



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

2021–2022

**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.

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Stephan Ahadi, Shuqin Tao, Elizabeth Xiaoxin Wei, Maryam Pezeshki, Kevin Clayton, Christina Sneed, and Jessica Singh. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and program leads.

TABLE OF CONTENTS

1.	INTRODUCTION	1
1.1	Test Background and Historical Context	1
1.2	Purpose and Intended Uses of the ILEARN Assessments	1
1.3	Participants in ILEARN Development and Analysis	2
1.4	Available Test Formats and Special Versions	3
1.5	Student Participation	4
2.	SUMMARY OF OPERATIONAL PROCEDURES	8
2.1	Administration Procedures	8
2.2	Universal Features, Designated Features, and Accommodations	8
3.	MAINTENANCE OF THE ITEM BANK.....	10
3.1	Overview of Item Development.....	10
3.2	Review of Operational Items	10
3.3	Field Testing.....	10
3.4	Operational Form Construction and Adaptive Simulations	11
4.	CLASSICAL ANALYSES OVERVIEW	12
4.1	Classical Item Analyses.....	12
4.1.1	<i>Item Discrimination</i>	13
4.1.2	<i>Distractor Analysis</i>	13
4.1.3	<i>Item Difficulty</i>	13
4.1.4	<i>Mean Total Score</i>	14
4.2	Differential Item Functioning Analysis.....	14
4.3	Item Analyses Results.....	17
5.	ITEM CALIBRATION	18
5.1	Item Response Theory Models.....	18
5.1.1	<i>ELA and Mathematics</i>	19
5.1.2	<i>Science</i>	19
5.1.3	<i>Social Studies</i>	19
5.2	IRT Analyses Results	20

5.2.1	<i>IRT Summaries</i>	20
5.2.2	<i>2021 ILEARN Test Characteristic Curves</i>	22
6.1	Maximum Likelihood Estimation	23
6.1.1	<i>Likelihood Function</i>	23
6.1.2	<i>Derivatives</i>	23
6.1.3	<i>Standard Errors of Estimates</i>	24
6.1.4	<i>Extreme Case Handling</i>	25
6.1.5	<i>Standard Errors of LOT/HOT Scores</i>	26
6.2	Transforming Theta Scores to Reporting Scale Scores.....	26
6.3	Overall Performance Classification.....	27
6.4	Reporting Category Scores	28
6.4.1	<i>MLE and MMLE Scoring</i>	28
6.4.2	<i>Strengths and Weaknesses</i>	29
6.4.3	<i>Standard-Level Aggregate Scores</i>	29
6.5	Lexile® and Quantile® Scores	31
6.6	Comparison of Scores to Previous Year.....	31
7.	QUALITY ASSURANCE PROCEDURES	32
7.1	Quality Assurance in Test Configuration	32
7.2	Quality Assurance in Computer-Delivered Test Production.....	33
7.2.1	<i>Production of Content</i>	33
7.2.2	<i>Web Approval of Content During Development</i>	33
7.2.3	<i>Platform Review</i>	34
7.2.4	<i>User Acceptance Testing and Final Review</i>	34
7.2.5	<i>Functionality and Configuration</i>	35
7.3	Quality Assurance in Data Preparation.....	36
7.4	Quality Assurance in Item Analysis and Equating	37
7.5	Quality Assurance in Scoring and Reporting	37

LIST OF TABLES

Table 1: Required Uses and Citations of ILEARN	2
Table 2: Number of Students Participating in ILEARN 2021–2022	5
Table 3: Distribution of Demographic Characteristics of Tested Population, ELA	5
Table 4: Distribution of Demographic Characteristics of Tested Population, Mathematics	6
Table 5: Distribution of Demographic Characteristics of Tested Population, Science	7
Table 6: Distribution of Demographic Characteristics of Tested Population, Social Studies.....	7
Table 7: 2021–2022 ILEARN Testing Windows	8
Table 8: Evaluative Criteria in Classical Item Analysis	12
Table 9: DIF Classification Rules	17
Table 10: Operational Item p-Value Five-Point Summary and Range, Social Studies ..	17
Table 11: N Students Used in Field-Test Calibrations.....	20
Table 12: Operational Item Parameter 5-Point Summary and Range, ELA	20
Table 13: Operational Item Parameter 5-Point Summary and Range, Mathematics	21
Table 14: Operational Item Parameter 5-Point Summary and Range, Science	21
Table 15: Operational Item Parameter 5-Point Summary and Range, Social Studies ..	22
Table 16: Theta and Scaled-Score Limits for Extreme Ability Estimates, ELA.....	25
Table 17: Theta and Scaled-Score Limits for Extreme Ability Estimates, Mathematics	25
Table 18: Theta and Scaled-Score Limits for Extreme Ability Estimates, Science	26
Table 19: Theta and Scaled-Score Limits for Extreme Ability Estimates, Social Studies	26
Table 20: Scaling Constants on the Reporting Metric	27
Table 21: Proficiency Levels, ELA.....	27
Table 22: Proficiency Levels, Mathematics	28
Table 23: Proficiency Levels, Science.....	28
Table 24: Proficiency Levels, Social Studies Grade 5.....	28
Table 25: Proficiency Levels, Social Studies, U.S. Government	28

Table 26: Overview of Quality Assurance Reports..... 39

LIST OF APPENDICES

- Appendix A: Operational Item Statistics
- Appendix B: Field Test Summaries
- Appendix C: Test Characteristic Curves
- Appendix D: Distribution of Scale Scores and Standard Deviations
- Appendix E: Distribution of Reporting Category Scores
- Appendix F: Operational Item Exposure and Blueprint Match
- Appendix G: Simulation Report

1. INTRODUCTION

The Indiana Learning Evaluation Assessment Readiness Network (ILEARN) 2021–2022 technical report documents and makes transparent all methods used in item development, test construction, psychometrics, standard setting, score reporting, student assessment result summaries, and supporting evidence for intended uses and interpretations of the test scores. The technical report is presented as six 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.
6. **Additional Studies.** This volume includes any additional studies that IDOE has requested. For the *ILEARN 2021–2022 Technical Report*, this includes Corporation-level Performance Regression Analysis.

The Indiana Department of Education (IDOE) communicates the quality of the ILEARN assessments to stakeholders and to the public by producing and providing these technical reports.

1.1 TEST BACKGROUND AND HISTORICAL CONTEXT

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+*).

1.2 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, which are aligned to the IAS in ELA and Mathematics grades 3–8, Science grades 4 and 6, and Social Studies grade 5. Indiana also develops two ILEARN end-of-course assessments (EOCs) 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 the principles of evidence-centered design; it is also developed to be accessible 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 the IAS, which can be used to inform instructional strategies and to remediate or enrich curriculum. An array of reporting metrics allows achievement to be monitored at both the student and aggregate levels and growth to be measured at both levels over time. 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 of this technical report) 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

1.3 PARTICIPANTS IN ILEARN DEVELOPMENT AND ANALYSIS

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

Indiana Department of Education

The Office of Student Assessment (OSA) oversees all aspects of the ILEARN program, including coordination with other IDOE offices, Indiana public and nonpublic 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, item writing, and committee reviews of test items and passages.

Technical Advisory Committee

SBOE and IDOE convene a panel three times a year to discuss psychometric, test development, administrative, and policy issues relevant to current and future Indiana assessments. This committee comprises several nationally recognized assessment experts.

Cambium Assessment, Inc.

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

Human Resources Research Organization

For the 2021–2022 ILEARN assessments, the Human Resources Research Organization (HumRRO) conducted independent verifications of scoring activities.

1.4 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. All online adaptive assessments make use of technology-enhanced item types. Students unable to participate in the online administrations have the option to use an online accommodated form or a paper-pencil form. Students participating in the computer-based ILEARN test can use standard online testing features in the Test Delivery System (TDS), which include a selection of font colors and sizes and the ability to zoom in and out and highlight text. In addition to the resources available to all students, ILEARN provides accommodated forms for braille and Spanish. Students with disabilities can take ILEARN with or without accommodations, or they can take the Indiana’s Alternate Measure (*I AM*) assessment. Visually impaired students can take the braille version of ILEARN ELA, Mathematics, Science, and Social Studies. English learners (ELs) can take the Spanish language version of ILEARN Mathematics, Science, and Social Studies. During test development, CAI ensured that scores obtained on the alternative modes of administrations were comparable to those received on the

standard online tests, which adhered to the same blueprints. Post-administration checks were also performed, and no concerns were found in the 2021–2022 administration. The test summary comparison between the standard online form and the alternative mode forms is provided in Volume 2 of this technical report.

1.5 STUDENT PARTICIPATION

All Indiana public and nonpublic school students in ELA and Mathematics grades 3–8; Science grades 4 and 6, and students taking the Biology End-of-Course (EOC) assessment; and Social Studies grade 5, are required to participate in the state assessments. U.S. Government is an optional EOC assessment. Table 2 shows the number of students tested and the number of students reported for the 2021–2022 ILEARN assessments. Table 3 through Table 6 present the distribution of students, in counts and percentages. The subgroup categories reported are gender, ethnicity, students with special education (SPED) status, students with Section 504 Plans, and ELs.

Table 2: Number of Students Participating in ILEARN 2021–2022

ELA			Mathematics			Science			Social Studies		
Grade	Number Tested	Number Reported	Grade	Number Tested	Number Reported	Grade	Number Tested	Number Reported	Grade	Number Tested	Number Reported
3	79,953	79,915	3	79,967	79,940	3			3		
4	81,034	81,003	4	81,028	80,990	4	80,871	80,848	4		
5	81,136	81,102	5	81,133	81,080	5			5	80,963	80,939
6	82,218	82,180	6	82,230	82,102	6	81,969	81,904	6		
7	83,391	83,346	7	83,426	83,262	7			7		
8	85,047	84,990	8	85,073	84,897	8			8		
						Biology (Spring)	81,972	81,292	U.S. Government	279	278
						Biology (Fall)	936	931			
						Biology (Winter)	1,387	1,381			

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

Grade	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 Plan	English Learner
3	N	79,953	40,826	39,127	51,822	9,976	2,502	10,989	123	73	4,468	14,058	1,797	8,103
	%	100	51.06	48.94	64.82	12.48	3.13	13.74	0.15	0.09	5.59	17.58	2.25	10.13
4	N	81,034	41,439	39,595	52,596	10,180	2,645	10,981	132	89	4,411	13,884	2,028	8,309
	%	100	51.14	48.86	64.91	12.56	3.26	13.55	0.16	0.11	5.44	17.13	2.50	10.25
5	N	81,136	41,371	39,765	52,674	10,100	2,416	11,273	135	80	4,458	13,535	2,473	6,884
	%	100	50.99	49.01	64.92	12.45	2.98	13.89	0.17	0.10	5.49	16.68	3.05	8.48
6	N	82,218	42,089	40,129	53,273	10,345	2,448	11,582	113	81	4,376	13,310	2,525	5,809

Grade	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 Plan	English Learner
	%	100	51.19	48.81	64.79	12.58	2.98	14.09	0.14	0.10	5.32	16.19	3.07	7.07
7	N	83,391	42,504	40,887	54,248	10,400	2,304	11,975	135	73	4,256	12,635	2,845	5,771
	%	100	50.97	49.03	65.05	12.47	2.76	14.36	0.16	0.09	5.10	15.15	3.41	6.92
8	N	85,047	43,278	41,769	55,616	10,493	2,254	12,182	124	81	4,297	12,683	2,976	5,481
	%	100	50.89	49.11	65.39	12.34	2.65	14.32	0.15	0.10	5.05	14.91	3.50	6.44

Table 4: Distribution of Demographic Characteristics of Tested Population, Mathematics

Grade	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 Plan	English Learner
3	N	79,967	40,836	39,131	51,831	9,973	2,500	10,999	124	72	4,468	14,076	1,876	8,120
	%	100	51.07	48.93	64.82	12.47	3.13	13.75	0.16	0.09	5.59	17.60	2.35	10.15
4	N	81,028	41,430	39,598	52,595	10,178	2,643	10,980	133	90	4,409	13,887	2,121	8,294
	%	100	51.13	48.87	64.91	12.56	3.26	13.55	0.16	0.11	5.44	17.14	2.62	10.24
5	N	81,133	41,371	39,762	52,664	10,101	2,414	11,278	135	80	4,461	13,541	2,523	6,879
	%	100	50.99	49.01	64.91	12.45	2.98	13.9	0.17	0.10	5.50	16.69	3.11	8.48
6	N	82,230	42,098	40,132	53,274	10,355	2,448	11,585	113	80	4,375	13,327	2,581	5,811
	%	100	51.20	48.80	64.79	12.59	2.98	14.09	0.14	0.10	5.32	16.21	3.14	7.07
7	N	83,426	42,530	40,896	54,255	10,414	2,305	11,983	135	73	4,261	12,658	2,952	5,783
	%	100	50.98	49.02	65.03	12.48	2.76	14.36	0.16	0.09	5.11	15.17	3.54	6.93
8	N	85,073	43,292	41,781	55,628	10,506	2,257	12,180	125	81	4,296	12,678	3,051	5,489
	%	100	50.89	49.11	65.39	12.35	2.65	14.32	0.15	0.10	5.05	14.90	3.59	6.45

Table 5: Distribution of Demographic Characteristics of Tested Population, Science

Grade	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 Plan	English Learner
4	N	80,871	41338	39,533	52,527	10,130	2,643	10,955	132	88	4,396	13,837	2,152	8,289
	%	100	51.12	48.88	64.95	12.53	3.27	13.55	0.16	0.11	5.44	17.11	2.66	10.25
6	N	81,969	41961	40,008	53,150	10,277	2,442	11,551	112	80	4,357	13,256	2,646	5,793
	%	100	51.19	48.81	64.84	12.54	2.98	14.09	0.14	0.10	5.32	16.17	3.23	7.07
Biology (Spring)	N	81,972	41918	40,054	53,684	9,832	2,223	12,104	144	88	3,897	10,999	3,097	4,364
	%	100	51.14	48.86	65.49	11.99	2.71	14.77	0.18	0.11	4.75	13.42	3.78	5.32
Biology (Fall)	N	936	480	456	641	81	18	139	1	0	56	170	19	67
	%	100	51.28	48.72	68.48	8.65	1.92	14.85	0.11	0	5.98	18.16	2.03	7.16
Biology (Winter)	N	1,387	738	649	993	190	15	133	4	0	52	219	38	24
	%	100	53.21	46.79	71.59	13.70	1.08	9.59	0.29	0	3.75	15.79	2.74	1.73

Table 6: Distribution of Demographic Characteristics of Tested Population, Social Studies

Grade	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 Plan	English Learner
4	N	80,963	41,258	39,705	52,590	10,053	2,409	11,249	135	80	4,447	13,501	2,565	6,863
	%	100	50.96	49.04	64.96	12.42	2.98	13.89	0.17	0.10	5.49	16.68	3.17	8.48
U.S. Government	N	279	139	140	182	59	4	26	1	0	7	57	16	9
	%	100	49.82	50.18	65.23	21.15	1.43	9.32	0.36	0	2.51	20.43	5.73	3.23

2. SUMMARY OF OPERATIONAL PROCEDURES

2.1 ADMINISTRATION PROCEDURES

Table 7 shows the testing window schedule for the 2021–2022 ILEARN assessments.

Table 7: 2021–2022 ILEARN Testing Windows

Assessment	Grade/Subject	Mode	Testing Window
ILEARN	ELA 3–8 Mathematics 3–8 Science 4 & 6 Social Studies 5	Online Paper	April 18–May 13, 2022
	Biology	Online Paper	November 29–December 16, 2021 (Fall window) November 29–December 9, 2021 (Fall window)
		Online Paper	February 7–24, 2022 (Winter window) February 7–17, 2022 (Winter window)
		Online Paper	April 18–May 20, 2022
	U.S. Government	Online Paper	April 18–May 20, 2022

The key personnel involved with ILEARN administration include the Corporation Test Coordinators (CTCs), the Co-Op role, Non-Public School Test Coordinators (NPSTCs), School Test Coordinators (STCs), the Principal (PR), and Test Administrators (TAs) who administered the test. Test Administrator’s Manuals (TAMs) were provided so that personnel involved with statewide assessment administrations could maintain both standardized administration conditions and test security.

CAI’s Secure Browser was required to access the online ILEARN assessments. The online browser provided a secure environment for student testing by disabling the hot keys, copy/paste, and screen-capture capabilities, and preventing access to the desktop (e.g., Internet, email, 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). ILEARN performance tasks did *not* have a pause rule.

2.2 UNIVERSAL FEATURES, DESIGNATED FEATURES, AND ACCOMMODATIONS

Accessibility supports discussed in 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 Learning Plan (ILP).

Scores achieved by students using designated features and accommodations are included for federal accountability purposes. All educators are provided training on accessibility supports and accommodations. Trainings include the range of designated features and accommodations available and the appropriate uses of the various supports.

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 [TTS]) are provided digitally through instructional or assessment technology, while non-embedded accommodations (e.g., scribe) are external to the TDS 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 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 TAs and STCs 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 (ASL), 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 of this technical report.

3. MAINTENANCE OF THE ITEM BANK

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 Readiness [ICCR]), Indiana-owned items from external sources, and Indiana custom-developed items. Volume 2 of this technical report 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 field testing. Several factors play into the development of new items; the item development team conducts a gap analysis for distributions of items across multiple dimensions, such as item counts, item types, item difficulty, and numbers in each strand or benchmark.

3.2 REVIEW OF OPERATIONAL ITEMS

During and after each operational test administration, a series of quality assurance reports is generated and used to evaluate whether operational items are performing as intended. These reports serve as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. Flagged items are reviewed by psychometricians and content experts. Details can be found in Chapter 7, Quality Assurance Procedures.

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 2021–2022 ILEARN assessments contained newly developed field-test items. The embedded field test (EFT) slots are randomly positioned for the online adaptive English/Language Arts (ELA), Mathematics, and Science assessments and are in fixed positions for the online fixed-form Social Studies assessments. To render high-quality responses to the EFT items, students were unaware of which were operational items and which were EFT items. For all assessments, field-test items were randomly distributed from the pool of available field-test items.

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

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

This design mimics the “spiraling-by-student within a classroom” model typically used with paper-pencil forms and ensures broad representation of the items across abilities and demographic groups. To describe the distribution of forms, consider that J total forms are available for administration and a total of N students are participating in the field test. The

probability that any one of the J forms can be assigned to one student is $1/J$. Thus, the distribution of forms would follow a uniform distribution with sample sizes per form equal to N/J . Therefore, field-test item exposure rates depend on the number of field-test slots and the number of field-test items.

3.4 OPERATIONAL FORM CONSTRUCTION AND ADAPTIVE SIMULATIONS

Prior to the operational testing window for adaptive tests, CAI psychometricians employed 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 G, Simulation Report, provides more details about the simulation approach and results.

Appendix F, Operational Item Exposure and Blueprint Match, contains the operational item exposure rates, as well as the operational blueprint-match results for the Fall and Winter Biology EOC assessments. Item exposure rates were calculated over all completed test cases. The location of the item on the form (e.g., first, last) did 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 and grade-repeaters 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 Appendix F, 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 of this technical report contains more detailed information about operational test form development.

4. CLASSICAL ANALYSES OVERVIEW

4.1 CLASSICAL ITEM ANALYSES

Classical item statistics are based on the classical test theory framework and have been widely applied to examine whether test items function as intended. Table 8 shows the types of classical test statistics along with the criteria used to evaluate ILEARN items.

Table 8: Evaluative Criteria in Classical Item Analysis

Analysis Type	Evaluative Criteria
Item Discrimination	Biserial/polyserial correlation statistic is less than 0.25 for multiple-choice or constructed-response items.*
Distractor Analysis	Biserial correlation statistic is greater than 0.00 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 (multiple-choice items)	Proportion correct value is less than 0.25 or greater than 0.95 for multiple-choice items.
Item Difficulty (non-multiple-choice items)	Proportion correct value is less than 0.25 or greater than 0.95 for constructed-response 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.
Differential Item Functioning (DIF)	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.

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

A minimum sample of 200 responses (Zwick, 2012) per item is applied for classical item analyses. Similarly, a minimum sample of 200 responses (Zwick, 2012) per item in each subgroup is applied for differential item functioning (DIF) analyses. However, CAI implements field-test designs to ensure at least 4,000 responses per item for 2PL/GPC item response theory (IRT; van der Linden & Hambleton, 1997) models.

It is important to note that classical item statistics are sample dependent, which means item difficulty and item discrimination indices are dependent on the sample of students selected to answer the items. If the same items are given to a different sample, they may vary substantially depending on the nature of the sample. This property is particularly important for ILEARN assessments because ELA, Mathematics, and Science assessments are administered via adaptive algorithms, while Social Studies assessments are fixed forms. For fixed-form tests, forms are randomly assigned to students, ensuring that each item is seen by a representative sample of participating students. By contrast, in an adaptive setting, items are selected to maximize test information near the student’s ability estimate, which causes the resulting data to include

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

4.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between test takers 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, Operational Item Statistics. Most of the operational items had higher biserial correlations than the evaluative 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 is 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 multiple-choice 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. For the 2021–2022 administration, most of the operational items had negative distractor correlations. CAI content experts reviewed items with positive distractor correlations and did not find any issues.

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 multiple-choice 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 answers. 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, Item Analyses Results. The p -values for operational items can be found in Appendix A, Operational Item Statistics. Most of the operational items had p -values within the expected range. Flagged items were verified by CAI content experts and psychometricians, who reported that all items behaved as expected.

4.1.4 Mean Total Score

For multi-point, constructed-response items, mean total score was calculated as the average ability estimate of the students in the score point category. Multi-point 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 3 points on a constructed-response item scored lower, on average, on the total test than students who received only 2 points on the item, the item will be flagged for review. Flagged items were verified by CAI content experts and psychometricians, who reported that all of them behaved as expected.

4.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

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

DIF analyses were conducted for all items to detect potential item bias across major and special population groups (e.g., gender, 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
- Student with Special Education (SPED)/Not SPED
- Title 1/Not Title 1 (proxy for Free and Reduced-Price Lunch)
- 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, 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

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

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

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

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

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

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

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

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

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. 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, female), or negative DIF (i.e., -A, -B, or -C), signifying that the item favored the reference group (e.g., White, 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. Please refer to Table 9 to review DIF classification rules.

Table 9: DIF Classification Rules

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

4.3 ITEM ANALYSES RESULTS

This section includes a summary of results from the classical item analysis for the 2021–2022 ILEARN operational forms. For the reasons stated in Section 4.1, Classical Item Analyses, regarding the sample-dependent property of classical item statistics for adaptively administered items, this section provides results for only Social Studies, which are fixed-form assessments. The summaries here are aggregates; item-specific details are found in Appendix A, Operational Item Statistics.

Table 10 provides summaries of the p -values by percentile and summaries of the range by grade for operational Social Studies items. Note that the “Total OP Items” column shows the number of operational items that were used in the computation of the percentiles. Indiana students’ performance indicates the desired variability across the scale for both tests. The variability informs us that the constructed operational forms had a good discrimination factor for Indiana students.

Table 10: Operational Item p -Value Five-Point Summary and Range, Social Studies

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	40	0.24	0.26	0.35	0.46	0.52	0.70	0.82
U.S. Govt.	54	0.07	0.08	0.17	0.31	0.45	0.60	0.72

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; however, 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 that were consistent with the item banks and 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:

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

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

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:

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

where $b'_i = (b_{i,1}, \dots, b_{i,m_i})$ 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 person j , k indexes step of the item i , and $b_{i,k}$ is the k th step parameter for item i with $m_i + 1$ total categories.

Rasch Model

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

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

Rasch Testlet Model

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

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

where u_g is the nuisance dimension parameter for cluster g .

5.1.1 ELA and Mathematics

English/Language Arts (ELA) and Mathematics adopted the Smarter Balanced IRT framework. For 1-point items, the two-parameter logistic model was used, and for multi-point items, the GPCM was used.

5.1.2 Science

Science item banks were newly established for the 2021–2022 test administration. 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 PTs can be found in Volume 2 of this technical report.

5.1.3 Social Studies

Social Studies item banks were newly established for the 2021–2022 test administration. Grade 5 adopted a process consistent with ELA and Mathematics, and only used the 2PL and GPC models. U.S. Government returned low sample sizes, so the Rasch model was used to ensure reliable item parameter estimates.

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 these calibrations can be found in the *ILEARN 2018–2019 Technical Report*. As of 2021–2022, all assessments are pre-equated.

For field-test item calibrations, all operational items were anchored to their bank values and field-test item parameters were estimated. Table 11 presents the number of students used in field-test calibrations.

Table 11: N Students Used in Field-Test Calibrations

ELA		Mathematics		Science		Social Studies	
Grade	Calibration N Count	Grade	Calibration N Count	Grade	Calibration N Count	Grade	Calibration N Count
3	76396	3	79628				
4	77194	4	80703	4	80524		
5	77491	5	80702			5	80702
6	78194	6	81699	6	81517		
7	79499	7	82920				
8	81100	8	84532				
				Biology	78914	U.S. Government	265

5.2.1 IRT Summaries

The IRT statistical properties of the final operational test forms or online item pools used for ILEARN are summarized in Table 12 through Table 15. It is important to note that these summaries are based on items that appear on general education tests; items that appear only on accommodated forms are not included.

Table 12: Operational Item Parameter 5-Point Summary and Range, ELA

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.04	0.30	0.44	0.59	0.76	1.01	1.43
	b	-1.29	1.22	1.42	1.59	1.75	2.00	2.43
4	a	0.05	0.25	0.43	0.58	0.73	0.95	1.42
	b	-2.46	-1.76	-1.06	-0.15	0.74	1.82	6.23
5	a	0.06	0.26	0.41	0.56	0.70	0.94	1.25
	b	-2.28	-1.47	-0.62	0.08	1.11	2.61	12.57
6	a	0.17	0.27	0.40	0.58	0.73	0.94	1.30

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
	b	-1.45	-1.06	-0.21	0.65	1.43	2.80	4.27
7	a	0.06	0.25	0.41	0.55	0.68	0.89	1.21
	b	-2.02	-0.80	0.05	1.00	1.84	3.36	7.12
8	a	0.06	0.26	0.41	0.52	0.68	0.88	1.12
	b	-3.01	-0.63	0.16	1.13	1.99	3.29	5.60

Table 13: Operational Item Parameter 5-Point Summary and Range, Mathematics

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.22	0.39	0.64	0.84	1.10	1.32	1.52
	b	-4.34	-2.76	-1.89	-1.27	-0.45	0.70	2.87
4	a	0.18	0.37	0.63	0.83	1.06	1.35	1.64
	b	-3.26	-2.08	-0.99	-0.24	0.42	1.36	4.11
5	a	0.18	0.32	0.57	0.73	0.93	1.16	1.47
	b	-2.78	-1.57	-0.36	0.33	0.96	2.05	6.20
6	a	0.13	0.29	0.53	0.71	0.87	1.08	1.38
	b	-3.93	-1.6	-0.25	0.79	1.56	2.71	9.16
7	a	0.05	0.25	0.48	0.75	0.93	1.13	1.43
	b	-1.89	-0.67	0.80	1.50	2.33	3.53	7.80
8	a	0.12	0.23	0.39	0.53	0.72	1.00	1.20
	b	-1.87	-0.94	0.39	1.74	3.00	4.59	9.02

Table 14: Operational Item Parameter 5-Point Summary and Range, Science

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
4	b	-2.71	-1.71	-0.64	0	0.36	1.22	1.83
6	b	-2.21	-1.62	-0.66	0	0.33	1.46	3.70
Biology (Spring)	b	-2.86	-1.49	-0.45	0	0.57	1.9	4.15
Biology (Fall & Winter)	b	-2.86	-1.51	-0.54	0.03	0.57	1.52	3.92

Table 15: Operational Item Parameter 5-Point Summary and Range, Social Studies

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	a	0.19	0.23	0.41	0.53	0.67	1.05	1.24
	b	-2.21	-1.09	-0.39	-0.01	0.73	2.00	4.04
U.S. Govt.	b	-2.01	-1.60	-0.60	-0.01	0.76	1.66	1.76

5.2.2 2021 ILEARN Test Characteristic Curves

Another way to view the technical properties of ILEARN test forms is via the test characteristic curves (TCCs). Given that ELA, Mathematics, and Science are adaptive tests, which means each student receives a uniquely constructed form, these TCC plots (displayed in Appendix C, Test Characteristic Curves) are constructed only for Social Studies, which are fixed forms.

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 in this section.

6.1.1 Likelihood Function

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

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

where

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

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

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

$$q_i = 1 - p_i$$

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 $\log(L(\theta))$ 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:

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

where

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

and where θ_t denotes the estimated θ at iteration t . N_{CR} is the number of items that are scored using the Generalized Partial Credit Model (GPCM), and N_{2PL} is the number of items scored using the two-parameter logistic (2PL) model.

6.1.3 Standard Errors of Estimates

When the MLE or MMLE is available and within the lowest observable theta score (LOT) and the highest observable theta score (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

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

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

6.1.4 Extreme Case Handling

When students answer all items correctly or 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 the 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 the smallest a-parameter among the administered operational items for a student. Adding 0.5 to an incorrect item score with the 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 the smallest a-parameter penalizes less than selecting any other item, 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 LOSS = lowest observable scale score and HOSS = highest observable scale score. Estimated theta values lower than the LOT or higher than the HOT will be truncated to the LOT and HOT values and will be assigned the LOSS and HOSS associated with the LOT and HOT.

Table 16 through Table 19 give the LOT, LOSS, HOT, and HOSS for the ILEARN assessments.

Table 16: Theta and Scaled-Score Limits for Extreme Ability Estimates, ELA

Grade	Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable 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 17: Theta and Scaled-Score Limits for Extreme Ability Estimates, Mathematics

Grade	Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable 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

Grade	Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable Scale Score (HOSS)
7	-5.0667	5.6000	6120	6920
8	-5.0667	6.0000	6120	6950

Table 18: Theta and Scaled-Score Limits for Extreme Ability Estimates, Science

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

*The theta and scaled-score limits were identical for the Fall 2021, Winter 2022, and Spring 2022 Biology administrations.

Table 19: Theta and Scaled-Score Limits for Extreme Ability Estimates, Social Studies

Grade	Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable 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 and HOT scores, the LOT and HOT values replace the theta in the formula in Section 6.1.3, Standard Errors of Estimates. 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 2021–2022, scale scores were reported for each student who took the ILEARN assessments. The scale scores were based on the operational items presented to the student and did not include any field-test items. The scale score is a linear transformation of the IRT ability estimate, θ :

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

where a is the slope and b is the intercept. Table 20 lists the scaling constants a and b for the ILEARN assessments.

ELA and Mathematics were reported on a vertical scale. The IRT vertical scale was established by Smarter Balanced and formed by linking across grades using common items in adjacent grades. Grade 6 was used as the baseline, and each grade was successively linked onto the scale. More details about the vertical scaling methods can be found in Chapter 9 of the *ILEARN 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 D, Distribution of Scale Scores and Standard Deviations, and the summary of scale scores for each reporting category is provided in Appendix E, Distribution of Reporting Category Scores.

Table 20: 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 2021–2022 school year was assigned an overall performance category in accordance with his or her overall scale score. Table 21 through Table 25 provide the scale score range for performance standards for ILEARN. The lower bound of Level 3, At Proficiency, marks the minimum cut score for proficiency.

Table 21: Proficiency Levels, 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 22: Proficiency Levels, 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
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 23: Proficiency Levels, 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 24: Proficiency Levels, 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 25: Proficiency Levels, 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 and MMLE Scoring

Reporting category theta scores were calculated using either MLE or MMLE, depending on the assessment and based on the items contained in a particular reporting category.

The same rules for scoring all correct and all incorrect cases were applied to reporting category scores.

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 SE of the reporting category was used to determine the relative strengths and weaknesses.

The specific rules for mastery are as follows:

- Below (Code = 1): if $\text{round}(SS_{rc} + 1.5 * SE(SS_{rc}),0) < SS_p$
- 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
- Above (Code = 3): if $\text{round}(SS_{rc} - 1.5 * SE(SS_{rc}),0) \geq SS_p$

SS_{rc} is the student’s scale score on a reporting category; SS_p is the proficiency scale cut score (Level 3 cut score); 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 2021, standard-level information would have been reported for the ELA, Mathematics, and Science assessments.

First, $p_{ij} = p(z_{ij} = 1)$ was defined, representing the probability that student j responded correctly to item i (z_{ij} represents the j^{th} student’s score on the i^{th} item). For items with one score point, the 2PL IRT model was used to calculate the expected score on item i for student j with $\theta_{Level\ 3\ cut}$ as:

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

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

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

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,

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

and

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

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 (positive $\underline{\delta}_{Tg}$) or less effective (negative $\underline{\delta}_{Tg}$) in teaching a given standard.

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

6.5 LEXILE[®] AND QUANTILE[®] SCORES¹

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

As a quality assurance check for aberrant test administrations in the context of the COVID-19 pandemic, CAI conducted a study to confirm the integrity of the test administration prior to the final release of Spring 2022 test scores. In this study, a weighted linear regression model was run for each assessment to identify expected levels of achievement for corporations in Spring 2022, given their observed achievement levels in Spring 2021. Corporations with large deviations from expected levels of achievement were identified. IDOE investigated flagged schools prior to final score release.

After the release of test scores, CAI conducted further investigation to determine possible explanations for deviation from predicted performance through analysis of residuals. This was done by predicting residuals using corporation characteristics such as corporation size, participation rate, and changes in demographic variables between the two administrations. Details of this regression study can be found in Volume 6.

¹ Lexile[®] and Quantile[®] measures are the intellectual property of MetaMetrics, Inc.

7. QUALITY ASSURANCE PROCEDURES

Quality assurance procedures are enforced throughout all stages of ILEARN test development, administration, and scoring and reporting. This chapter describes quality assurance procedures associated with the following:

- Test configuration
- Test production
- Data preparation
- Equating and scaling
- Scoring and reporting

Because quality assurance procedures pervade all aspects of test development, we note that discussion of quality assurance procedures is not limited to this chapter, but is also included in chapters describing all phases of test development and implementation.

7.1 QUALITY ASSURANCE IN TEST CONFIGURATION

Each test administration is generated by the adaptive algorithm to exactly match the detailed test blueprint while targeting test information to student ability. The blueprint describes the content to be covered, the Depth of Knowledge (DOK) with which it will be covered, the type of items that will measure the constructs, and every other content-relevant aspect of the test.

The adaptive test configuration process is managed through CAI's Test Simulator. Immediately upon completion of a test simulation, the Test Simulator generates a blueprint match report to ensure that all elements of the test blueprint have been satisfied. In addition, the Test Simulator produces a statistical summary of form characteristics to ensure consistency of test characteristics across simulated test forms.

Prior to its implementation in the operational test administration, the CAI scoring engine and the accuracy of data files are checked using a simulated student response data file. The simulated data are used to check whether the student responses entered in the TDS were captured accurately and the scoring specifications were applied correctly. The simulated data file is scored independently by two programmers following the scoring rules.

In addition to checking the scoring accuracy, the test configuration file is checked thoroughly. For the operational test administration, a test configuration file is the key file that contains all specifications for the item selection algorithm and, eventually, for the scoring algorithm, such as the test blueprint specifications, slopes, and intercepts for theta-to-scale score transformation, and the item information (e.g., cut scores, answer keys, item attributes, item parameters, passage information). The accuracy of the

information in the configuration file is checked and confirmed numerous times independently by multiple staff members prior to the testing window.

7.2 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION

7.2.1 Production of Content

While the online workflow requires some additional steps, it removes a substantial amount of work from the time-critical path, reducing the likelihood of errors. Like a test book, an online system can deliver a sequence of items; however, the online system makes the layout of that sequence algorithmic. The appearance of the item screen can be known with certainty before the final test is configured.

The production of computer-based tests (CBTs) includes the following four key steps:

1. Final content is previewed and approved in a process called web approval. Web approval packages the item exactly as it will be displayed to the student.
2. Complete test configuration is approved, which gathers the content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
3. Tests are initially deployed to a test site where they undergo platform review, a process during which we ensure that each item displays properly on a large number of platforms representative of those used in the state for testing purposes.
4. The final system is deployed to a staging environment accessible to IDOE for user acceptance testing (UAT) and final review.

7.2.2 Web Approval of Content During Development

The Item Tracking System (ITSx) integrates directly with the TDS display module and displays each item exactly as it will appear to the student. This process is called *web preview* and is tied to specific item review levels. Upon approval at those levels, the system locks content as it will be displayed to the student, transforming the item representation to the exact representation that will be rendered to the student. No change to the display content can occur without a subsequent web preview. This process freezes the display code that will present the item to the student.

Web approval functions as an item-by-item blueline review. It is the final rendering of the item as the student will see it. Layout changes can be made after this process in the following two ways:

1. Content can be revised and re-approved for web display.
2. Online style sheets can be changed to revise the layout of all items on the test.

Both processes are subject to strict change-control protocols to ensure that accidental changes are not introduced. In the next section, we discuss automated quality control processes during content publication that raise warnings if item content has changed after the most recent web-approved content was generated. The web approval process offers the benefit of allowing final layout review much earlier in the process, reducing the work that must be performed during the very busy period just before tests go live.

7.2.3 Platform Review

Platform review is a process in which each item is checked to ensure that it is displayed appropriately on each tested platform. A platform is a combination of a hardware device and an operating system. In recent years, the number of platforms has proliferated, and platform review now takes place on approximately 15 platforms that are significantly different from one another.

A team conducts the platform review, where the team leader projects the item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to see that it renders as expected.

7.2.4 User Acceptance Testing and Final Review

Each release of every one of our systems goes through a complete testing cycle, including regression testing. With each release, and every time we publish a test, the system goes through UAT. During UAT, we provide our client with login information to an identical (though smaller scale) testing environment to which the system has been deployed. We provide recommended testing scenarios and constant support during the UAT period. Identified issues will be resolved before the opening of the test administration or noted for future review and resolution if a current resolution is not feasible within the timeline. IDOE signs off on the administration go-live date at the conclusion of UAT activities.

Deployments to the production environment all follow specific, approved deployment plans. Teams working together execute the deployment plan. Each step in the deployment plan is executed by one team member and verified by a second. Each deployment undergoes shakeout testing following the deployment.

This careful adherence to deployment procedures ensures that the operational system is identical to the system tested on the testing and staging servers. Upon completion of each deployment project, management approves the deployment log.

Some changes may need to be made to the production system during the year. Outside of routine maintenance, no change is made to the production system without approval of the Production Control Board (PCB). The PCB includes the director of CAI's Assessment Program or the chief operating officer, the director of our Computer and Statistical Sciences Center (CSSC), and the project director. Any request for a change to the production system requires the signature of the system's lead engineer. The PCB reviews risks, test plans, and test results. If any proposed change will affect client functionality or pose a risk to operation of a client system, the PCB ensures that the client is informed and in agreement with the decision.

The PCB approves a maintenance plan that includes every scheduled change to the system.

Deviations from the maintenance plan must be approved by the PCB, including server or driver patches that differ from those approved in the maintenance plan.

Every bug fix, enhancement, data correction, or new feature must be presented with the results of a quality assurance plan and approved by the PCB.

An emergency procedure is in place that allows rapid response in the event of a time-critical change that is needed to prevent the system being compromised. Under those circumstances, any member of the PCB can authorize the senior engineer to make a change, with the PCB reviewing the change retroactively.

Typically, deployments happen during a maintenance window, and deployments are scheduled at a time that can accommodate full regression testing on the production machines. Any changes to the database or procedures that in any way might affect performance are typically subject to a load test at this time.

Cutover and Parallel Processing

CAI maintains multiple environments to ensure smooth cutover and parallel processing. With a centralized hosting site in Washington, DC, multiple development environments and a test environment can be maintained. At Rackspace, we maintain a staging environment and the production environment.

The production environment runs independently of the other environments and is changed only with the approval of the PCB. Enhancements are initially developed and tested on the development and test environments in Washington, DC, before being deployed to the staging environment at Rackspace.

The staging environment is a scaled-down version of the production environment. It is in this environment that UAT takes place. Only when UAT is complete and the PCB signs off is the production environment updated. In this way, the system continues to function uninterrupted as testing takes place in parallel until a clean cutover takes place.

Prior to deployment, the testing system and content are deployed to a staging server where they are subject to UAT. UAT of the TDS serves both a software evaluation and content approval role. The UAT period provides IDOE with an opportunity to interact with the exact test with which the students will interact.

7.2.5 Functionality and Configuration

The items, both in themselves and as configured onto the tests, form one type of online product. The delivery of that test can be thought of as an independent service. Here, we document quality assurance procedures for delivering the online assessments.

One area of quality unique to online delivery is the quality of the test delivery system. The following three activities provide for the predictable, reliable, quality performance of our system:

1. Testing on the system itself to ensure function, performance, and capacity
2. Capacity planning
3. Continuous monitoring

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

In addition, latency data are captured for each assessed student (i.e., data about how long it takes to load, view, or respond to an item). All this information is logged as well, enabling us to automatically identify schools or districts experiencing unusual slowdowns, often before they even notice.

7.3 Quality Assurance in Data Preparation

When a student responds to test questions online, his or her response to each item is immediately captured and stored in the Database of Record (DOR) at CAI, a repository for all data relevant to a student's testing experience. Our quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at CAI.

When data are prepared for psychometric analyses, they undergo two phases: a data preparation phase and a psychometric phase. In the former phase, data are extracted from the DOR and provided to two independent SAS programmers. These two programmers are provided with the client-assigned business rules, and they independently prepare data files suitable for subsequent psychometric analysis. The data files prepared by the different programmers are formally compared for congruency. Any discrepancies identified are resolved through code review meetings with the programmer lead and the lead psychometrician.

When the two data files match exactly, they are then passed over to two independent psychometricians, who each perform classical and IRT analyses. Any discrepancies are identified and resolved. When all results match from the independent analysts, the final results are uploaded to CAI's ITS.

CAI's Test Delivery System (TDS) has a real-time quality-monitoring component built in. As students test, data flow through our Quality Monitor (QM) System. The QM System conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the

test record contains no data from items that have been invalidated. The QM System scores the test, recalculates performance-level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

The QM System also aggregates data to detect problems that become apparent only in the aggregate. For example, the QM System monitors item statistics and flags items that perform differently operationally than their item parameters predict. This functions as a sort of automated key or rubric check, flagging items where data suggest a potential problem. This automated process is similar to the checks that are done for data review, but they are done on operational data and they are conducted in real time so that our psychometricians can catch and correct any problems before they have an opportunity to do any harm.

Data pass directly from the QM System to the DOR, which serves as the repository for all test information, and from which all test information for reporting is pulled. The data extract generator is the tool that is used to pull data from the DOR for delivery to IDOE and their QA contractor. CAI psychometricians ensure that data in the extract files match the DOR prior to delivery to the IDOE.

