Indiana Learning Evaluation Readiness Network (ILEARN)

2018–2019

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 American Institutes for Research (AIR): Stephan Ahadi, Elizabeth Ayers-Wright, Kevin Clayton, and Xiaoxin Wei. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and Program Leads.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

1. INTRODUCTION

The ILEARN 2018–2019 technical report is provided to document and make transparent all methods used in item development, test construction, psychometric methods, standard setting, score reporting methods, creating summaries of student assessment results, and supporting evidence for intended uses and interpretations of the test scores. The technical report is presented as seven 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) *Standard Setting*. This volume documents the methods and results of the standard setting process.
(7) *Special Studies*. This volume compiles any special studies conducted; it is updated annually to reflect studies relevant to the respective administration. If no special studies were conducted, the volume is not published.

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

1.1 BACKGROUND AND HISTORICAL CONTEXT OF TESTS

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

1.2 PURPOSE AND INTENDED USES OF THE ILEARN ASSESSMENTS

ILEARN is comprised of criterion-referenced tests that apply principles of evidence-centered design to yield overall and reporting-category-level test scores at the student level and at other levels of aggregation that reflect student performance of the IAS. ILEARN supports instruction and student learning by providing immediate feedback to educators and parents which can be used to inform instructional strategies that remediate or enrich instruction. An array of reporting metrics allows achievement to be monitored at both student and aggregate levels and growth to be measured at both student and group levels over time.

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

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 THE DEVELOPMENT AND ANALYSIS OF ILEARN

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

Indiana Department of Education

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

Indiana Educators

Indiana educators participate in most aspects of the conceptualization and development of ILEARN. Educators participate in the development of the academic standards, clarification of how these standards will be assessed, creation of blueprints and test design, and committee reviews of test items and passages.

Technical Advisory Committee

IDOE convenes a panel three times a year to discuss psychometric, test development, administrative, and policy issues relevant to current and future Indiana assessments. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from multiple Indiana school corporations.

American Institutes for Research

AIR is the current vendor for assessment testing and was selected through the state-mandated competitive procurement process. In the Winter of 2017, AIR became the primary party responsible for developing test content, building test forms, conducting psychometric

analyses, administering and scoring test forms, and reporting test results for ILEARN described in this report. Additionally, AIR is responsible for developing and maintaining the ILEARN bank, which is used for test construction.

Assessment Systems Corporation

For the ILEARN assessments, the Assessment Systems Corporation conducts independent verifications of scoring activities for all grades and subjects and blueprint checks for the adaptive assessments.

1.4 Available Test Formats and Special Versions

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

1.5 Student Participation

All Indiana public school students in ELA and Mathematics grades 3–8, Science grades 4, 6, and end-of-course Biology, Social Studies grade 5, and end-of-course U.S. Government can participate in the state assessments. Table 2 shows the number of students tested and the number of students reported in the 2018-2019 ILEARN by grade and subject area. Table 3 through Table 6 present the distribution of students, in counts and percentages. The subgroup categories reported here are gender, ethnicity, students with special education (SPED), Section 504, and English Learners.

Table 2: Number of Students Participating in ILEARN 2018–2019

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	83096	83074	3	83111	83080						
4	84175	84147	4	84183	84144	4	84107	84068			
5	86407	86381	5	86420	86369				5	86274	86253
6	85880	85833	6	85895	85817	6	85710	85659			
7	84669	84591	7	84692	84580						
8	83079	82991	8	83066	82991						
						Biology	81179	80677	U.S. Government	1245	1230

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	English Learner
3	*N*	83096	42614	40482	54763	10486	2290	10851	128	76	4502	13764	1858	7866
	%	100	51.28	48.72	65.90	12.62	2.76	13.06	0.15	0.09	5.42	16.56	2.24	9.47
4	*N*	84175	42792	41383	55652	10506	2203	11230	135	63	4386	13738	2208	7517
	%	100	50.84	49.16	66.11	12.48	2.62	13.34	0.16	0.07	5.21	16.32	2.62	8.93
5	*N*	86407	43946	42461	57277	10851	2133	11500	137	66	4443	13875	2333	5790
	%	100	50.86	49.14	66.29	12.56	2.47	13.31	0.16	0.08	5.14	16.06	2.7	6.7
6	*N*	85880	43755	42125	57464	10518	1964	11454	146	68	4266	13003	2529	3721
	%	100	50.95	49.05	66.91	12.25	2.29	13.34	0.17	0.08	4.97	15.14	2.94	4.33
7	*N*	84669	43323	41346	57150	10243	2086	10839	164	70	4117	12447	2244	2987
	%	100	51.17	48.83	67.50	12.10	2.46	12.80	0.19	0.08	4.86	14.7	2.65	3.53
8	N	83079	42457	40622	56976	9777	1905	10289	160	73	3899	12085	2244	2796
	%	100	51.10	48.90	68.58	11.77	2.29	12.38	0.19	0.09	4.69	14.55	2.7	3.37

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	English Learner
3	*N*	83111	42615	40496	54758	10489	2290	10866	128	77	4503	13772	1860	7881
	%	100	51.27	48.73	65.89	12.62	2.76	13.07	0.15	0.09	5.42	16.57	2.24	9.48
4	*N*	84183	42792	41391	55655	10499	2204	11242	135	63	4385	13760	2197	7534
	%	100	50.83	49.17	66.11	12.47	2.62	13.35	0.16	0.07	5.21	16.35	2.61	8.95
5	*N*	86420	43951	42469	57274	10852	2134	11513	137	66	4444	13886	2331	5811
	%	100	50.86	49.14	66.27	12.56	2.47	13.32	0.16	0.08	5.14	16.07	2.7	6.72
6	*N*	85895	43763	42132	57461	10514	1964	11473	146	68	4269	13035	2523	3735
	%	100	50.95	49.05	66.9	12.24	2.29	13.36	0.17	0.08	4.97	15.18	2.94	4.35
7	*N*	84692	43337	41355	57166	10231	2084	10855	164	70	4122	12459	2238	3008
	%	100	51.17	48.83	67.5	12.08	2.46	12.82	0.19	0.08	4.87	14.71	2.64	3.55
8	N	83066	42452	40614	56963	9751	1906	10309	162	73	3902	12063	2243	2820
	%	100	51.11	48.89	68.58	11.74	2.29	12.41	0.20	0.09	4.70	14.52	2.7	3.39

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	English Learner
4	*N*	84107	42749	41358	55612	10484	2204	11232	135	63	4377	13772	2203	7525
	%	100	50.83	49.17	66.12	12.47	2.62	13.35	0.16	0.07	5.20	16.37	2.62	8.95
6	*N*	85710	43654	42056	57372	10474	1962	11441	146	68	4247	13001	2523	3716
	%	100	50.93	49.07	66.94	12.22	2.29	13.35	0.17	0.08	4.96	15.17	2.94	4.34
Biology	*N*	81179	41474	39705	55935	9041	2209	10247	152	54	3541	10407	2120	3448
	%	100	51.09	48.91	68.9	11.14	2.72	12.62	0.19	0.07	4.36	12.82	2.61	4.25

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	English Learner
4	*N*	86274	43864	42410	57224	10797	2134	11480	136	66	4437	13885	2337	5785
	%	100	50.84	49.16	66.33	12.51	2.47	13.31	0.16	0.08	5.14	16.09	2.71	6.71
U.S. Government	*N*	1245	690	555	848	179	13	160	3	0	42	161	15	41
	%	100	55.42	44.58	68.11	14.38	1.04	12.85	0.24	0	3.37	12.93	1.20	3.29

2. SUMMARY OF OPERATIONAL PROCEDURES

2.1 ADMINISTRATION PROCEDURES

Table 7 shows the testing window schedule for the 2018–2019 ILEARN administrations by assessment.

Table 7: 2018–2019 ILEARN Testing Windows

Assessment	Grade/Subject	Mode	Testing Window
ILEARN	ELA 3–8 Mathematics 3–8 Science 4 & 6 Social Studies 5	Online	April 22–May 17, 2019
		Paper	April 22–May 10, 2019
	Biology	Online	December 4 – December 20, 2018 (Fall window)
		Online Paper	February 11–February 28, 2019 (Winter window)
		Online	April 22–May 24, 2019
		Paper	April 22–May 17, 2019
	U.S. Government	Online	April 22–May 24, 2019
		Paper	April 22–May 17, 2019

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

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

2.2 UNIVERSAL FEATURES, DESIGNATED FEATURES, AND ACCOMMODATIONS

Accessibility supports discussed within this document include both embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features that are available to all students as they access instructional or assessment content, designated features that are available to students for whom a need has been identified by an informed educator or team of educators, and accommodations that are generally available for

students for whom there is documentation on an Individualized Education Program (IEP), Section 504 Plan, or Individual Language Plan (ILP).

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

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

The test administrators and school test coordinators in Indiana are responsible for ensuring that arrangements for accommodations are made before the test administration dates. The available accommodation options for eligible students include braille, American Sign Language, closed captioning, streamline, assistive technology (e.g., adaptive keyboards, touch screen, switches), calculation device, print-on-demand, multiplication table, and scribe. Detailed descriptions for each of these accommodations can be found in Appendix J of Volume 5.

3. ITEM BANK AND TEST CONSTRUCTION

3.1 OVERVIEW OF ITEM DEVELOPMENT

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

3.2 FIELD TESTING

The 2019 ILEARN test forms contained newly developed field test items. The ELA and Mathematics ILEARN test forms also contained a collection of items from MetaMetrics used to establish a link with MetaMetrics Lexile and Quantile scales in ELA and Mathematics, respectively. The EFT slots are embedded in segments for adaptive ELA and Mathematics forms and in fixed positions across fixed-form test forms in all subjects, such that item location and motivation effects, if they exist, would not propagate into the estimates of the item parameters. To obtain high-quality responses to the EFT items, students were unaware of which items were operational and which were EFT. Items licensed from Smarter, ICCR, and Hawaii were commonly used for scoring across all adaptive and fixed-form test forms.

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

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

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

Thus, field test item exposure rates depend on the number of field test slots and the number of field test items in the segment. AIR confirmed expected exposure rates after the administration.

3.3 OPERATIONAL FORM CONSTRUCTION

Items from licensed item banks and previous ISTEP+ legacy items were marked as operational. In some instances, it was necessary to use newly developed custom Indiana items to meet blueprint. These items were marked as operational-field test and went through an expedited educator review before being used to score students.

Operational test form development (see Volume 2) includes an array of item types used to measure the IAS. Table 8 describes the item types used in the operational forms that were developed during the operational form construction, and Table 9 through Table 12 show the number of items by item type. The description and examples for each of the item types are also provided in Appendix E of Volume 2.

Table 8: ILEARN Item Types and Descriptions

Response Type	Description
Edit Task with Choice (ETC)*	Student identifies an incorrect word or phrase and chooses the replacement from a number of options.
Equation Response (EQ)	Student is directed to enter an equation, number, fraction, or expression.
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. Both with and without passage.
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 drag-and-drop feature to use text to support an analysis or make an inference.
Table Matching (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Multiple Choice (MC)	Student selects one correct answer from a number of options.
Multi 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.
Table Input (TI)	Student is directed to respond in a table.
Text Entry (TE)	Student is directed to type their response in a text box.
Simulation (SIM)	Student selects inputs to “run” trials. Data is presented in a table after trials are run. A simulation is typically only used within a Performance Task.
Extended Response (ER)	Student is directed to provide a longer, written response.

*Note: Three legacy ISTEP IC items were approved for inclusion in the pool by IDOE content specialists; however, AIR did not develop any custom IC items for ELA.

Table 9: ELA Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
TE	25	25	28	18	35	36
ETC	-	-	1	-	-	1
EBSR	59	32	30	40	21	31
HT	40	40	37	21	49	44
MI	21	13	7	11	3	8
MC	173	151	121	93	141	154
MS	72	48	60	40	64	75

ER	3	3	3	3	3	3

Table 10: Mathematics Operational Items by Item Type and Grade

Item Type	3	4	5	6	7	8
TE	6	9	5	6	4	10
EQ	259	286	236	291	316	112
GI	66	44	23	71	43	59
MI	33	73	73	57	37	69
MC	117	90	75	84	82	96
MS	10	8	15	82	90	63
TI	2	16	6	18	2	5

Table 11: Science Operational Items by Item Type and Grade

Item Type	4	6	Biology
TE	2	1	1
ETC	2	1	0
EQ	1	2	0
GI	-	1	14
HT	1	3	-
MI	-	3	1
MC	36	37	70
MS	6	10	3
PT*	1	2	2
TI	3	2	1

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

Table 12: Social Studies Operational Items by Item Type and Grade

Item Type	5	U.S. Government
TE	4	-
EBSR	-	18
MC	50	13
MS	-	22

Prior to the operational testing window for adaptive tests, AIR employs 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 provides more details about the simulation approach and results.

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

4. CLASSICAL ANALYSES OVERVIEW

4.1 CLASSICAL ITEM ANALYSES

IDOE and the AIR psychometricians collectively monitored the behavior of items while test forms were administered in the live environment. This was accomplished using AIR's quality monitoring system, which yielded an item-analysis report on the performance of test items throughout the testing window. During administration of the 2018–2019 ILEARN, this system served as a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine the performance of test items, this report generated classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation. The report is configurable and was produced to flag only items with statistics falling outside a specified range or to generate reports based on all items in the pool. A minimum sample of 200 responses (Zwick, 2012) per item was applied for classical item analyses. The criteria for flagging and reviewing items is provided in Table 13, and a description of the statistics is provided below.

Table 13: Thresholds for Flagging Items in Classical Item Analysis

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

* IDOE made the decision to forego committee review for any item with an adjusted biserial/polyserial correlation less than 0.10. AIR shared these items with IDOE to make final determinations.

4.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those examinees who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for multiple-choice items was calculated as the correlation between the item score and the ability estimate for students. Biserial correlations for operational items can be found in Appendix A. Most of the operational items had a higher biserial correlation than the flagging criteria. Across all tested grades, less than 4% of ELA operational items, less than 3% of Mathematics

operational items, less than 5% of Science operational items, and less than 7% of Social Studies operational items were flagged. Items with low biserial correlations were reviewed by AIR content experts, and all items behaved as expected.

4.1.2 Distractor Analysis

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

4.1.3 Item Difficulty

Items that were either extremely difficult or extremely easy were flagged for review but were not necessarily removed if they were grade-level appropriate and aligned with the test specifications. For MC items, the proportion of students in the sample selecting the correct answer (the *p*-value) was computed in addition to the proportion of students selecting incorrect responses. For constructed-response items, item difficulty was calculated using the item's relative mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score divided by the maximum possible score points). Conventional item *p*-values are summarized in Section 4.3. The *p*-values for operational items can be found in Appendix A. Most of the operational items had *p*-values within the expected range. Across all tested grades and subjects, less than 1% of operational items were flagged. Flagged items were verified by AIR content experts and psychometricians reported that all items behaved as expected.

4.1.4 Mean Total Score

For multi-point constructed-response items, mean total score was calculated using the item's relative mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score divided by the maximum possible score points). Items were flagged when the proportion of students in any score point category was greater than 0.95. In addition, constructed-response items were flagged if the average ability estimate of students in a score-point category was lower than the average ability estimate of students in the next lower score-point category. For example, if students who received three points on a constructed-response item score lower, on average, on the total test than students who received only two points on the item, the item will be flagged for review. The *p*-values for operational items can be found in Appendix A. Most of the multi-point operational items had *p*-values following the expected mean total score. Across all tested grades and subjects, less than 1% of operational items were flagged. Flagged items were verified by AIR content experts and psychometricians reported that all of them behaved as expected.

4.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

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

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

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

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

A generalized Mantel-Haenszel (MH) procedure was applied to calculate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's raw score on the operational items on a given test is used as the ability-matching variable. That score is divided into 10 intervals to compute the $MH\ \chi^2$ DIF statistics for balancing the stability and sensitivity of the DIF scoring category

selection. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD) are computed for polytomous items.

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

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

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

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

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

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

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

The MH conditional odds ratio is calculated as

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. DIF classification rules are illustrated in Table 14. Items were also indicated as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American, Hispanic, or female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., White or male). If the DIF statistics fell into the "C" category for any group, the item showed significant DIF and was reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts reviewed all items flagged based on DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance.

Table 14: DIF Classification Rules

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

In addition to the classical item summaries described in this section, IRT based statistics were used during item review. These are described in Section 5.2.

4.3 CLASSICAL ANALYSES RESULTS

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

Table 15 through **Error! Reference source not found.** provide summaries of the p-values by percentile and range by grade and subject for operational items. Note that the "Total OP Items" column shows the number of operational items that were used in the computation of the percentiles. The two-dimension scores for writing items are counted as two items in ELA. Indiana students' performance indicates the desired variability across the scale in all grades and subjects. The variability informs us that the constructed operational forms had a good discrimination for Indiana students.

Table 15: Operational Item p-*Value Five-Point Summary and Range, ELA*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	396	0.06	0.12	0.27	0.39	0.56	0.74	0.89
4	315	0.09	0.17	0.29	0.42	0.59	0.77	0.90
5	290	0.04	0.15	0.31	0.43	0.59	0.79	0.92
6	228	0.06	0.13	0.30	0.41	0.58	0.76	0.87
7	319	0.06	0.15	0.29	0.45	0.59	0.76	0.94
8	355	0.06	0.18	0.30	0.47	0.62	0.76	0.87

Table 16: Operational Item p-*Value Five-Point Summary and Range, Mathematics*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	493	0.02	0.29	0.46	0.56	0.65	0.75	0.91
4	526	0.08	0.28	0.42	0.50	0.56	0.70	0.95
5	433	0.03	0.22	0.37	0.46	0.56	0.69	0.89
6	609	0.03	0.26	0.42	0.49	0.56	0.68	0.96
7	574	0.02	0.15	0.28	0.41	0.52	0.66	0.87
8	414	0.02	0.11	0.29	0.40	0.49	0.65	0.87

Table 17: Operational Item p-*Value Five-Point Summary and Range, Science*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
4	58	0.24	0.29	0.43	0.54	0.64	0.75	0.92
6	80	0.05	0.14	0.44	0.58	0.66	0.83	0.85
Biology	104	0.05	0.16	0.37	0.49	0.62	0.78	0.92

Table 18: 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	54	0.28	0.34	0.44	0.52	0.60	0.73	0.77
U.S. Government	53	0.09	0.10	0.20	0.32	0.42	0.60	0.70

DIF summary tables based on Indiana students can be found in Appendix A. Across all operational items and DIF comparison groups, less than 5% of ELA operational items, less than 2% of Mathematics operational items, less than 5% of Science operational items, and less than 6% of Social Studies operational items were flagged as C DIF. Flagged items were reviewed by AIR content specialists and psychometricians to ensure that they were free of bias. The review of the flagged items did not reveal any serious issues with items.

5. ITEM CALIBRATION

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

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

5.1 ITEM RESPONSE THEORY MODELS

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

Two-Parameter Logistic Model

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

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

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

Generalized Partial Credit Model

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

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

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

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

Rasch Model

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

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

Rasch Testlet Model

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

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

where u_g is the nuisance dimension parameter for cluster g.

5.1.1 ELA, Mathematics, and Social Studies

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

5.1.2 Science

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

5.1.3 Social Studies

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

5.2 IRT Analyses Results

Table 18 displays the number of students in the operational calibrations. For ELA and Mathematics, all Smarter items in the bank used their previously calibrated item parameters, which are on a vertical scale. The Smarter items were anchored to their bank values and remaining items were calibrated so they were placed on the Smarter IRT vertical scale using Indiana data from the spring 2019 administration. While some items in Science and Social Studies had item parameters, a new IRT scale was established using Indiana data from the Spring 2019 administration.

Table 18: N *Students Used in Operational Calibrations*

ELA		Mathematics		Science		Social Studies	
Grade	Calibration N Count	Grade	Calibration N Count	Grade	Calibration N Count	Grade	Calibration N Count
3	72959	3	82316				
4	83916	4	83398	4	83236		
5	85810	5	85706			5	85469
6	75415	6	84953	6	84765		
7	85810	7	83586				
8	81975	8	81963				
				Biology	75745	U.S. Government	1217

5.2.1 IRT Summaries

The IRT statistical properties of the final operational test forms used for ILEARN are summarized in Table 19 through Table 22.

Table 19: Operational Item Parameter Five-Point Summary and Range, ELA

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.18	0.33	0.44	0.59	0.74	0.98	1.39
	b	-2.92	-2.09	-1.17	-0.34	0.41	1.63	4.12
4	a	0.05	0.27	0.45	0.60	0.75	0.95	1.25
	b	-2.46	-1.73	-0.94	-0.07	0.80	1.82	6.23
5	a	0.13	0.27	0.43	0.57	0.70	0.95	1.25
	b	-2.28	-1.56	-0.61	0.36	1.31	2.56	5.19
6	a	0.19	0.26	0.39	0.57	0.72	1.02	1.35
	b	-1.45	-1.06	0.00	0.97	1.64	2.87	4.27
7	a	0.01	0.26	0.44	0.56	0.68	0.86	1.17

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
	b	-2.02	-0.99	0.16	1.09	1.87	3.30	5.88
8	a	0.03	0.25	0.41	0.53	0.69	0.88	1.12
	b	-3.01	-0.64	0.14	1.20	2.09	3.56	5.60

Table 20: Operational Item Parameter Five-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	0.88	1.10	1.31	1.52
	b	-4.34	-2.77	-1.84	-1.25	-0.40	0.85	2.87
4	a	0.18	0.36	0.64	0.82	1.05	1.36	1.80
	b	-3.26	-1.93	-0.96	-0.24	0.40	1.34	4.11
5	a	0.18	0.34	0.58	0.75	0.94	1.21	1.47
	b	-2.53	-1.07	-0.20	0.36	1.05	2.17	6.20
6	a	0.13	0.29	0.53	0.70	0.87	1.11	1.40
	b	-3.93	-1.61	-0.17	0.80	1.58	2.68	9.16
7	a	0.05	0.25	0.49	0.76	0.94	1.17	1.49
	b	-2.02	-0.50	0.91	1.58	2.41	3.54	7.80
8	a	0.14	0.22	0.39	0.55	0.73	1.00	1.20
	b	-1.87	-0.95	0.52	2.00	3.07	5.09	9.02

Table 21: Operational Item Parameter Five-Point Summary and Range, Science

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
4	b	-2.71	-1.57	-0.67	-0.20	0.27	0.95	1.35
6	b	-2.21	-2.03	-1.13	-0.43	0.33	2.60	3.41
Biology	b	-2.86	-1.36	-0.56	0.25	0.97	2.05	3.51

Table 22: Operational Item Parameter Five-Point Summary and Range, Social Studies

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
5	a	0.19	0.21	0.42	0.57	0.69	0.98	1.19
	b	-1.35	-1.05	-0.52	-0.12	0.30	1.65	1.95
U.S. Government	b	-2.01	-1.53	-0.62	-0.11	0.64	1.55	1.65

5.2.2 2019 ILEARN Test Characteristic Curves

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

6. SCORING AND REPORTING

6.1 MAXIMUM LIKELIHOOD ESTIMATION

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

6.1.1 Likelihood Function

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

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

where

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

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

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

$$q_i = 1 - p_i$$

and where a_i is the slope of the item response curve (i.e., the discrimination parameter), b_i is the location parameter, z_i is the observed response to the item, i indexes item, h indexes step of the item, m_i is the maximum possible score point, b_{il} is the *i*th step for item i with m total categories, and $D = 1.7$.

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

6.1.2 Derivatives

Finding the maximum of the likelihood requires an iterative method, such as Newton-Raphson iterations. The estimated MLE is found via the following maximization routine:

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

where

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

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

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

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

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

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

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

6.1.3 Standard Errors of Estimates

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

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

where

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

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

6.1.4 Extreme Case Handling

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

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

Table 23 through Table 26 give the LOT/LOSS and HOT/HOSS for the ILEARN assessments.

Table 23: ELA Theta and Scaled-Score Limits for Extreme Ability Estimates

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
3	-5.8667	3.4667	5060	5760
4	-5.4667	4.1333	5090	5810
5	-5.2000	4.6667	5110	5850
6	-4.9333	4.9333	5130	5870
7	-4.9333	5.2000	5130	5890
8	-4.6667	5.6000	5150	5920

Table 24: Mathematics Theta and Scaled-Score Limits for Extreme Ability Estimates

Grade	Lowest of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of Scale Score (HOSS)
3	-5.6000	3.0667	6080	6730
4	-5.3333	4.0000	6100	6800
5	-5.2000	4.6667	6110	6850
6	-5.2000	4.9333	6110	6870
7	-5.0667	5.6000	6120	6920

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

Table 25: Science Theta and Scaled-Score Limits for Extreme Ability Estimates

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

Table 26: Social Studies Theta and Scaled-Score Limits for Extreme Ability Estimates

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

6.1.5 Standard Errors of LOT/HOT Scores

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

6.2 Transforming Theta Scores to Reporting Scale Scores

For 2018-2019, 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 or MetaMetrics linking 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 27 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 and formed by linking across grades using common items in adjacent grades. Grade 6 was used as the baseline and each grade was successively linked onto the scale. More details about the vertical scaling methods can be found in Chapter 9 of the 2013–2014 Technical Report (Smarter Balanced, 2016). The slope and

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

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

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

Table 27: 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 was assigned an overall performance category in accordance with his or her overall scale score. Table 28 through Table 32 provide the scale score range for performance standards for ILEARN. The lower bound of the Level 3, At Proficiency, marks the minimum cut score for proficiency.

Table 28: Proficiency Levels for ELA

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

Table 29: Proficiency Levels for Mathematics

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

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

Table 30: Proficiency Levels for Science

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

Table 31: Proficiency Levels for Social Studies Grade 5

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

Table 32: Proficiency Levels for Social Studies U.S. Government

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

6.4 REPORTING CATEGORY SCORES

6.4.1 MLE/MMLE Scoring

Reporting category theta scores were calculated using either MLE or MMLE, depending on the assessment, based on the items contained in a particular reporting category. The same rules for scoring all correct and all incorrect cases were applied to reporting category scores.

6.4.2 Strengths and Weaknesses

For reporting categories, relative strengths and weaknesses were reported for each student at the reporting category level. The difference between the proficiency cut score

and the reporting category score plus or minus 1.5 times standard error of the reporting category was used to determine the relative strengths and weaknesses.

The specific rules for mastery are as follows:

- Below (Code = 1): if $round(SS_{rc} + 1.5 * SE(SS_{rc}),0) < SS_p$;
- At/Near (Code = 2): if $round(SS_{rc} + 1.5 * SE(SS_{rc}),0) \geq SS_p$ and $round(SS_{rc} - 1.5 * SE(SS_{rc}),0) < SS_p$, a strength or weakness is indeterminable; and
- Above (Code = 3): if $round(SS_{rc} - 1.5 * SE(SS_{rc}),0) \geq SS_p$,

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

6.4.3 Standard Level Aggregate Scores

Standard level information was reported relative to the proficiency standard for tests that were adaptively administered. In Spring 2019 standard level information was reported for the ELA and Mathematics assessments.

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

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

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

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

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

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

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

$$\bar{\delta}_{Sg} = \frac{1}{\mathrm{n}_g}\sum_{j\in g}\delta_{jS},$$

and

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

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

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

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

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

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

6.5 LEXILE AND QUANTILE SCORES

ILEARN reports Lexile and Quantile measures with ELA and Mathematics test scores. MetaMetrics provided conversion tables between ELA scale scores and Lexile measures and between Mathematics scale scores and Quantile measures for each grade and subject. A linking study for ELA and Mathematics took place at the end of July 2019 to determine final conversions.

7. QUALITY CONTROL PROCEDURES

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

7.1 SCORING QUALITY CHECK

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

Additionally, the Assessment Systems Corporation provided replication of the psychometric scoring process for ILEARN. Scores were approved and published by the IDOE only when all three independent systems matched.

Despite the implementation of the established quality control processes, a small number of data issues resulted that were not immediately identified. Those issues were subsequently resolved, and the quality control processes have been updated for 2020.

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

2018–2019

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 American Institutes for Research (AIR): Stephan Ahadi, Elizabeth Ayers-Wright, Kevin Clayton, Christopher Johnston, and Gabriel Martinez. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and Program Leads.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION

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

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

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

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

Table 1: Sources of Items for the ILEARN 2018–2019 Assessments

Subject and Grade(s)	Licensed Bank(s)	Legacy ISTEP+ Items	Custom Development	Notes
ELA 3–8	Smarter ICCR	Yes	Yes	ICCR items were used to augment the pool where the Smarter item pool could not provide items or provided items only to a limited extent. ISTEP+ items were used only when required to ensure blueprint was met.
Mathematics 3–8	Smarter ICCR	Yes	Yes	ICCR items were used to augment the pool where the Smarter item pool could not provide items or provided items to a limited extent only. ISTEP+ items were used only when required to ensure blueprint was met.
Science 4 and 6	ICCR	Yes	Yes	Very few ICCR items were used operationally in 2018–2019, but additional newly

Subject and Grade(s)	Licensed Bank(s)	Legacy ISTEP+ Items	Custom Development	Notes
				developed ICCR items were field tested.
Science Biology	Hawaii EOC ICCR	Yes	Yes	Very few ICCR items were used operationally in 2018–2019, but additional newly developed ICCR items were field tested.
Social Studies 5	No	Yes	Yes	
U.S. Government	No	No	Yes	

1.1 CLAIM STRUCTURE

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

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

Table 2: ELA Claims

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

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

Table 3: Mathematics Categories

Grade	Reporting Categories				
Grade 3	Algebraic Thinking and Data Analysis	Computation	Geometry and Measurement	Number Sense	Process Standards
Grade 4	Algebraic Thinking and Data Analysis	Computation	Geometry and Measurement	Number Sense	Process Standards

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

1.2 UNDERLYING PRINCIPLES GUIDING DEVELOPMENT

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

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

Item development efforts support the goal of high-quality items through rigorous development processes managed and tracked by a content development platform that ensures that every item flows through the correct sequence of reviews and captures every comment and change to the item.

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

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

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

1.3 ORGANIZATION OF THIS VOLUME

This volume is organized in three sections:

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

2. ILEARN ITEM BANK SUMMARY

The ILEARN item bank is quite robust, containing licensed items which have been constructed explicitly to support multiple statewide assessment programs. As described above, all items used on ILEARN assessments are aligned to the IAS. The ILEARN item banks supported an adaptive assessment in Spring 2019 for ELA and Mathematics, a fixed-form assessment in all three grades of Science, and a fixed-form assessment in Social Studies grade 5 and U.S. Government. Summaries of current item inventories are provided in this section.

2.1 ITEM BANKS

Table 4 provides the count of items, by source, used on the 2018–2019 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. When gaps in coverage existed, AIR's ICCR item bank was used to augment the ILEARN item bank. Across grades, some gaps in IAS coverage existed, and legacy ISTEP+ items were used as needed to fill these gaps. In addition, in a few small instances, new, custom Indiana item development was needed to complete the item bank and ensure complete coverage of the IAS.

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

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

Table 4: Item Counts by Source

Subject and Grade	# of Smarter Items	# of ICCR Items	# of ISTEP+ Legacy Items	# of Custom Items	# of Hawaii EOC items
ELA 3	369	25	8	-	-
ELA 4	272	30	15	8	-
ELA 5	248	23	14	4	-
ELA 6	188	37	9	-	-
ELA 7	273	40	14	-	-
ELA 8	335	13	16	15	-
Mathematics 3	418	46	25	4	-
Mathematics 4	490	17	19	-	-
Mathematics 5	361	50	21	1	-

Subject and Grade	# of Smarter Items	# of ICCR Items	# of ISTEP+ Legacy Items	# of Custom Items	# of Hawaii EOC items
Mathematics 6	571	19	15	4	-
Mathematics 7	512	32	17	13	-
Mathematics 8	366	23	14	12	-
Science 4	-	1	33	20	-
Science 6	-	-	29	36	-
Biology	-	-	17	6	71
Social Studies 5	-	-	51	4	-
U.S. Government	-	-	-	54	-

Additionally, all assessments other than Social Studies included one performance task per grade. Table 5 presents the counts of performance tasks in the 2018–2019 item pool.

Table 5: Performance Task Counts by Source

Subject and Grade	# of Smarter Performance Tasks	# of Custom Indiana Performance Tasks
ELA 3	3	-
ELA 4	3	-
ELA 5	3	-
ELA 6	3	-
ELA 7	3	-
ELA 8	3	-
Mathematics 3	5	-
Mathematics 4	15	-
Mathematics 5	5	-
Mathematics 6	6	-
Mathematics 7	15	-
Mathematics 8	20	-
Science 4	-	2*
Science 6	-	2
Biology	-	2

*Note: While both Grade 4 Science performance tasks were administered to students, one was suppressed from scoring and reporting. Scores for students who received the suppressed performance task were calculated based on the non-performance task segment of the fixed-form. The non-performance task segment of the form met blueprint requirements for the overall test.

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 by AIR and IDOE. During two Item Acceptance Review meetings (April 2018 – Smarter and ICCR Mathematics and ELA; July 2018 – Hawaii EOC for Biology), educators reviewed items from these licensed banks and determined their levels of agreement with the alignment of items to the IAS. A short description of these meetings follows, and a full agenda can be found in Appendix L, Item Acceptance Review Meeting Plan.

AIR and IDOE worked to determine a crosswalk between the IAS and the standards for the licensed banks. During the 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.

2.3 SPRING 2019 ITEM BANK COMPOSITION

Table 6 through Table 9 list the ELA, Mathematics, Science, and Social Studies item types and provide a brief description of each. Examples of various item types can be found in Appendix E, Example Item Types.

Table 6: ELA 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.
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.
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 several options.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Text Entry (TE)	Student is directed to type their response in a text box.

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

Table 7: Mathematics Item Types and Descriptions

Response Type	Description
Equation Response (EQ)	Student uses a keypad with a variety of mathematical symbols to create a response. Responses can include numbers, fractions, expressions, inequalities, functions, and equations.

Response Type	Description
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.
Multiple-Choice (MC)	Student selects one correct answer from four options.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Table Input (TI)	Student types numeric values into a given table.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Text Entry (TE)	Student is directed to type their response in a text box.

Table 8: Science 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.
Graphic Response (GI)	Student selects numbers, words, phrases, or images and uses the drag-and-drop feature to place them into a graphic. This item type may also require the student to use the point, line, or arrow tools to create a response on a graph.
Hot Text (HT)	Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference.
Multiple-Choice (MC)	Student selects one correct answer from four options.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Performance Task (PT)	Student works through a group of items measuring multiple standards and using various item types to demonstrate the ability to integrate knowledge and skills.
Simulation (SIM)	Student selects inputs to “run” trials. Data is presented in a table after trials are run.
Table Input (TI)	Student types numeric values into a given table.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Text Entry (TE)	Student is directed to type their response in a text box.

Table 9: Social Studies Item Types and Descriptions

Response Type	Description
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.
Multiple-Choice (MC)	Student selects one correct answer from several options.
Table Match (MI)	Student checks a box to indicate if information from a column header matches information from a row.
Multiple Select (MS)	Student selects all correct answers from a number of options.
Text Entry (TE)**	Student is directed to type their response in a text box.

***This item type is not used in the optional U.S. Government assessment.*

3. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

3.1 OVERVIEW

Both Smarter and AIR 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 AIR'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 AIR reviewers and stakeholders in committee meetings, reviewed items in ITS as they would appear to the student, with all accessibility features and tools.

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

- Selection and training of item writers;
- Writing and internal review of items;
- Review by state personnel and stakeholder committees;
- Markup for translation and accessibility features;
- Field testing; and
- Post field-test reviews.

Each of these steps had a role in ensuring that the items could support the claims on which they were based. Table 10 describes how each step contributed to these goals. Each step in the process is discussed in more detail below.

Table 10: Summary of How Each Step of Development Supports the Validity of Claims

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

Per the recommendations of the 2016 ISTEP+ Panel, the Indiana Department of Education is leveraging quality content from third-party item banks for use on ILEARN assessments. These item banks are accompanied by item specifications which will be utilized where alignment was confirmed by Indiana educators. The specifications available are described in Table 11 below.

Table 11: 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 American Institutes for Research (AIR) 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 Standard are used for ILEARN assessments. Indiana educators approved alignment of items to each Indiana Standard.*

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

ICCR items and passage specifications were developed in collaboration between content experts in one of AIR's partner states and AIR 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 AIR test developers and IDOE content specialists.

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

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

3.2.1 Passage Specifications

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

3.2.2 Item Specifications

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

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

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

Table 12 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 12: Sample ELA Item Specification for Grade 4

Content Standard	**4.RV.2.2:** Identify relationships among words, including more complex homographs, homonyms, synonyms, antonyms, and multiple meanings.
Content Limits	Items should ask students not to define the type of word that is being used but rather to demonstrate its meaning between the words. Items may refer only to synonym and antonym in the stimuli. All words should be provided with sufficient context for support.
Construct-Relevant Vocabulary	antonyms, meaning, opposite, phrase, relationship, replace, similar/same as, synonyms,
Recommended Response Mechanisms (Item Types)	Drag and Drop Evidence-Based Selected Response Hot Text Multiple Choice Multi-Select
DOK	2
Evidence Statements	
Students replace a given word with synonyms, antonyms, homographs, homonyms, and multiple-meaning words.	
Students use context to determine or support meaning.	
Students identify a word, sentence, or phrase that uses a given word in the same way.	
(NOTE: Level of difficulty will depend on subtlety/amount of text and/or complexity of interpretation required.)	
Sample Item	
Why is “[word X]” a better word to use from paragraph 4 than “[word Y]”? A. [Word X] suggests [something more formal] B. [Word X] suggests [something more precise] C. [Word X] suggests [something more aligned to the tone] D. [Word X] suggests [something more audience appropriate]	
Literacy Considerations	Word List: Content can select construct-irrelevant words for glossing, which gives students access to the definition and an audio clip of those words. Considerations will include the question/task, standard, and construct-relevant words necessary for the item.
Visual and Auditory Considerations (NOTE: These considerations generally refer to the passage/media source rather than the item.)	American Sign Language: Allows a student to see a video of an ASL interpreter. This option will be included only if the media contains audio. Audio Transcriptions: Written transcripts of audio for students of varying auditory and visual abilities can be provided as needed. The same transcripts will be used for ASL videos.

	Closed Captioning: Captions media so that audio is available for students who are hearing impaired. Can be used for both audio-only and video media. Graphics: Graphics will be provided in formats that are accessible to students with varying abilities, including students who are blind or visually impaired. Graphics should contain only content that will help students understand or process information; those that do not contribute to the student's understanding should not be included. Graphics should be braillable whenever possible; those that cannot be brailled will be provided to blind/visually impaired students through a verbal or written description.
Linguistic Complexity	Rating to be completed after all final edits have been applied and approved by IDOE.

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

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

3.3 SELECTION AND TRAINING OF ITEM WRITERS

All AIR 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 G, Item Writer Training Materials. These include:

- AIR'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.

Sample specifications for passages, Mathematics, and ELA are presented in Exhibits C, D, and E, respectively.

3.4 INTERNAL REVIEW

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

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

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

3.4.1 Preliminary Review

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

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

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

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

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

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

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 F. The specialist also ensures that the revisions made during the Preliminary Review did not introduce errors or content inaccuracies. This reviewer approaches the item both from the perspective of potential clients as well as the specialist's own experience in test development.

3.4.3 Edit Review 1

During Edit Review 1, editors have four primary tasks.

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

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

Third, editors review all material for fairness and language accessibility issues, using AIR's Language Accessibility, Bias, and Sensitivity (LABS) Guidelines shown in Appendix G, Item Writer Training Materials.

