff from Cambium Assessment Inc. (CAI): Stephan Ahadi, Elizabeth Ayers-Wright, Elizabeth Xiaoxin Wei, Kevin Clayton, Aleah Pepper, and Kyra Bilenki. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and Program Leads.

## TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE .....	1
1.1 Reliability .....	2
1.2 Validity.....	4
2. PURPOSE OF <i>ILEARN</i> .....	7
3. EVIDENCE OF CONTENT VALIDITY .....	8
3.1 Content Standards .....	8
4. RELIABILITY .....	11
4.1 Marginal Reliability .....	11
4.2 Test Information Curves and Standard Error of Measurement.....	11
4.3 Reliability of Performance Classification .....	13
4.3.1 <i>Classification Accuracy</i> .....	14
4.3.2 <i>Classification Consistency</i> .....	16
4.4 Precision at Cut Scores.....	18
5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE .....	19
5.1 Correlations Among Reporting Category Scores.....	19
5.2 Confirmatory Factor Analysis .....	21
5.2.1 <i>Factor Analytic Methods</i> .....	22
5.2.2 <i>Results</i> .....	24
5.3 Local Independence .....	24
5.4 Convergent and Discriminant Validity .....	25
6. FAIRNESS IN CONTENT.....	27
6.1 Statistical Fairness in Item Statistics.....	27
7. SUMMARY .....	29
8. REFERENCES .....	30

## LIST OF TABLES

Table 1: Test Administration.....	1
Table 2: Number of Items for Each Reporting Category (ELA) .....	8
Table 3: Number of Items for Each Reporting Category (Mathematics) .....	9
Table 4: Number of Items for Each Reporting Category (Science) .....	10
Table 5: Number of Items for Each Reporting Category (Social Studies).....	10
Table 6: Marginal Reliability Coefficients.....	11
Table 7: Descriptive Statistics .....	14
Table 8: Classification Accuracy Index.....	16
Table 9: False Classification Rates (Biology) .....	17
Table 10: Classification Accuracy and Consistency (Cut 1 and Cut 2).....	17
Table 11: Classification Accuracy and Consistency (Cut 2 and Cut 3).....	17
Table 12: Classification Accuracy and Consistency (Cut 3 and Cut 4).....	18
Table 13: Performance Levels and Associated Conditional Standard Error of Measurement, Biology .....	18
Table 15: Observed Correlation Matrix Among Reporting Categories (Biology) .....	20
Table 16: Disattenuated Correlation Matrix Among Reporting Categories (Biology).....	20
Table 17: Biology Q <sub>3</sub> Statistic.....	25

## LIST OF FIGURES

Figure 1: Sample Test Information Function .....	12
Figure 2: Conditional Standard Error of Measurement by Biology Administration .....	13
Figure 3: Second-Order Factor Model (Biology).....	24

## LIST OF APPENDICES

Appendix A: <i>Reliability Coefficients</i>
Appendix B: <i>Conditional Standard Error of Measurement</i>

# 1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

*ILEARN* is an online, adaptive assessment for ELA, Mathematics, and Science and an online, fixed-form assessment for Social Studies. Online accommodated and paper-and-pencil versions of the assessments are available to students whose Individualized Education Programs (IEPs) or Section 504 Plans indicated such a need. Table 1 displays the complete list of test administration methods for the 2019–2020 school year.

*Table 1: Test Administration*

Subject	Administration*	Grade
ELA	Online census tests	3–8
Mathematics	Online census tests	3–8
Science	Online census tests	4, 6, Biology
Social Studies	Online census tests	5, U.S. Government

\*Accommodated versions, including braille and Spanish, were delivered online. Paper-and-pencil versions were also available. Full descriptions of available accommodations are listed in Volume 3, Section 1.2.

With the administration of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic performance from *ILEARN* scores. This volume provides empirical evidence about the reliability and validity of the 2019–2020 *ILEARN* assessments.

As detailed in Volume 1, only the Fall and Winter Biology assessments were administered during the 2019–2020 school year. This volume summarizes the reliability and validity of those assessments.

The purpose of this volume is to provide empirical evidence to support a validity argument regarding the uses and inferences for the *ILEARN* assessment. This volume addresses the following:

- **Reliability.** Marginal reliability estimates for each test are reported in this volume. The reliability estimates are presented by grade and subject in the main body and by demographic subgroups in Appendix A. This section also includes conditional standard errors of measurement (CSEMs), classification accuracy and consistency results by grade and subject.
- **Content Validity.** Evidence is provided to show that test forms were constructed to measure the Indiana Academic Standards (IAS) with a sufficient number of items targeting each area of the blueprint.
- **Internal Structure Validity.** Evidence is provided regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and disattenuated Pearson correlations among reporting categories per grade. Confirmatory factor analysis has also been performed using the second-order



factor model. Additionally, local item independence, an assumption of unidimensional IRT, was tested using the  $Q_3$  statistic.

- *Test Fairness.* Fairness is statistically analyzed using differential item functioning (DIF) in tandem with content alignment reviews by specialists.

## 1.1 RELIABILITY

Reliability refers to consistency in test scores. Reliability can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_X^2}.$$

There are various approaches for estimating the reliability of scores. The conventional approaches used are characterized as follows:

- The *test-retest* method measures stability over time. With this method, the same test is administered twice to the same group at two different points in time. If test scores from the two administrations are highly correlated, then the test scores are deemed to have a high level of stability. For example, if the result is highly stable, those who scored high on the first test administration tend to obtain a high score on the second administration. The critical factor, however, is the time interval. The time interval should not be too long, which could allow for changes in the test takers' true scores. Likewise, it should not be too short, or memory and practice may confound the results. The test-retest method is most effective for measuring constructs that are stable over time, such as intelligence or personality traits. This was not used for *ILEARN* assessments as there was a single test for all students.
- The *parallel-forms* method is used for measuring equivalence. With this design, two parallel forms of the test are administered to the same group. This method requires two similar forms of a test. However, it is difficult to create two strictly parallel forms. When this method is applied, the effects of memory or practice can be eliminated or reduced, since the tests are not purely identical as is the case with the test-retest method. The reliability coefficient from this method indicates the degree to which the two tests are measuring the same construct. While there are many possible items to administer to measure any particular construct, it is feasible to administer only a sample of items on any given test. If there is a high correlation between the scores of the two tests, then inferences regarding high reliability of scores can be substantiated. This method is commonly used to estimate the reliability of performance of aptitude tests. Since this method also requires two scores for students, this was also not used for *ILEARN* assessments.

- The *split-half* method uses one test divided into two halves within a single test administration. It is crucial to make the two half-tests as parallel as possible, as the correlation between the two half-tests is used to estimate the reliability of the whole test. In general, this method produces a coefficient that underestimates the reliability of the full test. To correct the estimate, the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910) can be applied. While this method is convenient, varying splits of items may yield different reliability estimates.
- The *internal consistency* method can be employed when it is not possible to conduct repeated test administrations. Whereas other methods often compute the correlation between two separate tests, this method considers each item within a test to be a one-item test. There are several other statistical methods based on this idea: coefficient *alpha* (Cronbach, 1951), Kuder-Richardson Formula 20 (Kuder & Richardson, 1937), Kuder-Richardson Formula 21 (Kuder & Richardson, 1937), stratified coefficient *alpha* (Qualls, 1995), and the Feldt-Raju coefficient (Feldt & Brennan, 1989; Feldt & Qualls, 1996).
- *Inter-rater reliability* is the extent to which two or more individuals (coders or raters) agree. Inter-rater reliability addresses the consistency of the implementation of a rating system. Inter-rater reliability in the form of percent agreement and weighted kappa was used to summarize writing prompt hand-scoring reliability.

The first four methods discussed above are classical methods of calculating reliability and are not optimal for computer adaptive testing. 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 along the scale. For example, most fixed-form assessments target test information near important cut scores or near the population mean, so that test scores are most precise in targeted locations. Because adaptive tests target test information near each student's ability level, the precision of test scores may increase, especially for lower- and higher-ability students. 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. In addition, the first two methods require multiple testing opportunities which are not available for *ILEARN*.

Another way to view reliability is to consider its relationship with the standard errors of measurement (SEMs)—the smaller the standard error, the higher the precision of the test scores. For example, classical test theory assumes that an observed score ( $X$ ) of any individual can be expressed as a true score ( $T$ ) plus some error as ( $E$ ),  $X = T + E$ . The variance of  $X$  can be shown as the sum of two orthogonal variance components:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

Returning to the definition of reliability as the ratio of true score variance to observed score variance, we arrive at:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}.$$

As the fraction of error variance to observed score variance tends toward zero, the reliability then tends toward 1. The classical test theory (CTT) SEM, which assumes a homoscedastic error, is derived from the classical notion expressed previously as  $\sigma_X \sqrt{1 - \rho_{XX'}}$ , where  $\sigma_X$  is the standard deviation of the scaled score and  $\rho_{XX'}$  is a reliability coefficient. Based on the definition of reliability, the following formula can be derived:

$$\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2},$$

$$\frac{\sigma_E^2}{\sigma_X^2} = 1 - \rho_{XX'},$$

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX'}),$$

$$\sigma_E = \sigma_X \sqrt{(1 - \rho_{XX'})}.$$

In general, the SEM is relatively constant across samples as the group dependent term,  $\sigma_X$ , and can be cancelled out as:

$$\sigma_E = \sigma_X \sqrt{(1 - \rho_{XX'})} = \sigma_X \sqrt{(1 - (1 - \frac{\sigma_E^2}{\sigma_X^2}))} = \sigma_X \sqrt{\frac{\sigma_E^2}{\sigma_X^2}} = \sigma_X \cdot \frac{\sigma_E}{\sigma_X} = \sigma_E.$$

This shows that the SEM in the CTT is assumed to be homoscedastic irrespective of the standard deviation of a group.

In contrast, the SEMs in IRT vary over the ability continuum. These heterogeneous errors are a function of a TIF that provides different information about test takers depending on their estimated abilities. Often, TIF is maximized over an important performance cut, such as the proficient cut score.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution, or near an important classification cut, and have less information at the tails of the score distribution. See Section 4.2, Test Information Curves and Standard Error of Measurement, for the derivation of heterogeneous errors in IRT.

## 1.2 VALIDITY

Validity refers to the degree to which “evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on

Measurement in Education [NCME], 2014). Messick (1989) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment.” Both of these definitions emphasize evidence and theory to support inferences and interpretations of test scores. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggest five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of evidence for validity is the relationship between the test content and the intended test construct (see Section 3.1, Content Standards). In order 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 (see Volume 2 of this technical report for details). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance.

For example, a Mathematics item targeting a specific mathematics skill that requires advanced reading proficiency and vocabulary has a high level of construct-irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores. Statistical analyses, such as factor analysis or multidimensional scaling, are also used to evaluate content relevance. Results from factor analysis for the *ILEARN* assessment are presented in Section 5.2, Confirmatory Factor Analysis. Evidence based on test content is a crucial component of validity, because construct underrepresentation or irrelevancy could result in unfair advantages or disadvantages to one or more groups of test takers.

In addition, technology-enhanced items should be examined to ensure that no construct-irrelevant variance is introduced. If some aspect of the technology impedes or advantages a student in his or her responses to items, this could affect item responses and inferences regarding abilities on the measured construct (see Volume 2).

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

The third source of evidence for validity is based on 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. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method of analyzing the

internal structure of tests (see Volume 1, Section 5.2). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (see Section 4, Reliability, and Section 5, Evidence of Internal-External Structure, for details).

A fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards* (AERA, APA, & NCME, 2014) divide this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs; conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multi-trait-multimethod matrix can be used (see Section 5.4, Convergent and Discriminant Validity, for details). Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends upon 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 evidence for validity 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; this process should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is in fact due to an unintended, confounding aspect of the test, this would interfere with the test's validity. As described in Volume 1 and in this volume, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This then allows for one to evaluate whether sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test 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.

## 2. PURPOSE OF *ILEARN*

The primary purpose of the *ILEARN* assessments is to yield test scores at the student level and other levels of aggregation that reflect student performance relative to the IAS. *ILEARN* supports instruction and student learning by measuring growth in student performance and providing feedback to educators and parents that can be used to form instructional strategies to remediate or enrich instruction. Assessments can be used to determine whether students in Indiana have the knowledge and skills essential for college-and-career-readiness.

Indiana’s education assessments also help fulfill the requirements for state and federal accountability systems. Test scores can be employed to evaluate students’ learning progress and help teachers improve their instruction, which in turn will have a positive effect on student learning over time.

The tests are constructed to measure student proficiency on the IAS in ELA, Mathematics, Science, and Social Studies. The tests were developed using principles of evidence-centered design and adhering to the principles of universal design to ensure that all students have access to the test content. Volume 2, Test Development, describes the IAS and test blueprints in more detail. This volume provides evidence of content validity in Section 3, Evidence of Content Validity. The *ILEARN* test scores are useful indicators for understanding individual students’ academic performance regarding the IAS and whether students are progressing in their performance over time. Additionally, individual test scores can be used to measure test reliability, which is described in Section 4, Reliability.

*ILEARN* assessments are criterion-referenced tests designed to measure student performance on the IAS in ELA, Mathematics, Science, and Social Studies. As a comparison, norm-referenced tests are designed to compare or rank all students to one another.

The scale score and relative strengths and weaknesses at the reporting category (domain) level were provided for each student to indicate student strengths and weaknesses in different content areas of the test relative to the other areas and to the district and state. These scores help teachers tailor their instruction, provided that the scores are viewed with the usual caution that accompanies the use of reporting category scores. Thus, we must examine the reliability coefficients for these test scores and the validity of the test scores to support practical use of these tests across the state. Volume 5 of this technical report is the score interpretation guide and provides details on all generated scores and their appropriate uses and limitations.

### 3. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the *ILEARN* are representative of the content standards of the larger knowledge domain. We describe the content standards for *ILEARN* and discuss the test development process, mapping *ILEARN* tests to the standards. A complete description of the test development process can be found in Volume 2, Test Development.

#### 3.1 CONTENT STANDARDS

The IAS were approved by the Indiana State Board of Education in April 2014 for ELA and Mathematics and in March 2015 for Social Studies. The IAS for Science were originally revised in 2010 and updated in 2016 to reflect changes in Science content. The IAS are intended to implement more rigorous standards, with the goal of challenging and motivating Indiana’s students to acquire stronger critical thinking, problem solving, and communications skills promoting college-and-career-readiness.

*ILEARN* blueprints are available in Volume 2’s appendices. Blueprints were developed to ensure that the test and the items were aligned to the prioritized standards they were intended to measure. A complete description of the blueprint and test form construction process can be found in Volume 2, Section 4.

Table 2 through Table 5 present the reporting categories by grade and test, as well as the number of items measuring each category on the 2019-2020 tests. Reading Foundations in ELA Grade 3, Speaking and Listening in ELA Grades 3-8, and Process Standards in Mathematics Grades 3-8 were not reported as a separate reporting category, but were included only in the overall aggregate scale score calculations.