7.4 QUALITY ASSURANCE IN ITEM ANALYSIS AND EQUATING

Prior to operational work, CAI produces simulated datasets for testing software and analysis procedures. The quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are independently replicated by two CAI psychometricians. Two psychometricians complete a dry run calibration, and linking activities, and compare results. The practice runs serve the following two functions:

1. To verify the accuracy of program code and procedures
2. To evaluate the communication and work flow among participants (if necessary, the team will reconcile differences and correct production or verification programs)

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

7.5 QUALITY ASSURANCE IN SCORING AND REPORTING

CAI implements a series of quality control steps to ensure error-free production of score reports in an online format. The quality of the information produced in the TDS is tested thoroughly before, during, and after the testing window.

7.5.1 Quality Assurance in Test Scoring

CAI verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each simulated student is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they also provide a check of the full range of item responses and test scores in fixed-form tests. Additionally, these simulations ensure that students at all performance levels are exposed to the full range of test item content as dictated by the ILEARN test blueprints. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the Reporting System, we merge item response data with the demographic information taken either from previous-year assessment data or, if current year enrollment data are available by the time simulated data files are created, we can verify online reporting using current-year testing information. By populating the simulated data files with real school information, it is possible to verify that special school types and special districts are being handled properly in the Reporting System.

Specifications for generating simulated data files are included in the analysis output student data file specifications document submitted to IDOE each year. Review of all simulated data is scheduled to be completed prior to the opening of the test administration, so that the integrity of item administration, data capture, and item and test scoring and reporting can be verified before the system goes live.

To monitor the performance of the assessment system during the test administration window, a series of quality assurance reports can be generated at any time during the online assessment window. For example, item analysis reports allow psychometricians to ensure that items are performing as intended and serve as an empirical key check through the operational testing window. In the context of adaptive test administrations, other reports, such as blueprint match and item exposure reports, allow psychometricians to verify that test administrations conform to specifications.

The quality assurance reports are generated on a regular schedule. Item analysis and blueprint match reports are evaluated frequently at the opening of the testing window to ensure that test administrations conform to blueprint and items are performing as anticipated.

Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Table 26 presents an overview of the quality assurance reports.

Table 26: Overview of Quality Assurance Reports

Quality Assurance Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology items)
Item Exposure Rates	To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages)	Early detection of any oversight in the blueprint specification
Blueprint Match Rates	To monitor unexpected low blueprint match rates	Early detection of unexpected blueprint match issues

Item Analysis Report

The item analysis report is used to monitor the performance of test items throughout the testing window and serves as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine test items for changes in performance, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as IRT-based item-fit statistics. The report is configurable and can be produced so that only items with statistics falling outside a specified range are flagged for reporting or to generate reports based on all items in the pool.

Item p-Value. For multiple-choice items, the proportion of students selecting each response option is computed; for constructed-response, performance, and technology items, the proportion of student responses classified at each score point is computed. For multiple-choice items, if the keyed response is not the modal response, the item is also flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

Item Discrimination. Biserial correlations for the keyed response for selected-response items and polyserial correlations for polytomous constructed-response, performance, and technology items are computed. CAI psychometric staff evaluates all items with biserial correlations below a target level, even if the obtained values are consistent with past item performance.

Item Fit. In addition to the item difficulty and item discrimination indices, an item-fit index is produced for each item. For each student, a residual between observed and expected score given the student’s ability is computed for each item. The residuals for each are averaged across all students, and the average residual is used to flag an item.

7.5.2 Quality Assurance in Reporting

Scores for the ILEARN online assessments are assigned by automated systems in real time. For machine-scored portions of the assessments, the machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field testing. The review process “locks down” the item and rubric when the item is approved for web display (web approval). During operational testing, actual item responses are compared to expected item responses (given the IRT parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports that are available to psychometricians.

The handscoring processes include rigorous training, validity and reliability monitoring, and back-reading to ensure accurate scoring. Handscored items are married up with the machine-scored items by our Test Integration System (TIS). The integration is based on identifiers that are never separated from their data and are further checked by the QM System, where the integrated record is passed for scoring. Once the integrated scores are sent to the QM System, the records are rescored in the test-scoring system that applies the ILEARN scoring rules and assigns scores from the calibrated items, including calculating performance-level indicators, subscale scores, and other features, which then pass automatically to the Reporting System and the DOR. The scoring system is tested extensively prior to deployment, including checks of scored tests and large-scale simulations to ensure that point estimates and standard errors are correct.

After passing through the series of validation checks in the QM System, data are passed to the DOR, which serves as the centralized location for all student scores and responses, ensuring that there is only one place where the official record is stored. Only after scores have passed the QM System checks and are uploaded to the DOR are they passed to the Reporting System, which is responsible for presenting individual-level results and calculating and presenting aggregate results. No score is reported in the Reporting System until it passes all the QM System’s validation checks.

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*)**

2021–2022

**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.

Major contributors to this technical report include the following staff from Cambium Assessment, Inc. (CAI): Stephan Ahadi, Shuqin Tao, Elizabeth Xiaoxin Wei, Maryam Pezeshki, Kevin Clayton, Christina Sneed, Jessica Singh, 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..... 2

 1.3 Organization of this Volume 3

2. ILEARN ITEM BANK SUMMARY..... 4

 2.1 Item Banks 4

 2.2 Item Acceptance Meetings..... 5

 2.3 Item Bank Composition 6

3. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS 9

 3.1 Overview 9

 3.2 Passage and Item Specifications 10

 3.2.1 *Passage Specifications*..... 11

 3.2.2 *Item Specifications*..... 12

 3.3 Selection and Training of Item Writers 15

 3.4 Internal Review 15

 3.4.1 *Preliminary Review* 15

 3.4.2 *Content Review 1*..... 16

 3.4.3 *Edit Review 1* 17

 3.4.4 *Senior Content Review* 17

 3.5 Review by State Personnel and Stakeholder Committees 18

 3.5.1 *State (Client) Review* 18

 3.5.2 *Content/Fairness Committee Review*..... 18

 3.5.3 *Markup for Translation and Accessibility Features*..... 19

 3.5.4 *Indiana Educator Review of Licensed Item Banks* 19

 3.6 Field Testing..... 19

 3.7 Post-Field-Test Review 19

 3.7.1 *Key Verification* 20

 3.7.2 *Rubric Validation*..... 20

 3.7.3 *Rangefinding*..... 21

 3.7.4 *Data Review*..... 21

4. ILEARN BLUEPRINTS AND STATE ASSESSMENT TEST CONSTRUCTION... 22

 4.1 Test Blueprints 22

 4.1.1 *Blueprint Construction Meeting* 22

 4.1.2 *ILEARN Test Specifications* 22

 4.1.3 *ELA Blueprints* 26

 4.1.4 *Mathematics Blueprints*..... 26

 4.1.5 *Science Blueprints* 27

 4.1.6 *Social Studies Blueprints* 27

 4.2 Test Form Construction..... 27

 4.3 Test Form Assembly 28

4.4	Roles and Responsibilities	29
4.4.1	Role of the CAI Content Team	29
4.4.2	Role of the CAI Technical Team	30
4.4.3	Role of IDOE	30
4.5	Target Guidelines	30
4.6	Accommodated Form Construction	30
4.6.1	Test Characteristic Curve	32
4.6.2	Test Characteristic Curve Difference	33
4.6.3	Conditional Standard Error of Measurement Curve	33
5.	PERFORMANCE LEVEL DESCRIPTORS	35
6.	REFERENCES	36

LIST OF TABLES

Table 1:	Sources of Items for the ILEARN 2021–2022 Assessments	1
Table 2:	Operational Item Counts by Source	4
Table 3:	Operational Performance Task Counts by Source	5
Table 4:	ILEARN Item Types and Descriptions	6
Table 5:	ELA Operational Items by Item Type and Grade	7
Table 6:	Mathematics Operational Items by Item Type and Grade	7
Table 7:	Science Operational Items by Item Type and Grade	8
Table 8:	Social Studies Operational Items by Item Type and Grade	8
Table 9:	How Each Step of Development Supports the Validity of Claims	9
Table 10:	ILEARN Item Specifications	10
Table 11:	Sample ELA Item Specification for Grade 4	13
Table 12:	Number of Hand-Scored Items by Form	23
Table 13:	Number of Embedded Field-Test Items by Form	23
Table 14:	Blueprint Percentage of Test Items Assessing Each Reporting Category in ELA	24
Table 15:	Blueprint Percentage of Test Items Assessing Each Reporting Category in Mathematics	24
Table 16:	Blueprint Percentage of Test Items Assessing Each Reporting Category in Science	25
Table 17:	Blueprint Percentage of Test Items Assessing Each Reporting Category in Social Studies	25
Table 18:	Statistical Test Summary Comparison for Grade 5 Social Studies Online and Paper Forms	31

LIST OF FIGURES

Figure 1:	Features of the REVISE Software	21
-----------	---------------------------------------	----

Figure 2: TCC Comparisons of Grade 5 Social Studies Online and Paper Forms 32
Figure 3: TCC Differences of Grade 4 Science Online and Accommodated Forms..... 33
Figure 4: CSEM Comparisons of Grade 4 Science Online and Accommodated Forms 34

LIST OF APPENDICES

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

1. INTRODUCTION

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

The IAS were approved by the Indiana State Board of Education in April 2014 for English/Language Arts (ELA) and Mathematics, and in March 2015 for Social Studies. The IAS for Science were originally revised in 2010 but were updated in 2016 to reflect changes in Science content. The IAS were most recently updated in 2020. The IAS are intended to implement more rigorous standards that promote college-and-career readiness, with the goal of challenging and motivating Indiana’s students to acquire stronger critical thinking, problem solving, and communications skills.

ILEARN assessments were created using a variety of item types from several sources. Table 1 denotes the sources of the items used in 2021-2022, including licensed item banks (Smarter Balanced Assessment Consortium [Smarter] and Independent College and Career Ready [ICCR], and custom Indiana development. Each item source is outlined in more detail in Section 2.

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

Table 1: Sources of Items for the ILEARN 2021–2022 Assessments

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

1.1 CLAIM STRUCTURE

The ILEARN assessments are designed to measure college-and-career readiness as defined by the IAS and support the claim that students in grades 3 through 8 are demonstrating progress toward college-and-career readiness in ELA, Mathematics, Science, and Social Studies. Expected student performance across all ILEARN-assessed contents is defined through Indiana’s ILEARN Policy Performance Level Descriptors (PLDs). The ILEARN Policy PLDs are high-level statements that reflect the varying degrees to which students may demonstrate proficiency on each grade-level ILEARN assessment. A panel of Indiana educators devised these PLDs by considering many factors such as Indiana’s diverse student populations, the Indiana Academic Standards, and national reference points. The Policy PLDs were used to develop more specific content area Range PLDs to inform item development, instructional practices, and standard setting. The Range PLDs are extensive documents that provide content specific claims across each Indiana Academic Standard to represent the range of expectations for student performance within each proficiency level. They are available for each content and grade assessed through ILEARN on Indiana’s Assessment Website for ILEARN <https://www.in.gov/doe/students/assessment/ilearn/>.

1.2 2UNDERLYING 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 Indiana custom development, began with detailed item specifications. The specifications, discussed in a later section, identified specific evidence statements for each standard, described the interaction types that could be used, provided guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and presented 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, translation, or assistive technologies. This goal is supported by the delivery of the items on CAI’s test delivery platform, which has received the Web Content Accessibility Guidelines (WCAG) 2.0 AA certification, an internationally recognized accessibility standard. The platform 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 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 a data review as well as *rubric validation*, a process that refines rule-based rubrics upon review of student responses.

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, while 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 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 2 provides the count of items, by source, used on the 2021–2022 *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 custom Indiana-developed items were also used to ensure complete coverage of the IAS and support a more robust item pool for the computer-adaptive assessment.

For Science grades 4 and 6, the item banks consisted mostly of Indiana-developed items. In Biology, the Indiana owned item pool was used primarily and was augmented by ICCR.

The Social Studies grade 5 item pool and the U.S. Government item pool contain solely custom Indiana items.

Table 2: Operational Item Counts by Source

Subject and Grade	# of Smarter Items	# of ICCR Items	# of Indiana Owned Items
ELA 3	366	25	34
ELA 4	288	44	43
ELA 5	265	17	42
ELA 6	203	37	19
ELA 7	259	52	40
ELA 8	325	13	27
Mathematics 3	380	47	55
Mathematics 4	423	17	43
Mathematics 5	339	50	50
Mathematics 6	501	19	29
Mathematics 7	477	32	37
Mathematics 8	329	23	31

Subject and Grade	# of Smarter Items	# of ICCR Items	# of Indiana Owned Items
Science 4		19	108
Science 6		13	131
Biology		17	225
Social Studies 5			68
U.S. Government			54

Additionally, all assessments other than Social Studies included one performance task per grade. Table 3 lists the counts of performance tasks in the 2021–2022 item pool.

Table 3: 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	5	
ELA 6	4	
ELA 7	3	
ELA 8	3	
Mathematics 3	2	
Mathematics 4	3	
Mathematics 5	1	
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. Results of that meeting can be found in Volume 2 of the 2019-2020 Technical Reports.

In November 2021, a subsequent item acceptance review was convened where alignment was considered for ELA performance tasks.

2.3 ITEM BANK COMPOSITION

Table 4 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 5 through Table 8 list the number of items by type for each grade and subject.

Table 4: 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.
Simulation (SIM)	Student selects inputs to “run” trials. Data is presented in a table after trials are run.

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

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

Table 5: ELA Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
TE	22	22	27	19	29	28
ETC	1	1	1		1	
EBSR	65	39	39	43	28	23
HT	38	45	38	23	50	27
MI	23	13	12	14	4	2
MC	200	202	143	114	165	100
MS	74	50	58	42	71	52
ER	2	3	5	4	3	3

Table 6: Mathematics Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
TE	6	6	5	6	3	10
EQ	254	269	242	270	308	110
GI	52	24	16	29	19	32
MI	32	70	76	53	38	69
MC	129	84	79	80	81	96
MS	11	11	15	95	93	61
HT		1		1		
TI	2	14	6	17	2	5

Table 7: Science Operational Items by Item Type and Grade

Item Type	4	6	Biology**
TE	10	4	4
ETC	9	9	5
EBSR		1	1
EQ	2	2	2
GI		2	33
HT	1	3	
MI	1	6	4
MC	86	93	177
MS	12	18	8
PT*	1	1	2
SIM			1
TI	2	3	3
ETC & MC**		1	
ETC & MS**	2	1	1
EQ & 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 8: Social Studies Operational Items by Item Type and Grade

Item Type	5	U.S. Government
TE	4	
EBSR	1	19
MC	59	10
MI	2	1
MS	2	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 the items could support the claims on which they were based. Table 9 describes how each step contributed to these goals. Each step in the process is discussed in more detail below.

Table 9: How Each Step of Development Supports the Validity of Claims

	Supports alignment to the standards	Reduces construct-irrelevant variance through universal design	Expands access through linguistic and other supports
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.	

	Supports alignment to the standards	Reduces construct-irrelevant variance through universal design	Expands access through linguistic and other supports
	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

The Indiana Department of Education leveraged quality content from third-party item banks for use on *ILEARN* assessments. These item banks were accompanied by item specifications which were utilized when alignment was confirmed by Indiana educators. The available specifications are described in Table 10 below.

Table 10: ILEARN Item Specifications

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

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 through a collaboration between content experts in one of CAI’s partner states and CAI content experts. The specifications align to nationally recognized standards. Over time, the specifications have been expanded to reflect continuous improvement and the availability of new interaction types.

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

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 specialists’ 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, 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 11 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 11: Sample ELA Item Specification for Grade 4

Content Standard	4.RV.2.2: Identify relationships among words, including more complex homographs, homonyms, synonyms, antonyms, and multiple meanings.
Content Limits	Items should ask students not to define the type of word that is being used but rather to demonstrate its meaning between the words. Items may refer only to synonym and antonym in the stimuli. All words should be provided with sufficient context for support.
Construct-Relevant Vocabulary	antonyms, meaning, opposite, phrase, relationship, replace, similar/same as, synonyms,
Recommended Response Mechanisms (Item Types)	Drag and Drop Evidence-Based Selected Response Hot Text Multiple Choice Multi-Select
DOK	2
Evidence Statements	
Students replace a given word with synonyms, antonyms, homographs, homonyms, and multiple-meaning words.	
Students use context to determine or support meaning.	
Students identify a word, sentence, or phrase that uses a given word in the same way.	
(NOTE: Level of difficulty will depend on subtlety/amount of text and/or complexity of interpretation required.)	
Sample Item	
Why is “[word X]” a better word to use from paragraph 4 than “[word Y]”?	
<ul style="list-style-type: none"> A. [Word X] suggests [something more formal] B. [Word X] suggests [something more precise] C. [Word X] suggests [something more aligned to the tone] D. [Word X] suggests [something more audience appropriate] 	
Literacy Considerations	Word List: Content can select construct-irrelevant words for glossing, which gives students access to the definition and an audio clip of those words. Considerations will include the question/task, standard, and construct-relevant words necessary for the item.
Visual and Auditory Considerations (NOTE: These considerations generally refer to the passage/media source rather than the item.)	American Sign Language: Allows a student to see a video of an ASL interpreter. This option will be included only if the media contains audio. Audio Transcriptions: Written transcripts of audio for students of varying auditory and visual abilities can be provided as needed. The same transcripts will be used for ASL videos.

	<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, 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 the preliminary review is conducted in a group setting, led by a senior test developer. During

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

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

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

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

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

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

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 and that all information in the item is correct. For mathematics items, editors perform all calculations to ensure accuracy.

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

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

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 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, in order 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. Clients can simply select 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 combines 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 members can include representatives from the special education, low

vision, hearing impaired, and other student populations, including English Learners. Further, diverse members of this committee represent students of various ethnic and economic backgrounds to ensure that all items are free of bias and sensitivity concerns.

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 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 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. Results of that meeting can be found in Volume 2 of the 2019-2020 Technical Reports.

3.6 FIELD TESTING

Custom Indiana development and licensed content from Smarter Balanced was field tested as embedded field-test items in Spring 2022.

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 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 for Item Number 17185. It is divided into three main sections:

- Sample Details:** Shows sample information and a table of rules.

Rule Short Name	Rule Description	Number of Responses
HighGridScore	Sample of responses that scored unusually high on this grid item (given overall score)	15
LowGridScore	Sample of responses that scored unusually low on this grid item (given overall score)	13
NormalResponses	Sample of responses with grid scores that are neither low nor high	17
- Responses:** A table listing individual responses with columns for Mark as Reviewed, Original Score, Process Score, Current Score, Proposed Score, Response ID, and Sample Type. A callout indicates: "Responses in the sample are listed here."

Mark as Reviewed	Original Score	Process Score	Current Score	Proposed Score	Response ID	Sample Type
0	0	0	0	0	18259	LowGridScore
1	1	1	1	1	52098	NormalResponse
1	1	1	1	1	52099	HighGridScore
1	1	1	1	1	52706	HighGridScore
1	1	1	1	1	54126	HighGridScore
0	0	0	0	0	55217	NormalResponse
- Test Item and Response:** Shows the actual test item for Item 17185: "When traveling at a constant speed, the distance that a plane travels, d , is proportional to the time, t . The table shows the relationship between the time and distance the plane travels."

Time (Hours)	Distance (Miles)
2	1,140
3	1,710
4	2,280

Create an equation that represents the relationship between the time and distance the plane travels.

The student response is: $570d$
 $1t$

Below the response, a callout indicates: "Users can see the actual student response here." To the right, a callout indicates: "The committee records its comments and consensus score here." The interface shows a "Response: 18259 Score: 0" and a "Proposed Score" field set to 0.

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 were 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 students graduate from high school ready to succeed in their college and career endeavors. To ensure that the *ILEARN* assessments provide a valid assessment of college-and-career-readiness, the test blueprints were constructed to ensure that the assessments represent the range of content defined in the IAS and result in accurate classification of student achievement as college-and-career-ready.

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

ILEARN is designed to support the claims described at the outset of this 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; and
- Number of hand-scored items.

Table 12 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 item(s). 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 12: 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. Indiana educators determined or confirmed the DOK expectations as item specifications were created and accepted. Table 13 denotes the number of EFT items per form.

Table 13: 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	All	5

Note that ELA EFT items were divided between segment 1 (Reporting Categories 1 and 2) and segment 2 (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 2019 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 14 through Table 17 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 14: 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 15: 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%
5	Algebraic Thinking	Computation	Geometry and Measurement, Data Analysis, and Statistics	Number Sense	Process Standards
	20—26%	22—28%	18—23%	22—28%	8—13%
	Algebra and Functions	Computation	Geometry and Measurement,	Number Sense	Process Standards

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

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

Grade	Reporting Categories				
	Questioning and Modeling	Investigating	Analyzing, Interpreting, and Computational Thinking	Explaining Solutions, Reasoning, and Communicating	
4	25—29%	25—29%	21—25%	21—25%	
6	21—25%	21—25%	25—29%	25—29%	
	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
Biology	18—22%	18—22%	18—22%	18—22%	18—22%

Table 17: Blueprint Percentage of Test Items Assessing Each Reporting Category in Social Studies

Grade	Reporting Categories		
	Civics and Government	Geography and Economics	History
5	38—43%	28—33%	28—33%
	Functions of Government	Historical Foundations of American Government	Institutions and Processes of Government
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 and up to four Media Literacy items;
- Stand-alone writing and/or research items; and
- One PT which includes two “precursor” items leading up to a text-based writing task.

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

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 2021, CAI psychometricians and content experts worked with IDOE to build forms for the Spring 2022 administration. *ILEARN* assessment test form construction utilized test construction guidelines, explicit blueprints, and collaborative participation from all parties. The Spring 2022 *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. Additionally, items on the ELA Grade 6, ELA Grade 7, and U.S. Government forms were replaced to remove content deemed to be sensitive for COVID or Social Justice reasons.

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 2022 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 18. 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 18: 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

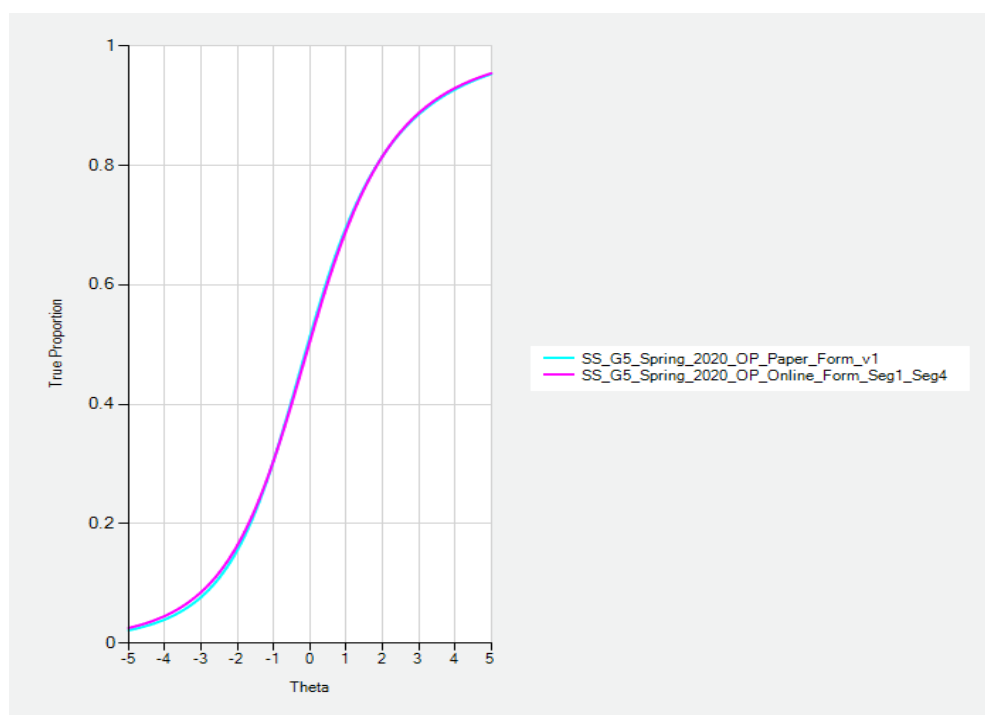
Type	Statistics	Online Form	Paper Form
	P-Value StDev	0.14	0.13
	P-Value Minimum	0.09	0.28
	P-Value Maximum	0.75	0.86
	Lowest Bi/Poly-Serial	0.22	0.25

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

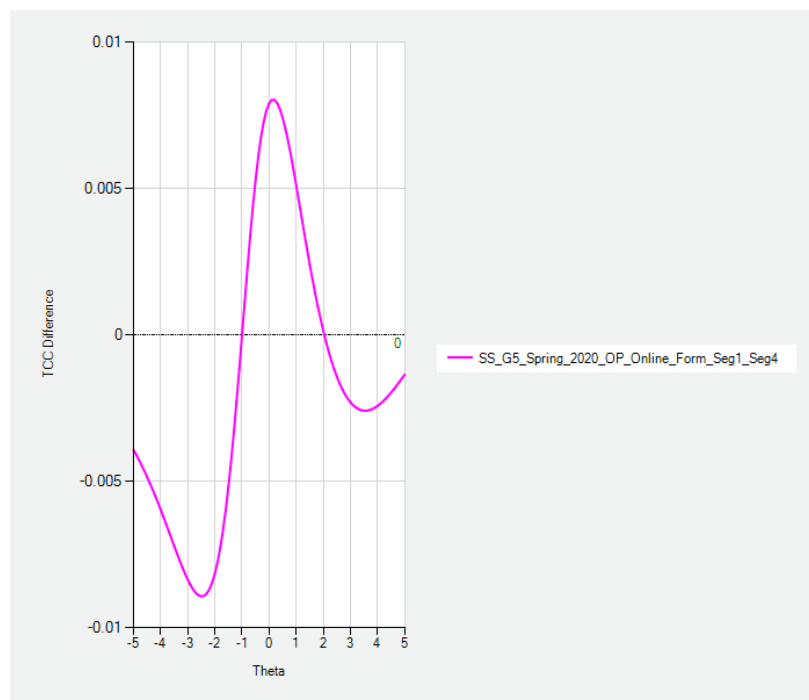
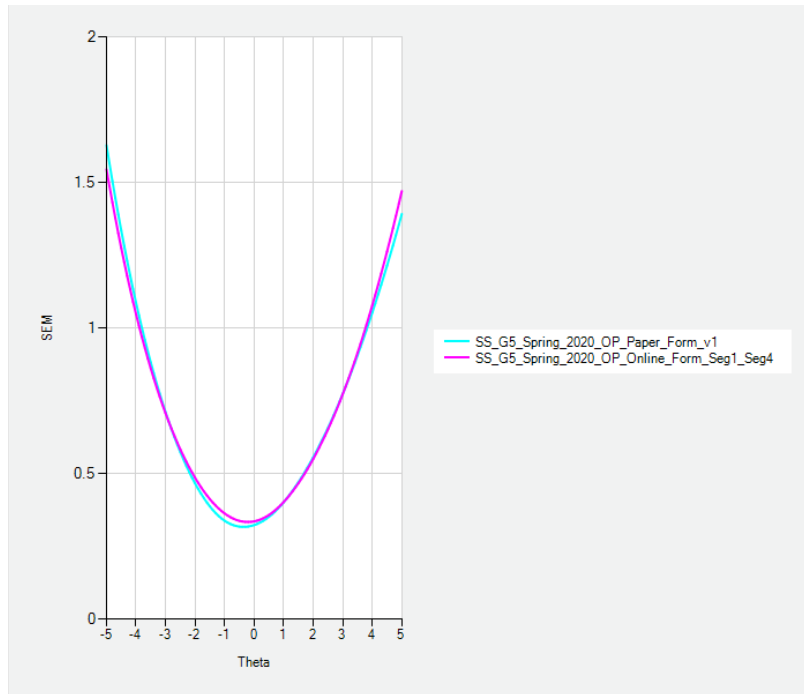


Figure 4: CSEM Comparisons of Grade 4 Science Online and Accommodated Forms



5. PERFORMANCE LEVEL DESCRIPTORS

The Indiana Department of Education (IDOE) held an educator workshop 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 use 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 meeting 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)**

2021-2022

**Volume 3
Test Administration**

TABLE OF CONTENTS

1. INTRODUCTION.....	1
1.1. Testing Procedures and Test Windows.....	2
1.2. Eligible Students	3
1.3. Testing Accommodations and Designated Features.....	4
2. ADMINISTRATOR TRAINING.....	8
2.1. Online Administration	8
2.2. Roles and Responsibilities in the Online Testing Systems.....	9
2.3. Test Administration Resources	10
3. DEPARTMENT RESOURCES AND SUPPORT.....	13
3.1. ILEARN Released Items Repository	13
3.2. ILEARN Practice Tests	14
4. TEST SECURITY PROCEDURES	15
4.1. Security of Test Materials.....	15
4.2. Identifying Test Irregularities or Potential Test Security Concerns	17
4.3. Tracking and Resolving Test Irregularities	17
4.4. CAI's System Security.....	19
REFERENCES	20

LIST OF TABLES

<i>Table 1. Designated Features and Accommodations Available in 2020-2021 for ILEARN</i>	5
<i>Table 2. User Guides and Manuals</i>	11
<i>Table 3. Examples of Test Irregularities and Test Security Violations</i>	18

LIST OF APPENDICES

- Appendix A: *Released Items Repository Quick Guide*
- Appendix B: *ILEARN Test Administrator's Manual Grades 3 through 8*
- Appendix C: *Test Information Distribution Engine User Guide*
- Appendix D: *Online Test Delivery System (TDS) User Guide*
- Appendix E: *Listing of Read-Aloud Scripts for ILEARN*
- Appendix F: *Indiana Assessments Policy Manual*
- Appendix G: *Technology Setup for Online Testing Quick Guide*
- Appendix H: *Test Administrator Certification Course Storyboard*
- Appendix I: *Accessibility and Accommodations Implementation and Setup Module*
- Appendix J: *Computer-Adaptive Tests Webinar Module*
- Appendix K: *Why It Is Important to Assess Webinar Module*
- Appendix L: *Request an Item Rescore Webinar Module*
- Appendix M: *Test Administration Overview Webinar Module*
- Appendix N: *Test Information Distribution Engine (TIDE) Webinar Module*
- Appendix O: *Test Delivery System (TDS) Webinar Module*
- Appendix P: *Online Reporting System (ORS) Webinar Module*
- Appendix Q: *Technology Requirements for Online Testing Webinar Module*
- Appendix R: *How the Scoring Process Works Webinar Module*
- Appendix S: *Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux*
- Appendix T: *Online Practice Test User Guide*
- Appendix U: *Assistive Technology Manual 2020-2021*
- Appendix V: *Online Reporting System User Guide*
- Appendix W: *Accessibility and Accommodations Guidance Manual*
- Appendix X: *ILEARN 3-8 Test Administrator's Manual (TAM) with Spanish Scripted Instructions*
- Appendix Y: *ILEARN Biology End-of-Course (ECA) Test Administrator's Manual (TAM)*
- Appendix Z: *ILEARN Biology End-of-Course (ECA) Test Administrator's Manual (TAM) with Spanish Scripted Instructions*
- Appendix AA: *ILEARN U.S. Government End-of-Course (ECA) Test Administrator's Manual*
- Appendix AB: *ILEARN U.S. Government End-of-Course (ECA) Test Administrator's Manual with Spanish Scripted Instructions*
- Appendix AC: *ILEARN Test Coordinator's Manual (TCM)*

1. INTRODUCTION

The State of Indiana implemented an online assessment for operational use beginning with the 2018–2019 school year. This assessment program, referred to as the ILEARN assessments, replaced Indiana Statewide Testing for Educational Progress-Plus (ISTEP+). ILEARN comprises English/Language Arts (ELA), Mathematics, Science, and Social Studies assessments for students ranging from third grade through the end of high school. ELA and Mathematics assessments are administered in grades 3–8. Science is administered in grades 4 and 6, and Biology is administered as an end-of-course assessment, typically in high school. Social Studies is administered in grade 5, and U.S. Government is administered in high school. The U.S. Government assessment is optional. During the 2021-2022 ILEARN administrations, ELA, Mathematics, Science, and Biology assessments were offered as computer-adaptive tests (CATs), while the Social Studies and U.S. Government tests were offered as fixed-form online assessments. The ELA, Mathematics, Science, and Biology assessments consist of a CAT (or an accommodated online fixed form in some cases) segment and a performance task segment. Students needed to complete the CAT segment of the test to receive a final overall scale score and both the CAT segment and the performance task segment 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. TESTING PROCEDURES AND TEST WINDOWS

Administering the 2021-2022 ILEARN assessments required coordination, detailed specifications, and proper training. In addition, several individuals in each corporation and school were involved in the administration process, from those setting up secure testing environments to those administering the tests. IDOE worked with CAI to develop and provide the training and documentation necessary for the administration of ILEARN under standardized conditions within all testing environments, both online and on paper-and-pencil tests.

All students were required to take a practice test at their school prior to taking the 2021-2022 ILEARN assessments. These practice tests contained sample test items similar to the test items that students would encounter on the ILEARN assessments to help students become familiar with the item types that would be presented on the online or paper-and-pencil assessments. Indiana students also had the opportunity to interact with released, non-secure items on public-facing [Released Items Repository](#) (RIR) assessments available on the [ILEARN portal](#). The ILEARN RIR was deployed for ILEARN Biology ECA in October 2021 and allowed students online access to released items two months prior to the opening of the fall test administration. A completely updated ILEARN RIR was deployed for all tests in late January, 2022. A quick guide for the RIR is available to the public (Appendix A).

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

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

The ILEARN assessments were untimed, but timing estimates were included in the ILEARN Test Administrator's Manuals (TAM) (Appendix B) to ensure that schools had resources available to create local testing schedules. The fall Biology test was available from November 29 through December 16, 2021, and the winter Biology test was available February 7 through February 24, 2022. The spring ILEARN test window for grades 3–8 was held from April 18 through May 13, 2022. The spring Biology and U.S. Government tests were available from April 18 through May 20, 2022.

1.2. ELIGIBLE STUDENTS

All students enrolled in tested grade levels and courses participated in the Spring 2022 ILEARN administration with or without accommodations, with the exception of students with significant cognitive disabilities (approximately 1% of the student population) who participated in the alternate assessment (I AM). I AM has a distinct administration that is described in a separate technical report. Students took the fall, winter, or spring Biology ECA upon completion of the respective high school course to coincide with one of the three test windows. Section 1111(b)(2)(A) of the Elementary and Secondary Education Act of 1965 (as amended by the Every Student Succeeds Act [ESSA]) requires the implementation of high-quality student academic assessments in Mathematics, Reading or Language Arts, and Science. Section 1111(b)(2)(B)(i)(II) requires that these assessments be administered to all elementary and secondary school students. In addition, Section 1111(c)(4)(E) requires participation rates in statewide assessments of at least 95% for all students and each subgroup of students and factors this percentage into the state's federal accountability system. Students' failure to take Indiana's assessments may result in a lower federal accountability rating. Students must take the tests appropriate for the grade level and subject in which they are receiving instruction. All testing is administered on the basis of the student's enrolled grade. Off-grade testing is not available for ILEARN.

- **Public and Nonpublic School Students.** Students enrolled in accredited Indiana public (including including charter schools) and nonpublic schools (including Choice schools) were required to participate in course-level appropriate ILEARN assessment(s).
- **English Learners (ELs).** All ELs enrolled in tested grade levels were expected to participate in all ILEARN assessments, including English/Language Arts, regardless of how long these students had been enrolled in a U.S. school. Mathematics, Science, and Social Studies assessments are available in stacked Spanish in the online Test Delivery System (TDS). Stacked Spanish is represented on the screen with the stimulus/passage and item appearing in both Spanish and English for students whose test setting language is Spanish. Translated glossaries are also available as a support for the top 5 student home languages in Indiana: Arabic, Burmese, Mandarin, Vietnamese, and Spanish.

Students with Disabilities. Indiana established procedures to ensure the inclusion in statewide testing of all public elementary and secondary school students with disabilities. Federal and state laws require that all students participate in the state testing system. In Indiana, a student with an Individualized Education Program (IEP) will participate in ILEARN with the appropriate testing supports and accommodations prescribed by the IEP. If required by the student's IEP, the student will participate in Indiana's Alternate Measure (I AM). Per the Individuals with Disabilities Education Improvement Act (IDEA) and Title 5 Article 7-Special Education, published December 2014 by the Indiana State Board of Education, decisions regarding the appropriate assessment for a student with disabilities are made annually by the student's IEP team. These decisions are based on the student's curriculum, present levels of academic achievement, functional

performance, and learning characteristics. Decisions cannot be based on program setting, category of disability, percentage of time in a particular placement or classroom, or any considerations regarding a school's Adequate Yearly Progress (AYP) designation.

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

1.3. TESTING ACCOMMODATIONS AND DESIGNATED FEATURES

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

Universal features are available in TDS to all students taking ILEARN assessments. These features include. During the tests, students can zoom in and zoom out to increase or decrease the size of text and images, highlight items and passages (or sections of items and passages), cross out response options by using the strikethrough function, use a notepad to make notes, and mark a question for review using the flag function.

Designated features, such as the ability to select an alternate background and font color, mouse pointer size and color, and font size before testing, as well as glossaries that provide definitions for approved words in a second language, are available for use by any student for whom the need has been indicated by an educator, or team of educators with parent/guardian and student.

Accommodations are supports provided to students with disabilities enrolled in public schools with current IEPs or Section 504 Plans, as well as to students identified as ELs. All Indiana state assessments have appropriate accommodations available to make test content accessible to students with disabilities and ELs, including ELs with disabilities. The accommodations available for eligible students participating in the ILEARN assessments are described in the ILEARN TAMs (Appendix B), which were accessible to schools before and during testing in the [Resources](#) section of the [ILEARN Portal](#). A comprehensive list of accommodations available for eligible students with IEPs, Section 504 Plans, or Individual Learning Plans participating in online assessments is given in the *Test Information Distribution Engine (TIDE) User Guide* (Appendix C).

Table 1 provides a list of the designated features and accommodations and that were offered in the 2021-2022 administration. The *Online Test Delivery System (TDS) User Guide* can be found on the ILEARN portal (Appendix D of this report volume) and provides instructions on how to access and use these features.

Table 1. Designated Features and Accommodations Available in 2021-2022 for ILEARN

Designated Features	Accommodations
Embedded	
Color contrast (Onscreen) Glossaries (Language) Spanish Masking Mouse pointer Print size Translation Stacked Spanish	American Sign Language (ASL) Audio Transcriptions Calculator Closed Captioning Permissive Mode Print on Demand Streamline Text-to-Speech Except Reading Comprehension Text-to-speech Including Reading Comprehension Refreshable Braille
Non-Embedded	
Assistive technology to Magnify/Enlarge Access to Sound Amplification Program Special Furniture or Equipment for Viewing Test Special Lighting Conditions Time of Day for Testing Altered Color Acetate Film for Paper Assessments	Braille Transcript for Audio Items Paper Booklet Large Print Booklet Read-Aloud to Self Read-Aloud Script for Paper Booklet* Scribe Speech-to-Text Tested Individual Interpreter for Sign Language Braille Booklet Multiplication Table Hundreds Chart Additional Breaks Bilingual Word-to-Word Dictionary Spanish Booklet Calculator Multiplication Table

*See **Appendix E** for a complete list of the Read-Aloud Scripts available to students during the 2021-2022 ILEARN assessments.

The TA and the School Test Coordinator (STC) were responsible for ensuring that arrangements for appropriate accommodations were made before the test administration dates. Requests for any non-standard accommodations were recorded under a Special Requests section in the Test Information Distribution Engine (TIDE) and required IDOE approval. IDOE provided a separate, supplemental accessibility manual – the *Indiana Assessments Policy Manual* (Appendix F) – for individuals involved in administering tests to students who required accommodations.

Students who required online accommodations (e.g., text-to-speech) were provided the opportunity to participate in practice activities for the statewide assessments with appropriate allowable accommodations. Test administrators identified test settings and accommodations in TIDE before students could start an online test session. Some

settings and accommodations could not be changed once a student started a test. IDOE approved updates to incorrectly assigned accommodations before any updates were applied to subsequent student testing. IDOE also determined which testing attempts to invalidate prior to score reporting.

Starting in the 2020-2021 school year, TTS was expanded and split into two separate accommodations for ELA. The 2021-2022 tests continued this pattern of accommodations wherein one accommodation read aloud only content that was not designed to assess reading comprehension. The second accommodation read aloud all test content, including those items and passages designed to assess reading comprehension. As a result, students who participated in ILEARN ELA in grades 3 through 8 could be assigned to either of two TTS modalities:

- TTS **except** for items and passages measuring reading comprehension; or
- TTS **including** items and passages measuring reading comprehension.

Case conference committees determined which of these accommodation modalities was appropriate for their students requiring TTS. Guidance to schools and case conference committees on assigning TTS for all items including reading comprehension was provided in the *2021-2022 Accessibility and Accommodations Guidance* manual (Appendix W), as well as in periodic communications with the field.

If an EL or a student with an IEP or Section 504 Plan used any accommodations during the test administration, this information was recorded by the Test Administrator (TA) in the required administration information and was captured by CAI in the database of record (DoR). CAI included this data in the state output student data score files (SDFs) provided to IDOE at the end of each test administration. Guidelines recommended for making accommodation decisions included the following:

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

Students with disabilities not enrolled in public schools or receiving services through public school programs who required accommodations to participate in a test administration were permitted access to accommodations if the following information was provided:

- Evidence that the student had been found eligible as a student with a disability as defined by Individuals with Disabilities Education Improvement Act (IDEA).

Documentation that the requested accommodations had been regularly used for instruction. The following accommodations were available for eligible students with IEPs or Section 504 Plans participating in paper-based assessments:

- Contracted UEB braille and Nemeth Code for Mathematics.
- Uncontracted braille and Nemeth Code for Mathematics.

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

2. ADMINISTRATOR TRAINING

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

TAs were required to complete CAI's online TA Certification Course before administering any tests. There were also several training modules developed by CAI in collaboration with IDOE to facilitate test administration. These modules included topics on CAI systems, test administration, and accessibility and accommodations. These modules are included in this volume's appendices.

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

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

The User Support sections of both the *Online Test Delivery System (TDS) User Guide* (Appendix D) and the *Test Information Distribution Engine (TIDE) User Guide* (Appendix C 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.

2.1. ONLINE ADMINISTRATION

The *Online Test Delivery System (TDS) User Guide* (Appendix D) 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 G) 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.

2.2. 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 D) for their specific responsibilities before, during, and after testing.

CTCs

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

CITCs

CITCs were responsible for ensuring that testing devices were properly configured to support testing and for coordinating participation in the 2021-2022 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 C) for more information.

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

TAs

In order to be certified as a TA, educators need to complete an online Test Administrator Certification Course (Appendix H). 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.