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

3.4.4 Senior Content Review

By the time an item arrives at Senior Content Review, it has been thoroughly vetted by both content reviewers and editors. Senior reviewers (in particular, Senior Content Specialists) look back at the item's entire review history, making sure that all the issues identified in that item have been adequately addressed. Senior reviewers verify the overall content of each item, confirming its accuracy and alignment to the

standard. For machine-scored, constructed-response items, senior reviewers carefully check the rubric and scoring logic by responding to the task just as the student would in the testing environment. They check full-credit, partial-credit, and zero-credit responses to verify that the scoring is working as intended and that the scoring assertions adequately address the evidence the student provides with each type of response.

3.5 Review by State Personnel and Stakeholder Committees

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

3.5.1 State (Client) Review

After items have been developed in the AIRCore 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. An AIR director for Mathematics or ELA reviews all client-requested edits in light of the AIRCore item specifications, other clients' requests, and existing items in the bank to determine whether the requested edits will be made. At this stage, clients have the option to present these items to committee (based on the edits made) or withhold them from committee review.

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

Once items have been accepted by IDOE and are ready for CFC, Linguistic complexity ratings are applied in ITS. For AIR-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 Indiana development was taken to the Content and Fairness Committee Review. This committee combined the functions of the Content Advisory

Committee and the Language Accessibility, Bias, and Sensitivity (LABS) Committee, as described in the following section.

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

3.5.3 Markup for Translation and Accessibility Features

After all approved state and committee recommended edits have been applied, the items are considered "locked" and ready for all accessibility tagging. Accessibility markup is embedded into each item as part of the item development process rather than as a post-hoc process applied to completed test forms.

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

3.5.4 Indiana Educator Review of Licensed Item Banks

Because ILEARN relies heavily on licensed banks, a process for ensuring alignment of those items to the Indiana Academic Standards was developed by AIR and IDOE. During two Item Acceptance Review meetings (April 2018 – Smarter and ICCR Mathematics and ELA; July 2018 – Hawaii EOC Biology), educators reviewed items from these licensed item banks. Appendix L, Item Acceptance Review Meeting Plan, contains the plans for these meetings.

3.6 FIELD TESTING

All Smarter and ICCR items were field tested embedded in operational, summative, accountability assessments in participating states. Previously operational ISTEP+ legacy items were field tested in Indiana prior to Spring 2019. Custom Indiana development was field tested (either as embedded field-test items or operational field-test items) in Spring 2019. The field testing is described in detail in Volume 1, Section 3.2.

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 that they received is created. These are reviewed by qualified AIR content staff to ensure that all correct responses, and only correct responses, receive a score.

3.7.2 Rubric Validation

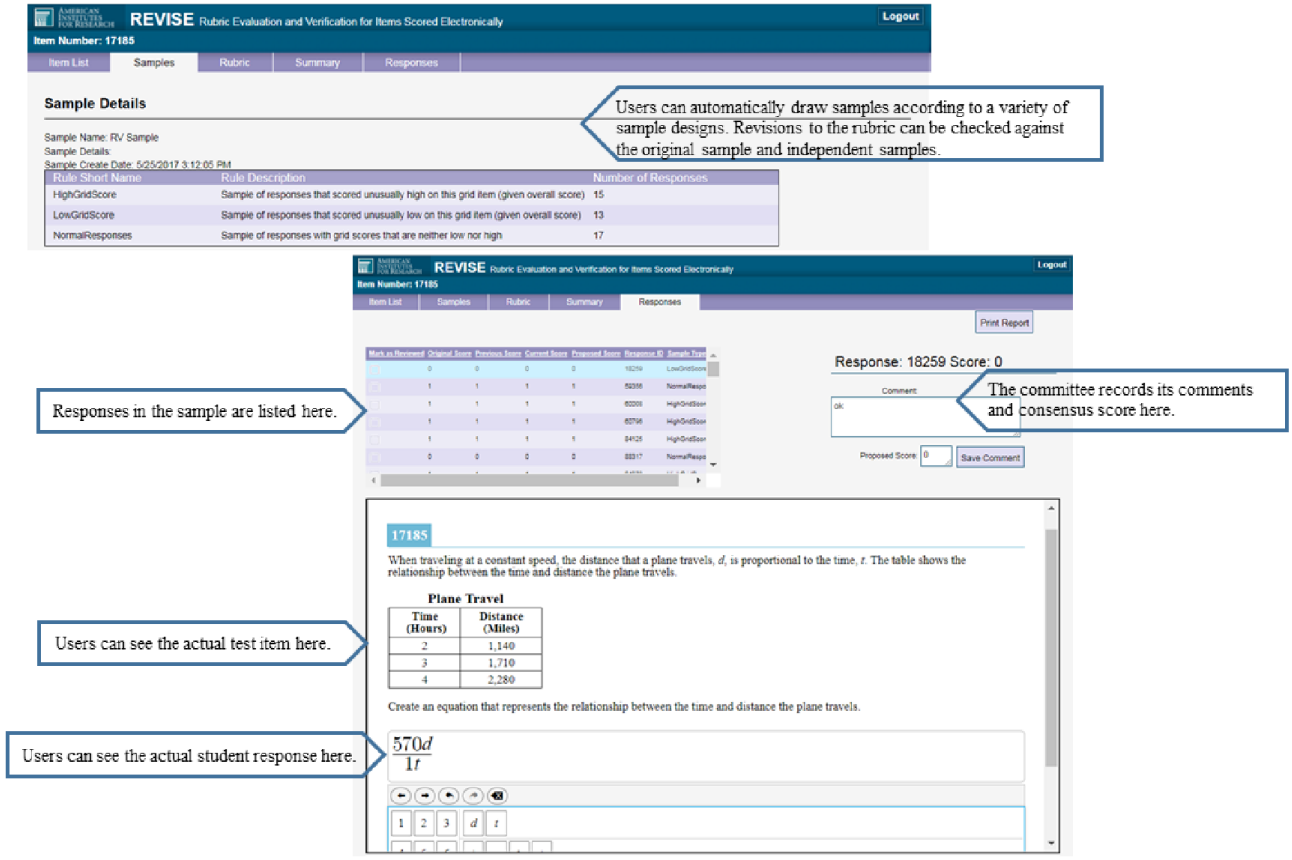
More complex selected-response items, as well as machine-scored constructed-response items, undergo rubric validation, which occurs in two phases. During the first phase, AIR content experts draw one or more samples to identify anomalous or unforeseen responses and ensure that they are scored correctly.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

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

During these reviews, experts review responses and scores in an AIR 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.

Figure 1: Features of the REVISE Software



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 AIR 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.

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 4, Section 6.1, of this technical report describes in detail the statistical flags that send items to data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues. Committee members were taught to

interpret these flags and given guidelines for examining the items for content or fairness issues. A sample of the training materials used for these data review meetings appears in Appendix J, Sample Data Review Training Materials.

4. ILEARN BLUEPRINTS AND STATE ASSESSMENT TEST CONSTRUCTION

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

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

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

4.1 TEST BLUEPRINTS

4.1.1 Blueprint Construction Meeting

In February 2018, IDOE and AIR worked closely with Indiana educators to create blueprints that guided the item development process for all subjects and grades.

IDOE conducted a formal recruitment window in Winter 2017-2018 to identify potential educator and stakeholder participants in the blueprint and performance-level descriptor (PLD) process. From this pool, a sample of participants were invited to represent north, central, and south; urban, rural, and suburban; and other distinct state student subpopulations to ensure accessibility of the content. Each subject-area panel was comprised of grade-band subpanels. Each grade-band subpanel included approximately eight panelists, with four panelists representing each grade-level assessment, for a total of 80 panelists across the full range of ILEARN assessments.

Participants worked in subject-area, grade-band, and grade-level panels, cycling back and forth to ensure that assessment-level panels were continuously receiving feedback from subject-area educators across grades, and that final recommendations were aligned across the full system of ILEARN assessments.

The workshop began with a large group session to orient participants to the workshop objective (produce test blueprints) and review the activities to meet those objectives. The meeting was structured around three segments.

In the first segment, educators defined essential evidence as identified in their rigorous review of grade-level content standards. This activity began with a review of standards and culminated with high-level evidentiary statements created by educators. Initial review of standards and production of essential evidence occurred in the grade-level subpanels.

After drafting essential evidence statements, grade-band panels met to discuss similarities and differences between adjacent grade levels. Grade-band panels worked to ensure vertical articulation of the essential evidence across the grade levels within the grade band.

The full subject-area panel then reconvened to ensure vertical articulations of essential evidence across the grade-band panels. Deliberation of essential elements, especially across grades, helped to inform panelists about the most useful reporting frameworks for ILEARN assessment results. These statements informed Segment 2, which took these statements and aggregated them into Reporting Categories.

While the ILEARN subject-area assessments are unidimensional, measuring student achievement in the subject area overall, educators benefit from more fine-grained feedback about student achievement. How that feedback is structured can have important implications for how educators use assessment results to guide instruction. Understanding of overall test performance can be augmented by reporting back to educators on student performance along any of the dimensions on which assessment items are aligned.

In the second segment, panelists reviewed their evidentiary statements and discussed potential reporting frameworks that best supported instruction in Indiana. While IDOE and AIR staff were present in the room to answer questions, educators were encouraged to discuss and propose the framework they determined to best support instruction and coverage of the IAS. Following panel-level deliberations, discussions were extended to the full subject-area panel. Because it was important to adopt a reporting framework that was coherent across grade levels, the subject-area panels worked collaboratively to achieve consensus on a common reporting framework. Each panel appointed a representative to report the basis for consensus within each grade-level panel, but all panelists were allowed to participate in the subject-area deliberations.

Once panelists agreed upon a reporting structure, in Segment 3, the IAS were aligned within the adopted reporting structure. Educators first weighted the relative importance of each reporting category, and then they weighted standards within reporting categories with respect to priority for ensuring that students are on track for college-and-career-readiness. Although test blueprints were constructed to yield test administrations that assess a representative sample of subject-area standards, standards are not of equal importance, with mastery of some standards far more essential for college-and-career-readiness than others. Within each subject-area and grade-level panel, panelists worked independently to classify each standard into a reporting category and each standard as less important, important, or critically important. Standards were considered less important or critically important if the majority of panelists (e.g., four of the six) agreed. After making their initial classifications, panelists were provided feedback about their initial ratings and worked through each of the standards to discuss why, for example, some panelists classified the standard as critically important while others did not. Panelists focused most of their discussions on standards where there was disagreement

about the importance of the standards. Based on these deliberations, panelists rated the importance of each standard a second time. Again, standards were rated as less important or critically important based on agreement of the majority of panelists.

AIR psychometricians and content experts incorporated the results of the educator meeting to create high-level, public-facing blueprints for all grades and subjects. There were also important constraints in the construction of test blueprints. Restrictions on testing time, for example, placed important constraints on overall test length. In addition, although some reporting categories were represented by fewer standards than others, each reporting category included a minimum number of items to yield reliable performance-level classifications. The presence of so many constraints limits the degree of freedom available for variation in test blueprints.

Subject-area panels reconvened via a webinar the week following the workshop and were provided with drafts for each of the grade-level test blueprints. A guided review of the initial blueprints illustrated how each of the blueprint elements was generated from the panelist feedback during the meeting and how the blueprints were based on constraints of the assessment system, reporting framework, and the standard importance ratings. Panels evaluated whether the recommended blueprints satisfied all constraints for the ILEARN assessments, including overall testing time. Subject-area panels were asked to deliberate about whether revisions should be made to the proposed grade-level blueprints to better support assessment goals. Following subject-area review and moderation of blueprints across grade-level panels, the subject-area panels made a recommendation to IDOE for the system of test blueprints. IDOE considered the draft blueprints and educator recommendations in order to finalize the blueprints.

Thus, ILEARN blueprints were designed to meet the following objectives:

- Provide full coverage of the breadth and depth of the IAS;
- Provide weight to the standards and reporting categories as identified by educators;
- Minimize testing time; and
- Include a Performance Task in all subjects except Social Studies.

The ILEARN item bank contains several different item types, such as traditional multiple-choice items, technology-enhanced items, and machine-scored constructed-response items. Any assessment built from this item bank could have a wide variety of item types represented. Thus, artificial restrictions were not placed on the number of items aligned to specific item types.

It is important to note that DOK ranges were not included in the blueprints because each IAS includes a target DOK. Other than U.S. Government, all IAS target DOK values were determined during the ISTEP+ administrations.

4.1.2 ILEARN Test Specifications

Test blueprints provided the following guidelines:

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

Table 13 presents the number of operational or operational field-test hand-scored items per form. Note that in ELA and Mathematics, all PTs included one or more hand-scored items. In Science, most of the PTs included one hand-scored interaction. Additionally, Indiana educators were invited to participate in the hand-scoring of these items in a partnership with Measurement Incorporated (MI).

Table 13: Number of Hand-Scored Items by Form

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

In addition to operational and non-operational field-test Items, each form included embedded field-test (EFT) items. It is important to note that DOK ranges were not included in the blueprints because each IAS includes a target DOK. Other than U.S.

Government, all IAS target DOK values were determined during the ISTEP+ administrations. Table 14 denotes the number of EFT items per form.

Table 14: Number of Embedded Field-Test Items by Form

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

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

The Spring 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.

Table15 through Table 18 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 15: 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 16: Blueprint Percentage of Test Items Assessing Each Reporting Category in Mathematics

Grade	Reporting Category				
	Algebraic Thinking and Data Analysis	**Computation**	**Geometry and Measurement**	**Number Sense**	**Process Standards**
3	19—24%	23—28%	19—24%	23—28%	8—13%
4	19—24%	23—28%	19—24%	23—28%	8—13%
	Algebraic Thinking	**Computation**	**Geometry and Measurement, Data Analysis, and Statistics**	**Number Sense**	**Process Standards**
5	20—26%	22—28%	18—23%	22—28%	8—13%
	Algebra and Functions	**Computation**	**Geometry and Measurement, Data Analysis, and Statistics**	**Number Sense**	**Process Standards**
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 17: 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 18: 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;
- 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 2018, AIR psychometricians and content experts worked with IDOE to build forms for the Spring 2019 administration. ILEARN assessment test form construction utilized test construction guidelines, explicit blueprints, and collaborative participation from all parties. The Spring 2019 ILEARN test forms were built by AIR test developers to match exactly the detailed test blueprint and target distributions of item difficulty and assessment information when information was available and to the extent possible.

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

For the online ELA and Mathematics computer-adaptive test (CAT), item pools of available items were used, and there was no single test form constructed. For online Science and 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, AIR 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. Spring 2019 is a first-year form and there is no baseline form for comparison, but in subsequent years, this review will ensure that new forms match the information curve and test characteristic curves from this first-year form.

The ILEARN forms were created using AIR'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 AIR Content Team

AIR 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 AIR content staff,
- Revision of the operational item sets according to feedback from the AIR 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 AIR Technical Team

The AIR 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 AIR'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

Because Spring 2019 was the first operational year, there was not a reference curve or statistical targets with which to compare. Instead, 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 second 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 19. 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 2: Statistical Test Summary Comparison for Grade 5 Social Studies Online and Paper Forms

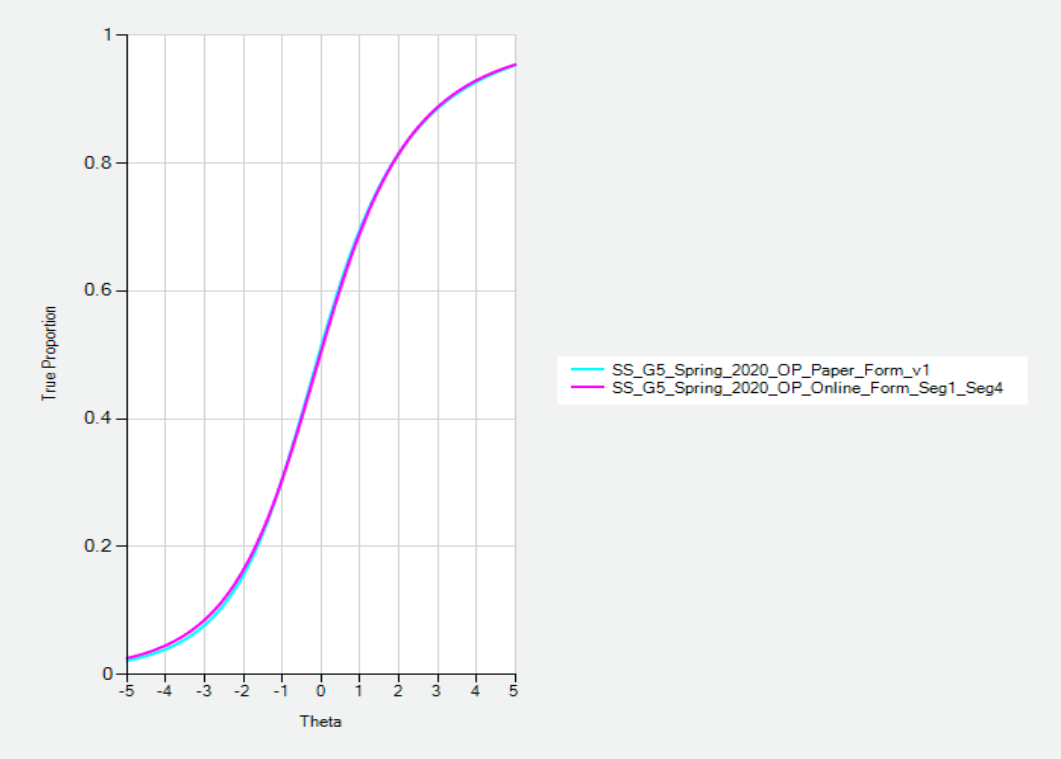
Type	Statistics	Paper Form	Online Form
Overall	Number of Items	40	40
	Possible Score	42	42
	Difficulty Mean	0.18	0.13
	Difficulty StDev	1.02	0.89
	Difficulty Minimum	-1.21	-2.21
	Difficulty Maximum	4.04	2.06
	Parameter-A Mean	0.56	0.53
	Parameter-A StDev	0.24	0.21
	Parameter-A Minimum	0.19	0.19
	Parameter-A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

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 braille/breach 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 braille/breach forms of grade 3 ELA Reading. Appendix C of Volume 1 provides the TCC for all grades in both subjects.

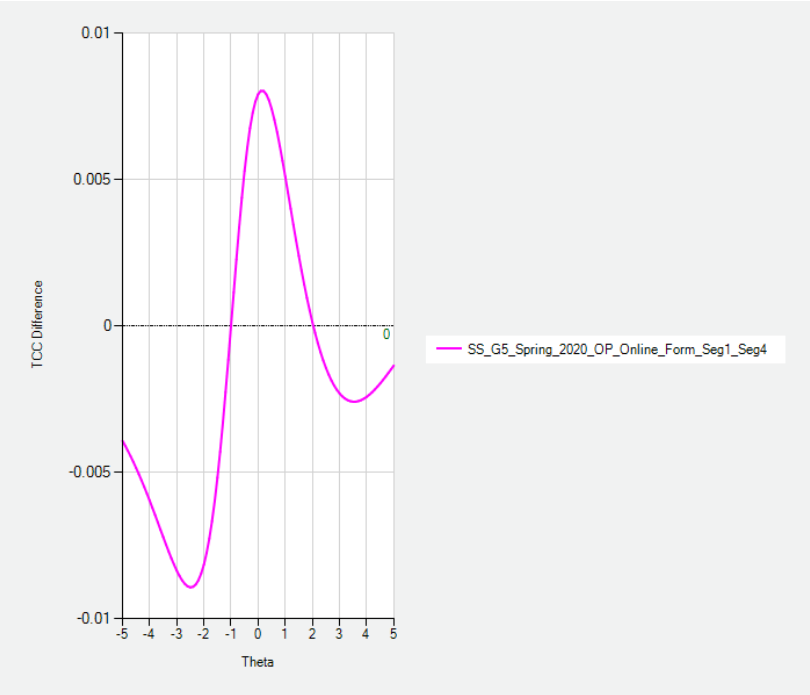
Figure 2: TCC Comparisons of Grade 5 Social Studies Online and Paper Forms



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.

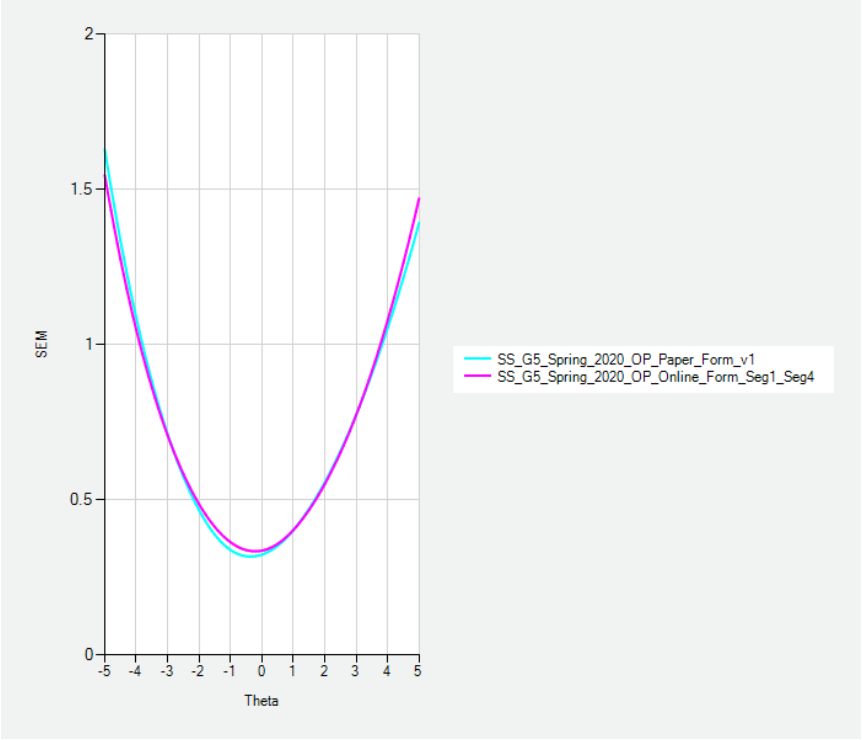
Figure 3: TCC Differences of Grade 4 Science Online and Accommodated Forms



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



5. PERFORMANCE LEVEL DESCRIPTORS

The Indiana Department of Education (IDOE) held a meeting with Indiana educators the week of June 18–21, 2018 to develop performance level descriptors (PLDs). The main purpose of the meeting was for educators to develop Range PLDs for each grade and content area and recommend proficiency level names to be used for reporting following their review of the policy PLDs.

Performance level descriptors (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. This section focuses on Policy and Range PLDs, as they were the subject of the June 2018 meetings with Indiana educators.

- **Policy PLDs:** Policy PLDs articulate the overall claims about a student's performance in each performance level. The policy PLDs are used by policymakers to broadly articulate the goals and rigor for the state's performance standards.
- **Range PLDs:** Range PLDs describe the expectations for students across each standard and proficiency level, demonstrating how the content represents a progression of knowledge, skills, and processes across performance levels and across grades. For licensed banks, range PLDs specific to each grade and content area were used by test developers to guide item writing within proficiency levels to ensure content discriminates by mastery of essential content with the range of proficiency. Range PLDs were created for each Indiana Academic Standard (IAS) for use in standard setting, as well as to guide future item writing.

5.1.1 Policy PLDs

Policy PLDs define, at a broad policy level, what it means to be proficient across the performance levels. Policy PLDs must convey an appropriate sense of rigor, clearly setting Indiana's expectations for a progression toward college and career readiness. Prior to the Range PLD meeting in June 2018, AIR and IDOE drafted Policy PLDs for educator review. The Policy PLDs were informed by Department leadership for educators to consider in light of the new assessment. During the first part of May 2018, IDOE sent a survey to educators to inform the labels for performance levels. On May 15, 2018, IDOE convened a stakeholder panel to make recommendations for ILEARN Policy PLDs. IDOE provided panelists with a background in the purpose and role of PLDs within the ILEARN assessment system. IDOE shared the educator survey information with the panelists and asked for their input on proficiency level names.

Panelists agreed with the educators' top choice for the following proficiency level names:

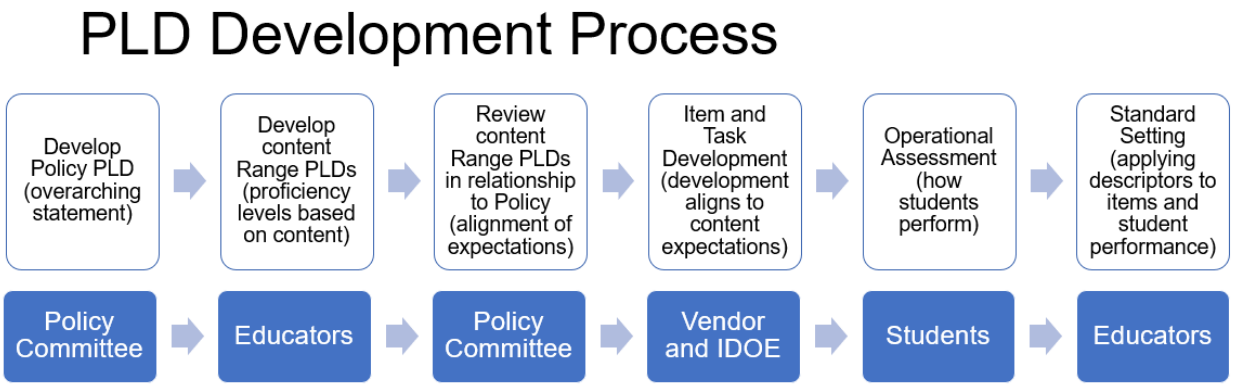
- Level 1: Below Proficiency
- Level 2: Approaching Proficiency
- Level 3: At Proficiency
- Level 4: Above Proficiency

After discussion, panelists unanimously agreed that PLDs should represent proficiency as on track for college and career readiness. During the meeting, the committee drafted recommended wording for each performance level that is reflected in the final Policy PLDs (see Volume 6, Appendix D).

5.1.2 June 2018 Range PLD Workshop

Panelists created Range PLDs during the June 2018 workshop. These Range PLDs were informed by two sample PLDs for each content area and grade level drafted by AIR to model PLD creation for workshop participants. AIR also created a large group PowerPoint (PPT) presentation to further articulate the purpose and process for Range PLD creation. The process followed is described in Figure 5:

Figure 5: PLD Development Process



IDOE approved the sample PLDs and PPT presentation prior to the workshop. Once IDOE approved these materials, each room facilitator adapted the PPT to their content area and grade band.

The workshop was organized as follows for each content area:

- ELA and Mathematics were each divided into grade band groups (Grades 3, 4, 5 and Grades 6, 7, 8);
- Science was divided into two groups (Grades 4 and 6 and Biology); and

- Social Studies had one group of Grade 5 educators. (Note: See Volume 7, "U.S. Government Standard Setting," for a description of how the Range PLDs were drafted for that content area.)

ELA and Mathematics had nine educators per grade band, enabling facilitators to divide the rooms into subgroups to complete work. Each subgroup was assigned a reporting category or set of standards. Recruitment targeted two teachers per grade, plus one access teacher per grade representing special populations (English Language Learners, Special Education). Science and Social Studies had six educators per room, with at least one special education teacher or English Learner teacher to represent special populations.

During the meeting, educators reviewed Policy PLDs and created Range PLDs. Facilitators were trained in advance on the following points:

- Eliciting educator input on the Policy PLDs;
- Creating Range PLDs for each standard and performance level, demonstrating how the content represents a progression of knowledge, skills, and processes across performance levels and across grades;
- Stressing to panelists that the Policy and Range PLDs are recommendations; and
- Asking panelists to provide recommendations on proficiency level names.

The workshop began one Day One with a welcome from IDOE and AIR staff, who provided an overview of the policy aspects of the workshop, including how this process contributes to the overall test development and standard setting processes. IDOE then discussed the Policy PLDs, providing an outline of the process used by the Policy PLD panel to draft Policy PLDs. IDOE shared the draft Policy PLDs with panelists. Then, AIR staff provided training on the processes to be used during the workshop.

After a break, the meeting shifted to Range PLD training within each room. Facilitators described the process for creating Range PLDs and shared the tools used for creating them. Facilitators began by asking panelists to consider the level of rigor described by each Policy PLD. Facilitators introduced panelists to Hess' Cognitive Rigor Matrix, asking them to think about the terms that best convey the rigor articulated by the different Policy descriptors. The panelists used the matrix as a resource to help form a common language around each proficiency level. Facilitators emphasized that there is not a direct correlation between DOK and proficiency levels.

Using an example Range PLD for one standard, facilitators then modeled how to parse out the Indiana Academic standards to create a Range PLD, focusing on the key words used in each performance level. In modeling how to parse the standards, the facilitator noted the importance of defining the level 3 (at proficiency) PLD as an anchor for the other descriptors. Next, the facilitator led the group through developing a Range PLD for

one standard. Each group developed a level 3 PLD, then moved to level 2, level 4, and level 1 for the first standard. Once the facilitators observed calibration across panelists, the panelists were split into groups to create Range PLDs for all standards. At the end of Day One, AIR and IDOE staff reviewed the panelists' work to check on subject area and content area coherence and consistency with expectations outlined in the Policy PLDs.

Based on results of the review at the end of Day One, room facilitators and IDOE staff spent some time recalibrating groups as necessary during the morning of Day Two. Once panelists completed the first assigned grade for content areas in subgroups, groups presented their standards to the room to calibrate the entire grade at the room level. When the Grades 3—5 and the Grades 6—8 rooms for Math and ELA each completed their first grade, they met for a cross-grade articulation to ensure coherence between groups. Once each group came to consensus on PLDs for their first grade, they moved to their second grade. Math and ELA worked on the grades that bridged the two groups: Grades 5 and 6. The Math and ELA groups followed the same process they used for the initial grade but referenced the Range PLDs for that grade to ensure coherence and consistency. Once science completed Grade 4, they moved on to Grade 6. Biology and Social Studies adjourned when they completed their Range PLDs for their sole grade.

On Day Three, the ELA and Math groups completed the second grade of Range PLDs. When both groups within a content area finished working, they met together to vertically articulate their Range PLDs to ensure coherence across grades. The groups paid attention to standards that overlapped between the two grades. Once they completed this task, ELA and Math groups completed Grades 3 and 8. They followed the same process used for the other grades, referencing those Range PLDs to ensure coherence and consistency. Once Science completed Grades 4 and 6, they met to ensure coherence across grades, and then adjourned.

On Day Four, Math and ELA content area groups met for a cross-grade articulation discussion. They compared the expectations for similar standards to ensure a sensible progression of rigor. The committee primarily focused on examining Level 3, since this level is considered the entry point for college-readiness. The group first conducted articulation across Grades 3, 4, and 5, then across Grades 6, 7, and 8.

On the afternoon of Day Four, the policy review panel convened to review the Range PLDs. The panelists met to ensure that the Range PLDs were consistent with the goals of the Policy PLDs. The panel consisted of Indiana stakeholders, including:

- Representatives from the State Board of Education;
- Indiana Department of Education leadership;
- State ELL Director;
- Special Education Director; and

- Representatives from higher education.

AIR provided a bundle of materials to the panel on the morning of the meeting, including:

- Policy PLDs for each content area and grade;
- Range PLDs for each content area and grade;
- Applicable notes about discussions and potential revisions; and
- Guiding questions to focus discussion. For example:
    - Do the level 3 Range PLDs convey an appropriate sense of rigor consistent with the Policy PLDs?
    - Do the Range PLDs for each grade reflect an appropriate progression of rigor across proficiency levels?
    - Do the PLDs reflect the increase in complexity of the standards across the grade levels?

After the June 2018 educator workshop, AIR and IDOE revised the PLDs based on feedback from the policy review panel. AIR 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.

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)

2018–2019

Volume 3
Test Administration

ACKNOWLEDGMENTS

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

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

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

1. INTRODUCTION

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

Assessment instruments should 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 American Institutes for Research (AIR)'s quality monitoring (QM) system and test irregularity investigation process to detect cheating, monitor item quality in real-time, and evaluate test integrity.

2. TESTING PROCEDURES AND TESTING WINDOWS

Administering the 2018–2019 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. Without the proper training and coordination of these individuals, the standardization of the test administration could have been compromised. IDOE worked with AIR 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 2018-2019 ILEARN assessments. These practice tests contained sample test items similar to the test items that students would encounter on the ILEARN assessments in order to help students become familiar with the item types that would be presented to the students on the online or paper-and-pencil assessments. Indiana students also had the opportunity to interact with released, non-secure items on a public-facing Released Items Repository (RIR) assessments available on the ILEARN portal. The ILEARN RIR was deployed in May 2018, which allowed students to have online access to the items for nine months prior to the opening of the testing window.

The ILEARN assessments were administered in multiple segments over multiple days. The test segments administered for each content area were as follows:

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

The ILEARN assessments were untimed, but timing estimates were included in the ILEARN Test Administrator's Manuals (TAMs) (Appendices L through N in this volume) to ensure that schools had resources available to create local testing schedules. The ILEARN testing window for grades 3–8 was April 16 through May 17, 2019. The fall Biology test was available from December 4 through December 20, 2018, and the winter Biology test was available February 11 through February 28, 2019. The spring Biology and U.S. Government tests were available April 16 through May 24, 2019.

2.1 ELIGIBLE STUDENTS

All students enrolled in tested grade levels/courses participated in the Spring 2019 ILEARN administration with or without accommodations. 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/subject in which they are receiving instruction. All testing is administered on the basis of the student's enrolled grade, and off-grade testing is not available for ILEARN.

- **Public and Private School Students.** Students enrolled in Indiana public, charter, accredited nonpublic, and Choice schools were required to participate in grade- and course-level appropriate ILEARN assessment(s).
- **Home Education Program Students.** Students who received instruction at home and were registered appropriately with their corporation office as Home Education Program students were eligible to participate in statewide assessments. If parents or guardians identified an ILEARN assessment as a selected measure of their child's annual progress, students could participate in an ILEARN administration, as directed by the Corporation Test Coordinator (CTC).
- **English Learners (ELs).** All ELs enrolled in tested grade levels and courses were expected to participate in all ILEARN assessments, including English/Language Arts, regardless of how long these students had been enrolled in a U.S. school. Mathematics, Science, and Social Studies assessments are all available in stacked Spanish in the online Test Delivery System (TDS). Stacked Spanish is represented on the screen with the stimuli, passage, and item all appearing in both Spanish and English for students whose test setting language is Spanish.
- **Students with Disabilities.** Indiana has established procedures to ensure the inclusion in statewide testing of all public elementary and secondary school students with disabilities. Federal and state laws require that all students participate in the state testing system. In Indiana, a student on an Individualized Education Program (IEP) participates under one of these three general options:
    1. ILEARN without accommodations
    2. ILEARN with approved accommodations
    3. Indiana Alternate Measure (I AM) Alternate Assessment

Per the Individuals with Disabilities Education Improvement Act (IDEA) and Title 511 Article 7-Special Education, published December 2014 by the Indiana State Board of

Education, decisions regarding which assessment option a student will participate in are made annually by the student's IEP team and are based on the student's curriculum, present levels of academic achievement, functional performance, and learning characteristics. Decisions cannot be based on program setting, category of disability, percentage of time in a particular placement or classroom, or any considerations regarding a school's Adequate Yearly Progress (AYP) designation.

If a student requested an extraordinary exemption option due to a medical complexity, he or she may have been exempt from participating in statewide, standardized assessments pursuant to the provisions of School Accountability, a letter requesting the exemption is required.

2.2 TESTING ACCOMMODATIONS

Students participating in the online ILEARN assessment were able to use the designated standard online testing features in the TDS. These features included the ability to select an alternate background and font color, mouse pointer size and color, and font size before testing. During the tests, students could zoom in and zoom out to increase or decrease the size of text and images; highlight items and passages (or sections of items and passages); cross out response options by using the strikethrough function; use a notepad to make notes; and mark a question for review using the flag function.

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

The accommodations available for eligible students participating in the ILEARN assessments are described in the various test administrator manuals (TAMs) (Appendices J, K, and L of this report volume), which were accessible to schools before and during testing in the Resources section of the ILEARN Portal.

The ILEARN assessments provide two categories of assessment features to students. These include designated features and accommodations, both embedded and non-embedded in the TDS. Section 3.2 of Volume 1 of this technical report lists the allowed accommodations and the number of students who were provided with accommodations during the 2018-2019 ILEARN test administration.

Table 1 provides a list of designed features and accommodations that were offered in the 2018-2019 administration. Designated features for the ILEARN are those supports that are available for use by any student for whom the need has been indicated by an educator (or team of educators with parent/guardian and student). The *Online Test Delivery System (TDS) User Guide* at the ILEARN portal (Appendix A of this report volume) provides instructions on how to access and use these features.

Table 1: Designated Features and Accommodations Available in Spring 2019

Designated Features	Accommodations
Embedded	
Color contrast (Onscreen) Glossaries (Language) Language Masking Mouse pointer Print size Translation Stacked Spanish	American Sign Language (ASL) Audio Transcriptions Closed Captioning Permissive Mode Print on Demand Streamline Text-to-Speech
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	Alternate Indication of a Response Paper Booklet Braille Transcript for Audio Items Large Print Booklet Read-Aloud 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 Color Acetate Film for Paper Assessments Calculator

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

Students who required online accommodations (e.g., text-to-speech) were provided the opportunity to participate in practice activities for the statewide assessments with appropriate allowable accommodations. Test settings and accommodations were required to be identified in TIDE before starting an online test session. Some settings and accommodations could not be changed once a student started the test.

If an EL or a student with an IEP or Section 504 Plan used any accommodations during the test administration, this information was recorded by the Test Administrator (TA) in his or her required administration information and captured by AIR in the database of record (DoR). AIR included this data in the state output student data score files (SDFs) provided to IDOE.

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; and
- 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); and
- Documentation that the requested accommodations had been regularly used for instruction.

2.3 AVAILABLE ACCOMMODATIONS

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

For eligible students with IEPs or Section 504 Plans participating in paper-based assessments, the following accommodations were available:

- Contracted UEB braille and UEB Nemeth for Math.

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

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

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

3. ADMINISTRATOR TRAINING

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

TAs were required to complete the online AIR TA Certification Course before administering the test. There were also several training modules developed by AIR in collaboration with IDOE to facilitate test administration. The modules included topics on AIR systems, test administration, and accessibility and accommodations. These modules are included in the appendices to this volume of the technical report.

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

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

The User Support sections of both the *Online Test Delivery System (TDS) User Guide* (Appendix A of this report volume) and the *Test Information Distribution Engine (TIDE) User Guide* (Appendix E of this report volume) provided instructions that addressed technology challenges that could occur during test administration. The AIR Help Desk collaborated with IDOE to provide support to Indiana schools as they administered the state assessment.

3.1 ONLINE ADMINISTRATION

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

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

Roles and Responsibilities in the Online Testing Systems

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

CTCs

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

CITCs

CITCs were responsible for ensuring that testing devices were properly configured to support testing and coordinating participation in the January 2019 statewide 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 AIR'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 User Guide* (Appendix E of this report volume) for more information.

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

TAs

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

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

assessment according to the directions found in the guide. TAs were required to report to the STC any deviation in test administration, at which time the STC was required to report it to the CTC. Then, if necessary, the CTC was to report it to IDOE. TAs also ensured that the only available resources were those allowed for specific tests were available tests, and no additional resources were being used during administration of the ILEARN assessment.

For the ELA component of the online ELA assessment, students in grades 3–8 were required to have headphones or earbuds. There were no technical specifications for either device. IDOE did not provide headphones or earbuds; rather, the schools provided them, or students could use their own. Headphones should have been checked prior to the first day of testing to ensure they functioned properly with the computer or device the students would use for the assessment. TAs were also instructed to make sure that the students used their headphones or earbuds on the ILEARN practice test. On the day of ILEARN testing, to further verify that headphones were functional, a sound check was built into the sign-in process of the online assessment, and students were asked to confirm that headphones or earbuds were working prior to entering the test.