*Table 2: Number of Items for Each Reporting Category (ELA)*

Reporting Category	Grade					
	3	4	5	6	7	8
Key Ideas and Textual Support/Vocabulary	12-15	11-14	11-14	10-13	10-13	10-12
Structural Elements and Organization/Connection of Ideas/Media Literacy	10-12	11-14	11-14	10-13	10-13	10-12
Writing	6-8	7-8	6-8	7-8	7-8	6-8
Speaking and Listening	2-3	2-3	2-3	2-3	2-3	2-3

*Table 3: Number of Items for Each Reporting Category (Mathematics)*

Grade	Reporting Category	Number of Items
3	Algebraic Thinking and Data Analysis	9-11
	Computation	11-13
	Geometry and Measurement	9-11
	Number Sense	11-13
	Process Standards	4-6
4	Algebraic Thinking and Data Analysis	9-11
	Computation	11-13
	Geometry and Measurement	9-11
	Number Sense	11-13
	Process Standards	4-6
5	Algebraic Thinking	10-12
	Computation	11-13
	Geometry and Measurement, Data Analysis, and Statistics	9-11
	Number Sense	11-13
	Process Standards	4-6
6	Algebra and Functions	11-13
	Computation	10-12
	Geometry and Measurement, Data Analysis, and Statistics	9-11
	Number Sense	10-12
	Process Standards	4-6
7	Algebra and Functions	11-13
	Data Analysis, Statistics, and Probability	9-11
	Geometry and Measurement	9-11
	Number Sense and Computation	12-13
	Process Standards	4-6
8	Algebra and Functions	11-13
	Data Analysis, Statistics, and Probability	10-12
	Geometry and Measurement	10-12
	Number Sense and Computation	9-11
	Process Standards	4-6



*Table 4: Number of Items for Each Reporting Category (Science)*

Grade	Reporting Category	Number of Items
4	Questioning and Modeling	12-14
	Investigating	12-14
	Analyzing, Interpreting, and Computational Thinking	10-12
	Explaining Solutions, Reasoning, and Communicating	10-12
6	Questioning and Modeling	10-12
	Investigating	10-12
	Analyzing, Interpreting, and Computational Thinking	12-14
	Explaining Solutions, Reasoning, and Communicating	12-14
Biology*	Developing and Using Models to Describe Structure and Function	10-12
	Developing and Using Models to Explain Processes	10-12
	Analyzing Data and Mathematical Thinking	10-12
	Constructing and Communicating an Explanation	10-12
	Evaluating Claims with Evidence	10-12

\*The operational item pool and blueprint for the fall, winter, and spring windows were identical.

*Table 5: Number of Items for Each Reporting Category (Social Studies)*

Grade	Reporting Category	Number of Items
5	Civics and Government	17
	Geography and Economics	11
	History	12
U.S. Government	Functions of Government	20
	Historical Foundations of American Government	14
	Institutions and Processes of Government	20

## 4. RELIABILITY

### 4.1 MARGINAL RELIABILITY

Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the performance scale, for all students. The marginal reliability coefficients are nearly identical or close to the coefficient *alpha*. For our analysis, the marginal reliability coefficients were computed using operational items.

Within the IRT framework, measurement error varies across the range of ability. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the TIF represents the SEM. 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, as opposed to students with more extreme scores. Conversely, measurement error is minimal for the part of the underlying scale that is at the middle of the test distribution and greater on scaled values farther away from the middle.

The marginal reliability of a test is computed by integrating  $\theta$  out of the TIF as follows:

$$\rho = \frac{\sigma_{\theta}^2 - \bar{\sigma}_e^2}{\sigma_{\theta}^2},$$

where  $\sigma_{\theta}^2$  is the true score variance of  $\theta$  and

$$\bar{\sigma}_e^2 = \int_{-\infty}^{\infty} \frac{1}{I(\theta)} g(\theta) d\theta,$$

where  $g(\theta)$  is a density function. Population parameters are assumed normal,  $g(\theta) \sim N(0,1)$ .

Table 6 presents the marginal reliability coefficients for all students. The marginal reliability coefficients for the fall and winter administrations were 0.927 and 0.926, respectively.

Table 6: Marginal Reliability Coefficients

Grade	Marginal Reliability
Fall	0.927
Winter	0.926

### 4.2 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

Within the IRT framework, measurement error varies across the range of ability as a result of the test, providing varied information across the range of ability 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.

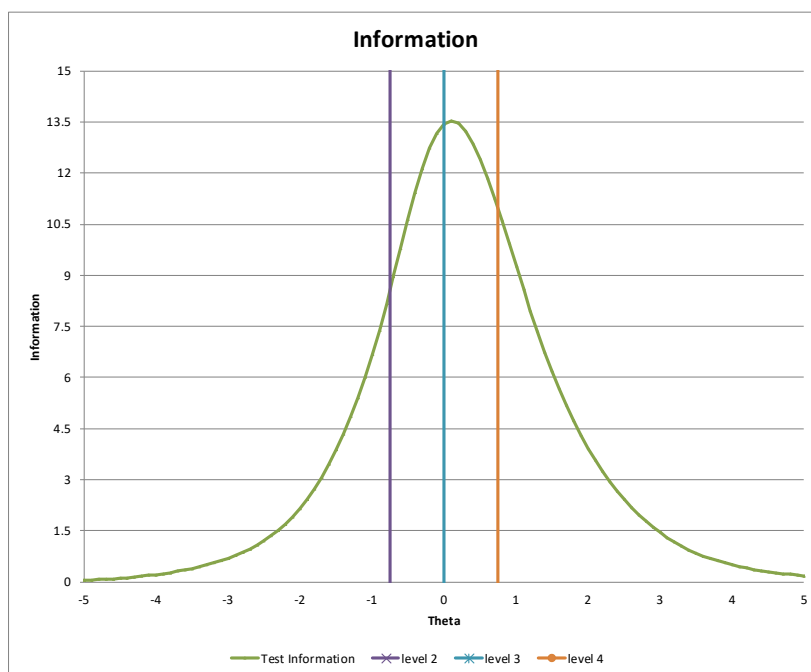
Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most-precise scores in this range. Where the curve is lower at the tails indicates that the test provides less information about test takers at the tails relative to the center.

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

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

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

Figure 1: Sample Test Information Function

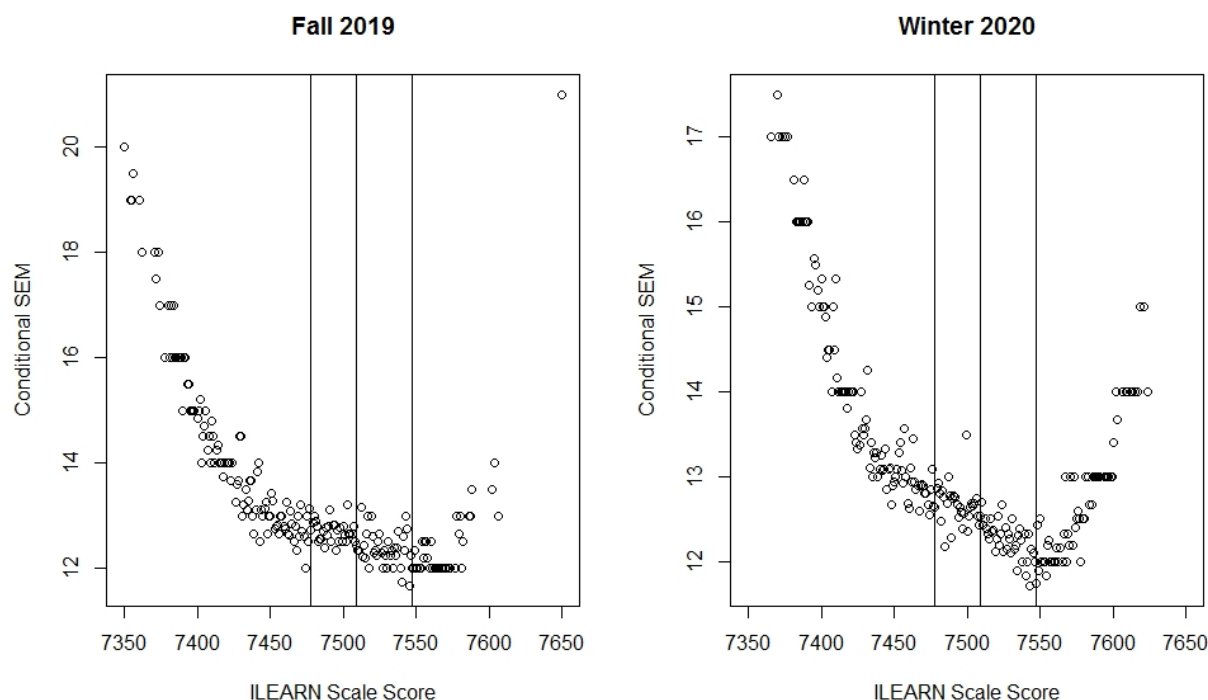


The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

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

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented in Figure 2. These plots are based on the scaled scores reported in 2020. Vertical lines represent the performance category cut scores.

Figure 2: Conditional Standard Error of Measurement by Biology Administration



For most tests, the standard error curves follow the typical expected trends, with more test information regarding scores observed near the middle of the score scale.

Reliability coefficients and SEM for each reporting category are also presented in Appendix A, and Appendix B includes the average CSEM by scale score and corresponding performance levels for each scale score.

### 4.3 RELIABILITY OF PERFORMANCE CLASSIFICATION

When students complete *ILEARN* assessments, they are placed into performance levels by their observed scaled score. The cut scores for student classification into the different performance levels were determined after the *ILEARN* standard-setting process. A complete description of the standard-setting process can be found in Volume 6, Setting Performance Standards.

Misclassification probabilities are computed for all performance-level standards (i.e., for the cuts between levels 1 and 2, levels 2 and 3, and levels 3 and 4). The performance level- cut between level 2 and level 3 is of primary interest, because students are classified as At Proficiency or Approaching Proficiency using this cut. Students with observed scores far from the level 3 cut are expected to be classified more accurately as At Proficiency or Approaching Proficiency than students with scores near this cut.

This report estimates classification reliabilities using two different methods: one based on observed abilities and a second based on estimating a latent posterior distribution for the true scores.

Two approaches for estimating classification probabilities are provided. The first is an observed score approach to computing misclassification probabilities and is designed to explore the following research questions:

1. What is the overall classification accuracy index (CAI) of the total test?
2. What is the classification accuracy rate index for each individual performance cut within the test?

The second approach computes misclassification probabilities using an IRT-based method for students scoring at each score point. This approach is designed to explore the following research questions:

1. What is the probability that the student's true score is below the cut point?
2. What is the probability that the student's true score is above the cut point?

Both approaches yield student-specific classification probabilities that can be aggregated to form overall misclassification rates for the test. The former estimates the classification accuracy, and the latter estimates the classification consistency.

For these analyses, we used students from the fall and winter *ILEARN* Biology population data files that had an overall score reported. Table 7 provides the sample size, mean, and standard deviation of the observed theta data. The theta scores are based on the maximum likelihood estimates (MLEs) obtained from CAI's scoring engine.

Table 7: Descriptive Statistics

Administration	Sample Size	Mean Theta	Standard Deviation of Theta	Mean Scale Score	Standard Deviation of Scale Scores
Fall	870	-0.41	0.98	7479.47	48.80
Winter	1713	-0.23	0.95	7488.54	47.60

### 4.3.1 Classification Accuracy

The observed score approach (Rudner, 2001), implemented to assess classification accuracy, is based on the probability that the true score,  $\theta$ , for student  $j$  is within

performance level  $l = 1, 2, \dots, L$ . This probability can be estimated from evaluating the integral

$$p_{jl} = \Pr(c_{lower} \leq \theta_j < c_{upper} | \hat{\theta}_j, \hat{\sigma}_j^2) = \int_{c_{lower}}^{c_{upper}} f(\theta_j | \hat{\theta}_j, \hat{\sigma}_j^2) d\theta_j,$$

where  $c_{upper}$  and  $c_{lower}$  denote the score corresponding to the upper and lower limits of the performance level, respectively.  $\hat{\theta}_j$  is the ability estimate of the  $j$ th student with SEM of  $\hat{\sigma}_j$ , and using the asymptotic property of normality of the maximum likelihood estimate (MLE),  $\hat{\theta}_j$ , we take  $f(\cdot)$  as asymmetrically normal, so the previous probability can be estimated by

$$p_{jl} = \Phi\left(\frac{c_{upper} - \hat{\theta}_j}{\hat{\sigma}_j}\right) - \Phi\left(\frac{c_{lower} - \hat{\theta}_j}{\hat{\sigma}_j}\right),$$

where  $\Phi(\cdot)$  denotes the standard normal cumulative distribution function. The expected number of students at level  $l$  based on students from observed level  $v$  can be expressed as

$$E_{vl} = \sum_{p_{l_i} \in v} p_{jl},$$

where  $p_{l_j}$  is the  $j$ th student's performance level and the values of  $E_{vl}$  are the elements used to populate the matrix  $\mathbf{E}$ , a  $4 \times 4$  matrix of conditionally expected numbers of students to score within each performance-level bin based on their true scores. The overall CAI of the test can then be estimated from the diagonal elements of the matrix

$$CAI = \frac{tr(\mathbf{E})}{N},$$

where  $N = \sum_{v=1}^4 N_v$  and  $N_v$  is the observed number of students scoring in performance level  $v$ . The classification accuracy index for the individual cut  $p$ , ( $CAIC_p$ ), is estimated by forming square partitioned blocks of the matrix  $\mathbf{E}$  and taking the summation over all elements within the block as follows:

$$CAIC_p = \left( \sum_{v=1}^p \sum_{l=1}^p E_{vl} + \sum_{v=p+1}^4 \sum_{l=p+1}^4 E_{vl} \right) / N,$$

where  $p(p = 1, 2, 3)$  is the  $p$ th cut.

Table 8 provides the overall CAI and the classification accuracy index for the individual cuts (CAIC) based on the observed score approach. There is no industry standard, but these numbers suggest that misclassification would not be frequent in the population data.

The cut accuracy rates are much higher, denoting that the degree to which we can reliably differentiate between students of adjacent performance levels is above 0.9.

Table 8: Classification Accuracy Index

Administration	Overall Accuracy Index	Cut Accuracy Index		
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4
Fall	0.820	0.925	0.929	0.966
Winter	0.810	0.917	0.929	0.963

### 4.3.2 Classification Consistency

Classification accuracy refers to the degree to which a student's true score and observed score would fall within the same performance level (Rudner, 2001). Classification consistency refers to the degree to which test takers are classified into the same performance level, assuming the test is administered twice independently (Lee, Hanson, & Brennan, 2002)—that is, the percentages of students who are consistently classified in the same performance levels on two equivalent test forms. In reality, the true ability is unknown, and students do not take an alternate, equivalent form; therefore, classification consistency is estimated based on students' item scores, the item parameters, and the assumed underlying latent ability distribution.

The IRT-based approach (Guo, 2006) makes use of student-level item response data from the 2019 test administration. For the  $j$ th student, we can estimate a posterior probability distribution for the latent true score and, from this, estimate the probability that a true score is above the cut as

$$p(\theta_j \geq c) = \frac{\int_c^{\infty} p(\mathbf{z}_j|\theta_j)f(\theta_j|\mu, \sigma)d\theta_j}{\int_{-\infty}^{\infty} p(\mathbf{z}_j|\theta_j)f(\theta_j|\mu, \sigma)d\theta_j},$$

where  $c$  is the cut score required for passing in the same assigned metric,  $\theta_j$  is true ability in the true-score metric,  $\mathbf{z}_j$  is the item score,  $\mu$  is the mean, and  $\sigma$  is the standard deviation of the population distribution. The function  $p(\mathbf{z}_j|\theta_j)$  is the probability of a particular pattern of responses given the theta, and  $f(\theta)$  is the density of the proficiency  $\theta$  in the population.

Similarly, we can estimate the probability that a true score is below the cut as

$$p(\theta_j < c) = \frac{\int_{-\infty}^c p(\mathbf{z}_j|\theta_j)f(\theta_j|\mu, \sigma)d\theta_j}{\int_{-\infty}^{\infty} p(\mathbf{z}_j|\theta_j)f(\theta_j|\mu, \sigma)d\theta_j}.$$

From these misclassification probabilities, we can estimate the overall false positive rate (FPR) and false negative rate (FNR) of the test. The FPR is expressed as the proportion of individuals who scored above the cut based on their observed score but whose true score would otherwise have classified them as below the cut. The FNR is expressed as the proportion of individuals who scored below the cut based on their observed score but who otherwise would have been classified as above the cut based on their true scores. These rates are estimated as follows:

$$FPR = \sum_{j \in \bar{\theta}_j \geq c} p(\theta_j < c) / N$$

$$FNR = \sum_{j \in \bar{\theta}_j < c} p(\theta_j \geq c) / N.$$

Table 9 provides the FPR and FNR by administration. The FPR and FNR rates for the level 2/3 cut are between 0.01 and 0.11.