2.3. TEST ADMINISTRATION RESOURCES

The list of webinars and training resources available to corporations and schools for the 2021-2022 ILEARN administration is provided below. All training materials were available online at the [ILEARN Portal](#). PDFs of these resources have also been included as appendices in this technical report. Test administration resources comprising various tutorials and documents (e.g., user guides, manuals, quick guides) also were available through the [ILEARN Portal](#).

- **Test Administrator Certification Course:** All educators who administered the ILEARN assessment were required to complete the online TA Certification Course (Appendix H).
- **Accessibility and Accommodations Implementation and Setup Module:** This online module provided information on accessibility and accommodations available for use on the ILEARN assessments (Appendix I).
- **Computer-Adaptive Tests Webinar Module:** This online module described computer-adaptive-testing and the student test experience (Appendix J).
- **Why It Is Important to Assess Webinar Module:** This online module illustrated the importance of statewide testing (Appendix K).
- **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 assessments (Appendix L).
- **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 test window (Appendix M).
- **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 (Appendix N).
- **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 (Appendix O).
- **Online Reporting System (ORS) Webinar Module:** This module provided a general overview of ORS where student scores, including individual scores and aggregate scores, are displayed after students complete the ILEARN assessments (Appendix P).

- **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 (Appendix Q).
- **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 (Appendix R).

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.

Table 2. User Guides and Manuals

Resource	Description
<i>Online Test Delivery System (TDS) User Guide (Appendix D)</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 (Appendix G)</i>	This document explains in four steps how to set up technology in Indiana corporations and schools.
<i>2020-2021 Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux (Appendix S)</i>	This manual provides information about hardware, software, and network configurations for running various testing applications provided by CAI.
<i>Online Practice Test User Guide (Appendix T)</i>	This user guide provides an overview of the ILEARN Practice Test.
<i>Test Information Distribution Engine (TIDE) User Guide (Appendix C)</i>	This user guide describes the tasks performed in the Test Information Distribution Engine (TIDE) for ILEARN assessments.
<i>Assistive Technology Manual (Appendix U)</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 (Appendix V)</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>Accessibility and Accommodations Guidance (Appendix W)</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.
<i>ILEARN 3-8 Test Administrator's Manual (TAM) (Appendix B)</i>	The ILEARN 3 through 8 Test Administrator's Manual (TAM) provides an overview of the specific roles and responsibilities required before, during, and after testing.

<i>ILEARN 3-8 Test Administrator's Manual (TAM) with Spanish Scripted Instructions (Appendix X)</i>	The ILEARN 3-8 Test Administrator's Manual (TAM) with Spanish Scripted Instructions provides an overview of the specific roles and responsibilities required before, during, and after testing. The scripted instructions read by Test Administrators to students are in Spanish.
<i>ILEARN Biology End-of-Course (ECA) Test Administrator's Manual (TAM) (Appendix Y)</i>	The ILEARN Biology ECA Test Administrator's Manual (TAM) provides an overview of the specific roles and responsibilities required before, during, and after testing.
<i>ILEARN Biology End-of-Course (ECA) Test Administrator's Manual (TAM) with Spanish Scripted Instructions (Appendix Z)</i>	The ILEARN Biology ECA Test Administrator's Manual (TAM) with Spanish Scripted Instructions provides an overview of the specific roles and responsibilities required before, during, and after testing. The scripted instructions read by Test Administrators to students are in Spanish.
<i>ILEARN U.S. Government End-of-Course (ECA) Test Administrator's Manual (TAM) (Appendix AA)</i>	The ILEARN U.S. Government ECA Test Administrator's Manual (TAM) provides an overview of the specific roles and responsibilities required before, during, and after testing.
<i>ILEARN U.S. Government End-of-Course (ECA) Test Administrator's Manual (TAM) with Spanish Scripted Instructions (Appendix AB)</i>	ILEARN U.S. Government Test Administrator's Manual (TAM) with Spanish Scripted Instructions provides an overview of the specific roles and responsibilities required before, during, and after testing. The scripted instructions read by Test Administrators to students are in Spanish.
<i>ILEARN Test Coordinators Manual (TCM) (Appendix AC)</i>	The ILEARN Test Coordinator's Manual (TCM) provides an overview of test administration activities intended for Test Coordinators.

3. 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 IDOE’s Office of Assessment to all officially designated CTCs in IDOE’s database. The newsletter was titled “ILEARN Assessment Update” and included new announcements relevant to the ILEARN assessment, reminders of upcoming milestones, and a “Planning Ahead” section with important dates specific to the ILEARN program. The 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 with 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. Such memos were distributed to superintendents, principals, and school leaders.
- General information about the assessments was posted on the Office of Assessment website (<https://www.in.gov/doi/>), including approved test windows for all state-administered assessments. The Accessibility and Accommodations Guidance in the ILEARN Policy and Guidance section of their website was often referenced to address questions pertaining to accommodations and overall accessibility.
- Pretest workshops, presentations at annual state conferences (choice, non-public, HASTI, ICTM, etc.), Questions and Answers sessions (pre-administration and during for CTCs), and results webinars were provided by the IDOE Office of Student Assessment staff. Sessions included topics regarding the ILEARN administration, test security, and results for instructional decision making.

3.1. 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 would encounter when taking the ILEARN assessment. The ILEARN RIR was deployed on January 24, 2022 and remained available throughout the test 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.

3.2. ILEARN PRACTICE TESTS

The purpose of the practice tests was to familiarize students with 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, but 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 January 24, 2022 and remained available throughout the spring test window. Schools accessed the ILEARN practice tests via 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 aligned with the student experience for the operational test, including accommodations and accessibility features. TAs used scripts to administer the practice tests. Scripts provided instruction for all aspects of the practice test and described the presentation of items and tools in TDS. Online practice test scripts were available in English, Spanish, and for students with a hard of hearing accommodation who required an approved sign language interpreter. 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 on the pages immediately preceding the first operational test segment inside 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 their students understood how to respond to the different question types represented on the ILEARN assessment. Separate paper-and-pencil scripts were developed to support the administration of English, Spanish, and braille forms.

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

The test security procedures for ILEARN included the following:

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

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 AC) and the ILEARN TAMs (Appendix B), IDOE and CAI ensured that all test security procedures were available to everyone involved in test administration. Each manual included protocols for reporting any deviations in test administration.

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

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

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. Maintaining security of all test content was of high priority before, during, and after committee meetings. Printed copies of items and performance task content were not provided to educator participants. Any secure materials created or distributed during the meetings were collected and destroyed following the meetings.

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

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

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

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

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

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

Students were not able to print from the CAI Secure Browser unless testing with the Print-on-Demand accommodation. Print-on-Demand allows students to participate in computer-adaptive assessments while using paper to read and respond to items when necessary. This accommodation requires a one-on-one testing environment in a secure location and additional test security management. Printed content is securely destroyed at the local level once testing is complete, in accordance with established protocols.

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. Review Appendix A of the *Online Test Delivery System (TDS) User Guide* 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 2021-2022 ILEARN test windows. In addition, response change analyses for multiple-choice and multiple-select items were conducted. The last and next to last (if it existed) responses were compared and students or aggregates were flagged if the number or average number of wrong to right response changes was above the flagging criteria.

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

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

CAI also contracts with a third party vendor, Caveon, to detect security breaches.

4.3. TRACKING AND RESOLVING TEST IRREGULARITIES

Throughout the test window, TAs were instructed to report breaches of protocol and testing irregularities to the appropriate STC. Test irregularity requests were submitted, as appropriate, through the irregularities module under Administering Tests in TIDE.

TIDE allowed CTCs, NPSTCs, and STCs to request action to a test (e.g., re-open test, re-open test segment) in response to a test irregularity that occurred in the testing environment. In many cases, schools were required by IDOE to provide formal documentation of test irregularities before creating an Irregularity Request in TIDE.

CTCs, NPSTCs, STCs, and TAs had to discuss the details of a test irregularity to determine whether test invalidation was appropriate. CTCs, NPSTCs, and STCs were

required to submit to IDOE a *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 test window, TAs were also required to immediately report any test incidents (e.g., disruptive students, loss of Internet connectivity, student improprieties) to the STC. A test incident could include testing that was interrupted for an extended period due to a local technical malfunction or severe weather. STCs notified CTCs or NPSTCs of any test irregularities that were reported. CTCs or NPSTCs were responsible for completing test invalidations via TIDE. Schools managed the invalidation process based on local decisions or guidance from IDOE regarding test irregularities or test security concerns. This information was stored in TIDE for the school year and remained available until TIDE was updated for the 2021-2022 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 test window.
TA or teacher coaching or providing any other type of assistance to students that may affect their responses. This includes both verbal cues (e.g., interpreting, explaining, or paraphrasing the test items or prompts) and nonverbal cues (e.g., voice inflection, pointing, or nodding head) to the correct answer. This also includes leading students through instructional strategies such as think-aloud, asking students to point to the correct answer or otherwise identify the source of their answer, requiring students to show their work to the TA, or reminding students of a recent lesson on a topic.
TA providing students with unallowable materials or devices during test administration or allowing inappropriate designated features and/or accommodations during test administration.
TA providing a student access to another student’s work/responses.
TA or Test Coordinator modifying student responses or records at any time.

TA providing students with access to a calculator during a portion of the assessment that does not allow the use of a calculator.

TA uses another staff member's username and/or password to access vendor systems or administer tests.

TA uses a student's login information to access practice tests or operational tests.

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
Assessment Readiness Network
(*ILEARN*)**

2021–2022

**Volume 4
Evidence of Reliability and
Validity**

ACKNOWLEDGMENTS

This technical report was produced on behalf of the Indiana Department of Education (IDOE). For additional information concerning this technical report or the associated appendices, please contact IDOE at INassessments@doe.in.gov.

The major contributors to this technical report from Cambium Assessment, Inc. (CAI) include Stephan Ahadi, Shuqin Tao, Elizabeth Xiaoxin Wei, Maryam Pezeshki, Kevin Clayton, Christina Sneed, and Jessica Singh. The 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 THE <i>ILEARN</i> ASSESSMENTS.....	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.....	13
4.3 Reliability of Performance Classification	25
4.3.1 <i>Classification Accuracy</i>	25
4.3.2 <i>Classification Consistency</i>	30
4.4 Precision at Cut Scores.....	33
4.5 Writing Prompts Inter-Rater Reliability.....	36
5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE	39
5.1 Correlations among Reporting Category Scores	39
5.2 Confirmatory Factor Analysis	45
5.2.1 <i>Factor Analytic Methods</i>	46
5.2.2 <i>Results</i>	49
5.2.3 <i>Discussion</i>	52
5.3 Local Independence	53
5.4 Convergent and Discriminant Validity.....	55
6. FAIRNESS IN CONTENT.....	65
6.1 Statistical Fairness in Item Statistics.....	65
7. SUMMARY	67
8. REFERENCES	68

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.....	12
Table 7: Descriptive Statistics	26
Table 8: Classification Accuracy Index (ELA).....	28
Table 9: Classification Accuracy Index (Mathematics).....	28
Table 10: Classification Accuracy Index (Science).....	29
Table 11: Classification Accuracy Index (Social Studies).....	29
Table 12: False Classification Rates (ELA).....	29
Table 13: False Classification Rates (Mathematics).....	29
Table 14: False Classification Rates (Science).....	30
Table 15: False Classification Rates (Social Studies).....	30
Table 16: Classification Accuracy and Consistency (Cut 1 and Cut 2).....	31
Table 17: Classification Accuracy and Consistency (Cut 2 and Cut 3).....	32
Table 18: Classification Accuracy and Consistency (Cut 3 and Cut 4).....	32
Table 19: Performance Levels and Associated Conditional Standard Error of Measurement (ELA).....	33
Table 20: Performance Levels and Associated Conditional Standard Error of Measurement (Mathematics)	34
Table 21: Performance Levels and Associated Conditional Standard Error of Measurement (Science).....	35
Table 22: Performance Levels and Associated Conditional Standard Error of Measurement (Social Studies).....	35
Table 23: Percentage Agreement Example.....	36
Table 24: Inter-Rater Reliability	37
Table 25: Weighted Kappa Coefficients	38
Table 26: Observed Correlation Matrix Among Reporting Categories (ELA)	40
Table 27: Observed Correlation Matrix Among Reporting Categories (Mathematics)...	40
Table 28: Observed Correlation Matrix Among Reporting Categories (Science)	41
Table 29: Observed Correlation Matrix Among Reporting Categories (Social Studies).42	
Table 30: Disattenuated Correlation Matrix Among Reporting Categories (ELA).....	43
Table 31: Disattenuated Correlation Matrix Among Reporting Categories (Mathematics)	43
Table 32: Disattenuated Correlation Matrix Among Reporting Categories (Science)....	44
Table 33: Disattenuated Correlation Matrix Among Reporting Categories (Social Studies)	45
Table 34: Goodness-of-Fit Second-Order CFA	49
Table 35: Correlations Among Factors (ELA).....	50
Table 36: Correlations Among Factors (Mathematics)	51
Table 37: Correlations Among Factors (Science).....	51
Table 38: Correlations Among Factors (Social Studies).....	52

Table 39: Q ₃ Statistic (ELA)	54
Table 40: Q ₃ Statistic (Mathematics)	54
Table 41: Q ₃ Statistic (Science).....	54
Table 42: Q ₃ Statistic (Social Studies).....	55
Table 43: Grade 3 Observed Score Correlations	56
Table 44: Grade 3 Disattenuated Score Correlations.....	56
Table 45: Grade 4 Observed Score Correlations	57
Table 46: Grade 4 Disattenuated Score Correlations.....	58
Table 47: Grade 5 Observed Score Correlations	59
Table 48: Grade 5 Disattenuated Score Correlations.....	60
Table 49: Grade 6 Observed Score Correlations	61
Table 50: Grade 6 Disattenuated Score Correlations.....	62
Table 51: Grade 7 Observed Score Correlations	63
Table 52: Grade 7 Disattenuated Score Correlations.....	63
Table 53: Grade 8 Observed Score Correlations	64
Table 54: Grade 8 Disattenuated Score Correlations.....	64

LIST OF FIGURES

Figure 1: Sample Test Information Function	14
Figure 2: Conditional Standard Error of Measurement (ELA).....	15
Figure 3: Conditional Standard Error of Measurement (Mathematics)	18
Figure 4: Conditional Standard Error of Measurement (Science).....	21
Figure 5: Conditional Standard Error of Measurement (Social Studies).....	24
Figure 6: Second-Order Factor Model (Biology).....	48

LIST OF APPENDICES

Appendix A: *Reliability Coefficients*

Appendix B: *Conditional Standard Error of Measurement*

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The Indiana Learning Evaluation Assessment Readiness Network (*ILEARN*) is an online, adaptive assessment for English/Language Arts (ELA), Mathematics, and Science and an online, fixed-form assessment for Social Studies. For the 2021–2022 school year, accommodated and paper-and-pencil versions of the assessments were available to students whose Individualized Education Programs (IEPs) or Section 504 Plans indicated that need.

Table 1 displays the complete list of available test administration methods for the 2021–2022 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 1.4, Testing Accommodations and Designated Features.

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

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

- **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, Reliability Coefficients. This section also includes Conditional Standard Errors of Measurement (CSEMs), classification accuracy, and classification 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 the observed and disattenuated Pearson correlations among reporting categories by grade. Confirmatory factor analysis (CFA) 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 the true score variance to the 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 of test takers at two different points in time. If the test scores from the two test 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 test administration. The critical factor, however, is the time interval. The time interval should not be too long, in order to avoid potential 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 method was not used for the *ILEARN* assessments, as there was a single test for all students.
- The *parallel-forms* method is used for measuring equivalence. This method involves administering two parallel forms of a test to the same group of test takers. 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 measure 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 the inferences regarding high reliability of scores can be substantiated. This method is commonly used to estimate the reliability of performance on aptitude tests. Since this method also requires two sets of student scores, it was also not used for the *ILEARN* assessments.

- The *split-half* method uses one test divided into two halves within a single test administration. It is crucial to construct 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 item splits 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 percentage agreement and weighted kappa was used to summarize writing prompt handscoring reliability.

The first four methods just discussed 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. The precision of individual test scores is critically important to valid test score interpretation and is provided along with test scores as part of all student-level reporting. In addition, the test-retest and parallel-forms 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, the classical test theory (CTT) 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 the true score variance to the 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 CTT SEM, which assumes a homoscedastic error, is derived from the classical notion expressed previously as $\sigma_X\sqrt{1 - \rho_{XX'}}$, where σ_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:

$$\begin{aligned}\rho_{XX'} &= 1 - \frac{\sigma_E^2}{\sigma_X^2}, \\ \frac{\sigma_E^2}{\sigma_X^2} &= 1 - \rho_{XX'}, \\ \sigma_E^2 &= \sigma_X^2(1 - \rho_{XX'}), \\ \sigma_E &= \sigma_X\sqrt{(1 - \rho_{XX'})}.\end{aligned}$$

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

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

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 the IRT vary over the ability continuum. These heterogeneous errors are a function of a test information function (TIF) that provides different information about test takers depending on their estimated abilities. Often, the 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 the 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 definitions emphasize evidence and theory to support the 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 validity evidence is the relationship between the test content and the intended test construct (see Section 3.1, Content Standards). 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, Test Development 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* assessments 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, Test Development, for details).

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

The third source of validity evidence is based on the internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. 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 4.2, Differential Item Functioning Analysis, for details). 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 on Internal-External Structure, for details).

A fourth source of validity evidence is the relationship of the test scores to external variables. The *Standards* (AERA, APA, & NCME, 2014) divide this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs; conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. A multi-trait multi-method matrix can be used to analyze both convergent and discriminant evidence (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 on the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

The fifth source of validity evidence is that the intended and unintended consequences of test use should be included in the test validation process. Determining the validity of the test should depend upon evidence directly related to the test; external factors should not influence this process. For example, if an employer administers a test to determine the hiring rates for different groups of people and the results indicate an unequal distribution of skills related to the measurement construct, that would not necessarily imply a lack of test validity. However, if the unequal distribution of scores is, in fact, due to an unintended, confounding aspect of the test, that would interfere with the test's validity. As described in Volume 1 of this technical report and here in Volume 4, test use should align with the test's intended purpose.

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

2. PURPOSE OF THE *ILEARN* ASSESSMENTS

The primary purpose of the Indiana Learning Evaluation Assessment Readiness Network (*ILEARN*) assessments is to yield test scores at the student level and other levels of aggregation that reflect student performance relative to the Indiana Academic Standards (IAS). *ILEARN* supports instruction and student learning by measuring proficiency and growth in student performance and providing feedback to educators and parents that can be used to inform instructional strategies to remediate or enrich instruction. The 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 can positively affect student learning over time.

The tests are constructed to measure student proficiency on the IAS in English/Language Arts (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 of the technical report 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, as described in Section 4, Reliability.

The *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 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, Score Interpretation Guide, of this technical report 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 Indiana Learning Evaluation Assessment Readiness Network (*ILEARN*) assessments 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 the *ILEARN* tests to the standards. A complete description of the test development process is available in Volume 2, Test Development.

3.1 CONTENT STANDARDS

The Indiana Academic Standards (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. These IAS were most recently updated in 2020; however these minimally updated versions of the standards will not be assessed until Spring 2023. The Science IAS 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 communication skills that promote college and career readiness.

ILEARN blueprints are available in the appendices in Volume 2, Test Development. The blueprints were developed to ensure that both 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 is available in Volume 2, Test Development, Section 4, *ILEARN* Blueprints and State Assessment Test Construction.

Table 2 through Table 5 present the domains by grade and test and the number of items measuring each domain on the 2021–2022 assessments. 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 Domain (ELA)

Domain	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
Reading Foundations	0-2					

*Writing item ranges do not include performance task items (one per grade) to account for adjustments made to ensure that all students meet blueprint minimums.

Table 3: Number of Items for Each Domain (Mathematics)

Grade	Domain	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	11-12
	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	Analyzing, Interpreting, and Computational Thinking	10-12
	Explaining Solutions, Reasoning, and Communicating	10-12
	Investigating	12-14
	Questioning and Modeling	12-14
6	Analyzing, Interpreting, and Computational Thinking	12-14
	Explaining Solutions, Reasoning, and Communicating	12-14
	Investigating	10-12
	Questioning and Modeling	10-12
Biology*	Analyzing Data and Mathematical Thinking	10-12
	Constructing and Communicating an Explanation	10-12
	Developing and Using Models to Explain Processes	10-12
	Developing and Using Models to Describe Structure and Function	10-12
	Evaluating Claims with Evidence	10-12

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

The marginal reliability is defined as:

$$\bar{\rho} = 1 - \frac{\int \sigma_e^2(\hat{\theta})f(\hat{\theta})d\hat{\theta}}{\sigma_x^2}$$

where $\sigma_e^2(\hat{\theta})$ is the function generating the standard error of measurement and $f(\hat{\theta})$ is the assumed population density.

The marginal reliability can be calculated using two approaches: the theoretical approach and the empirical approach. For the theoretical approach, the marginal reliability of a test is computed by integrating θ out of the test information function as follows:

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

where σ_θ^2 is the true score variance of θ and

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

where $g(\theta)$ is a density function. If population parameters are assumed normal, then $g(\theta) \sim N(\mu, \sigma^2)$. In the absence of information about the population distribution of θ , a uniform prior is available such that $g(\theta) \sim U[a, b]$ where a and b are the lower and upper limits of the uniform distribution, respectively. The integral is evaluated using Gauss-Hermite quadrature:

$$\bar{\sigma}_e^2 \approx \sum_{q=1}^Q \frac{1}{I(\theta_q)} w_q$$

where θ_q is the value at node q and w_q is the weight at node q . The true score variance of θ can be obtained from the marginal maximum likelihood (MML) means procedure.

In IRT, the marginal likelihood is typically maximized to estimate item parameters by integrating θ out of the function and treating population parameters as known. However, suppose the item parameters are treated as fixed but the population parameters are treated as latent. Then, the following marginal likelihood can be maximized with respect to the two latent parameters associated with the normal population distribution:

$$\arg \max L(\mu, \sigma) = \prod_{i=1}^N \int_{-\infty}^{\infty} \prod_{j=1}^K p(x_j | \theta_i, \mathbf{Y}_j) g(\theta | \mu, \sigma) d\theta$$

where in this context $p(x_j | \theta_i, \mathbf{Y}_j)$ is used to mean the probability of individual $i = \{1, 2, \dots, N\}$ having observed response x to item $j = \{1, 2, \dots, K\}$ given the vector of item parameters, \mathbf{Y} . The integral has no closed form and so the function is evaluated using a fixed quadrature routine. Rather than using Gauss-Hermite, Q nodes are chosen from the normal distribution at fixed points and then the integral is evaluated by summation over the Q nodes as:

$$\arg \max L(\mu, \sigma) = \prod_{i=1}^N \sum_{q=1}^Q \prod_{j=1}^K p(x_j | \theta_q, \mathbf{Y}_j) g(\theta_q | \mu, \sigma)$$

where θ_q is node q . In this instance, fixed quadrature points allow a smaller number of likelihood evaluations because the values for θ_q are fixed. If Gauss-Hermite were used, the nodes would change as each value of μ and σ are updated and the likelihood calculations would need to be performed at each iteration.

The empirical approach of the marginal reliability can be calculated using the following formulae:

$$\bar{\rho} = 1 - \frac{\sum_{i=1}^N CSEM_i^2 / N}{\sigma_x^2}$$

where N is the number of students, $CSEM_i$ is the conditional SEM of the scaled score of student i , and σ_x^2 is the variance in observed scaled scores of students. Marginal reliability coefficients reported in the technical report are calculated using the empirical approach.

Table 6 presents the marginal reliability coefficients for all students. The marginal reliability coefficients for all subjects and grades ranged from 0.871 to 0.960, which is similar to other statewide standardized tests.

Table 6: Marginal Reliability Coefficients

Grade	Marginal Reliability
ELA 3	0.895
ELA 4	0.899

Grade	Marginal Reliability
ELA 5	0.896
ELA 6	0.889
ELA 7	0.894
ELA 8	0.902
Mathematics 3	0.960
Mathematics 4	0.955
Mathematics 5	0.952
Mathematics 6	0.951
Mathematics 7	0.945
Mathematics 8	0.944
Science 4	0.907
Science 6	0.912
Biology (Fall)	0.919
Biology (Winter)	0.917
Biology (Spring)	0.927
Social Studies 5	0.871
U.S. Government	0.885

4.2 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

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

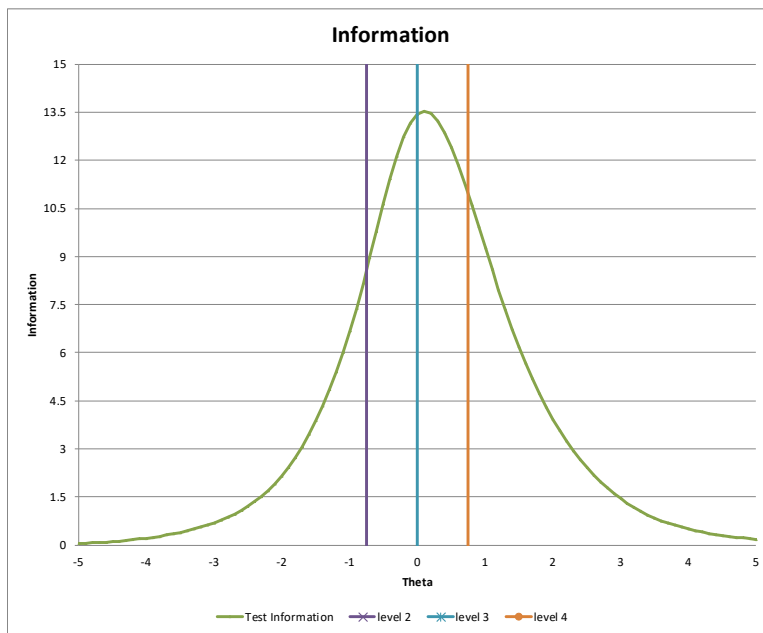
Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. Where the curve is lower at the tails indicates that 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 Indiana Learning Evaluation Assessment Readiness Network (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 (GPC) items, N_{2PL} is the number of items scored using the two-parameter logistic (2PL) model, i indicates item i ($i \in \{1, 2, \dots, N\}$), m_i is the maximum possible score of the item, s indicates student s , and θ_s is the ability of student s .

Figure 1: Sample Test Information Function

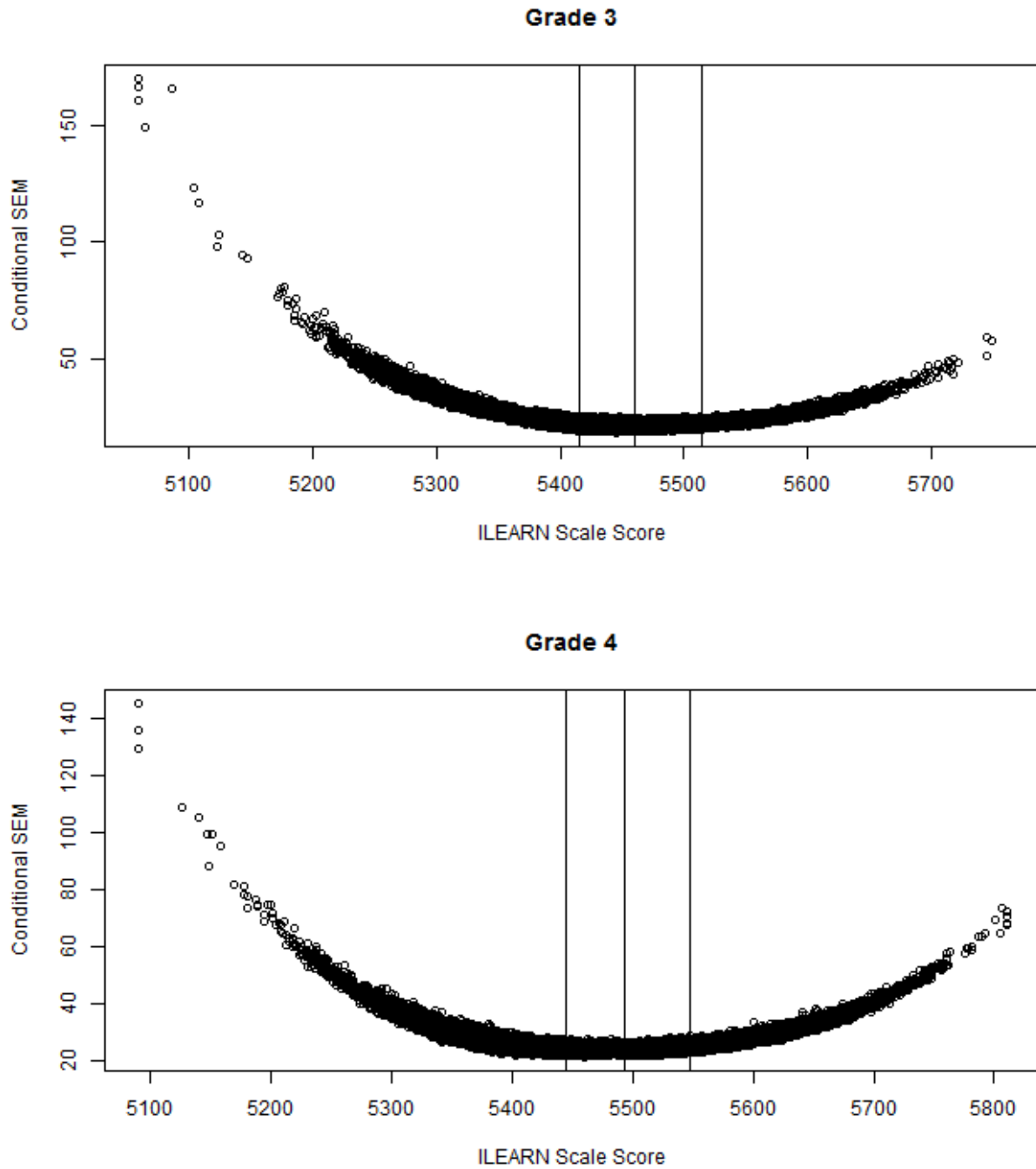


The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

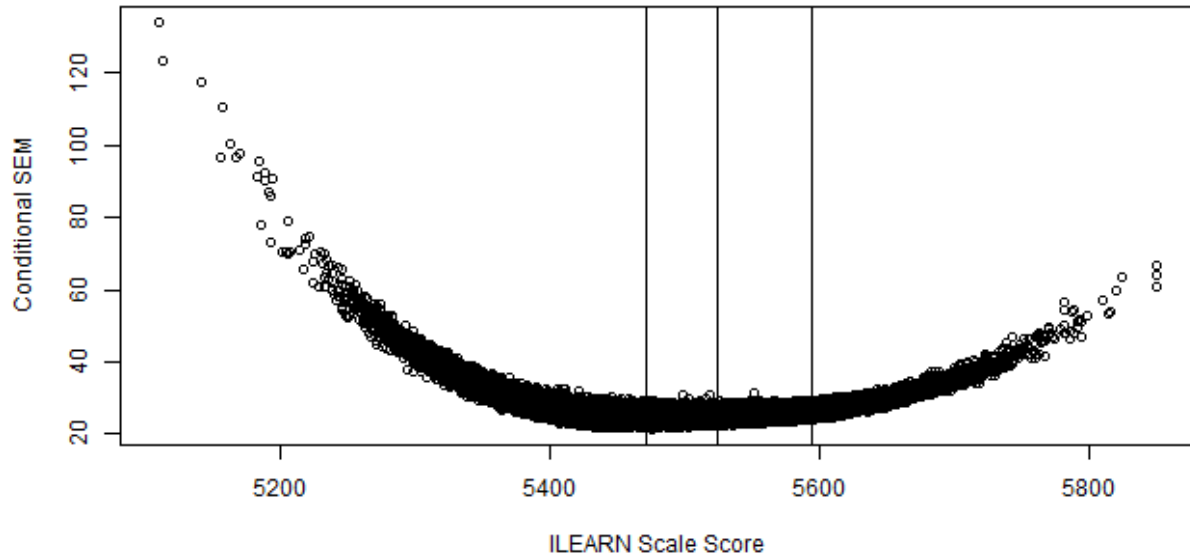
$$se(\theta_s) = \frac{1}{\sqrt{TIF(\theta_s)}}$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the SEMs are more useful for score interpretation. For this reason, standard error plots are presented in Figures 2–5. These plots are based on the scaled scores reported in 2021–2022. Vertical lines represent the performance category cut scores.

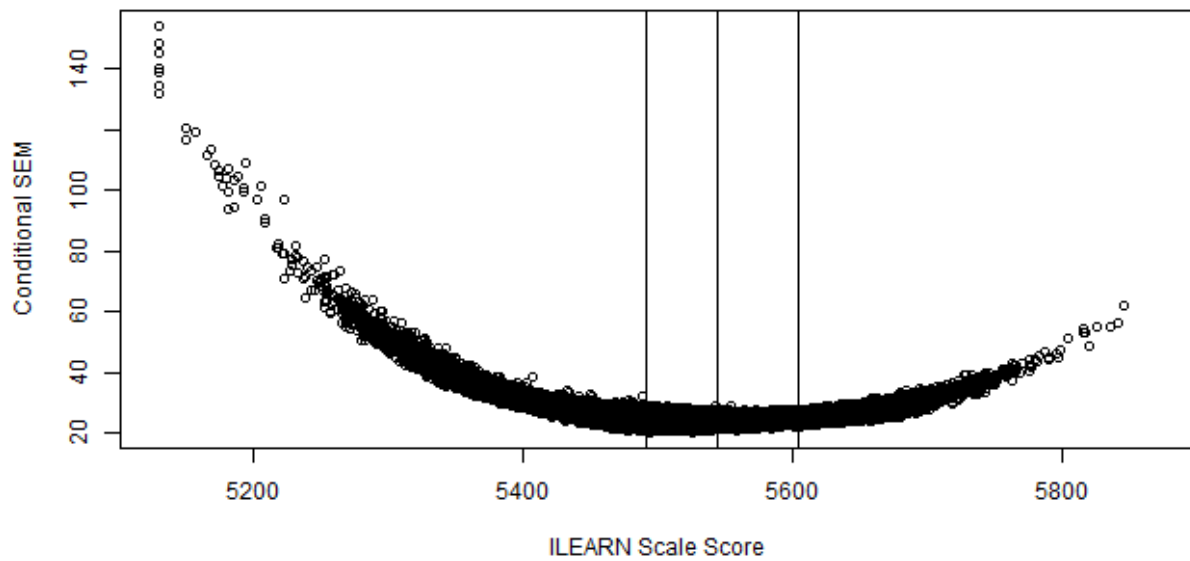
Figure 2: Conditional Standard Error of Measurement (ELA)



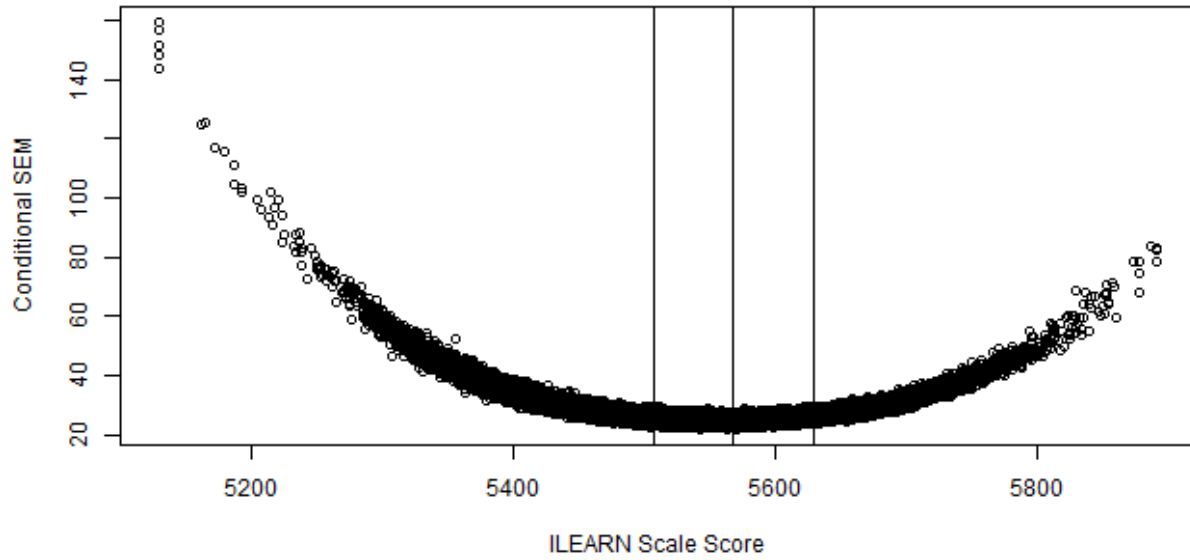
Grade 5



Grade 6



Grade 7



Grade 8

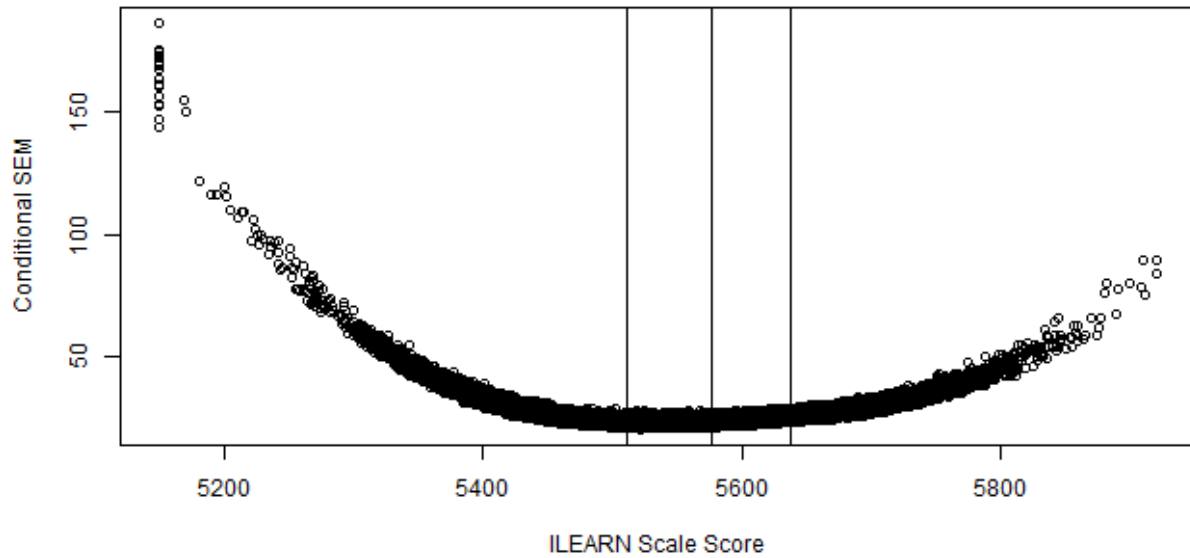
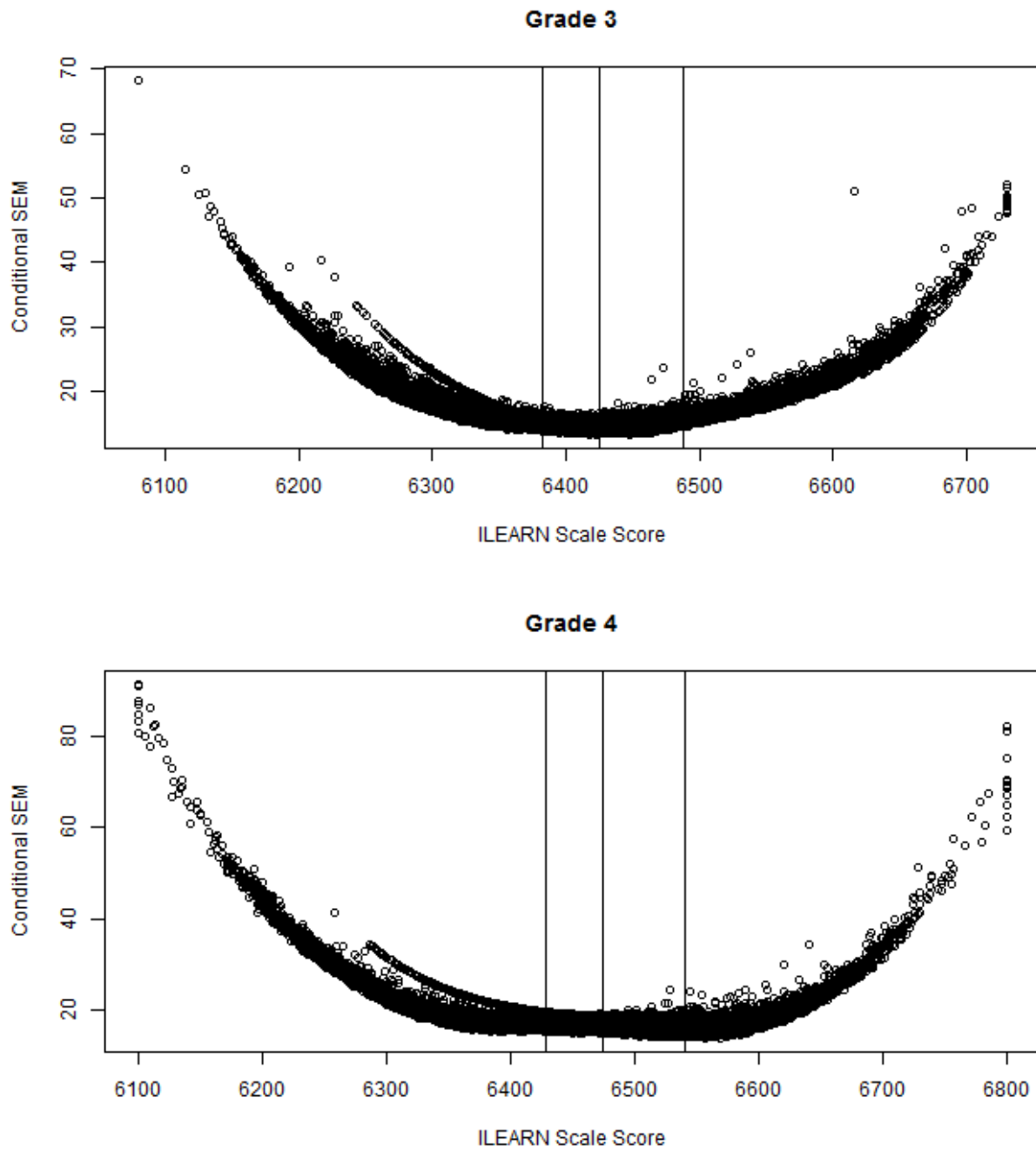
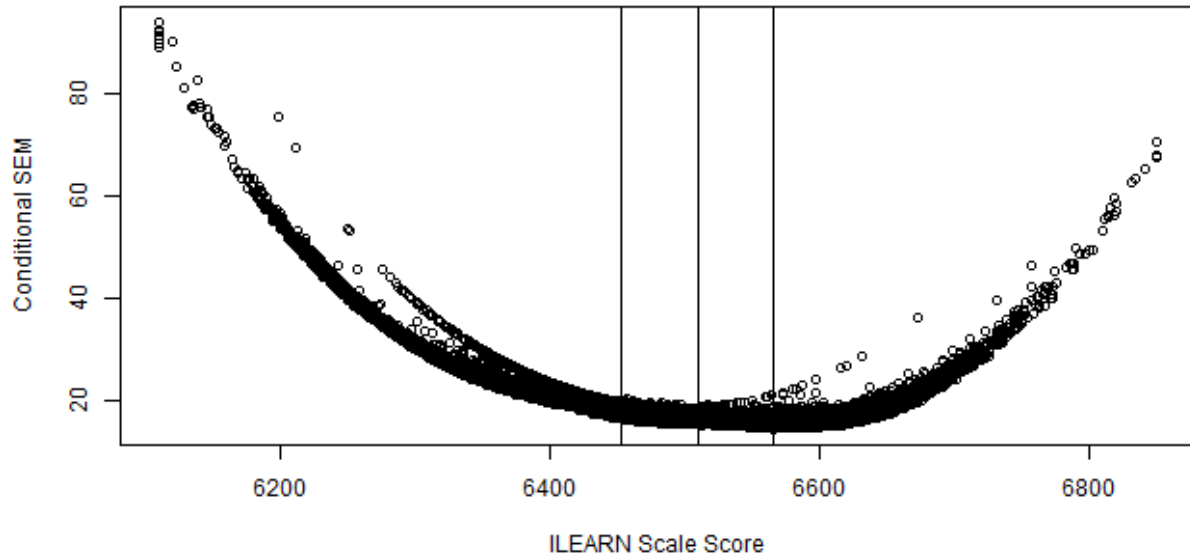