3.2 TEST ADMINISTRATION RESOURCES

The list of webinars and training resources available to corporations and schools for the 2018-2019 ILEARN administration is provided below. All training materials were available online at the ILEARN Portal. (PDFs of these resources have also been included in this technical report as Appendices J, Q–V, and S–AC, respectively.) Test administration resources comprising various tutorials and documents (user guides, manuals, quick guides, etc.) were available through the ILEARN Portal.

- **Test Administrator Certification Course:** All educators who administered the ILEARN assessment were required to complete an online TA Certification Course.
- **Accessibility and Accommodations Implementation and Setup Module:** This online module provided information on accessibility and accommodations in Indiana for the ILEARN tests.
- **Understanding Indiana's New Assessment System Webinar Module:** This online module provided an overview of the new ILEARN assessment to prepare parents, educators, and administrators for what to expect from the 2018-2019 assessments.
- **Computer-Adaptive Tests Webinar Module:** This online module described computer-adaptive-testing and the student test experience.
- **Why It Is Important to Assess Webinar Module:** This online module illustrated the importance of statewide testing.
- **Student Interface Training Webinar Module:** This online module provided information and a step-by-step guide through the Student Interface in the TDS.
- **Test Administrator Training Webinar Module:** This online module provided information and a step-by-step guide through the TA Interface in the TDS.

- **Request an Item Rescore Webinar Module:** This online module provided additional information regarding Indiana legislation that allows a principal or parent/guardian to request an item rescore for handscored items on the ILEARN tests.
- **Test Administration Overview Webinar Module:** This module provided a general overview of the TA role in the test administration process, including key responsibilities before, during, and after the testing window.
- **Test Information Distribution Engine (TIDE) Webinar Module:** This module provided a general overview of TIDE and the features applicable to educators and administrators before, during, and after testing.
- **Test Delivery System (TDS) Webinar Module:** This module provided a general overview of AIR's TDS and the features available in both the TA Interface and the Student Interface within TDS.
- **Online Reporting System (ORS) Webinar Module:** This module provided a general overview of the ORS where student scores, including individual scores and aggregate scores, are displayed after students complete the ILEARN assessments.
- **Technology Requirements for Online Testing Webinar Module:** This module provided technology requirements for corporation and school technology coordinators to ensure that their testing devices are set up properly before testing.
- **How the Scoring Process Works Webinar Module:** This module provided information for educators to better understand the scoring process that the tests go through prior to reporting.

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

Table 2: User Guides and Manuals

Resource	Description
Online Test Delivery System (TDS) User Guide	This user guide supports TAs who manage testing for students participating in the ILEARN practice tests, released item repository tests, operational tests.
Technology Setup for Online Testing Quick Guide	This document explains in four steps how to set up technology in Indiana corporations and schools.
2018–2019 Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux	This manual provides information about hardware, software, and network configurations for running various testing applications provided by American Institutes for Research (AIR).
Indiana Online Practice Test User Guide	This user guide provides an overview of the ILEARN Practice Test.
Test Information Distribution Engine (TIDE)	This user guide describes the tasks performed in the Test Information Distribution Engine (TIDE) for ILEARN assessments.

Braille Requirements Manual for Online Testing	This manual provides an overview of how to ensure your computer devices are set up properly to successfully administer the online Braille assessments for ILEARN.
Online Reporting System (ORS) User Guide	This user guide provides an overview of the different features available to educators to support viewing student scores for the ILEARN assessment.
2018–2019 Indiana Accessibility and Accommodations Guidance	The accessibility manual establishes the guidelines for the selection, administration, and evaluation of accessibility supports for instruction and assessment of all students, including students with disabilities, English learners (ELs), ELs with disabilities, and students without an identified disability or EL status.

Department Resources and Support

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

- Weekly newsletter distributed via email from the IDOE Office of Assessment to all officially designated CTCs in IDOE's database. The newsletter was titled "ILEARN Assessment Update" and included information on new announcements relevant to the ILEARN assessment, reminders of upcoming milestones, and a planning ahead section with important dates in the ILEARN program. The IDOE Office of Assessment contact information was also available at the end of each weekly newsletter so that corporations and schools could contact the IDOE directly if there were any questions.
- Communications via email memos took place on an "as needed" basis. These messages generally addressed specific issues that needed to be transmitted quickly to administrators and teachers in the field or important information that the IDOE wanted to ensure was clearly outlined due to its importance to the ILEARN program. An example of this was a memo the IDOE sent in Fall 2018 that contained extensive information about ILEARN scheduling and timing guidance, which was intended to help schools develop their ILEARN testing schedules. The distribution was to superintendents, principals, and school leaders.
- General information about the assessments was posted on the IDOE Office of Assessment website (https://www.doe.in.gov/assessment), such as dates of testing windows for all state-administered assessments. The Accessibility and Accommodations Guidance in the ILEARN Policy and Guidance section of IDOE's website was often referenced to address questions pertaining to accommodations and overall accessibility.

ILEARN Released Items Repository

The ILEARN Released Item Repository (RIR) is a collection of non-secure items and performance tasks that were available to the public via the ILEARN Portal and were intended to allow students, parents, and educators access to content that would be similar to what the student encountered when taking the ILEARN assessment. The ILEARN RIR was deployed on May 15, 2018, and remained available throughout the testing window.

A scoring guide accompanied the RIR, which provided educators the opportunity to see how their students performed on the assessment and where to focus efforts to improve student performance prior to the administration of the ILEARN assessment.

ILEARN Practice Tests

The purpose of the practice tests was to familiarize students with the TDS functionality and item types that students would experience on the ILEARN tests. The practice tests did not contain performance tasks and were not computer-adaptive. The items provided a grade-specific testing experience, including a variety of question types. The practice tests were not intended to guide classroom instruction. Users could also use the tutorials on each item to familiarize themselves with the different features and response instructions for each item type.

The ILEARN practice tests were deployed on October 1, 2018, and remained available throughout the testing window. The ILEARN practice tests were designed for use with the AIR Secure Browser and a supported web browser. The portal provided a list of supported web browsers on which to administer the practice tests. AIR's TDS delivered the practice tests in secure mode and used the same test delivery engine as the operational test to ensure that the student testing experience on the practice test matches the student experience for the operational test. IDOE required all students to take the practice test before taking the operational ILEARN test.

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

4. TEST SECURITY PROCEDURES

Test security involves maintaining the confidentiality of test questions and answers, and is critical in ensuring the integrity of a test and the validity of test results. Indiana has developed an appropriate set of policies and procedures to prevent test irregularities and ensure test result integrity. These include maintaining the security of test materials, assuring adequate trainings for everyone involved in test administration, outlining appropriate incident-reporting procedures, detecting test irregularities, and planning for investigation and handling of test security violations.

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

The test security procedures for ILEARN included the following:

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

To support these policies and procedures, IDOE leveraged security measures within AIR 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 AIR Secure Browser after 20 minutes of inactivity.

In developing the *ILEARN Test Coordinator's Manual* (Appendix O of this report volume) and the ILEARN TAMs (Appendices L through M of this report volume), IDOE and AIR 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 the breach form.

4.1 SECURITY OF TEST MATERIALS

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

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

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

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

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

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

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

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

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

4.2 Identifying Test Irregularities or Potential Test Security Concerns

AIR'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. AIR 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 2018-2019 ILEARN testing windows. In addition, response change analyses for multiple-choice and multi-select items were conducted. The last and next to last (if it existed) responses were compared and students or aggregates were flagged if the number or average number of wrong to right responses changes was above the flagging criteria.

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

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

4.3 Tracking and Resolving Test Irregularities

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

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

CTCs, NPSTCs, STCs, and TAs had to discuss the details of a test irregularity to determine whether test invalidation was appropriate. CTCs, NPSTCs, and STCs had to submit to IDOE a *Testing Concerns and Security Violations Report* when invalidating any

student test in response to a test security breach or interaction that compromised the integrity of the student's test administration.

During the testing window, TAs were also required to immediately report any test incidents (e.g., disruptive students, loss of Internet connectivity, student improprieties) to the STC. A test incident could include testing that was interrupted for an extended period due to a local technical malfunction or severe weather. STCs notified CTCs or NPSTCs of any test irregularities that were reported. CTCs or NPSTCs were responsible for completing test invalidations via TIDE. Schools managed the invalidation process based on local decisions or guidance from IDOE regarding test irregularities or test security concerns. This information was stored in TIDE for the school year and remained available until TIDE was updated for the 2019-2020 school year.

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

Table 3: Examples of Test Irregularities and Test Security Violations

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

4.4 AIR'S SYSTEM SECURITY

AIR 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 AIR's systems encrypt data at rest and in transit. ILEARN data resides on servers at Rackspace, AIR's online hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. Staff at both AIR 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 AIR networks from intrusion. AIR's systems maintain security and access logs that are regularly audited for login failures, which may indicate intrusion attempts. All of AIR'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).

AIR's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. AIR 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, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*.

Indiana Learning Evaluation and Readiness Network (ILEARN)

2018–2019

Volume 4
Evidence of Reliability and Validity

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

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, Kevin Clayton, and Kyra Bilenki. Major contributors from IDOE include the Assessment Director, Assistant Assessment Director, and Program Leads.

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The State of Indiana implemented a new assessment program for operational use during the 2018–2019 school year: the Indiana Learning Evaluation Assessment Readiness Network (LEARN). The ILEARN replaced the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) in English/Language Arts (ELA), Mathematics, Science, and Social Studies. The assessments were delivered as online adaptive assessments for Mathematics and ELA and online fixed-form assessments for Science and Social Studies. Online accommodated and paper-and-pencil versions of the assessments were available to students whose Individualized Education Programs (IEPs) or Section 504 Plans indicated such a need. Table 1 displays the complete list of test administration methods for the 2018–2019 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 5, Section 1.2. The number of students who were provided with accommodations is presented in Volume 1, Section 2.2.

With the implementation of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic performance from ILEARN scores. This volume provides empirical evidence about the reliability and validity of the 2018–2019 ILEARN assessments.

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

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

disattenuated Pearson correlations among reporting categories per grade. Confirmatory factor analysis has also been performed using the second-order factor model. Additionally, local item independence, an assumption of unidimensional IRT, was tested using the Q_3 statistic.

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

1.1 RELIABILITY

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

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

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

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

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

The first four methods discussed above are classical methods of calculating reliability, and are not optimal for computer adaptive testing. While classical indicators provide a single estimate of the reliability of test forms, the precision of test scores varies with respect to the information value of the test at each location along the scale. For example, most fixed-form assessments target test information near important cut scores or near the population mean, so that test scores are most precise in targeted locations. Because adaptive tests target test information near each student's ability level, the precision of test scores may increase, especially for lower- and higher-ability students. Precision of individual test scores is critically important to valid test score interpretation and is provided along with test scores as part of all student-level reporting. In addition, the first two methods require multiple testing opportunities which are not available for ILEARN.

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

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

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

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

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

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

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

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

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

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

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

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

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

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the lack of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution, or near an important classification cut, and have less information at the tails of the score distribution. See Section 3.3, Test Information Curves and Standard Error of Measurement, for the derivation of heterogeneous errors in IRT.

1.2 VALIDITY

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

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

The first source of evidence for validity is the relationship between the test content and the intended test construct (see Section 4.2, Alignment of ILEARN Test Forms to the Content Standards and Benchmarks). In order for test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies, in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct (see Volume 2 of this technical report for details). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance.

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

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

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

The third source of evidence for validity is based on internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the

internal structure of tests (see Volume 1, Section 5.2). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (see Section 3, Reliability, and Section 5, Evidence of Internal-External Structure, for details).

A fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards* (AERA, APA, & NCME, 2014) divide this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs; conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multi-trait-multimethod matrix can be used (see Section 5.4, Convergent and Discriminant Validity, for details). Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends upon the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation specific or can be generalized across different settings and times. For example, sampling errors or range restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

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

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

2. PURPOSE OF ILEARN

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

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

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

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

The scale score and relative strengths and weaknesses at the reporting category (domain) level were provided for each student to indicate student strengths and weaknesses in different content areas of the test relative to the other areas and to the district and state. These scores help teachers tailor their instruction, provided that the scores are viewed with the usual caution that accompanies the use of reporting category scores. Thus, we must examine the reliability coefficients for these test scores and the validity of the test scores to support practical use of these tests across the state. Volume 6 of this technical report is the score interpretation guide and provides details on all generated scores and their appropriate uses and limitations.

3. EVIDENCE OF CONTENT VALIDITY

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

3.1 CONTENT STANDARDS

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

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

Table 2 through Table 5 present the reporting categories by grade and test, as well as the number of items measuring each category on the 2018-2019 tests. Reading Foundations in ELA Grade 3, Speaking and Listening in ELA Grades 3-8, and Process Standards in Mathematics Grades 3-8 were not reported as a separate reporting category, but were included only in the overall aggregate scale score calculations.

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

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

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

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

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

Grade	Reporting Category	Number of Items
4	Questioning and Modeling	12-13
	Investigating	12
	Analyzing, Interpreting, and Computational Thinking	11-12
	Explaining Solutions, Reasoning, and Communicating	11
6	Questioning and Modeling	11-12
	Investigating	11-12
	Analyzing, Interpreting, and Computational Thinking	12
	Explaining Solutions, Reasoning, and Communicating	13
Biology	Developing and Using Models to Describe Structure and Function	11
	Developing and Using Models to Explain Processes	10-11
	Analyzing Data and Mathematical Thinking	11-12
	Constructing and Communicating an Explanation	11
	Evaluating Claims with Evidence	11

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

Grade	Reporting Category	Number of Items
5	Civics and Government	16-17
	Geography and Economics	11
	History	12
U.S. Government	Functions of Government	19
	Historical Foundations of American Government	14
	Institutions and Processes of Government	20

4. RELIABILITY

4.1 MARGINAL RELIABILITY

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

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

The marginal reliability of a test is computed by integrating θ out of the TIF 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. Population parameters are assumed normal, $g(\theta) \sim N(0,1)$.

Table 6 presents the marginal reliability coefficients for all students. The marginal reliability coefficients for all subjects and grades range from 0.872 to 0.947, which is similar to other statewide standardized tests.

Table 6: Marginal Reliability Coefficients

Grade	Marginal Reliability
ELA 3	0.872
ELA 4	0.880
ELA 5	0.878
ELA 6	0.881
ELA 7	0.880
ELA 8	0.879
Mathematics 3	0.943
Mathematics 4	0.944

Grade	Marginal Reliability
Mathematics 5	0.938
Mathematics 6	0.947
Mathematics 7	0.934
Mathematics 8	0.940
Science 4	0.875
Science 6	0.899
Biology	0.915
Social Studies 5	0.874
U.S. Government	0.880

4.2 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

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

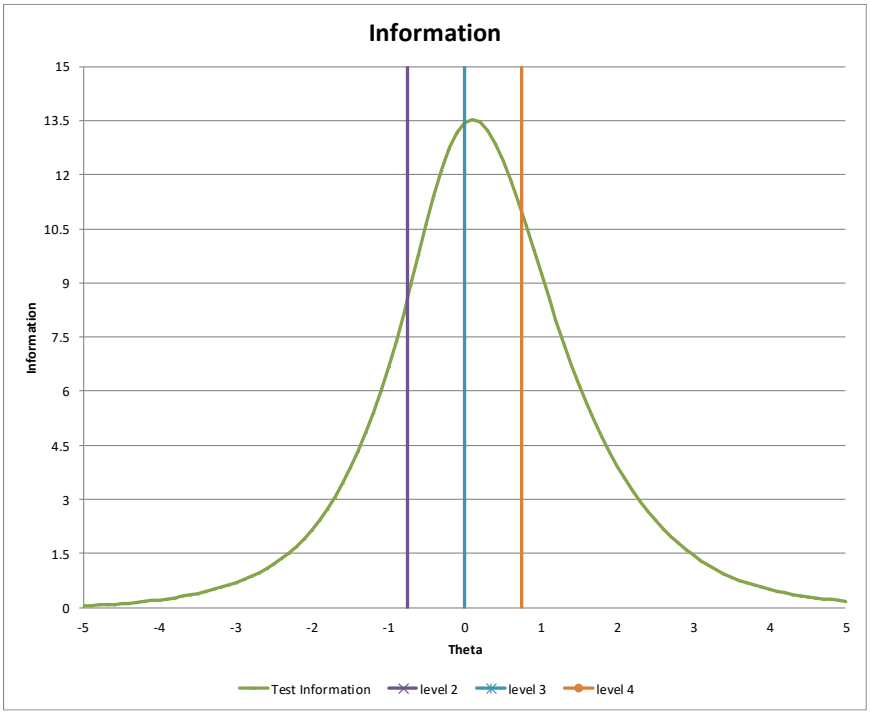
Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most-precise scores in this range. Where the curve is lower at the tails indicates that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

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

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

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

Figure 1: Sample Test Information Function



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

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

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented in Figure 2: Conditional Standard Error of Measurement (ELA) through Figure 4, instead of the TIFs for ELA, Mathematics, Science, and Social Studies. These plots are based on the scaled scores reported in 2019. Vertical lines represent the three performance category cut scores.

Figure 2: Conditional Standard Error of Measurement (ELA)

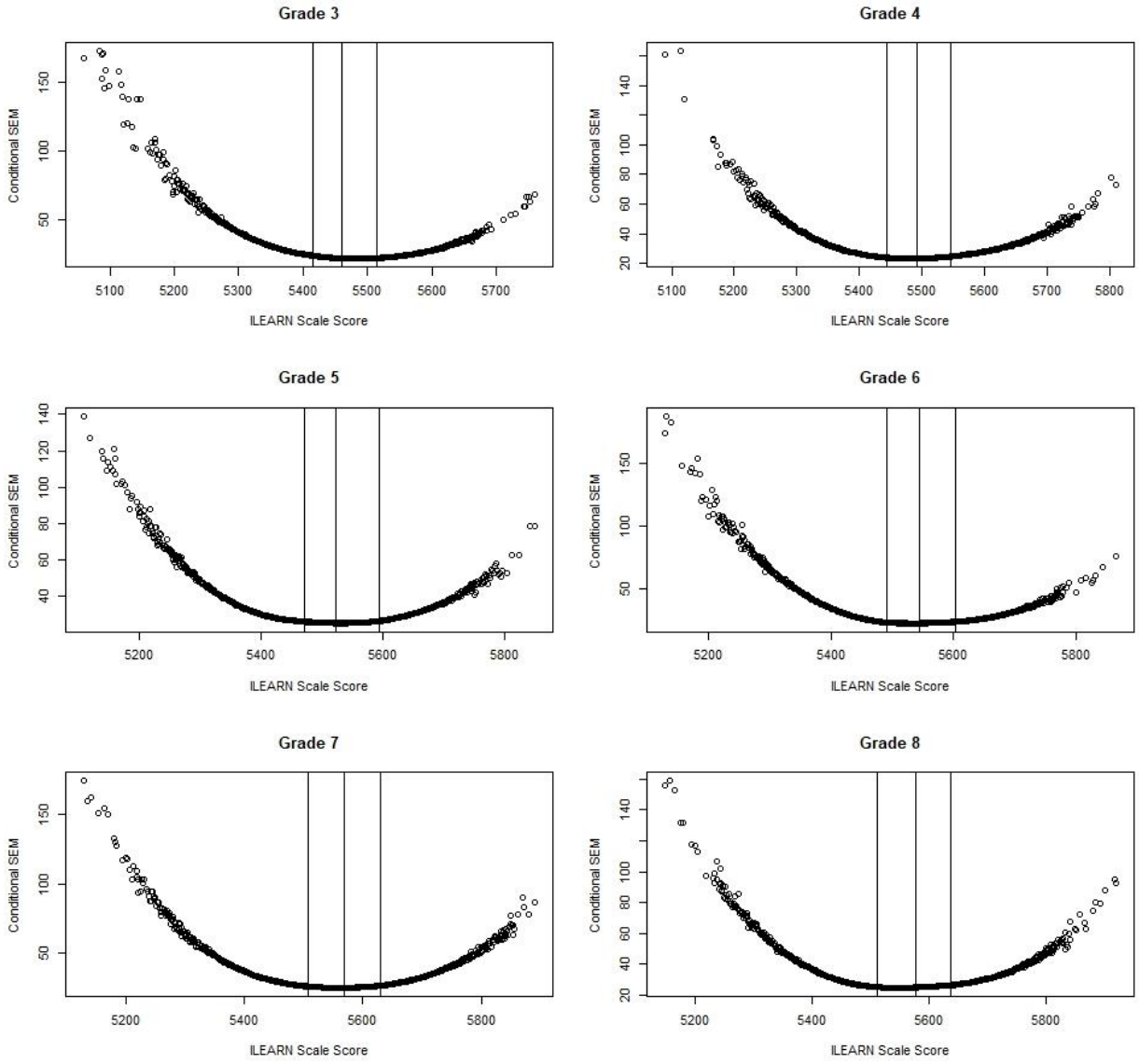


Figure 2: Conditional Standard Error of Measurement (Mathematics)

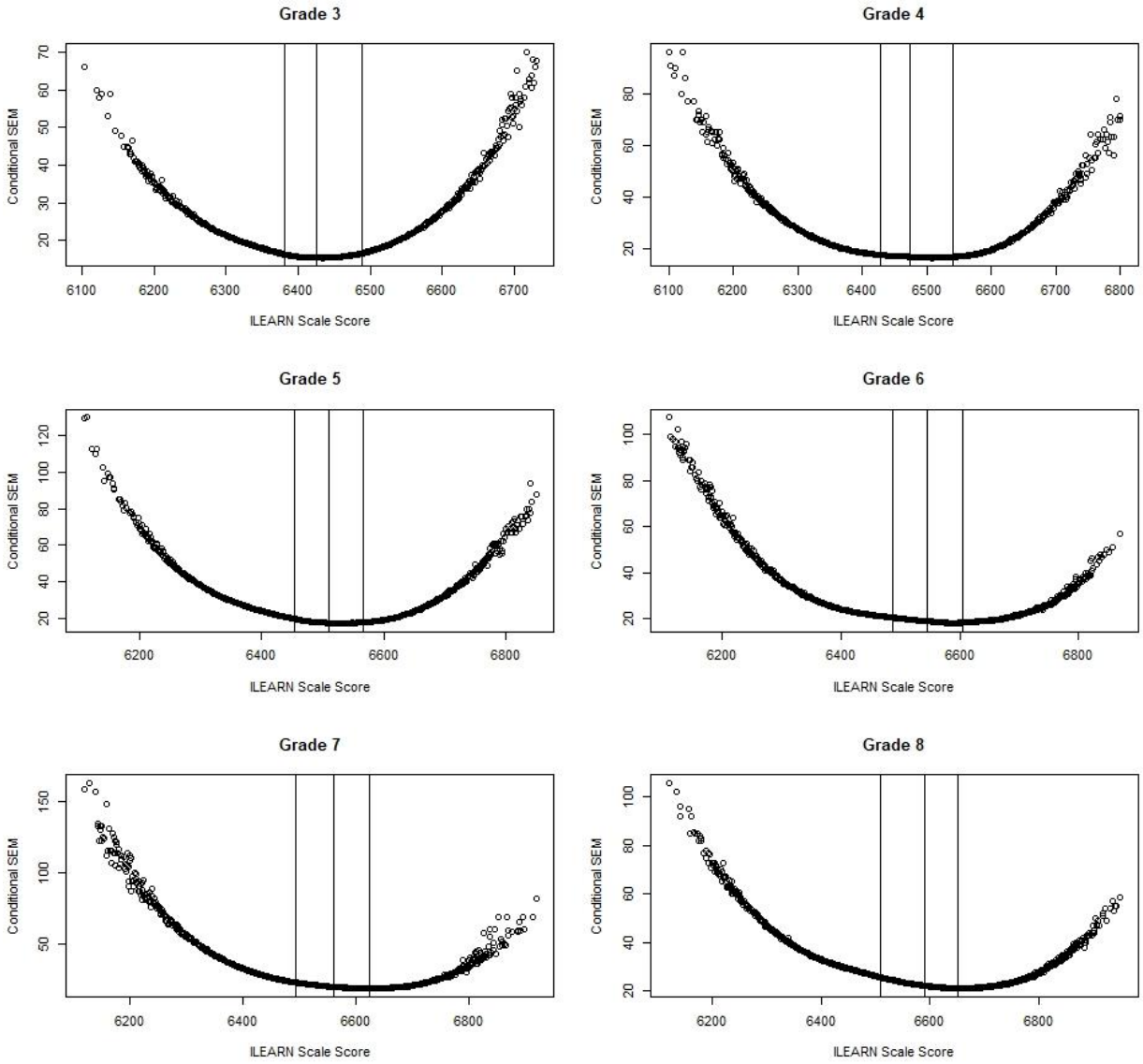


Figure 3: Conditional Standard Error of Measurement (Science)

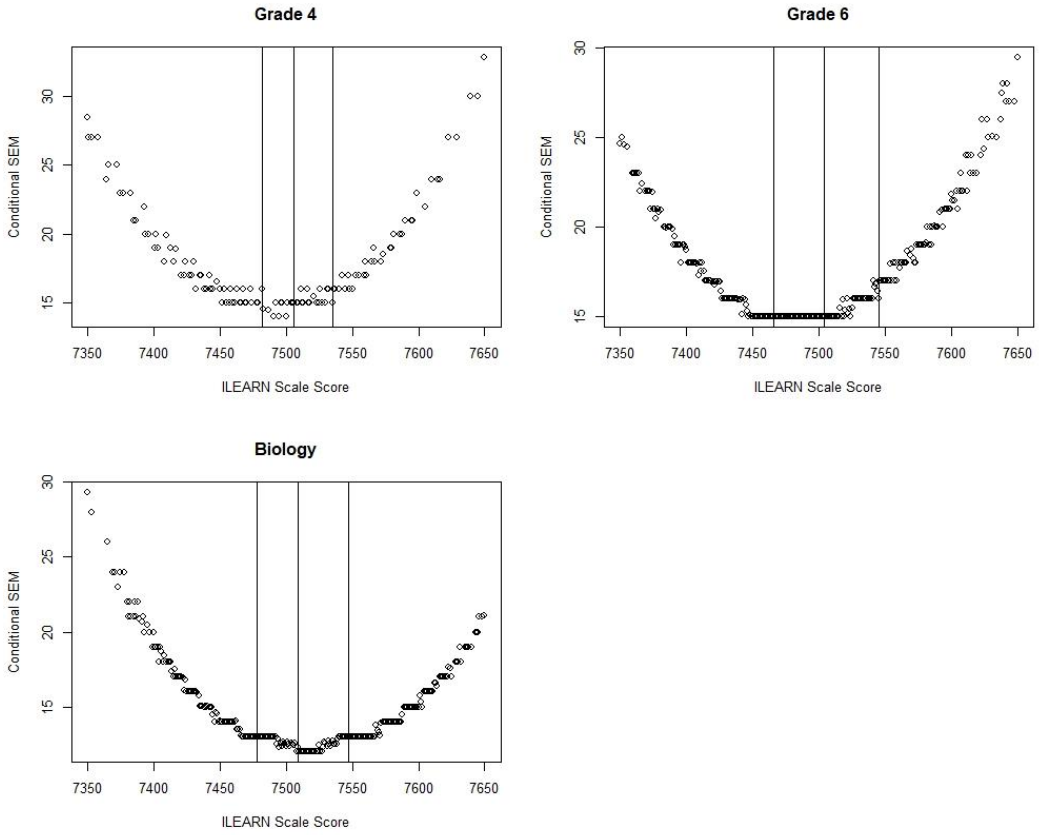
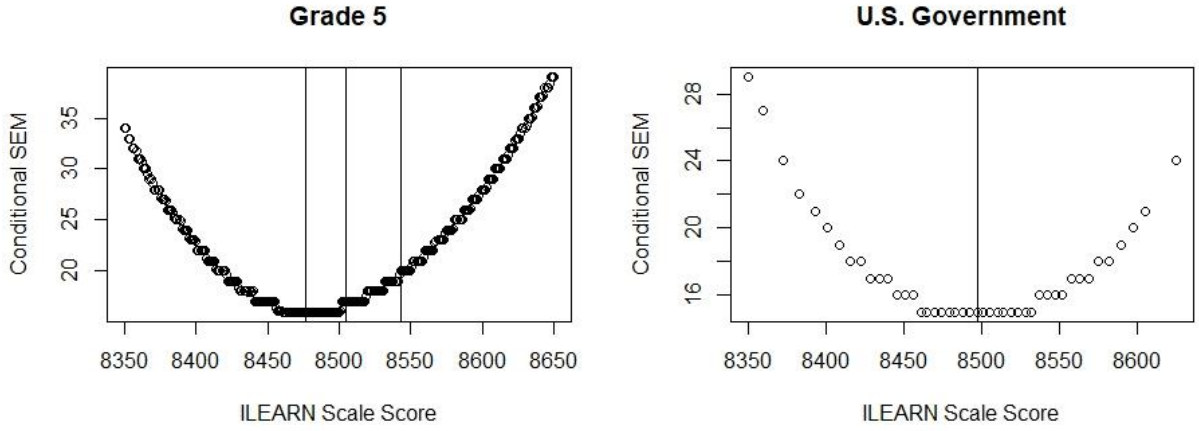


Figure 4: Conditional Standard Error of Measurement (Social Studies)



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

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

4.3 Reliability of Performance Classification

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

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

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

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

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

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

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

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

For these analyses, we used students from the Spring 2019 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 AIR's scoring engine.

Table 7: Descriptive Statistics

ELA Grade	Sample Size	Mean Theta	Standard Deviation of Theta	Mean Scale Score	Standard Deviation of Scale Scores
ELA 3	83,074	-0.67	0.92	5449.74	69.13
ELA 4	84,147	-0.25	1.01	5481.24	75.49
ELA 5	86,381	0.18	1.06	5513.26	79.85
ELA 6	85,833	0.46	0.98	5534.31	73.36
ELA 7	84,591	0.80	1.10	5559.97	82.16
ELA 8	82,991	0.97	1.06	5572.88	79.23
Mathematics 3	83,080	-0.83	1.01	6437.16	75.70
Mathematics 4	84,144	-0.31	1.04	6476.73	77.78
Mathematics 5	86,369	0.02	1.13	6501.15	84.83
Mathematics 6	85,817	0.36	1.24	6527.18	93.34
Mathematics 7	84,580	0.47	1.30	6535.57	97.61
Mathematics 8	82,991	0.67	1.44	6550.37	108.11
Science 4	84,068	-0.003	0.92	7499.94	46.14
Science 6	85,659	-0.002	1.04	7499.90	51.76
Biology	80,677	-0.03	0.93	7498.46	46.33
Social Studies 5	86,253	0.02	1.10	8500.82	54.94
U.S. Government	1,230	-1.01	1.03	8449.44	51.55

4.3.1 Classification Accuracy

The observed score approach (Rudner, 2001), implemented to assess classification accuracy, is based on the probability that the true score, θ, for student j is within performance level $l = 1,2,\cdots,\mathrm{L}$. This probability can be estimated from evaluating the integral

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

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

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

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

$$E_{vl} = \sum_{pl_i \, \epsilon \, v} p_{jl},$$

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 $\boldsymbol{E}$, 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

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

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

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

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

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. Here, the overall classification accuracy of the test ranges from 0.750 to 0.757 for ELA, from 0.821 to 0.831 for Mathematics, from 0.741 to 0.798 for Science, and was 0.754 for Social Studies grade 5 and 0.954 for U.S. Government. There is no industry standard, but these numbers suggest that misclassification would not be frequent in the population data.

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

Table 8: Classification Accuracy Index (ELA)

Grade	Overall Accuracy Index	Cut Accuracy Index		
		Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4
3	0.751	0.912	0.906	0.931
4	0.750	0.918	0.904	0.925
5	0.750	0.917	0.901	0.931
6	0.757	0.922	0.908	0.926
7	0.750	0.928	0.902	0.918
8	0.751	0.931	0.903	0.917

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	0.829	0.951	0.938	0.940
4	0.826	0.946	0.932	0.948
5	0.823	0.942	0.933	0.948
6	0.821	0.943	0.930	0.948
7	0.830	0.937	0.937	0.955
8	0.831	0.938	0.938	0.955

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	0.741	0.910	0.904	0.919
6	0.772	0.930	0.912	0.930
Biology	0.798	0.912	0929	0.956

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	0.754	0.907	0.908	0.930
U.S. Government*	0.954	0.954	--	--

*U.S. Government has only one cut.

4.3.2 Classification Consistency

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

The IRT-based approach (Guo, 2006) makes use of student-level item response data from the 2019 test administration. For the *j*th student, we can estimate a posterior

probability distribution for the latent true score and, from this, estimate the probability that a true score is above the cut as

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

where c is the cut score required for passing in the same assigned metric, θ_j is 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 12: False Classification Rates (ELA) through Table 15: False Classification Rates (Social Studies) provide the FPR and FNR for the ILEARN assessments. The FPR and FNR rates for the level 2/3 cut are between 0.09 and 0.12 in ELA, between 0.05 and 0.09 in Mathematics, between 0.06 and 0.11 in Science, and between 0.03 and 0.13 in Social Studies.

Table 12: False Classification Rates (ELA)

	1/2 cut		2/3 cut		3/4 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**	**FPR**	**FNR**
3	0.16	0.06	0.09	0.11	0.04	0.23
4	0.15	0.06	0.09	0.11	0.04	0.23
5	0.16	0.06	0.09	0.12	0.03	0.27
6	0.16	0.05	0.09	0.11	0.05	0.22

7	0.17	0.04	0.09	0.11	0.04	0.24
8	0.19	0.04	0.10	0.10	0.05	0.24

Table 13: False Classification Rates (Mathematics)

	1/2 cut		2/3 cut		3/4 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**	**FPR**	**FNR**
3	0.12	0.03	0.07	0.05	0.04	0.14
4	0.12	0.03	0.07	0.07	0.03	0.14
5	0.11	0.04	0.06	0.08	0.03	0.14
6	0.10	0.04	0.06	0.08	0.03	0.15
7	0.11	0.04	0.05	0.08	0.02	0.13
8	0.10	0.05	0.05	0.09	0.02	0.14

Table 14: False Classification Rates (Science)

	1/2 cut		2/3 cut		3/4 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**	**FPR**	**FNR**
4	0.11	0.07	0.08	0.11	0.05	0.21
6	0.16	0.04	0.08	0.10	0.04	0.21
Biology	0.14	0.06	0.06	0.10	0.03	0.17

Table 15: False Classification Rates (Social Studies)

	1/2 cut		2/3 cut		3/4 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**	**FPR**	**FNR**
5	0.15	0.06	0.09	0.10	0.03	0.21
U.S. Government	--	--	0.03	0.13	--	--

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

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

Classification consistency with classification accuracy results are presented in Table 12 through Table 14. In cut 1 and cut 2 and in cut 2 and cut 3 results, all accuracy values are higher than 0.90, and consistency values are around 0.85. Across all grades and subjects and in all performance levels, classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy because the consistency is based on two tests with measurement errors, while the accuracy is based on one test with a measurement error and the true score. The accuracy

and consistency rates for each performance level are higher for the levels with smaller standard error.

Table 12. Classification Accuracy and Consistency (Cut 1 and Cut 2)

Grade	Accuracy	Consistency
ELA 3	0.912	0.872
ELA 4	0.918	0.882
ELA 5	0.917	0.881
ELA 6	0.922	0.890
ELA 7	0.928	0.898
ELA 8	0.931	0.904
Mathematics 3	0.951	0.930
Mathematics 4	0.946	0.923
Mathematics 5	0.942	0.917
Mathematics 6	0.943	0.918
Mathematics 7	0.937	0.910
Mathematics 8	0.938	0.909
Science 4	0.910	0.883
Science 6	0.930	0.900
Biology	0.912	0.871
Social Studies 5	0.907	0.866
U.S. Government	0.954	0.932

Table 13. Classification Accuracy and Consistency (Cut 2 and Cut 3)

Grade	Accuracy	Consistency
ELA 3	0.906	0.863
ELA 4	0.904	0.861
ELA 5	0.901	0.856
ELA 6	0.908	0.865
ELA 7	0.902	0.858
ELA 8	0.903	0.858
Mathematics 3	0.938	0.912
Mathematics 4	0.932	0.902
Mathematics 5	0.933	0.905
Mathematics 6	0.930	0.899
Mathematics 7	0.937	0.911

Grade	Accuracy	Consistency
Mathematics 8	0.938	0.911
Science 4	0.904	0.883
Science 6	0.912	0.872
Biology	0929	0.896
Social Studies 5	0.908	0.866

Table 14. Classification Accuracy and Consistency (Cut 3 and Cut 4)

Grade	Accuracy	Consistency
ELA 3	0.931	0.863
ELA 4	0.925	0.861
ELA 5	0.931	0.856
ELA 6	0.926	0.865
ELA 7	0.918	0.858
ELA 8	0.917	0.858
Mathematics 3	0.940	0.913
Mathematics 4	0.948	0.926
Mathematics 5	0.948	0.927
Mathematics 6	0.948	0.927
Mathematics 7	0.955	0.938
Mathematics 8	0.955	0.937
Science 4	0.919	0.884
Science 6	0.930	0.900
Biology	0.956	0.935
Social Studies 5	0.930	0.898

4.4 Precision at Cut Scores

Table 15 through Table 18 present mean CSEM at each performance level by grade and subject. These tables also include performance-level cut scores and associated CSEM. The ILEARN test scores are somewhat more precise for test scores near the middle of the scale, especially around the At Proficiency performance standard cut. The following tables also show that test scores remain precise even for students in the lowest and highest performance levels.

Table 15: 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	28.699	--	--
	2	22.133	5416	23.244
	3	21.244	5460	21.249
	4	23.319	5515	21.586
4	1	28.271	--	--
	2	23.407	5444	23.814
	3	23.774	5493	23.237
	4	27.279	5547	24.678
5	1	30.198	--	--
	2	25.662	5472	26.025
	3	25.628	5524	25.395
	4	29.229	5595	26.664
6	1	28.698	--	--
	2	22.248	5492	22.813
	3	22.564	5544	22.172
	4	25.313	5604	23.365
7	1	31.583	--	--
	2	25.382	5507	26.205
	3	25.646	5568	25.185
	4	29.993	5629	26.731
8	1	29.808	--	--
	2	24.982	5511	25.289
	3	25.704	5577	25.157
	4	29.172	5638	26.705

Table 16: 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	19.198	--	--
	2	15.686	6382	16.278
	3	15.701	6425	15.415
	4	20.058	6488	16.609
4	1	20.471	--	--

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
	2	16.996	6429	17.310
	3	16.415	6474	16.716
	4	18.905	6541	16.591
5	1	25.014	--	--
	2	18.081	6453	19.509
	3	17.251	6510	17.261
	4	20.960	6566	17.500
6	1	24.893	--	--
	2	19.638	6488	20.500
	3	18.434	6545	18.947
	4	19.655	6605	18.318
7	1	31.086	--	--
	2	20.923	6493	22.764
	3	18.907	6562	19.423
	4	19.944	6625	18.558
8	1	31.421	--	--
	2	23.757	6509	25.683
	3	21.621	6590	22.273
	4	22.394	6651	21.204

Table 17: 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	16.284	--	--
	2	14.566	7482	17.333
	3	15.424	7506	15
	4	17.915	7535	15
6	1	16.203	--	--
	2	15.001	7466	15
	3	15.598	7504	15.006
	4	19.211	7545	16.042
Biology	1	14.303	--	--
	2	12.763	7478	13.007
	3	12.333	7509	12.325
	4	14.094	7547	13.000

Table 18: 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	17.541	--	--
	2	15.995	8477	15.997
	3	17.750	8502	16.989
	4	25.377	8543	19.995
U.S. Government	1	18.050	--	--
	2	15.692	8497	15.000

4.5 WRITING PROMPTS INTER-RATER RELIABILITY

All ELA writing prompts were hand-scored by a human with a 15% second read. The basic method to compute inter-rater reliability is percentage agreement. As seen in Table 19, 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 exact agreement was 2/4, 50%, and the adjacent and non-adjacent percentages were 25% each.