*Table 9: False Classification Rates (Biology)*

	1/2 cut		2/3 cut		3/4 cut	
Administration	FPR	FNR	FPR	FNR	FPR	FNR
Fall	0.08	0.07	0.01	0.11	0.01	0.14
Winter	0.11	0.07	0.01	0.04	0.01	0.19

The classification consistency index for the individual cut  $c$ , ( $CICC_c$ ), was estimated using the following equation:

$$CICC_c = \frac{\sum_j \{p^2(\theta_j \geq c) + p^2(\theta_j < c)\}}{N}.$$

Classification consistency with classification accuracy results are presented in Table 10 through Table 12. All accuracy values are higher than 0.91, and consistency values are around 0.90. Across all grades and subjects and in all performance levels, classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy because the consistency is based on two tests with measurement errors, while the accuracy is based on one test with a measurement error and the true score. The accuracy and consistency rates for each performance level are higher for the levels with smaller standard error.

*Table 10. Classification Accuracy and Consistency (Cut 1 and Cut 2)*

Administration	Accuracy	Consistency
Fall	0.925	0.888
Winter	0.917	0.882

*Table 11. Classification Accuracy and Consistency (Cut 2 and Cut 3)*

Administration	Accuracy	Consistency
Fall	0.929	0.894
Winter	0.929	0.897



Table 12. Classification Accuracy and Consistency (Cut 3 and Cut 4)

Administration	Accuracy	Consistency
Fall	0.966	0.952
Winter	0.963	0.949

## 4.4 PRECISION AT CUT SCORES

Table 13: Performance Levels and Associated Conditional Standard Error of Measurement, Biology presents mean CSEM at each performance level by administration. The table also includes performance-level cut scores and associated CSEM. The *ILEARN* test scores are somewhat more precise for test scores near the middle of the scale, especially around the At Proficiency performance standard cut. The following table also shows that test scores remain precise even for students in the lowest and highest performance levels.

Table 13: Performance Levels and Associated Conditional Standard Error of Measurement, Biology

Administration	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Fall	1	13.720	--	--
	2	12.707	7478	12.727
	3	12.379	7509	12.444
	4	12.478	7547	12.000
Winter	1	13.410	--	--
	2	12.668	7478	12.647
	3	12.300	7509	12.533
	4	12.507	7547	11.750

## 5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE

In this section, we explore the internal structure of the assessment using the scores provided at the reporting category level. The relationship of the subscores is just one indicator of the test dimensionality.

In ELA grades, there are three reporting categories per grade: Key Ideas and Textual Support/Vocabulary, Structural Elements and Organization/Connection of Ideas/Media Literacy, and Writing. In Mathematics, Science, and Social Studies, reporting categories differ in each grade or course (see Table 3 through Table 5 for reporting category information).

Scale scores and relative strengths and weaknesses based on each reporting category were provided to students. Evidence is needed to verify that scale scores and relative strengths and weaknesses for each reporting category provide both different and useful information for student performance.

It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional IRT model difficult, although we could then easily justify reporting these separate scores. On the contrary, if the reporting categories were perfectly correlated, we could justify a unidimensional model, but we could not justify the reporting of separate scores.

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

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. Both observed correlations and disattenuated correlations are provided in the following section.

### 5.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 14 presents the observed correlation matrix of the reporting category scores for each Biology administration. The correlations among reporting categories ranged from 0.63 to 0.73.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the strand level, given the limited number of items from which the scores were derived.

Consequently, over-interpretation of these correlations, as either high or low, should be made cautiously.

Table 15 displays disattenuated correlations. Disattenuated values greater than 1.00 are reported as 1.00\*. The overall average disattenuated correlation was 0.97 for fall and 0.95 for winter. These values suggest that validity evidence of internal structure is supported.

*Table 14: Observed Correlation Matrix Among Reporting Categories (Biology)*

Administration	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
Fall	Developing and Using Models to Describe Structure and Function (Cat1)	10-12	1.00				
	Developing and Using Models to Explain Processes (Cat2)	10-12	0.64	1.00			
	Analyzing Data and Mathematical Thinking (Cat3)	10-12	0.70	0.65	1.00		
	Constructing and Communicating an Explanation (Cat4)	10-12	0.67	0.63	0.72	1.00	
	Evaluating Claims with Evidence (Cat5)	10-12	0.67	0.63	0.71	0.70	1.00
Winter	Developing and Using Models to Describe Structure and Function (Cat1)	10-12	1.00				
	Developing and Using Models to Explain Processes (Cat2)	10-12	0.65	1.00			
	Analyzing Data and Mathematical Thinking (Cat3)	10-12	0.70	0.66	1.00		
	Constructing and Communicating an Explanation (Cat4)	10-12	0.69	0.63	0.73	1.00	
	Evaluating Claims with Evidence (Cat5)	10-12	0.67	0.63	0.71	0.68	1.00

*Table 15: Disattenuated Correlation Matrix Among Reporting Categories (Biology)*

Administration	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
Fall	Developing and Using Models to Describe Structure and Function (Cat1)	10-12	1.00				
	Developing and Using Models to Explain Processes (Cat2)	10-12	0.97	1.00			
	Analyzing Data and Mathematical Thinking (Cat3)	10-12	0.98	0.95	1.00		
	Constructing and Communicating an Explanation (Cat4)	10-12	0.96	0.94	0.99	1.00	
	Evaluating Claims with Evidence (Cat5)	10-12	0.99	0.97	0.99	1.00*	1.00

Winter	Developing and Using Models to Describe Structure and Function (Cat1)	10-12	1.00				
	Developing and Using Models to Explain Processes (Cat2)	10-12	0.96	1.00			
	Analyzing Data and Mathematical Thinking (Cat3)	10-12	0.95	0.93	1.00		
	Constructing and Communicating an Explanation (Cat4)	10-12	0.97	0.92	0.99	1.00	
	Evaluating Claims with Evidence (Cat5)	10-12	0.96	0.93	0.97	0.96	1.00

## 5.2 CONFIRMATORY FACTOR ANALYSIS

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

The results in this section are intended to provide evidence that the methods for reporting *ILEARN* strand scores align with the underlying structure of the test and provide evidence for appropriateness of the selected IRT models. This section is based on a second-order confirmatory factor analysis, in which the first-order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto factors they are intended to measure. The underlying structure of the *ILEARN* assessments was common across all grades, which is useful for comparing the results of our analyses across the grades.

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

The measurement model and the score reporting method assume a single underlying factor, with separate factors representing each of the reporting categories. Consequently, it is important to collect validity evidence on the internal structure of the assessment to determine the rationality of conducting concurrent calibrations, as well as to use these scoring and reporting methods.

### 5.2.1 Factor Analytic Methods

A series of confirmatory factor analyses (CFA) were conducted using the statistical program Mplus, version 7.31 (Muthén & Muthén, 2012) for each grade and subject assessment. Mplus is commonly used for collecting validity evidence on the internal structure of assessments. The estimation method, weighted least squares means and variance adjusted (WLSMV), was employed because it is less sensitive to the size of the sample and the model and is also shown to perform well with categorical variables (Muthén, du Toit, & Spisic, 1997).

As previously stated, the method of reporting scores used for the *ILEARN* assessments implies separate factors for each reporting category, connected by a single underlying factor. This model is subsequently referred to as the implied model. In factor analytic terms, this suggests that test items load onto separate first-order factors, with the first-order factors connected to a single underlying second-order factor. The use of the CFA in this section establishes some validity evidence for the degree to which the implied model is reasonable.

A chi-square difference test is often applied to assess model fit. However, it is sensitive to sample size, almost always rejecting the null hypothesis when the sample size is large. Therefore, instead of conducting a chi-square difference test, other goodness-of-fit indices were used to evaluate the implied model for *ILEARN*.

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

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

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

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

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

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

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

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

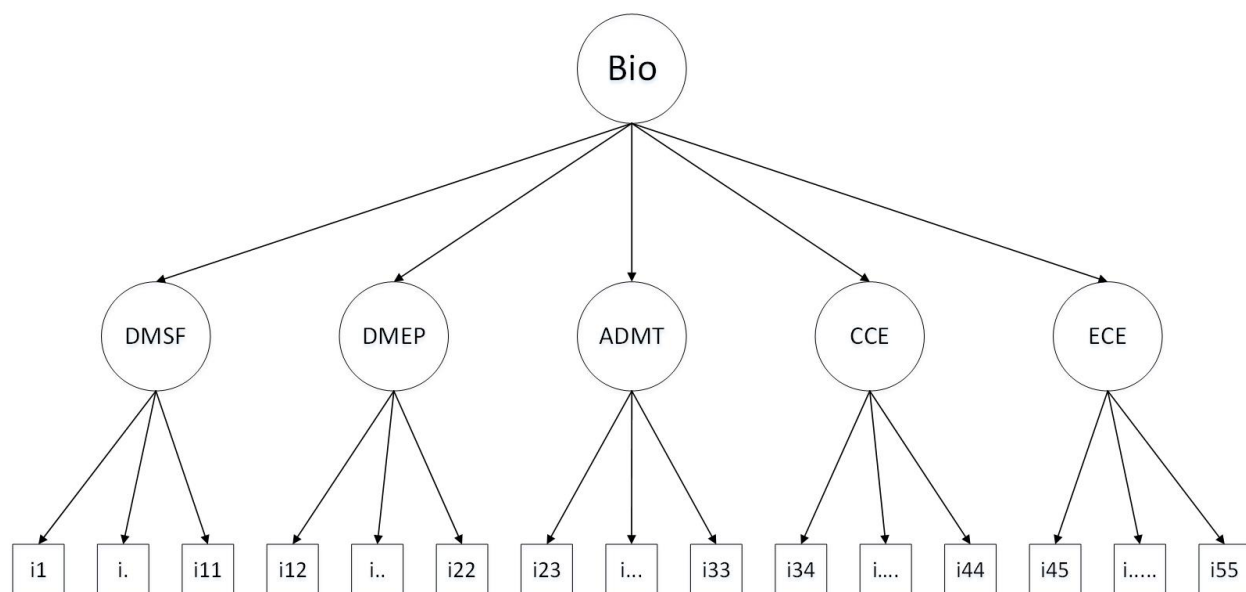
where  $\hat{\Sigma}$  is the implied correlation matrix among test items,  $\Lambda$  is the  $p \times k$  matrix of first-order factor loadings relating item scores to first-order factors,  $\Gamma$  is the  $k \times 1$  matrix of second-order factor loadings relating the first-order factors to the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

There is a separate factor for each reporting category for ELA, Mathematics, Science, and Social Studies. Therefore, the number of rows in  $\Gamma$  ( $k$ ) differs among subjects, but the general structure of the factor analysis is consistent across ELA, Mathematics, Science, and Social Studies.

The second-order factor model can also be represented graphically, and a sample of the generalized approaches is provided on the following page. The general structure of the second-order factor analysis for Science is illustrated in Figure 3. This figure is generally representative of the factor analyses performed for all grades and subjects, with the understanding that the number of items within each reporting category could vary across the grades.

The purpose of conducting confirmatory factor analysis for *ILEARN* was to provide evidence that each individual assessment in *ILEARN* implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

Figure 3: Second-Order Factor Model (Biology)



### 5.2.2 Results

CAI ran CFA for both the Fall and Winter Biology administrations. Due to small sample sizes and the number of items in the pool, the response matrices were sparse. This led to inconclusive results for both administrations.

### 5.3 LOCAL INDEPENDENCE

The validity of the application of IRT depends greatly on meeting the underlying assumptions of the models. One such assumption is local independence, which means that for a given proficiency estimate, the marginal likelihood is maximized, assuming that the probability of correct responses is the product of independent probabilities over all items (Chen & Thissen, 1997):

$$L(\theta) = \int \prod_{i=1}^I \Pr(z_i|\theta) f(\theta) d\theta.$$

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that “local independence follows automatically from unidimensionality” (as cited in Bejar, 1980, p.5). From a dimensionality perspective, there may be nuisance factors that are influencing relationships among certain items, after accounting for the intended construct of interest. These nuisance factors can be influenced by a number of testing features, such as speededness, fatigue, item chaining, and item or response formats (Yen, 1993).

Yen’s  $Q_3$  statistic (Yen, 1984) was used to measure local independence, which was derived from the correlation between the performances of two items. Simply, the  $Q_3$  statistic is the correlation among IRT residuals and is computed using the equation,

$$d_{ij} = u_{ij} - T_i(\hat{\theta}_j),$$

where  $u_{ij}$  is the item score of the  $j$ th test taker for item  $i$ ,  $T_i(\hat{\theta}_j)$  is the estimated true score for item  $i$  of test taker  $j$ , which is defined as

$$T_i(\hat{\theta}_j) = \sum_{l=1}^m y_{il} P_{il}(\hat{\theta}_j),$$

where  $y_{il}$  is the weight for response category  $l$ ,  $m$  is the number of response categories, and  $P_{il}(\hat{\theta}_j)$  is the probability of response category  $l$  to item  $i$  by test taker  $j$  with the ability estimate  $\hat{\theta}_j$ .

The pairwise index of local dependence  $Q_3$  between item  $i$  and item  $i'$  is

$$Q_{3ii'} = r(d_i, d_{i'}),$$

where  $r$  refers to the Pearson product-moment correlation.

When there are  $n$  items,  $n(n-1)/2$ ,  $Q_3$  statistics will be produced. The  $Q_3$  values are expected to be small. Table 16: Biology  $Q_3$  Statistic presents summaries of the distributions of  $Q_3$  statistics—minimum, 5th percentile, median, 95th percentile, and maximum values from each Biology administration. The results show that less than 2% of the items were greater than a critical value of 0.2 for  $|Q_3|$  (Chen & Thissen, 1997).

Table 16: Biology  $Q_3$  Statistic

Administration	Q <sub>3</sub> Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
Fall	-0.264	-0.120	-0.019	0.111	0.238
Winter	-0.459	-0.136	-0.014	0.010	0.669

## 5.4 CONVERGENT AND DISCRIMINANT VALIDITY

According to Standard 1.14 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), it is necessary to provide evidence of convergent and discriminant validity evidence. It is a part of demonstrating validity evidence that assessment scores are related as expected with criteria and other variables for all student groups. However, a second, independent test measuring the same constructs as ELA and Mathematics in Indiana, which could easily permit for a cross-test set of correlations, was not available. Therefore, the correlations between subscores within and across tests were examined alternatively. The *a priori* expectation is that subscores within the same subject (e.g., ELA) will correlate more positively than subscore correlations across subjects (e.g., ELA and Mathematics). These correlations are based on a small number of items, typically around eight to 18; consequently, the observed score correlations will be smaller in magnitude as a result of the very large measurement error at the subscore level. For this reason, both the observed score and the disattenuated correlations are provided.



No other *ILEARN* assessments overlapped with the fall and winter Biology test windows, and these tables are not presented for the 2019-2020 Technical Report.

## 6. FAIRNESS IN CONTENT

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

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

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

### 6.1 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is not sufficient to determine 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 that was used was differential item functioning (DIF). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF, according to the DIF classification convention illustrated in Volume 1 of this technical report. 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, or Female), or negatively (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., White or Male). Items were flagged if their DIF statistics indicated the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. 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 surrounding this review of items for bias is further described in Volume 2, Test Development.

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

- Male/Female;

- White/African-American;
- White/Hispanic;
- White/Asian;
- White/Native American;
- Text-to-Speech (TTS)/Not TTS;
- Student with Special Education (SPED)/Not SPED;
- Title 1/Not Title 1; and
- English Learners (ELs)/Not ELs.

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 4.2, of the *2019–2020 ILEARN Annual Technical Report*. The DIF statistics for each operational test item are presented in the appendix A of Volume 1 of the *2019–2020 ILEARN Annual Technical Report*.

## 7. SUMMARY

This report is intended to provide a collection of reliability and validity evidence to support appropriate inferences from the observed test scores. The overall results can be summarized as follows:

- *Reliability.* Various measures of reliability are provided at the aggregate and subgroup levels, showing the reliability of all tests is in line with acceptable industry standards.
- *Content validity.* Evidence is provided to support the assertion that content coverage on each form was consistent with test specifications of the blueprint across testing modes.
- *Internal structural validity.* Evidence is provided to support the selection of the measurement model, the tenability of local independence, and the reporting of subscores and an overall score at the reporting category levels.