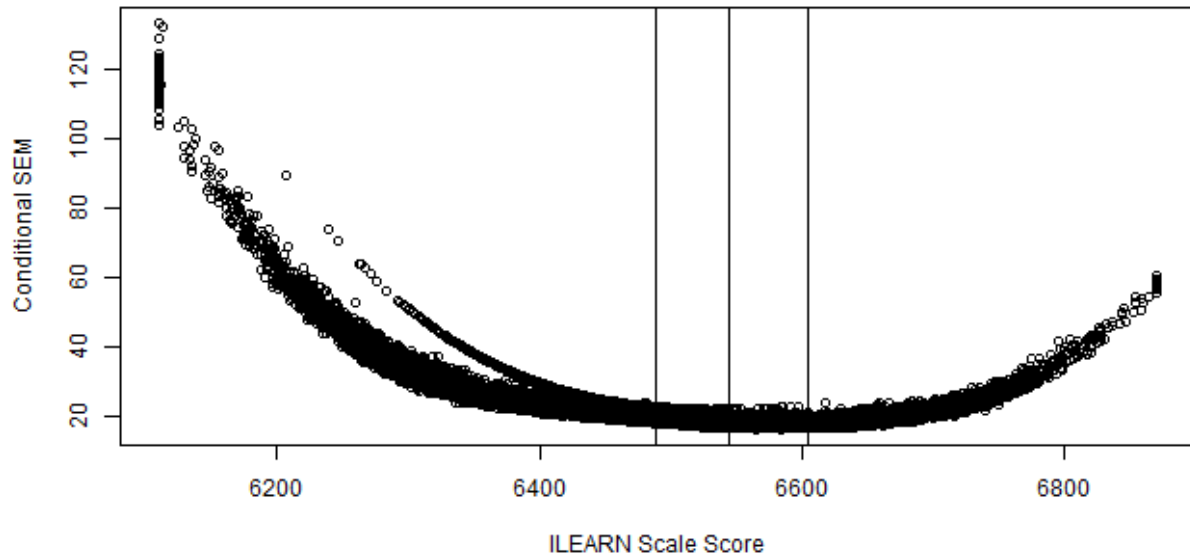
Figure 3: Conditional Standard Error of Measurement (Mathematics)



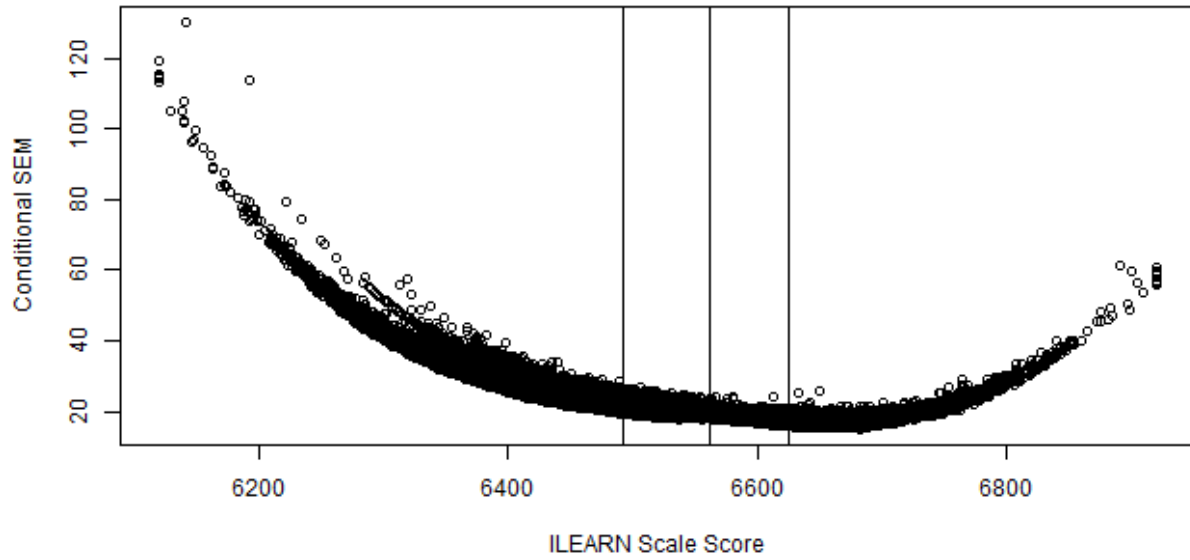
Grade 5



Grade 6



Grade 7



Grade 8

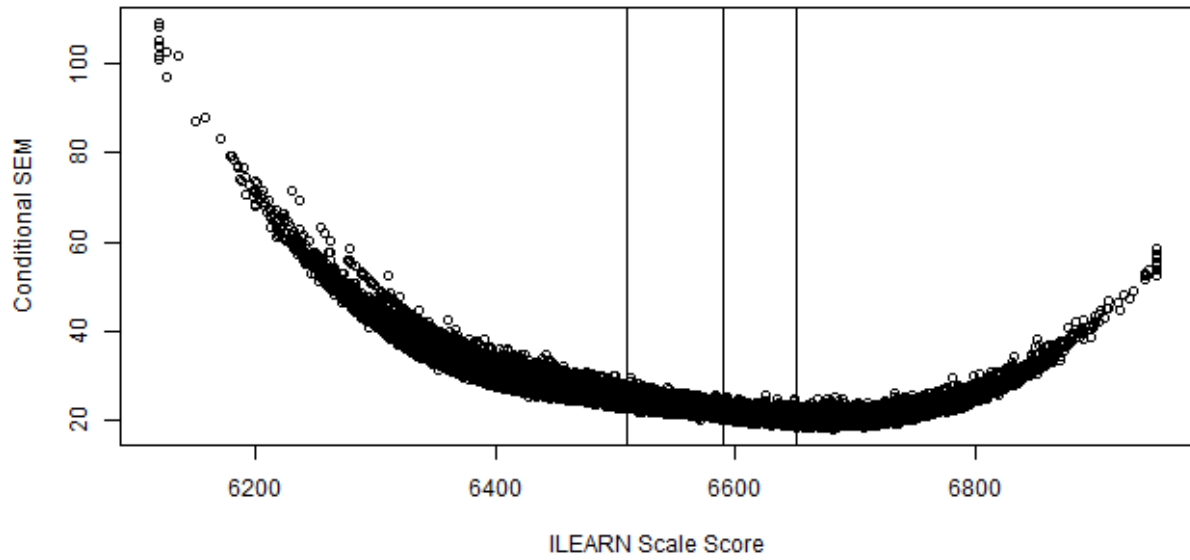
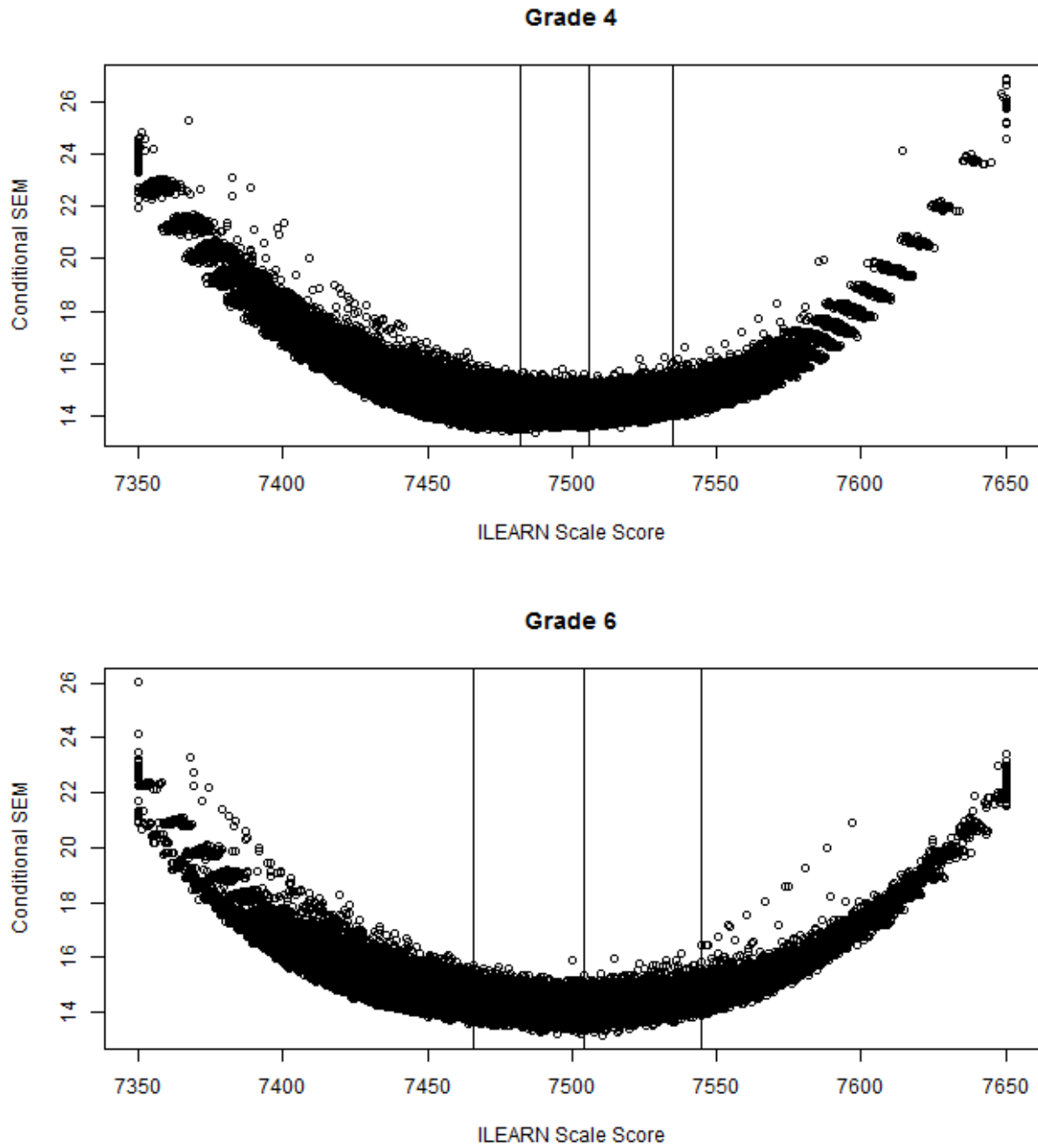
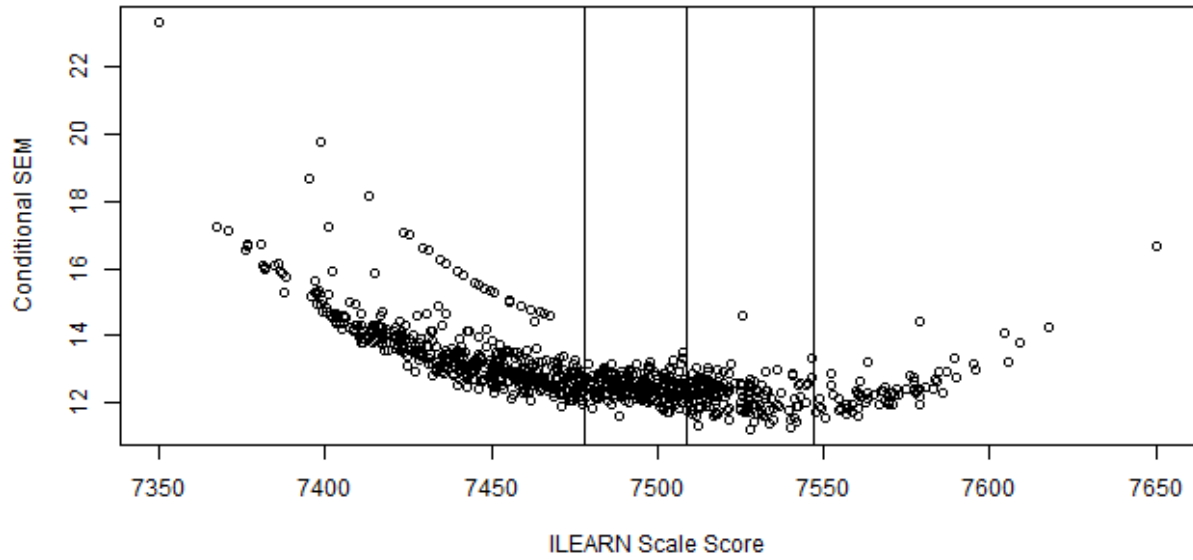


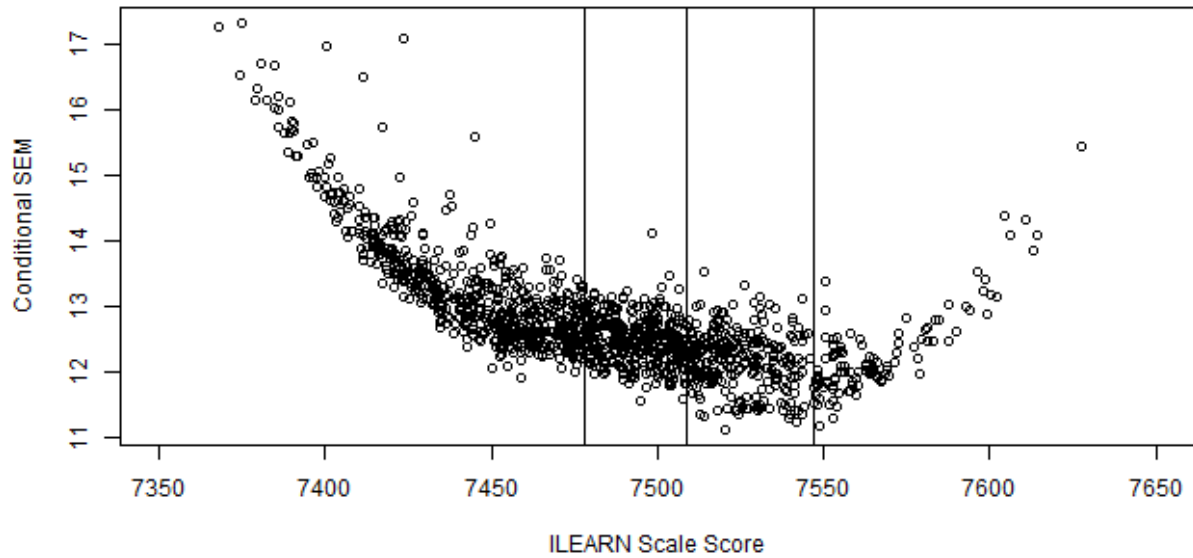
Figure 4: Conditional Standard Error of Measurement (Science)



Biology Fall



Biology Winter



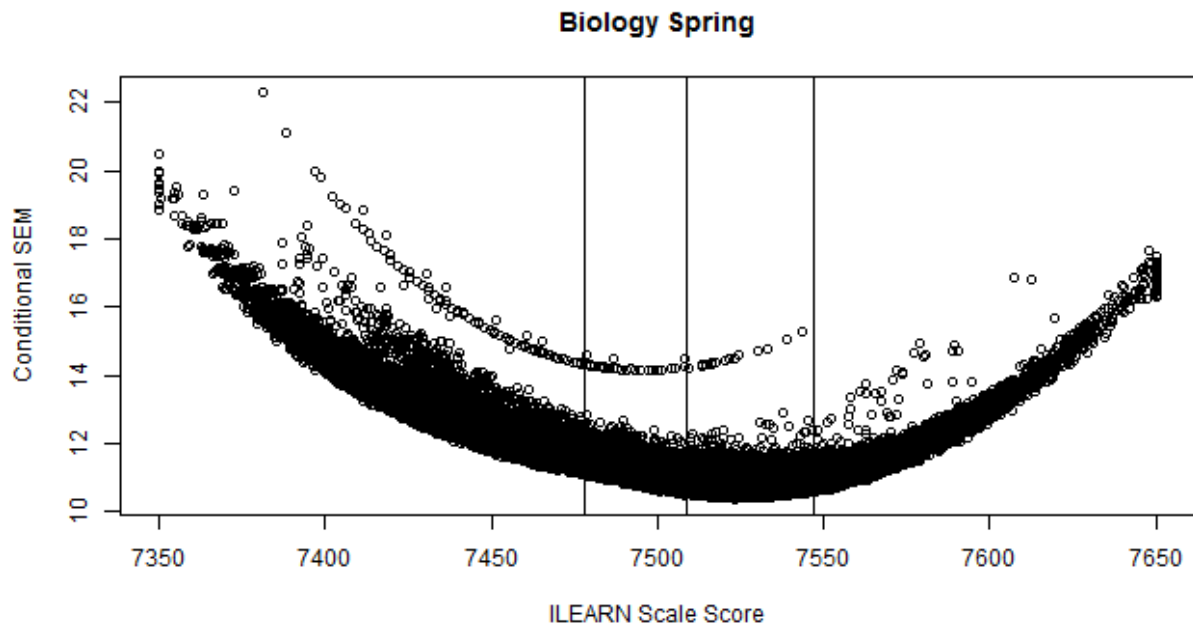
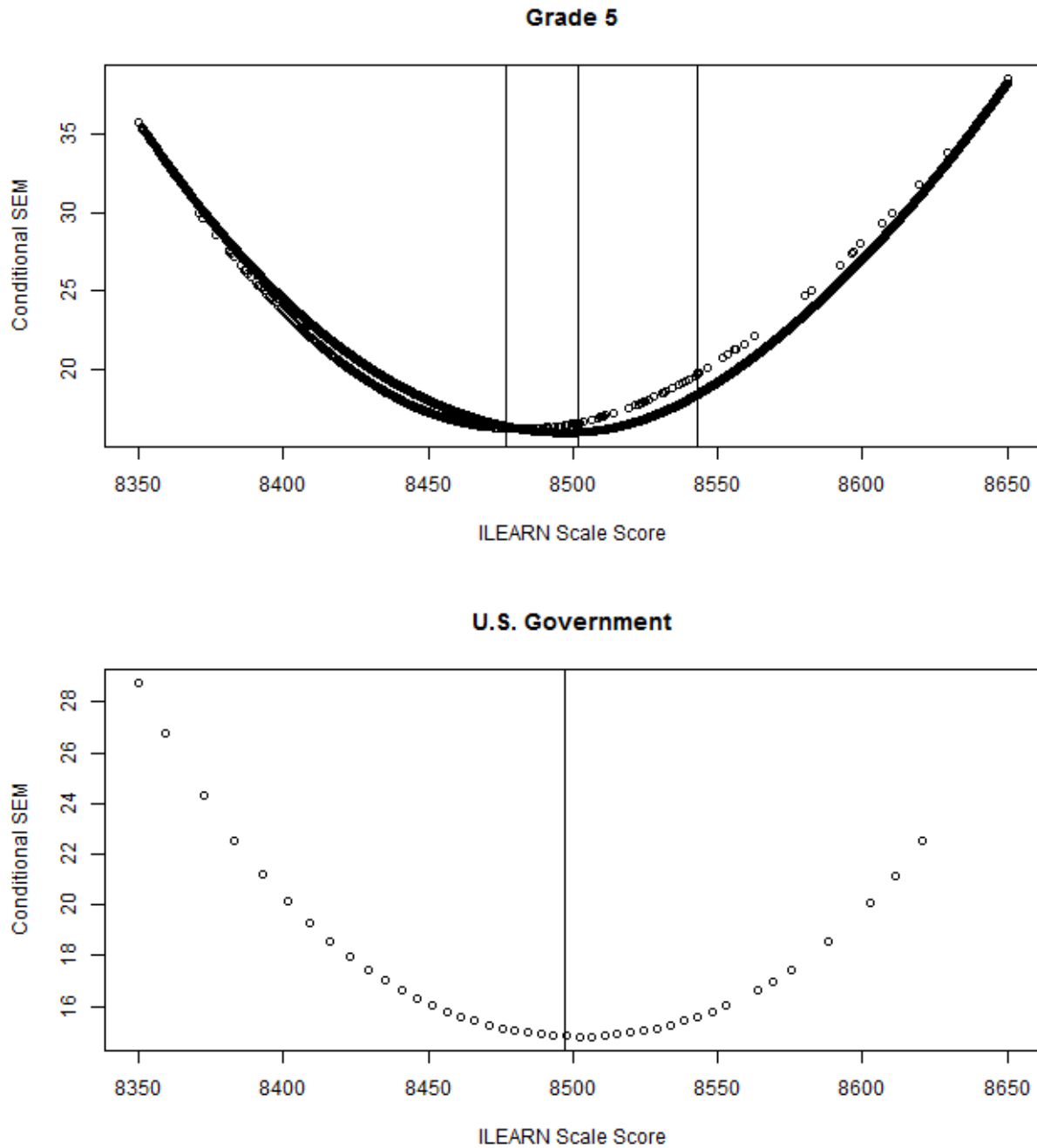


Figure 5: Conditional Standard Error of Measurement (Social Studies)



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

The reliability coefficients and SEM for each reporting category are also presented in Appendix A, Reliability Coefficients, and Appendix B, Conditional Standard Error of

Measurement, and include the average Conditional Standard Error of Measurement (CSEM) by scale score and corresponding performance levels for each scale score.

4.3 RELIABILITY OF PERFORMANCE CLASSIFICATION

Students who complete the *ILEARN* assessments are placed into performance levels by their observed scaled scores. The cut scores for classifying students into different performance levels were determined after the *ILEARN* standard-setting process. A complete description of the standard-setting process is available in Volume 6, Setting Performance Standards.

4.3.1 Classification Accuracy

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 (Rudner, 2001) 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 (Lee, Hanson, & Brennan, 2002; Guo, 2006) 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.

For these analyses, we used students from the *ILEARN* 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 the scoring engine.

Table 7: Descriptive Statistics

Grade	Sample Size	Mean Theta	Standard Deviation of Theta	Mean Scale Score	Standard Deviation of Scale Scores
ELA 3	79,915	-0.81	1.00	5439.07	74.77
ELA 4	81,003	-0.36	1.11	5473.04	83.07
ELA 5	81,102	-0.01	1.12	5499.60	84.22
ELA 6	82,180	0.23	1.08	5517.12	80.85
ELA 7	83,346	0.62	1.14	5546.30	85.85
ELA 8	84,990	0.76	1.14	5557.31	85.86
Mathematics 3	79,940	-1.00	1.10	6425.07	82.87
Mathematics 4	80,990	-0.48	1.11	6464.18	83.35
Mathematics 5	81,080	-0.23	1.19	6483.05	89.16
Mathematics 6	82,102	0.05	1.32	6503.45	99.33
Mathematics 7	83,262	0.16	1.34	6512.11	100.42
Mathematics 8	84,897	0.35	1.49	6526.30	111.82
Science 4	80,848	-0.26	1.01	7486.96	50.49
Science 6	81,904	-0.25	1.01	7487.27	50.38
Biology (Fall)	931	-0.44	0.92	7477.84	46.11
Biology (Winter)	1,381	-0.35	0.89	7482.35	44.63
Biology (Spring)	81,292	-0.27	0.89	7486.65	44.34
Social Studies 5	80,939	-0.20	1.07	8490.24	53.28
U.S. Government	278	-1.04	1.07	8447.94	53.33

The observed score approach (Rudner, 2001), implemented to assess classification accuracy, is based on the probability that the true score, θ , 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 a SEM of $\hat{\sigma}_j$. Using the asymptotic property of normality of the 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

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

where pl_j is the j th student's performance level and the values of E_{vl} are the elements used to populate the matrix E , a 4×4 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

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

where $N = \sum_{v=1}^4 N_v$ and N_v is the observed number of students scoring in performance level v . The CAI for the individual cut p , ($CAIC_p$), is estimated by forming square partitioned blocks of the matrix 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.

The IRT-based approach (Lee, Hanson, & Brennan, 2002; Guo, 2006) uses student-level item response data from the 2022 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, θ_j is the true ability in the true-score metric, \mathbf{z}_j is the item score, μ is the mean, and σ 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 θ 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:

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

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

Table 8 through Table 11 provide the overall CAI and the classification accuracy index for the individual cuts (CAIC) based on the observed score approach (Rudner, 2001). There is no industry standard, but these numbers suggest that misclassification would not be frequent in the population data.

The cut accuracy rates were much higher, denoting that the degree to which we can reliably differentiate between students of adjacent performance levels is mostly in 0.90s.

Table 8: Classification Accuracy Index (ELA)

Grade	Overall Accuracy Index	Cut Accuracy Index		
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4
3	0.779	0.920	0.918	0.939
4	0.771	0.918	0.916	0.934
5	0.774	0.921	0.913	0.940
6	0.767	0.915	0.912	0.938
7	0.767	0.924	0.910	0.932
8	0.777	0.929	0.916	0.931

Table 9: Classification Accuracy Index (Mathematics)

Grade	Overall Accuracy Index	Cut Accuracy Index		
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4
3	0.849	0.953	0.945	0.951
4	0.840	0.945	0.939	0.956
5	0.845	0.945	0.942	0.958
6	0.843	0.940	0.942	0.961
7	0.851	0.938	0.947	0.966
8	0.853	0.939	0.947	0.966

Table 10: Classification Accuracy Index (Science)

Grade	Overall Accuracy Index	Cut Accuracy Index		
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4
4	0.771	0.920	0.911	0.933
6	0.792	0.929	0.913	0.950
Biology (Fall)	0.831	0.921	0.929	0.980
Biology (Winter)	0.811	0.920	0.922	0.969
Biology (Spring)	0.821	0.920	0.931	0.970

Table 11: Classification Accuracy Index (Social Studies)

Grade	Overall Accuracy Index	Cut Accuracy Index		
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4
5	0.768	0.902	0.917	0.943
U.S. Government*	0.958	0.958		

*U.S. Government has only one cut.

Table 12 through Table 15 provide the FPR, FNR, and accuracy index based on the IRT-based method (Lee, Hanson, & Brennan, 2002; Guo, 2006). The FPR and FNR rates for the level 2/3 cut were between 0.022 and 0.048. The accuracy rates for the level 2/3 cut were between 0.906 and 0.950.

Table 12: False Classification Rates (ELA)

Grade	1/2 cut			2/3 cut			3/4 cut		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
3	0.042	0.037	0.921	0.042	0.041	0.917	0.027	0.036	0.937
4	0.044	0.036	0.920	0.042	0.044	0.914	0.029	0.040	0.931
5	0.043	0.035	0.922	0.045	0.045	0.910	0.025	0.038	0.937
6	0.045	0.039	0.916	0.044	0.046	0.910	0.027	0.039	0.934
7	0.042	0.033	0.925	0.048	0.046	0.906	0.028	0.042	0.930
8	0.038	0.031	0.931	0.044	0.042	0.914	0.032	0.041	0.927

Table 13: False Classification Rates (Mathematics)

Grade	1/2 cut			2/3 cut			3/4 cut		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
3	0.025	0.023	0.952	0.028	0.026	0.946	0.023	0.026	0.951
4	0.029	0.026	0.945	0.031	0.030	0.939	0.020	0.025	0.955
5	0.028	0.027	0.945	0.029	0.029	0.942	0.018	0.024	0.958

Grade	1/2 cut			2/3 cut			3/4 cut		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
6	0.030	0.028	0.942	0.029	0.030	0.941	0.018	0.022	0.960
7	0.036	0.028	0.936	0.024	0.028	0.948	0.014	0.019	0.967
8	0.032	0.030	0.938	0.022	0.028	0.950	0.011	0.019	0.970

Table 14: False Classification Rates (Science)

Grade	1/2 cut			2/3 cut			3/4 cut		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
4	0.043	0.036	0.921	0.044	0.044	0.912	0.028	0.039	0.933
6	0.041	0.030	0.929	0.044	0.043	0.913	0.021	0.030	0.949
Biology (Fall)	0.045	0.033	0.922	0.033	0.037	0.930	0.008	0.012	0.980
Biology (Winter)	0.044	0.036	0.920	0.037	0.040	0.923	0.014	0.017	0.969
Biology (Spring)	0.045	0.035	0.920	0.032	0.037	0.931	0.013	0.017	0.970

Table 15: False Classification Rates (Social Studies)

Grade	1/2 cut			2/3 cut			3/4 cut		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy	FPR	FNR	Accuracy
5	0.047	0.049	0.904	0.039	0.043	0.918	0.022	0.034	0.944
U.S. Government*	0.024	0.020	0.956	NA	NA	NA	NA	NA	NA

*U.S. Government has only one cut.

4.3.2 Classification Consistency

Classification accuracy refers to the degree to which a student’s true and observed scores falls 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 classification consistency index for the individual cut c , ($CICC_c$), was estimated using the following equation:

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

The classification consistency with classification accuracy results based on the IRT-based method (Lee, Hanson, & Brennan, 2002) are presented in Table 16 through Table 18. All accuracy values were in the 0.90s, and the consistency values ranged in the high 0.80s and low 0.90s. The classification accuracy was slightly higher than the classification consistency across all grades and subjects and in all performance levels. The classification consistency rates can be lower than the classification accuracy because the consistency is based on two tests with measurement errors. In contrast, 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 16: Classification Accuracy and Consistency (Cut 1 and Cut 2)

Grade	Accuracy	Consistency
ELA 3	0.921	0.889
ELA 4	0.920	0.887
ELA 5	0.922	0.889
ELA 6	0.916	0.882
ELA 7	0.925	0.894
ELA 8	0.931	0.903
Mathematics 3	0.952	0.933
Mathematics 4	0.945	0.922
Mathematics 5	0.945	0.922
Mathematics 6	0.942	0.919
Mathematics 7	0.936	0.910
Mathematics 8	0.938	0.912
Science 4	0.921	0.889
Science 6	0.929	0.901
Biology (Fall)	0.920	0.888
Biology (Winter)	0.922	0.889
Biology (Spring)	0.920	0.888
Social Studies 5	0.904	0.865
U.S. Government	0.956	0.936

Table 17: Classification Accuracy and Consistency (Cut 2 and Cut 3)

Grade	Accuracy	Consistency
ELA 3	0.917	0.884
ELA 4	0.914	0.878
ELA 5	0.910	0.874
ELA 6	0.910	0.873
ELA 7	0.906	0.868
ELA 8	0.914	0.878
Mathematics 3	0.946	0.922
Mathematics 4	0.939	0.913
Mathematics 5	0.942	0.918
Mathematics 6	0.941	0.917
Mathematics 7	0.948	0.930
Mathematics 8	0.950	0.934
Science 4	0.912	0.877
Science 6	0.913	0.878
Biology (Fall)	0.930	0.903
Biology (Winter)	0.923	0.892
Biology (Spring)	0.931	0.903
Social Studies 5	0.918	0.884

Table 18: Classification Accuracy and Consistency (Cut 3 and Cut 4)

Grade	Accuracy	Consistency
ELA 3	0.937	0.911
ELA 4	0.931	0.903
ELA 5	0.937	0.911
ELA 6	0.934	0.908
ELA 7	0.930	0.904
ELA 8	0.927	0.898
Mathematics 3	0.951	0.931
Mathematics 4	0.955	0.937
Mathematics 5	0.958	0.940
Mathematics 6	0.960	0.943
Mathematics 7	0.967	0.956
Mathematics 8	0.970	0.962

Grade	Accuracy	Consistency
Science 4	0.933	0.905
Science 6	0.949	0.928
Biology (Fall)	0.980	0.970
Biology (Winter)	0.969	0.956
Biology (Spring)	0.970	0.957
Social Studies 5	0.944	0.923

4.4 PRECISION AT CUT SCORES

Table 19 through Table 22 present the mean CSEM at each performance level by administration. The tables include performance-level cut scores and associated CSEM. The *ILEARN* test scores are somewhat more precise for scores near the middle of the scale, especially around the At Proficiency performance standard cut. The following tables also show that test scores remain precise even for students in the lowest and highest performance levels.

Table 19: Performance Levels and Associated Conditional Standard Error of Measurement (ELA)

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	26.699	-	-
	2	21.337	5416	21.880
	3	21.257	5460	20.983
	4	23.881	5515	22.000
4	1	27.282	-	-
	2	24.104	5444	24.084
	3	24.520	5493	24.241
	4	28.052	5547	25.143
5	1	28.278	-	-
	2	25.034	5472	24.868
	3	25.716	5524	25.191
	4	28.865	5595	26.553
6	1	29.282	-	-
	2	24.218	5492	24.452
	3	24.446	5544	24.248
	4	26.734	5604	24.881
7	1	30.512	-	-
	2	24.955	5507	25.509

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
	3	25.357	5568	24.791
	4	29.442	5629	26.415
8	1	28.541	-	-
	2	23.666	5511	23.824
	3	24.952	5577	24.130
	4	28.865	5638	26.175

Table 20: Performance Levels and Associated Conditional Standard Error of Measurement (Mathematics)

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	17.328	-	-
	2	14.602	6382	15.019
	3	14.856	6425	14.547
	4	17.904	6488	15.744
4	1	18.986	-	-
	2	16.621	6429	16.778
	3	16.070	6474	16.429
	4	17.265	6541	15.791
5	1	22.884	-	-
	2	17.095	6453	18.071
	3	16.318	6510	16.573
	4	17.144	6566	16.000
6	1	24.278	-	-
	2	19.127	6488	19.800
	3	18.268	6545	18.490
	4	19.818	6605	18.317
7	1	27.475	-	-
	2	20.262	6493	21.045
	3	18.745	6562	19.632
	4	17.954	6625	17.684
8	1	29.943	-	-
	2	23.473	6509	24.872
	3	21.578	6590	22.203
	4	21.758	6651	20.978

Table 21: Performance Levels and Associated Conditional Standard Error of Measurement (Science)

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
4	1	15.828	-	-
	2	14.676	7482	14.652
	3	14.734	7506	14.722
	4	15.641	7535	14.799
6	1	15.581	-	-
	2	14.290	7466	14.508
	3	14.353	7504	14.214
	4	15.583	7545	14.801
Biology (Fall)	1	13.730	-	-
	2	12.477	7478	13.000
	3	12.252	7509	12.333
	4	12.423	7547	13.000
Biology (Winter)	1	13.370	-	-
	2	12.470	7478	12.714
	3	12.147	7509	12.308
	4	12.315	7547	12.000
Biology (Spring)	1	12.744	-	-
	2	11.291	7478	11.668
	3	11.035	7509	11.074
	4	11.835	7547	11.082

Table 22: Performance Levels and Associated Conditional Standard Error of Measurement (Social Studies)

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	19.030	-	-
	2	16.000	8477	16.000
	3	16.657	8502	16.006
	4	23.537	8543	18.025
U.S. Government	1	18.321	-	-
	2	15.870	8497	15.000

4.5 WRITING PROMPTS INTER-RATER RELIABILITY

The basic method to compute inter-rater reliability (IRR) is percentage agreement. All English/Language Arts (ELA) writing prompts were hand-scored by a human with a 15% second read. As shown in Table 23, the percentage of exact agreement (when two raters gave the same score), the percentage of adjacent ratings (when the difference between two raters was 1), and the percentage of non-adjacent ratings (when the difference was greater than 1) were all computed. In this example, the percentage of exact agreement was 2/4, or 50%, and the adjacent and non-adjacent percentages were 25% each.

Table 23: Percentage Agreement Example

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

Likewise, IRR monitors how often scorers are in exact agreement with each other and ensures that an acceptable agreement rate is maintained. In cases where IRR begins to track below acceptable values, IDOE content experts review the items and scoring materials to consider improvements that might increase reliability. CAI further reviews cases of low IRR with the scoring manager to gain insights into where any score discrepancies may be occurring, based on feedback from scorers received during scoring. The calculations for the IRR in this report are as follows:

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

Table 24 displays the rater-agreement percentages. The percentage of exact agreement between two raters ranged from 59% to 73%. The percentage of adjacent rating was between 25% and 39%. The non-adjacent percentages fell between 1% and 4%. The total number of processed responses does not necessarily correspond to the number of student responses selected to be second read by a human reader. These numbers could potentially be higher, as some students might request rescoring and have their responses rescored as requested.

Table 24: Inter-Rater Reliability

Grade	Dimension	% Exact	% Adjacent	% Not Adjacent	Total Number of Processed Responses
3	Purpose, Focus, & Organization	60	36	3	7,504
	Evidence & Elaboration	60	36	4	
	Conventions	61	37	2	
4	Purpose, Focus, & Organization	62	35	3	9,692
	Evidence & Elaboration	62	35	3	
	Conventions	59	37	4	
5	Purpose, Focus, & Organization	60	38	2	9,661
	Evidence & Elaboration	59	39	2	
	Conventions	62	37	1	
6	Purpose, Focus, & Organization	63	34	2	10,360
	Evidence & Elaboration	63	35	2	
	Conventions	63	33	4	
7	Purpose, Focus, & Organization	62	37	2	10,488
	Evidence & Elaboration	61	37	2	
	Conventions	67	32	2	
8	Purpose, Focus, & Organization	61	38	2	10,660
	Evidence & Elaboration	61	38	2	
	Conventions	73	25	2	

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

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

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

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

where

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

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

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

Weighted kappa coefficients for operational writing prompts by dimension are presented in Table 25.

Table 25: Weighted Kappa Coefficients

Grade	<i>N</i>	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
3	6,408	0.656	0.649	0.400
4	7,916	0.678	0.683	0.365
5	8,622	0.709	0.702	0.398
6	9,018	0.674	0.670	0.371
7	9,819	0.675	0.667	0.380
8	10,331	0.678	0.683	0.402

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 English/Language Arts (ELA), there are three reporting categories per grade: Key Ideas and Textual Support/Vocabulary, Structural Elements and Organization/Connection of Ideas/Media Literacy, and Writing. The reporting categories in Mathematics, Science, and Social Studies differed in each grade or course (see Table 2 through Table 5 for reporting category information).

The scale scores and relative strengths and weaknesses from each reporting category were provided to students. Evidence is needed to verify that the 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 Item Response Theory (IRT) model difficult. However, we could then easily justify reporting these separate scores. On the contrary, if the reporting categories were perfectly correlated, we could justify using a unidimensional model, but we could not justify reporting separate scores.

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 and reporting subscores.

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. The observed correlations and disattenuated correlations are provided in Section 5.1, Correlations among Reporting Category Scores.

5.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 26 through Table 29 present the observed correlation matrix of the reporting category scores for each subject area. In ELA, the correlations among the reporting categories ranged from 0.59 to 0.71. In Mathematics, the correlations were between 0.64 and 0.81. In Science, the correlations among reporting categories ranged from 0.58 to 0.73. In Social Studies, the correlations ranged from 0.65 to 0.74.

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 avoided cautiously.

Table 30 through Table 33 display disattenuated correlations. The overall average disattenuated correlation was 0.91 for ELA, 0.94 for Mathematics, 0.95 for Science, and 1.00* for Social Studies. These values suggest that validity evidence of internal structure is supported.