Table 19: 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, inter-rater reliability monitors how often scorers are in exact agreement with each other and ensures that an acceptable agreement rate is maintained. The calculations for inter-rater reliability 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 20 displays rater-agreement percentages. The percentage of exact agreement between two raters ranged from 61% to 79%. The percentage of adjacent rating was between 20% and 37%. The non-adjacent percentages fell between 1% and 2%. The total number of processed responses does not necessarily correspond to the number of students participating in the Writing portion. These numbers could potentially be higher, as some students are scored more than once when rescoring for some responses, as requested.

Table 20: Inter-Rater Reliability

Grade	Dimension	% Exact	% Adjacent	% Not Adjacent	Total Number of Processed Responses
3	Purpose, Focus, & Organization	67	31	2	8743
	Evidence & Elaboration	67	31	2	
	Conventions	68	31	1	
4	Purpose, Focus, & Organization	66	32	2	9683
	Evidence & Elaboration	66	32	2	
	Conventions	69	30	1	
5	Purpose, Focus, & Organization	64	34	1	11,534
	Evidence & Elaboration	65	34	1	
	Conventions	68	32	1	
6	Purpose, Focus, & Organization	66	33	1	11,543
	Evidence & Elaboration	67	32	1	
	Conventions	76	23	1	
7	Purpose, Focus, & Organization	63	36	1	11,412
	Evidence & Elaboration	64	35	1	
	Conventions	72	27	1	
8	Purpose, Focus, & Organization	62	37	2	11,749
	Evidence & Elaboration	61	37	2	
	Conventions	79	20	1	

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 21.

Table 21: Weighted Kappa Coefficients

Grade	*N*	Purpose, Focus, & Organization	Evidence & Elaboration	Conventions
3	8743	0.696	0.691	0.464
4	9683	0.704	0.701	0.472
5	11534	0.719	0.717	0.418
6	11543	0.647	0.665	0.398
7	11412	0.658	0.664	0.425
8	11749	0.652	0.656	0.417

5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE

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

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

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

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

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

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

5.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 22 through Table 25 present the observed correlation matrix of the reporting category raw scores for each subject area. In ELA, the correlations among the reporting categories ranged from 0.55 to 0.67. In Mathematics, the correlations were between 0.60 and 0.78. In Science, the correlations among reporting categories ranged from 0.59 to 0.70. In Social Studies, the correlations ranged from 0.65 to 0.73.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error

at the strand level, given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations, as either high or low, should be made cautiously.

Table 26 through Table 29 display disattenuated correlations. Disattenuated values greater than 1.00 are reported as 1.00*. The overall average disattenuated correlation was 0.89 for ELA, 0.95 for Mathematics, 0.99 for Science, and 1.03 for Social Studies. These values suggest that validity evidence of internal structure is supported.

Table 22: Observed Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	12-13	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-11	0.67	1.00	
	Writing (Cat3)	7-8	0.60	0.55	1.00
4	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11	0.64	1.00	
	Writing (Cat3)	7-8	0.63	0.57	1.00
5	Key Ideas and Textual Support/Vocabulary (Cat1)	11-13	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	12-13	0.60	1.00	
	Writing (Cat3)	7-8	0.64	0.57	1.00
6	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10	0.66	1.00	
	Writing (Cat3)	6-8	0.62	0.60	1.00
7	Key Ideas and Textual Support/Vocabulary (Cat1)	11-13	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-13	0.62	1.00	0.60
	Writing (Cat3)	6-8	0.65	0.60	1.00
8	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-12	0.60	1.00	0.57
	Writing (Cat3)	7-8	0.66	0.57	1.00

Table 23: Observed Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	10	1.00			

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	Computation (Cat2)	13	0.78	1.00		
	Geometry and Measurement (Cat3)	9	0.74	0.73	1.00	
	Number Sense (Cat4)	11	0.73	0.71	0.71	1.00
4	Algebraic Thinking and Data Analysis (Cat1)	9-10	1.00			
	Computation (Cat2)	11	0.73	1.00		
	Geometry and Measurement (Cat3)	10-11	0.71	0.70	1.00	
	Number Sense (Cat4)	11	0.73	0.74	0.73	1.00
5	Algebraic Thinking (Cat1)	12	1.00			
	Computation (Cat2)	11	0.74	1.00	0.66	0.70
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.69	0.66	1.00	
	Number Sense (Cat4)	11	0.73	0.70	0.66	1.00
6	Algebra and Functions (Cat1)	12	1.00			
	Computation (Cat2)	10	0.76	1.00		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.71	0.67	1.00	
	Number Sense (Cat4)	10	0.76	0.71	0.68	1.00
7	Algebra and Functions (Cat1)	11	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	9	0.68	1.00		
	Geometry and Measurement (Cat3)	10	0.61	0.60	1.00	
	Number Sense and Computation (Cat4)	11-12	0.75	0.74	0.65	1.00
8	Algebra and Functions (Cat1)	12	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	10	0.74	1.00		
	Geometry and Measurement (Cat3)	10	0.72	0.70	1.00	
	Number Sense and Computation (Cat4)	9	0.70	0.65	0.65	1.00

Table 24: Observed Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Questioning and Modeling (Cat1)	12-13	1.00				--
	Investigating (Cat2)	12	0.59	1.00			--
	Analyzing, Interpreting, and Computational Thinking (Cat3)	11-12	0.61	0.67	1.00		--
	Explaining Solutions, Reasoning, and Communicating (Cat4)	11	0.59	0.65	0.68	1.00	--
6	Questioning and Modeling (Cat1)	11-12	1.00				--
	Investigating (Cat2)	11-12	0.70	1.00			--

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Analyzing, Interpreting, and Computational Thinking (Cat3)	12	0.67	0.68	1.00		--
	Explaining Solutions, Reasoning, and Communicating (Cat4)	13	0.68	0.71	0.67	1.00	--
Biology	Developing and Using Models to Describe Structure and Function (Cat1)	11	1.00				
	Developing and Using Models to Explain Processes (Cat2)	10-11	0.66	1.00			
	Analyzing Data and Mathematical Thinking (Cat3)	11-12	0.61	0.62	1.00		
	Constructing and Communicating an Explanation (Cat4)	11	0.65	0.66	0.64	1.00	
	Evaluating Claims with Evidence (Cat5)	11	0.65	0.66	0.67	0.68	1.00

Table 25: Observed Correlation Matrix Among Reporting Categories (Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	16-17	1.00		
	Geography and Economics (Cat2)	11	0.67	1.00	
	History (Cat3)	12	0.71	0.65	1.00
U.S. Government	Functions of Government (Cat1)	19	1.00		
	Historical Foundations of American Government (Cat2)	14	0.69	1.00	
	Institutions and Processes of Government (Cat3)	20	0.73	0.65	1.00

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

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	12-13	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-11	0.99	1.00	
	Writing (Cat3)	7-8	0.91	0.89	1.00
4	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11	0.92	1.00	
	Writing (Cat3)	7-8	0.90	0.85	1.00
5	Key Ideas and Textual Support/Vocabulary (Cat1)	11-13	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	12-13	0.94	1.00	
	Writing (Cat3)	7-8	0.89	0.85	1.00

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
6	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10	1.00*	1.00	
	Writing (Cat3)	6-8	0.88	0.91	1.00
7	Key Ideas and Textual Support/Vocabulary (Cat1)	11-13	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-13	0.93	1.00	
	Writing (Cat3)	6-8	0.92	0.87	1.00
8	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1.00		
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10-12	0.91	1.00	
	Writing (Cat3)	7-8	0.91	0.86	1.00

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

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	10	1.00			
	Computation (Cat2)	13	1.00*	1.00		
	Geometry and Measurement (Cat3)	9	1.00*	1.00*	1.00	
	Number Sense (Cat4)	11	0.95	0.95	0.98	1.00
4	Algebraic Thinking and Data Analysis (Cat1)	9-10	1.00			
	Computation (Cat2)	11	0.98	1.00		
	Geometry and Measurement (Cat3)	10-11	0.98	0.92	1.00	
	Number Sense (Cat4)	11	0.97	0.95	0.95	1.00
5	Algebraic Thinking (Cat1)	12	1.00			
	Computation (Cat2)	11	0.99	1.00		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.98	0.97	1.00	
	Number Sense (Cat4)	11	0.98	0.96	0.96	1.00
6	Algebra and Functions (Cat1)	12	1.00			
	Computation (Cat2)	10	0.95	1.00		
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	9	0.93	0.89	1.00	
	Number Sense (Cat4)	10	0.97	0.92	0.92	1.00
7	Algebra and Functions (Cat1)	11	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	9	0.90	1.00		
	Geometry and Measurement (Cat3)	10	0.85	0.86	1.00	
	Number Sense and Computation (Cat4)	11-12	0.99	1.00*	0.87	1.00

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
8	Algebra and Functions (Cat1)	12	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	10	0.94	1.00		
	Geometry and Measurement (Cat3)	10	0.94	0.94	1.00	
	Number Sense and Computation (Cat4)	9	0.95	0.91	0.95	1.00

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

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
4	Questioning and Modeling (Cat1)	12-13	1.00				-
	Investigating (Cat2)	12	0.98	1.00			-
	Analyzing, Interpreting, and Computational Thinking (Cat3)	11-12	0.99	1.03	1.00		-
	Explaining Solutions, Reasoning, and Communicating (Cat4)	11	1.00	1.00*	1.00*	1.00	-
6	Questioning and Modeling (Cat1)	11-12	1.00				-
	Investigating (Cat2)	11-12	1.00*	1.00			-
	Analyzing, Interpreting, and Computational Thinking (Cat3)	12	1.00*	1.00*	1.00		-
	Explaining Solutions, Reasoning, and Communicating (Cat4)	13	1.00	1.00*	1.00	1.00	-
Biology	Developing and Using Models to Describe Structure and Function (Cat1)	11	1.00				
	Developing and Using Models to Explain Processes (Cat2)	10-11	0.94	1.00			
	Analyzing Data and Mathematical Thinking (Cat3)	11-12	0.92	0.94	1.00		
	Constructing and Communicating an Explanation (Cat4)	11	0.96	0.97	0.99	1.00	
	Evaluating Claims with Evidence (Cat5)	11	0.93	0.95	1.00*	0.99	1.00

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

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government (Cat1)	16-17	1.00		
	Geography and Economics (Cat2)	11	1.00*	1.00	
	History (Cat3)	12	1.00	1.00*	1.00
U.S. Government	Functions of Government (Cat1)	19	1.00		
	Historical Foundations of American Government (Cat2)	14	1.00*	1.00	
	Institutions and Processes of Government (Cat3)	20	1.00*	1.00*	1.00

5.2 CONFIRMATORY FACTOR ANALYSIS

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

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

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

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

5.2.1 Factor Analytic Methods

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

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

in this section establishes some validity evidence for the degree to which the implied model is reasonable.

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

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

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

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

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

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

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

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

For the implied model, we posit a second-order factor analysis 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

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

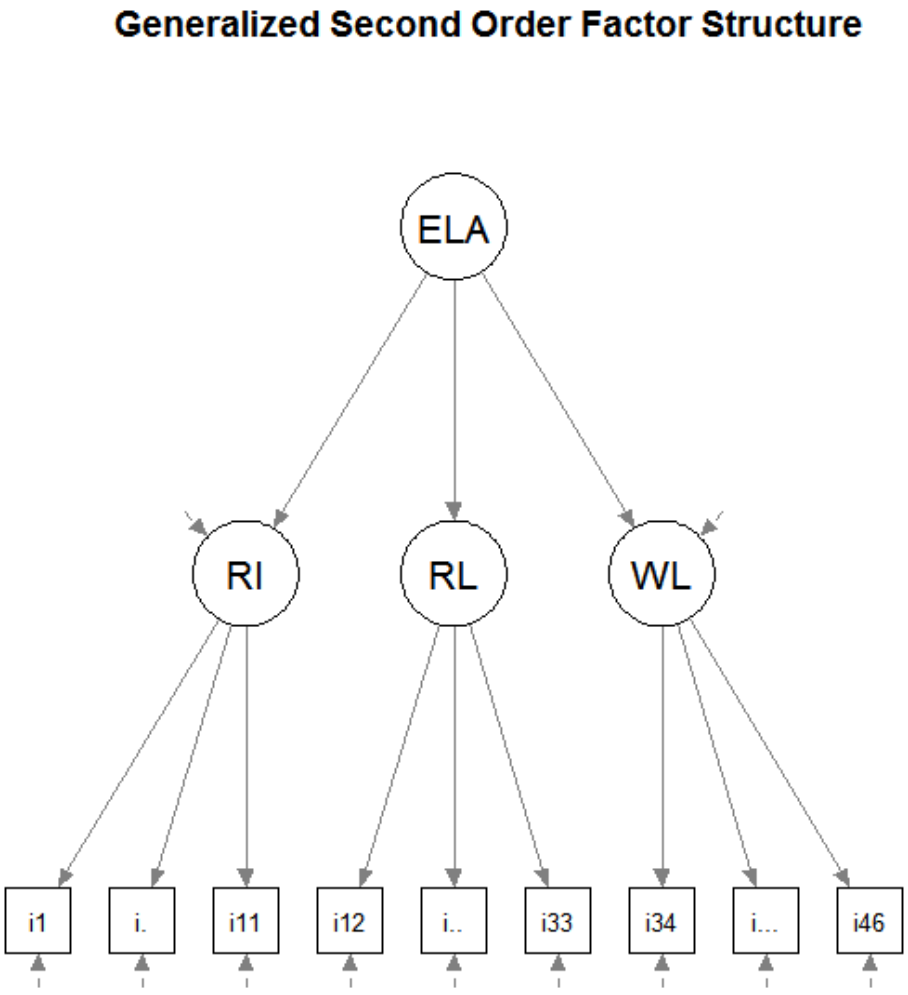
where $\hat{\boldsymbol{\Sigma}}$ is the implied correlation matrix among test items, $\boldsymbol{\Lambda}$ is the *p x k* matrix of first-order factor loadings relating item scores to first-order factors, $\boldsymbol{\Gamma}$ is the *k x 1* matrix of second-order factor loadings relating the first-order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

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

The second-order factor model can also be represented graphically, and a sample of the generalized approaches is provided on the following page. The general structure of the second-order factor analysis for ELA is illustrated in Figure 5. This figure is generally representative of the factor analyses performed for all grades and subjects, with the understanding that the number of items within each reporting category could vary across the grades.

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

Figure 5: Second-Order Factor Model (ELA)



5.2.2 Results

Several goodness-of-fit statistics from each of the analyses are presented in Table 30, which shows the summary results obtained from confirmatory factor analysis. 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 cut-off 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, RMSEA was below 0.05, and CFI and TLI were equal to or greater than 0.95.

Table 30: 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	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 31 to Table 34, 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 31: Correlations Among ELA Factors

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
3	Key Ideas and Textual Support/Vocabulary (Cat1)	13	1		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
	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 32: Correlations Among Mathematics Factors

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	9	1			
	Computation (Cat2)	13	0.989	1		
	Geometry and Measurement (Cat3)	10	0.969	0.959	1	
	Number Sense (Cat4)	11	0.908	0.898	0.880	1
4	Algebraic Thinking and Data Analysis (Cat1)	9	1			
	Computation (Cat2)	12	0.963	1		
	Geometry and Measurement (Cat3)	10	0.929	0.894	1	
	Number Sense (Cat4)	12	0.934	0.900	0.868	1
5	Algebraic Thinking (Cat1)	11	1			
	Computation (Cat2)	11	0.888	1		

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
	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 33: Correlations Among Science Factors

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	
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	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	

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4	Cat5
	Evaluating Claims with Evidence (Cat5)	11	0.971	0.945	0.977	0.991	1

Table 34: Correlations Among Social Studies Factors

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

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):

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

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

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

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

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

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

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

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

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

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

When there are *n* items, *n*(*n*-1)/2, Q_3 statistics will be produced. The Q_3 values are expected to be small. Table 35 through Table 38 present summaries of the distributions of Q_3 statistics—minimum, 5th percentile, median, 95th percentile, and maximum values from each grade and subject. The results show that less than 3% of the items were greater than a critical value of 0.2 for $|Q_3|$ (Chen & Thissen, 1997).

Table 35: ELA Q_3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	-0.210	-0.093	-0.020	0.055	0.205
4	-0.335	-0.089	-0.021	0.045	0.221
5	-0.290	-0.093	-0.019	0.053	0.203
6	-0.192	-0.086	-0.017	0.051	0.197
7	-0.235	-0.088	-0.018	0.050	0.270
8	-0.236	-0.095	-0.021	0.050	0.190

Table 36: Mathematics Q_3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	-0.356	-0.092	-0.018	0.061	0.961
4	-0.250	-0.098	-0.023	0.057	0.881
5	-0.324	-0.097	-0.023	0.058	0.914
6	-0.335	-0.096	-0.020	0.065	0.769
7	-0.265	-0.095	-0.02	0.059	0.825
8	-0.516	-0.097	-0.02	0.058	0.785

Table 37: Science Q_3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
4	-0.206	-0.078	-0.009	0.044	0.266
6	-0.378	-0.149	-0.012	0.069	0.590
Biology	-0.240	-0.136	-0.014	0.121	0.571

Table 38: Social Studies Q_3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
5	-0.107	-0.55	-0.024	0.008	0.206
U.S. Government	-0.158	-0.016	0.104	0.278	0.419

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 39 through Table 50 show the observed and disattenuated score correlations between 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 39: 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)	1.00						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.67	1.00					
	Writing (Cat3)	0.60	0.55	1.00				
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.63	0.58	0.58	1.00			
	Computation (Cat2)	0.63	0.58	0.57	0.78	1.00		
	Geometry and Measurement (Cat3)	0.59	0.54	0.54	0.74	0.73	1.00	
	Number Sense (Cat4)	0.59	0.55	0.54	0.73	0.71	0.71	1.00

Table 40: 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)	1.00						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.99	1.00					
	Writing (Cat3)	0.91	0.89	1.00				
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.84	0.83	0.85	1.00			
	Computation (Cat2)	0.85	0.85	0.86	1.02	1.00		
	Geometry and Measurement (Cat3)	0.83	0.83	0.84	1.01	1.02	1.00	
	Number Sense (Cat4)	0.80	0.80	0.80	0.95	0.95	0.98	1.00

Table 41: 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)	1.00										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.64	1.00									
	Writing (Cat3)	0.63	0.57	1.00								
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.62	0.57	0.61	1.00							
	Computation (Cat2)	0.58	0.52	0.57	0.73	1.00						
	Geometry and Measurement (Cat3)	0.57	0.52	0.57	0.71	0.70	1.00					
	Number Sense (Cat4)	0.59	0.54	0.59	0.73	0.74	0.73	1.00				
Science	Questioning and Modeling (Cat1)	0.56	0.52	0.53	0.57	0.53	0.55	0.56	1.00			
	Investigating (Cat2)	0.63	0.59	0.58	0.64	0.60	0.61	0.62	0.59	1.00		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.65	0.60	0.60	0.64	0.60	0.61	0.63	0.61	0.67	1.00	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.62	0.57	0.57	0.62	0.57	0.59	0.60	0.59	0.65	0.68	1.00

Table 42: 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)	1.00										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.92	1.00									
	Writing (Cat3)	0.90	0.85	1.00								
Mathematics	Algebraic Thinking and Data Analysis (Cat1)	0.86	0.82	0.88	1.00							
	Computation (Cat2)	0.76	0.73	0.80	0.98	1.00						
	Geometry and Measurement (Cat3)	0.78	0.75	0.81	0.98	0.92	1.00					
	Number Sense (Cat4)	0.77	0.75	0.80	0.97	0.94	0.95	1.00				
Science	Questioning and Modeling (Cat1)	0.87	0.86	0.86	0.90	0.81	0.86	0.84	1.00			
	Investigating (Cat2)	0.92	0.91	0.89	0.94	0.85	0.89	0.87	0.98	1.00		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.93	0.91	0.89	0.93	0.83	0.87	0.85	0.99	1.00*	1.00	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.93	0.90	0.89	0.94	0.83	0.88	0.86	1.00	1.00*	1.00*	1.00

Table 43: 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)	1.00									
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.60	1.00								
	Writing (Cat3)	0.64	0.57	1.00							
Mathematics	Algebra and Functions (Cat1)	0.60	0.55	0.64	1.00						
	Computation (Cat2)	0.55	0.50	0.60	0.74	1.00					
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.53	0.49	0.56	0.69	0.66	1.00				
	Number Sense (Cat4)	0.55	0.52	0.58	0.73	0.70	0.66	1.00			
Social Studies	Civics and Government (Cat1)	0.61	0.58	0.59	0.61	0.54	0.54	0.58	1.00		
	Geography and Economics (Cat2)	0.55	0.52	0.53	0.57	0.51	0.51	0.55	0.67	1.00	
	History (Cat3)	0.63	0.58	0.61	0.61	0.55	0.54	0.57	0.71	0.65	1.00

Table 44: 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)	1.00									
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.94	1.00								
	Writing (Cat3)	0.89	0.85	1.00							
Mathematics	Algebra and Functions (Cat1)	0.84	0.82	0.84	1.00						
	Computation (Cat2)	0.78	0.76	0.80	0.99	1.00					
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.80	0.79	0.80	0.98	0.97	1.00				
	Number Sense (Cat4)	0.79	0.78	0.77	0.98	0.96	0.96	1.00			
Social Studies	Civics and Government (Cat1)	0.87	0.88	0.79	0.82	0.75	0.79	0.80	1.00		
	Geography and Economics (Cat2)	0.89	0.89	0.80	0.86	0.79	0.85	0.84	1.00*	1.00	
	History (Cat3)	0.92	0.90	0.85	0.84	0.78	0.82	0.81	1.00*	1.00*	1.00

Table 45: 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)	1.00										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.66	1.00									
	Writing (Cat3)	0.62	0.60	1.00								
Mathematics	Algebra and Functions (Cat1)	0.61	0.60	0.63	1.00							
	Computation (Cat2)	0.56	0.56	0.59	0.76	1.00						
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.57	0.55	0.58	0.71	0.67	1.00					
	Number Sense (Cat4)	0.60	0.59	0.60	0.76	0.71	0.68	1.00				
Science	Questioning and Modeling (Cat1)	0.61	0.60	0.56	0.64	0.57	0.59	0.63	1.00			
	Investigating (Cat2)	0.66	0.64	0.61	0.67	0.61	0.63	0.66	0.70	1.00		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.60	0.59	0.55	0.62	0.56	0.58	0.62	0.67	0.68	1.00	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.62	0.61	0.59	0.66	0.61	0.62	0.65	0.68	0.70	0.67	1.00

Table 46: Grade 6 Disattenuated Score Correlations

Subject		ELA			Mathematics				Science			
	Reporting Category	**Cat1**	**Cat2**	**Cat3**	**Cat1**	**Cat2**	**Cat3**	**Cat4**	**Cat1**	**Cat2**	**Cat3**	**Cat4**
ELA	Key Ideas and Textual Support/Vocabulary (Cat1)	1.00										
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	1.00*	1.00									
	Writing (Cat3)	0.88	0.91	1.00								
Mathematics	Algebra and Functions (Cat1)	0.81	0.86	0.82	1.00							
	Computation (Cat2)	0.77	0.81	0.79	0.95	1.00						
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	0.80	0.83	0.80	0.93	0.89	1.00					
	Number Sense (Cat4)	0.83	0.86	0.81	0.97	0.92	0.92	1.00				
Science	Questioning and Modeling (Cat1)	0.90	0.94	0.81	0.87	0.80	0.85	0.88	1.00			
	Investigating (Cat2)	0.94	0.98	0.86	0.89	0.83	0.88	0.91	1.00*	1.00		
	Analyzing, Interpreting, and Computational Thinking (Cat3)	0.89	0.93	0.80	0.85	0.78	0.84	0.87	1.00*	1.00*	1.00	
	Explaining Solutions, Reasoning, and Communicating (Cat4)	0.91	0.95	0.85	0.89	0.83	0.88	0.90	1.00	1.00*	1.00*	1.00

Table 47: 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)	1.00						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.62	1.00					
	Writing (Cat3)	0.65	0.59	1.00				
Mathematics	Algebra and Functions (Cat1)	0.59	0.55	0.60	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	0.61	0.56	0.61	0.68	1.00		
	Geometry and Measurement (Cat3)	0.51	0.47	0.51	0.61	0.60	1.00	
	Number Sense and Computation (Cat4)	0.63	0.58	0.64	0.75	0.74	0.65	1.00

Table 48: 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)	1.00						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.93	1.00					
	Writing (Cat3)	0.92	0.87	1.00				
Mathematics	Algebra and Functions (Cat1)	0.83	0.81	0.82	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	0.89	0.86	0.87	0.97	1.00		
	Geometry and Measurement (Cat3)	0.72	0.70	0.71	0.85	0.86	1.00	
	Number Sense and Computation (Cat4)	0.86	0.82	0.85	0.99	1.02	0.87	1.00

Table 49: 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)	1.00						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.60	1.00					
	Writing (Cat3)	0.66	0.57	1.00				
Mathematics	Algebra and Functions (Cat1)	0.61	0.55	0.63	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	0.59	0.53	0.62	0.74	1.00		
	Geometry and Measurement (Cat3)	0.55	0.50	0.58	0.72	0.70	1.00	
	Number Sense and Computation (Cat4)	0.54	0.49	0.56	0.69	0.65	0.65	1.00

Table 50: 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)	1.00						
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	0.91	1.00					
	Writing (Cat3)	0.91	0.86	1.00				
Mathematics	Algebra and Functions (Cat1)	0.80	0.79	0.82	1.00			
	Data Analysis, Statistics, and Probability (Cat2)	0.79	0.77	0.82	0.94	1.00		
	Geometry and Measurement (Cat3)	0.77	0.76	0.80	0.94	0.94	1.00	
	Number Sense and Computation (Cat4)	0.78	0.77	0.80	0.95	0.91	0.95	1.00

6. FAIRNESS IN CONTENT

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

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

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

6.1 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is not sufficient to determine the fairness of a test. Rather, it must be accompanied by statistical processes. While a variety of item statistics were reviewed during form building to evaluate the quality of items, one notable statistic that was used was differential item functioning (DIF). Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF, according to the DIF classification convention illustrated in Volume 1 of this technical report. Furthermore, items were categorized positively (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American/Black, Hispanic, or Female), or negatively (i.e., –A, –B, or–C), signifying that the item favored the reference group (e.g., White or Male). Items were flagged if their DIF statistics indicated the “C” category for any group. A DIF classification of “C” indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness. Items were reviewed by the Bias and Sensitivity Committee regardless of whether the DIF statistic favored the focal group or the reference group. The details surrounding this review of items for bias is further described in Volume 2, Test Development.

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

- Male/Female;

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

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 4.2, of the *2018–2019 ILEARN Annual Technical Report*. The DIF statistics for each operational test item are presented in the appendix A of Volume 1 of the *2018–2019 ILEARN Annual Technical Report*.

7. SUMMARY

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

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

8. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87*(3), 513–524.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Chen, F., Kenneth, A., Bollen, P., Paxton, P., Curran, P. J., & Kirby, J. B. 2001. Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies. *Sociological Methods & Research, 29*, 468–508.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), 105–146. New York: Macmillan.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education, 9*, 277–286.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6).

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*(3), 381–389.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237–255.

Lee, W., Hanson, B., & Brennan, R. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 412–432.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, L. K., & Muthén, B. O. (2012). Mplus user's guide, 7th Edition. Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8*, 111–120.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42*, 549–565.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14).

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

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

van Driel, O. P. 1978. ‘‘On Various Causes of Improper Solutions in Maximum Likelihood Factor Analysis.’’ *Psychometrika, 43*, 225–243.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Yoon, B., & Young, M. J. (2000). *Estimating the reliability for test scores with mixed item formats: Internal consistency and generalizability*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Indiana Learning Evaluation Readiness Network (ILEARN)

2018–2019

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, Eugenia Kim, Kevin Clayton, and Kyra Bilenki. Major contributors from the Indiana Department of Education include the Assessment Director, Assistant Assessment Director, and Program Leads.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

1. INDIANA SCORE REPORTS

In Spring 2019, 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 2018-2019 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 2018-2019 school year. Test scores from each 2018-2019 assessment were provided to corporations and schools through the ORS on August 15, 2019, after the standard setting that occurred July 15–17, 2019. The ORS provides information on student performance and aggregated summaries at several levels—state, corporation, school, and roster. During future administrations, real-time reporting will allow the ORS to report scores within 12 business days after assessments have been scored.

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

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

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

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

1.2 Overall Scores and Reporting Categories

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

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

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

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

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

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

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

Table 3: Reporting Categories for Science

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

Table 4: Reporting Categories for Social Studies

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

1.3 ONLINE REPORTING SYSTEM

ORS generates a set of online score reports that describes student performance for students, parents, educators, and other stakeholders. The online score reports are produced after the tests are submitted by the students, hand-scored and machine-scored, and processed into the ORS. In 2019, scores were not immediately available due to the need for standard setting. However, in future years, results will be available as soon as hand-scored items are processed. 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* is included in Appendix A.

Table 5: Indiana Score Reports Summary

Report	Description	Level of Availability				
		State	Corporation	School	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	✓	✓	✓	✓	
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	✓	✓	✓	✓	
Student-Level Subject Report	List of all students who belong to a school, teacher, or roster with their			✓	✓	

Report	Description	Level of Availability				
		State	Corporation	School	Roster	Student/ Parent
	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

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

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

Table 6: Indiana List of Sub-Groups

Sub-Group	Sub-Group Category
Ethnicity	White
	Black/African American
	Hispanic
	Asian
	American Indian/Alaska Native
	Native Hawaiian/Other Pacific Islander
	Multiracial/Two or More Races
Gender	Male
	Female

Sub-Group	Sub-Group Category
English Learner	English Learner
	Not English Learner
Special Education	Special Education
	Not Special Education
Section 504 Plan	Section 504 Plan
	Not Section 504 Plan
Grade	Grade 3
	Grade 4
	Grade 5
	Grade 6
	Grade 7
	Grade 8
	Grade 9
	Grade 10
	Grade 11
	Grade 12

1.6 REPORTS

1.6.1 Summary Performance Report

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

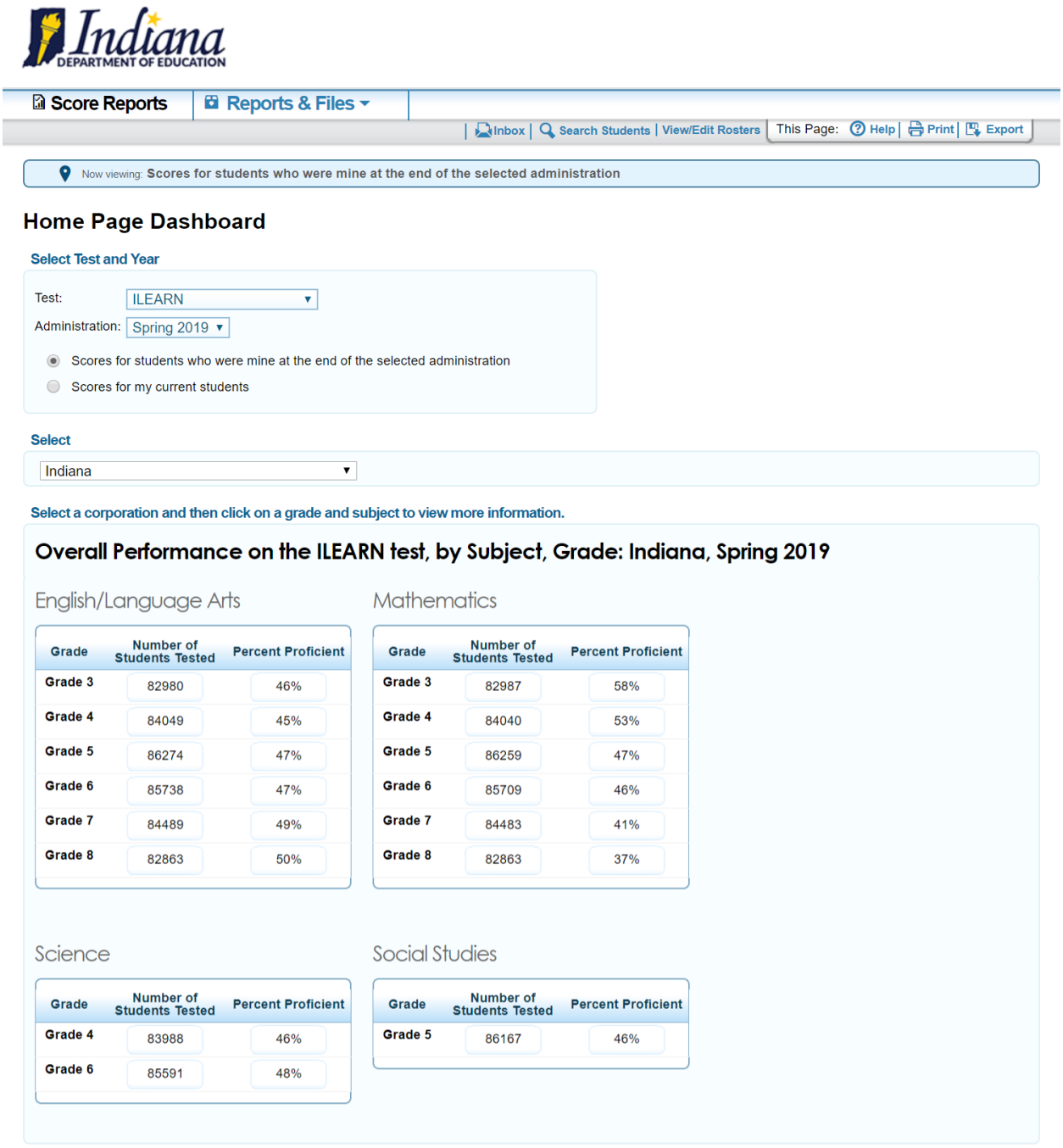
- Displays summary data separated by grade and subject
- Bases the level of aggregation on a user’s role
- Reports the number of students tested and percentage proficient

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

- Number of students tested
- 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



Indiana DEPARTMENT OF EDUCATION

Score Reports | Reports & Files

Inbox | Search Students | View/Edit Rosters | This Page: Help | Print | Export

Now viewing: **Scores for students who were mine at the end of the selected administration**

Home Page Dashboard

Select Test and Year

Test: ILEARN

Administration: Spring 2019

- (•) Scores for students who were mine at the end of the selected administration
- ( ) Scores for my current students

Select

Indiana

Select a corporation and then click on a grade and subject to view more information.

Overall Performance on the ILEARN test, by Subject, Grade: Indiana, Spring 2019

English/Language Arts

Grade	Number of Students Tested	Percent Proficient
Grade 3	82980	46%
Grade 4	84049	45%
Grade 5	86274	47%
Grade 6	85738	47%
Grade 7	84489	49%
Grade 8	82863	50%

Mathematics

Grade	Number of Students Tested	Percent Proficient
Grade 3	82987	58%
Grade 4	84040	53%
Grade 5	86259	47%
Grade 6	85709	46%
Grade 7	84483	41%
Grade 8	82863	37%

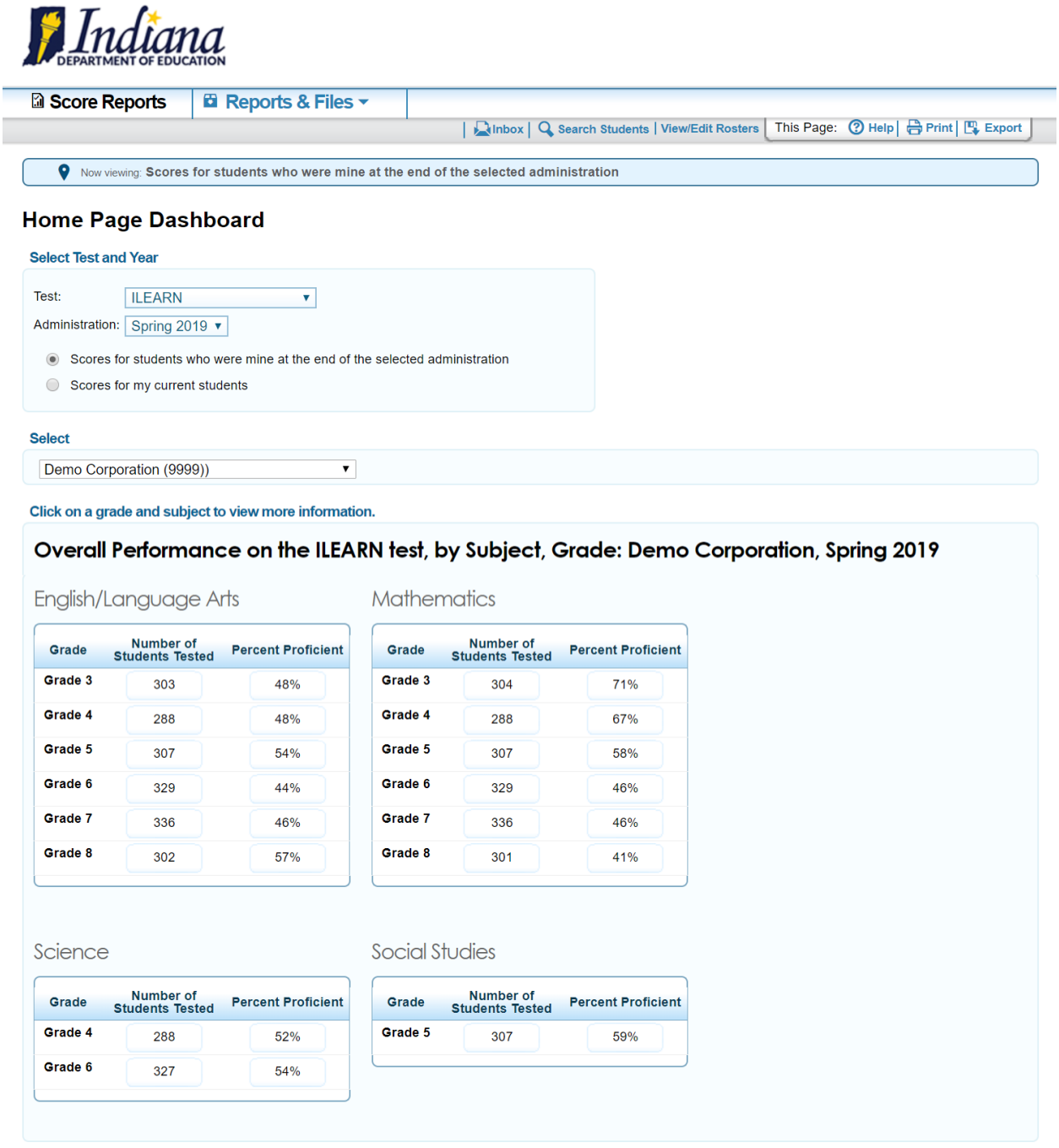
Science

Grade	Number of Students Tested	Percent Proficient
Grade 4	83988	46%
Grade 6	85591	48%

Social Studies

Grade	Number of Students Tested	Percent Proficient
Grade 5	86167	46%

Figure 2: Corporation-Level Summary Performance Report



Indiana Department of Education

Score Reports | Reports & Files

Inbox | Search Students | View/Edit Rosters | This Page: Help | Print | Export

Now viewing: Scores for students who were mine at the end of the selected administration

Home Page Dashboard

Select Test and Year

Test: ILEARN

Administration: Spring 2019

(•) Scores for students who were mine at the end of the selected administration

( ) Scores for my current students

Select

Demo Corporation (9999))

Click on a grade and subject to view more information.