## 8. REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin*, 87(3), 513–524.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Chen, F., Kenneth, A., Bollen, P., Paxton, P., Curran, P. J., & Kirby, J. B. 2001. Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies. *Sociological Methods & Research*, 29, 468–508.
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), 105–146. New York: Macmillan.
- Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, 9, 277–286.
- Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, 11(6).
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.

- Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59(3), 381–389.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, 2(3), 151–160.
- Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12, 237–255.
- Lee, W., Hanson, B., & Brennan, R. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26(4), 412–432.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132.
- Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*, 7th Edition. Los Angeles, CA: Muthén & Muthén.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443–460.
- Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, 8, 111–120.
- Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika*, 42, 549–565.
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14).
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>.
- van Driel, O. P. 1978. "On Various Causes of Improper Solutions in Maximum Likelihood Factor Analysis." *Psychometrika*, 43, 225–243.

- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.
- Yoon, B., & Young, M. J. (2000). *Estimating the reliability for test scores with mixed item formats: Internal consistency and generalizability*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.



**Indiana Learning Evaluation  
Readiness Network  
(*ILEARN*)**

**2019–2020**

**Volume 5  
Score Interpretation Guide**



## **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 the IDOE at [INassessments@doe.in.gov](mailto:INassessments@doe.in.gov).

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Stephan Ahadi, Elizabeth Ayers-Wright, Elizabeth Xiaoxin Wei, Grace Chung, Kevin Clayton, Aleah Pepper, and Kyra Bilenki. Major contributors from the Indiana Department of Education include the Assessment Director, Assistant Assessment Director, and Program Leads.

## TABLE OF CONTENTS

1. INDIANA SCORE REPORTS .....	1
1.1 Overview of Indiana's Score Reports.....	1
1.2 Overall Scores and Reporting Categories.....	2
1.3 Online Reporting System .....	5
1.4 Available Reports on the Indiana Online Reporting System .....	5
1.5 Reporting by Sub-Group .....	6
1.6 Reports .....	8
1.6.1 Summary Performance Report .....	8
1.6.2 Aggregate-Level Subject Report .....	11
1.6.3 Aggregate-Level Reporting Category Report .....	14
1.6.4 Aggregate-Level Standards Report.....	16
1.6.5 Student-Level Subject Report .....	18
1.6.6 Student-Level Reporting Category Report .....	20
1.6.7 Individual Student Report.....	22
1.6.8 Interpretive Guide.....	24
1.6.9 Reports by Sub-Group .....	25
1.6.10 Data File.....	26
2. INTERPRETATION OF REPORTED SCORES .....	27
2.1 Appropriate Uses for Scores and Reports .....	27
2.2 Scale Score .....	28
2.3 Standard Error Measurement .....	29
2.4 Performance Level.....	29
2.5 Performance Category for Reporting Categories.....	29
2.6 Cut Scores .....	30
2.7 Aggregated Scores .....	31
2.8 Writing Performance .....	31
2.9 Relative Strength and Weakness .....	32
2.10 Lexile® Measure.....	32
2.11 Quantile® Measure .....	33
3. SUMMARY .....	34

## LIST OF APPENDICES

Appendix A: Data File Layout

## LIST OF TABLES

Table 1: Reporting Categories for ELA .....	4
Table 2: Reporting Categories for Mathematics .....	4
Table 3: Reporting Categories for Science .....	4
Table 4: Reporting Categories for Social Studies .....	5
Table 5: Indiana Score Reports Summary .....	6
Table 6: Indiana List of Sub-Groups .....	7
Table 7: <i>ILEARN</i> ELA Assessment Proficiency Cut Scores .....	30
Table 8: <i>ILEARN</i> Mathematics Assessment Proficiency Cut Scores .....	30
Table 9: <i>ILEARN</i> Science Assessment Proficiency Cut Scores .....	31
Table 10: <i>ILEARN</i> Social Studies Grade 5 Assessment Proficiency Cut Scores.....	31
Table 11: <i>ILEARN</i> U.S. Government Assessment Proficiency Cut Scores .....	31
Table 12: Writing Scoring Dimensions.....	32

## LIST OF FIGURES

Figure 1: Sample State Summary Performance Report .....	9
Figure 2: Sample Corporation-Level Summary Performance Report .....	10
Figure 3: Corporation Aggregate-Level Subject Report, Biology .....	12
Figure 4: Corporation Aggregate-Level Subject Report with Sub-group, Biology .....	13
Figure 5: Corporation Aggregate-Level Reporting Category Report, Biology .....	15
Figure 6: Sample Corporation Aggregate-Level Standards Report, Biology.....	17
Figure 7: Student-Level Subject Report, Biology .....	19
Figure 8: Student-Level Reporting Category Report, Biology .....	21
Figure 9: Individual Student Report, Biology.....	23
Figure 10: Supplemental Interpretive Guide .....	24
Figure 11: Corporation Aggregate-Level Subject Report by Gender, Biology .....	25
Figure 12: Data File .....	26

## 1. INDIANA SCORE REPORTS

During school year 2019-2020, pursuant to IC 20-32-5, Biology *ILEARN* assessments were administered to Indiana students during fall and winter test windows. Due to school closures resulting from the COVID-19 pandemic, no spring assessments were administered in grades 3–8 English/Language Arts (ELA) and Mathematics; grades 4 and 6 Science and Biology; and grade 5 Social Studies and U.S. Government.

The purpose of this volume is to document the features of the Indiana Online Reporting System (ORS), which is designed to assist stakeholders in reviewing and downloading the test results and in understanding and appropriately using the results of the state assessments. Additionally, this volume of the technical report describes the score types reported for the 2019-2020 assessments, the features of the score reports, and the appropriate uses and inferences that can be drawn from those score types.

This volume describes the features of ORS for all administrations. No reporting occurred for the Spring 2020 assessments due to the cancellation of that test administration. However, results from the fall and winter Biology test administrations were reported as described below.

### 1.1 OVERVIEW OF INDIANA’S SCORE REPORTS

Test scores from the fall and winter Biology assessment windows were provided to corporations and schools through the ORS. The ORS provides information on student performance and aggregated summaries at several levels—state, corporation, school, and teacher/roster.

The ORS (<https://in.reports.cambiumast.com>) is a web-based application that provides *ILEARN* results at various, privileged levels. Test results are available for users based on their roles and the privileges determined by the authentication granted to them. There are three basic levels of user roles: the corporation, school, and teacher (classroom) levels. Each user is granted drill-down access to reports in the system based on his or her assigned role. This means that teachers can access data for only their roster(s) of students, schools can access data for only the students in their school, and corporations can access data for all schools and students in their corporation.

To access ORS, users must be added to the Test Information Distribution Engine (TIDE). Test coordinators add users to TIDE at the corporation and school level. The following user roles have access to ORS:

- State users: access to all state, corporation, school, teacher, and student test data.
- Co-Op role and Corporation Test Coordinator (CTC): access to all test data for their corporation and for the schools and students in their corporation.
- School Test Coordinator (STC) and Principal (PR): access to all test data for their school and the students in their school.
- Test Administrator (TA): access to all aggregated test data for their rosters and the students within their rosters.

Access to reports is password protected, and users can access data at their assigned level and below. For example, an STC user can access the school report of students for their school but not for another school.

## 1.2 OVERALL SCORES AND REPORTING CATEGORIES

Each student receives a single scale score for each subject tested if there is a valid score to report. Normally, a student takes a test in the Test Delivery System (TDS) and then submits it. TDS then forwards the test for scoring before the ORS reports the scores. However, tests may also be manually invalidated before reaching the ORS if testing irregularities occur (e.g., cheating, unscheduled interruptions, loss of power or Internet).

The validity of a score is determined using invalidation rules, which define a set of parameters under which a student's assessment may be counted. A student's score will be automatically invalidated if they fail to respond to at least five test items. When a student receives an accommodation for which he or she is not eligible or is otherwise impacted by an irregularity that affects the validity of the student's assessment attempt, the student's test is invalidated. Within ORS, "Invalidated" will appear in lieu of score data for the student.

A student's score is based on the operational items on the assessment that they attempted. A scale score is used to describe how well a student performed on a test and is an estimate of a student's knowledge and skills measured. The scale score is transformed from a theta score, which is estimated based on Item Response Theory (IRT) models as described in Volume 1 of this technical report. Lower scale scores indicate less mastery of the grade-level knowledge and skills measured by the test. Conversely, higher scale scores indicate more mastery of the grade-level knowledge and skills measured by the test. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and performance-level descriptors.

Performance-level descriptors (PLDs) define the content area knowledge and skills that students at each performance level are expected to demonstrate. PLDs exist at different levels of precision for different uses. Policy PLDs are overarching, high-level statements that reflect the varying degrees to which students may demonstrate proficiency on each grade-level *ILEARN* assessment. The policy PLDs were written first, and a diverse panel of Indiana educators was convened to consider many factors as they defined each Policy PLD. Educators were also enlisted to develop Range PLDs for the *ILEARN* assessments. Range PLDs are content-specific statements that reflect the varying degrees to which students may demonstrate proficiency on grade-level standards on the *ILEARN* assessments. The Indiana Policy and grade and subject Range PLDs can be found on the IDOE website (<https://www.doe.in.gov/assessment/ilearn-sample-items-and-scoring>).

Based on the scale score, a student will receive an overall performance level. The *ILEARN* scale has been divided into four performance levels, defined by descriptors and cut scores that indicate four levels of proficiency as follows:

- Level 1: Below Proficiency;
- Level 2: Approaching Proficiency;
- Level 3: At Proficiency; or
- Level 4: Above Proficiency.

The *ILEARN* U.S. Government scale scores are mapped into two performance levels:

- Level 1: Below Proficiency; or
- Level 2: At Proficiency.

Each student is assigned a performance level based on their score compared to the cut scores and defined by the PLDs. Cut points are listed in Section 2.5. Generally, students performing on *ILEARN* at Levels 3 and 4 are considered on track to demonstrate progress toward mastery of the knowledge, application, and analytical skills necessary for college and career readiness.

In addition to an overall score, students will receive reporting category scores. Reporting categories (also known as strands or subscores) represent distinct groups of knowledge within each grade and subject. For *ILEARN*, students' performance on each reporting category is reported using three performance categories:

- Below
- At/Near
- Above

Unlike the performance levels for the overall test, student performance on each of the reporting categories is evaluated entirely with respect to meeting the reporting category proficiency cut score. Performance-level classifications are computed to classify student performance levels for each of the domain or reporting category subscales. For each subscale, the band is generally defined as a range extending 1.5 Standard Error of Measurement (SEM) below to 1.5 SEM above the proficiency cut score used on the overall test.

Students performing at either Below or Above can be interpreted as “student performance clearly below or above the Meets Standard cut score for a specific reporting category.” Students performing at At/Near can be interpreted as “student performances that do not provide enough information to tell whether students reached the Meets Standard mark for the specific reporting category.”

Table 1 through Table 4 display the reporting categories by grade and subject.

*Table 1: Reporting Categories for ELA*

Grade	Reporting Category
3–5	Key Ideas and Textual Support/Vocabulary Structural Elements and Organization/Connection of Ideas/Media Literacy Writing
6–8	Key Ideas and Textual Support/Vocabulary Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy Writing

*Table 2: Reporting Categories for Mathematics*

Grade	Reporting Category
3–4	Algebraic Thinking and Data Analysis Computation Geometry and Measurement Number Sense
5	Algebraic Thinking Computation Geometry and Measurement, Data Analysis, and Statistics Number Sense
6	Algebra and Functions Data Analysis, Statistics, and Probability Geometry and Measurement Number Sense and Computation
7–8	Algebra and Functions Data Analysis, Statistics, and Probability Geometry and Measurement Number Sense and Computation

*Table 3: Reporting Categories for Science*

Grade	Reporting Category
4, 6	Questioning and Modeling Investigating Analyzing, Interpreting, and Computational Thinking Explaining Solutions, Reasoning, and Communicating
Biology	Developing and Using Models to Describe Structure and Function Developing and Using Models to Explain Processes Analyzing Data and Mathematical Thinking Constructing and Communicating an Explanation Evaluating Claims with Evidence

*Table 4: Reporting Categories for Social Studies*

Grade	Reporting Category
5	Civics and Government Geography and Economics History
U.S. Government	Functions of Government Historical Foundations of American Government Institutions and Processes of Government

### 1.3 ONLINE REPORTING SYSTEM

ORS generates a set of online score reports that describes student performance for students, parents, educators, and other stakeholders. The online score reports are produced after the tests are submitted by the students, hand-scored and machine-scored, and processed into the ORS. In addition to each individual student's score report, the ORS produces aggregate score reports for teachers, schools, corporations, and states. The timely accessibility of aggregate score reports helps users monitor student group performance in each subject and grade area, evaluate the effectiveness of instructional strategies, and inform the adoption of strategies to improve student learning and teaching during the school year.

Furthermore, to facilitate comparisons, each aggregate report contains the summary results for the selected aggregate unit, as well as all aggregate units above the selected aggregate. For example, if a school is selected, the summary results of the corporations to which the school belongs and the summary results of the state are also provided. This occurs so that the school's performance can be compared with the corporation's performance and the state's performance. If a teacher is selected, the summary results for the school, corporation, and state above the teacher are also provided for comparison purposes. Table 5 (in Section 1.4) lists the types of online reports and the levels at which they can be viewed (student, roster, teacher, school, and corporation).

### 1.4 AVAILABLE REPORTS ON THE INDIANA ONLINE REPORTING SYSTEM

ORS is hierarchically structured. An authorized user can view reports at their own aggregated unit and any lower level of aggregation. For example, a school user can view only the reports and data at the school and student levels of his or her school. Co-Op and CTC users can view the reports and data for their corporations and the student-level results for all their schools.

Table 5 summarizes the types of score reports that are available in the ORS and the levels at which the reports can be viewed. A description of each report is also provided. Data files are also accessible for corporations to download.



For detailed information on available reports and features, educators can refer to the ORS user guide. The *Indiana State Assessment Online Reporting System User Guide* can be found online at:

[https://ilearn.portal.cambiumast.com/core/fileparse.php/4152/urlt/ORS\\_Guide\\_FINAL\\_Rebranded\\_Approved.pdf](https://ilearn.portal.cambiumast.com/core/fileparse.php/4152/urlt/ORS_Guide_FINAL_Rebranded_Approved.pdf)

*Table 5: Indiana Score Reports Summary*

Report	Description	Level of Availability				
		State	Corporation	School	Roster	Student/ Parent
<b>Summary Performance</b>	Summary of performance (to date) across grades and subjects or courses for the current administration	✓	✓	✓	✓	
<b>Aggregate-Level Subject Report</b>	Summary of overall performance for a subject and a grade for all students in the defined level of aggregation	✓	✓	✓	✓	
<b>Aggregate-Level Reporting Category Report</b>	Summary of overall performance on each reporting category for a given subject and grade across all students within the selected level of aggregation	✓	✓	✓	✓	
<b>Student-Level Subject Report</b>	List of all students who belong to a school, teacher/roster with their associated subject or course scores for the current administration			✓	✓	
<b>Student-Level Reporting Category Report</b>	List of all students who belong to a school, teacher/roster with their associated reporting category performance for the current administration			✓	✓	
<b>Individual Student Report (ISR)</b>	Detailed information about a selected student's performance in a specified subject or course; includes overall subject and reporting category results					✓
<b>Data Files</b>	Text/CSV files containing overall and reporting category scale scores and performance levels along with demographic information		✓	✓	✓	

## 1.5 REPORTING BY SUB-GROUP

The aggregate score reports at the overall subject level and reporting category level provide overall student results by default but can at any time be analyzed by sub-groups based on demographic data. When used on aggregate-level reports, an additional level of analysis will be provided by aggregating students based on sub-group. For example, when the “Gender” sub-group is selected, the ORS will display aggregate results by *all* students, *male* students, and *female* students. When used on student-level reports, sub-groups can instead filter individual results. For example, a user will have the option to select “Male” or “Female” after the “Gender” sub-group is selected.

Users can see student assessment results by any sub-group at any time by selecting the desired sub-group from the “Breakdown By” drop-down list available. Table 6 presents the types of sub-groups and sub-group categories provided in the ORS.