Table 26: Observed Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	12–15	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.69	1	
	Writing (Cat3)	6–8	0.63	0.59	1
4	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.64	1	
	Writing (Cat3)	7–8	0.68	0.63	1
5	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.62	1	
	Writing (Cat3)	6–8	0.68	0.59	1
6	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.65	1	
	Writing (Cat3)	7–8	0.64	0.63	1
7	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.65	1	
	Writing (Cat3)	7–8	0.69	0.62	1
8	Key Ideas and Textual Support/Vocabulary (Cat1)	10–12	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.65	1	
	Writing (Cat3)	6–8	0.71	0.61	1

Table 27: Observed Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	9–11	1			
	Computation (Cat2)	11–13	0.81	1		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Geometry and Measurement (Cat3)	9–11	0.79	0.77	1	
	Number Sense (Cat4)	11–13	0.79	0.77	0.78	1
4	Algebraic Thinking and Data Analysis (Cat1)	9–11	1			
	Computation (Cat2)	11–13	0.78	1		
	Geometry and Measurement (Cat3)	9–11	0.77	0.75	1	
	Number Sense (Cat4)	11–13	0.77	0.77	0.75	1
5	Algebraic Thinking (Cat1)	10–12	1			
	Computation (Cat2)	11–13	0.79	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9–11	0.76	0.76	1	
	Number Sense (Cat4)	11–13	0.76	0.75	0.73	1
6	Algebra and Functions (Cat1)	11–13	1			
	Computation (Cat2)	10–12	0.74	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9–11	0.75	0.67	1	
	Number Sense (Cat4)	10–12	0.80	0.72	0.73	1
7	Algebra and Functions (Cat1)	11–12	1			
	Data Analysis, Statistics, and Probability (Cat2)	9–11	0.74	1		
	Geometry and Measurement (Cat3)	9–11	0.66	0.64	1	
	Number Sense and Computation (Cat4)	12–13	0.78	0.75	0.67	1
8	Algebra and Functions (Cat1)	11–13	1			
	Data Analysis, Statistics, and Probability (Cat2)	10–12	0.75	1		
	Geometry and Measurement (Cat3)	10–12	0.73	0.72	1	
	Number Sense and Computation (Cat4)	9–11	0.69	0.67	0.68	1

Table 28: Observed Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Questioning and Modeling (Cat1)	10–12	1				
	Investigating (Cat2)	10–12	0.68	1			
	Analyzing, Interpreting, and Computational Thinking (Cat3)	12–14	0.69	0.69	1		
	Explaining Solutions, Reasoning, and Communicating (Cat4)	12–14	0.71	0.70	0.72	1	
6	Questioning and Modeling (Cat1)	12–14	1				
	Investigating (Cat2)	12–14	0.69	1			

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Analyzing, Interpreting, and Computational Thinking (Cat3)	10–12	0.68	0.70	1		
	Explaining Solutions, Reasoning, and Communicating (Cat4)	10–12	0.70	0.72	0.71	1	
Biology (Fall)	Developing and Using Models to Describe Structure and Function (Cat1)	10–12	1				
	Developing and Using Models to Explain Processes (Cat2)	10–12	0.64	1			
	Analyzing Data and Mathematical Thinking (Cat3)	10–12	0.68	0.68	1		
	Constructing and Communicating an Explanation (Cat4)	10–12	0.65	0.64	0.71	1	
	Evaluating Claims with Evidence (Cat5)	10–12	0.63	0.63	0.66	0.66	1
Biology (Winter)	Developing and Using Models to Describe Structure and Function (Cat1)	10–12	1				
	Developing and Using Models to Explain Processes (Cat2)	10–12	0.58	1			
	Analyzing Data and Mathematical Thinking (Cat3)	10–12	0.67	0.63	1		
	Constructing and Communicating an Explanation (Cat4)	10–12	0.68	0.61	0.71	1	
	Evaluating Claims with Evidence (Cat5)	10–12	0.60	0.58	0.68	0.68	1
Biology (Spring)	Developing and Using Models to Describe Structure and Function (Cat1)	10–12	1				
	Developing and Using Models to Explain Processes (Cat2)	10–12	0.66	1			
	Analyzing Data and Mathematical Thinking (Cat3)	10–12	0.71	0.68	1		
	Constructing and Communicating an Explanation (Cat4)	10–12	0.70	0.66	0.73	1	
	Evaluating Claims with Evidence (Cat5)	10–12	0.64	0.63	0.70	0.67	1

Table 29: Observed Correlation Matrix Among Reporting Categories (Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	17	1		
	Geography and Economics (Cat2)	11	0.67	1	
	History (Cat3)	12	0.71	0.65	1
U.S. Government	Functions of Government (Cat1)	20	1		
	Historical Foundations of American Government (Cat2)	14	0.70	1	
	Institutions and Processes of Government (Cat3)	20	0.74	0.69	1

Table 30: Disattenuated Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	12-15	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-12	0.98	1	
	Writing (Cat3)	6-8	0.89	0.85	1
4	Key Ideas and Textual Support/Vocabulary (Cat1)	11-14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11-14	0.92	1	
	Writing (Cat3)	7–8	0.92	0.88	1
5	Key Ideas and Textual Support/Vocabulary (Cat1)	11–14	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11–14	0.94	1	
	Writing (Cat3)	6–8	0.91	0.86	1
6	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.99	1	
	Writing (Cat3)	7–8	0.89	0.90	1
7	Key Ideas and Textual Support/Vocabulary (Cat1)	10–13	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–13	0.95	1	
	Writing (Cat3)	7–8	0.92	0.89	1
8	Key Ideas and Textual Support/Vocabulary (Cat1)	10–12	1		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–12	0.93	1	
	Writing (Cat3)	6–8	0.92	0.88	1

Table 31: Disattenuated Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	9–11	1			
	Computation (Cat2)	11–13	1.00*	1		
	Geometry and Measurement (Cat3)	9–11	0.97	0.97	1	
	Number Sense (Cat4)	11–13	0.95	0.95	0.95	1
4	Algebraic Thinking and Data Analysis (Cat1)	9–11	1			
	Computation (Cat2)	11–13	0.96	1		
	Geometry and Measurement (Cat3)	9–11	0.96	0.93	1	
	Number Sense (Cat4)	11–13	0.96	0.95	0.94	1
5	Algebraic Thinking (Cat1)	10–12	1			

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

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

Table 32: Dissattenuated Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Questioning and Modeling (Cat1)	10–12	1				
	Investigating (Cat2)	10–12	0.98	1			
	Analyzing, Interpreting, and Computational Thinking (Cat3)	12–14	0.99	0.99	1		
	Explaining Solutions, Reasoning, and Communicating (Cat4)	12–14	1.00*	0.99	1.00*	1	
6	Questioning and Modeling (Cat1)	12–14	1				
	Investigating (Cat2)	12–14	0.99	1			
	Analyzing, Interpreting, and Computational Thinking (Cat3)	10–12	0.98	0.98	1		
	Explaining Solutions, Reasoning, and Communicating (Cat4)	10–12	0.98	0.99	0.98	1	
Biology (Fall)	Developing and Using Models to Describe Structure and Function (Cat1)	10–12	1				
	Developing and Using Models to Explain Processes (Cat2)	10–12	0.96	1			
	Analyzing Data and Mathematical Thinking (Cat3)	10–12	0.96	0.97	1		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Constructing and Communicating an Explanation (Cat4)	10–12	0.94	0.93	0.98	1	
	Evaluating Claims with Evidence (Cat5)	10–12	0.95	0.95	0.95	0.97	1
Biology (Winter)	Developing and Using Models to Describe Structure and Function (Cat1)	10–12	1				
	Developing and Using Models to Explain Processes (Cat2)	10–12	0.91	1			
	Analyzing Data and Mathematical Thinking (Cat3)	10–12	0.96	0.93	1		
	Constructing and Communicating an Explanation (Cat4)	10–12	0.98	0.91	0.97	1	
	Evaluating Claims with Evidence (Cat5)	10–12	0.90	0.91	0.96	0.96	1
Biology (Spring)	Developing and Using Models to Describe Structure and Function (Cat1)	10–12	1				
	Developing and Using Models to Explain Processes (Cat2)	10–12	0.92	1			
	Analyzing Data and Mathematical Thinking (Cat3)	10–12	0.94	0.92	1		
	Constructing and Communicating an Explanation (Cat4)	10–12	0.95	0.92	0.96	1	
	Evaluating Claims with Evidence (Cat5)	10–12	0.92	0.93	0.96	0.95	1

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

Table 33: Disattenuated Correlation Matrix Among Reporting Categories (Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	17	1		
	Geography and Economics (Cat2)	11	1.00*	1	
	History (Cat3)	12	1.00*	1.00*	1
U.S. Government	Functions of Government (Cat1)	20	1		
	Historical Foundations of American Government (Cat2)	14	1.00*	1	
	Institutions and Processes of Government (Cat3)	20	1.00*	1.00*	1

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

5.2 CONFIRMATORY FACTOR ANALYSIS

The Indiana Learning Evaluation Assessment Readiness Network (*ILEARN*) test items were designed to measure different standards and higher-level reporting categories. Test scores were reported as an overall performance measure. Additionally, scores on the various reporting categories were also provided as indices of strand-specific

performance. The strand scores were reported in a fashion that aligned with the theoretical structure of the test derived from the test blueprint.

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 (CFA), in which the first-order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto factors they are intended to measure. The underlying structure of the *ILEARN* assessments was common across all grades, which is useful for comparing the results of our analyses across the grades.

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

The 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.

The results in this section were based on the data collected from the initial administration of the *ILEARN* assessments, which was the Spring 2019 administration. The purpose is to provide validity evidence regarding the dimensionality of the assessments. Given there is no major change in test design, this analysis does not need to be conducted in subsequent test administrations.

5.2.1 Factor Analytic Methods

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

The factor analysis models are based on the matrix \mathbf{S} of tetrachoric and polychoric sample correlations among the item scores (Olsson, 1979), and the matrix \mathbf{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:

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

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

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

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

where $\mathbf{\Lambda}$ is the matrix of item factor loadings (with $\mathbf{\Lambda}'$ representing its transpose), and $\mathbf{\Theta}$ is the uniqueness, or measurement error. The matrix $\mathbf{\Phi}$ is the correlation among the separate factors. For the base model, items are thought only to load onto a single underlying factor. Hence $\mathbf{\Lambda}'$ is a $p \times 1$ vector, where p is the number of test items and $\mathbf{\Phi}$ is a scalar equal to 1. Therefore, it is possible to drop the matrix $\mathbf{\Phi}$ 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

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

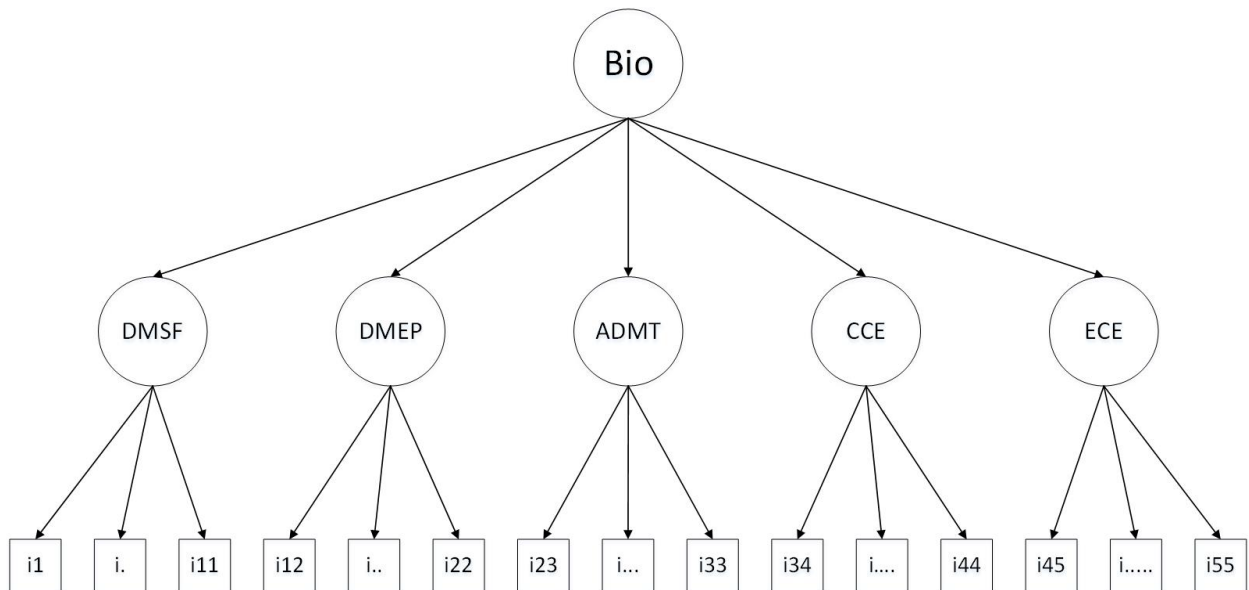
where $\hat{\Sigma}$ is the implied correlation matrix among test items, Λ is the $p \times k$ matrix of first-order factor loadings relating item scores to first-order factors, Γ is the $k \times 1$ matrix of second-order factor loadings relating the first-order factors to the second-order factor with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

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

The second-order factor model can also be represented graphically. A sample of the generalized approaches is provided below in Figure 6. This sample illustrates the general structure of the second-order factor analysis for Biology, and 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 CFA 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 6: Second-Order Factor Model (Biology)



5.2.2 Results

Several goodness-of-fit statistics from each of the analyses are presented in Table 34, which shows the summary results obtained from CFA. Three goodness-of-fit indices were used to evaluate model fit of the item parameters to the manner in which students actually responded to the items. The root mean square error of approximation (RMSEA) is referred to as a badness-of-fit index so that a value closer to 0 implies better fit and a value of 0 implies best fit. In general, RMSEA below 0.05 is considered as good fit and RMSEA over 0.1 suggests poor fit (Browne & Cudeck, 1993). The Tucker-Lewis index (TLI) and the comparative fit index (CFI) are incremental goodness-of-fit indices. These indices compare the implied model to the baseline model where no observed variables are correlated (i.e., there are no factors). Values greater than 0.9 are recognized as acceptable, and values over 0.95 are considered as good fit (Hu & Bentler, 1999). As Hu and Bentler (1999) suggest, the selected cutoff values of the fit index should not be overgeneralized and should be interpreted with caution.

Based on the fit indices, the model showed good fit across content domains. For all tests, the RMSEA was below 0.05, and the CFI and TLI were equal to or greater than 0.95.

Table 34: Goodness-of-Fit Second-Order CFA

ELA					
Grade	df	RMSEA	CFI	TLI	Convergence
3	524	0.014	0.983	0.981	Yes
4	557	0.014	0.983	0.982	Yes
5	591	0.009	0.984	0.983	Yes
6	492	0.014	0.984	0.983	Yes
7	460	0.012	0.982	0.981	Yes
8	557	0.010	0.985	0.984	Yes
Mathematics					
Grade	df	RMSEA	CFI	TLI	Convergence
3	1076	0.017	0.983	0.982	Yes
4	1076	0.014	0.958	0.955	Yes
5	1076	0.015	0.977	0.976	Yes
6	1075	0.019	0.942	0.939	Yes
7	1075	0.013	0.983	0.982	Yes
8	1075	0.025	0.916	0.912	Yes
Science					
Grade	df	RMSEA	CFI	TLI	Convergence
4	1032	0.019	0.975	0.974	Yes
6	1031	0.019	0.981	0.98	Yes
Biology (Spring)	1321	0.021	0.975	0.974	Yes
Social Studies					

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

In Table 35 through Table 38, we provide the estimated correlations between the reporting categories from the second-order factor model for ELA, Mathematics, Science, and Social Studies, respectively. In all cases, these correlations are very high. However, the results provide empirical evidence that there is some detectable dimensionality among reporting categories.

Table 35: Correlations Among Factors (ELA)

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

Table 36: Correlations Among Factors (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	9	1			
	Computation (Cat2)	13	0.989	1		
	Geometry and Measurement (Cat3)	10	0.969	0.959	1	
	Number Sense (Cat4)	11	0.908	0.898	0.880	1
4	Algebraic Thinking and Data Analysis (Cat1)	9	1			
	Computation (Cat2)	12	0.963	1		
	Geometry and Measurement (Cat3)	10	0.929	0.894	1	
	Number Sense (Cat4)	12	0.934	0.900	0.868	1
5	Algebraic Thinking (Cat1)	11	1			
	Computation (Cat2)	11	0.888	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.890	0.790	1	
	Number Sense (Cat4)	11	0.926	0.823	0.825	1
6	Algebra and Functions (Cat1)	11	1			
	Computation (Cat2)	11	0.820	1		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.763	0.645	1	
	Number Sense (Cat4)	11	0.973	0.823	0.766	1
7	Algebra and Functions (Cat1)	11	1			
	Data Analysis, Statistics, and Probability (Cat2)	10	0.865	1		
	Geometry and Measurement (Cat3)	10	0.891	0.859	1	
	Number Sense and Computation (Cat4)	11	0.912	0.880	0.906	1
8	Algebra and Functions (Cat1)	11	1			
	Data Analysis, Statistics, and Probability (Cat2)	10	0.748	1		
	Geometry and Measurement (Cat3)	12	0.821	0.712	1	
	Number Sense and Computation (Cat4)	10	0.815	0.707	0.775	1

Table 37: Correlations Among Factors (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Questioning and Modeling (Cat1)	12	1				
	Investigating (Cat2)	12	0.990	1			
	Analyzing, Interpreting, and Computational Thinking (Cat3)	12	0.990	1	1		
	Explaining Solutions, Reasoning, and Communicating (Cat4)	11	0.987	0.997	0.997	1	

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
6	Questioning and Modeling (Cat1)	11	1				
	Investigating (Cat2)	11	0.994	1			
	Analyzing, Interpreting, and Computational Thinking (Cat3)	12	0.988	0.983	1		
	Explaining Solutions, Reasoning, and Communicating (Cat4)	13	0.995	0.989	0.984	1	
Biology (Spring)	Developing and Using Models to Describe Structure and Function (Cat1)	10	1				
	Developing and Using Models to Explain Processes (Cat2)	10	0.934	1			
	Analyzing Data and Mathematical Thinking (Cat3)	11	0.966	0.940	1		
	Constructing and Communicating an Explanation (Cat4)	11	0.980	0.953	0.986	1	
	Evaluating Claims with Evidence (Cat5)	11	0.971	0.945	0.977	0.991	1

Table 38: Correlations Among Factors (Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	16	1		
	Geography and Economics (Cat2)	11	0.982	1	
	History (Cat3)	12	0.947	0.950	1
U.S. Government	Functions of Government (Cat1)	19	1		
	Historical Foundations of American Government (Cat2)	14	0.962	1	
	Institutions and Processes of Government (Cat3)	20	0.957	0.971	1

5.2.3 Discussion

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

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

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

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 and $T_i(\hat{\theta}_j)$ 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 39 through Table 42 present summaries of the distributions of Q_3 statistics: minimum, 5th percentile, median, 95th percentile, and maximum values from each ILEARN subject. The results show that a very small percentage of ILEARN items were greater than a critical value of 0.2 for $|Q_3|$ (Chen & Thissen, 1997).

Table 39: Q₃ Statistic (ELA)

Grade	Q ₃ Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	-0.215	-0.091	-0.022	0.046	0.228
4	-0.236	-0.099	-0.023	0.048	0.235
5	-0.229	-0.100	-0.024	0.050	0.267
6	-0.281	-0.102	-0.022	0.049	0.254
7	-0.241	-0.097	-0.023	0.044	0.282
8	-0.304	-0.097	-0.023	0.052	0.215

Table 40: Q₃ Statistic (Mathematics)

Grade	Q ₃ Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	-0.244	-0.097	-0.022	0.061	0.900
4	-0.314	-0.093	-0.021	0.058	0.864
5	-0.275	-0.092	-0.020	0.061	0.753
6	-0.348	-0.101	-0.022	0.065	0.573
7	-0.282	-0.095	-0.020	0.059	0.774
8	-0.270	-0.092	-0.019	0.061	0.785

Table 41: Q₃ Statistic (Science)

Grade	Q ₃ Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
4	-0.279	-0.075	-0.020	0.035	0.485
6	-0.322	-0.070	-0.019	0.038	0.435
Biology (Spring)	-0.579	-0.075	-0.008	0.081	0.585
Biology (Fall)	-0.643	-0.160	-0.014	0.129	0.475
Biology (Winter)	-0.322	-0.117	-0.014	0.095	0.447

Table 42: Q₃ Statistic (Social Studies)

Grade	Q ₃ Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
5	-0.101	-0.053	-0.023	0.010	0.082
U.S. Government	-0.296	-0.136	-0.019	0.104	0.290

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.

Observed and disattenuated subscore correlations were calculated both within subjects and across subjects for grades 3–8 ELA and Mathematics. In grades 4 and 6, Science was included and in grade 5, Social Studies was included. Table 43 through Table 54 show the observed and disattenuated score correlations among ELA, Mathematics, Science, and Social Studies subscores for grades 3–8, where students took included subjects. In general, the pattern is consistent with the *a priori* expectation that subscores within a test correlate more highly than correlations between tests measuring a different construct.

Table 43: Grade 3 Observed Score Correlations

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

Table 44: Grade 3 Disattenuated Score Correlations

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

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

Table 45: Grade 4 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.64	1									
	Writing (Cat3)	0.68	0.63	1								
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.65	0.61	0.68	1							
	Computation (Cat2)	0.60	0.57	0.64	0.78	1						
	Geometry and Measurement (Cat3)	0.60	0.58	0.64	0.77	0.75	1					
	Number Sense (Cat4)	0.59	0.58	0.63	0.77	0.77	0.75	1				
Science	Questioning and Modeling (Cat1)	0.64	0.61	0.63	0.66	0.62	0.64	0.62	1			
	Investigating (Cat2)	0.63	0.60	0.62	0.67	0.62	0.65	0.63	0.68	1		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.65	0.62	0.64	0.68	0.63	0.64	0.63	0.69	0.69	1	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.67	0.64	0.66	0.69	0.65	0.66	0.65	0.71	0.70	0.72	1

Table 46: Grade 4 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.92	1									
	Writing (Cat3)	0.92	0.89	1								
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.84	0.83	0.87	1							
	Computation (Cat2)	0.78	0.77	0.81	0.96	1						
	Geometry and Measurement (Cat3)	0.79	0.79	0.82	0.96	0.93	1					
	Number Sense (Cat4)	0.78	0.79	0.81	0.96	0.95	0.94	1				
Science	Questioning and Modeling (Cat1)	0.90	0.89	0.87	0.88	0.82	0.86	0.84	1			
	Investigating (Cat2)	0.89	0.88	0.86	0.90	0.83	0.87	0.85	0.98	1		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.91	0.91	0.88	0.90	0.83	0.86	0.85	0.99	0.99	1	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.93	0.93	0.90	0.91	0.85	0.88	0.85	1.00	0.99	1.00*	1

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

Table 47: Grade 5 Observed Score Correlations

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

Table 48: Grade 5 Disattenuated Score Correlations

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

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

Table 49: Grade 6 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.65	1									
	Writing (Cat3)	0.63	0.63	1								
Mathematics	Algebra and Functions (Cat1)	0.62	0.62	0.66	1							
	Computation (Cat2)	0.54	0.54	0.59	0.74	1						
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.58	0.57	0.61	0.75	0.67	1					
	Number Sense (Cat4)	0.62	0.61	0.64	0.80	0.72	0.73	1				
Science	Questioning and Modeling (Cat1)	0.60	0.61	0.59	0.66	0.57	0.61	0.65	1			
	Investigating (Cat2)	0.64	0.63	0.62	0.69	0.60	0.65	0.69	0.69	1		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.62	0.62	0.60	0.68	0.60	0.63	0.68	0.68	0.70	1	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.64	0.64	0.62	0.68	0.59	0.65	0.68	0.70	0.72	0.71	1

Table 50: Grade 6 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.99	1									
	Writing (Cat3)	0.89	0.90	1								
Mathematics	Algebra and Functions (Cat1)	0.84	0.86	0.84	1							
	Computation (Cat2)	0.74	0.76	0.76	0.92	1						
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.82	0.83	0.82	0.96	0.87	1					
	Number Sense (Cat4)	0.84	0.86	0.82	0.98	0.91	0.96	1				
Science	Questioning and Modeling (Cat1)	0.89	0.92	0.82	0.88	0.78	0.87	0.89	1			
	Investigating (Cat2)	0.92	0.94	0.85	0.90	0.79	0.90	0.92	0.99	1		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.90	0.92	0.82	0.88	0.79	0.87	0.90	0.99	0.98	1	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.91	0.93	0.83	0.87	0.77	0.87	0.89	0.99	0.99	0.98	1

Table 51: Grade 7 Observed Score Correlations

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

Table 52: Grade 7 Disattenuated Score Correlations

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

Table 53: Grade 8 Observed Score Correlations

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

Table 54: Grade 8 Disattenuated Score Correlations

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

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 and provides 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
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 insufficient for determining the fairness of a test. Rather, it must be accompanied by statistical processes. While a variety of item statistics were reviewed during form building to evaluate the quality of items, one notable statistic used was differential item functioning (DIF). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe evidence of DIF, according to the DIF classification convention illustrated in Volume 1 of this technical report. Furthermore, items were categorized positively (i.e., +A, +B, +C), signifying that the item favored the focal group (e.g., African American/Black, Hispanic, Female), or negatively (i.e., –A, –B, –C), signifying that the item favored the reference group (e.g., White, Male). Items across all groups were flagged if their DIF statistics indicated the “C” category. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Items were reviewed by the Bias and Sensitivity Committee regardless of whether the DIF statistic favored the focal group or the reference group. The details about how these items were reviewed for bias is further described in 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. These 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
- English Learners (ELs)/Not ELs

A detailed description of the DIF analysis performed is presented in Volume 1, Section 4.2, of the *ILEARN 2021–2022 Annual Technical Report*. The DIF statistics for each operational test item are presented in the Appendix A, Operational Item Statistics, of Volume 1 of the *2021–2022 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*.
- Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin*, 87(3), 513–524. <https://doi.org/10.1037/0033-2909.87.3.513>
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3), 296–322. <https://www.gwern.net/docs/statistics/1910-brown.pdf>
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1), 62–83. <https://doi.org/10.1111/j.2044-8317.1984.tb00789.x>
- 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). Sage.
- 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(4), 213–220. <https://doi.org/10.1037/h0026256>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. http://cda.psych.uiuc.edu/psychometrika_highly_cited_articles/cronbach_1951.pdf
- Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.
- Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, 9(3), 277–286. https://doi.org/10.1207/s15324818ame0903_5
- Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, 11(6), 1–9. <https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1192&context=pare>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. http://expsylab.psych.uoa.gr/fileadmin/expsylab.psych.uoa.gr/uploads/papers/Hu_Bentler_1999.pdf

- 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, W.-C., Hanson, B. A., & Brennan, R. L. (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*. Lawrence Erlbaum Associates.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132. <https://doi.org/10.1007/BF02294210>
- Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*, 7th Edition.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460. <https://doi.org/10.1007/BF02296207>
- Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, 8(2), 111–120. https://doi.org/10.1207/s15324818ame0802_1
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14). <https://doi.org/10.7275/an9m-2035>
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295. <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes. <https://nceo.info/Resources/publications/onlinepubs/Synthesis44.html>
- 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(2), 125–145. <https://doi.org/10.1177/014662168400800201>

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>



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

2021–2022

**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.

Major contributors to this technical report include the following staff from American Institutes for Research (AIR): Stephan Ahadi, Elizabeth Ayers-Wright, Xiaoxin Wei, Grace Chung, Kevin Clayton, and Aleah Pepper. 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	4
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 Trend Reports.....	11
1.6.4 Aggregate-Level Reporting Category Report.....	16
1.6.5 Aggregate-Level Standards Report.....	21
1.6.6 Student-Level Subject Report	26
1.6.7 Student-Level Reporting Category Report	30
1.6.8 Individual Student Report.....	35
1.6.9 Interpretive Guide.....	42
1.6.10 Reports by Sub-Group	43
1.6.11 Data File.....	45
2. INTERPRETATION OF REPORTED SCORES	46
2.1 Appropriate Uses for Scores and Reports	46
2.2 Scale Score	47
2.3 Performance Level.....	48
2.4 Performance Category for Reporting Categories.....	48
2.5 Cut Scores.....	48
2.6 Aggregated Scores	50
2.7 Writing Performance	50
2.8 Relative Strength and Weakness	51
2.9 Lexile® Measure.....	51
2.10 Quantile® Measure.....	51
3. SUMMARY	52

LIST OF APPENDICES

Appendix A: Data File Layout

LIST OF TABLES

Table 1: Reporting Categories for ELA	3
Table 2: Reporting Categories for Mathematics	4
Table 3: Reporting Categories for Science	4
Table 4: Reporting Categories for Social Studies	4
Table 5: Indiana Score Reports Summary	5
Table 6: Indiana List of Sub-Groups	7
Table 7: ILEARN ELA Assessment Proficiency Cut Scores	49
Table 8: ILEARN Mathematics Assessment Proficiency Cut Scores	49
Table 9: ILEARN Science Assessment Proficiency Cut Scores	49
Table 10: ILEARN Social Studies Grade 5 Assessment Proficiency Cut Scores.....	49
Table 11: ILEARN U.S. Government Assessment Proficiency Cut Scores	50
Table 12: Writing Scoring Dimensions.....	50

LIST OF FIGURES

Figure 1: Sample State Summary Performance Report	9
Figure 2: Corporation-Level Summary Performance Report.....	10
Figure 3: Corporation Aggregate-Level Subject Report, Grade 8 ELA	12
Figure 4: Corporation Aggregate-Level Subject Report, Grade 8 Mathematics.....	13
Figure 5: Corporation Aggregate-Level Subject Report, Grade 6 Science	14
Figure 6: Corporation Aggregate-Level Subject Report, Grade 5 Social Studies	15
Figure 7: Corporation Aggregate-Level Reporting Category Report, Grade 8 ELA ...	17
Figure 8: Corporation Aggregate-Level Reporting Category Report, Grade 8 Mathematics	18
Figure 9: Corporation Aggregate-Level Reporting Category Report, Grade 6 Science 19	
Figure 10: Corporation Aggregate-Level Reporting Category Report, Grade 5 Social Studies.....	20
Figure 11: Sample District Aggregate-Level Standards Report, Grade 8 ELA.....	22
Figure 12: Sample District Aggregate-Level Standards Report, Grade 8 Mathematics 24	
Figure 13: Student-Level Subject Report, Grade 8 ELA	26
Figure 14: Student-Level Subject Report, Grade 8 Mathematics.....	27
Figure 15: Student-Level Subject Report, Grade 6 Science	28
Figure 16: Student-Level Subject Report, Grade 5 Social Studies	29
Figure 17: Student-Level Reporting Category Report, Grade 8 ELA	31

Figure 18: Student-Level Reporting Category Report, Grade 8 Mathematics..... 32

Figure 19: Student-Level Reporting Category Report, Grade 6 Science 33

Figure 20: Student-Level Reporting Category Report, Grade 5 Social Studies 34

Figure 21: Individual Student Report, Grade 8 ELA..... 36

Figure 22: Individual Student Report, Grade 8 Mathematics 38

Figure 23: Individual Student Report, Grade 6 Science..... 40

Figure 24: Individual Student Report, Grade 5 Social Studies..... 41

Figure 25: Supplemental Interpretive Guide 42

Figure 26: Corporation Aggregate-Level Subject Report by Gender, Grade 8 ELA... 43

Figure 27: Corporation Aggregate-Level Reporting Category Report by Section 504
Plan Status, Grade 8 Mathematics 44

Figure 28: Data File 45

1. INDIANA SCORE REPORTS

During school year 2021-2022, pursuant to IC 20-32-5, ILEARN assessments were administered to Indiana students 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 2021-2022 assessments, the features of the score reports, and the appropriate uses and inferences that can be drawn from those score types.

1.1 OVERVIEW OF INDIANA’S SCORE REPORTS

ILEARN assessments were administered during the 2021-2022 school year. Test scores from each assessment 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 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. 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 they attempted. For online tests, the student must attempt at least five or more items but less than 32 items on the test to get Undetermined for both overall score and reporting category scores. For paper tests, the student must attempt at least five or more items but less than 32 items on the test to get Undetermined for reporting category scores. 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.in.gov/doe/students/assessment/ilearn/>).

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; and
- Level 4: Above Proficiency.

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

- Level 1: Below Proficiency; and
- 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 and additional details can be found in Volume 6 of this report. 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 subscores) represent distinct groups of knowledge within each grade subject. For *ILEARN*, students’ performance on each reporting category is reported using three performance categories:

- Below;
- At/Near; and
- 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 Computation Geometry and Measurement, Data Analysis, and Statistics Number Sense
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

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, corporations, 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 corporations).

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/-/media/project/client-portals/indiana-ilearn/pdf/sy21-22-documents/ors_guide_spring-2021-2022.pdf

Table 5: Indiana Score Reports Summary

Report	Description	Level of Availability					
		State	Corporation	School	Teacher	Roster	Student/Parent
Summary Performance	Summary of performance (to date) across grades and subjects or courses for the current administration	✓	✓	✓	✓	✓	
Aggregate-Level Subject Report	Summary of overall performance for a subject and a grade for all students in the defined level of aggregation	✓	✓	✓	✓	✓	

Report	Description	Level of Availability					
		State	Corporation	School	Teacher	Roster	Student/ Parent
Aggregate-Level Reporting Category Report	Summary of overall performance on each reporting category for a given subject and grade across all students within the selected level of aggregation	✓	✓	✓	✓	✓	
Aggregate-Level Standards Report	Presents data on the performance of aggregate entities on each standard of a subject for the current window. (Only available for adaptive ILEARN assessments)		✓	✓			
Student-Level Subject Report	List of all students who belong to a school, teacher, or roster with their associated subject or course scores for the current administration			✓	✓	✓	
Student-Level Reporting Category Report	List of all students who belong to a school, teacher, or roster with their associated reporting category performance for the current administration			✓	✓	✓	
Individual Student Report (ISR)	Detailed information about a selected student's performance in a specified subject or course; includes overall subject and reporting category results						✓
Data Files	Text/CSV files containing overall and reporting category scale scores and performance levels along with demographic information		✓	✓	✓	✓	

1.5 REPORTING BY SUB-GROUP

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

Subgroup	Subgroup 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
Special Education	Special Education
	Not Special Education
Section 504 Plan	Section 504 Plan
	Not Section 504 Plan
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
	Grade 13

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 see 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

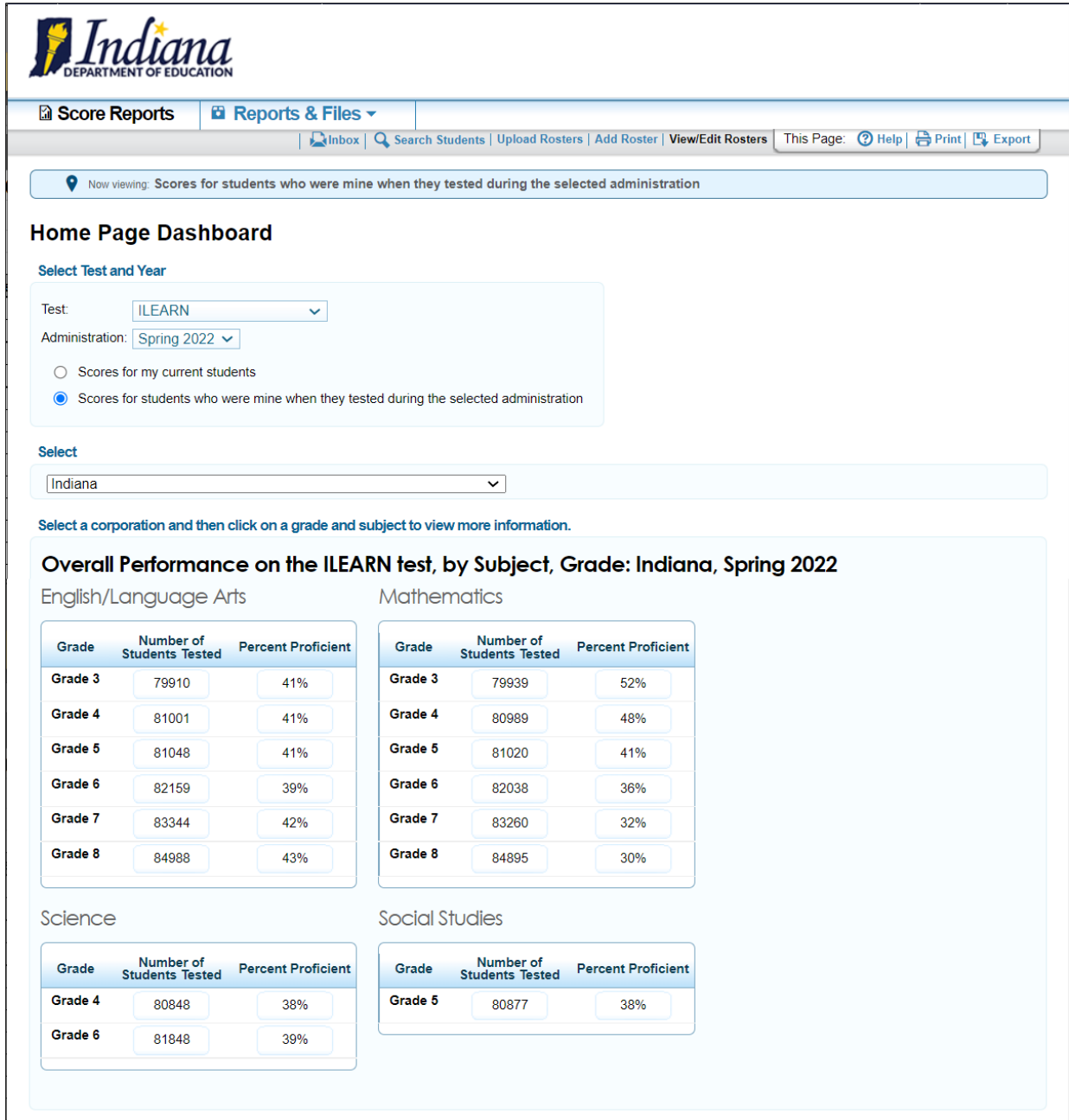
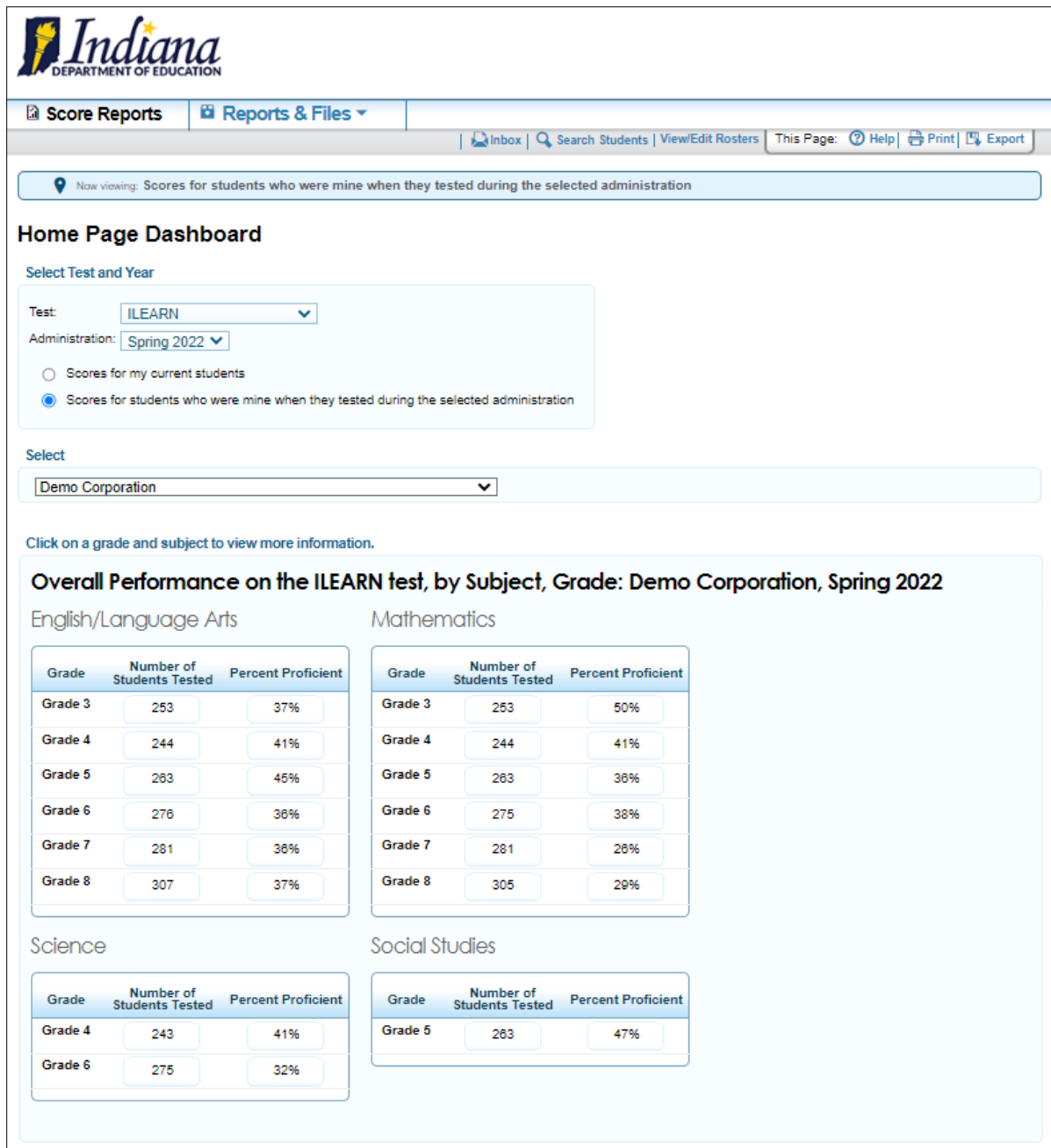


Figure 2: Corporation-Level Summary Performance Report



The Corporation Summary Report is similar to the State Summary Report, except that summary data for the Corporation Summary Report 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 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 and standard error of the average scale score;
- Percentage proficient;
- Number of students in each performance level; and
- Percentage of students in each performance level.

The summaries are also presented for overall students and by sub-groups. Figure 3 presents an example of Aggregate-Level Subject Reports for grade 8 ELA at the corporation level without sub-groups. Figure 4 highlights grade 8 Mathematics at the corporation level when a user selects a sub-group of gender. Figure 5 and 6 presents Science and Social Studies subject reports at the corporation level.

1.6.3 Trend Reports

Trend reports are available only on the ILEARN ELA and Mathematics assessments. Trend reports display the overall performance of a student or group of students in the selected subject over time. For each testing window, the report displays either the average scale score and associated standard error or the percentage of proficient students. Scores from previous years represent either a group's average score or a student's individual score from that year's testing window. All ELA and Math CAT and Performance Task or Fixed Form and Performance Task taken within the school year will appear on the individual student trend report.

The trend report shows the performance progress for the entity or individual being analyzed. The graph plots the data points for the selected groups of students or individual students at each point in time (across test administrations and school years). Trend reports are interactive, allowing users to specify which data to plot on the graph.