Overall Performance on the ILEARN test, by Subject, Grade: Demo Corporation, Spring 2019

English/Language Arts

Grade	Number of Students Tested	Percent Proficient
Grade 3	303	48%
Grade 4	288	48%
Grade 5	307	54%
Grade 6	329	44%
Grade 7	336	46%
Grade 8	302	57%

Mathematics

Grade	Number of Students Tested	Percent Proficient
Grade 3	304	71%
Grade 4	288	67%
Grade 5	307	58%
Grade 6	329	46%
Grade 7	336	46%
Grade 8	301	41%

Science

Grade	Number of Students Tested	Percent Proficient
Grade 4	288	52%
Grade 6	327	54%

Social Studies

Grade	Number of Students Tested	Percent Proficient
Grade 5	307	59%

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

1.6.2 Aggregate-Level Subject Report

Detailed summaries of student performance within a grade 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
- 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.

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

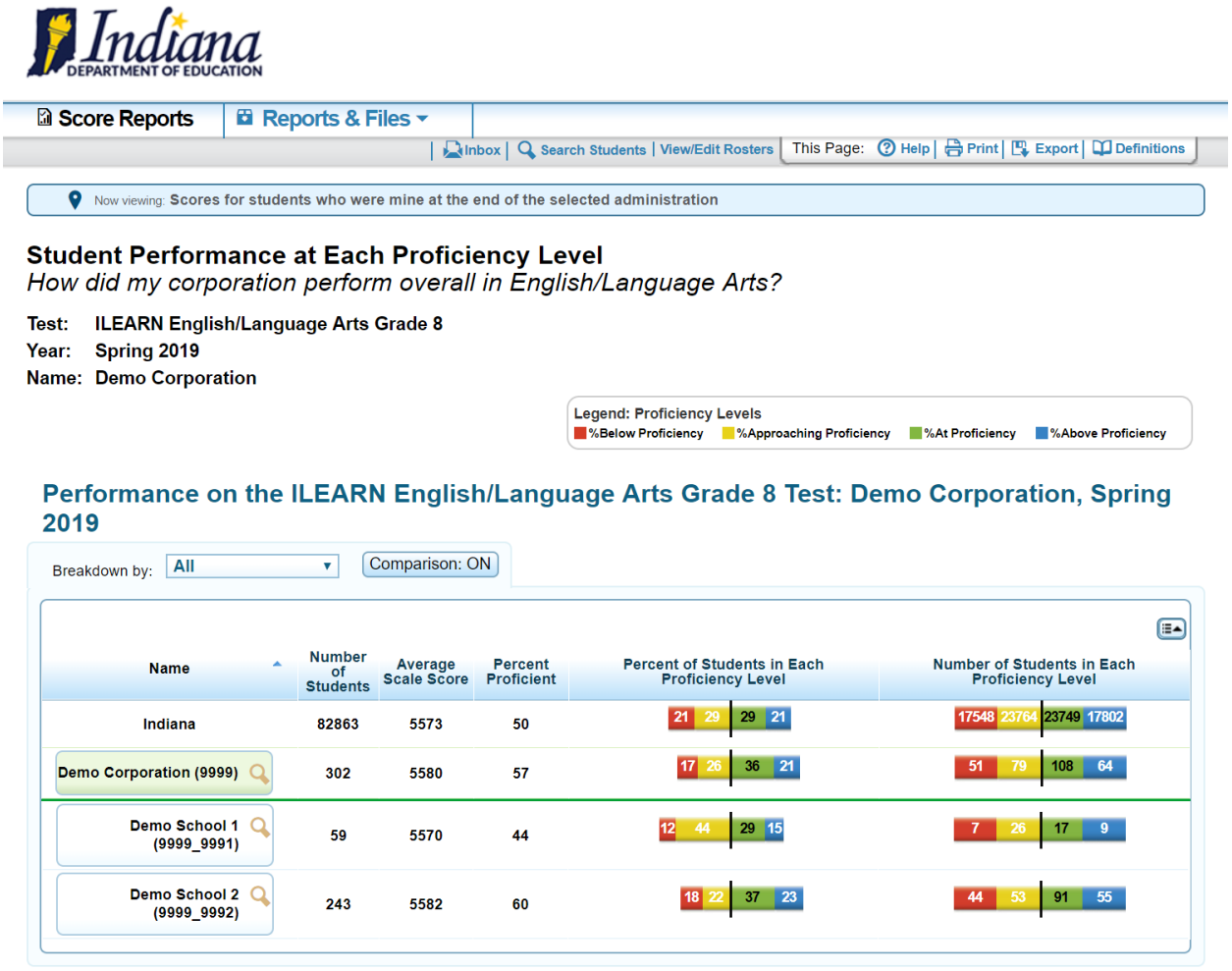
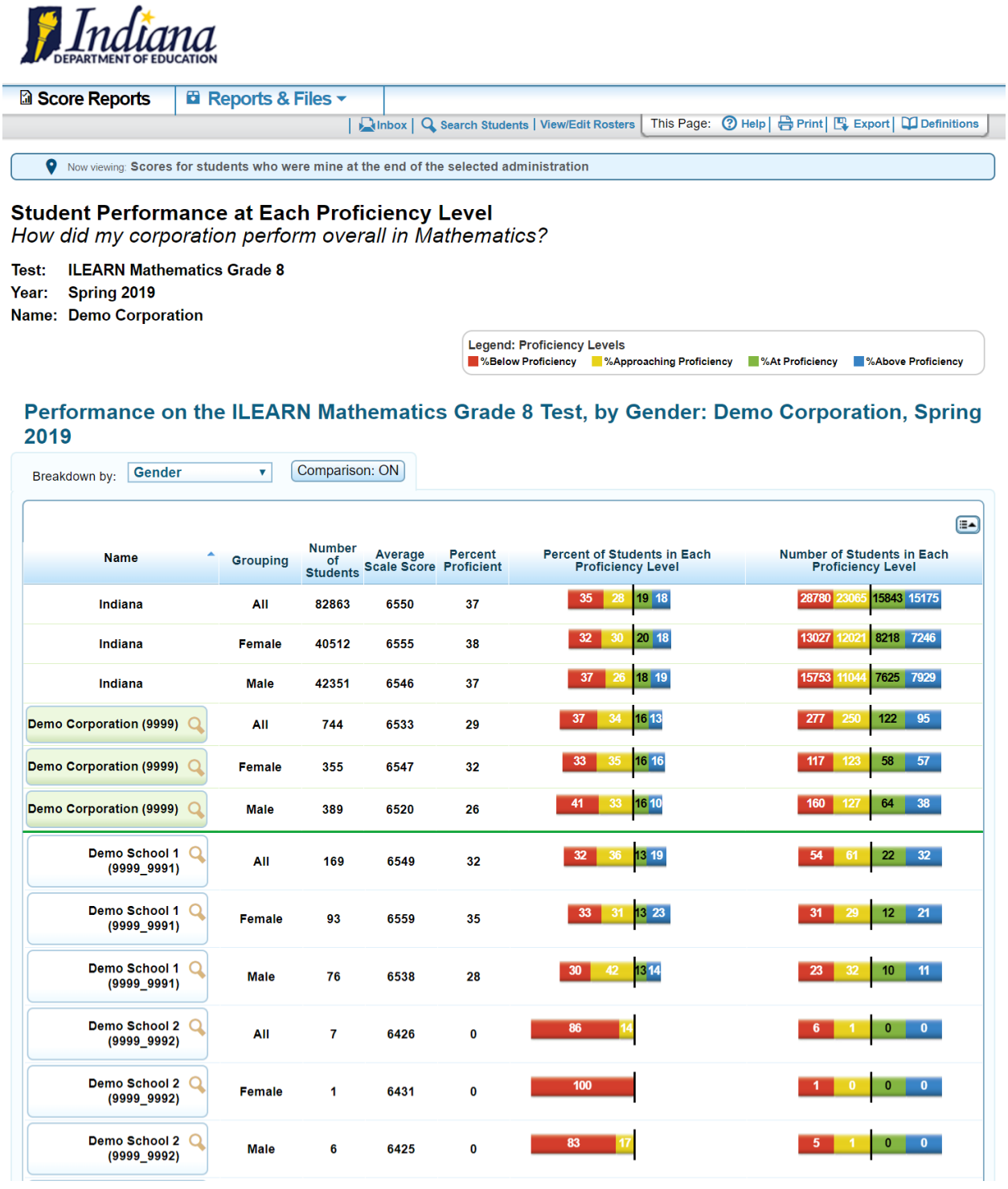


Figure 4: Corporation Aggregate-Level Subject Report, Grade 8 Mathematics



Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Percent of Students in Each Proficiency Level	Number of Students in Each Proficiency Level
Indiana	All	82863	6550	37	35 28 19 18	28780 23065 15843 15175
Indiana	Female	40512	6555	38	32 30 20 18	13027 12021 8218 7246
Indiana	Male	42351	6546	37	37 26 18 19	15753 11044 7625 7929
Demo Corporation (9999)	All	744	6533	29	37 34 16 13	277 250 122 95
Demo Corporation (9999)	Female	355	6547	32	33 35 16 16	117 123 58 57
Demo Corporation (9999)	Male	389	6520	26	41 33 16 10	160 127 64 38
Demo School 1 (9999_9991)	All	169	6549	32	32 36 13 19	54 61 22 32
Demo School 1 (9999_9991)	Female	93	6559	35	33 31 13 23	31 29 12 21
Demo School 1 (9999_9991)	Male	76	6538	28	30 42 13 14	23 32 10 11
Demo School 2 (9999_9992)	All	7	6426	0	86 14	6 1 0 0
Demo School 2 (9999_9992)	Female	1	6431	0	100	1 0 0 0
Demo School 2 (9999_9992)	Male	6	6425	0	83 17	5 1 0 0

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

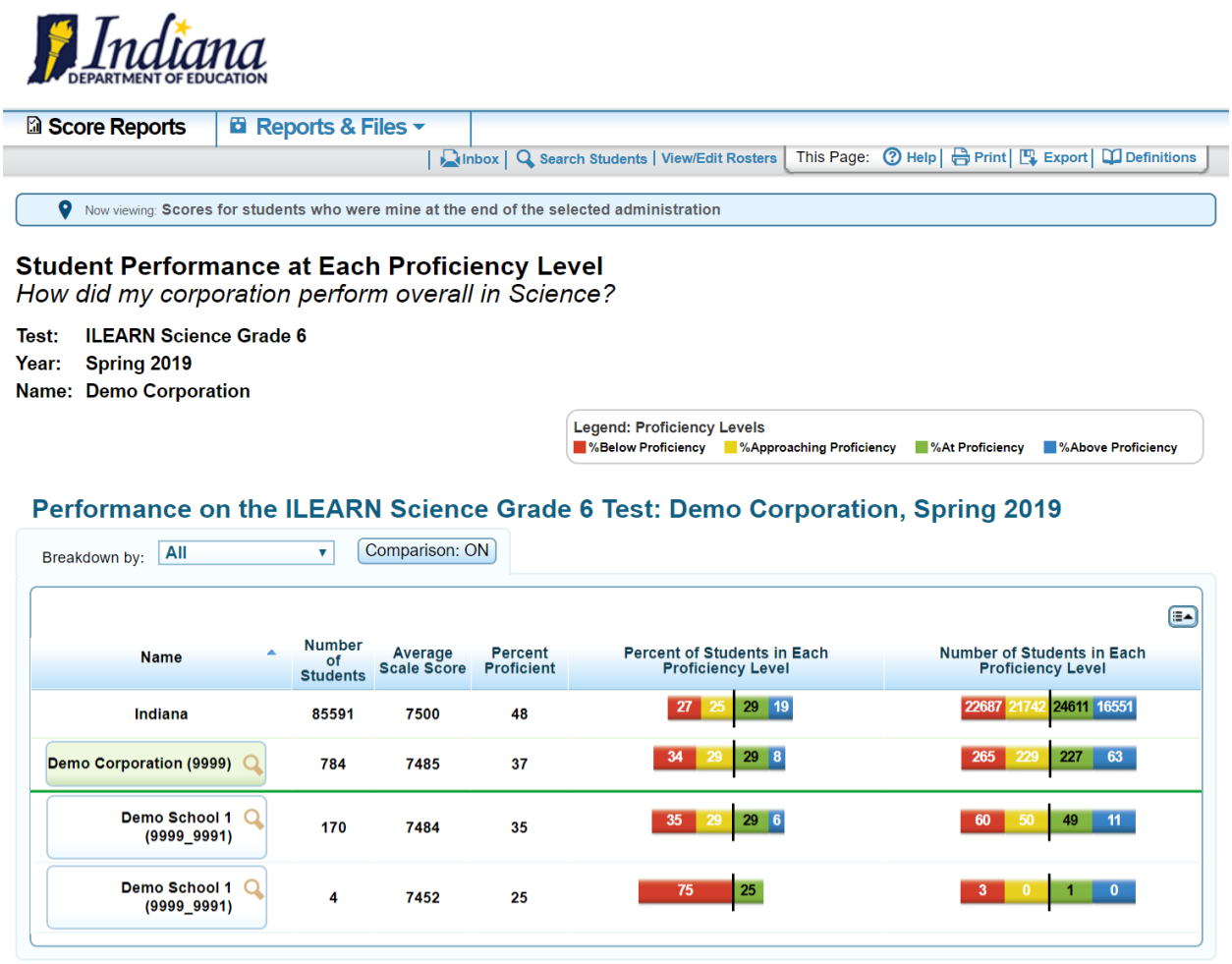
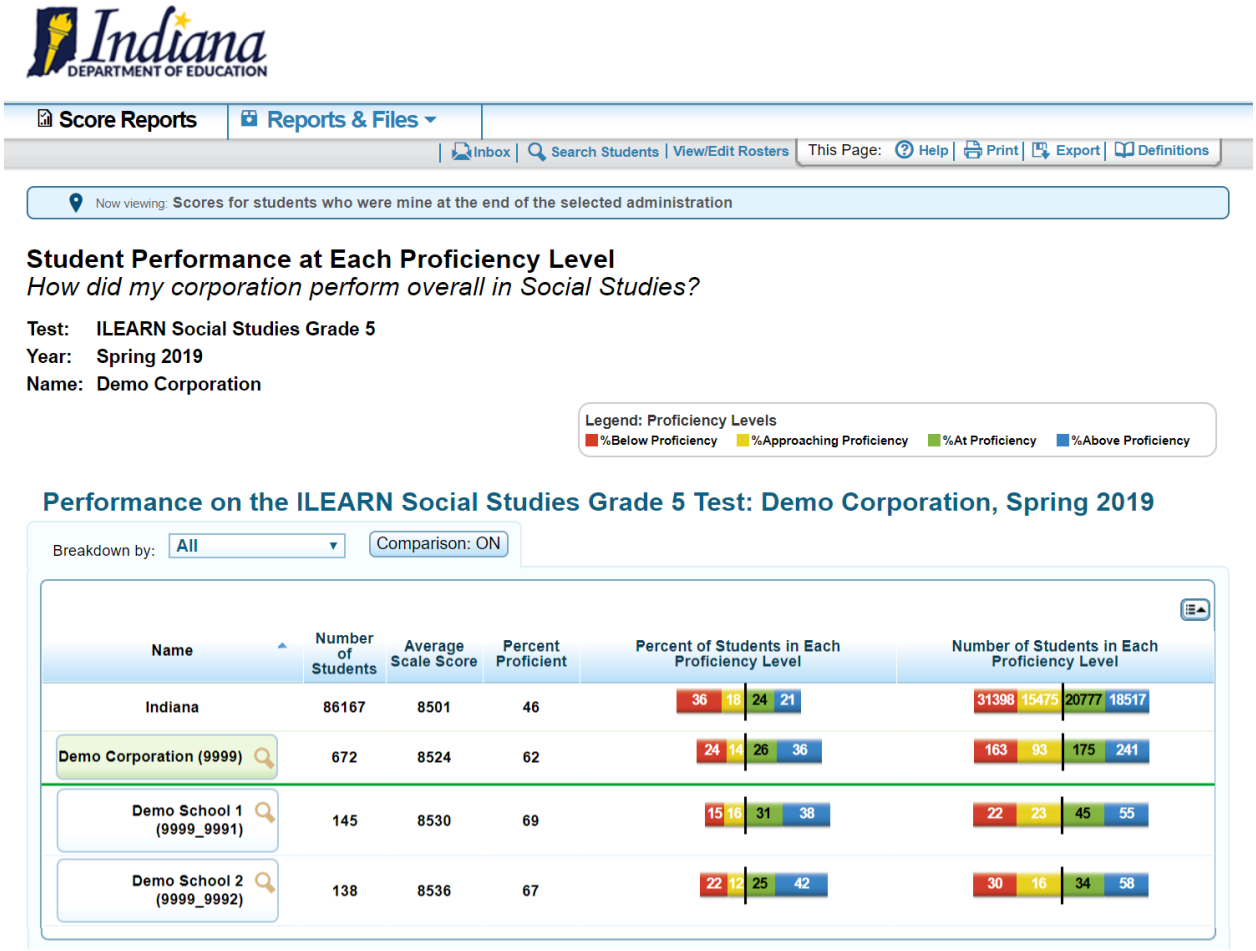


Figure 6: Corporation Aggregate-Level Subject Report, Grade 5 Social Studies



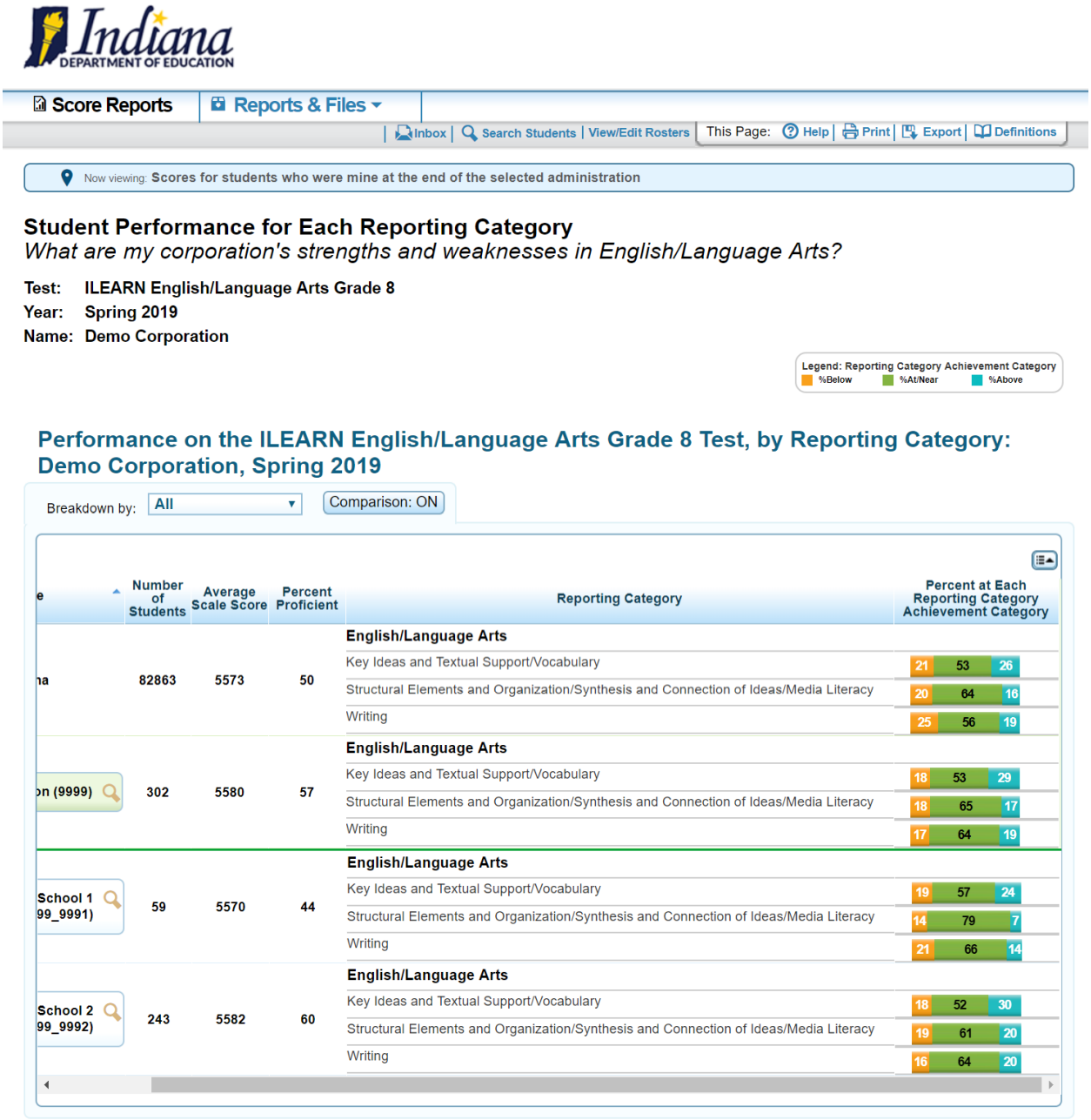
1.6.3 Aggregate-Level Reporting Category Report

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

- Number of students
- Average scale score and standard error of the average scale score
- Percentage proficient
- 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



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 at the end of the selected administration**

Student Performance for Each Reporting Category

What are my corporation's strengths and weaknesses in English/Language Arts?

Test: ILEARN English/Language Arts Grade 8
Year: Spring 2019
Name: Demo Corporation

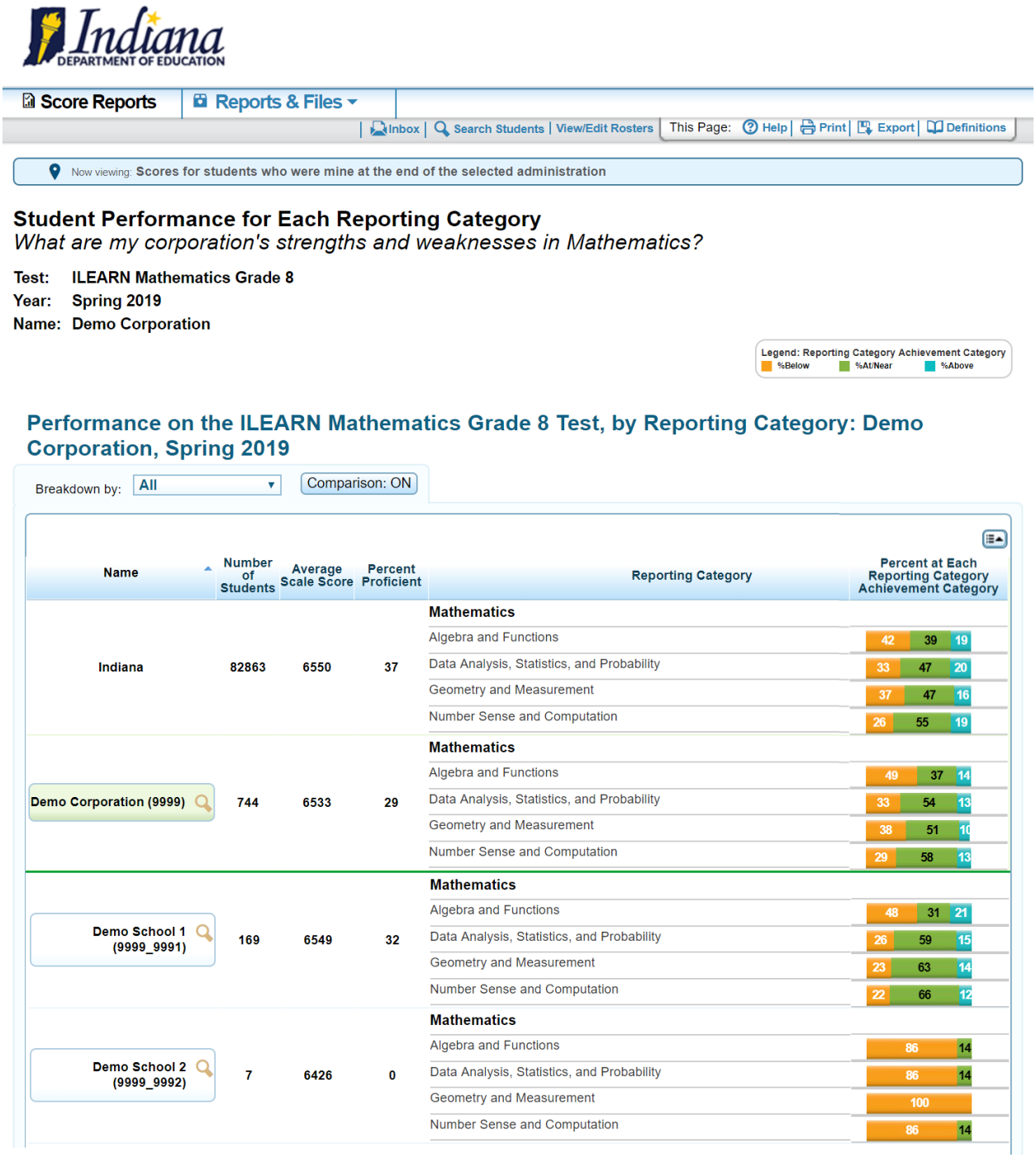
Legend: Reporting Category Achievement Category — %Below | %At/Near | %Above

Performance on the ILEARN English/Language Arts Grade 8 Test, by Reporting Category: Demo Corporation, Spring 2019

Breakdown by: All | Comparison: ON

e	Number of Students	Average Scale Score	Percent Proficient	Reporting Category	Percent at Each Reporting Category Achievement Category
ıa	82863	5573	50	**English/Language Arts**	
				Key Ideas and Textual Support/Vocabulary	21 53 26
				Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy	20 64 16
				Writing	25 56 19
ɔn (9999)	302	5580	57	**English/Language Arts**	
				Key Ideas and Textual Support/Vocabulary	18 53 29
				Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy	18 65 17
				Writing	17 64 19
School 1 99_9991)	59	5570	44	**English/Language Arts**	
				Key Ideas and Textual Support/Vocabulary	19 57 24
				Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy	14 79 7
				Writing	21 66 14
School 2 99_9992)	243	5582	60	**English/Language Arts**	
				Key Ideas and Textual Support/Vocabulary	18 52 30
				Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy	19 61 20
				Writing	16 64 20

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



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 at the end of the selected administration

Student Performance for Each Reporting Category

What are my corporation's strengths and weaknesses in Mathematics?

Test: ILEARN Mathematics Grade 8
Year: Spring 2019
Name: Demo Corporation

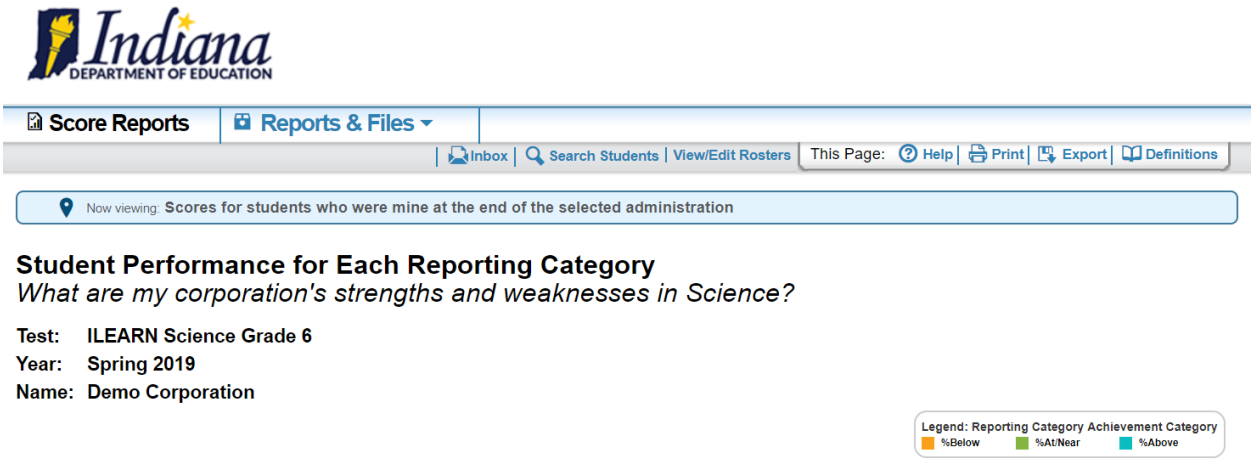
Legend: Reporting Category Achievement Category — %Below | %At/Near | %Above

Performance on the ILEARN Mathematics Grade 8 Test, by Reporting Category: Demo Corporation, Spring 2019

Breakdown by: All | Comparison: ON

Name	Number of Students	Average Scale Score	Percent Proficient	Reporting Category	Percent at Each Reporting Category Achievement Category
Indiana	82863	6550	37	**Mathematics**	
				Algebra and Functions	42 / 39 / 19
				Data Analysis, Statistics, and Probability	33 / 47 / 20
				Geometry and Measurement	37 / 47 / 16
				Number Sense and Computation	26 / 55 / 19
Demo Corporation (9999)	744	6533	29	**Mathematics**	
				Algebra and Functions	49 / 37 / 14
				Data Analysis, Statistics, and Probability	33 / 54 / 13
				Geometry and Measurement	38 / 51 / 10
				Number Sense and Computation	29 / 58 / 13
Demo School 1 (9999_9991)	169	6549	32	**Mathematics**	
				Algebra and Functions	48 / 31 / 21
				Data Analysis, Statistics, and Probability	26 / 59 / 15
				Geometry and Measurement	23 / 63 / 14
				Number Sense and Computation	22 / 66 / 12
Demo School 2 (9999_9992)	7	6426	0	**Mathematics**	
				Algebra and Functions	86 / 14
				Data Analysis, Statistics, and Probability	86 / 14
				Geometry and Measurement	100
				Number Sense and Computation	86 / 14

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



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 at the end of the selected administration**

Student Performance for Each Reporting Category

What are my corporation's strengths and weaknesses in Science?

Test: ILEARN Science Grade 6
Year: Spring 2019
Name: Demo Corporation

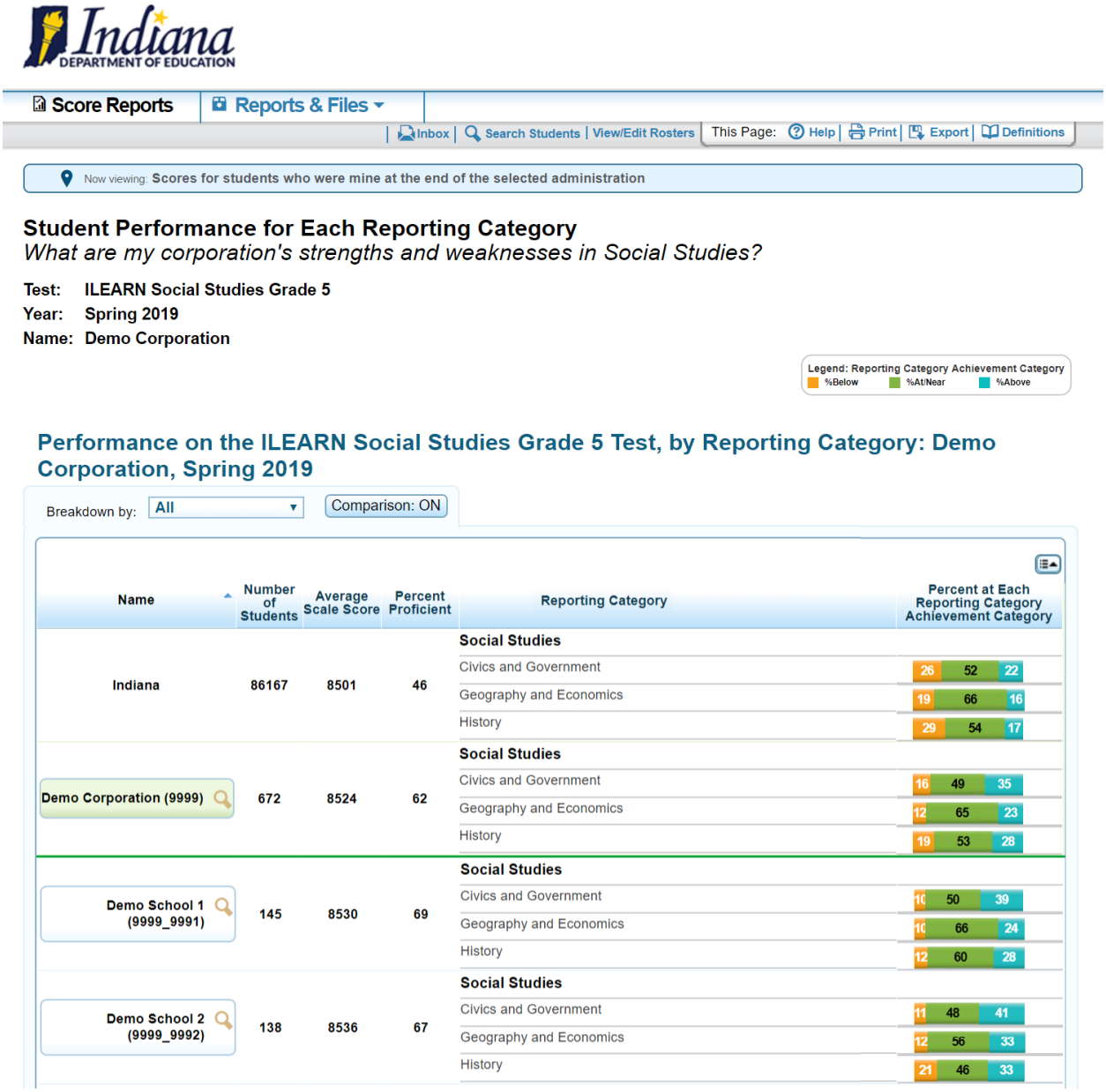
Legend: Reporting Category Achievement Category — %Below | %At/Near | %Above

Performance on the ILEARN Science Grade 6 Test, by Reporting Category: Demo Corporation, Spring 2019

Breakdown by: All | Comparison: ON

Name	Number of Students	Average Scale Score	Percent Proficient	Reporting Category	Percent at Each Reporting Category Achievement Category
Indiana	85591	7500	48	**Science**	
				Questioning and Modeling	24 / 57 / 18
				Investigating	24 / 56 / 20
				Analyzing, Interpreting, and Computational Thinking	26 / 58 / 16
				Explaining Solutions, Reasoning, and Communicating	27 / 57 / 16
Demo Corporation (9999)	784	7485	37	**Science**	
				Questioning and Modeling	31 / 62 / 7
				Investigating	32 / 56 / 13
				Analyzing, Interpreting, and Computational Thinking	32 / 60 / 8
				Explaining Solutions, Reasoning, and Communicating	34 / 57 / 9
Demo School 1 (9999_9991)	170	7484	35	**Science**	
				Questioning and Modeling	29 / 65 / 6
				Investigating	34 / 51 / 15
				Analyzing, Interpreting, and Computational Thinking	34 / 61 / 6
				Explaining Solutions, Reasoning, and Communicating	35 / 56 / 9
Demo School 2 (9999_9992)	4	7452	25	**Science**	
				Questioning and Modeling	75 / 25
				Investigating	75 / 25
				Analyzing, Interpreting, and Computational Thinking	25 / 75
				Explaining Solutions, Reasoning, and Communicating	75 / 25

Figure 10: Corporation Aggregate-Level Reporting Category 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 at the end of the selected administration

Student Performance for Each Reporting Category

What are my corporation's strengths and weaknesses in Social Studies?

Test: ILEARN Social Studies Grade 5
Year: Spring 2019
Name: Demo Corporation

Legend: Reporting Category Achievement Category — %Below | %At/Near | %Above

Performance on the ILEARN Social Studies Grade 5 Test, by Reporting Category: Demo Corporation, Spring 2019

Breakdown by: All | Comparison: ON

Name	Number of Students	Average Scale Score	Percent Proficient	Reporting Category	Percent at Each Reporting Category Achievement Category
Indiana	86167	8501	46	**Social Studies**	
				Civics and Government	26 / 52 / 22
				Geography and Economics	19 / 66 / 16
				History	29 / 54 / 17
Demo Corporation (9999)	672	8524	62	**Social Studies**	
				Civics and Government	16 / 49 / 35
				Geography and Economics	12 / 65 / 23
				History	19 / 53 / 28
Demo School 1 (9999_9991)	145	8530	69	**Social Studies**	
				Civics and Government	10 / 50 / 39
				Geography and Economics	10 / 66 / 24
				History	12 / 60 / 28
Demo School 2 (9999_9992)	138	8536	67	**Social Studies**	
				Civics and Government	11 / 48 / 41
				Geography and Economics	12 / 56 / 33
				History	21 / 46 / 33

1.6.4 Aggregate-Level Standards Report

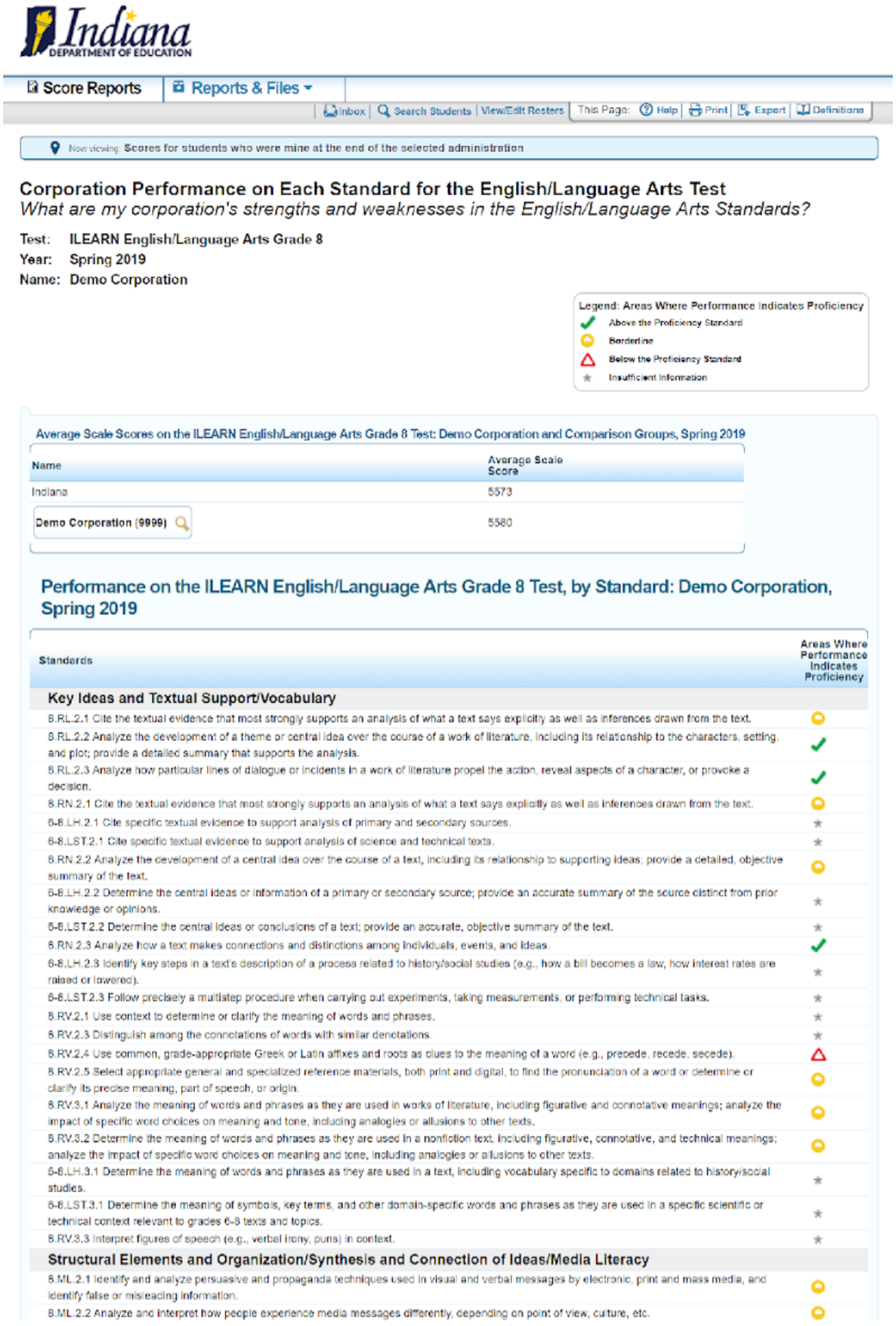
The Aggregate-Level Standards Report lists data on the performance of student groups on each standard of a subject for the current testing window and reports the following measures for the selected level of aggregation:

- Areas Where Performance Indicates Proficiency

For adaptive assessments, a standard performance indicator produces information on how a group of students in a 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



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 at the end of the selected administration

Corporation Performance on Each Standard for the English/Language Arts Test

What are my corporation's strengths and weaknesses in the English/Language Arts Standards?