*Table 6: Indiana List of Sub-Groups*

Sub-Group	Sub-Group Category
Ethnicity	White
	Black/African American
	Hispanic
	Asian
	American Indian/Alaska Native
	Native Hawaiian/Other Pacific Islander
	Multiracial/Two or More Races
Gender	Male
	Female
English Learner	English Learner
	Not English Learner
Special Education	Special Education
	Not Special Education
Section 504 Plan	Section 504 Plan
	Not Section 504 Plan
Socioeconomic Status	Yes
	No
Home Language	English
	Arabic
	Burmese
	Mandarin
	Spanish
	Vietnamese
Grade	Grade 3
	Grade 4
	Grade 5
	Grade 6
	Grade 7
	Grade 8
	Grade 9
	Grade 10
	Grade 11
	Grade 12

## **1.6 REPORTS**

### **1.6.1 Summary Performance Report**

The home page allows authorized users to log in to the ORS and select “Score Reports,” which contains summaries of student performance across grades and subjects. State personnel can view state summaries, corporation personnel see corporation summaries, school personnel see school summaries, and teachers see student summaries. State users can view a summary of students’ performance within each corporation, as well. The Summary Performance Report:

- Displays summary data separated by grade and subject;
- Bases the level of aggregation on a user’s role; and
- Reports the number of students tested and percentage proficient.

The Summary Performance Report provides summaries of student performance, including:

- Number of students tested; and
- Percentage proficient.

Figure 1 and Figure 2 present sample Summary Performance Reports at the state and corporation level.

Figure 1: Sample State Summary Performance Report

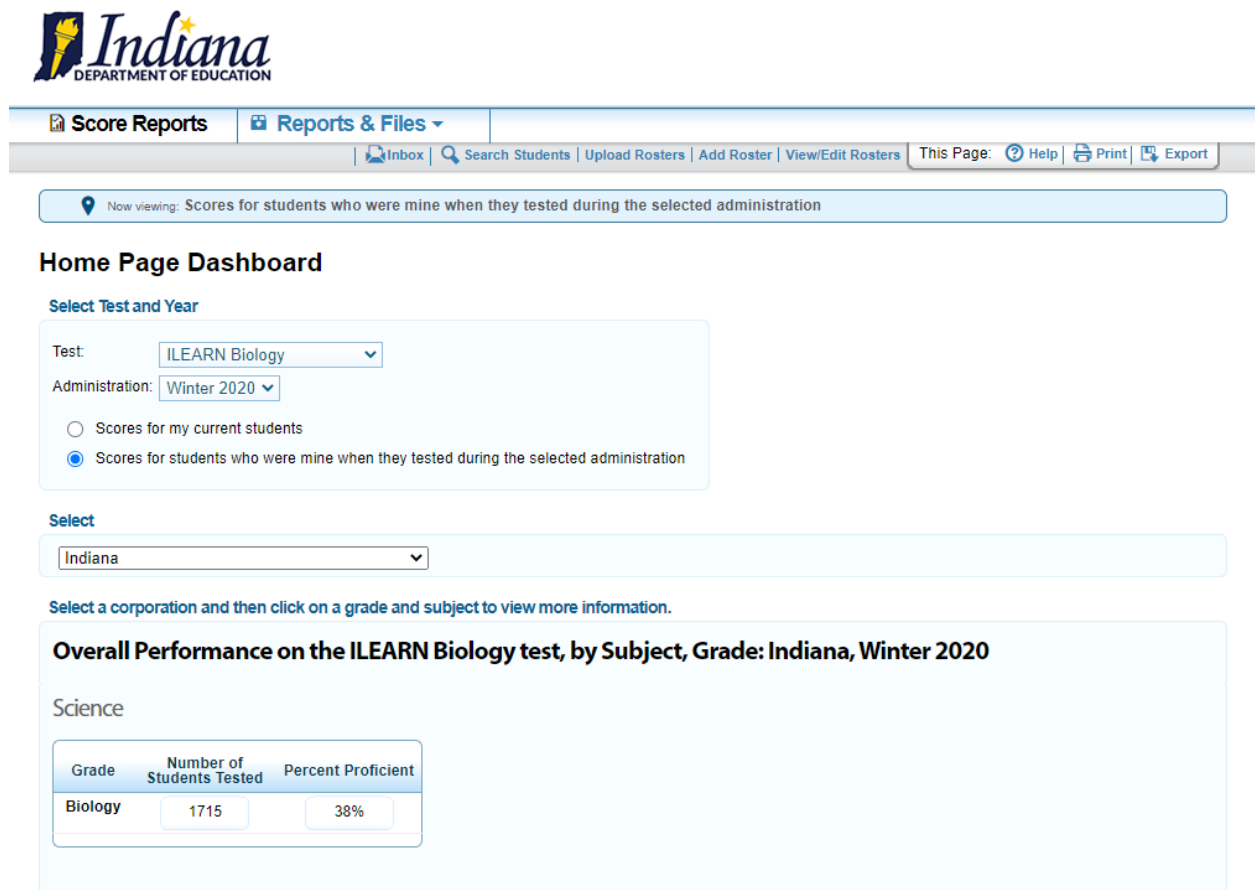
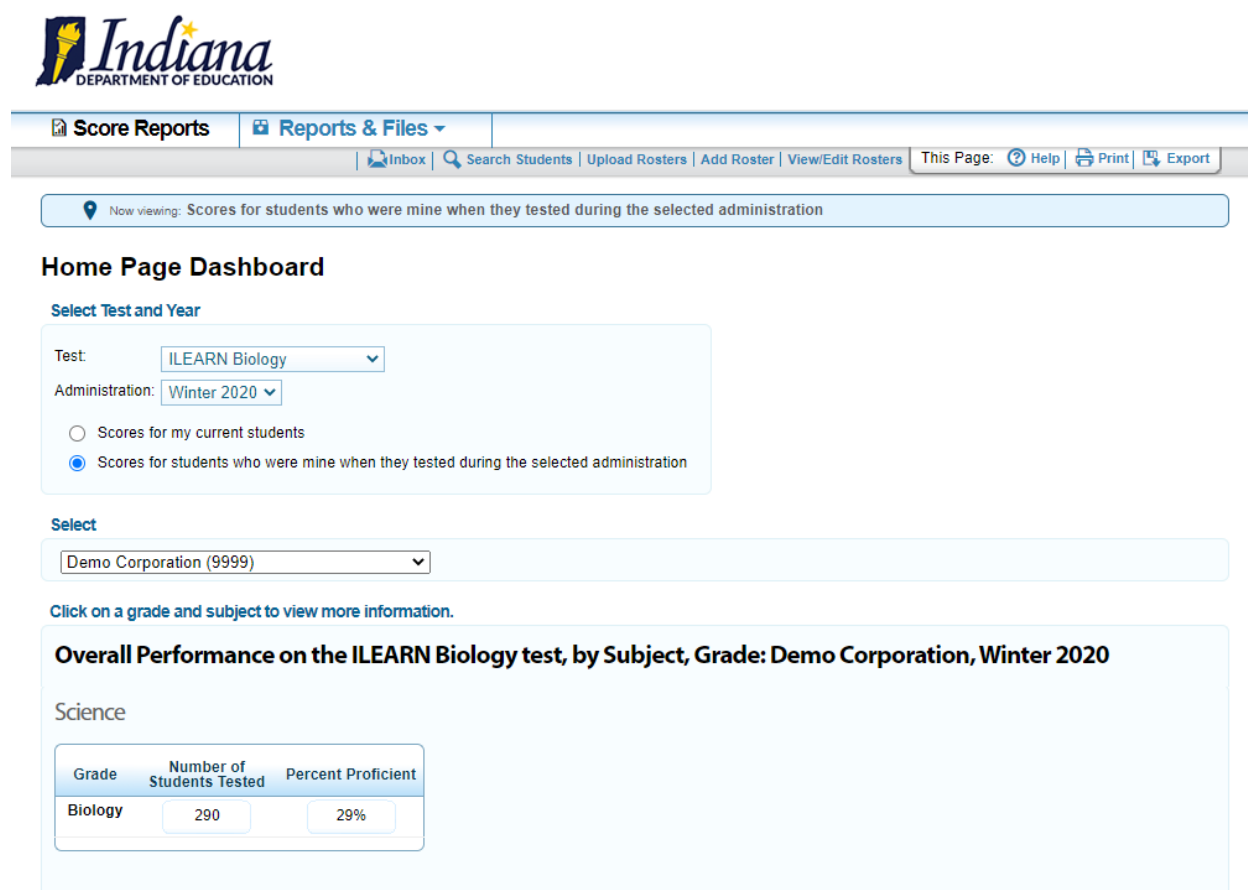


Figure 2: Sample Corporation-Level Summary Performance Report



**Indiana**  
DEPARTMENT OF EDUCATION

**Score Reports** **Reports & Files**

[Inbox](#) [Search Students](#) [Upload Rosters](#) [Add Roster](#) [View/Edit Rosters](#) This Page: [Help](#) [Print](#) [Export](#)

Now viewing: Scores for students who were mine when they tested during the selected administration

### Home Page Dashboard

**Select Test and Year**

Test: ILEARN Biology

Administration: Winter 2020

☐ Scores for my current students

☒ Scores for students who were mine when they tested during the selected administration

**Select**

Demo Corporation (9999)

[Click on a grade and subject to view more information.](#)

### Overall Performance on the ILEARN Biology test, by Subject, Grade: Demo Corporation, Winter 2020

Science

Grade	Number of Students Tested	Percent Proficient
Biology	290	29%

The Corporation Summary Report is similar to the State Summary Report, except that summary data are displayed for all students in the selected corporation who have completed the selected test with a valid reported score.

## **1.6.2 Aggregate-Level Subject Report**

Detailed summaries of student performance within a grade and subject area are available within the Aggregate-Level Subject Report. The Aggregate-Level Subject Report presents results for the aggregate unit as well as the results for the state and any higher-level aggregate units. For example, a school Aggregate-Level Subject Report will also contain the summary results of the state and school corporation so that school performance can be compared with the above aggregate levels.

The Aggregate-Level Subject Report provides the aggregate summaries on a specific subject area, including:

- Number of students;
- Average scale score;
- Percentage proficient;
- Percent of students in each proficiency level; and
- Number of students in each proficiency level.

The summaries are also presented for overall students and by sub-groups. Figure 3 presents an example of Aggregate-Level Subject Reports for Biology at the corporation level without sub-groups. Figure 4 highlights Biology at the corporation level when a user selects a sub-group of gender.

Figure 3: Corporation Aggregate-Level Subject Report, Biology

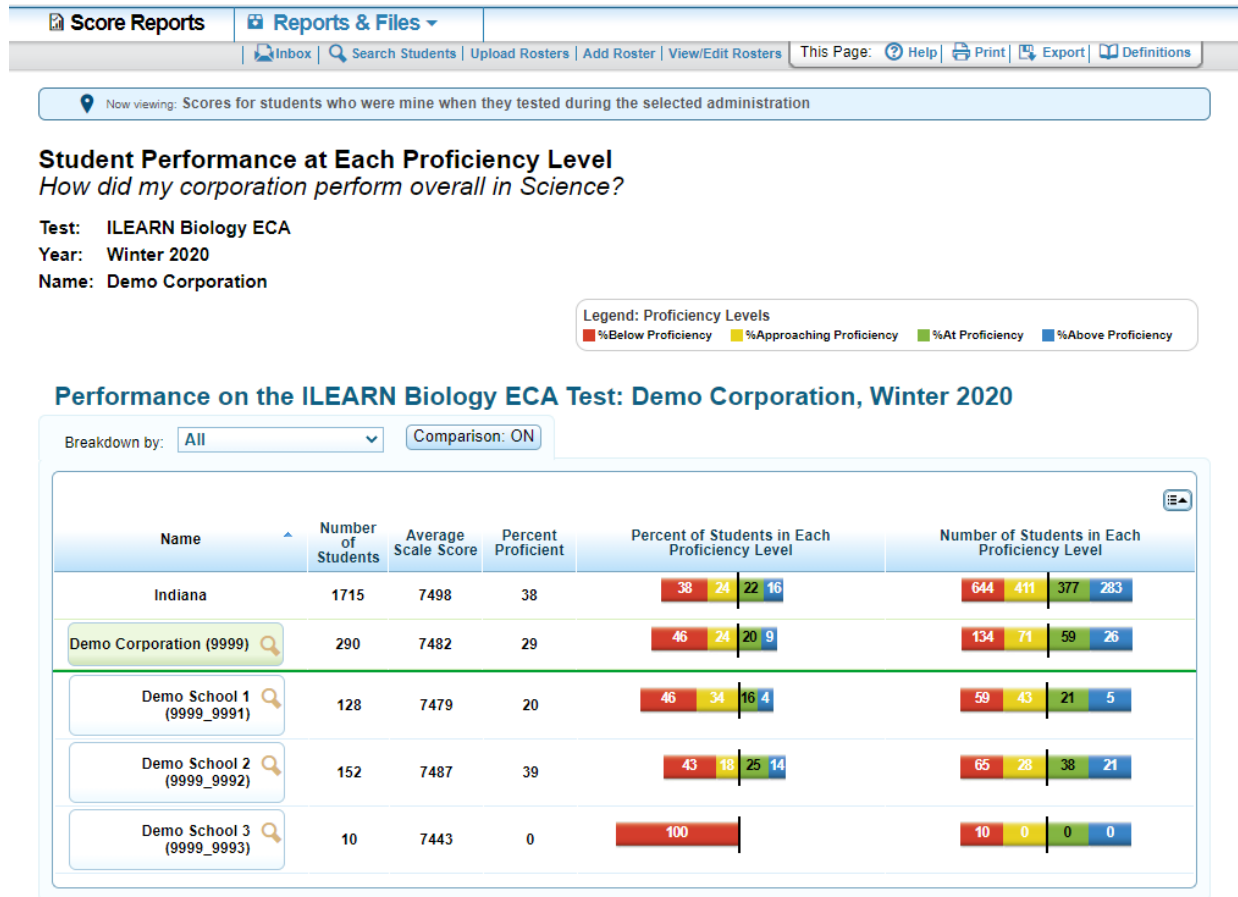
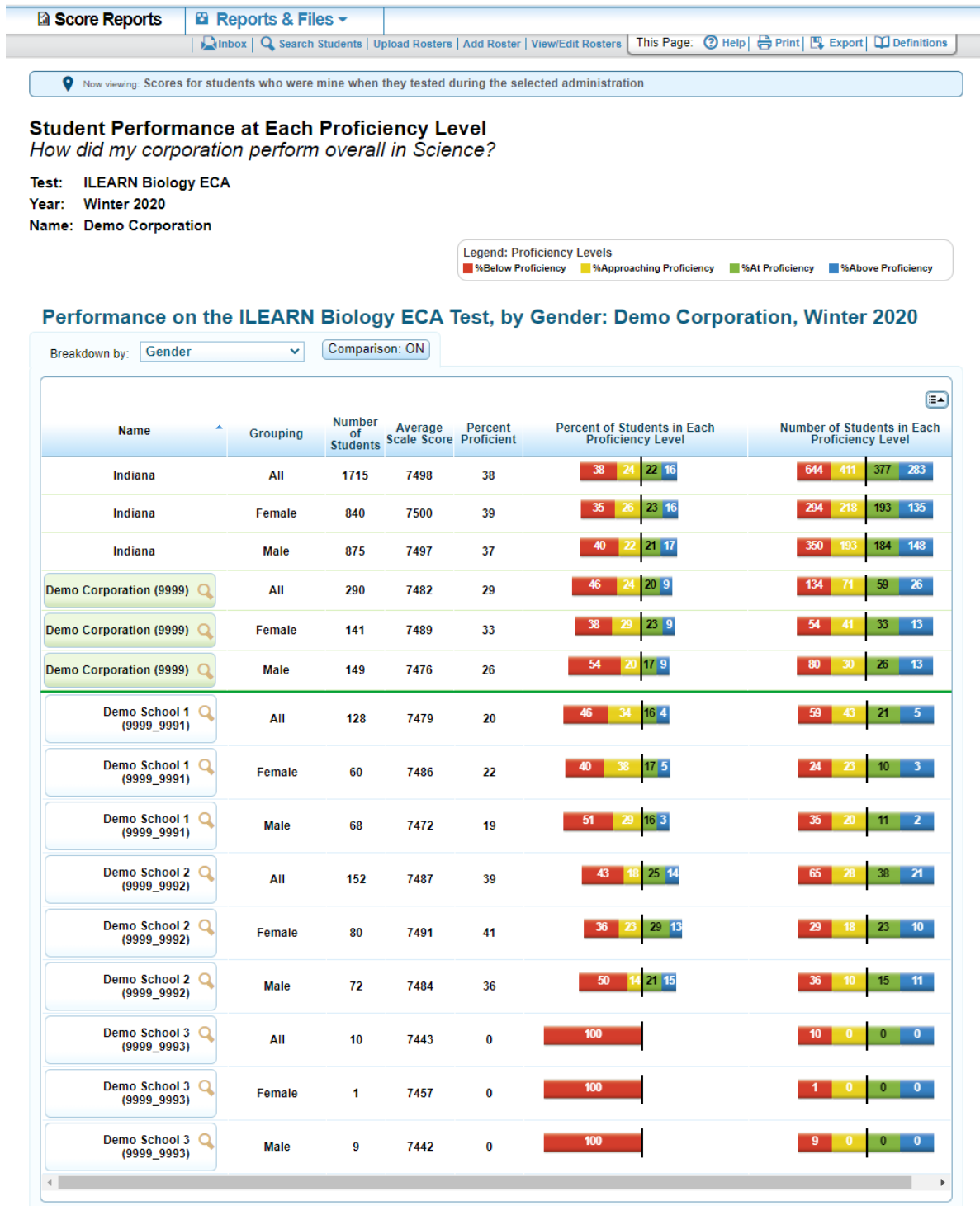