Figure 3: Corporation Aggregate-Level Subject Report, Grade 8 ELA

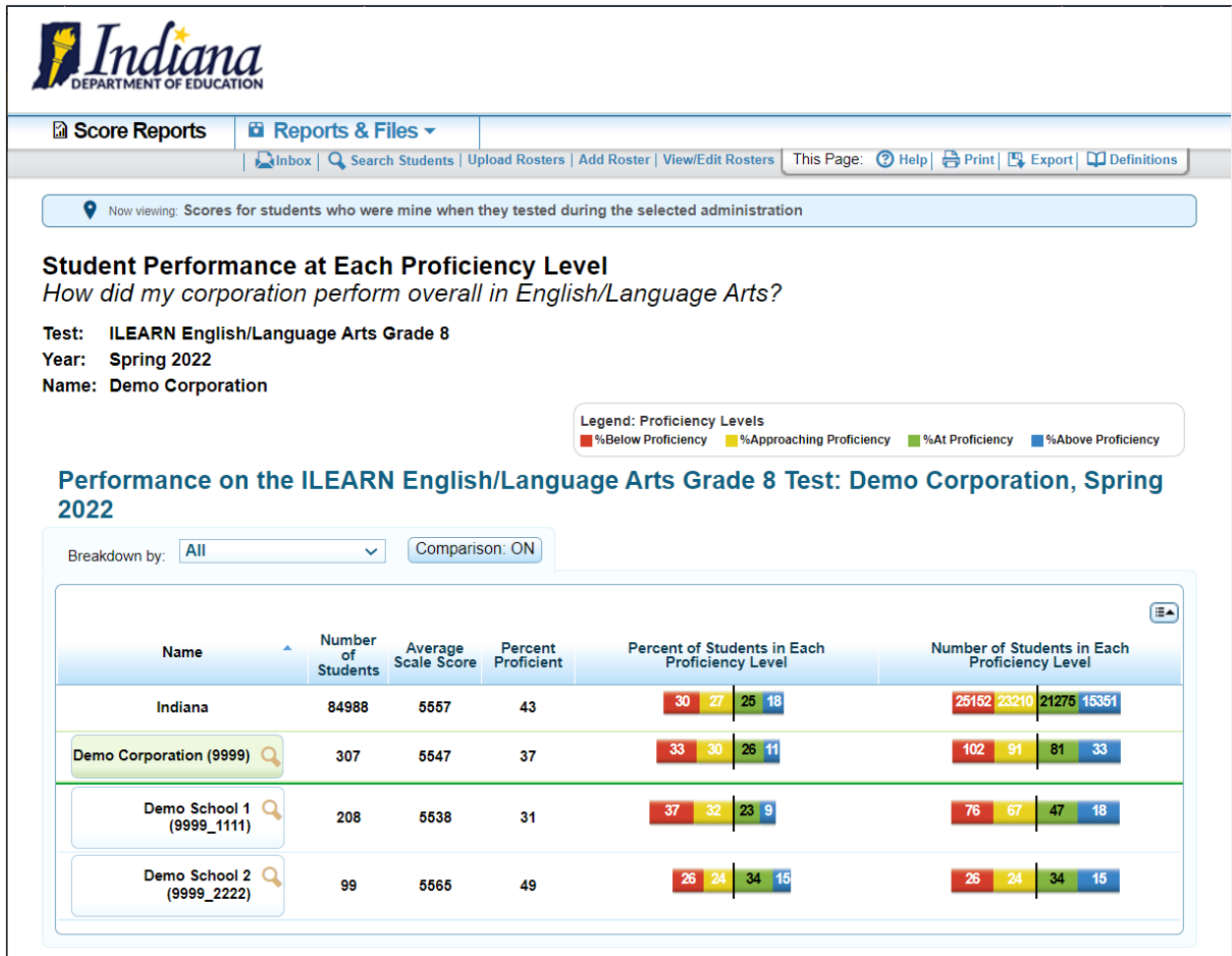


Figure 4: Corporation Aggregate-Level Subject Report, Grade 8 Mathematics

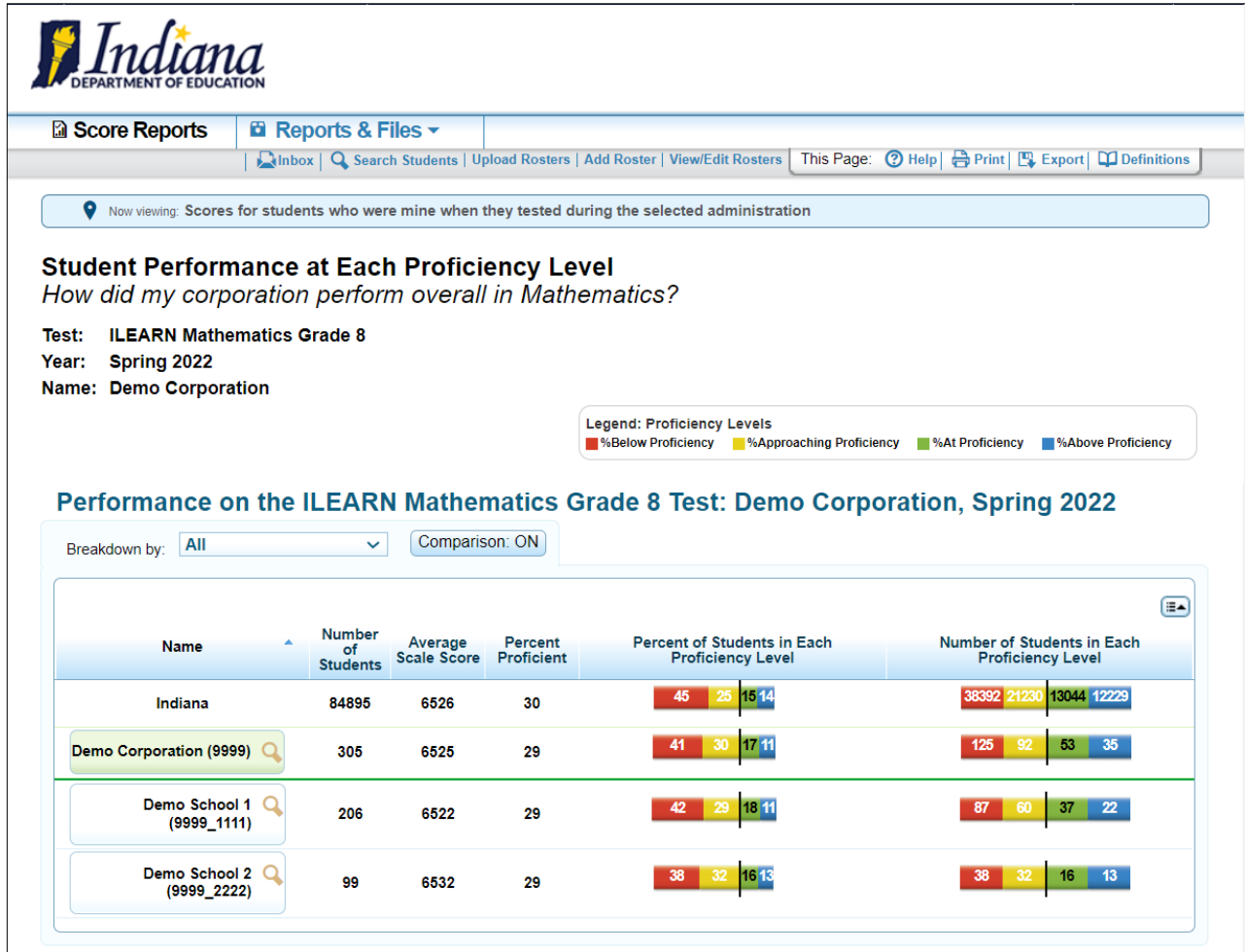


Figure 5: Corporation Aggregate-Level Subject Report, Grade 6 Science

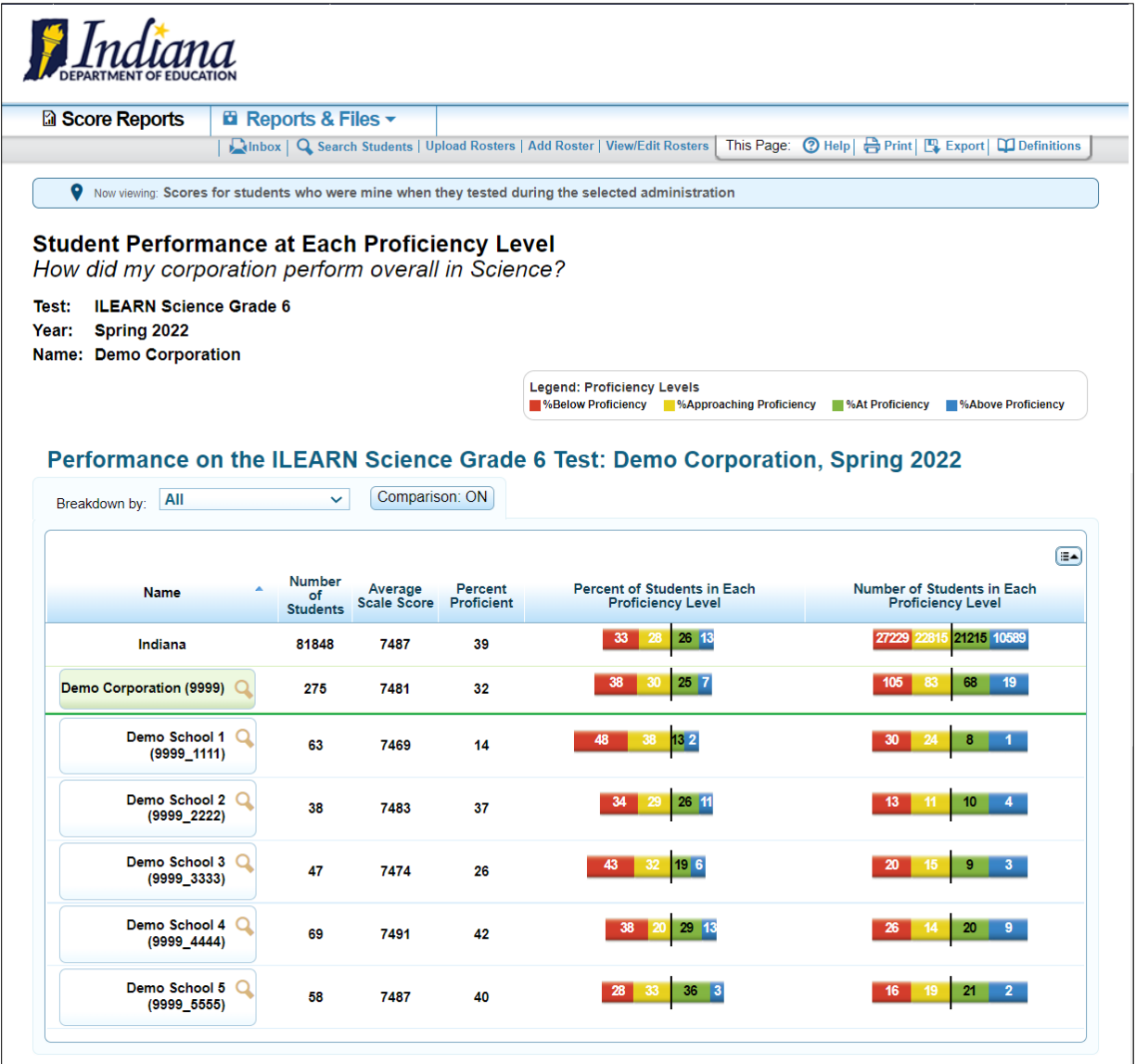
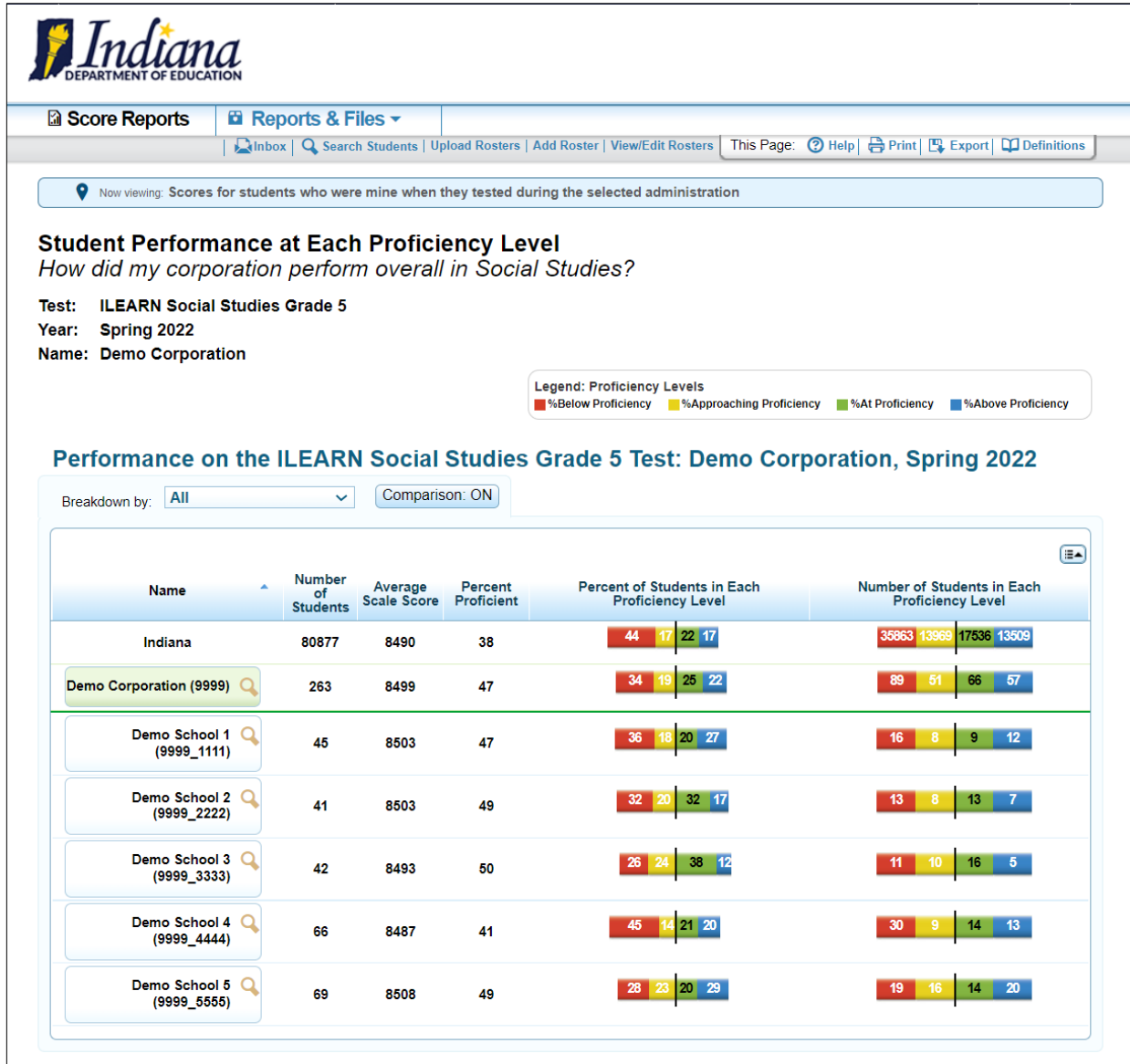


Figure 6: Corporation Aggregate-Level Subject Report, Grade 5 Social Studies



1.6.4 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 and standard error of the average scale score;
- Percentage proficient; and
- For each reporting category, the percentage of students in each performance 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 7 through Figure 10 present examples of the Corporation Aggregate-Level Reporting Category Report for ILEARN.

Figure 7: Corporation Aggregate-Level Reporting Category Report, Grade 8 ELA

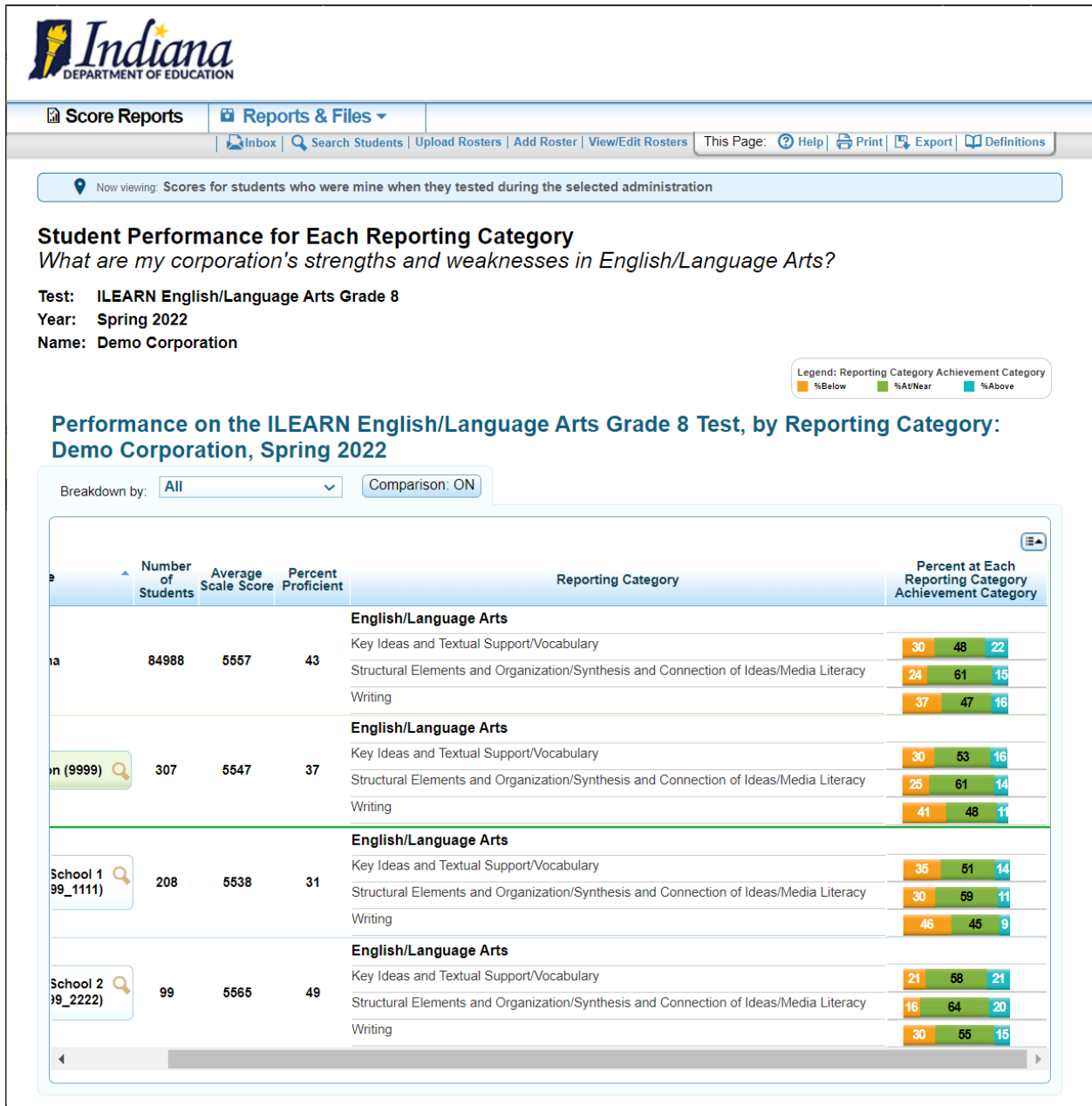


Figure 8: Corporation Aggregate-Level Reporting Category Report, Grade 8 Mathematics

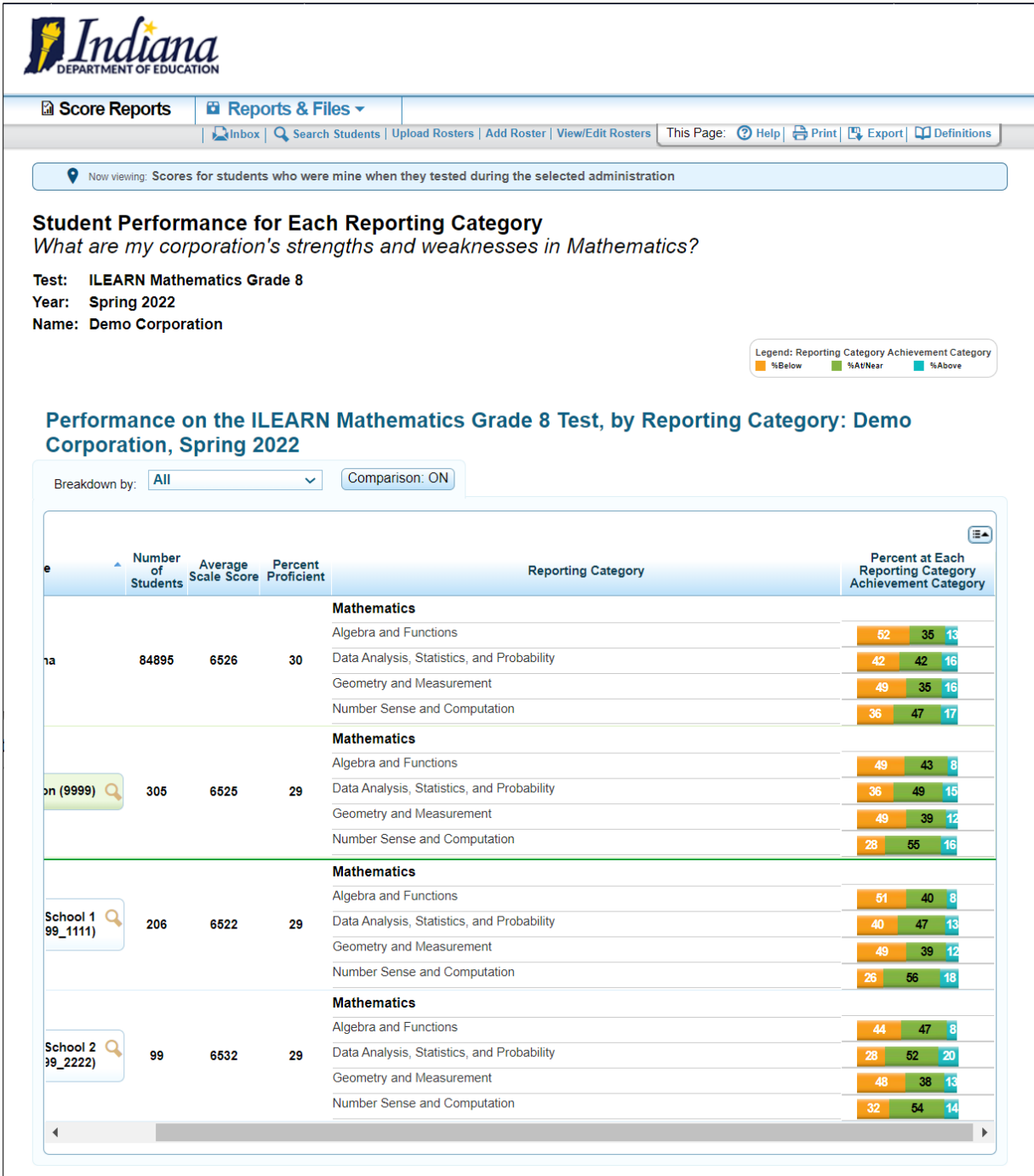


Figure 9: Corporation Aggregate-Level Reporting Category Report, Grade 6 Science

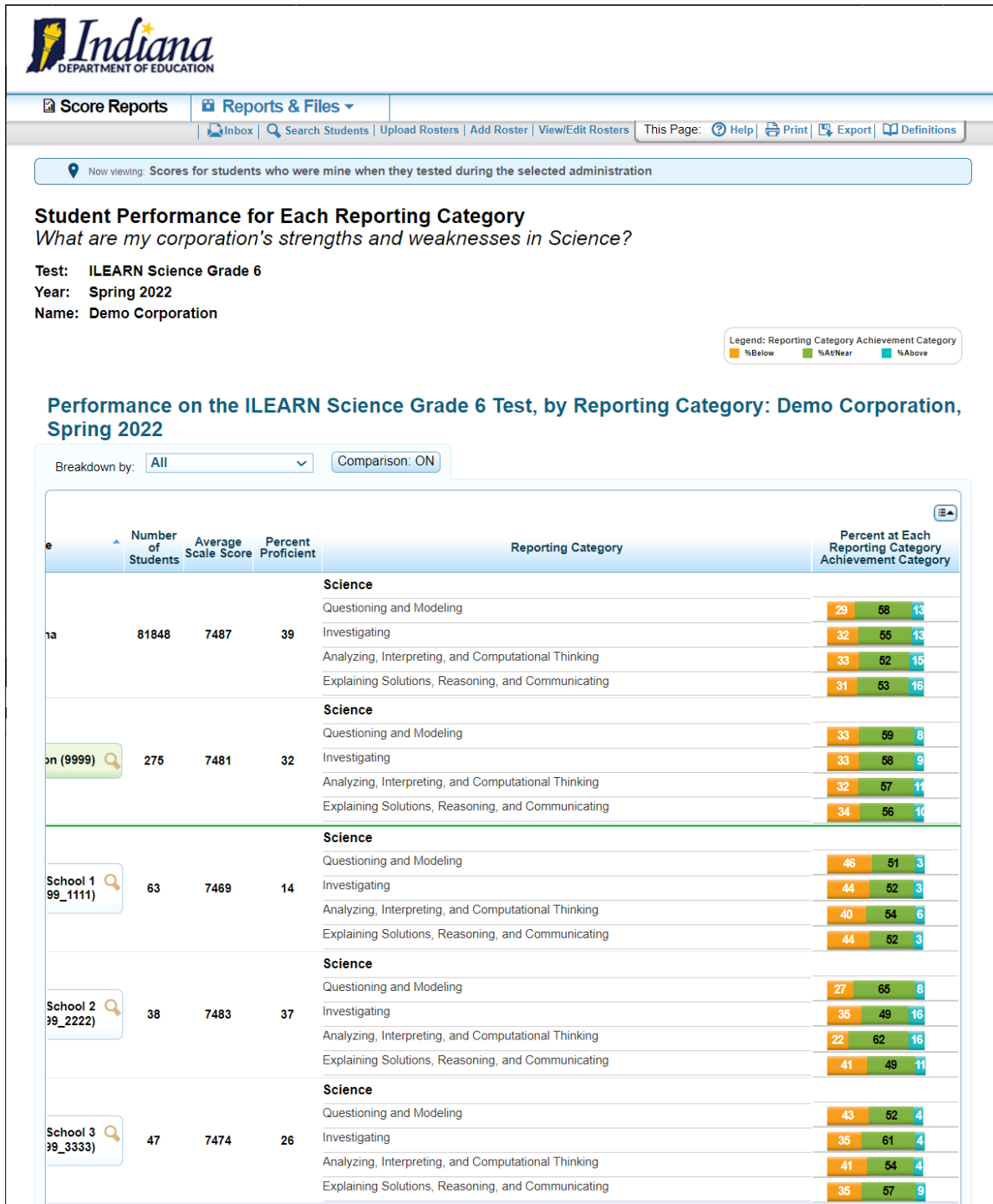
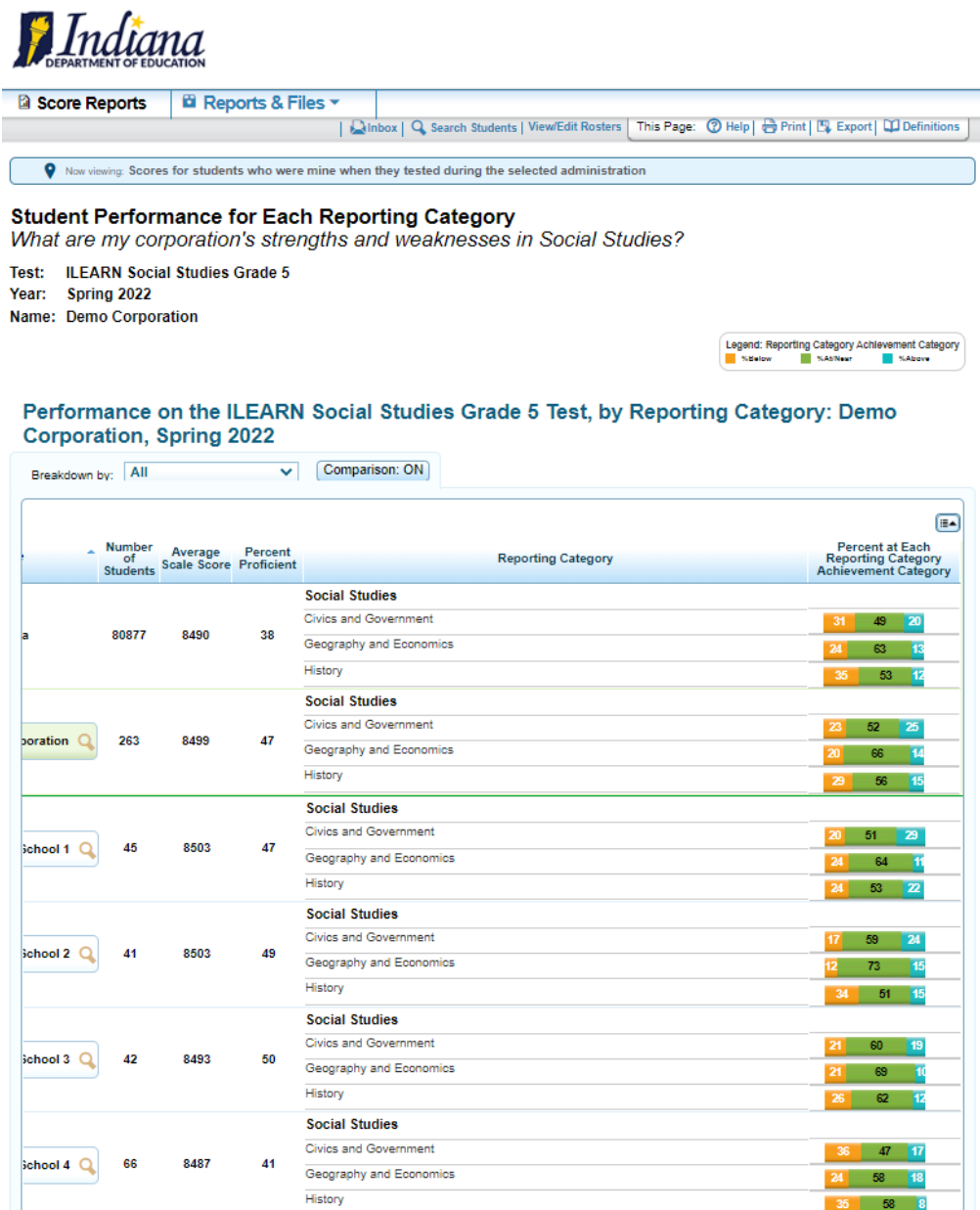


Figure 10: Corporation Aggregate-Level Reporting Category Report, Grade 5 Social Studies



1.6.5 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 class, 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 roster, school, or district 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. 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.

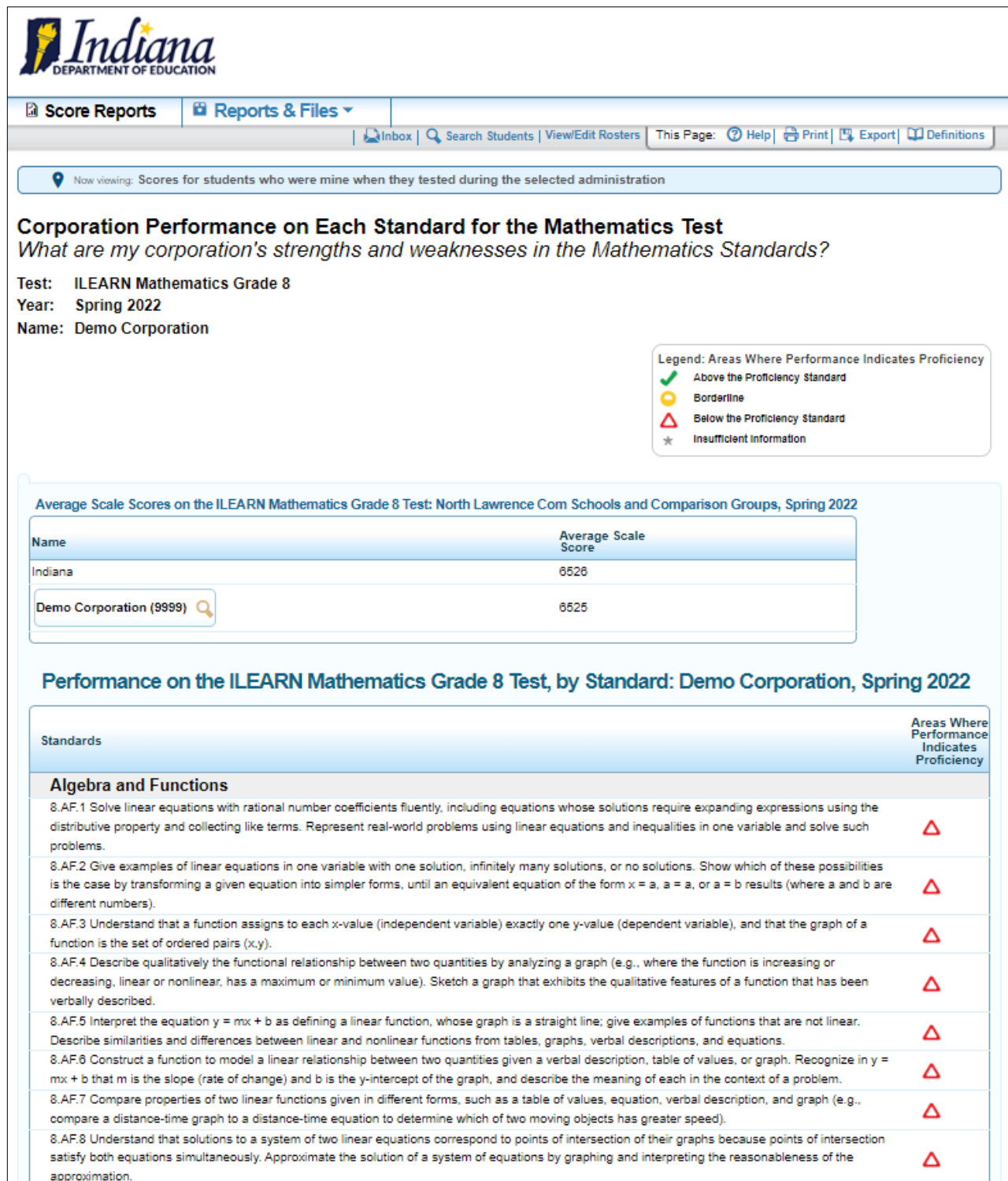
Figure 11 and Figure 12 present examples of the Aggregate-Level Standards Report for ELA and Mathematics, respectively.

Figure 11: Sample District Aggregate-Level Standards Report, Grade 8 ELA



8.RL.3.2 Analyze a particular point of view or cultural experience in a work of world literature considering how it reflects heritage, traditions, attitudes, and beliefs.	*
8.RL.4.1 Analyze the extent to which a filmed or live production of a story or play stays faithful to or departs from the text or script, evaluating the choices made by the director or actors.	*
8.RL.4.2 Analyze how works of literature draw on and transform earlier texts.	*
8.RN.3.2 Analyze in detail the structure of a specific paragraph in a text, including the role of particular sentences in developing and refining a key concept.	△
6-8.LH.3.2 Describe how a text presents information (e.g., sequentially, comparatively, causally).	*
6-8.LST.3.2 Analyze the structure an author uses to organize a text, including how the major sections contribute to the whole and to an understanding of the topic.	*
8.RN.3.3 Determine an author's perspective or purpose in a text, and analyze how the author acknowledges and responds to conflicting evidence or viewpoints.	△
6-8.LH.3.3 Identify aspects of a text that reveal an author's perspective or purpose (e.g., loaded language, inclusion or avoidance of particular facts).	*
6-8.LST.3.3 Analyze the author's purpose in providing an explanation, describing a procedure, or discussing an experiment in a text.	*
8.RN.4.1 Delineate and evaluate the argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient; recognize when irrelevant evidence is introduced.	△
6-8.LH.4.2 Distinguish among fact, opinion, and reasoned judgment in a text.	*
6-8.LST.4.2 Distinguish among facts, reasoned judgment based on research findings, and speculation in a text.	*
8.RN.4.2 Evaluate the advantages and disadvantages of using different mediums (e.g., print or digital text, video, multimedia) to present a particular topic or idea.	*
6-8.LH.4.1 Integrate visual information (e.g., charts, graphs, photographs, videos, or maps) with other information in print and digital texts.	*
6-8.LST.4.1 Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table).	*
8.RN.4.3 Analyze a case in which two or more texts provide conflicting information on the same topic and identify where the texts disagree on matters of fact or interpretation.	△
6-8.LH.4.3 Compare and contrast treatments of the same topic in a primary and secondary source.	*
6-8.LST.4.3 Compare and contrast the information gained from experiments, simulations, video, or multimedia sources with that gained from reading a text on the same topic.	*
Writing	
8.W.3.1 Write arguments in a variety of forms that: introduce claim(s), acknowledge and distinguish the claim(s) from alternate or opposing claims, and organize the reasons and evidence logically, support claim(s) with logical reasoning and relevant evidence, using accurate, credible sources and demonstrating an understanding of the topic or text, use effective transitions to create cohesion and clarify the relationships among claim(s), counterclaims, reasons, and evidence, establish and maintain a consistent style and tone appropriate to purpose and audience, provide a concluding statement or section that follows from and supports the argument presented.	△
6-8.LH.5.1 Write arguments focused on discipline-specific content.	*
6-8.LST.5.1 Write arguments focused on discipline-specific content.	*
8.W.3.2 Write informative compositions in a variety of forms that: introduce a topic clearly, previewing what is to follow; organize ideas, concepts, and information into broader categories; include formatting (e.g., headings), graphics (e.g., charts, tables), and multimedia when useful to aiding comprehension, develop the topic with relevant, well-chosen facts, definitions, concrete details, quotations, or other information and examples from various sources and texts, use appropriate and varied transitions to create cohesion and clarify the relationships among ideas and concepts, choose language and content-specific vocabulary that express ideas precisely and concisely, recognizing and eliminating wordiness and redundancy, establish and maintain a style appropriate to the purpose and audience, provide a concluding statement or section that follows from and supports the information or explanation presented.	△
6-8.LH.5.2 Write informative texts, including analyses of historical events.	*
6-8.LST.5.2 Write informative texts, including scientific procedures/experiments or technical processes that include precise descriptions and conclusions drawn from data and research.	*
8.W.3.3 Write narrative compositions in a variety of forms that: engage and orient the reader by establishing a context and point of view and introducing a narrator and/or characters, organize an event sequence (e.g., conflict, climax, resolution) that unfolds naturally and logically, using a variety of transition words, phrases, and clauses to convey sequence and signal shifts from one time frame or setting to another, use narrative techniques, such as dialogue, pacing, description, and reflection, to develop experiences, events, and/or characters, use precise words and phrases, relevant descriptive details, and sensory language to capture the action and convey experiences and events, provide an ending that follows from and reflects on the narrated experiences or events.	△
8.W.4 Apply the writing process to: plan and develop; draft; revise using appropriate reference materials; rewrite; try a new approach; and edit to produce and strengthen writing that is clear and coherent, with some guidance and support from peers and adults, use technology to interact and collaborate with others to generate, produce, and publish writing and present information and ideas efficiently.	△
6-8.LH.6.1 Plan and develop; draft; revise using appropriate reference materials; rewrite; try a new approach; and edit to produce and strengthen writing that is clear and coherent, with some guidance and support from peers and adults.	*
6-8.LST.6.1 Plan and develop; draft; revise using appropriate reference materials; rewrite; try a new approach; and edit to produce and strengthen writing that is clear and coherent, with some guidance and support from peers and adults.	*
8.W.5 Conduct short research assignments and tasks to build knowledge about the research process and the topic under study: formulate a research question, gather relevant information from multiple sources, using search terms effectively, and annotate sources, assess the credibility and accuracy of each source, quote or paraphrase the information and conclusions of others, avoid plagiarism and follow a standard format for citation, present information, choosing from a variety of formats.	△
6-8.LH.7.1 Conduct short research assignments and tasks to answer a question (including a self-generated question), drawing on several sources and generating additional related, focused questions that allow for multiple avenues of exploration.	*
6-8.LH.7.2 Gather relevant information from multiple sources, using search terms effectively; annotate sources; assess the credibility and accuracy of each source, and quote or paraphrase the data and conclusions of others while avoiding plagiarism and following a standard format for citation (e.g., APA or Chicago).	*
6-8.LST.7.3 Draw evidence from informational texts to support analysis, reflection, and research.	*
8.W.6.1b Demonstrate command of English grammar and usage, focusing on: Verbs: Explaining the function of verbals (gerunds, participles, infinitives) in general and their function in particular sentences; forming and using active and passive voice; recognizing and correcting inappropriate shifts in verb voice.	△
8.W.6.2b Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling focusing on: Punctuation: Using punctuation (comma, ellipsis, dash) to indicate a pause, break, or omission.	△
Others	
8.SL.3.1 Analyze the purpose of information presented in diverse media and formats (e.g., visually, quantitatively, orally) and evaluate the motives (e.g., social, commercial, political) behind its presentation.	△
8.SL.3.2 Delineate a speaker's argument and specific claims, evaluating the soundness of the reasoning and relevance and sufficiency of the evidence and identifying when irrelevant evidence is introduced.	△

Figure 12: Sample District Aggregate-Level Standards Report, Grade 8 Mathematics



Data Analysis, Statistics, and Probability	
8.DSP.1 Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantitative variables. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.	△
8.DSP.2 Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and describe the model fit by judging the closeness of the data points to the line.	△
8.DSP.3 Write and use equations that model linear relationships to make predictions, including interpolation and extrapolation, in real-world situations involving bivariate measurement data; interpret the slope and y-intercept.	△
8.DSP.4 Understand that, just as with simple events, the probability of a compound event is the fraction of outcomes in the sample space for which the compound event occurs. Understand and use appropriate terminology to describe independent, dependent, complementary, and mutually exclusive events.	●
8.DSP.5 Represent sample spaces and find probabilities of compound events (independent and dependent) using methods, such as organized lists, tables, and tree diagrams.	△
8.DSP.6 For events with a large number of outcomes, understand the use of the multiplication counting principle. Develop the multiplication counting principle and apply it to situations with a large number of outcomes.	△
Geometry and Measurement	
8.GM.1 Identify, define and describe attributes of three-dimensional geometric objects (right rectangular prisms, cylinders, cones, spheres, and pyramids). Explore the effects of slicing these objects using appropriate technology and describe the two-dimensional figure that results.	●
8.GM.2 Solve real-world and other mathematical problems involving volume of cones, spheres, and pyramids and surface area of spheres.	△
8.GM.3 Verify experimentally the properties of rotations, reflections, and translations, including: lines are mapped to lines, and line segments to line segments of the same length; angles are mapped to angles of the same measure; and parallel lines are mapped to parallel lines.	△
8.GM.4 Understand that a two-dimensional figure is congruent to another if the second can be obtained from the first by a sequence of rotations, reflections, and translations. Describe a sequence that exhibits the congruence between two given congruent figures.	△
8.GM.5 Understand that a two-dimensional figure is similar to another if the second can be obtained from the first by a sequence of rotations, reflections, translations, and dilations. Describe a sequence that exhibits the similarity between two given similar figures.	△
8.GM.6 Describe the effect of dilations, translations, rotations, and reflections on two-dimensional figures using coordinates.	△
8.GM.7 Use inductive reasoning to explain the Pythagorean relationship.	*
8.GM.8 Apply the Pythagorean Theorem to determine unknown side lengths in right triangles in real-world and other mathematical problems in two dimensions.	△
8.GM.9 Apply the Pythagorean Theorem to find the distance between two points in a coordinate plane.	●
Number Sense and Computation	
8.C.1 Solve real-world problems with rational numbers by using multiple operations.	△
8.C.2 Solve real-world and other mathematical problems involving numbers expressed in scientific notation, including problems where both decimal and scientific notation are used. Interpret scientific notation that has been generated by technology, such as a scientific calculator, graphing calculator, or excel spreadsheet.	△
8.NS.1 Give examples of rational and irrational numbers and explain the difference between them. Understand that every number has a decimal expansion; for rational numbers, show that the decimal expansion terminates or repeats, and convert a decimal expansion that repeats into a rational number.	●
8.NS.2 Use rational approximations of irrational numbers to compare the size of irrational numbers, plot them approximately on a number line, and estimate the value of expressions involving irrational numbers.	△
8.NS.3 Given a numeric expression with common rational number bases and integer exponents, apply the properties of exponents to generate equivalent expressions.	●
8.NS.4 Use square root symbols to represent solutions to equations of the form $x^2 = p$, where p is a positive rational number.	△
Others	
PS.1: Make sense of problems and persevere in solving them.	△
PS.2: Reason abstractly and quantitatively.	●
PS.3: Construct viable arguments and critique the reasoning of others.	△
PS.4: Model with mathematics.	△
PS.5: Use appropriate tools strategically.	*
PS.6: Attend to precision.	△
PS.7: Look for and make use of structure.	●
PS.8: Look for and express regularity in repeated reasoning.	△

1.6.6 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; and
- Lexile® (for ELA) or Quantile® (for Mathematics) measure.