Test: ILEARN English/Language Arts Grade 8
Year: Spring 2019
Name: Demo Corporation

Legend: Areas Where Performance Indicates Proficiency
- Above the Proficiency Standard
- Borderline
- Below the Proficiency Standard
- Insufficient Information

Average Scale Scores on the ILEARN English/Language Arts Grade 8 Test: Demo Corporation and Comparison Groups, Spring 2019

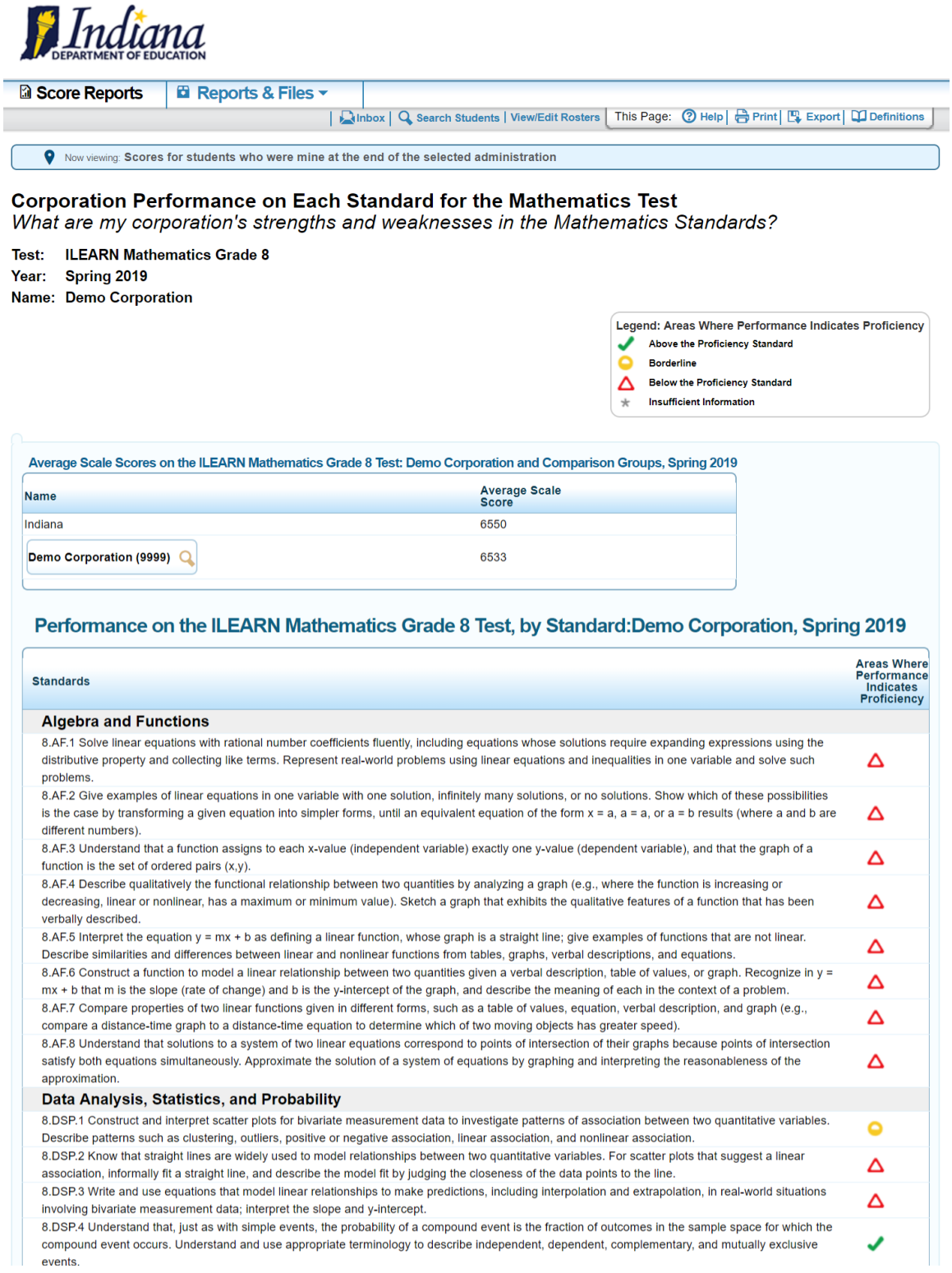
Name	Average Scale Score
Indiana	5573
Demo Corporation (9999)	5580

Performance on the ILEARN English/Language Arts Grade 8 Test, by Standard: Demo Corporation, Spring 2019

Standards	Areas Where Performance Indicates Proficiency
Key Ideas and Textual Support/Vocabulary	
8.RL.2.1 Cite the textual evidence that most strongly supports an analysis of what a text says explicitly as well as inferences drawn from the text.	○
8.RL.2.2 Analyze the development of a theme or central idea over the course of a work of literature, including its relationship to the characters, setting, and plot; provide a detailed summary that supports the analysis.	✓
8.RL.2.3 Analyze how particular lines of dialogue or incidents in a work of literature propel the action, reveal aspects of a character, or provoke a decision.	✓
8.RN.2.1 Cite the textual evidence that most strongly supports an analysis of what a text says explicitly as well as inferences drawn from the text.	○
6-8.LH.2.1 Cite specific textual evidence to support analysis of primary and secondary sources.	★
6-8.LST.2.1 Cite specific textual evidence to support analysis of science and technical texts.	★
8.RN.2.2 Analyze the development of a central idea over the course of a text, including its relationship to supporting ideas; provide a detailed, objective summary of the text.	○
6-8.LH.2.2 Determine the central ideas or information of a primary or secondary source; provide an accurate summary of the source distinct from prior knowledge or opinions.	★
6-8.LST.2.2 Determine the central ideas or conclusions of a text; provide an accurate, objective summary of the text.	★
8.RN.2.3 Analyze how a text makes connections and distinctions among individuals, events, and ideas.	✓
6-8.LH.2.3 Identify key steps in a text's description of a process related to history/social studies (e.g., how a bill becomes a law, how interest rates are raised or lowered).	★
6-8.LST.2.3 Follow precisely a multistep procedure when carrying out experiments, taking measurements, or performing technical tasks.	★
8.RV.2.1 Use context to determine or clarify the meaning of words and phrases.	★
8.RV.2.3 Distinguish among the connotations of words with similar denotations.	★
8.RV.2.4 Use common, grade-appropriate Greek or Latin affixes and roots as clues to the meaning of a word (e.g., precede, recede, secede).	△
8.RV.2.5 Select appropriate general and specialized reference materials, both print and digital, to find the pronunciation of a word or determine or clarify its precise meaning, part of speech, or origin.	○
8.RV.3.1 Analyze the meaning of words and phrases as they are used in works of literature, including figurative and connotative meanings; analyze the impact of specific word choices on meaning and tone, including analogies or allusions to other texts.	○
8.RV.3.2 Determine the meaning of words and phrases as they are used in a nonfiction text, including figurative, connotative, and technical meanings; analyze the impact of specific word choices on meaning and tone, including analogies or allusions to other texts.	○
6-8.LH.3.1 Determine the meaning of words and phrases as they are used in a text, including vocabulary specific to domains related to history/social studies.	★
6-8.LST.3.1 Determine the meaning of symbols, key terms, and other domain-specific words and phrases as they are used in a specific scientific or technical context relevant to grades 6-8 texts and topics.	★
8.RV.3.3 Interpret figures of speech (e.g., verbal irony, puns) in context.	★
Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy	
8.ML.2.1 Identify and analyze persuasive and propaganda techniques used in visual and verbal messages by electronic, print and mass media, and identify false or misleading information.	○
8.ML.2.2 Analyze and interpret how people experience media messages differently, depending on point of view, culture, etc.	○

8.RL.3.1 Compare and contrast the structure of two or more related works of literature (e.g., similar topic or theme), and analyze and evaluate how the differing structure of each text contributes to its meaning and style.	★
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.LH.6.2 Use technology to produce and publish writing and present the relationships between information and ideas clearly and efficiently.	★
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.	★
6-8.LST.6.2 Use technology to produce and publish writing and present the relationships between information and ideas clearly and efficiently.	★
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.LH.7.3 Draw evidence from informational texts to support analysis, reflection, and research.	★
6-8.LST.7.1 Conduct short research assignments and tasks to answer a question (including a self-generated question), or test a hypothesis, drawing on several sources and generating additional related, focused questions that allow for multiple avenues of exploration.	★
6-8.LST.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 CSE).	★
6-8.LST.7.3 Draw evidence from informational texts to support analysis, reflection, and research.	★
8.W.6.1b 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 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



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 at the end of the selected administration**

Corporation Performance on Each Standard for the Mathematics Test

What are my corporation's strengths and weaknesses in the Mathematics Standards?

Test: ILEARN Mathematics Grade 8
Year: Spring 2019
Name: Demo Corporation

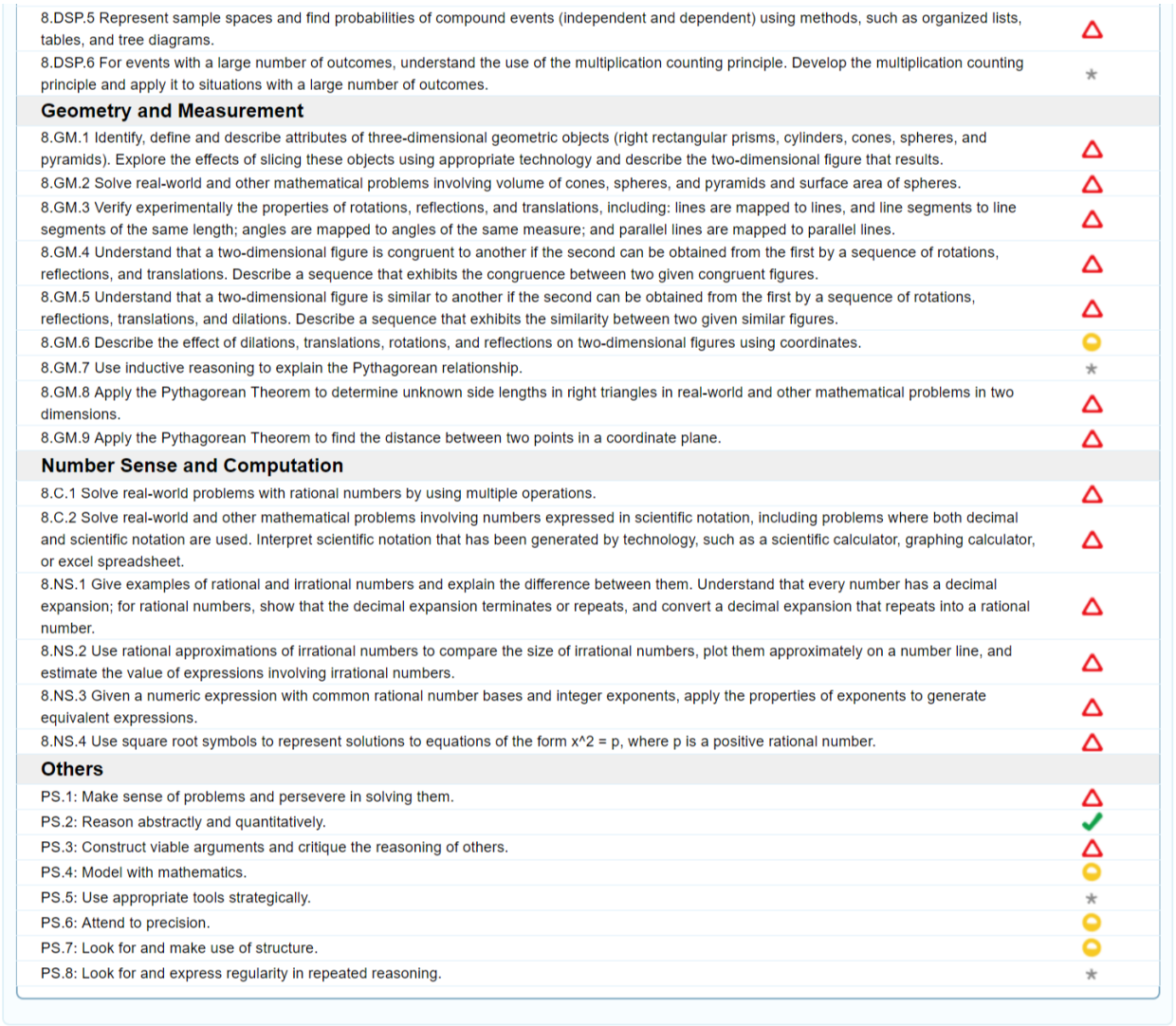
Legend: Areas Where Performance Indicates Proficiency
- Above the Proficiency Standard
- Borderline
- Below the Proficiency Standard
- Insufficient Information

Average Scale Scores on the ILEARN Mathematics Grade 8 Test: Demo Corporation and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	6550
Demo Corporation (9999)	6533

Performance on the ILEARN Mathematics Grade 8 Test, by Standard:Demo Corporation, Spring 2019

Standards	Areas Where Performance Indicates Proficiency
Algebra and Functions	
8.AF.1 Solve linear equations with rational number coefficients fluently, including equations whose solutions require expanding expressions using the distributive property and collecting like terms. Represent real-world problems using linear equations and inequalities in one variable and solve such problems.	△
8.AF.2 Give examples of linear equations in one variable with one solution, infinitely many solutions, or no solutions. Show which of these possibilities is the case by transforming a given equation into simpler forms, until an equivalent equation of the form x = a, a = a, or a = b results (where a and b are different numbers).	△
8.AF.3 Understand that a function assigns to each x-value (independent variable) exactly one y-value (dependent variable), and that the graph of a function is the set of ordered pairs (x,y).	△
8.AF.4 Describe qualitatively the functional relationship between two quantities by analyzing a graph (e.g., where the function is increasing or decreasing, linear or nonlinear, has a maximum or minimum value). Sketch a graph that exhibits the qualitative features of a function that has been verbally described.	△
8.AF.5 Interpret the equation y = mx + b as defining a linear function, whose graph is a straight line; give examples of functions that are not linear. Describe similarities and differences between linear and nonlinear functions from tables, graphs, verbal descriptions, and equations.	△
8.AF.6 Construct a function to model a linear relationship between two quantities given a verbal description, table of values, or graph. Recognize in y = mx + b that m is the slope (rate of change) and b is the y-intercept of the graph, and describe the meaning of each in the context of a problem.	△
8.AF.7 Compare properties of two linear functions given in different forms, such as a table of values, equation, verbal description, and graph (e.g., compare a distance-time graph to a distance-time equation to determine which of two moving objects has greater speed).	△
8.AF.8 Understand that solutions to a system of two linear equations correspond to points of intersection of their graphs because points of intersection satisfy both equations simultaneously. Approximate the solution of a system of equations by graphing and interpreting the reasonableness of the approximation.	△
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.	★

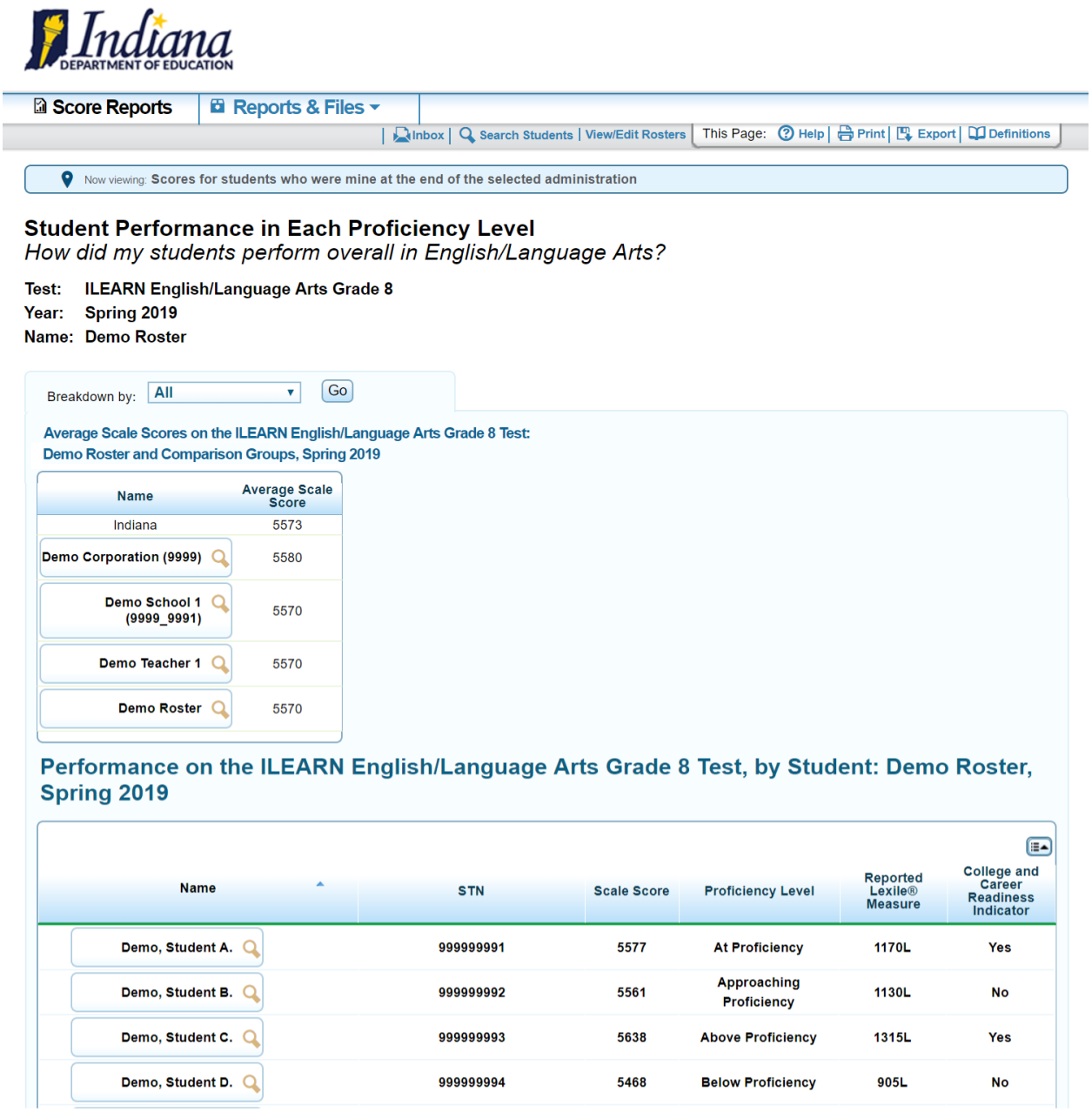
1.6.5 Student-Level Subject Report

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

- Scale score
- Overall subject performance level
- Lexile® (for ELA) or Quantile® (for Mathematics) measure

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

Inbox | Search Students | View/Edit Rosters | This Page: Help | Print | Export | Definitions

Now viewing: **Scores for students who were mine at the end of 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 2019
Name: Demo Roster

Breakdown by: All Go

Average Scale Scores on the ILEARN English/Language Arts Grade 8 Test: Demo Roster and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	5573
Demo Corporation (9999)	5580
Demo School 1 (9999_9991)	5570
Demo Teacher 1	5570
Demo Roster	5570

Performance on the ILEARN English/Language Arts Grade 8 Test, by Student: Demo Roster, Spring 2019

Name	STN	Scale Score	Proficiency Level	Reported Lexile® Measure	College and Career Readiness Indicator
Demo, Student A.	999999991	5577	At Proficiency	1170L	Yes
Demo, Student B.	999999992	5561	Approaching Proficiency	1130L	No
Demo, Student C.	999999993	5638	Above Proficiency	1315L	Yes
Demo, Student D.	999999994	5468	Below Proficiency	905L	No

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

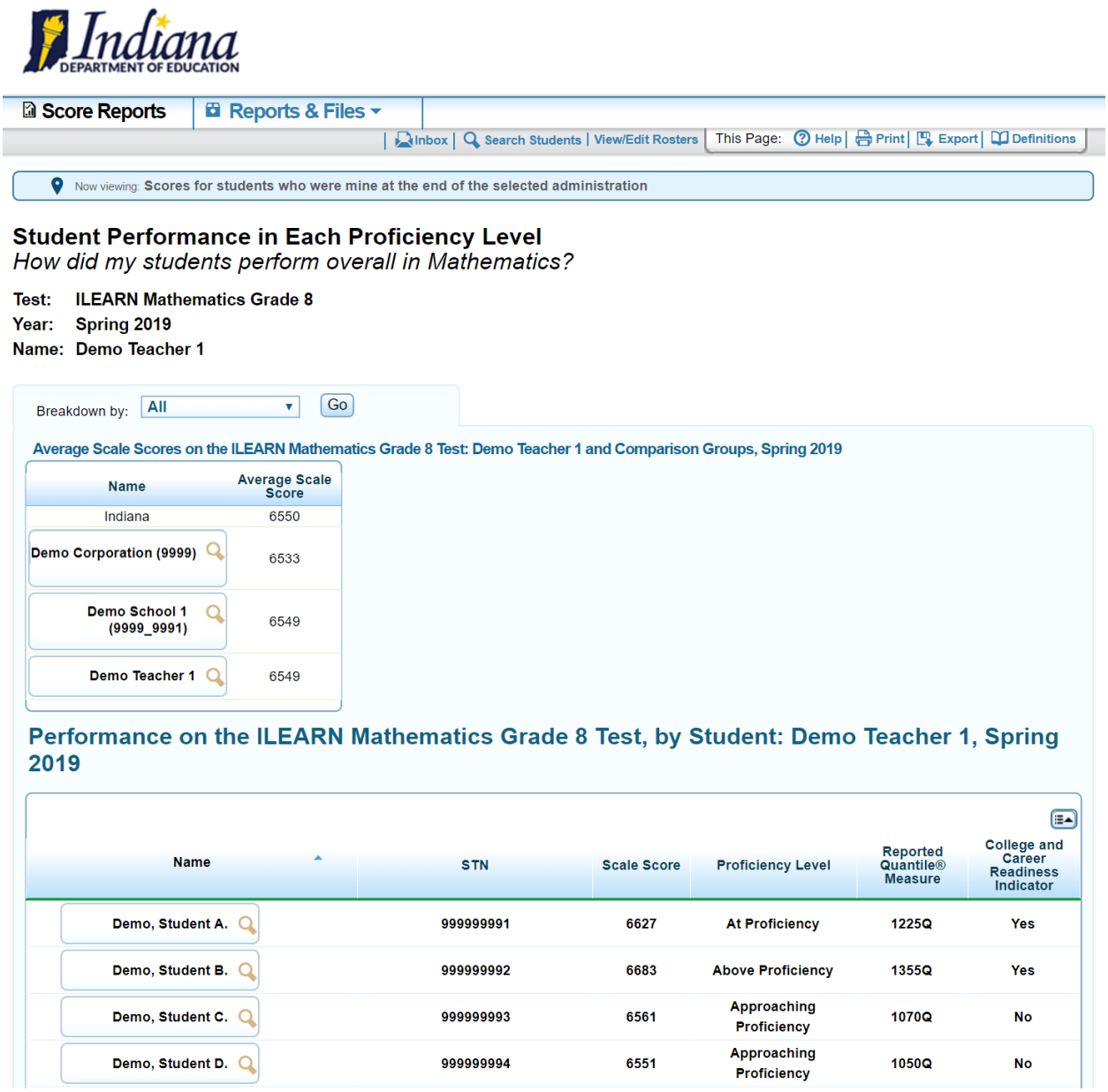


Figure 15: Student-Level Subject Report, Grade 6 Science

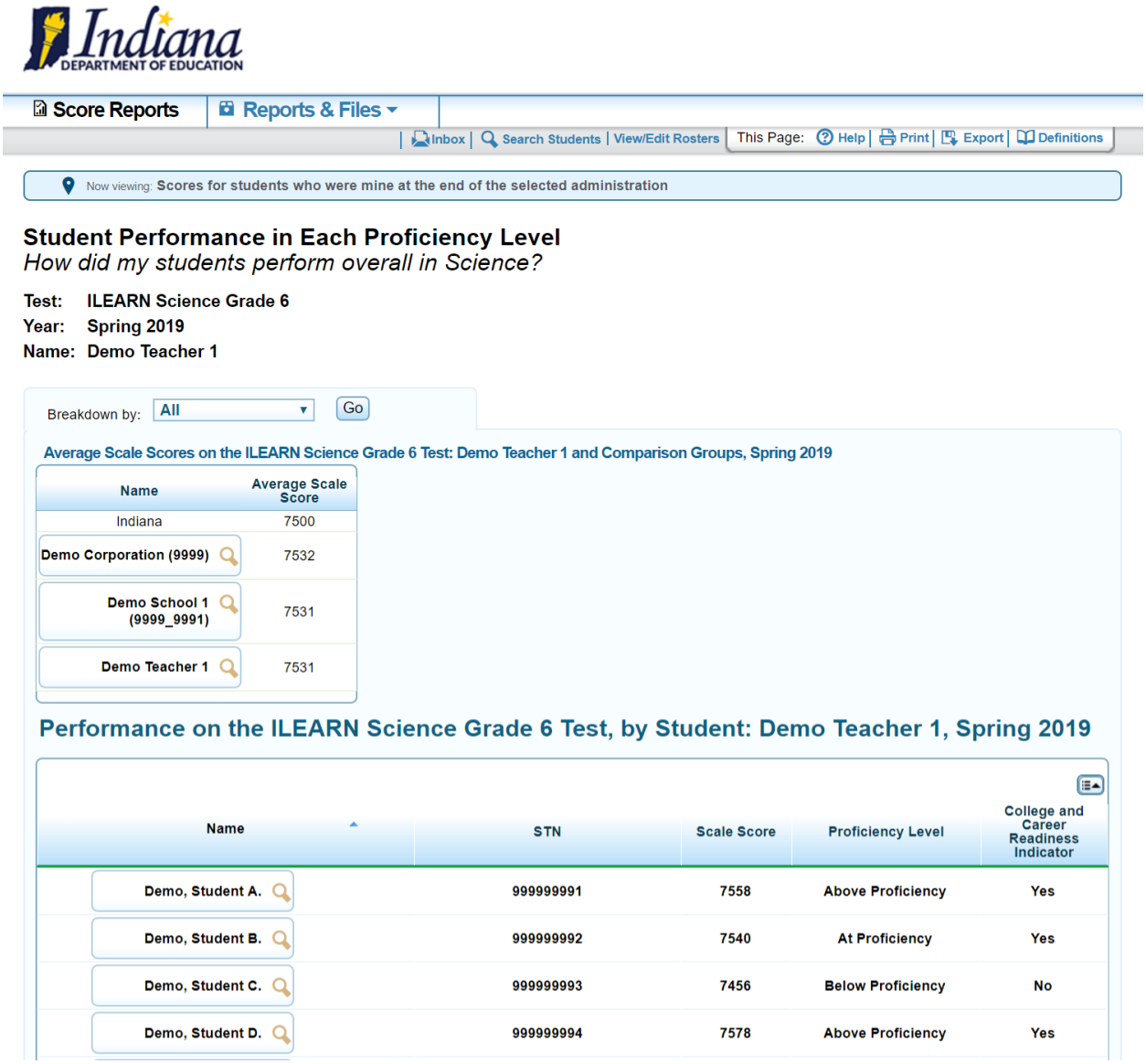
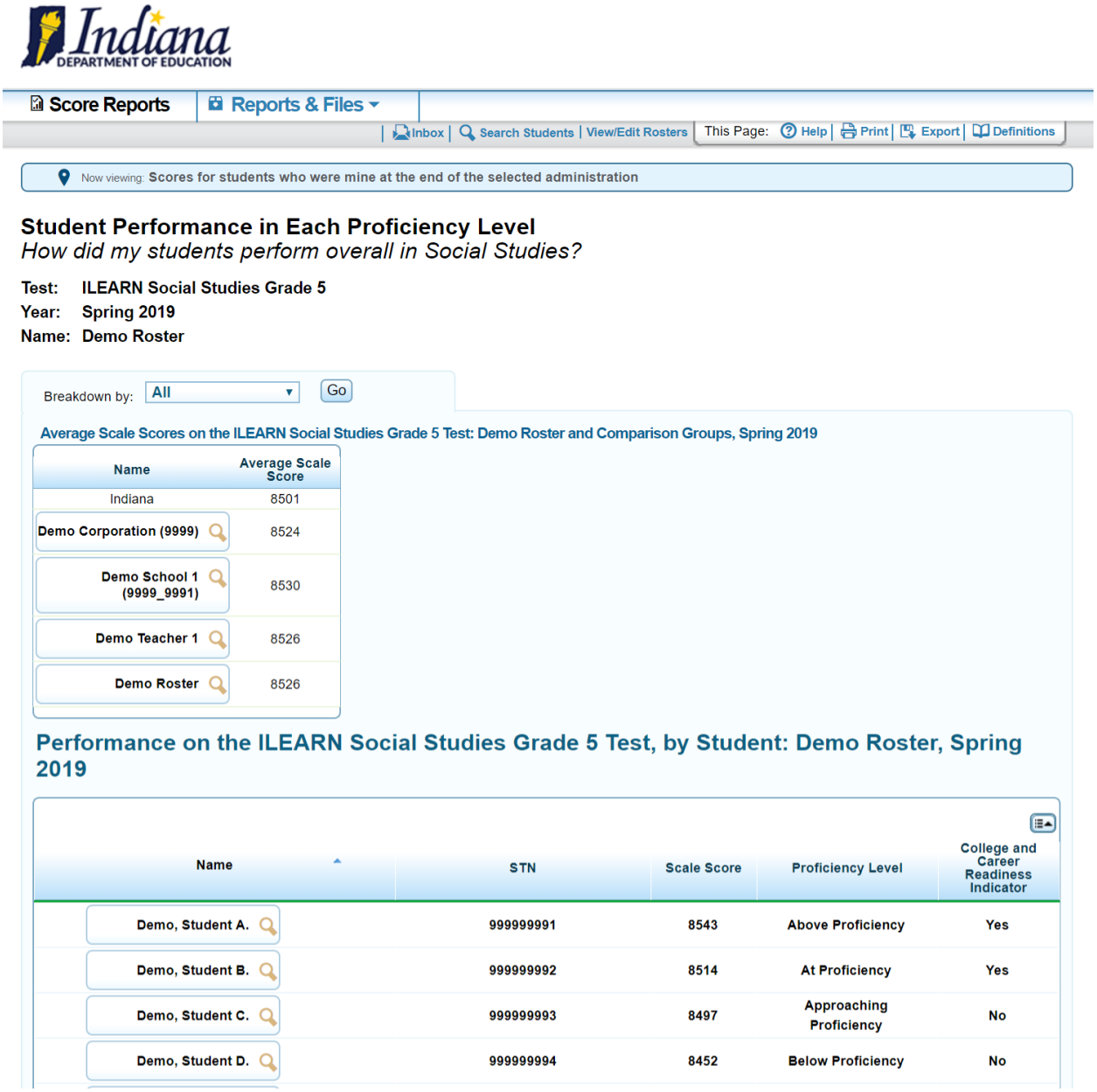


Figure 16: Student-Level Subject Report, Grade 5 Social Studies



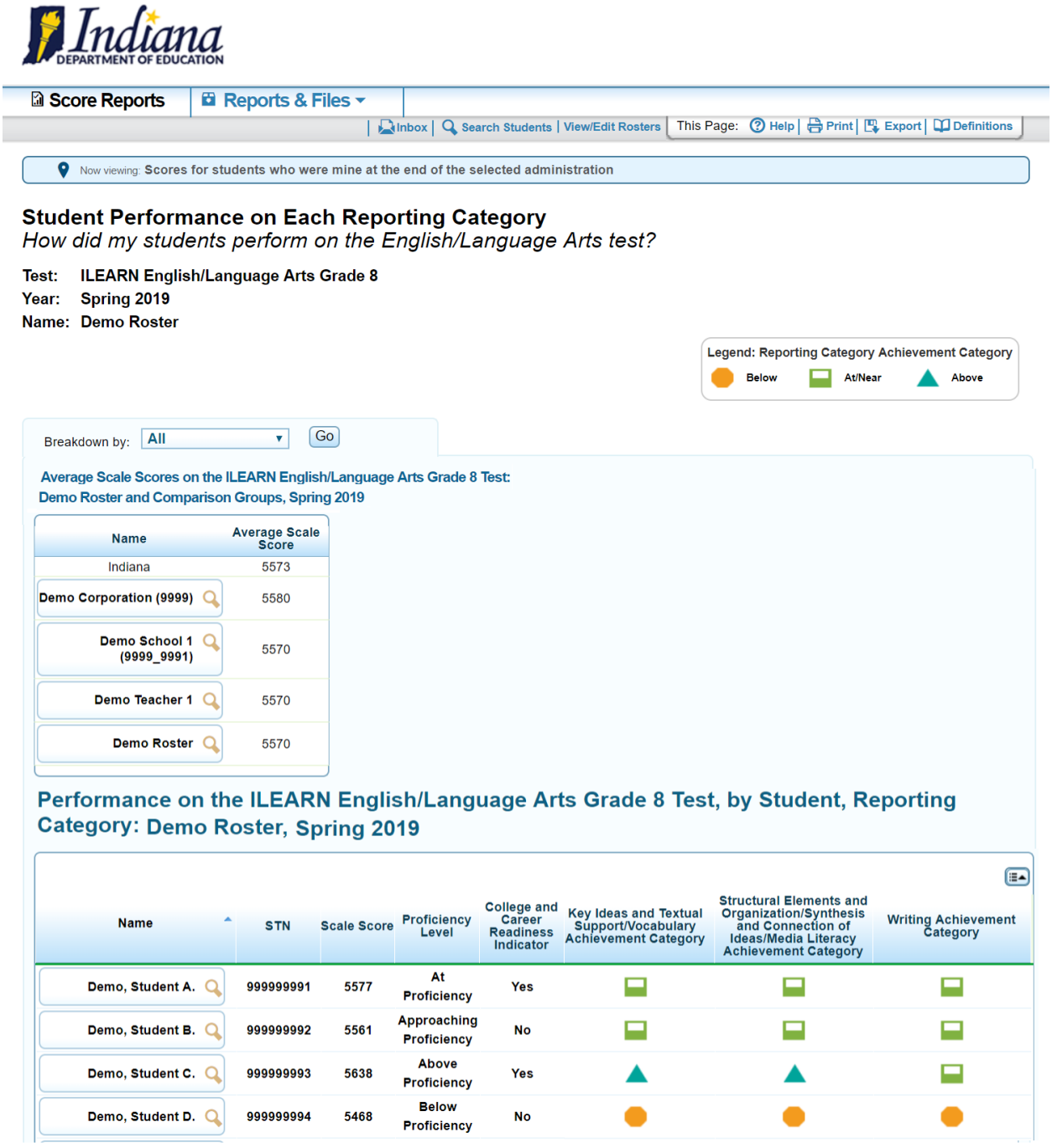
1.6.6 Student-Level Reporting Category Report

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

- Scale score
- Overall subject performance level
- Reporting category
- Performance category

Figure 17 through Figure 20 displays this information for ILEARN.

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



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 at the end of the selected administration

Student Performance on Each Reporting Category

How did my students perform on the English/Language Arts test?

Test: ILEARN English/Language Arts Grade 8
Year: Spring 2019
Name: Demo Roster

Legend: Reporting Category Achievement Category — Below | At/Near | Above

Breakdown by: All Go

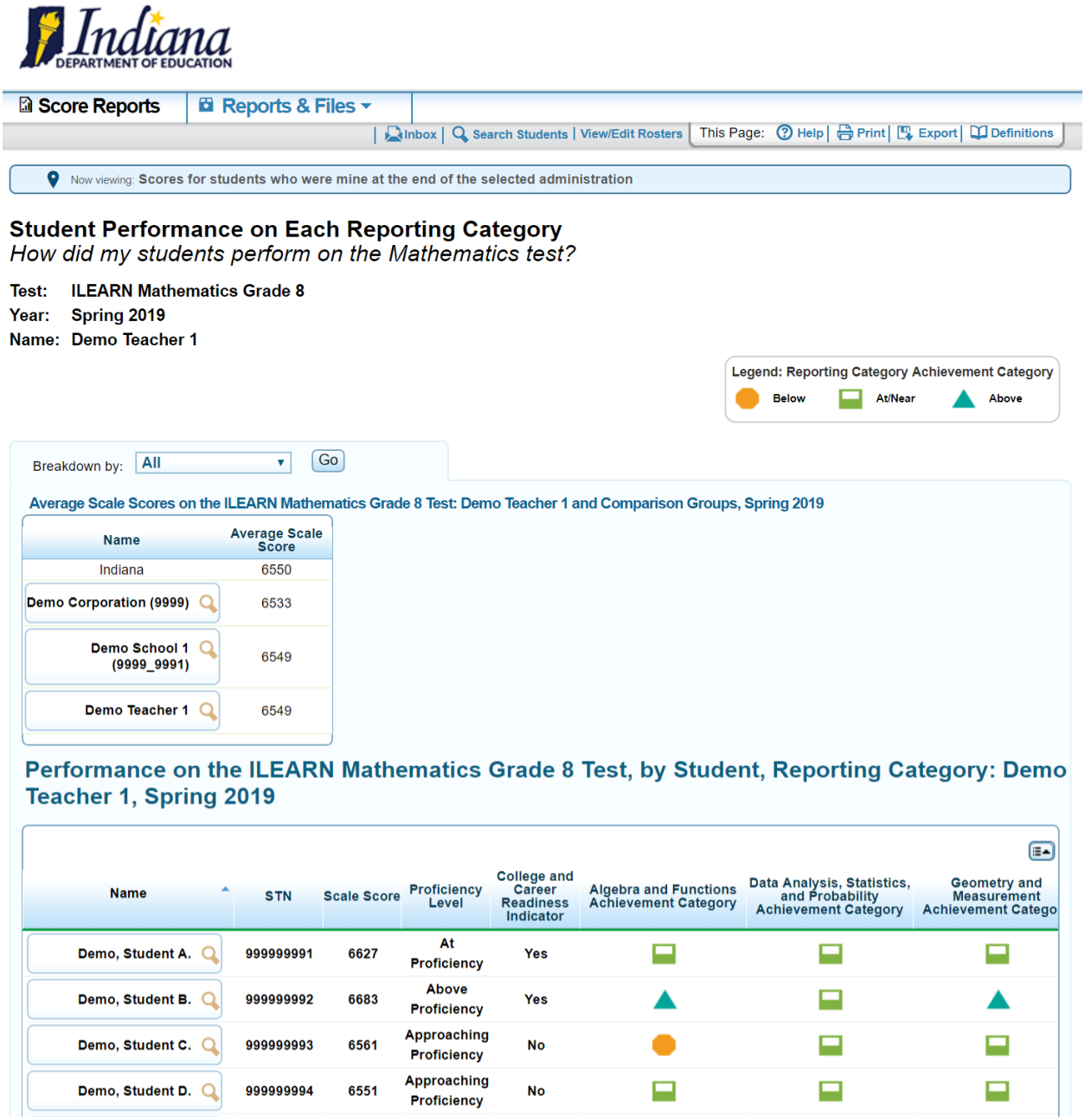
Average Scale Scores on the ILEARN English/Language Arts Grade 8 Test: Demo Roster and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	5573
Demo Corporation (9999)	5580
Demo School 1 (9999_9991)	5570
Demo Teacher 1	5570
Demo Roster	5570

Performance on the ILEARN English/Language Arts Grade 8 Test, by Student, Reporting Category: Demo Roster, Spring 2019

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator	Key Ideas and Textual Support/Vocabulary Achievement Category	Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy Achievement Category	Writing Achievement Category
Demo, Student A.	999999991	5577	At Proficiency	Yes	At/Near	At/Near	At/Near
Demo, Student B.	999999992	5561	Approaching Proficiency	No	At/Near	At/Near	At/Near
Demo, Student C.	999999993	5638	Above Proficiency	Yes	Above	Above	At/Near
Demo, Student D.	999999994	5468	Below Proficiency	No	Below	Below	Below

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



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 at the end of the selected administration**

Student Performance on Each Reporting Category

How did my students perform on the Mathematics test?