Figure 4: Corporation Aggregate-Level Subject Report with Sub-group, Biology





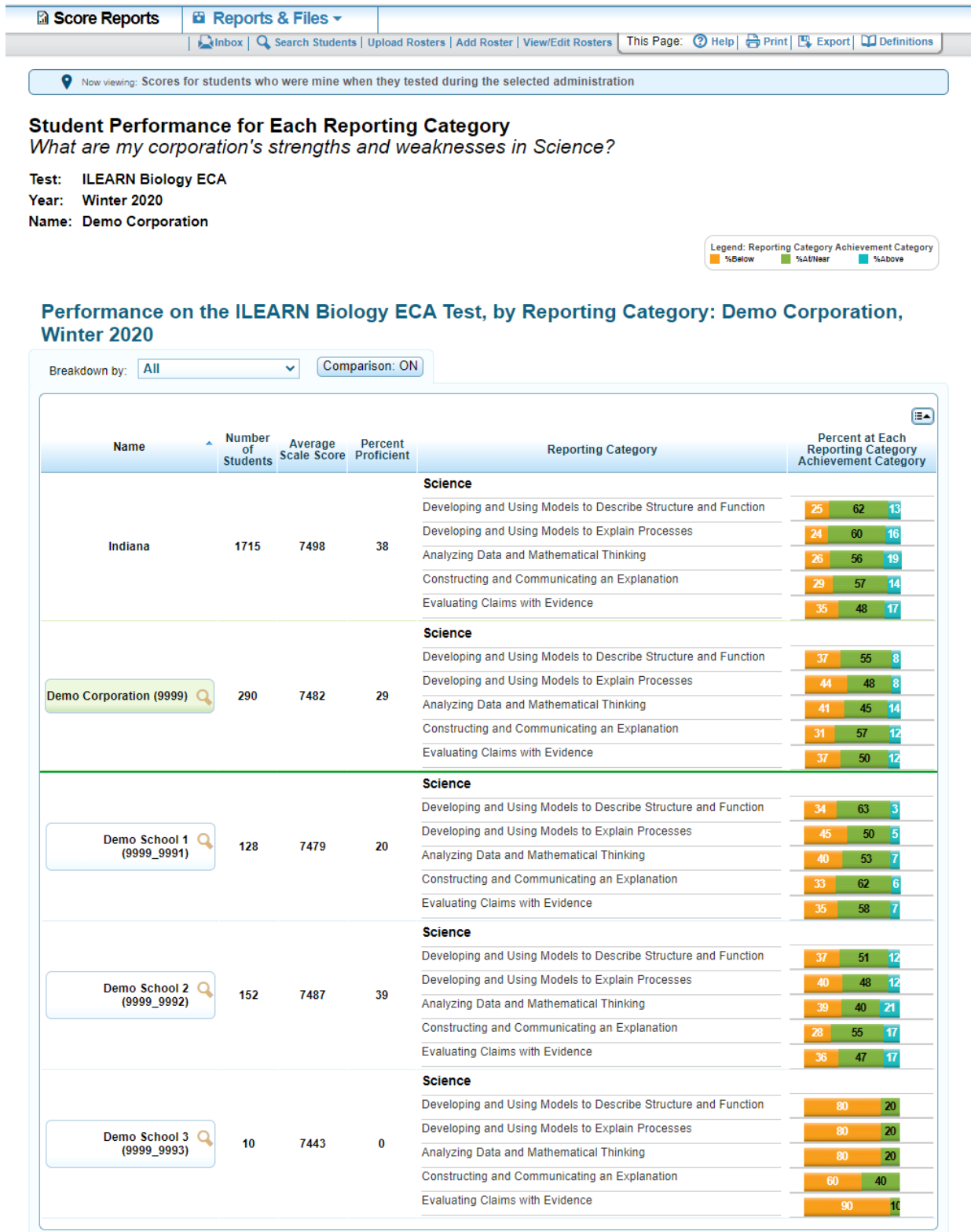
### **1.6.3 Aggregate-Level Reporting Category Report**

The Aggregate-Level Reporting Category Report provides the aggregate summaries on student performance in each reporting category for a particular grade and subject. The summaries on the Aggregate-Level Reporting Category Report include:

- Number of students;
- Average scale score;
- Overall percentage proficient; and
- For each reporting category, the percentage of students in each achievement category.

Similar to the Aggregate-Level Subject Report, this report presents the summary results for the selected aggregate unit as well as the summary results for the state and the aggregate unit above the selected aggregate. In addition, summaries can be presented for all students within an aggregate and by students within a defined sub-group. Figure 5 presents an example of the Corporation Aggregate-Level Reporting Category Report for *ILEARN* Biology.

Figure 5: Corporation Aggregate-Level Reporting Category Report, Biology



### 1.6.4 Aggregate-Level Standards Report

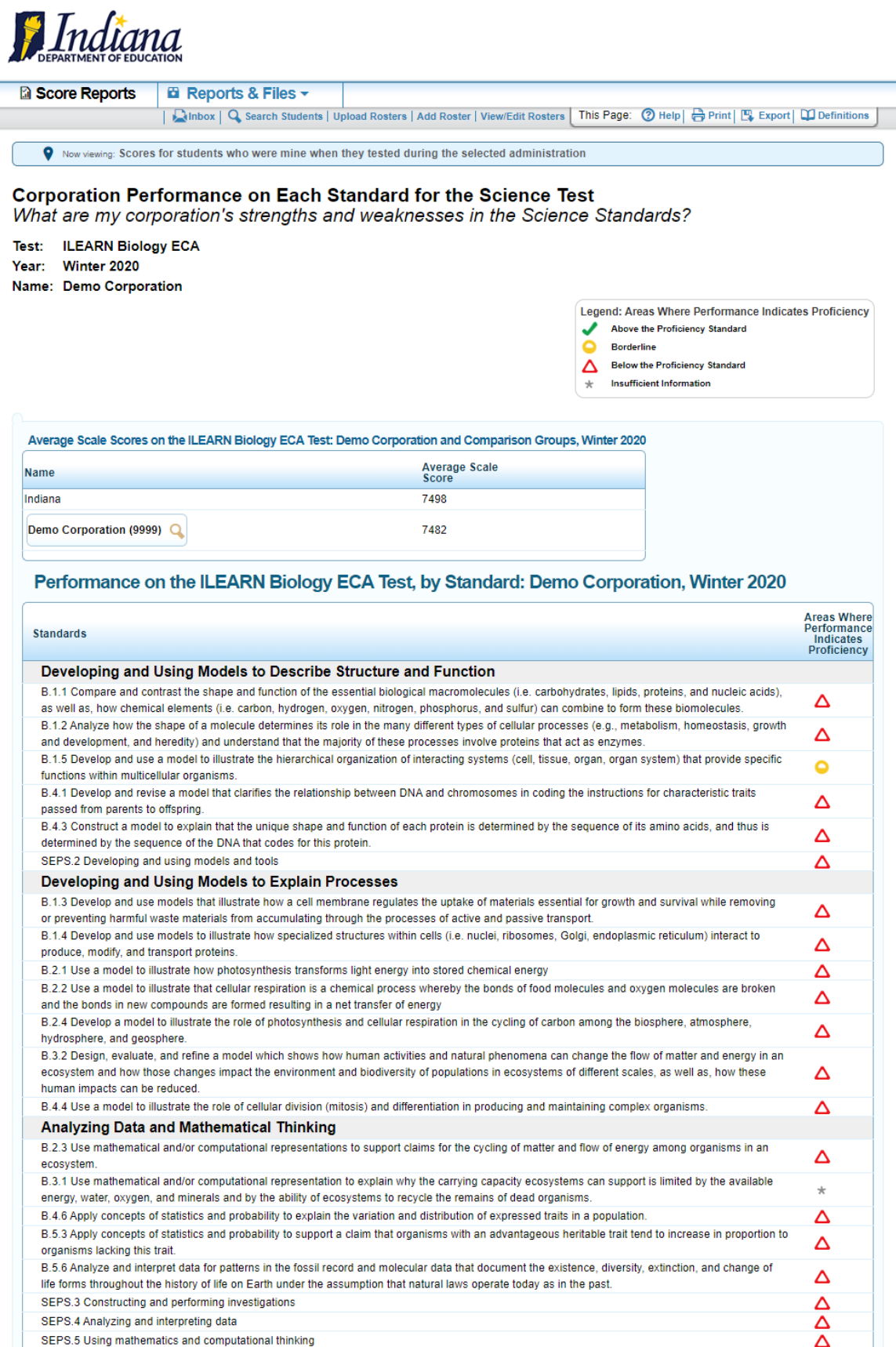
The Aggregate-Level Standards Report lists data on the performance of student groups on each standard of a subject for the current testing window and reports the following measures for the selected level of aggregation:

- Areas Where Performance Indicates Proficiency.

For adaptive assessments, a standard performance indicator produces information on how a group of students in a school or corporation performed on the standard compared to the proficiency cut. For “Areas Where Performance Indicates Proficiency,” a performance indicator produces information on how a group of students in a school or corporation performed on the standard compared to the proficiency cuts. It shows whether performance on this standard for this group was above, no different from, or below what is expected of students at the proficient level. It also indicates (with an asterisk) if enough information was available to determine whether performance on the standard was above, no different from, or below the proficiency standard. An asterisk will appear if the standard was not assessed within the aggregate group or not enough students received and responded to items that measured the standard. This indicator shows strengths and weaknesses for a group of students and is provided only at an aggregate level, because it is unstable at the individual level.

Figure 6 presents an example of the Aggregate-Level Standards Report for Biology.

Figure 6: Sample Corporation Aggregate-Level Standards Report, Biology



Constructing and Communicating an Explanation	
B.4.2 Construct an explanation for how the structure of DNA determines the structure of proteins which carry out the essential functions of life through systems of specialized cells	△
B.5.2 Communicate scientific information that common ancestry and biological evolution are supported by multiple lines of empirical evidence including both anatomical and molecular evidence.	△
B.5.5 Construct an explanation based on evidence that the process of evolution primarily results from four factors: (1) the potential for a species to increase in number, (2) the heritable genetic variation of individuals in a species due to mutation and sexual reproduction, (3) competition for limited resources, and (4) the proliferation of those organisms that are better able to survive and reproduce in the environment.	△
SEPS.1 Posing questions (for science) and defining problems (for engineering)	△
SEPS.6 Constructing explanations (for science) and designing solutions (for engineering)	△
SEPS.8 Obtaining, evaluating, and communicating information	△
Evaluating Claims with Evidence	
B.3.3 Evaluate the claims, evidence, and reasoning that the complex interactions in ecosystems maintain relatively consistent numbers and types of organisms in stable conditions, and identify the impact of changing conditions or introducing non-native species into that ecosystem.	△
B.4.5 Make and defend a claim based on evidence that inheritable genetic variations may result from: (1) new genetic combinations through meiosis, (2) viable errors occurring during replication, and (3) mutations caused by environmental factors.	△
B.5.1 Evaluate anatomical and molecular evidence to provide an explanation of how organisms are classified and named based on their evolutionary relationships into taxonomic categories.	△
B.5.4 Evaluate evidence to explain the role of natural selection as an evolutionary mechanism that leads to the adaptation of species, and to support claims that changes in environmental conditions may result in: (1) increases in the number of individuals of some species, (2) the emergence of new species over time, and/or (3) the extinction of other species.	△
SEPS.7 Engaging in argument from evidence	△

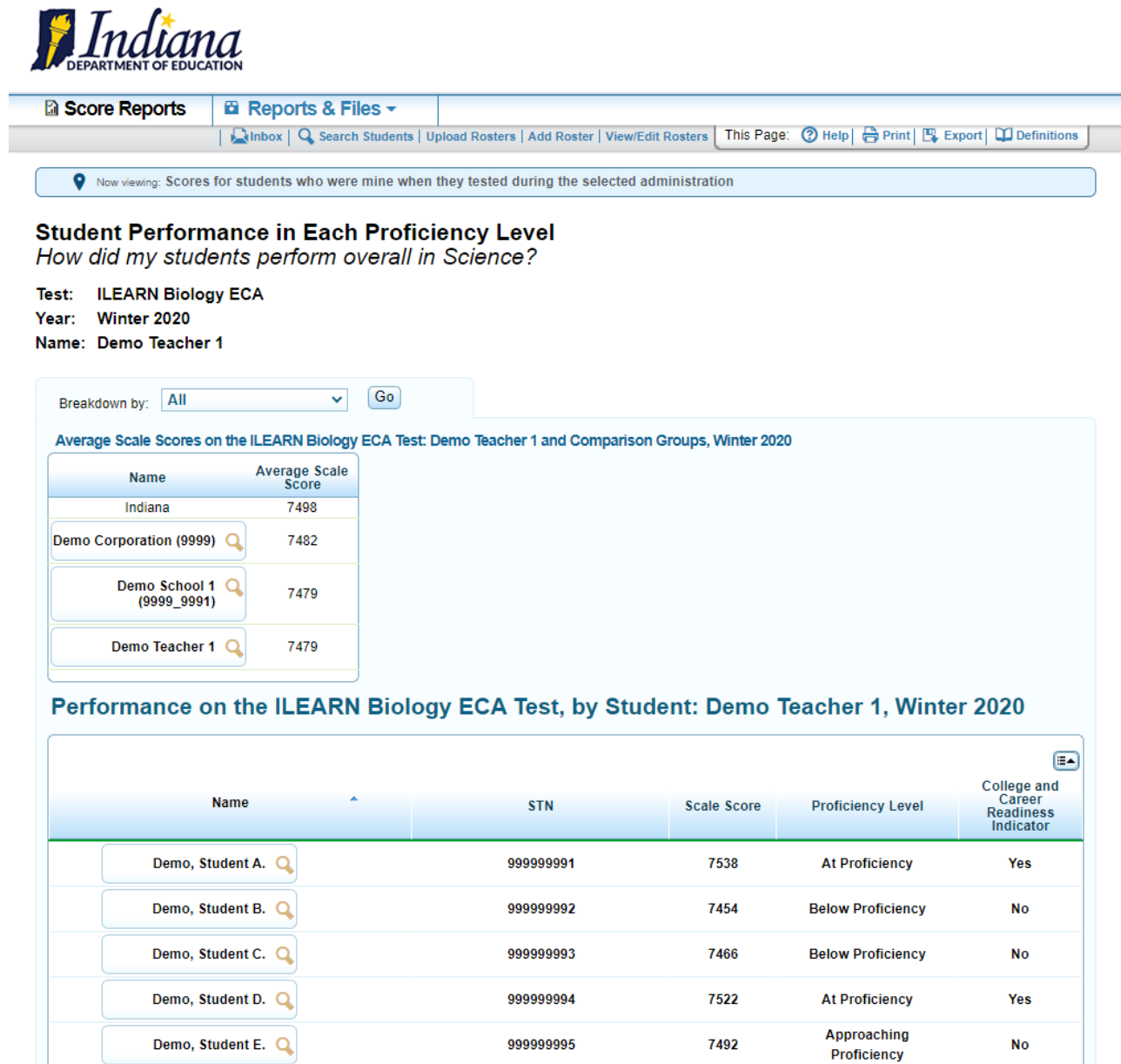
## 1.6.5 Student-Level Subject Report

The Student-Level Subject Report lists all students who belong to the selected aggregate level, such as a school, and reports the following measures for each student:

- Scale score;
- Overall subject performance level;
- Lexile® (for ELA) or Quantile® (for Mathematics) measure; and
- College and Career Readiness indicator.