Figure 13 through Figure 16 demonstrate examples of the Student-Level Subject Report for ILEARN.

Figure 13: Student-Level Subject Report, Grade 8 ELA

Indiana
DEPARTMENT OF EDUCATION

Score Reports | Reports & Files

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

Student Performance in Each Proficiency Level

How did my students perform overall in English/Language Arts?

Test: ILEARN English/Language Arts Grade 8
Year: Spring 2022
Name: Demo School

Breakdown by: All

Average Scale Scores on the ILEARN English/Language Arts Grade 8 Test: Demo School and Comparison Groups, Spring 2022

Name	Average Scale Score
Indiana	5557
Demo Corporation (9999)	5547
Demo School (9999_9999)	5538

Performance on the ILEARN English/Language Arts Grade 8 Test, by Student: Demo School, Spring 2022

Name	STN	Scale Score	Proficiency Level	Reported Lexile® Measure	College and Career Readiness Indicator
Demo, Student A.	123456789	5578	At Proficiency	1170L	Yes
Demo, Student B.	234567890	5455	Below Proficiency	870L	No
Demo, Student C.	345678901	5636	At Proficiency	1310L	Yes
Demo, Student D	456789012	5492	Below Proficiency	960L	No

Figure 14: Student-Level Subject Report, Grade 8 Mathematics

Indiana
DEPARTMENT OF EDUCATION

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 in Each Proficiency Level

How did my students perform overall in Mathematics?

Test: ILEARN Mathematics Grade 8
Year: Spring 2022
Name: Demo School

Breakdown by: **All**

Average Scale Scores on the ILEARN Mathematics Grade 8 Test: Demo School and Comparison Groups, Spring 2022

Name	Average Scale Score
Indiana	6526
Demo Corporation (9999)	6479
Demo School (9999_9999)	6441

Performance on the ILEARN Mathematics Grade 8 Test, by Student: Demo School, Spring 2022

Name	STN	Scale Score	Proficiency Level	Reported Quantile® Measure	College and Career Readiness Indicator
Demo, Student A.	123456789	6421	Below Proficiency	750Q	No
Demo, Student B.	234567890	6601	At Proficiency	1165Q	Yes
Demo, Student C.	345678901	6505	Below Proficiency	945Q	No
Demo, Student D.	456789012	6409	Below Proficiency	720Q	No
Demo, Student E.	567890123	6529	Approaching Proficiency	1000Q	No

Figure 15: Student-Level Subject Report, Grade 6 Science

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

Student Performance in Each Proficiency Level

How did my students perform overall in Science?

Test: ILEARN Science Grade 6
 Year: Spring 2022
 Name: Demo School

Breakdown by: All Go

Average Scale Scores on the ILEARN Science Grade 6 Test: Demo School and Comparison Groups, Spring 2022

Name	Average Scale Score
Indiana	7487
Demo Corporation (9999)	7481
Demo School (9999_9999)	7491

Performance on the ILEARN Science Grade 6 Test, by Student: Demo School, Spring 2022

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator
Demo, Student A.	123456789	7457	Below Proficiency	No
Demo, Student B.	234567890	7486	Approaching Proficiency	No
Demo, Student C.	345678901	7534	At Proficiency	Yes
Demo, Student D.	456789012	7543	At Proficiency	Yes

Figure 16: Student-Level Subject Report, Grade 5 Social Studies

Indiana
DEPARTMENT OF EDUCATION

Score Reports | **Reports & Files** | [Inbox](#) | [Search Students](#) | [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 in Each Proficiency Level

How did my students perform overall in Social Studies?

Test: ILEARN Social Studies Grade 5
Year: Spring 2022
Name: Demo Roster

Breakdown by: **All**

Average Scale Scores on the ILEARN Social Studies Grade 5 Test: Demo Roster and Comparison Groups, Spring 2022

Name	Average Scale Score
Indiana	8490
Demo Corporation (9999)	8499
Demo Roster	8508

Performance on the ILEARN Social Studies Grade 5 Test, by Student: Demo Roster, Spring 2022

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator
Demo Student A.	99999991	8474	Below Proficiency	No
Demo Student B.	99999992	8485	Approaching Proficiency	No
Demo Student C.	99999993	8500	Approaching Proficiency	No
Demo Student D.	99999994	8486	Approaching Proficiency	No
Demo Student E.	99999995	8582	Above Proficiency	Yes

1.6.7 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;
- Reporting category; and
- Performance category.

Figure 17 through Figure 20 displays this information for ILEARN.

Figure 17: Student-Level Reporting Category Report, Grade 8 ELA

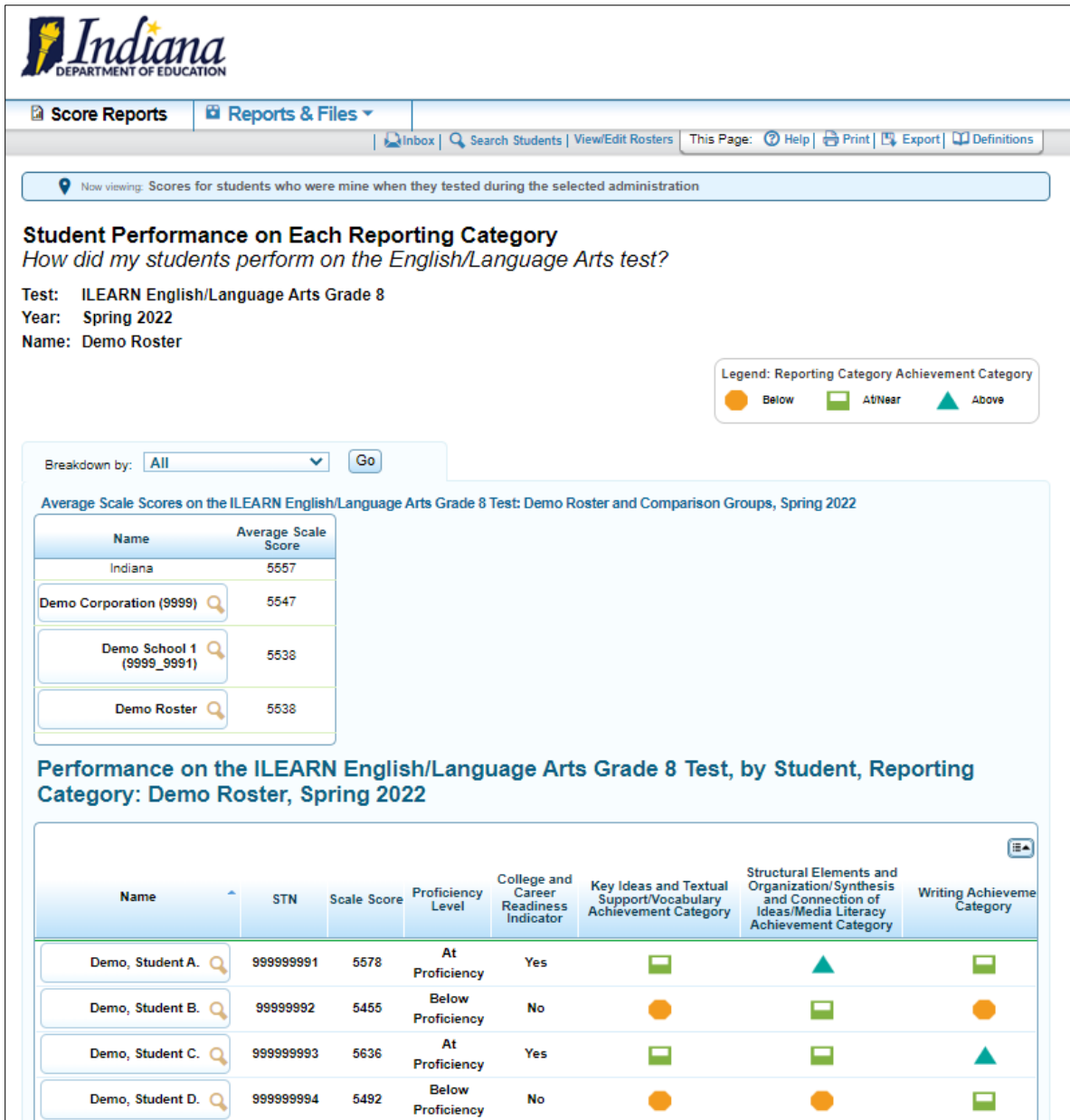


Figure 18: Student-Level Reporting Category Report, Grade 8 Mathematics

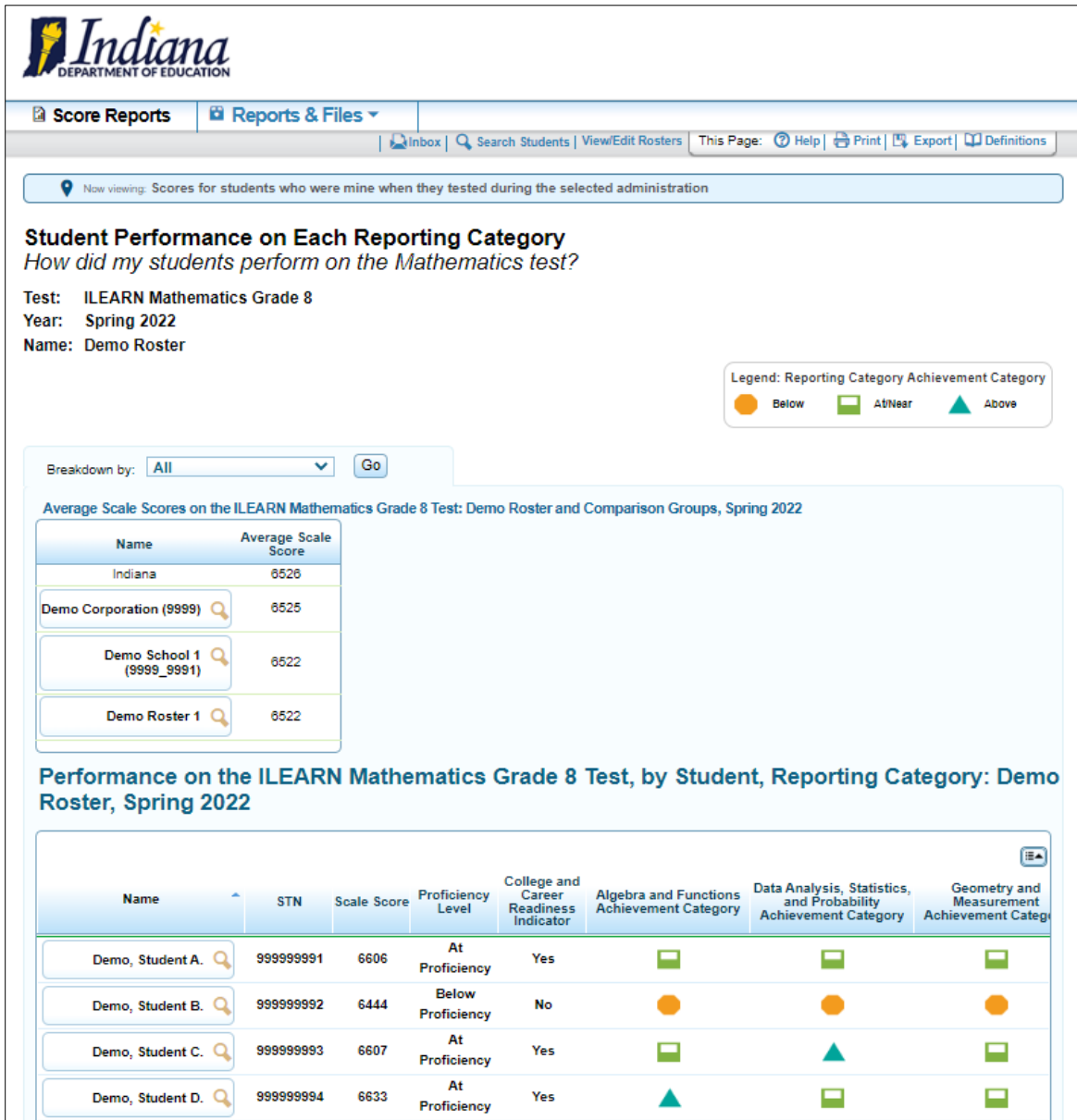


Figure 19: Student-Level Reporting Category Report, Grade 6 Science

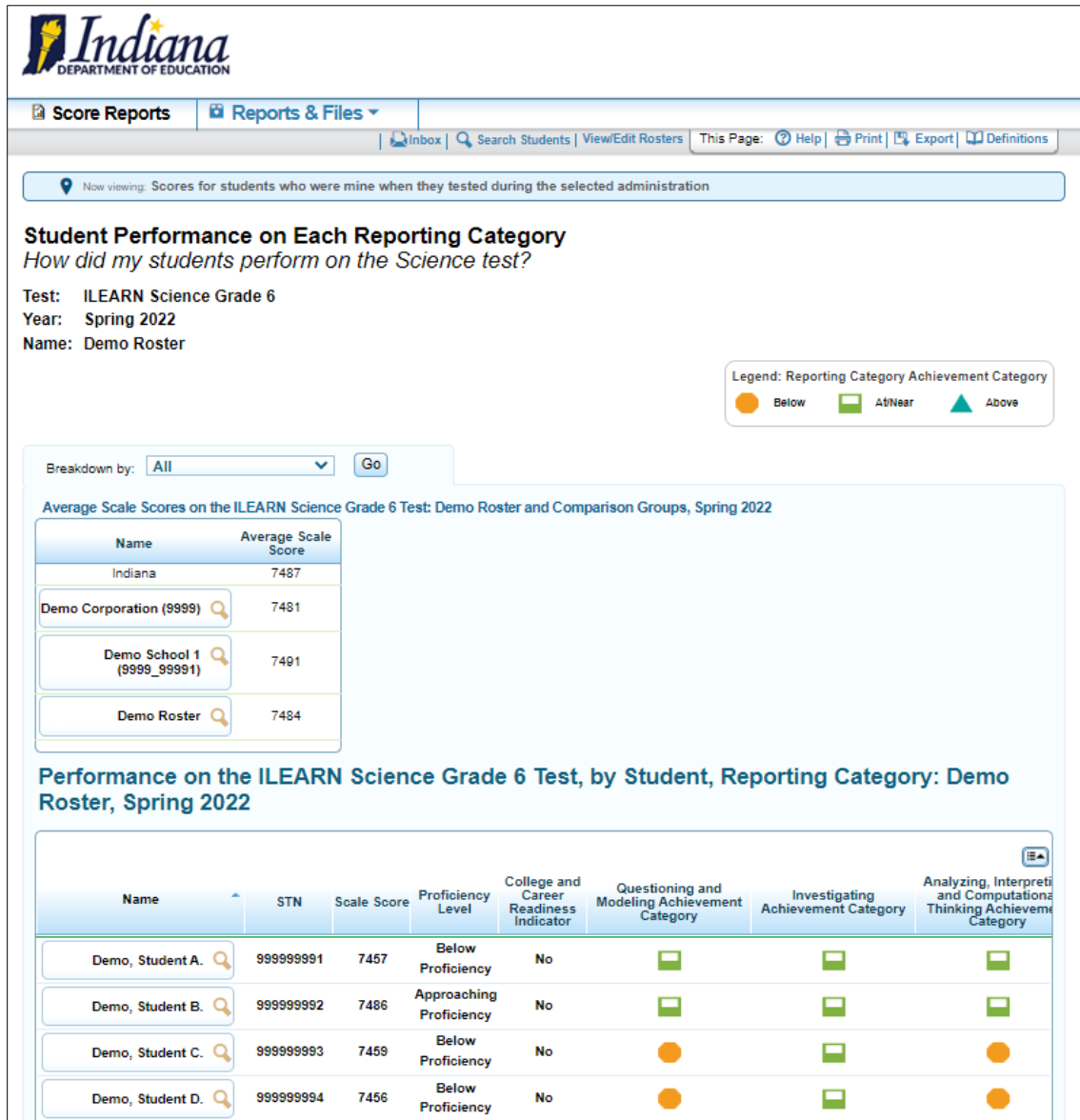
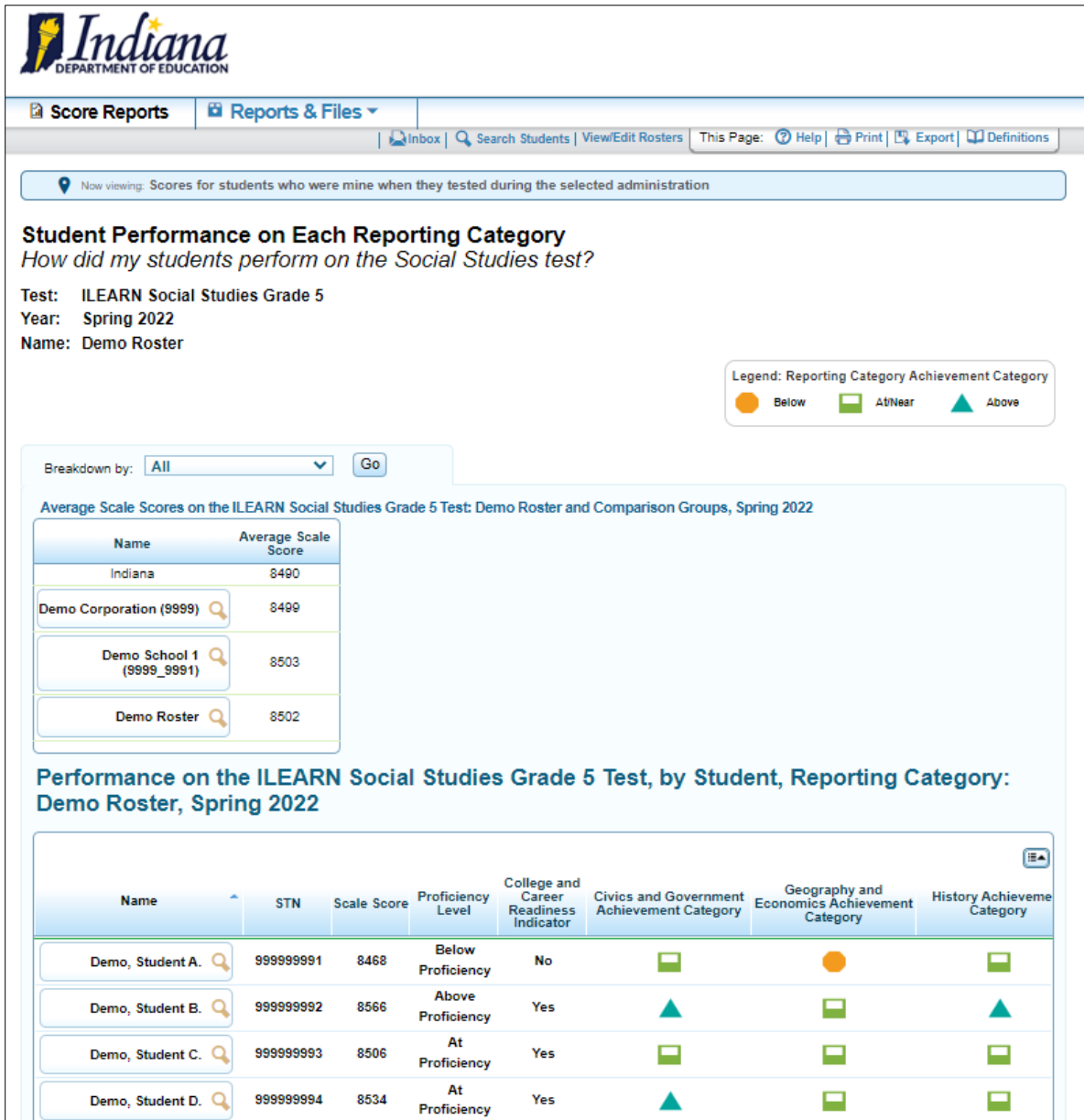


Figure 20: Student-Level Reporting Category Report, Grade 5 Social Studies



1.6.8 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 and SEM;
- Overall subject performance level;
- Average scale scores for a student’s state, corporation, and school;
- Performance category in each reporting category; and
- Writing performance descriptors in each dimension (ELA only).

The top of the report includes:

- Student’s name;
- Scale score with SEM;
- Performance level; and
- Lexile® (ELA only) or Quantile® (Mathematics only).

The middle section includes:

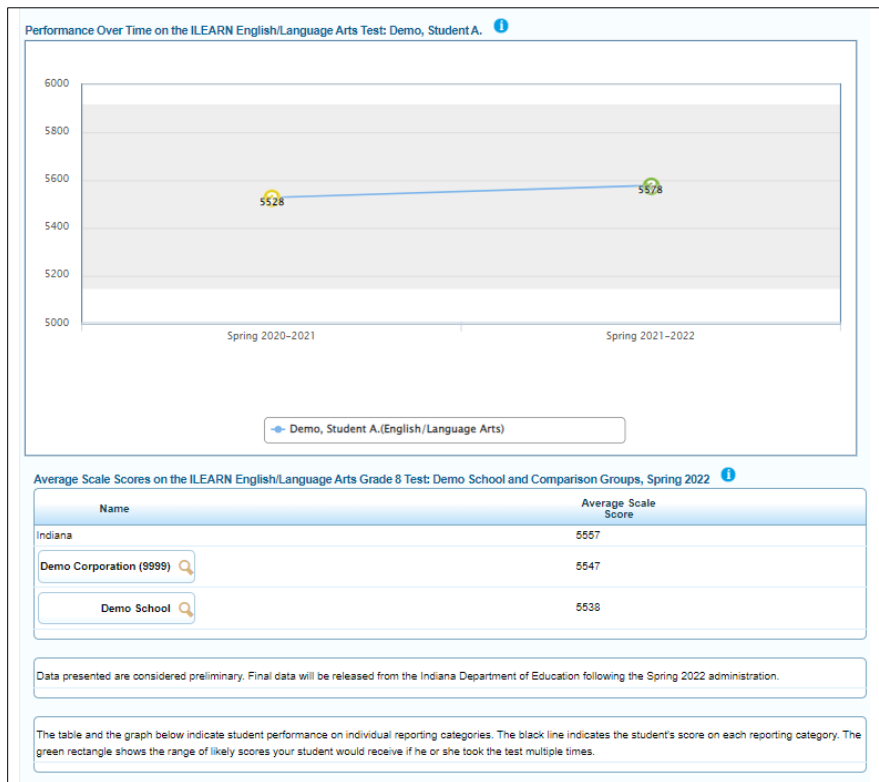
- Bar chart with the student’s scale score;
- Performance-level descriptors with cut scores at each performance level; and
- Average scale scores for state, corporation, and school aggregation levels.

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.
- ILEARN ELA and Mathematics reports will include trend reports that represent student performance over time. Note that trend performance over time data is available in Spring 2021 and Spring 2022.

Figure 21 through Figure 24 present examples of ISRs for ILEARN.

Figure 21: Individual Student Report, Grade 8 ELA



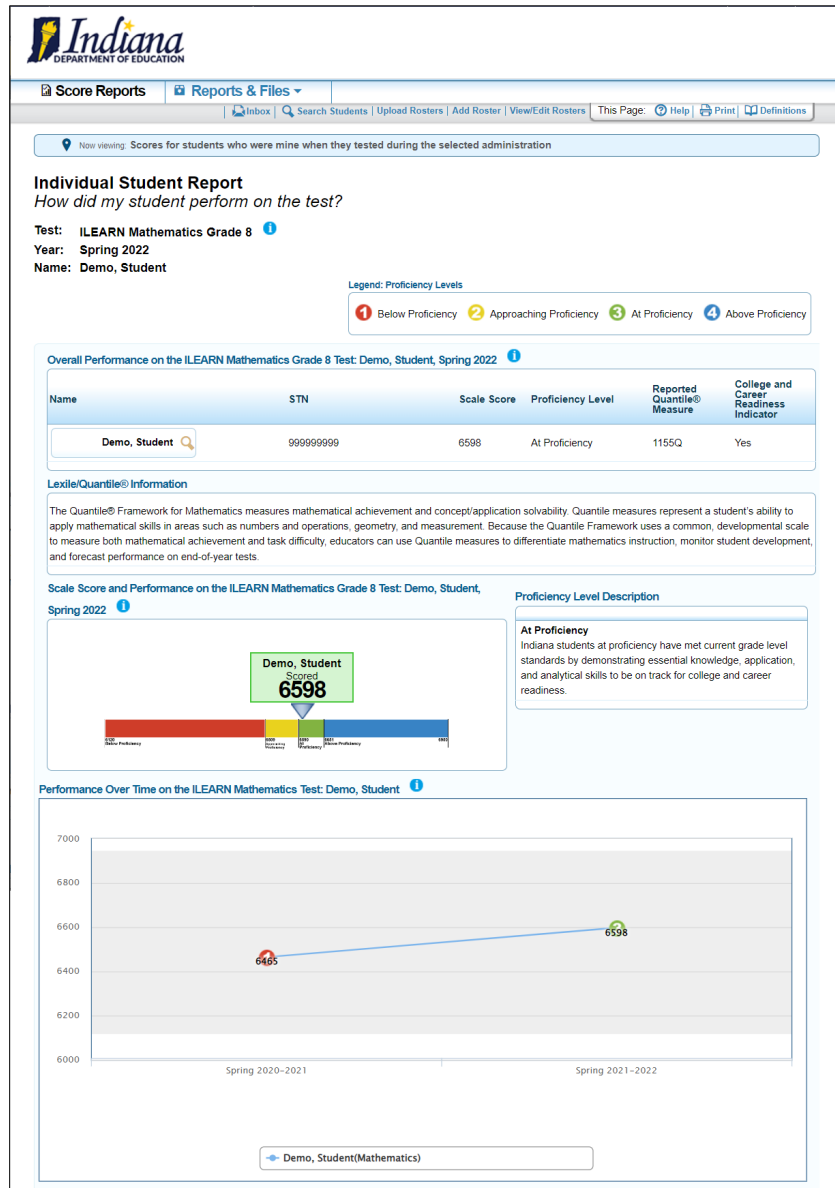
Performance on the ILEARN English/Language Arts Grade 8 Test, by Reporting Category: Demo, Student A., Spring 2022 ⓘ

Reporting Category	Reporting Category Performance	Reporting Category Description
Key Ideas and Textual Support/Vocabulary		<p>What These Results Mean Your student can often independently interact with literary, informational, historical, and scientific texts to explain how central ideas develop, describe how dialogue affects plot and characters, cite strong and relevant evidence, and interpret figures of speech.</p> <p>Next Steps Ask your student to read a literary or nonfiction text and explain the central idea and how it develops. Discuss how specific pieces of dialogue impact the characters and plot. Interpret any figures of speech and analogies in context with your student.</p>
Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy		<p>What These Results Mean Your student can almost always independently compare the structures of related texts, analyze points of view/cultural experiences, infer authors' purposes/positions, analyze literary works influenced by older texts, and evaluate persuasive techniques used by different media.</p> <p>Next Steps Ask your student to read two related texts about different cultural traditions. Discuss with your student how culture influences the points of view expressed. Read/listen to different media with your student and evaluate the effectiveness of the persuasive techniques used.</p>
Writing		<p>What These Results Mean Your student can often independently organize and develop writing for argumentative, informative, and narrative purposes; clearly distinguish a topic/claim; support ideas with relevant details; use transitions to clarify ideas; establish style; and use correct punctuation.</p> <p>Next Steps Ask your student to examine a text of his or her choice and discuss how the author organizes ideas in a logical way. Discuss how relevant details are used to support ideas. Ask your student to determine the text's style/voice and identify how the transitions clarify ideas.</p>

Writing Performance on the ILEARN English/Language Arts Grade 8 Test, Based on the Performance Task Writing Rubric: Demo, Student A., Spring 2022 ⓘ

Writing Prompt	Organization/Purpose	Evidence/Development & Elaboration	Conventions
Argumentative	The argumentative response has an inconsistent structure including an unclear claim, uneven development, few transitions, and loosely connected ideas. If present, the introduction or conclusion may be weak. The response may address the opposing argument. (2 out of 4 points)	The argumentative response provides uneven elaboration to support the claim including few facts and details cited from sources, weak elaborative techniques and ineffective language for the audience and purpose. (2 out of 4 points)	The argumentative response shows an adequate understanding of correct sentence formation, punctuation, capitalization, grammar usage, and spelling. (2 out of 2 points)

Figure 22: Individual Student Report, Grade 8 Mathematics



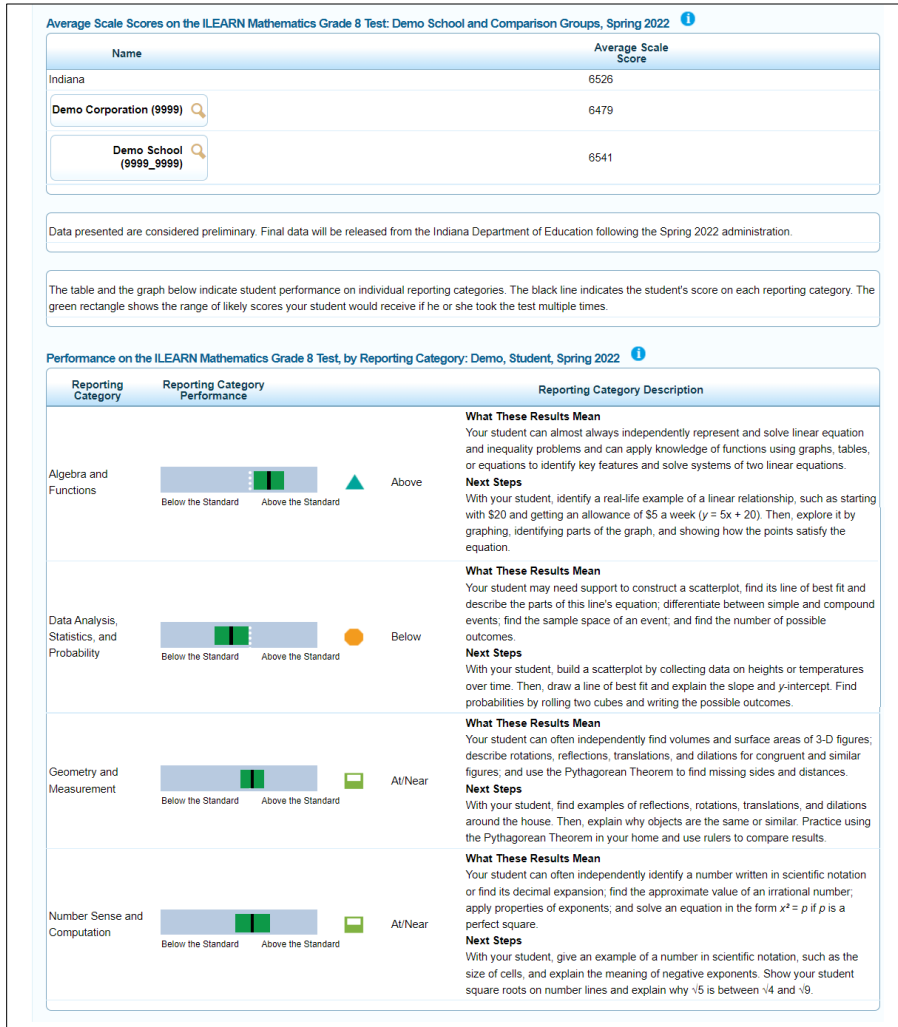



Figure 23: Individual Student Report, Grade 6 Science



[Score Reports](#) | [Reports & Files](#)

Inbox | Search Students | Upload Rosters | Add Roster | View/Edit Rosters | This Page: Help | Print | Definitions

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

Individual Student Report

How did my student perform on the test?


Test: ILEARN Science Grade 6 i
Year: Spring 2022
Name: Demo, Student

Overall Performance on the ILEARN Science Grade 6 Test: Demo, Student, Spring 2022 i

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator
Demo, Student	999999999	7529	At Proficiency	Yes

Scale Score and Performance on the ILEARN Science Grade 6 Test: Demo, Student, Spring 2022 i

Demo, Student
Scored
7529



Proficiency Level Description

At Proficiency
Indiana students at proficiency have met current grade level standards by demonstrating essential knowledge, application, and analytical skills to be on track for college and career readiness.

Average Scale Scores on the ILEARN Science Grade 6 Test: Demo School and Comparison Groups, Spring 2022 i

Name	Average Scale Score
Indiana	7487
Demo Corporation (9999)	7467
Demo School (9999_9999)	7489

Data presented are considered preliminary. Final data will be released from the Indiana Department of Education following the Spring 2022 administration.

The table and the graph below indicate student performance on individual reporting categories. The black line indicates the student's score on each reporting category. The green rectangle shows the range of likely scores your student would receive if he or she took the test multiple times.

Performance on the ILEARN Science Grade 6 Test, by Reporting Category: Demo, Student, Spring 2022 i



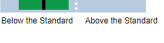


Reporting Category	Reporting Category Performance	Reporting Category Description
Questioning and Modeling		<p>What These Results Mean Your student can almost always independently construct models to develop questions, predictions, and explanations about the natural and designed worlds. He or she can independently identify constraints in a design problem and determine the criteria for possible solutions.</p> <p>Next Steps Ask your student to analyze a model of a natural system or engineering problem to find constraints or flaws. Then, ask your student to explain the model's limitations and describe how the model could be refined to better address questions.</p>
Investigating		<p>What These Results Mean Your student can often independently use investigations to produce and analyze data, use mathematics and computational tools to construct simulations, solve equations, make predictions, ask questions, and identify solutions efficiently and effectively.</p> <p>Next Steps Ask your student to mathematically explain the relationships among variables and to analyze the reliability of the data/results of a given investigation. Also, ask your student to explore proposed design solutions using simulations or models.</p>
Analyzing, Interpreting, and Computational Thinking		<p>What These Results Mean Your student may need support in planning and conducting scientific and engineering investigations. He or she may need additional support identifying variables, tools, and procedures associated with scientific experiments.</p> <p>Next Steps Ask your student to observe a phenomenon that can be studied using scientific or engineering processes. Then, ask your student to record questions that he or she has about the phenomenon and to identify variables that need to be studied to answer those questions.</p>
Explaining Solutions, Reasoning, and Communicating		<p>What These Results Mean Your student can almost always independently make reasoned arguments and cite supporting data to explain scientific and engineering ideas. He or she communicates scientific and engineering ideas orally and in writing and uses logic and evidence to analyze competing ideas.</p> <p>Next Steps Ask your student to collaborate with other students to research and evaluate the evidence that supports a given scientific theory. Then, ask your student to evaluate the relevance and validity of that evidence.</p>

Figure 24: Individual Student Report, Grade 5 Social Studies



Score Reports
Reports & Files

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

Individual Student Report

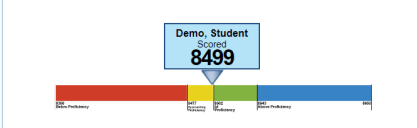
How did my student perform on the test?

Test: ILEARN Social Studies Grade 5 1
Year: Spring 2022
Name: Demo, Student

Overall Performance on the ILEARN Social Studies Grade 5 Test: Demo, Student, Spring 2022 1

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator
Demo, Student	999999999	8499	Approaching Proficiency	No

Scale Score and Performance on the ILEARN Social Studies Grade 5 Test: Demo, Student, Spring 2022 1



Approaching Proficiency

Indiana students approaching proficiency have nearly met current grade level standards by demonstrating some basic knowledge, application, and limited analytical skills. Students may require support to be on track for college and career readiness.


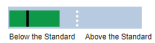
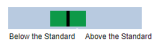
Average Scale Scores on the ILEARN Social Studies Grade 5 Test: Demo School and Comparison Groups, Spring 2022 1

Name	Average Scale Score
Indiana	8490
Demo Corporation (9999)	8471
Demo School (9999_9999)	8498

Data presented are considered preliminary. Final data will be released from the Indiana Department of Education following the Spring 2022 administration.

The table and the graph below indicate student performance on individual reporting categories. The black line indicates the student's score on each reporting category. The green rectangle shows the range of likely scores your student would receive if he or she took the test multiple times.


Performance on the ILEARN Social Studies Grade 5 Test, by Reporting Category: Demo, Student, Spring 2022 1

Reporting Category	Reporting Category Performance	Reporting Category Description
Civics and Government	 Above	<p>What These Results Mean Your student can almost always independently explain key ideas and concepts related to the founding of the U.S., the U.S. Constitution, elections, and the branches of government. Your student can identify and explain ways that citizens can bring about political change.</p> <p>Next Steps Ask your student to investigate a political issue that he or she feels strongly about. Then, ask your student to think about this issue in constitutional terms. Finally, discuss the issue and productive actions your student could take as a responsible citizen.</p>
Geography and Economics	 Below	<p>What These Results Mean Your student may need support using maps to locate places and regions and to identify physical and human systems from both today and the past. Your student may know that we have an economy but may need support defining market economies and describing how they work.</p> <p>Next Steps With your student, locate places or regions in the United States on modern and historical maps. With your student, discuss what market economies are. Define important terms with your student (supply, demand, price) and discuss how the terms apply to market economies.</p>
History	 At/Near	<p>What These Results Mean Your student can often identify early cultures and settlements in North America and major leaders who influenced the American Revolution. Your student often thinks chronologically and can use sources to examine historical events.</p> <p>Next Steps Ask your student to use multiple sources to examine early European settlements and Native American Indian cultures. Use multiple sources to examine major American Revolution leaders, then use these sources to describe how major leaders influenced the American Revolution.</p>

1.6.9 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 25) to help teachers, administrators, parents, and students better understand the data presented in the ISR. The ISRs and the supplemental “interpretive guide” are available in five different languages: Arabic, Chinese, Burmese, Spanish, and Vietnamese.

Figure 25: Supplemental Interpretive Guide



Indiana Learning Evaluation and Readiness Network ILEARN Assessment Results

Dear Parent/Guardian,
This report provides information about your child's performance on the Indiana Learning Evaluation and Readiness Network (ILEARN) assessment. ILEARN is the summative accountability assessment for Indiana students to measure student growth and proficiency in English/ language Arts, mathematics, science, and social studies according to the Indiana Academic Standards.

Please read this report closely and discuss the results with your child and his/her teacher. Thank you for supporting your child's education.

Indiana Department of Education

INFORMATION ON INDIANA'S ILEARN ASSESSMENT

ILEARN is Indiana's new online computer-adaptive assessment designed to measure your child's proficiency based on the Indiana Academic Standards. Overall student results in ILEARN are reported as four-digit scale scores. The overall scale scores for Indiana students align with the four proficiency levels (Below Proficiency, Approaching Proficiency, At Proficiency, and Above Proficiency). This report provides your family with useful information, such as how your child scored on the assessment, whether the scores meet state proficiency standards, and how your child's scores compare with students in his/her school, corporation, and state.

UNDERSTANDING THE ILEARN ASSESSMENT

Individual Student Report
How did my student perform on the test?
Test: ILEARN English/Language Arts Grade 6
Year: Spring 2019
Name: Demo, Student A

Overall Performance on the ILEARN English/Language Arts Grade 6 Test: Demo, Student A, Spring 2019				
Name	STN	Scale Score	Proficiency Level	Reported Lexile® Measure
Demo, Student A	99999901	2710	Above Proficiency	750L

Scale Score: Represents your child's overall numerical score placed on an alternative scale rather than just using percent correct or a raw score.

Proficiency Level: Indicates which proficiency level your child is placed into based on the overall scale scores.

Reported Lexile® Measure (English/Language Arts only): Represents your child's reading ability, and serves as a guide in selecting books for your child.

Reported Quantile® Measure (Mathematics only): Represents your child's mathematical skills, and helps you identify activities to support your child in gaining mathematical skills and understanding.

College and Career Readiness Indicator: Indicates whether your child meets the college-and-career readiness standards.

We encourage you to review these results with your child and his/her teacher. If you have questions about the contents of this report, contact your local school or corporation.

Questions to consider with your child's teacher:

- ▶ What are strengths?
- ▶ What are areas of growth?
- ▶ What strategies can we use to support growth?
- ▶ What instructional materials do you recommend for my child?

Based on your child's ILEARN scale score, he/she is placed into one of the four proficiency levels: Below Proficiency, Approaching Proficiency, At Proficiency, or Above Proficiency. **Students performing At or Above Proficiency are on track for college and career readiness.**

Your child's test score can vary if the test is taken several times. His/her knowledge and skills likely fall within a score range rather than a precise number. Scores are an estimation of your child's ability.

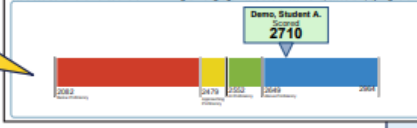
The comparison scores table shows how your child's scale score compares with peers at the school, corporation, and state levels.

The reporting category performance table shows your child's performance across domains within a content area. Reporting category performance is reported as Below (●), At/Near (■), or Above (▲).

Bar charts indicate how your child performed. The black bar shows your child's performance. The white bar shows the expectation by domain. The green band shows the range of performance expected over time typically associated with the assessment's small measurement error.

English/Language Arts reports include descriptions of your child's performance on the Performance Task (i.e., writing portion). If a condition code appears, your child's response could not be scored. Unscorable responses include responses that are blank, insufficient, written in a non-scorable language, off-topic, off-purpose, or illegible.

Scale Score and Performance on the ILEARN English/Language Arts Grade 6 Test: Demo, Student A, Spring 2019



Average Scale Scores on the ILEARN English/Language Arts Grade 6 Test: Demo School 9991 and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	2427
Demo Corporation 9999 (9999)	2466
Demo School 9991 (9999, 9991)	2484

Performance on the ILEARN English/Language Arts Grade 6 Test by Reporting Category: Demo, Student A, Spring 2019

Reporting Category	Reporting Category Performance	Reporting Category Description
Key Ideas and Details	At/Near	Your student can almost always independently identify with literary information, historical, and scientific texts, but at other times requires change, close attention, and interpretation of the impact of words.
Text Structure and Organization	Approaching	Your student can often independently explain how authors structure information, develop points of view, and support these with details. He or she can compare text features and analyze texts from different sources, genres, or media approach similar themes and topics.
Writing	Below	Your student may need regular guidance and help developing writing for personal, social, and academic purposes. He or she may need help supporting ideas with facts and details, choosing appropriate words, and using correct punctuation.