Test: ILEARN Mathematics Grade 8
Year: Spring 2019
Name: Demo Teacher 1

Legend: Reporting Category Achievement Category — Below | At/Near | Above

Breakdown by: All Go

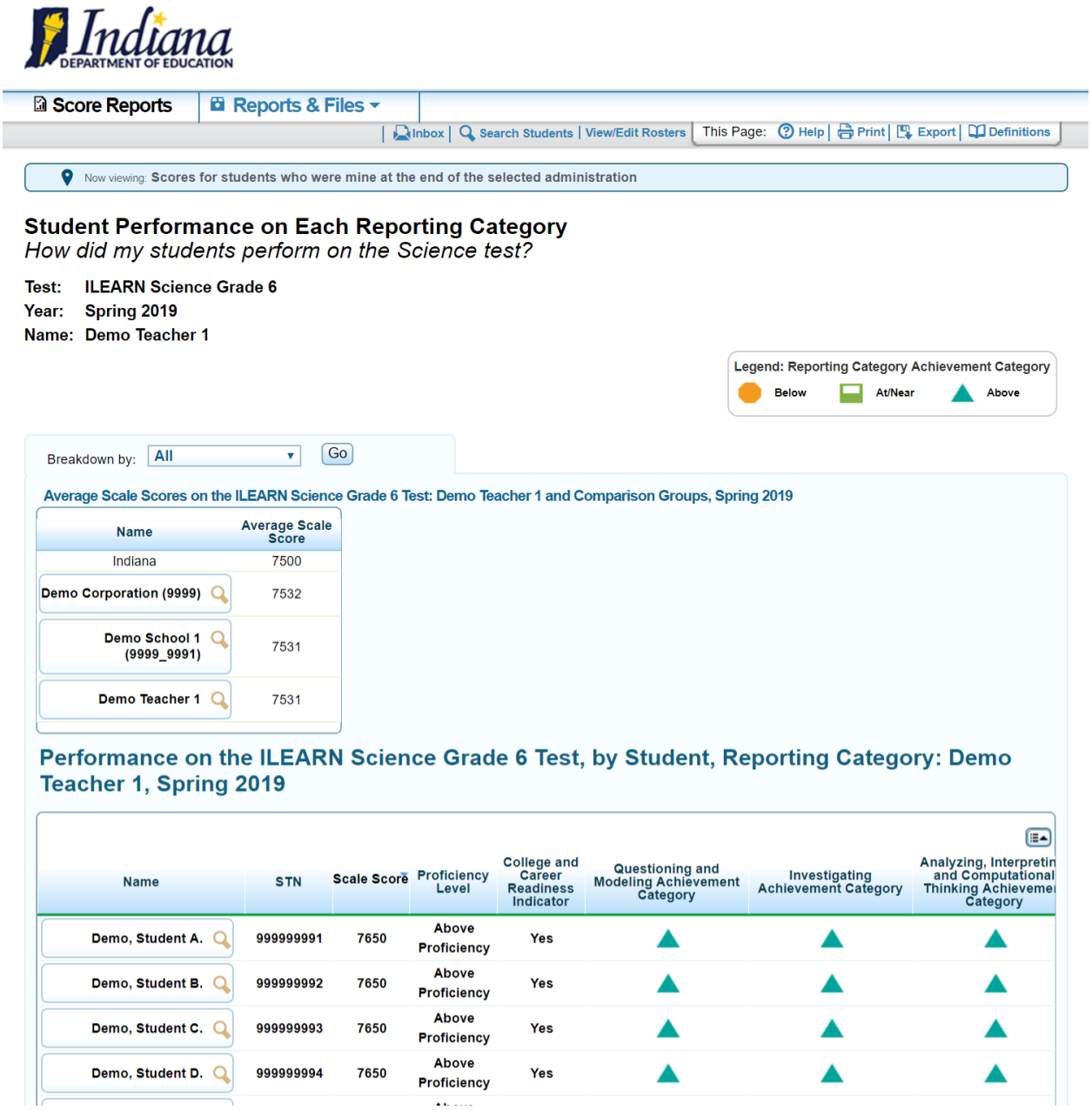
Average Scale Scores on the ILEARN Mathematics Grade 8 Test: Demo Teacher 1 and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	6550
Demo Corporation (9999)	6533
Demo School 1 (9999_9991)	6549
Demo Teacher 1	6549

Performance on the ILEARN Mathematics Grade 8 Test, by Student, Reporting Category: Demo Teacher 1, Spring 2019

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator	Algebra and Functions Achievement Category	Data Analysis, Statistics, and Probability Achievement Category	Geometry and Measurement Achievement Catego
Demo, Student A.	999999991	6627	At Proficiency	Yes	At/Near	At/Near	At/Near
Demo, Student B.	999999992	6683	Above Proficiency	Yes	Above	At/Near	Above
Demo, Student C.	999999993	6561	Approaching Proficiency	No	Below	At/Near	At/Near
Demo, Student D.	999999994	6551	Approaching Proficiency	No	At/Near	At/Near	At/Near

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



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 at the end of the selected administration

Student Performance on Each Reporting Category

How did my students perform on the Science test?

Test: ILEARN Science Grade 6
Year: Spring 2019
Name: Demo Teacher 1

Legend: Reporting Category Achievement Category — Below | At/Near | Above

Breakdown by: All Go

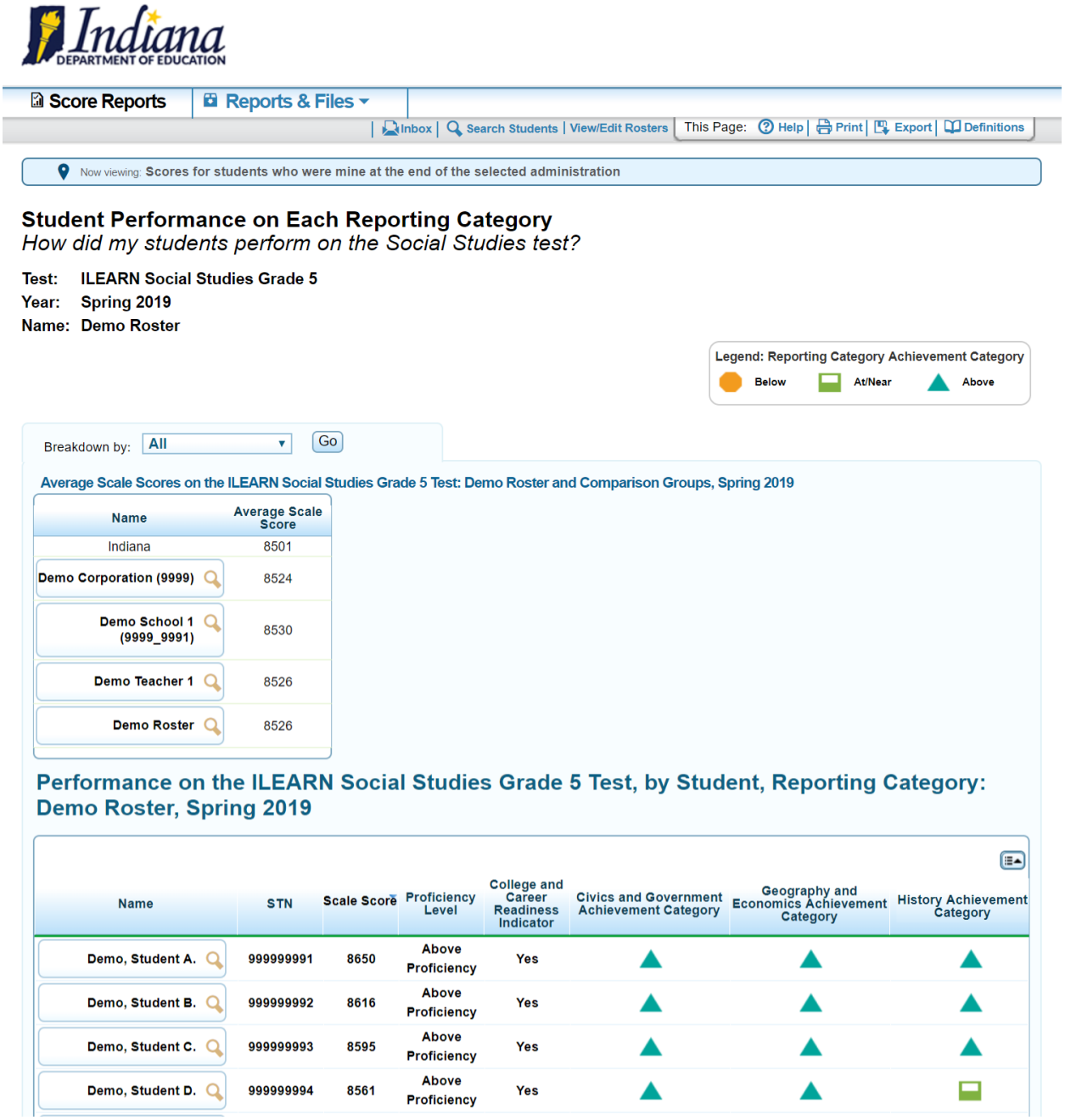
Average Scale Scores on the ILEARN Science Grade 6 Test: Demo Teacher 1 and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	7500
Demo Corporation (9999)	7532
Demo School 1 (9999_9991)	7531
Demo Teacher 1	7531

Performance on the ILEARN Science Grade 6 Test, by Student, Reporting Category: Demo Teacher 1, Spring 2019

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator	Questioning and Modeling Achievement Category	Investigating Achievement Category	Analyzing, Interpretin and Computational Thinking Achieveme Category
Demo, Student A.	999999991	7650	Above Proficiency	Yes	▲	▲	▲
Demo, Student B.	999999992	7650	Above Proficiency	Yes	▲	▲	▲
Demo, Student C.	999999993	7650	Above Proficiency	Yes	▲	▲	▲
Demo, Student D.	999999994	7650	Above Proficiency	Yes	▲	▲	▲

Figure 20: Student-Level Reporting Category 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 at the end of the selected administration

Student Performance on Each Reporting Category

How did my students perform on the Social Studies test?

Test: ILEARN Social Studies Grade 5
Year: Spring 2019
Name: Demo Roster

Legend: Reporting Category Achievement Category — Below | At/Near | Above

Breakdown by: All Go

Average Scale Scores on the ILEARN Social Studies Grade 5 Test: Demo Roster and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	8501
Demo Corporation (9999)	8524
Demo School 1 (9999_9991)	8530
Demo Teacher 1	8526
Demo Roster	8526

Performance on the ILEARN Social Studies Grade 5 Test, by Student, Reporting Category: Demo Roster, Spring 2019

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator	Civics and Government Achievement Category	Geography and Economics Achievement Category	History Achievement Category
Demo, Student A.	999999991	8650	Above Proficiency	Yes	▲	▲	▲
Demo, Student B.	999999992	8616	Above Proficiency	Yes	▲	▲	▲
Demo, Student C.	999999993	8595	Above Proficiency	Yes	▲	▲	▲
Demo, Student D.	999999994	8561	Above Proficiency	Yes	▲	▲	▬

1.6.7 Individual Student Report

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

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

The top of the report includes:

- Student's name
- Scale score with SEM
- Performance level
- Lexile® (ELA only) or Quantile® (Mathematics only)

The middle section includes:

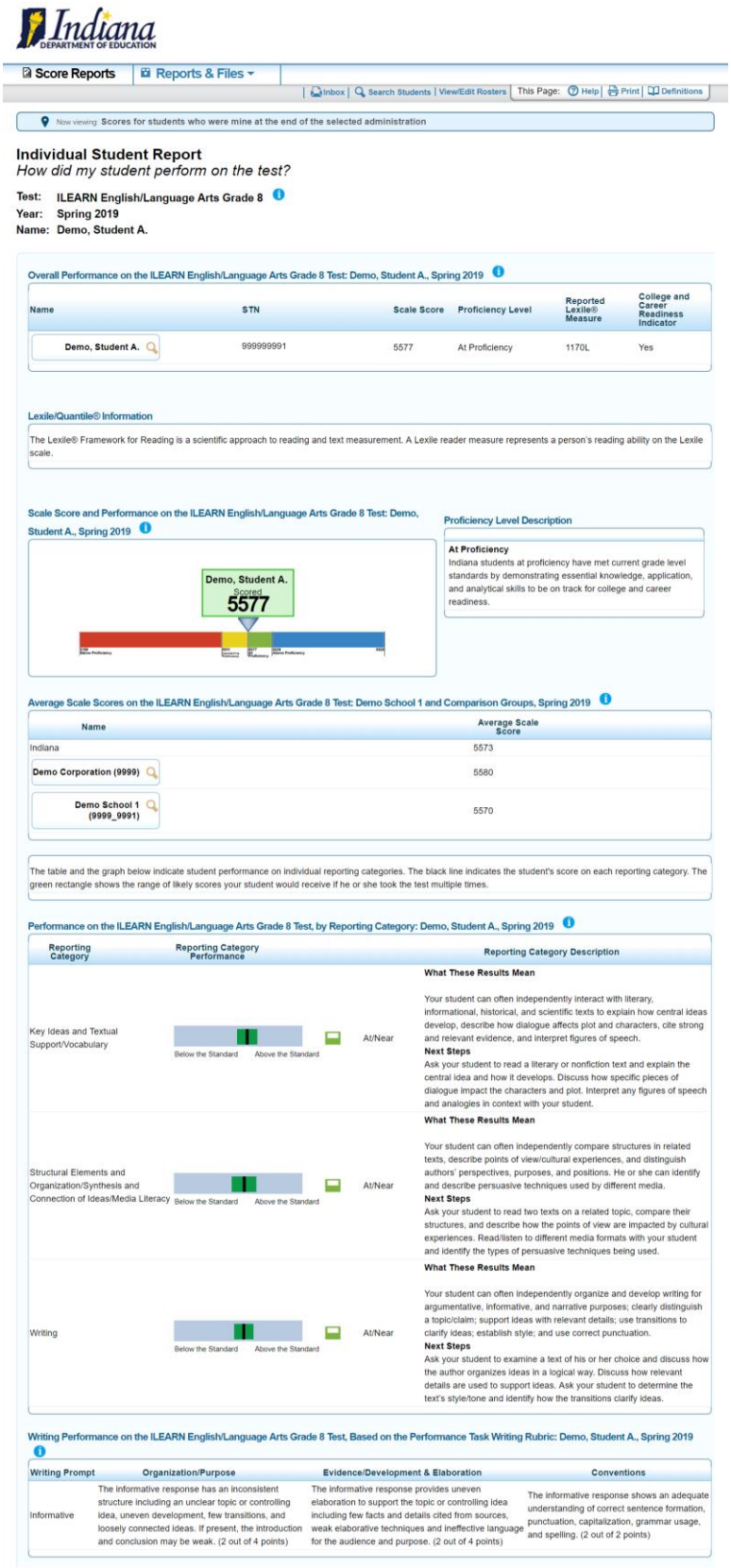
- Bar chart with the student's scale score
- Performance-level descriptors with cut scores at each performance level
- Average scale scores for state, corporation, and school aggregation levels

The bottom of the report includes:

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

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

Figure 21: Individual Student Report, Grade 8 ELA



Indiana Department of Education

Score Reports | Reports & Files

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

Now viewing: Scores for students who were mine at the end of the selected administration

Individual Student Report

How did my student perform on the test?

Test: ILEARN English/Language Arts Grade 8
Year: Spring 2019
Name: Demo, Student A.

Overall Performance on the ILEARN English/Language Arts Grade 8 Test: Demo, Student A., Spring 2019

Name	STN	Scale Score	Proficiency Level	Reported Lexile® Measure	College and Career Readiness Indicator
Demo, Student A.	999999991	5577	At Proficiency	1170L	Yes

Lexile/Quantile® Information

The Lexile® Framework for Reading is a scientific approach to reading and text measurement. A Lexile reader measure represents a person's reading ability on the Lexile scale.

Scale Score and Performance on the ILEARN English/Language Arts Grade 8 Test: Demo, Student A., Spring 2019

Demo, Student A.
Scored
5577

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 English/Language Arts Grade 8 Test: Demo School 1 and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	5573
Demo Corporation (9999)	5580
Demo School 1 (9999_9991)	5570

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.

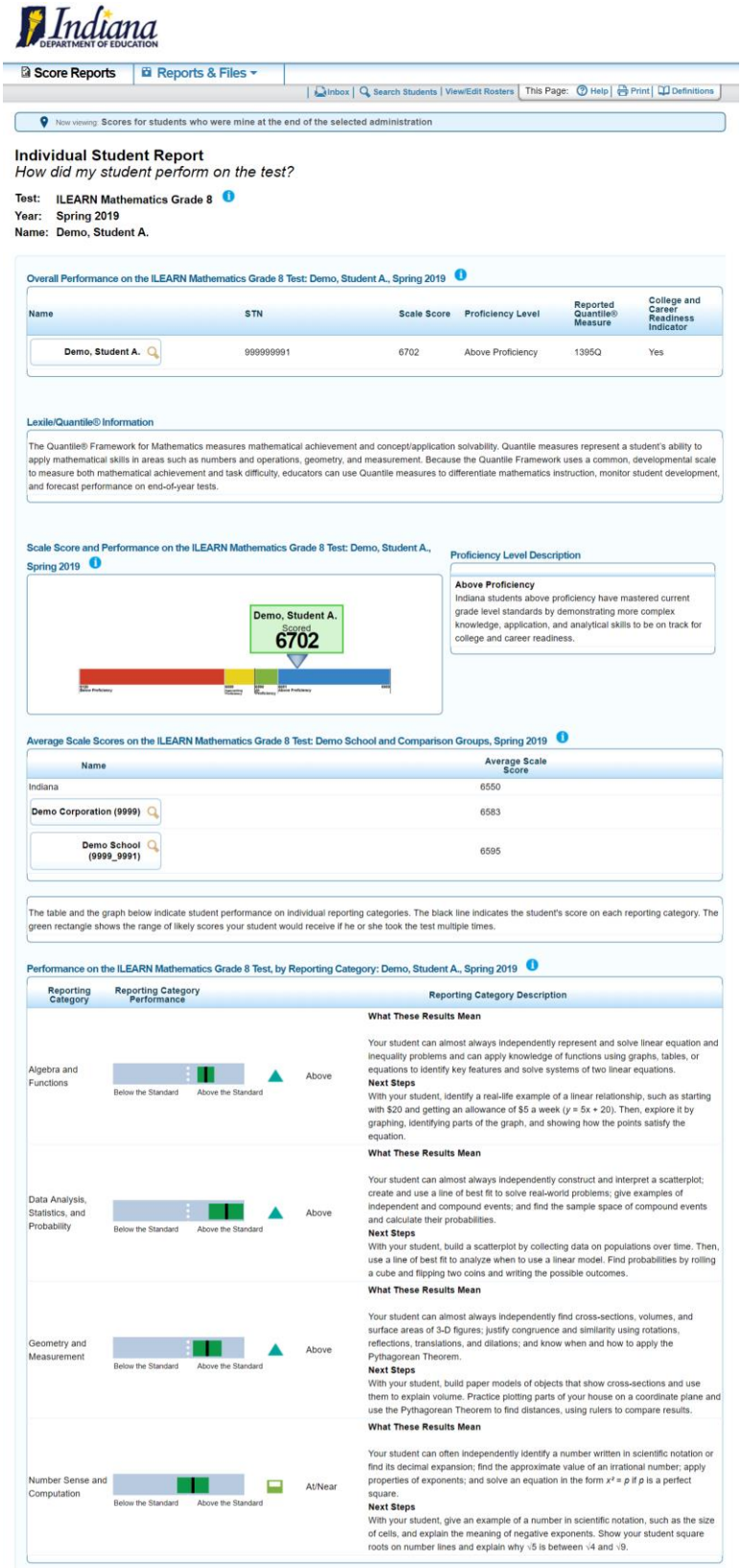
Performance on the ILEARN English/Language Arts Grade 8 Test, by Reporting Category: Demo, Student A., Spring 2019

Reporting Category	Reporting Category Performance	Reporting Category Description
Key Ideas and Textual Support/Vocabulary	Below the Standard / Above the Standard — At/Near	**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. **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.
Structural Elements and Organization/Synthesis and Connection of Ideas/Media Literacy	Below the Standard / Above the Standard — At/Near	**What These Results Mean** Your student can often independently compare structures in related texts, describe points of view/cultural experiences, and distinguish authors' perspectives, purposes, and positions. He or she can identify and describe persuasive techniques used by different media. **Next Steps** Ask your student to read two texts on a related topic, compare their structures, and describe how the points of view are impacted by cultural experiences. Read/listen to different media formats with your student and identify the types of persuasive techniques being used.
Writing	Below the Standard / Above the Standard — At/Near	**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. **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/tone and identify how the transitions clarify ideas.

Writing Performance on the ILEARN English/Language Arts Grade 8 Test, Based on the Performance Task Writing Rubric: Demo, Student A., Spring 2019

Writing Prompt	Organization/Purpose	Evidence/Development & Elaboration	Conventions
Informative	The informative response has an inconsistent structure including an unclear topic or controlling idea, uneven development, few transitions, and loosely connected ideas. If present, the introduction and conclusion may be weak. (2 out of 4 points)	The informative response provides uneven elaboration to support the topic or controlling idea 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 informative 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



Indiana Department of Education

Score Reports | Reports & Files

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

Now viewing: Scores for students who were mine at the end of the selected administration

Individual Student Report

How did my student perform on the test?

Test: ILEARN Mathematics Grade 8
Year: Spring 2019
Name: Demo, Student A.

Overall Performance on the ILEARN Mathematics Grade 8 Test: Demo, Student A., Spring 2019

Name	STN	Scale Score	Proficiency Level	Reported Quantile® Measure	College and Career Readiness Indicator
Demo, Student A.	999999991	6702	Above Proficiency	1395Q	Yes

Lexile/Quantile® Information

The Quantile® Framework for Mathematics measures mathematical achievement and concept/application solvability. Quantile measures represent a student's ability to apply mathematical skills in areas such as numbers and operations, geometry, and measurement. Because the Quantile Framework uses a common, developmental scale to measure both mathematical achievement and task difficulty, educators can use Quantile measures to differentiate mathematics instruction, monitor student development, and forecast performance on end-of-year tests.

Scale Score and Performance on the ILEARN Mathematics Grade 8 Test: Demo, Student A., Spring 2019

Demo, Student A. Scored 6702

Proficiency Level Description

Above Proficiency
Indiana students above proficiency have mastered current grade level standards by demonstrating more complex knowledge, application, and analytical skills to be on track for college and career readiness.

Average Scale Scores on the ILEARN Mathematics Grade 8 Test: Demo School and Comparison Groups, Spring 2019

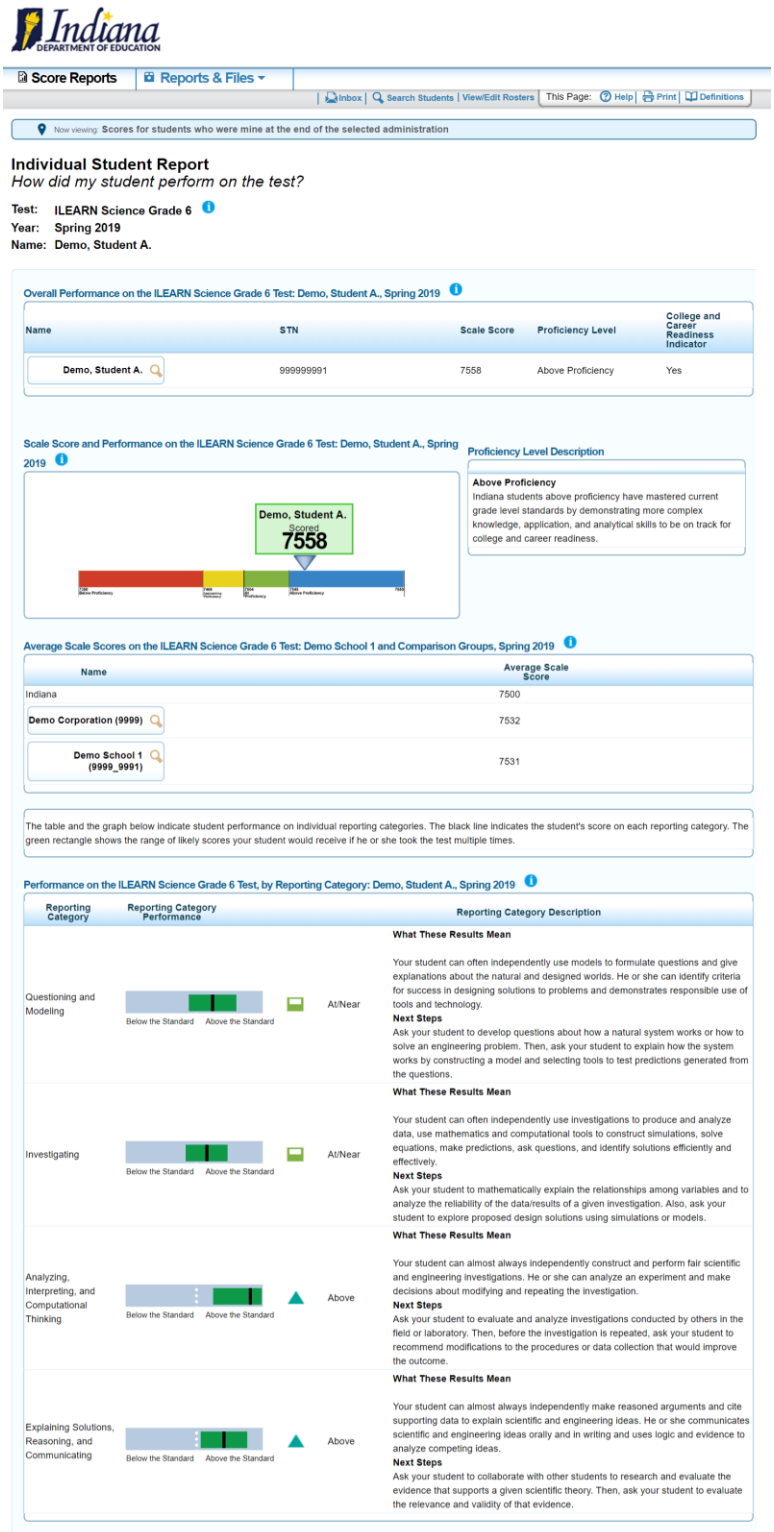
Name	Average Scale Score
Indiana	6550
Demo Corporation (9999)	6583
Demo School (9999_9991)	6595

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 Mathematics Grade 8 Test, by Reporting Category: Demo, Student A., Spring 2019

Reporting Category	Reporting Category Performance	Reporting Category Description
Algebra and Functions	Below the Standard / Above the Standard — Above	**What These Results Mean** Your student can almost always independently represent and solve linear equation and inequality problems and can apply knowledge of functions using graphs, tables, or equations to identify key features and solve systems of two linear equations. **Next Steps** With your student, identify a real-life example of a linear relationship, such as starting with $20 and getting an allowance of $5 a week ($y = 5x + 20$). Then, explore it by graphing, identifying parts of the graph, and showing how the points satisfy the equation.
Data Analysis, Statistics, and Probability	Below the Standard / Above the Standard — Above	**What These Results Mean** Your student can almost always independently construct and interpret a scatterplot; create and use a line of best fit to solve real-world problems; give examples of independent and compound events; and find the sample space of compound events and calculate their probabilities. **Next Steps** With your student, build a scatterplot by collecting data on populations over time. Then, use a line of best fit to analyze when to use a linear model. Find probabilities by rolling a cube and flipping two coins and writing the possible outcomes.
Geometry and Measurement	Below the Standard / Above the Standard — Above	**What These Results Mean** Your student can almost always independently find cross-sections, volumes, and surface areas of 3-D figures; justify congruence and similarity using rotations, reflections, translations, and dilations; and know when and how to apply the Pythagorean Theorem. **Next Steps** With your student, build paper models of objects that show cross-sections and use them to explain volume. Practice plotting parts of your house on a coordinate plane and use the Pythagorean Theorem to find distances, using rulers to compare results.
Number Sense and Computation	Below the Standard / Above the Standard — At/Near	**What These Results Mean** Your student can often independently identify a number written in scientific notation or find its decimal expansion; find the approximate value of an irrational number; apply properties of exponents; and solve an equation in the form $x^2 = p$ if p is a perfect square. **Next Steps** With your student, give an example of a number in scientific notation, such as the size of cells, and explain the meaning of negative exponents. Show your student square roots on number lines and explain why $\sqrt{5}$ is between $\sqrt{4}$ and $\sqrt{9}$.

Figure 23: Individual Student Report, Grade 6 Science



Indiana Department of Education

Score Reports | Reports & Files

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

Now viewing: Scores for students who were mine at the end of the selected administration

Individual Student Report

How did my student perform on the test?

Test: ILEARN Science Grade 6
Year: Spring 2019
Name: Demo, Student A.

Overall Performance on the ILEARN Science Grade 6 Test: Demo, Student A., Spring 2019

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator
Demo, Student A.	999999991	7558	Above Proficiency	Yes

Scale Score and Performance on the ILEARN Science Grade 6 Test: Demo, Student A., Spring 2019

Demo, Student A. Scored 7558

Proficiency Level Description

Above Proficiency
Indiana students above proficiency have mastered current grade level standards by demonstrating more complex knowledge, application, and analytical skills to be on track for college and career readiness.

Average Scale Scores on the ILEARN Science Grade 6 Test: Demo School 1 and Comparison Groups, Spring 2019

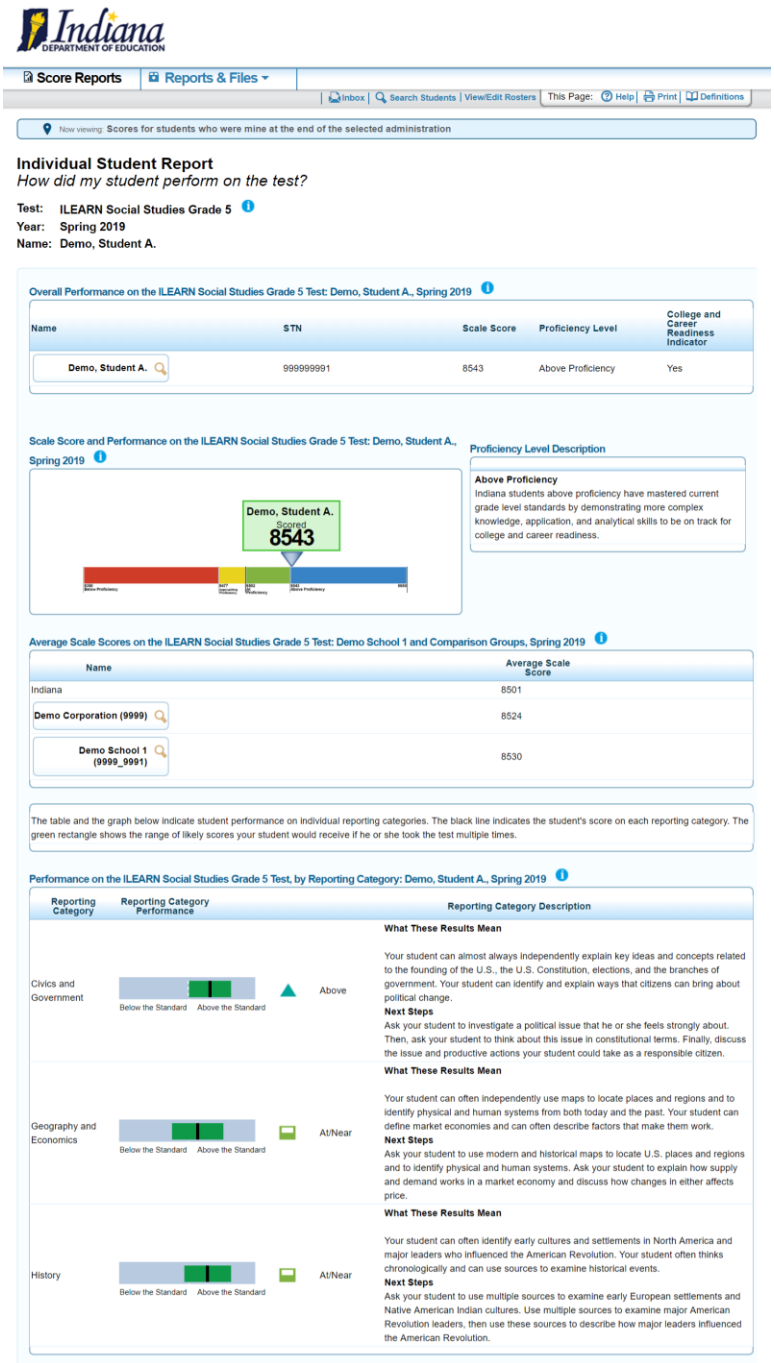
Name	Average Scale Score
Indiana	7500
Demo Corporation (9999)	7532
Demo School 1 (9999_9991)	7531

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 A., Spring 2019

Reporting Category	Reporting Category Performance	Reporting Category Description
Questioning and Modeling	Below the Standard / Above the Standard — At/Near	**What These Results Mean** Your student can often independently use models to formulate questions and give explanations about the natural and designed worlds. He or she can identify criteria for success in designing solutions to problems and demonstrates responsible use of tools and technology. **Next Steps** Ask your student to develop questions about how a natural system works or how to solve an engineering problem. Then, ask your student to explain how the system works by constructing a model and selecting tools to test predictions generated from the questions.
Investigating	Below the Standard / Above the Standard — At/Near	**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. **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.
Analyzing, Interpreting, and Computational Thinking	Below the Standard / Above the Standard — Above	**What These Results Mean** Your student can almost always independently construct and perform fair scientific and engineering investigations. He or she can analyze an experiment and make decisions about modifying and repeating the investigation. **Next Steps** Ask your student to evaluate and analyze investigations conducted by others in the field or laboratory. Then, before the investigation is repeated, ask your student to recommend modifications to the procedures or data collection that would improve the outcome.
Explaining Solutions, Reasoning, and Communicating	Below the Standard / Above the Standard — Above	**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. **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.

Figure 24: Individual Student Report, Grade 5 Social Studies



Indiana Department of Education

Score Reports | Reports & Files

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

Now viewing: Scores for students who were mine at the end of the selected administration

Individual Student Report

How did my student perform on the test?

Test: ILEARN Social Studies Grade 5
Year: Spring 2019
Name: Demo, Student A.

Overall Performance on the ILEARN Social Studies Grade 5 Test: Demo, Student A., Spring 2019

Name	STN	Scale Score	Proficiency Level	College and Career Readiness Indicator
Demo, Student A.	999999991	8543	Above Proficiency	Yes

Scale Score and Performance on the ILEARN Social Studies Grade 5 Test: Demo, Student A., Spring 2019

Demo, Student A.
Scored
8543

Proficiency Level Description

Above Proficiency
Indiana students above proficiency have mastered current grade level standards by demonstrating more complex knowledge, application, and analytical skills to be on track for college and career readiness.

Average Scale Scores on the ILEARN Social Studies Grade 5 Test: Demo School 1 and Comparison Groups, Spring 2019

Name	Average Scale Score
Indiana	8501
Demo Corporation (9999)	8524
Demo School 1 (9999_9991)	8530

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.

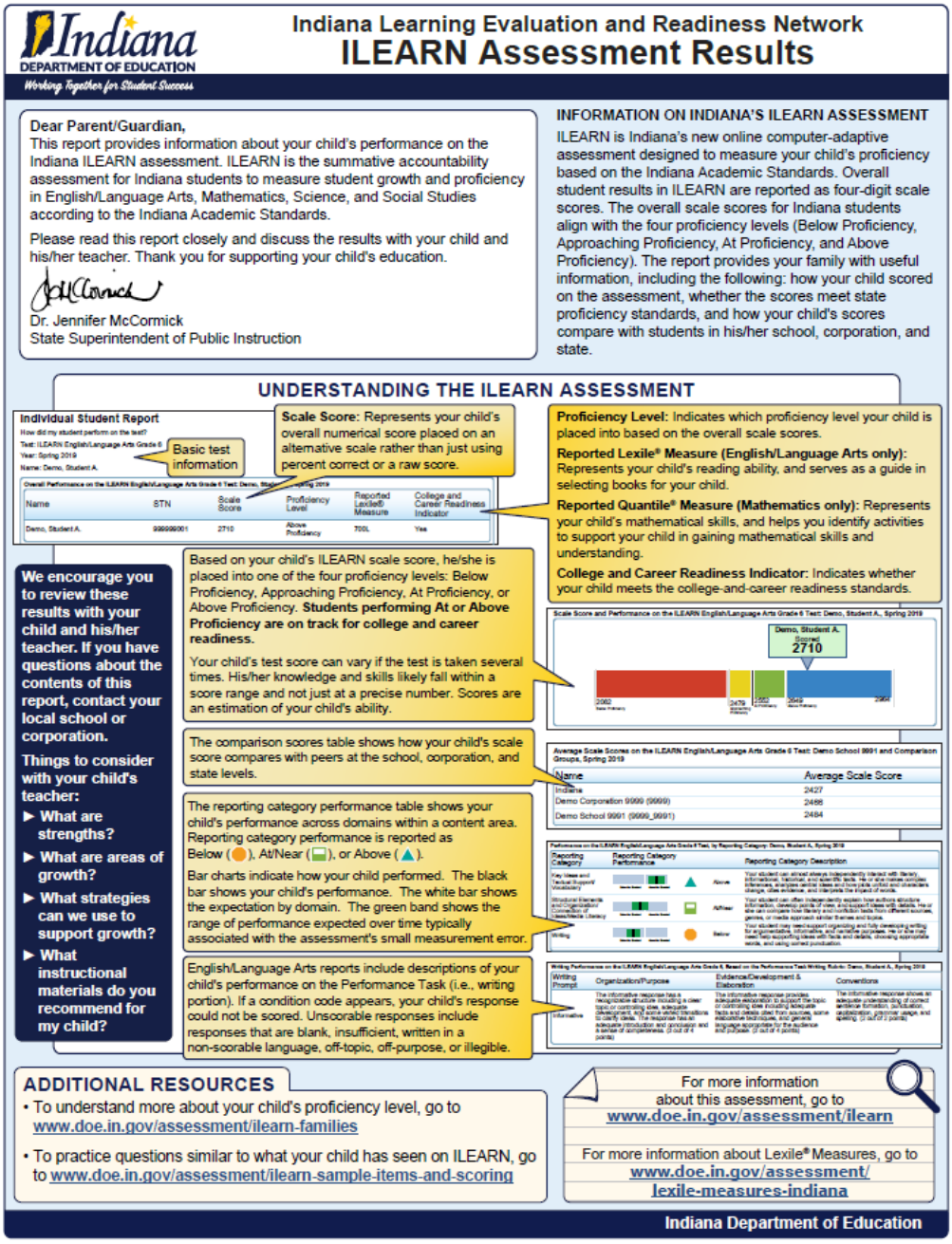
Performance on the ILEARN Social Studies Grade 5 Test, by Reporting Category: Demo, Student A., Spring 2019

Reporting Category	Reporting Category Performance	Reporting Category Description
Civics and Government	Below the Standard / Above the Standard — Above	**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. **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.
Geography and Economics	Below the Standard / Above the Standard — At/Near	**What These Results Mean** Your student can often independently use maps to locate places and regions and to identify physical and human systems from both today and the past. Your student can define market economies and can often describe factors that make them work. **Next Steps** Ask your student to use modern and historical maps to locate U.S. places and regions and to identify physical and human systems. Ask your student to explain how supply and demand works in a market economy and discuss how changes in either affects price.
History	Below the Standard / Above the Standard — At/Near	**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. **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.