Figure 7 demonstrates an example of the Student-Level Subject Report for Biology. Please note that Lexile® (for ELA) or Quantile® (for Mathematics) measures are not applicable to Biology and are not included in the screenshot below.

Figure 7: Student-Level Subject Report, Biology



### **1.6.6 Student-Level Reporting Category Report**

The Student-Level Reporting Category Report lists all students who belong to the selected aggregate level, such as a school, and reports the following measures for each student:

- Scale score;
- Overall subject performance level;
- College and Career Readiness indicator;
- Reporting category; and
- Performance category.

Figure 8 displays this information for *ILEARN* Biology.

Figure 8: Student-Level Reporting Category Report, Biology



**Score Reports** | **Reports & Files** ▾

[Inbox](#) | [Search Students](#) | [Upload Rosters](#) | [Add Roster](#) | [View/Edit Rosters](#) | This Page: [Help](#) | [Print](#) | [Export](#) | [Definitions](#)

Now viewing: Scores for students who were mine when they tested during the selected administration

### Student Performance on Each Reporting Category

How did my students perform on the Science test?

Test: ILEARN Biology ECA

Year: Winter 2020

Name: Demo Teacher 1

Legend: Reporting Category Achievement Category

Below At/Near Above

Breakdown by: **All** Go

#### Average Scale Scores on the ILEARN Biology ECA Test: Demo Teacher 1 and Comparison Groups, Winter 2020

Name	Average Scale Score
Indiana	7498
Demo Corporation (9999)	7482
Demo School 1 (9999_9991)	7479
Demo Teacher 1	7479

#### Performance on the ILEARN Biology ECA Test, by Student, Reporting Category: Demo Teacher 1, Winter 2020

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator	Developing and Using Models to Describe Structure and Function Achievement Category	Developing and Using Models to Explain Processes Achievement Category	Analyzing Data and Mathematical Thinking Achievement Category	Constructing and Communicating an Explanation Achievement Category
Demo, Student A.	999999991	7538	At Proficiency	Yes	At/Near	At/Near	Above	At/Near
Demo, Student B.	999999992	7454	Below Proficiency	No	Below	Below	Below	At/Near
Demo, Student C.	999999993	7466	Below Proficiency	No	At/Near	Below	Below	At/Near
Demo, Student D.	999999994	7522	At Proficiency	Yes	At/Near	At/Near	Above	At/Near
Demo, Student E.	999999995	7492	Approaching Proficiency	No	At/Near	At/Near	At/Near	At/Near



## 1.6.7 Individual Student Report

When a student receives a valid test score, an ISR can be generated in the ORS. The ISR contains the following measures:

- Scale score;
- Overall subject performance level;
- College and Career Readiness Indicator;
- Lexile® (ELA only) or Quantile® (Mathematics only);
- Average scale scores for the state and the student's corporation and school;
- Performance level in each reporting category; and
- Writing performance descriptors in each dimension (ELA only).

The top of the report includes:

- Student's name;
- Scale score;
- Proficiency level;
- Lexile® (ELA only) or Quantile® (Mathematics only); and
- College and Career Readiness Indicator.

The middle section includes:

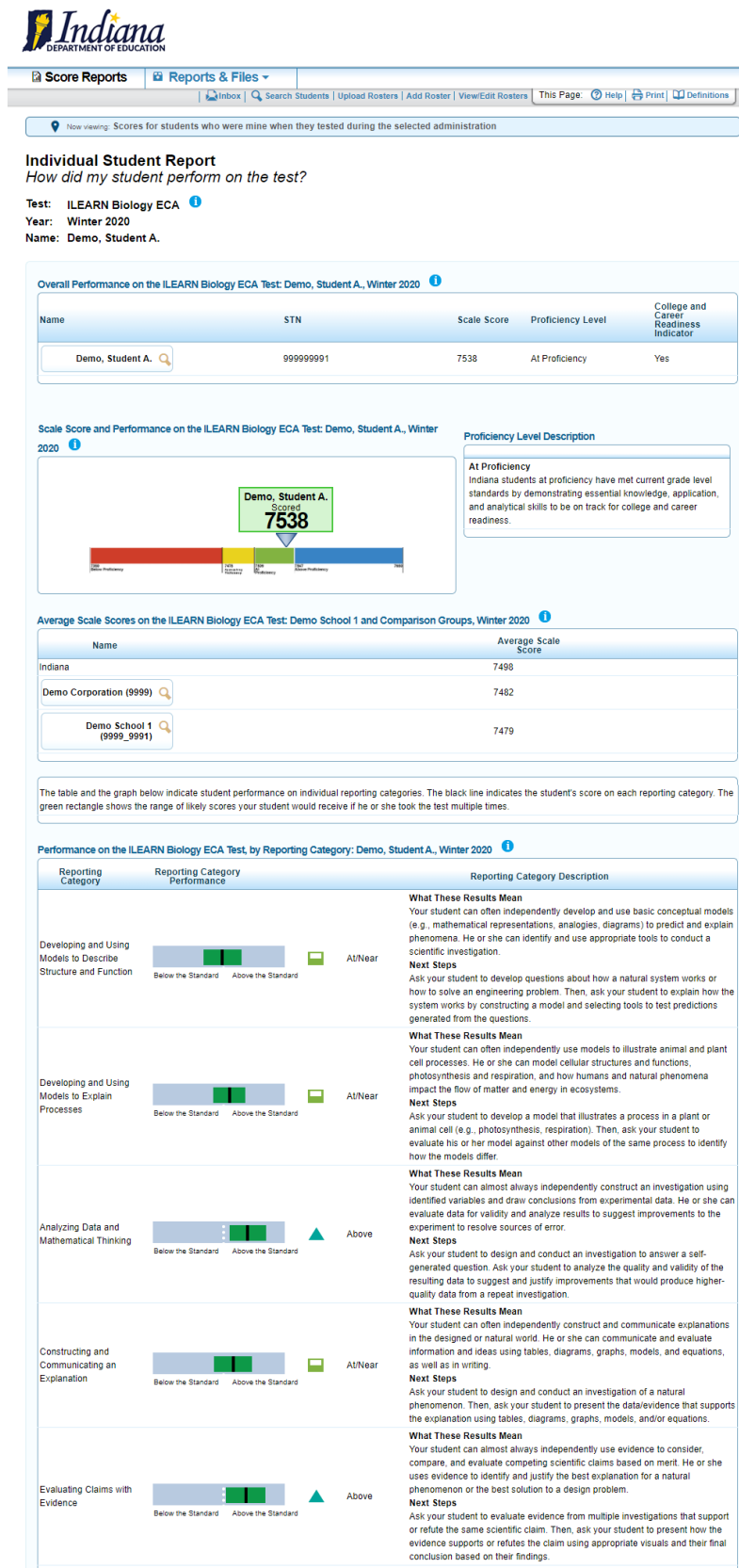
- Bar chart with the student's scale score;
- Performance-level descriptors with cut scores at each performance level; and
- Average scale scores for state, corporation, and school aggregation levels.

The bottom of the report includes:

- Detailed information on student performance on each reporting category.
  - *Note: Bar charts in the reporting category table show how students performed on each reporting category (black bar) relative to the reporting category performance standard (dashed white line). Green boxes show the score range the student would likely fall within if he or she took the test multiple times.*
- Writing dimension scores (ELA only) along with a performance description for each writing dimension.

Figure 9 presents an example ISR for ILEARN Biology. Please note that Lexile® (for ELA) or Quantile® (for Mathematics) measures are not applicable to Biology and are not included in the screenshot below.

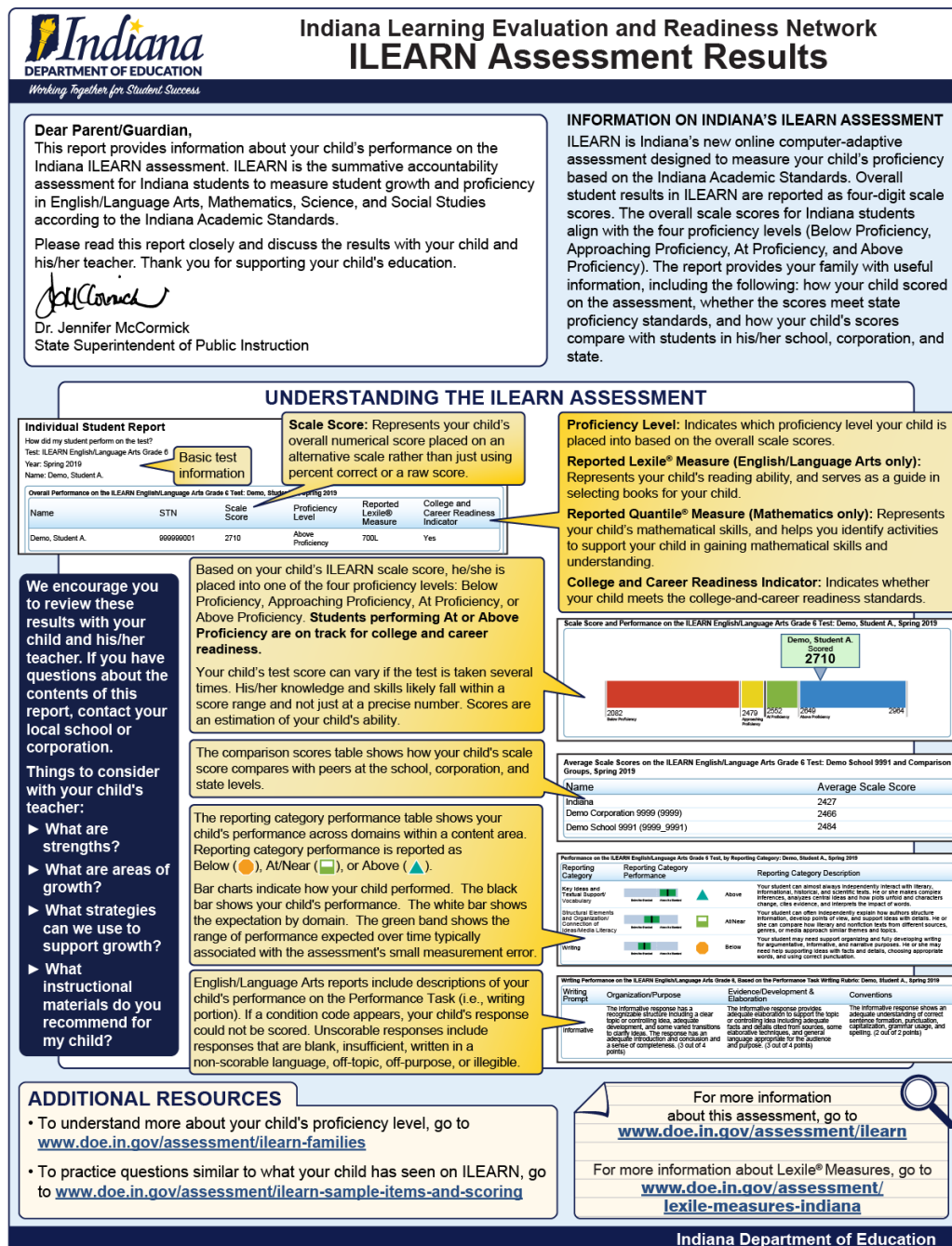
Figure 9: Individual Student Report, Biology



# 1.6.8 Interpretive Guide

When printing ISRs, users have the option to print a supplemental “interpretive guide” (also called an “Addendum” when printing a Simple ISR), which is intended to serve as a stand-alone document (see Figure 10) to help teachers, administrators, parents, and students better understand the data presented in the ISR. The ISRs and the supplemental “interpretive guide” are also available in five different languages: Arabic, Mandarin, Burmese, Spanish, and Vietnamese.

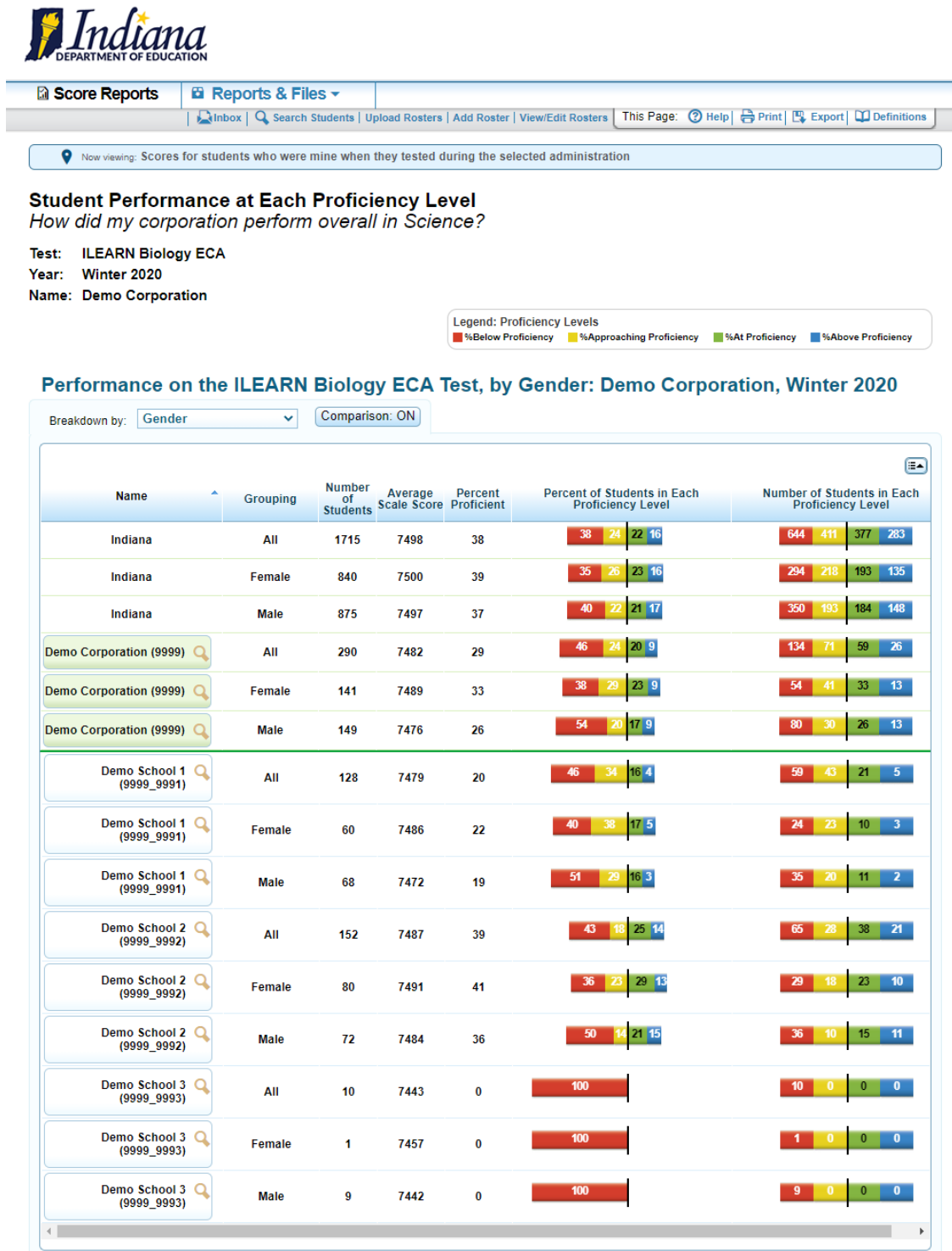
Figure 10: Supplemental Interpretive Guide



## 1.6.9 Reports by Sub-Group

At the aggregate level, student performance can be broken down by demographic sub-groups, such as gender (Figure 11).