Writing Performance on the ILEARN English/Language Arts Grade 6 Test on the Performance Task Writing Subtest: Demo, Student A, Spring 2019

Writing Prompt	Organizational/Purpose	Evidence/Development & Elaboration	Conventions
The informative response has a recognizable structure including a clear topic or controlling idea, adequate development, and some needed transitions to clarify points. The response has adequate textual detail and organization and a minor over-completeness. (2 out of 2 points)	The informative response provides adequate evidence to support the topic or controlling idea including adequate facts and relevant text evidence. Some elaboration is present, and general language appropriate for the audience and purpose. (2 out of 2 points)	The informative response shows an adequate understanding of correct sentence formation, punctuation, capitalization, grammar usage, and spelling. (2 out of 2 points)	

ADDITIONAL RESOURCES

- To practice questions similar to what your child has seen on ILEARN, go to <https://ilearn.portal.cambiumast.com/>
- For more information about this assessment, go to <https://www.in.gov/doe/students/assessment/ilearn/>

Indiana Department of Education

Score Interpretation Guide

42

Indiana Department of Education

1.6.10 Reports by Sub-Group

At the aggregate level, student performance can be broken down by demographic sub-groups, such as gender (Figure 26) or English language learner status (Figure 27).

Figure 26: Corporation Aggregate-Level Subject Report by Gender, Grade 8 ELA

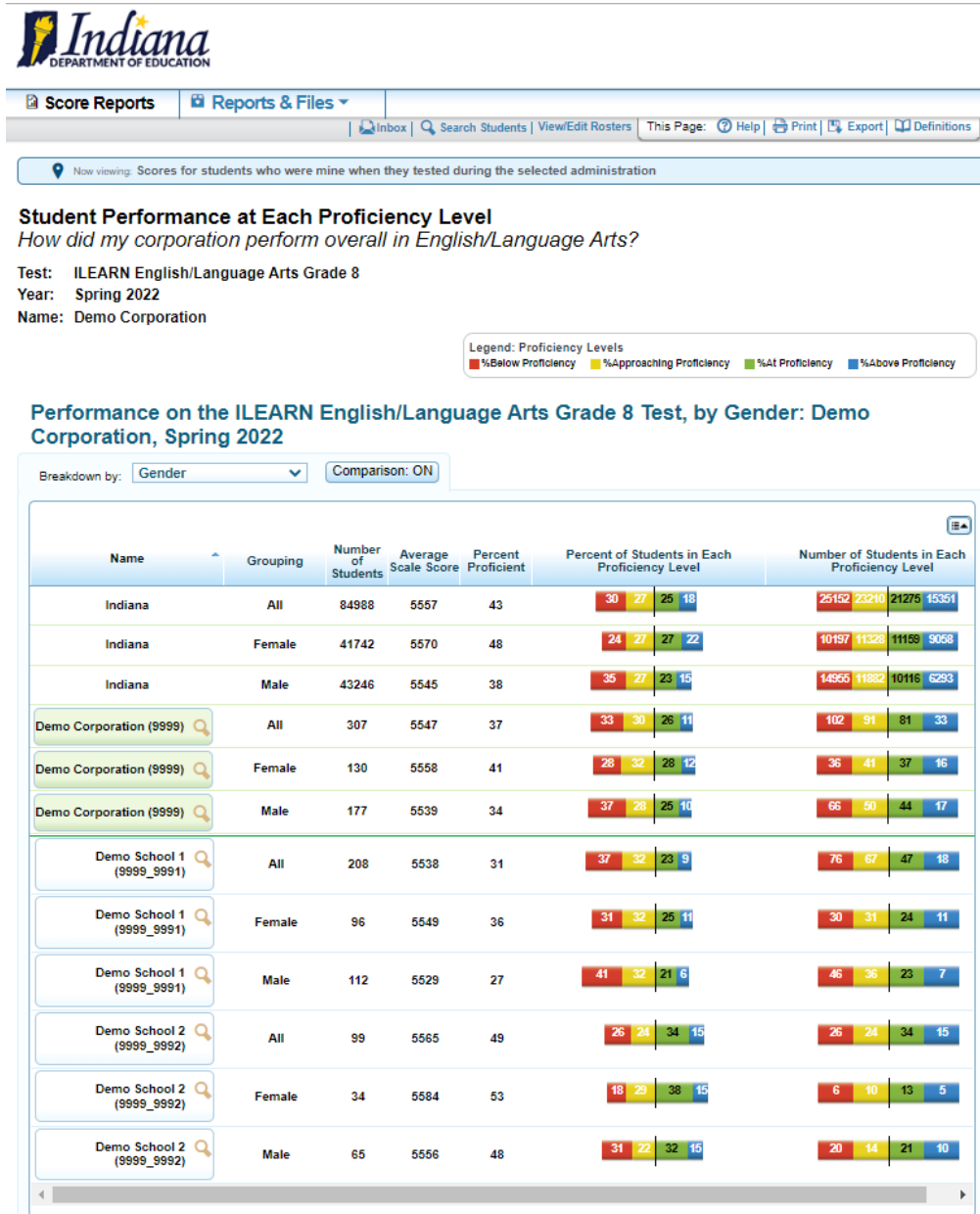


Figure 27: Corporation Aggregate-Level Reporting Category Report by Section 504 Plan Status, Grade 8 Mathematics



1.6.11 Data File

ORS users have the option to quickly generate a comprehensive data file of their students' scores. Data files (see Figure 28) can be downloaded in Microsoft Excel or CSV format and contain a wide variety of data, including scale 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 28: Data File

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG		
1	Gender	Ethnicity	Special Ed Identified	Section 5C	Socioecon	Enrolled C	Enrolled S	Enrolled S	Enrolled C	Enrolled C	Test name	Overall sc	Overall pr	Reported	College ar	Passing St	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	Reporting	
2	M	Hispanic	N	N	Y						8 Demo Sch 9999_9991 Demo Cor	9999 ILEARN En	3516	Approach	1020L	No		At/Near			Below			At/Near							
3	M	Hispanic	N	N	Y						8 Demo Sch 9999_9991 Demo Cor	9999 ILEARN Mi	6514	Approach	965Q	No		Below			At/Near			At/Near						At/Near	
4	F	White	N	N	Y						5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN En	5550	At Profici	995L	Yes		At/Near			At/Near			At/Near						Above	
5	F	White	N	N	Y						5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN Mi	6487	Approach	745Q	No		At/Near			At/Near			At/Near						At/Near	
6	F	White	N	N	Y						5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN So	8495	Approach	Profici	No		At/Near			At/Near			At/Near						At/Near	
7	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN En	5548	At Profici	990L	Yes		At/Near			At/Near			At/Near						Above	
8	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN Mi	6537	At Profici	890Q	Yes		At/Near			At/Near			At/Near						Above	
9	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN So	8527	At Proficiency		Yes		At/Near			At/Near			At/Near						At/Near	
10	M	White	N	N	N	N					5 Demo Sch 9999_9991 Demo Cor	9999 ILEARN En	5600	Above Prc	1105L	Yes		At/Near			At/Near			At/Near						Above	
11	M	White	N	N	N	N					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN Mi	6637	Above Prc	1125Q	Yes		Above			Above			Above							Above
12	M	White	N	N	N	N					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN So	8580	Above Proficiency		Yes		Above			Above			Above							
13	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN En	5506	Approach	900L	No		At/Near			At/Near			At/Near						At/Near	
14	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN Mi	6508	Approach	805Q	No		At/Near			At/Near			At/Near						Below	
15	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN So	8515	At Proficiency		Yes		At/Near			At/Near			At/Near						At/Near	
16	M	Hispanic	N	N	N	N					6 Demo Sch 9999_9992 Demo Cor	9999 ILEARN En	5473	Below Prc	870L	No		At/Near			At/Near			Below						Below	
17	M	Hispanic	N	N	N	N					6 Demo Sch 9999_9992 Demo Cor	9999 ILEARN Mi	6455	Below Prc	665Q	No		Below			Below			Below						Below	
18	M	Hispanic	N	N	N	N					6 Demo Sch 9999_9992 Demo Cor	9999 ILEARN Sc	7519	At Proficiency		Yes		At/Near			At/Near			At/Near						At/Near	
19	M	Hispanic	N	N	N	N					7 Demo Sch 9999_9992 Demo Cor	9999 ILEARN En	5551	Approach	1045L	No		At/Near			Below			At/Near						At/Near	
20	M	Hispanic	N	N	N	N					7 Demo Sch 9999_9992 Demo Cor	9999 ILEARN Mi	6577	At Profici	1070Q	Yes		At/Near			Above			At/Near						At/Near	
21	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN En	5368	Below Prc	650L	No		Below			Below			Below						Below	
22	M	Hispanic	N	Y	N	Y					5 Demo Sch 9999_9992 Demo Cor	9999 ILEARN Mi	6422	Below Prc	555Q	No		Below			At/Near			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 levels. 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 districts are providing effective instruction of the Indiana Academic Standards. For the adaptive assessments (ELA, Mathematics, and Science in Spring 2022), 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 item.

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 several 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 instruction 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 and 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 are able to 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 categories for reporting categories can be used to identify an individual student's relative strengths and weaknesses among reporting categories within a content 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 a precise measure of 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 students' 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 of 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 PERFORMANCE LEVEL

Based on their scale score, a student will receive an overall performance level. 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) using performance standards (or cut scores—see Section 2.5). 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 the 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.in.gov/doe/students/assessment/ilearn/>.

2.4 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.

2.5 CUT SCORES

For all grades and subjects within ILEARN, scale scores are mapped onto four performance levels (Level 1—Below Proficiency, Level 2—Approaching Proficiency, Level 3—At Proficiency, and Level 4—Above Proficiency). 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 show 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

2.6 AGGREGATED SCORES

Students' scale scores are aggregated at roster, teacher, 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.

2.7 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 and Elaboration, and Conventions, as shown in Table 12. Each of these dimensions is independently scored and reported on the student reports. For item analysis Organization/Purpose and Evidence 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 if it can 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 and Elaboration	1–4 points
Conventions	0–2 points

2.8 RELATIVE STRENGTH AND WEAKNESS

For standard performance, relative strengths and weaknesses at each standard are reported for aggregate levels only (e.g., classroom, 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' achievement 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.

2.9 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®¹ 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.

2.10 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.

¹ Lexiles and Quantiles are the intellectual property of MetaMetrics, Inc.

3. SUMMARY

ILEARN results are reported online via the Online Reporting System (ORS). The results are released after the testing window has closed and state quality control measures are completed. Starting with the 2019–2020 school year, the system can report results on tests as they are completed and hand-scores are available.

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.



**Indiana Learning
Evaluation Readiness
Network
(ILEARN)
2021–2022**

Volume 6

**Examining the Consistency of Corporation
Performance between Spring 2021 and Spring 2022
ILEARN Test Administrations**

Table of Contents

1. BACKGROUND	1
2. PURPOSE, METHODOLOGY, AND DATA.....	2
2.1 Purpose	2
2.2 Methodology	2
2.3 Data	3
3. RESULTS	5
3.1 Predicting Spring 2022 Scores	5
3.2 Predicting the Magnitude of Spring 2022 Score Residuals	9
3.3 Predicting the Direction of Spring 2022 Score Residuals.....	12
4. SUMMARY	15

List of Tables

Table 1: Prediction Regression of Spring 2022 Scores—Standardized Coefficients, ELA	6
Table 2: Prediction Regression of Spring 2022 Scores—Standardized Coefficients, Mathematics	6
Table 3: Prediction Regression of Spring 2022 Scores—Standardized Coefficients, Science and Social Studies	6
Table 4: Prediction Regression of Spring 2022 Scores—R ² and F Test	7
Table 5: Number of Corporations Flagged by Prediction Regression Standardized Residuals.....	8
Table 6: Frequency of the Same Corporations Being Flagged across Subjects and Grades.....	9
Table 7: Prediction Regression of Absolute Residual—Standardized Coefficients, ELA	9
Table 8: Prediction Regression of Absolute Residual—Standardized Coefficients, Mathematics	10
Table 9: Prediction Regression of Absolute Residual—Standardized Coefficients, Science and Social Studies	10
Table 10: Prediction Regression of Absolute Residual— R ² and F Test.....	11
Table 11: Prediction Regression of Residual—Standardized Coefficients, ELA	12
Table 12: Prediction Regression of Residual—Standardized Coefficients, Mathematics	12
Table 13: Prediction Regression of Residual—Standardized Coefficients, Science and Social Studies.....	13
Table 14: Prediction Regression of Residual—R ² and F Test.....	13

List of Appendices

Appendix A. Descriptive Statistics
Appendix B. Prediction Regression of Spring 2022 Scores
Appendix C. Prediction Regression of Absolute Residuals
Appendix D. Prediction Regression of Residuals

1. BACKGROUND

For the October 2022 Indiana Technical Advisory Committee (TAC) meeting, CAI proposed conducting a regression study to determine whether unexpected changes were present in corporation performance across administrations, as well as to identify factors related to changes in corporation performance. This document outlines the methodology, analysis results, and recommendations of the regression study when applied to the full test-taking populations of all ILEARN grade-level assessments (English Language Arts Grades 3-8, Mathematics Grades 3-8, Science Grades 4 and 6, and Social Studies Grade 5).

2. PURPOSE, METHODOLOGY, AND DATA

2.1 PURPOSE

The purpose of this study is to evaluate unexpected changes in corporation-level performance between Spring 2021 and Spring 2022 that could indicate irregularities in the Spring 2022 test administration. The findings of the study will mainly be used to identify any areas where corporations may need additional supports or whether they stay at their level, drop, or improve.

It is important to note that during a typical assessment administration, where mean test scores at the corporation level could be expected to be stable over time, this type of evaluation is typically accomplished by comparing corporation-level scale score means between two administrations. However, given the long-term impacts of pandemic-related disruptions in education, it is not expected that corporation achievement levels will be stable between administrations. Post-pandemic, it is necessary to evaluate consistency with respect to greater or lesser than expected declines in performance, following pandemic-related disruptions in education, as well as greater or lesser than expected gains during recovery as for the Spring 2022 administration.

2.2 METHODOLOGY

Given that some change in statewide student achievement is expected for the Spring 2022 test administration during the post-pandemic recovery, the aim of this study is to evaluate the consistency of corporation performance via the following two steps:

1. Identifying expected levels of achievement by regressing Spring 2022 corporation mean scale scores on Spring 2021 corporation mean scale scores and Spring 2022 corporation-level characteristics (e.g., demographic variables); and
2. Determining possible explanations for deviation from predicted performance through further analysis of residuals. This will be done by predicting residuals using corporation characteristics such as corporation size, participation rate, and changes in demographic variables between the two administrations.

This study uses weighted linear regression models for both of these steps. The weights used are sample sizes of Spring 2022 corporations.

It is important to note that weighted regression is a method used when the least squares assumption of constant variance in the residuals (homoscedasticity) is violated, as is the case with this study. As the variance of corporation means is inversely proportional to corporation size, smaller corporations are expected to have larger random fluctuations in performance, resulting in a larger variance in the residuals. Therefore, to avoid violating homoscedasticity, weighted regression is used so that it minimizes the sum of weighted squared residuals to produce residuals with a constant variance.

To achieve homoscedasticity, an analytical weight (sometimes called an inverse variance weight or a regression weight) is used, specifying that the i -th corporation comes from a sub-population with variance σ^2/w_i , where σ^2 is a common variance and w_i is the weight

of the i _th corporation set to the sample size associated with the i _th corporation. This is consistent with what is typically done in meta-analyses, where each "observation" is the mean of a sample and sample size is used as a weight. This way, observations based on varying sample sizes can influence a weighted analysis to the extent that they are precise. In the context of this study, using sample size as a weight allows for the inclusion of all corporations in the analysis while not allowing the large errors associated with small corporations to unduly influence the results.

Weighted standardized residuals are used to identify corporations that deviate from expected levels of achievement. In a usual linear regression, all observations are assumed to have the same standard deviation, so standardized residuals are calculated by dividing all the residuals by the same estimated standard deviation. With a weighted regression, however, the observations may all have different standard deviations, so weighted standardized residuals are calculated by dividing each residual by its "own" estimated standard deviation. This is particularly important for this study because using weighted standardized residuals allows for a flagging criterion that is sensitive to the size of the corporation to be applied. By doing so the over-flagging of small corporations, which tend to have larger residuals, can be avoided.

2.3 DATA

All students who attempted the test and were on grade level during the Spring 2022 and Spring 2021 testing windows were used as data for this study and are included within the analysis. All corporations with data from both Spring 2022 and Spring 2021 were included in the analysis.

Appendix A provides descriptive statistics for all corporation-level variables used in this study, which are defined in brief below.

- Corporation mean: the average of scale scores of eligible students within each corporation.
- Participation rate: the ratio between the observed N count in the Spring 2022 data and the expected N count in Roster Tracking System (RTS).
- Corporation-level demographic variables: percentages of students in important subgroups. The subgroups in this study include
 - Title 1,
 - Special Education,
 - Section 504,
 - English Learner,
 - Female,
 - White,
 - Black/African American,
 - Asian,
 - Hispanic,
 - American Indian/Alaska Native, and
 - Native Hawaiian/Other Pacific Islander.

- Differences in corporation-level demographic variables: differences in corporation-level demographic variables between the Spring 2022 and Spring 2021 test administrations

A few noteworthy patterns that emerge from the descriptive statistics.

Corporation size varied greatly for corporations, with 5th, 50th, 95th percentiles at each grade comprising student populations in the ranges of 5-7, 48-56, and 593-663, respectively. This is important because it is expected that smaller corporations will fluctuate more than larger corporations, purely as a result of large error associated with corporation means. This fluctuation does not necessarily indicate aberrant test administrations.

Participation rates also appeared to be very high, typically around 100%. Corporations with participation rates less than 95% were not excluded from the study in the hope that using participation rate as a predictor might shed light on some underlying factors and potentially explain the fluctuations in corporation performance.

As expected, corporations seemed to show a slight rebound in performance in Spring 2022 compared to Spring 2021. This is likely a result of improved instructional conditions due to reopening of schools under diminishing impacts of the pandemic. This rebound, however, seemed to be limited.

Lastly, corporation-level demographics appeared to be stable between administrations, with some slight differences observed for some subgroups.

3. RESULTS

CAI ran a weighted linear regression model for each assessment to identify expected levels of achievement for corporations in Spring 2022, given their observed achievement levels in Spring 2021. Important corporation-level demographic variables that have often been found to covary with academic achievement in the research literature were included in the model as predictors.

Next, raw residuals from the above regression models were analyzed to understand possible explanations for deviations from expected achievement levels. Raw residuals, rather than weighted residuals, were used, because the purpose of this analysis was to determine possible explanations for deviation from predicted performance, and it could best be accomplished by analyzing raw residuals, which are defined as deviation from predicted performance. Two sets of weighted regressions were run, with the outcome variables being observed residuals and absolute value of residuals, respectively. Both observed residuals and absolute residuals were analyzed, because deviation from predicted performance has two important aspects: (1) the magnitude of the deviation, which is captured by absolute residuals; and (2) the direction of the deviation, or over-versus under-prediction, which is captured by observed residuals. Predicting absolute residuals focuses on the magnitude of the deviation and helps identify factors that lead corporation results to be unreliable or volatile, while predicting observed residuals focuses on the direction of the deviation and helps identify factors related to over- or under-prediction of corporation performance.

Both analyses used a set of predictors that may further explain observed deviations from expected levels of achievement, such as corporation size, participation rate, and changes in demographic variables between the two administrations. Absolute residual was added as a predictor for observed residual. This covaries out the magnitude of the residual while leaving in the directionality of the observed residual, which allows for better understanding of the directional aspect of the residuals.

3.1 PREDICTING SPRING 2022 SCORES

To identify expected levels of achievement, a weighted linear regression model was run for each assessment, using Spring 2022 corporation mean scale scores as the outcome. Predictors included Spring 2021 corporation mean scale scores and Spring 2022 corporation-level demographic variables. Spring 2022 corporation size was used as the weight. The regression model is mathematically specified as follows:

$$SS_{2022} = \beta_0 + \beta_1 * SS_{2021} + \beta_2 * Percent_{Title1} + \beta_3 * Percent_{SpecEd} + \beta_4 * Percent_{Sec504} + \beta_5 * Percent_{ELL} + \beta_6 * Percent_{Female} + \beta_7 * Percent_{White} + \beta_8 * Percent_{Black} + \beta_9 * Percent_{Asian} + \beta_{10} * Percent_{Hispanic} + \beta_{11} * Percent_{AmeriIndian} + \beta_{12} * Percent_{Pacific}$$

Tables 1 through 3 present standardized coefficient estimates of predictors for the Spring 2022 corporation mean scores by subject. Only significant effects ($p < 0.05$) are shown here to depict patterns across grade-levels and/or subject-area assessments. Appendix B shows the regression model parameter estimates of the predictors for the Spring 2022

corporation mean scores, including standardized and unstandardized coefficients, the standard error of the unstandardized coefficient, and p value regardless of significance level. Tables 4 presents R^2 , adjusted R^2 , and corresponding F test statistics.

Table 1: Prediction Regression of Spring 2022 Scores–Standardized Coefficients, ELA

Predictors	G3E	G4E	G5E	G6E	G7E	G8E
Mean [Sp21]	0.76	0.76	0.73	0.75	0.73	0.72
Title1 [Sp22]	-0.19	-0.11	-0.15	-0.15	-0.1	-0.14
Special Education [Sp22]	-	-0.14	-0.12	-0.17	-0.13	-0.21
Section 504 [Sp22]	0.03	-	-	0.04	-	-
English Learner [Sp22]	-	-0.1	-0.08	-0.07	-	-
Female [Sp22]	-	-	-	-	-	-
White [Sp22]	-	-	-	-	0.23	-
Black / African American [Sp22]	-	-	-	-	-	-
Asian [Sp22]	-	-	-	0.07	0.12	0.08
Hispanic [Sp22]	-	-	-	-	0.14	-
American Indian / Alaska Native [Sp22]	-	-	-	-	-	-
Native Hawaiian / Other Pacific Islander [Sp22]	-	-	-	-	0.05	-

Table 2: Prediction Regression of Spring 2022 Scores–Standardized Coefficients, Mathematics

Predictors	G3M	G4M	G5M	G6M	G7M	G8M
Mean [Sp21]	0.8	0.75	0.78	0.8	0.79	0.81
Title1 [Sp22]	-0.19	-0.06	-0.13	-0.15	-0.13	-0.11
Special Education [Sp22]	-	-0.14	-0.09	-0.11	-0.07	-0.17
Section 504 [Sp22]	-	-	-	-	-	-
English Learner [Sp22]	-	-0.08	-	-	-	-
Female [Sp22]	-	-0.06	-	-	-0.06	-0.1
White [Sp22]	-	-	-	-	-	-
Black / African American [Sp22]	-	-	-	-	-	-
Asian [Sp22]	-	-	-	-	0.09	0.06
Hispanic [Sp22]	-	-	-	-	-	-
American Indian / Alaska Native [Sp22]	-	-	-	-	-	-
Native Hawaiian / Other Pacific Islander [Sp22]	-	-0.03	-	-	0.03	-

Table 3: Prediction Regression of Spring 2022 Scores–Standardized Coefficients, Science and Social Studies

Predictors	G4S	G6S	G5SS
Mean [Sp21]	0.7	0.75	0.76
Title1 [Sp22]	-0.14	-0.16	-0.17
Special Education [Sp22]	-0.14	-0.15	-0.08
Section 504 [Sp22]	0.03	0.03	0.04

Predictors	G4S	G6S	G5SS
English Learner [Sp22]	-0.12	-0.06	-0.12
Female [Sp22]	-0.06	-	-
White [Sp22]	-	-	-
Black / African American [Sp22]	-	-	-
Asian [Sp22]	0.07	-	0.07
Hispanic [Sp22]	-	-	-
American Indian / Alaska Native [Sp22]	-	-	-
Native Hawaiian / Other Pacific Islander [Sp22]	-0.05	-	-

Table 4: Prediction Regression of Spring 2022 Scores—R² and F Test

Subject	Grade	R ²	Adjusted R ²	F	F (Num. DF)	F (Den. DF)	P-value
ELA	3	0.88	0.88	381.9	12	623	<0.001
	4	0.89	0.88	407.1	12	624	<0.001
	5	0.89	0.88	404.4	12	624	<0.001
	6	0.86	0.86	321.5	12	609	<0.001
	7	0.89	0.89	407.8	12	597	<0.001
	8	0.87	0.87	328.6	12	594	<0.001
Mathematics	3	0.89	0.89	416.5	12	623	<0.001
	4	0.90	0.90	476.8	12	624	<0.001
	5	0.89	0.89	423.8	12	624	<0.001
	6	0.88	0.88	380.0	12	609	<0.001
	7	0.91	0.90	474.6	12	597	<0.001
	8	0.87	0.87	338.2	12	593	<0.001
Science	4	0.91	0.91	539.7	12	624	<0.001
	6	0.89	0.89	428.7	12	607	<0.001
Social Studies	5	0.89	0.89	435.5	12	623	<0.001

Taken altogether, the proportions of variance (R^2) in the outcome variable accounted for by the predictors for public corporations range from 0.86 to 0.89 for ELA, 0.87 to 0.91 for Mathematics, 0.89 to 0.91 for Science, and 0.89 for Social Studies. This suggests that corporation achievement levels in Spring 2022 can be predicted with high levels of accuracy by the corporation’s prior achievement and characteristics.

The standardized regression coefficients show that, as expected, Spring 2021 corporation-level achievement is the strongest predictor for future achievement across assessments. Title 1 seems to be a statistically significant predictor for all assessments, while special education seems to show statistical significance for many assessments. This suggests that, in addition to prior achievement, corporations with higher percentages of Title 1 and special education students showed lower corporation level achievement consistently across grades and subjects. Corporations with higher percentages of Black/African American students performed less well in some assessments. ELL, female,

and Asian subgroups show statistical significance for some assessments. The effects seem to be small to moderate as compared to other predictors.

To identify corporations that deviated from expected levels of achievement, weighted standardized residuals were generated based on the linear prediction regression models. Table 5 shows the number of corporations flagged by weighted standardized residuals greater than 3 and 4, respectively. Note that flagged corporations only indicate that a corporation is not performing as expected in relation to the performance of all other corporations in the state. Generally, very few corporations were flagged, indicating most corporations performed as expected in Spring 2022. Specifically, between zero and three corporations were identified as having weighted standardized residuals greater than four, and between one and eight corporations were identified as having weighted standardized residuals greater than three. It is noteworthy that with the use of weighted standardized residuals, as expected, the residuals required for flagging were smaller for larger corporations and larger for smaller corporations.

Table 5: Number of Corporations Flagged by Prediction Regression Standardized Residuals

Subject	Grade	N Corps	N Corps Flagged by Std. Residual > 4	N Corps Flagged by Std. Residual > 3
ELA	3	636	0	6
	4	637	0	5
	5	637	0	4
	6	622	1	2
	7	610	0	4
	8	607	1	5
Mathematics	3	636	3	4
	4	637	1	4
	5	637	3	8
	6	622	1	5
	7	610	1	5
	8	606	2	7
Science	4	637	0	1
	6	620	0	2
Social Studies	5	636	2	5

Flagged corporations were also evaluated to see whether the same corporations were flagged for multiple grades or multiple subjects. Table 6 provides a summary of the number of times the same corporations were flagged by using weighted standardized residual greater than 4 and 3, respectively. When a weighted standardized residual of 4 is used, only 3 corporations were flagged more than once. When a weighted standardized residual of 3 is used, 13 corporations were flagged more than once.

Table 6: Frequency of the Same Corporations Being Flagged across Subjects and Grades

N Flags	Corporations	
	Weighted Std. Residual > 4	Weighted Std. Residual > 3
1	7	21
2	2	11
3	0	2
4	1	0
5	0	0
6	0	0
7	0	0
Total	10	34

3.2 PREDICTING THE MAGNITUDE OF SPRING 2022 SCORE RESIDUALS

To identify possible explanations for the *magnitude* of deviation from predicted performance, a linear regression model was run for each, by using the *absolute* value of the residuals from the prediction model as the outcome and corporation characteristics (e.g., corporation size, participation rate, and the *absolute* value of changes in demographic variables between the two administrations) as predictors. The rationale for this analysis is that (1) small sample sizes result in more error (e.g., measurement error and sampling error) associated with corporation means, thus leading to larger residuals; and (2) significant changes to the corporation population would also be expected to make the regression model fit less well, resulting in larger residuals. The regression model is mathematically specified as follows:

$$\begin{aligned}
 AbsRes_{2022} = & \beta_0 + \beta_1 * ParticipationRate_{2022} + \beta_2 * N_{2022} + \beta_3 * AbsDiff_{Title1} + \beta_4 \\
 & * AbsDiff_{SpecEd} + \beta_5 * AbsDiff_{Sec504} + \beta_6 * AbsDiff_{ELL} + \beta_7 \\
 & * AbsDiff_{Female} + \beta_8 * AbsDiff_{White} + \beta_9 * AbsDiff_{Black} + \beta_{10} \\
 & * AbsDiff_{Asian} + \beta_{11} * AbsDiff_{Hisp} + \beta_{12} * AbsDiff_{AmeriIndian} + \beta_{13} \\
 & * AbsDiff_{Pacific}
 \end{aligned}$$

Tables 7 through 9 present standardized coefficient estimates of the predictors for residuals by subject. Only significant effects ($p < 0.05$) are shown here to depict patterns across grade-level and/or subject-area assessments. Appendix C shows the regression model parameter estimates of the predictors for residuals, including standardized and unstandardized coefficients, the standard error of the unstandardized coefficient, and p value regardless of significance level. Tables 10 presents R^2 , adjusted R^2 , and corresponding F test statistics.

Table 7: Prediction Regression of Absolute Residual—Standardized Coefficients, ELA

Predictors	G3E	G4E	G5E	G6E	G7E	G8E
Participation Rate	-	-	-	-0.13	-	-
N [Sp22]	-0.08	-0.06	-0.04	-0.04	-0.07	-0.06
Title1 [Abs Diff]	0.14	0.12	-	-	-	0.14
Special Education [Abs Diff]	-	0.14	0.17	-	-	-
Section 504 [Abs Diff]	-	-	-	-	-	-
English Learner [Abs Diff]	-	-	-	-	-	-
Female [Abs Diff]	-	0.17	0.15	0.14	0.17	0.14
White [Abs Diff]	-	-	0.12	-	-	0.2
Black / African American [Abs Diff]	0.19	0.11	0.18	-	-	0.15
Asian [Abs Diff]	-	-	-	0.12	-	-
Hispanic [Abs Diff]	-	-	-	0.24	-	-
American Indian / Alaska Native [Abs Diff]	-	-	-	-	-	-
Native Hawaiian / Other Pacific Islander [Abs Diff]	-	-	-	-	-	0.08

Table 8: Prediction Regression of Absolute Residual—Standardized Coefficients, Mathematics

Predictors	G3M	G4M	G5M	G6M	G7M	G8M
Participation Rate	-	-	-	-0.14	-	-
N [Sp22]	-0.07	-0.06	-0.04	-0.05	-0.05	-0.07
Title1 [Abs Diff]	0.12	-	-	-	-	-
Special Education [Abs Diff]	-	0.1	0.15	0.1	-	-
Section 504 [Abs Diff]	-	-	-	-	0.09	-
English Learner [Abs Diff]	-	-	-	-	-	0.11
Female [Abs Diff]	-	0.11	0.17	-	0.19	0.19
White [Abs Diff]	0.21	0.14	0.19	-	-	0.23
Black / African American [Abs Diff]	0.31	0.15	0.2	0.21	0.16	-
Asian [Abs Diff]	-	-	-	0.11	-	-
Hispanic [Abs Diff]	-	-	-	0.18	-	-
American Indian / Alaska Native [Abs Diff]	0.09	-	-	-	-	-
Native Hawaiian / Other Pacific Islander [Abs Diff]	-	-	-	-	-	-

Table 9: Prediction Regression of Absolute Residual—Standardized Coefficients, Science and Social Studies

Predictors	G4S	G6S	G5SS
Participation Rate	-	-	-
N [Sp22]	-0.06	-0.05	-0.06
Title1 [Abs Diff]	0.12	-	0.11
Special Education [Abs Diff]	0.11	0.09	0.14
Section 504 [Abs Diff]	-	-	-

Predictors	G4S	G6S	G5SS
English Learner [Abs Diff]	-	-	-
Female [Abs Diff]	0.13	0.1	0.17
White [Abs Diff]	-	-	0.19
Black / African American [Abs Diff]	0.11	0.13	-
Asian [Abs Diff]	-	-	-
Hispanic [Abs Diff]	-	-	-
American Indian / Alaska Native [Abs Diff]	-	-	-
Native Hawaiian / Other Pacific Islander [Abs Diff]	-	-	-

Table 10: Prediction Regression of Absolute Residual— R^2 and F Test

Subject	Grade	R^2	Adjusted R^2	F	F (Num. DF)	F (Den. DF)	P-value
ELA	3	0.26	0.25	17.2	13	622	<0.001
	4	0.26	0.25	17.0	13	623	<0.001
	5	0.29	0.28	20.0	13	623	<0.001
	6	0.23	0.21	14.1	13	608	<0.001
	7	0.25	0.23	15.1	13	596	<0.001
	8	0.23	0.22	13.9	13	593	<0.001
Mathematics	3	0.32	0.30	22.1	13	622	<0.001
	4	0.28	0.26	18.4	13	623	<0.001
	5	0.24	0.23	15.3	13	623	<0.001
	6	0.27	0.25	17.2	13	608	<0.001
	7	0.24	0.22	14.5	13	596	<0.001
	8	0.21	0.19	12.0	13	592	<0.001
Science	4	0.29	0.28	19.6	13	623	<0.001
	6	0.24	0.22	14.4	13	606	<0.001
Social Studies	5	0.30	0.29	20.6	13	622	<0.001

Taken altogether, the proportions of variance in absolute residuals accounted for by the predictors ranges from 0.23 to 0.29 for ELA, 0.21 to 0.32 for Mathematics, 0.24 to 0.29 for Science, and 0.30 for Social Studies. To reiterate, the purpose of predicting absolute residuals is to identify factors that lead corporation results to be unreliable or volatile.

Note that all corporations were included in this study, allowing for a wide range in corporation size. As a result, corporation means based on these varying corporation sizes vary greatly in the amount of error associated with them, thus affecting prediction accuracy. Corporation size seems to show substantial negative coefficients for all assessments. This suggests that smaller corporations tend to be associated with large absolute residuals, which is consistent with the expectation that smaller corporations tend to fluctuate more in performance over time than larger corporations due to the larger errors associated with corporation means. Similarly, absolute changes in demographics seem to show large positive coefficients, suggesting that the larger the changes in

corporation-level demographics, the larger the deviation between observed performance and predicted performance. This is also consistent with expectations.

3.3 PREDICTING THE DIRECTION OF SPRING 2022 SCORE RESIDUALS

To identify possible explanations for the direction of deviation from predicted performance, a linear regression model was run for each assessment and separately for public and non-public corporations by using the residuals from the prediction model as the outcome and corporation characteristics (e.g., corporation size, participation rate, and changes in demographic variables between the two administrations) and absolute residual as predictors. The regression model is mathematically specified as follows:

$$Res_{2022} = \beta_0 + \beta_1 * ParticipationRate_{2022} + \beta_2 * N_{2022} + \beta_3 * AbsRes_{2022} + \beta_4 * Diff_{Title1} + \beta_5 * Diff_{SpecEd} + \beta_6 * Diff_{Sec504} + \beta_7 * Diff_{ELL} + \beta_8 * Diff_{Female} + \beta_9 * Diff_{White} + \beta_{10} * Diff_{Black} + \beta_{11} * Diff_{Asian} + \beta_{12} * Diff_{Hisp} + \beta_{13} * Diff_{AmeriIndian} + \beta_{14} * Diff_{Pacific}$$

Tables 11 through 13 present standardized coefficient estimates of the predictors for residuals. Only significant effects ($p < 0.05$) are shown here to depict patterns across grade-level and/or subject-area assessments. Appendix D shows the regression model parameter estimates of the predictors for residuals, including standardized and unstandardized coefficients, the standard error of the unstandardized coefficient, and p value regardless of significance level. Table 14 presents R^2 , adjusted R^2 , and corresponding F test statistics.

Table 11: Prediction Regression of Residual—Standardized Coefficients, ELA

Predictors	G3E	G4E	G5E	G6E	G7E	G8E
Participation Rate	0.11	-	0.08	0.23	-	0.14
N [Sp22]	-	-	-	0.02	-	-
Absolute Residual	-	-	-	-	-	-
Title1 [Diff]	-	-	-	-	-	-
Special Education [Diff]	-0.14	-0.18	-0.16	-0.17	-0.16	-0.28
Section 504 [Diff]	-	-	-0.09	-	-	-
English Learner [Diff]	-	-	-	-	-	-
Female [Diff]	-	-	-	-	-	-
White [Diff]	-	-	-	-	-	-
Black / African American [Diff]	-	-	-0.15	-	-	-0.24
Asian [Diff]	-	-	-	-	-	-
Hispanic [Diff]	-	-	-	-	-	-
American Indian / Alaska Native [Diff]	-	-	-	-	-	-
Native Hawaiian / Other Pacific Islander [Diff]	-0.09	-	0.07	-	-	-

Table 12: Prediction Regression of Residual—Standardized Coefficients, Mathematics

Predictors	G3M	G4M	G5M	G6M	G7M	G8M
Participation Rate	-	0.15	0.09	0.16	0.13	-
N [Sp22]	-	-	-	0.02	-	-
Absolute Residual	-	-	-	-	-	-
Title1 [Diff]	-	-	-	-	0.1	-
Special Education [Diff]	-0.14	-0.2	-0.19	-0.12	-0.21	-0.2
Section 504 [Diff]	-	-	-0.13	-	-0.09	-0.09
English Learner [Diff]	-	-	-	-	-	-
Female [Diff]	-	-	-	-	-	-
White [Diff]	0.19	-	-	-	-	-
Black / African American [Diff]	-	-	-0.21	-0.24	-	-0.29
Asian [Diff]	-	-	-	-	-	-
Hispanic [Diff]	-	-	-	-	-	-
American Indian / Alaska Native [Diff]	-	-	-	-	-	-
Native Hawaiian / Other Pacific Islander [Diff]	-	-	-	-	-	-

Table 13: Prediction Regression of Residual—Standardized Coefficients, Science and Social Studies

Predictors	G4S	G6S	G5SS
Participation Rate	0.14	-	-
N [Sp22]	-	0.02	-
Absolute Residual	-	-	-
Title1 [Diff]	-	0.1	-
Special Education [Diff]	-0.13	-0.11	-0.09
Section 504 [Diff]	-	-	-0.08
English Learner [Diff]	-	-	-
Female [Diff]	-	-	-
White [Diff]	-	-	-
Black / African American [Diff]	-	-0.2	-
Asian [Diff]	-	-	-
Hispanic [Diff]	-	-	-
American Indian / Alaska Native [Diff]	-	-	-
Native Hawaiian / Other Pacific Islander [Diff]	-	-	0.08

Table 14: Prediction Regression of Residual— R^2 and F Test

Subject	Grade	R^2	Adjusted R^2	F	F (Num. DF)	F (Den. DF)	P-value
ELA	3	0.06	0.04	2.7	14	621	<0.001
	4	0.05	0.03	2.4	14	622	<0.01
	5	0.07	0.05	3.3	14	622	<0.001
	6	0.07	0.05	3.5	14	607	<0.001
	7	0.04	0.01	1.6	14	595	0.080
	8	0.08	0.06	3.9	14	592	<0.001
Mathematics	3	0.09	0.07	4.6	14	621	<0.001
	4	0.09	0.07	4.4	14	622	<0.001
	5	0.13	0.11	6.6	14	622	<0.001
	6	0.07	0.05	3.4	14	607	<0.001
	7	0.06	0.04	2.9	14	595	<0.01
	8	0.05	0.03	2.3	14	591	<0.01
Science	4	0.06	0.04	2.8	14	622	<0.001
	6	0.07	0.05	3.4	14	605	<0.001
Social Studies	5	0.05	0.02	2.1	14	621	<0.05

Taken altogether, the proportions of variance in residuals accounted for by the predictors range from 0.04 to 0.08 for ELA, 0.05 to 0.13 for Mathematics, 0.06 to 0.07 for Science, and 0.05 for Social Studies. These results are not surprising. The very large R^2 for the prediction model as specified in Section 3.1 indicates that the prediction model fits so well that variation in residuals is very small. The F tests for the overall models were statistically significant for most assessments.

An inspection of the standardized coefficients reveals some noteworthy patterns. Differences in percentages of special education, Section 504 and Black/African American students tend to show negative coefficients. This indicates that for these assessments, corporations with increases in percentages of special education, Section 504 or Black/African American students tend to perform worse than expected. Participation rate seems to show positive coefficients, suggesting that corporations with higher participation rates tend to perform better than expected. Given the very small R^2 s, however, it is advisable not to over-interpret these coefficients, even though they are statistically significant.

These results, combined with the results based on absolute residuals, appear to further indicate that deviations from expected performance, with respect to both direction and magnitude, are primarily due to small sample sizes and changes in the tested population.

4. SUMMARY

This study aimed to evaluate unexpected changes in corporation-level performance between Spring 2022 and Spring 2021 that could indicate irregularities in the Spring 2022 test administration. In this study, Spring 2022 corporation performance was predicted by using Spring 2021 corporation performance and corporation-level characteristics. Residuals from these prediction regression models were further examined by using predictors such as participation rate, corporation size, and changes in corporation demographic variables.

Results from this study suggest that Spring 2022 corporation-level achievement can be predicted with remarkably high levels of accuracy based on a corporation's prior achievement and characteristics. Only a few corporations were flagged for deviating from expected levels of achievement by weighted standardized residuals greater than 4 or 3. Some corporations were found to be flagged more than once across subjects and grades.

It is important to note that corporations flagged for potential irregularities only indicate that a corporation is not performing as expected in relation to the performance of all other corporations in the state. This could be attributed to a number of factors, which may or may not be reflected in the variables collected in the study. For the variables collected in the study, it was found that deviations from expected performance were related to both small sample sizes and shifts in the tested population for corporations. For the variables not collected in the study, educational practice, for example, could also lead to such deviations. In the context of post-pandemic K-12 education, corporations varied greatly in their capabilities to reopen school buildings, operate safely, and provide effective instruction with the lingering effects of the pandemic. The variation in the degree to which instruction can be delivered effectively without disruption may cause some corporations to perform better than expected and other corporations to perform worse than expected. IDOE noted that the intent of flagging was to try to identify any areas where corporations may need additional supports or whether they stay at their level, drop, or improve.

While the use of weighted standardized residuals avoids over-flagging small corporations, corporations may still be flagged due to shifts in the tested population. It is therefore recommended that, when flagging corporations for unexpected shifts in student performance, IDOE first evaluate any shifts in the distribution of the student population. These changes may cause unexpected changes in student performance. Where student populations are stable, IDOE may want to gather as much information regarding the corporations' educational practices as possible to understand the context of the unexpected changes in performance. Unexpected changes in performance may be associated with testing irregularities, if and only if all other plausible explanations are ruled out.