Figure 11: Corporation Aggregate-Level Subject Report by Gender, Biology



## 1.6.10 Data File

ORS users have the option to quickly generate a comprehensive data file of their students' scores. Data files (see Figure 12) can be downloaded in Microsoft Excel or CSV format and contain a wide variety of data, including overall subject and reporting category scores, demographic data, and performance levels. Data files can be useful as a resource for further analysis and can be generated at the corporation, school, teacher, or roster level. The data file layout can be found in Appendix A, and contains the data column names, descriptions, acceptable values, and indicates for which grades and subjects each data column appears.

Figure 12: Data File

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Gender	Ethnicity	Special Ed Identified	Section 504	Socioecon	Enrolled Gr	Enrolled School	Enrolled School ID	Enrolled Corp	Enrolled C	Test name	Overall sc	Overall pr	Reported	College ar	Passing St	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting
2	F	Multiracia	N	N	N	Y	9 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7504	Approaching Profici	No				At/Near			At/Near	
3	F	White	N	N	N	N	11 Demo School 2	9999_9992	Demo Corpor	9999	ILEARN Biology ECA	7526	At Proficiency	Yes				Above			At/Near	
4	F	Multiracia	N	N	N	Y	10 Demo School 2	9999_9992	Demo Corpor	9999	ILEARN Biology ECA	7448	Below Proficiency	No				At/Near			Below	
5	F	White	N	N	N	Y	9 Demo School 2	9999_9992	Demo Corpor	9999	ILEARN Biology ECA	7495	Approaching Profici	No				At/Near			At/Near	
6	M	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7466	Below Proficiency	No				At/Near			Below	
7	F	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7522	At Proficiency	Yes				At/Near			At/Near	
8	F	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7492	Approaching Profici	No				At/Near			At/Near	
9	F	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7506	Approaching Profici	No				At/Near			At/Near	
10	M	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7468	Below Proficiency	No				At/Near			Below	
11	F	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7484	Approaching Profici	No				At/Near			At/Near	
12	F	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7524	At Proficiency	Yes				At/Near			At/Near	
13	M	Black/Afri	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7431	Below Proficiency	No				Below			Below	
14	F	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7463	Below Proficiency	No				At/Near			Below	
15	F	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7463	Below Proficiency	No				At/Near			Below	
16	F	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7496	Approaching Profici	No				At/Near			Below	
17	F	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7519	At Proficiency	Yes				At/Near			At/Near	
18	M	Hispanic	Y	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7440	Below Proficiency	No				At/Near			Below	
19	F	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7480	Approaching Profici	No				Below			At/Near	
20	M	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7437	Below Proficiency	No				Below			Below	
21	F	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7453	Below Proficiency	No				Below			Below	
22	M	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7475	Below Proficiency	No				At/Near			At/Near	
23	M	White	N	N	N	N	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7454	Below Proficiency	No				At/Near			Below	
24	F	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7463	Below Proficiency	No				Below			Below	
25	M	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7459	Below Proficiency	No				Below			At/Near	
26	M	White	N	N	N	Y	10 Demo School 1	9999_9991	Demo Corpor	9999	ILEARN Biology ECA	7418	Below Proficiency	No				Below			Below	

## **2. INTERPRETATION OF REPORTED SCORES**

A student's performance on a test is reported as a scale score and a performance level for the overall test, and also as a separate performance level for each reporting category. Students' scores and performance levels are summarized at the aggregate level. This section describes how to interpret these scores.

### **2.1 APPROPRIATE USES FOR SCORES AND REPORTS**

The primary intended use of the *ILEARN* assessment system is for school accountability, to ensure that educators, schools, and corporations are providing effective instruction of the Indiana Academic Standards. For the adaptive assessments (ELA, Mathematics, and Science during school year 2019-2020), even though each individual student is administered only a sample of items measuring each subject area, at the aggregate levels of classroom, teacher, school, and corporation, student achievement is assessed across the full range of items measuring knowledge and skills of each subject.

Assessment results on student performance on the test can be used to help teachers or schools make decisions on how to support students' learning. Aggregate score reports on the teacher and school level provide information about the strengths and weaknesses of students and can be used to improve teaching and student learning. For example, a group of students may have performed well overall but not as well in one or more reporting categories. In this case, teachers or schools can identify the strengths and weaknesses of their students through the group performance by reporting category and promote instruction on specific areas where student performance is below overall performance. Furthermore, by narrowing the student performance result by sub-group, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from disadvantaged sub-groups. For example, teachers might see student assessment results by gender and observe that a particular group of students is struggling with literary response and analysis in reading. Teachers can then provide additional instructions for these students to enhance their performance on the benchmarks for literary response and analysis.

In addition, assessment results can be used to compare students' performance among different students and different groups. Teachers can evaluate how their students perform compared with other students in schools and corporations by overall scores and reporting category scores. Furthermore, scale scores can be used to measure the growth of individual students over time, if data are available. The *ILEARN* scale score is on a vertical scale for ELA and Mathematics, which means scales are vertically linked across grades, and scores across grades are on the same scale. Therefore, ELA and Mathematics scale scores are comparable across grades so that scale scores from one grade can be compared with the next. Science and Social Studies scale scores are reported on separate within-test scales, and cross-grade comparisons are not appropriate.

Assessment results can be used to provide information on individual students' performance on the test. Overall, assessment results demonstrate what students know and can do in certain subject areas and give further information on whether students are on track to demonstrate the knowledge and skills necessary for college and career readiness. Additionally, assessment results can be used to identify a student's relative strengths and

weaknesses in certain content areas. For example, performance levels for reporting categories can be used to identify an individual student's relative strengths and weaknesses among reporting categories within a content or subject area.

Although 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 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; users need to consider measurement error when using student scores to make decisions about student performance. 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 performance, such as classroom assessment and teacher evaluation, should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to take into account the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus requiring interpretation with more caution.

## **2.2 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 a student's knowledge and skills as measured by their performance on the test. A scale score is the student's overall numeric score. *ILEARN* scale scores are reported on a vertical scale for ELA and Mathematics based on the vertical scale established by Smarter Balanced, which means that scores from different grades can be compared within the same tested subject. The vertical scale was formed by linking tests across grades using common items, and a statistical relationship is then determined. A vertical linking study provides the relationship among adjacent grade levels, allowing for meaningful comparisons across grades and, by extension, tracking growth over time as a student or cohort advances through each grade level (see Section 6.2 in Volume 1 of this technical report for more information). Science and Social Studies scale scores are reported on separate within-test scales, and cross-grade comparisons are not appropriate.

Scale scores can be used to illustrate students' current levels of performance and are powerful when used to measure their growth over time. Lower scale scores can indicate that the student does not possess sufficient knowledge and skills measured by the test. Conversely, higher scale scores can indicate that the student has proficient knowledge and skills measured by the test. When combined across a student population, scale scores can also describe school and corporation-level changes in performance and reveal gaps in performance among different groups of students. In addition, scale scores can be averaged across groups of students, allowing educators to use group comparison. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and performance-level descriptors. It should be noted that the utility of scale scores is limited when comparing smaller differences among scores (or averaged group scores), particularly when the difference among scores is within the SEM. Furthermore, the scale score of individual students should be cautiously interpreted when comparing two scale scores, because small differences in scores may not reflect real differences in performance.



## 2.3 STANDARD ERROR MEASUREMENT

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and are not just precise numbers. A scale score (the observed score on any test) is an estimate of the true score. A test contains items that sample a student's knowledge and skills; if a student takes a similar test several times, the resulting scale scores would vary across administrations, sometimes being a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered several times. The SEM can be interpreted as the degree of uncertainty of a student's score based on a statistical analysis of the student's answers on a test. When interpreting scale scores, it is recommended to always consider the range of scale scores incorporating the SEM of the scale score.

## 2.4 PERFORMANCE LEVEL

Based on their scale score, a student will receive an overall performance level. Using performance standards (or cut scores—see Section 2.5) *ILEARN* scale scores are mapped into four performance levels:

- Level 1: Below Proficiency;
- Level 2: Approaching Proficiency;
- Level 3: At Proficiency; and
- Level 4: Above Proficiency.

U.S. Government scale scores are mapped into two performance levels:

- Level 1: Below Proficiency; and
- Level 2: At Proficiency.

Performance -level descriptors are descriptions of content area knowledge and skills that students at each performance level are expected to possess. Thus, performance levels can be interpreted based on performance-level descriptors. Students performing on *ILEARN* at Levels 3 and 4 are considered to have met or mastered current grade level standards by demonstrating essential knowledge, application, and analytical skills to be on track for college and career readiness. Because performance levels are for the classification of students into a small number of groups, such as those comprising four or five students, and based on the cut scores, they have limited use for measuring growth. Thus, the performance level is an indicator of whether a student has mastered the required skill for a given level.

Performance-level descriptors are available on the IDOE web page at <https://www.doe.in.gov/assessment/ilearn-sample-items-and-scoring>.

## 2.5 PERFORMANCE CATEGORY FOR REPORTING CATEGORIES

Students' performance on each reporting category is reported on three performance categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Students



performing at Below Standard or Above Standard can be interpreted as student performances clearly below or above the Meets Standard cut score for a specific reporting category. Students performing at At/Near Standard can be interpreted as student performances that are close to the cut score, but there is not enough information to determine if it is above or below. Performance levels for the reporting category are limited in their diagnostic ability based on the degree of the calculated SEM of the student’s scale score for the tested grade and subject.

## 2.6 CUT SCORES

For all grades and subjects within *ILEARN*, scale scores are mapped into four performance levels. U.S. Government scale scores are mapped into two performance levels. For each performance level, there is a minimum and maximum scale score that defines the range of scale scores students within each performance level have achieved. Collectively, these minimum and maximum scale scores are defined as “cut scores” and are the cutoff points for each performance level. Table 7 through Table 11 shows the cut scores for *ILEARN*.

*Table 7: ILEARN ELA Assessment Proficiency Cut Scores*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
3	5060–5415	5416–5459	5460–5514	5515–5760
4	5090–5443	5444–5492	5493–5546	5547–5810
5	5110–5471	5472–5523	5524–5594	5595–5850
6	5130–5491	5492–5543	5544–5603	5604–5870
7	5130–5506	5507–5567	5568–5628	5629–5890
8	5150–5510	5511–5576	5577–5637	5638–5920

*Table 8: ILEARN Mathematics Assessment Proficiency Cut Scores*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
3	6080–6381	6382–6424	6425–6487	6488–6730
4	6100–6428	6429–6473	6474–6540	6541–6800
5	6110–6452	6453–6509	6510–6565	6566–6850
6	6110–6487	6488–6544	6545–6604	6605–6870
7	6120–6492	6493–6561	6562–6624	6625–6920
8	6120–6508	6509–6589	6590–6650	6651–6950

*Table 9: ILEARN Science Assessment Proficiency Cut Scores*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
4	7350–7481	7482–7505	7506–7534	7535–7650
6	7350–7465	7466–7503	7504–7544	7545–7650
Biology	7350–7477	7478–7508	7509–7546	7547–7650

*Table 10: ILEARN Social Studies Grade 5 Assessment Proficiency Cut Scores*

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
5	8350–8476	8477–8501	8502–8542	8543–8650

*Table 11: ILEARN U.S. Government Assessment Proficiency Cut Scores*

Grade	Level 1 Below Proficiency	Level 2 At Proficiency
U.S. Government	8350–8496	8497–8650

## 2.7 AGGREGATED SCORES

Students' scale scores are aggregated at teacher/roster, school, corporation, and state levels to represent how a group of students performs on a test. When students' scale scores are aggregated, the aggregated scale scores can be interpreted as an estimate of knowledge and skills that a group of students possesses. This interpretation makes aggregated scores a powerful tool when comparing student performance across different groups of students, whether it be at a similar level of aggregation (e.g., school to school) or an analysis of a sub-group (e.g., comparing a teacher's roster to the overall school).

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 is reported at the aggregate level to represent how well a group of students performs overall and by reporting category.

## 2.8 WRITING PERFORMANCE

ELA reports include descriptions of the student's performance on the writing portion based on the performance task writing rubric for each criterion. Essay responses are scored on three dimensions: Organization/Purpose, Evidence/Development and Elaboration, and Conventions, as Table 12 shows. Each of these dimensions is independently scored and

reported on the student reports. For item analysis Organization/Purpose and Evidence/Development and Elaboration are averaged and rounded to an integer. Thus, the overall writing prompt score will range from 0 to 6.

A condition code is assigned to a student's written response that could not be scored, based on set criteria. Unscorable responses include responses that are blank, insufficient, written in a language other than English, off topic, illegible, or off-purpose. It should be noted that the reporting category score for writing consists of the overall writing score from the prompt and stand-alone writing items.

*Table 12: Writing Scoring Dimensions*

Dimension	Possible Scores
Organization/Purpose	1–4 points
Evidence/Development and Elaboration	1–4 points
Conventions	0–2 points

## 2.9 RELATIVE STRENGTH AND WEAKNESS

For standard performance, relative strengths and weaknesses at each standard are reported for aggregate levels only (e.g., school, or corporation). Because an individual student responds to too few items within a standard to generate reliable data, the standard performance is produced by aggregating all items within a standard across students at an aggregate level. Standard reports include data on Performance Relative to Proficiency for each standard.

The Performance Relative to Proficiency data for a standard show how a group of students performed in each standard relative to the expected performance for proficiency. For summative tests, this is the expected level of performance necessary to achieve Level 3 performance. This is a standards-based report with the group performance in each standard being compared to the performance standard for that standard. Similar to the performance levels provided for the total test, these data indicate students' achievements in the standard with respect to the standards. Because the Performance Relative to Proficiency data for each standard are comparable to the standards-based expectations, performance across groups can be compared.

## 2.10 LEXILE® MEASURE

The Lexile® framework uses quantitative methods, based on individual words and sentence lengths, rather than qualitative analysis of content to produce scores. A Lexile® measure is defined as “the numeric representation of an individual's reading ability or a text's readability (or difficulty), followed by an ‘L’ (Lexile®).” A Lexile® text measure is obtained by evaluating the readability of a piece of text, such as a book or an article. A Lexile® measure of a text can assist in selecting targeted materials that present an appropriate level of challenge for a reader—not too difficult to be frustrating, yet difficult enough to challenge a reader and encourage reading growth.

## **2.11 QUANTILE® MEASURE**

Quantile® measures provide an alternative—and possibly more useful—measure of Mathematics ability than grade-equivalent scores. Similar to the Lexile® framework, the Quantile® framework measures both the mathematics skill level of a student and the difficulty of Mathematics skills and concepts on the same developmental scale. Quantile® measures help educators, parents, and students determine which skills and concepts they are ready to learn next. Mathematics skills and concepts content, such as Mathematics textbooks and online instructional materials, also get a Quantile® measure. Using these two measures together, parents and teachers can match students with resources that help them connect the dots among different Mathematics skills and concepts and build on their learning.

### **3. SUMMARY**

*ILEARN* results are reported online via the Online Reporting System (ORS). The results are released after score QC has been completed to align with a date approved by the state.

The ORS is interactive. When educators or administrators log in, they see a summary of data about students for whom they are responsible (a principal would see the students in his or her school; a teacher would see students in his or her class). They can then drill down through various levels of aggregation all the way to individual reports. The system allows them to tailor the content more precisely, moving from subject area through reporting categories and even to standards-level reports for aggregates. Aggregate reports are available at every level, and authorized users can print these or download them (or the data on which they are based). Individual student reports (ISRs) can be produced individually or batched as PDF reports.

All authorized users can download files, including data about students for whom they are responsible, at any time. The various reports available may be used to inform stakeholders regarding student performance and instructional strategies.