1.6.8 Interpretive Guide

When printing ISRs, users have the option to print a supplemental "interpretive guide" (also called an "Addendum" when printing a Simple ISR), which is intended to serve as a stand-alone document (see Figure 25) to help teachers, administrators, parents, and students better understand the data presented in the ISR. The ISRs and the supplemental "interpretive guide" are also available in five different languages: Arabic, Chinese, Burmese, Spanish, and Vietnamese.

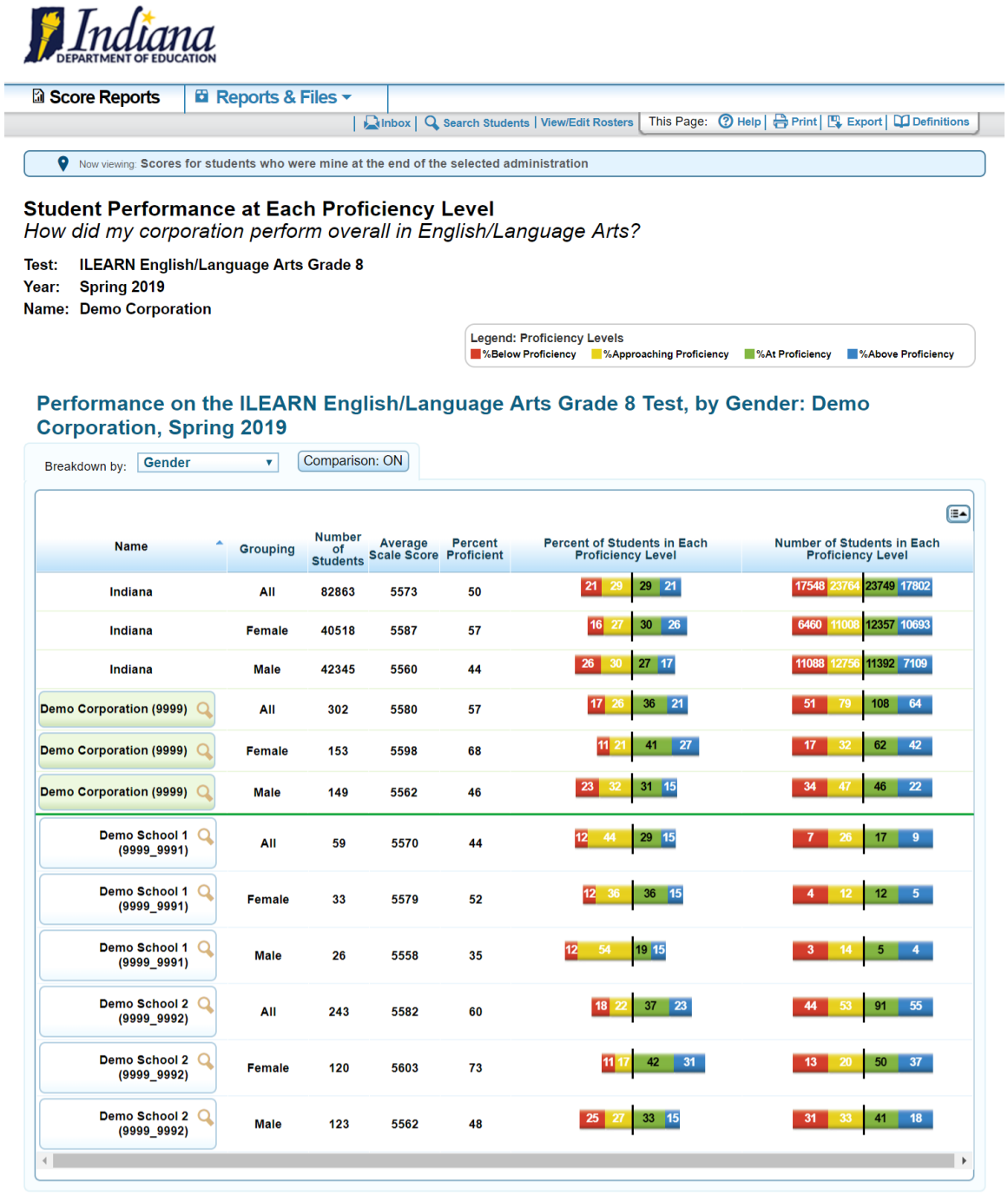
Figure 25: Supplemental Interpretive Guide



1.6.9 Reports by Sub-Group

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

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



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 at the end of the selected administration

Student Performance at Each Proficiency Level

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

Test: ILEARN English/Language Arts Grade 8
Year: Spring 2019
Name: Demo Corporation

Legend: Proficiency Levels
%Below Proficiency | %Approaching Proficiency | %At Proficiency | %Above Proficiency

Performance on the ILEARN English/Language Arts Grade 8 Test, by Gender: Demo Corporation, Spring 2019

Breakdown by: Gender | Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Percent of Students in Each Proficiency Level	Number of Students in Each Proficiency Level
Indiana	All	82863	5573	50	21 29 29 21	17548 23764 23749 17802
Indiana	Female	40518	5587	57	16 27 30 26	6460 11008 12357 10693
Indiana	Male	42345	5560	44	26 30 27 17	11088 12756 11392 7109
Demo Corporation (9999)	All	302	5580	57	17 26 36 21	51 79 108 64
Demo Corporation (9999)	Female	153	5598	68	11 21 41 27	17 32 62 42
Demo Corporation (9999)	Male	149	5562	46	23 32 31 15	34 47 46 22
Demo School 1 (9999_9991)	All	59	5570	44	12 44 29 15	7 26 17 9
Demo School 1 (9999_9991)	Female	33	5579	52	12 36 36 15	4 12 12 5
Demo School 1 (9999_9991)	Male	26	5558	35	12 54 19 15	3 14 5 4
Demo School 2 (9999_9992)	All	243	5582	60	18 22 37 23	44 53 91 55
Demo School 2 (9999_9992)	Female	120	5603	73	11 17 42 31	13 20 50 37
Demo School 2 (9999_9992)	Male	123	5562	48	25 27 33 15	31 33 41 18

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

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 at the end of the selected administration

Student Performance for Each Reporting Category

What are my corporation's strengths and weaknesses in Mathematics?

Test: ILEARN Mathematics Grade 8
Year: Spring 2019
Name: Demo Corporation

Legend: Reporting Category Achievement Category — %Below | %At/Near | %Above

Performance on the ILEARN Mathematics Grade 8 Test, by Reporting Category, by Section 504 Plan: Demo Corporation, Spring 2019

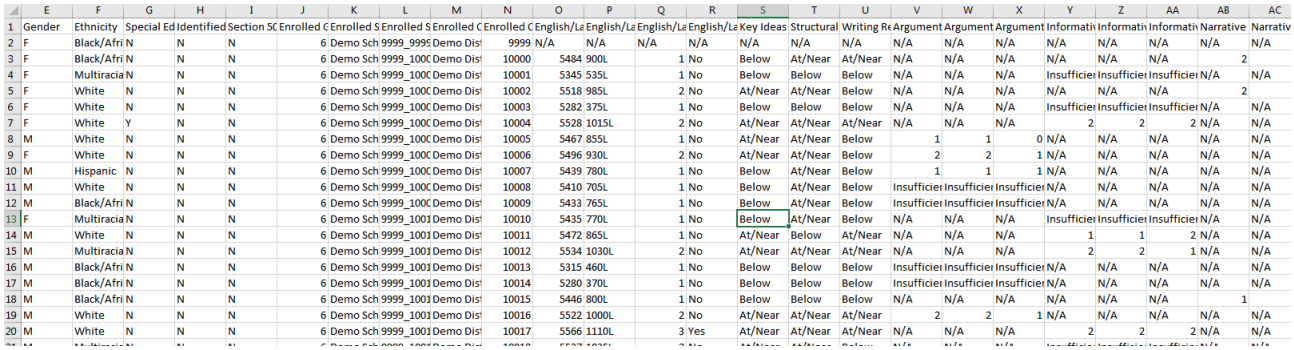
Breakdown by: Section 504 Plan | Comparison: ON

Name	Grouping	Number of Students	Average Scale Score	Percent Proficient	Reporting Category	Percent at Each Reporting Category Achievement Category
Indiana	All	82863	6550	37	Mathematics	
					Algebra and Functions	42 39 19
					Data Analysis, Statistics, and Probability	33 47 20
					Geometry and Measurement	37 47 16
					Number Sense and Computation	26 55 19
Indiana	Not Section 504 Plan	80604	6551	38	Mathematics	
					Algebra and Functions	42 39 20
					Data Analysis, Statistics, and Probability	33 47 20
					Geometry and Measurement	37 47 16
					Number Sense and Computation	26 55 19
Indiana	Section 504 Plan	2259	6530	26	Mathematics	
					Algebra and Functions	49 40 11
					Data Analysis, Statistics, and Probability	37 49 13
					Geometry and Measurement	43 47 10
					Number Sense and Computation	30 58 12
Demo Corporation (9999)	All	744	6533	29	Mathematics	
					Algebra and Functions	49 37 14
					Data Analysis, Statistics, and Probability	33 54 13
					Geometry and Measurement	38 51 10
					Number Sense and Computation	29 58 13
Demo Corporation (9999)	Not Section 504 Plan	717	6533	30	Mathematics	
					Algebra and Functions	48 37 15
					Data Analysis, Statistics, and Probability	33 54 13
					Geometry and Measurement	38 52 10
					Number Sense and Computation	28 59 13
Demo Corporation (9999)	Section 504 Plan	27	6527	19	Mathematics	
					Algebra and Functions	63 30 7
					Data Analysis, Statistics, and Probability	30 56 15
					Geometry and Measurement	48 48 4
					Number Sense and Computation	44 52 4
Demo School 1 (9999_9991)	All	169	6549	32	Mathematics	
					Algebra and Functions	48 31 21
					Data Analysis, Statistics, and Probability	26 59 15
					Geometry and Measurement	23 63 14
					Number Sense and Computation	22 66 12
Demo School 1 (9999_9991)	Not Section 504 Plan	167	6550	32	Mathematics	
					Algebra and Functions	47 32 21
					Data Analysis, Statistics, and Probability	26 58 16
					Geometry and Measurement	23 63 14
					Number Sense and Computation	22 65 13
Demo School 1 (9999_9991)	Section 504 Plan	2	6518	0	Mathematics	
					Algebra and Functions	100
					Data Analysis, Statistics, and Probability	100
					Geometry and Measurement	50 50
					Number Sense and Computation	100
Demo School 2 (9999_9992)	All	7	6426	0	Mathematics	
					Algebra and Functions	86 14
					Data Analysis, Statistics, and Probability	86 14
					Geometry and Measurement	100
					Number Sense and Computation	86 14
Demo School 2 (9999_9992)	Not Section 504 Plan	5	6426	0	Mathematics	
					Algebra and Functions	100
					Data Analysis, Statistics, and Probability	80 20
					Geometry and Measurement	100
					Number Sense and Computation	100
Demo School 2 (9999_9992)	Section 504 Plan	2	6424	0	Mathematics	
					Algebra and Functions	50 50
					Data Analysis, Statistics, and Probability	100
					Geometry and Measurement	100
					Number Sense and Computation	50 50

1.6.10 Data File

ORS users have the option to quickly generate a comprehensive data file of their students' scores. Data files (see Figure 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
1	Gender	Ethnicity	Special Ed	Identified	Section 50	Enrolled G	Enrolled S	Enrolled S	Enrolled C	Enrolled C	English/La	English/La	English/La	English/La	Key Ideas	Structural	Writing Re	Argument	Argument	Argument	Informativ	Informativ	Informativ	Narrative	Narrativ
2	F	Black/Afri	N	N	N	6	Demo Sch	9999_9995	Demo Dist	9999	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	F	Black/Afri	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10000	5484	900L		1	No	Below	At/Near	At/Near	N/A	N/A	N/A	N/A	N/A	N/A	2	
4	F	Multiracia	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10001	5345	535L		1	No	Below	Below	Below	N/A	N/A	N/A	Insufficier	Insufficier	Insufficier	N/A	N/A
5	F	White	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10002	5518	985L		2	No	At/Near	At/Near	Below	N/A	N/A	N/A	N/A	N/A	N/A	2	
6	F	White	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10003	5282	375L		1	No	Below	Below	Below	N/A	N/A	N/A	Insufficier	Insufficier	Insufficier	N/A	N/A
7	F	White	Y	N	N	6	Demo Sch	9999_1000	Demo Dist	10004	5528	1015L		2	No	At/Near	At/Near	At/Near	N/A	N/A	N/A	2	2	2	N/A	N/A
8	M	White	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10005	5467	855L		1	No	At/Near	At/Near	Below	1	1	0	N/A	N/A	N/A	N/A	N/A
9	F	White	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10006	5496	930L		2	No	At/Near	At/Near	Below	2	2	1	N/A	N/A	N/A	N/A	N/A
10	M	Hispanic	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10007	5439	780L		1	No	Below	At/Near	Below	1	1	1	N/A	N/A	N/A	N/A	N/A
11	M	White	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10008	5410	705L		1	No	Below	At/Near	Below	Insufficier	Insufficier	Insufficier	N/A	N/A	N/A	N/A	N/A
12	M	Black/Afri	N	N	N	6	Demo Sch	9999_1000	Demo Dist	10009	5433	765L		1	No	Below	At/Near	Below	Insufficier	Insufficier	Insufficier	N/A	N/A	N/A	N/A	N/A
13	F	Multiracia	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10010	5435	770L		1	No	Below	At/Near	Below	N/A	N/A	N/A	Insufficier	Insufficier	Insufficier	N/A	N/A
14	M	White	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10011	5472	865L		1	No	At/Near	Below	At/Near	N/A	N/A	N/A	1	1	2	N/A	N/A
15	M	Multiracia	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10012	5534	1030L		2	No	At/Near	At/Near	At/Near	N/A	N/A	N/A	2	2	1	N/A	N/A
16	M	Black/Afri	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10013	5315	460L		1	No	Below	Below	Below	Insufficier	Insufficier	Insufficier	N/A	N/A	N/A	N/A	N/A
17	M	Black/Afri	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10014	5280	370L		1	No	Below	Below	Below	Insufficier	Insufficier	Insufficier	N/A	N/A	N/A	N/A	N/A
18	M	Black/Afri	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10015	5446	800L		1	No	Below	Below	Below	N/A	N/A	N/A	N/A	N/A	N/A	1	
19	M	White	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10016	5522	1000L		2	No	At/Near	At/Near	At/Near	2	2	1	N/A	N/A	N/A	N/A	N/A
20	M	White	N	N	N	6	Demo Sch	9999_1001	Demo Dist	10017	5566	1110L		3	Yes	At/Near	At/Near	At/Near	N/A	N/A	N/A	2	2	2	N/A	N/A

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 and Mathematics in Spring 2019), 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 instructions for these students to enhance their performance on the benchmarks for literary response and analysis.

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

Assessment results can be used to provide information on individual students' performance on the test. Overall, assessment results demonstrate what students know and 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 the precise measure for student performance. A student's scale score is associated with measurement error; users need to consider measurement error when using student scores to make decisions about student performance. Moreover, although student scores may be used to help make important decisions about students' placement and retention or teachers' instructional planning and implementation, the assessment results should not be used as the only source of information. Given that assessment results measured by a test provide limited information, other sources on student performance, such as classroom assessment and teacher evaluation, should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to take into account the group size. The smaller the group size, the larger the measurement error related to these aggregate data, thus requiring interpretation with more caution.

2.2 SCALE SCORE

A scale score is used to describe how well a student performed on a test and can be interpreted as an estimate of a 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 growth over time as a student or cohort advances through each grade level (see Section 6.2 in Volume 1 of this technical report for more information). Science and Social Studies scale scores are reported on separate within-test scales, and cross-grade comparisons are not appropriate.

Scale scores can be used to illustrate students' current levels of performance and are powerful when used to measure their growth over time. Lower scale scores can indicate that the student does not possess sufficient knowledge and skills measured by the test. Conversely, higher scale scores can indicate that the student has proficient knowledge and skills measured by the test. When combined across a student population, scale scores can also describe school and corporation-level changes in performance and reveal gaps in performance among different groups of students. In addition, scale scores can be averaged across groups of students, allowing educators to use group comparison. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and performance-level descriptors. It should be noted that the utility of scale scores is limited when comparing smaller differences among scores (or averaged group scores), particularly when the difference among scores is within the SEM. Furthermore, the scale

score of individual students should be cautiously interpreted when comparing two scale scores, because small differences in scores may not reflect real differences in performance.

2.3 Standard Error Measurement

A student's score is best interpreted when recognizing that the student's knowledge and skills fall within a score range and are not just precise numbers. A scale score (the observed score on any test) is an estimate of the true score. A test contains items that sample a student's knowledge and skills; if a student takes a similar test several times, the resulting scale scores would vary across administrations, sometimes being a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered several times. The SEM can be interpreted as the degree of uncertainty of a student's score based on a statistical analysis of the student's answers on a test. When interpreting scale scores, it is recommended to always consider the range of scale scores incorporating the SEM of the scale score.

2.4 Performance Level

Based on their scale score, a student will receive an overall performance level. 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.doe.in.gov/assessment/ilearn-sample-items-and-scoring.

2.5 Performance Category for Reporting Categories

Students' performance on each reporting category is reported on three performance categories: (1) Below Standard, (2) At/Near Standard, and (3) Above Standard. Students performing at Below Standard or Above Standard can be interpreted as student performances clearly below or above the Meets Standard cut score for a specific reporting category. Students performing at At/Near Standard can be interpreted as student performances that are close to the cut score, but there is not enough information to determine if it is above or below. Performance levels for the reporting category are limited in their diagnostic ability based on the degree of the calculated SEM of the student's scale score for the tested grade and subject.

2.6 CUT SCORES

For all grades and subjects within ILEARN, scale scores are mapped 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 shows the cut scores for ILEARN.

Table 7: ILEARN ELA Assessment Proficiency Cut Scores

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

Table 8: ILEARN Mathematics Assessment Proficiency Cut Scores

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency	Level 4 Above Proficiency
3	6080–6381	6382–6424	6425–6487	6488–6730
4	6100–6428	6429–6473	6474–6540	6541–6800
5	6110–6452	6453–6509	6510–6565	6566–6850
6	6110–6487	6488–6544	6545–6604	6605–6870
7	6120–6492	6493–6561	6562–6624	6625–6920
8	6120–6508	6509–6589	6590–6650	6651–6950

Table 9: ILEARN Science Assessment Proficiency Cut Scores

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

Table 10: ILEARN Social Studies Grade 5 Assessment Proficiency Cut Scores

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

Table 11: ILEARN U.S. Government Assessment Proficiency Cut Scores

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

2.7 AGGREGATED SCORES

Students' scale scores are aggregated at 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.

2.8 WRITING PERFORMANCE

ELA reports include descriptions of the student's performance on the writing portion based on the performance task writing rubric for each criterion. Essay responses are scored on three dimensions: Organization/Purpose, Evidence and Elaboration, and Conventions, as Table 12 shows. Each of these dimensions is independently scored and reported on the student reports. For item analysis Organization/Purpose and Evidence and Elaboration are averaged and rounded to an integer. Thus, the overall writing prompt score will range from 0 to 6.

A condition code is assigned to a student's written response that could not be scored, based on set criteria. Unscorable responses include responses that are blank, insufficient, written in a language other than English, off topic, illegible, or off-purpose. It should be noted that the reporting category score for writing consists of the overall writing score from the prompt and stand-alone writing items.

Table 12: Writing Scoring Dimensions

Dimension	Possible Scores
Organization/Purpose	1–4 points
Evidence and Elaboration	1–4 points
Conventions	0–2 points

2.9 Relative Strength and Weakness

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

2.10 Lexile® Measure

The Lexile® framework uses quantitative methods, based on individual words and sentence lengths, rather than qualitative analysis of content to produce scores. A Lexile® measure is defined as "the numeric representation of an individual's reading ability or a text's readability (or difficulty), followed by an 'L' (Lexile®)." A Lexile® text measure is obtained by evaluating the readability of a piece of text, such as a book or an article. A Lexile® measure of a text can assist in selecting targeted materials that present an appropriate level of challenge for a reader—not too difficult to be frustrating, yet difficult enough to challenge a reader and encourage reading growth.

2.11 Quantile® Measure

Quantile® measures provide an alternative—and possibly more useful—measure of Mathematics ability than grade-equivalent scores. Similar to the Lexile® framework, the Quantile® framework measures both the mathematics skill level of a student and the difficulty of Mathematics skills and concepts on the same developmental scale. Quantile® measures help educators, parents, and students determine which skills and concepts they are ready to learn next. Mathematics skills and concepts content, such as Mathematics

textbooks and online instructional materials, also get a Quantile® measure. Using these two measures together, parents and teachers can match students with resources that help them connect the dots among different Mathematics skills and concepts and build on their learning.

3. SUMMARY

ILEARN results are reported online via the Online Reporting System (ORS). The results are released after the testing window has closed and standard setting has been 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)

2018–2019

Volume 6
Recommending ILEARN Performance Standards

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: Stephan Ahadi, Elizabeth Ayers-Wright, Xiaoxin Wei, Kevin Clayton, and Kyra Bilenki. Major contributors from the Indiana Department of Education include the Assessment Director, Assistant Assessment Director, and Program Leads.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. INTRODUCTION

The Indiana Academic Standards (IAS) are designed to ensure that students across grades are receiving the instruction they need to be on track for college and career by the time they graduate. The IAS were approved by the Indiana State Board of Education (SBOE) in April 2014 for English/Language arts (ELA) and Mathematics and 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. In Spring 2018, the blueprints were updated with the goal of challenging and motivating Indiana's students to acquire stronger critical thinking, problem solving, and communication skills. In Spring 2019, the IDOE administered ILEARN assessments for the first time to assess proficiency on the IAS via the new blueprints. ILEARN measures ELA in grades 3–8, Mathematics in grades 3–8, Science in grades 4, 6, and Biology, and Social Studies in grade 5 and U.S. Government.

ILEARN is a series of computer-adaptive (CAT) and fixed-form assessments that are intended to be administered online, although the assessment is offered as a dual mode, online and paper, to accommodate testing needs for Indiana students. ELA and Mathematics students are administered as a series of CAT assessments in grades 3–8. Science students in grades 4, 6, and Biology were administered as a series of fixed-form assessments during 2018-2019. Social Studies students in grade 5 and U.S. Government are administered a series of fixed-form assessments.

The first operational administration of the ILEARN assessments took place in December 2018 for Biology and Spring 2019 for all other grades and subjects. Online administration of the ILEARN occurred from December 4–20 and February 11–28 for Biology; April 22–May 17 for ELA and Mathematics grades 3–8, Science grades 4 and 6, and Social Studies grade 5; and through May 24 for Biology and U.S. Government. The paper version of the ILEARN was administered from February 11–29 for Biology and April 22–May 10 for all grades and subjects. Following the close of the test administration windows, the American Institutes for Research (AIR), under contract to IDOE, convened nine panels of Indiana educators to recommend performance standards on the assessments. This document describes the procedures used to conduct the standard setting workshops as well as the recommended performance standards and resulting impacts.

The U.S. Government assessment was to be used as a final exam for the Spring 2019 semester, and as a proficiency indicator was needed prior to the testing window, a standard setting for U.S. Government was held in February 2019. The technical report for the U.S. Government standard setting can be found as a special study in Volume 7 of this technical report.

1.1 PERFORMANCE STANDARDS AND VALIDITY OF TEST SCORE INTERPRETATIONS

Validity refers to the degree to which test score interpretations are supported by evidence, and speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is the most fundamental component of test design and evaluation.

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this framework, the Standards describe the range of evidence that may be brought to bear to support the validity of test score interpretations. [1]

The kinds of evidence required to support the validity of test score interpretations depend centrally on the claims made for how test scores may be interpreted. Moreover, the standards make it explicitly clear that validity is not an attribute of tests, but rather test score interpretations. Some test score interpretations may be supported by validity evidence, while others are not. Thus, the assessment itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

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

The IAS specify what students should know and be able to do by the end of each grade level in order for graduating students to be ready for post-secondary education or entry into the workforce. Because directly measuring student achievement against each benchmark in the IAS would result in an impractically long assessment, each test administration is designed to measure a representative sample of the content domain defined by the Standards. To ensure that each student is assessed on the intended breadth and depth of the Standards, test form construction is guided by a set of test specifications and blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark. Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the IAS is evaluated, alignment of test blueprints with the content standards is critical. IDOE has published the ILEARN test blueprints that specify the distribution of items across reporting strands.

[1] Responsive to Standards for Education and Psychological Testing: Standard 9.13

Alignment of test content to the IAS[2] ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the IAS. However, the interpretation of the ILEARN test scores rests fundamentally on how test scores relate to performance standards which define the extent to which students have achieved the expectations defined in the IAS. ILEARN test scores are reported with respect to four proficiency levels, demarcating the degree to which ILEARN students have achieved the learning expectations defined by the IAS. The cut score establishing the At Proficiency level of performance is the most critical, since it indicates that students are meeting grade-level expectations for achievement of the IAS that they are prepared to benefit from instruction at the next grade level, and that they are on track for college and career readiness. Procedures used to adopt performance standards for the ILEARN assessments are therefore central to the validity of test score interpretations.

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

Panelists were also provided with contextual information to help inform their primarily content-driven cut-score recommendations. The decision to provide panelists with contextual benchmark information was discussed during a meeting with the Indiana State Board of Education Technical Advisory Committee (SBOE TAC) and confirmed by the policy committee. Panelists recommending performance standards for the ELA and Mathematics grades 3–8 assessments were provided with the approximate location of relevant National Assessment of Educational Progress (NAEP) and Smarter Balanced Assessment Consortium (Smarter) performance standards. Panelists recommending performance standards for the Science grades 4, 6, and Biology assessments were provided with the approximate location of relevant NAEP performance standards. Panelists recommending performance standards for the Social Studies grade 5 assessment were provided with the approximate location of relevant Smarter

[2] Responsive to Standards for Education and Psychological Testing: Standards 12.8 and 12.10

performance standards for grade 5 ELA. Panelists were asked to consider the location of these benchmark locations when making their content-based cut-score recommendations. When panelists are able to use benchmark information to locate performance standards that converge across assessment systems, validity of test score interpretations is bolstered.

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

Based on the recommended cut scores, Table 1 shows the estimated percentage of students meeting the ILEARN proficient standard for each assessment in Spring 2019. Table 1 also shows the national percentages of students that meet the NAEP and Smarter proficient standards. Since NAEP is only delivered in Grades 4 and 8, the percentages in other grades were interpolated or extrapolated so estimated percentages were available in all grades. As Table 1 indicates, the performance standards recommended for ILEARN assessments are consistent with relevant NAEP and Smarter proficient benchmarks. Moreover, because the performance standards were vertically articulated in ELA and Mathematics, the proficiency rates across grade levels are generally consistent.

Table 1: Estimated Percentage of Students Meeting ILEARN and Benchmark Proficient Standards

Grade	ILEARN At Proficiency	NAEP Proficient	Smarter Proficient
ELA 3	46	41	45
ELA 4	45	41	47
ELA 5	47	41	50
ELA 6	47	41	48
ELA 7	49	41	50
ELA 8	50	41	50
Mathematics 3	58	51	47
Mathematics 4	53	48	43
Mathematics 5	47	46	36
Mathematics 6	46	43	38
Mathematics 7	41	41	38
Mathematics 8	37	38	37
Science 4	46	42	--
Science 6	47	39	--

Grade	ILEARN At Proficiency	NAEP Proficient	Smarter Proficient
Biology	39	35	--
Social Studies 5	45	--	50

2. OVERVIEW OF STANDARD SETTING APPROACH

The Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) was used to recommend performance standards for ILEARN. The IDOE previously used the Bookmark method to recommend performance standards for the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) assessments. The Bookmark method was implemented in three rounds, providing panelists with feedback and benchmark information prior to Round 2, and panelist feedback, benchmark, and impact data prior to Round 3. To facilitate vertical articulation of performance standards across grades for ELA and Mathematics, workshop panelists began by recommending performance standards for "anchor" grades 4, 6, and 8, following standard Bookmark procedures. For the remaining "adjacent" grades, following a vertical moderation session to articulate performance standards across grades, panelists were provided with interpolated performance standards based on the recommended standards from the anchor grades. Panelists were instructed to use this interpolated page as another source of information to guide their judgment task.

Panelists were tasked with recommending three performance standards (Approaching Proficiency, At Proficiency, and Above Proficiency) that resulted in four performance levels (Below Proficiency, Approaching Proficiency, At Proficiency, and Above Proficiency).

2.1 WORKSHOP DESIGN

To recommend performance standards for each of the ILEARN assessments, IDOE convened nine panels: six panels representing three grade bands for ELA and Mathematics (3–4, 5–6, and 7–8), one panel representing two grades for Science (4 and 6), and two panels representing one grade each for Biology and Social Studies grade 5. The panels consisted of educators and administrators from around Indiana. The panelists recommended performance standards based primarily on content considerations with additional context provided by relevant benchmark information from Smarter and NAEP exams, as well as estimated student performance on the recommended standards prior to Round 3. Panelists used ordered-item booklets (OIBs) and performance-level descriptors (PLDs) to place performance standards for the three performance cut scores, Approaching Proficiency, At Proficiency, and Above Proficiency, in three rounds. First, panelists recommended performance standards for the initial grades: 4, 6, and 8 for ELA and Mathematics, grade 4 and Biology for Science, and grade 5 for Social Studies. After recommending performance standards for the initial grades, a moderation session was conducted for ELA and Mathematics with the table leaders from each of the panels to review the vertical articulation of the performance standards, and to implement any adjustments to the anchor grade recommendations to facilitate vertical articulation. Following the vertical articulation session, panelists continued on to recommend performance standards for the remaining grade-level assessments, using the interpolated standards for ELA and Mathematics to provide further contextual information about the potential neighborhood of performance standards. Following the adjacent grade work, a final moderation session was conducted for ELA and Mathematics. Science panelists

continued on to recommend performance standards for grade 6 without any moderation or interpolation.

The ILEARN standard setting workshops were conducted over three days. A broad overview of the workshop calendar is presented in Table 2. Detailed agendas for the standard setting workshops are included in Appendix A, Workshop Agendas.

Table 2: Overview of Workshop Calendar

Workshop	Monday, July 15	Tuesday, July 16	Wednesday, July 17
ELA and Mathematics Grades 3–8 Science Grades 4 and 6	Standard Setting Day 1	Standard Setting Day 2	Standard Setting Day 3
Biology Social Studies Grade 5	Standard Setting Day 1	Standard Setting Day 2	n/a

A virtual table leader orientation to review with table leaders their roles and responsibilities was held one week prior to the workshop. The workshop began with a large group training to provide panelists with an overview of the workshop activities and initial training in the bookmarking procedures. Following the large group session, the workshop panels convened in their meeting rooms and began their work by participating in the same ILEARN online assessments that were administered to their students in the Spring. Panelists then spent several hours working through the PLDs developed by IDOE with guidance from their policy committee of educators, and developing modified descriptors to characterize the special subset of students who "just barely" qualify for entry into each of the performance levels. After developing descriptors for the "just barely" students, panelists spent the remainder of Day 1 reviewing their OIBs.

Panelists did not begin recommending performance standards until Day 2, which began with training on the bookmark placement task. Panelists then worked through their OIBs and placed their bookmarks for Round 1. After Round 1, panelists were provided feedback about the bookmark placements of the other panelists and discussed those bookmark placements at their tables and across the room more generally. Upon completion of panel discussions, panelists were provided with benchmark information prior to making a second round of bookmark placements. After Round 2, panelists were again provided feedback about the bookmark placements of the other panelists and discussed those bookmark placements at their tables and across the room. Upon completion of panel discussions, panelists were provided with impact information prior to making a final third round of bookmark placements. After Round 3, panelists began the process over again for the subsequent assessments.

2.2 WORKSHOP LOCATION

The workshops were held at:
Sheraton at Keystone Crossing
8787 Keystone Crossing
Indianapolis, Indiana 46240

The location provided meeting spaces to hold the ILEARN workshop panels, as well as a psychometric work room for completion of analysis activities and storage space for secure materials throughout the workshop.

2.3 WORKSHOP STAFFING

A senior workshop coordinator was tasked with leading the cross-workshop introductory training and vertical moderation meetings, working with each facilitator, and monitoring the flow of activities across workshops. American Institutes for Research (AIR) test development staff served as workshop facilitators, leading each panel through training activities and execution of the standard setting process. Additionally, an AIR test development staffer was assigned to each panel to support the workshop facilitator. Because test development staff served as workshop facilitators, they were highly qualified to facilitate the development of "just barely" PLDs and to serve as subject matter resources for panelists as they navigated the OIBs. A team of three AIR psychometricians managed psychometric activities in support of the workshop, including ensuring accurate data capture of bookmark placements, presentation of vertical articulation results for moderation meetings, and production of final results for the standard setting technical report. In addition, AIR project staff facilitated organization of meeting space and meals and provided support to panelists as necessary.

IDOE staff monitored all standard setting activities and also addressed any policy or test development questions from panelists. While IDOE staff answered specific, direct questions, they were not actively involved in the facilitation of the meeting.

2.4 WORKSHOP PANELISTS

IDOE worked to obtain broadly representative panels for the standard setting workshops that reflected the teacher population in the state of Indiana in terms of gender, race, ethnicity, and geographical representation. Diverse groups of panelists bring a wide range of perspectives and experience to the standard setting effort, ensuring that the recommendations that are forwarded to the SBOE are thoughtful and representative of broad educational constituencies, and represent the range of expertise and experiences found in the educator population across the state.

Within each of the panels, a total of 12 panelists per grade-band subpanel were recruited to recommend standards. IDOE targeted the number of male and female panelists to mirror the population of educators. In the same way, IDOE worked to include proportional representation of American Indian/Native American, Asian/Pacific Islander, Black (Non-Hispanic), Hispanic, and White (Non-Hispanic) panelists, and a proportional number of panelists from rural, urban, and suburban corporations.

Within each subpanel, tables were balanced to include panelists with varying content expertise and demographic representation in each group.

IDOE designated three table leaders for each panel. Table leaders attended an additional orientation meeting conducted via webinar and were tasked with assisting standard setting staff by

- facilitating discussions within their table;
- distributing and collecting readiness and recording sheets and secure materials;
- alerting workshop staff of confusion or concerns within their tables; and
- representing their tables and panels during vertical articulation meetings.

Letters containing logistical information and reminders about the purpose[3] of the workshop were emailed to confirmed panelists two weeks prior to the standard setting workshop. In the week prior, testing contractor staff contacted all panelists via phone to confirm receipt of information. Throughout the process, IDOE continued to recruit replacements for panelists who withdrew their participation.

Appendix B[4], Composition of Panels, presents the composition of the standard setting panels. For each panel, the table includes a record for each panelist and indicates the geographic region he or she represents and his or her gender, ethnicity, and main expertise. While it is critically important to include a range of stakeholders in the standard setting process, experience has shown that it is essential for panelists to have direct knowledge of academic standards and student grade-level performance to participate meaningfully in the bookmarking procedure. For this reason, panel recruitment was focused on classroom teachers and curriculum specialists with expertise in ELA, Mathematics, Science, and Social Studies curriculum and instruction.

2.5 WORKSHOP TRAINING

Thorough training is an essential element of a standard setting workshop. Training at the meetings helped panelists become familiar with the assessment system and the standard setting process. It also involved a review and discussion of the assessments, the student populations that participated in each assessment, and the PLDs. In addition, training included an in-depth discussion of concepts key to bookmark placement, such as the notion of what would constitute a student “just barely” in a performance level. All panelists were administered an operational assessment in order to understand the test content, the

[3] Responsive to Standards for Education and Psychological Testing: Standard 5.0, 5.21, 5.22, and 7.0

[4] Responsive to Standards for Education and Psychological Testing: Standard 7.5

testing interface, and various item types through which student knowledge and skills were assessed. A sample of the presentation slides used to conduct the introductory training, and those used to facilitate each workshop are provided in Appendix C[5], Workshop Slides.

To begin the workshop, the panelists were convened for a brief introductory training that focused on the purpose of the standard setting workshop and a review of the main workshop activities. Following this large group introduction, panelists joined their assigned workshop panels where the workshop leader for each assessment guided panelists through the standard setting activities and provided in-depth training throughout the course of the workshop.

Table leaders had the additional responsibilities of ensuring that table activities remained focused on the task at hand, helping to verify that panelists understood their tasks, and alerting workshop leaders to any issues encountered by panelists as they engaged in their workshop tasks. Table leaders were not expected to provide training to panelists, but rather serve as liaisons between the panelists and workshop leaders to ensure that workshop activities were implemented correctly, alerting workshop leaders to any issues that arose during the course of conducting workshop activities, and representing their tables in the cross-panel moderation deliberations. A table leader orientation meeting was convened prior to the standard setting workshop to familiarize table leaders with their roles and responsibilities, including suggestions on how to provide leadership at the tables during the standard setting process and how to manage the secure materials.

2.6 ONLINE STANDARD SETTING TOOL

During the standard setting meeting panelists used AIR's online standard setting tool to view and interact with items, set their bookmarks, and view feedback. Each panelist was provided with a laptop, and a secure connection to the online standard setting tool. Following the large group orientation, panelists received training on the standard setting tool. All AIR facilitators and room assistants were trained on the standard setting tool prior to the meeting and able to answer any questions that arose.

The online standard setting tool automates and standardizes all workshop activities. All steps in the standard setting workshop were assembled in the online tool and each step was configured according to IDOE specifications. For example, while all Bookmark

[5] Responsive to Standards for Education and Psychological Testing: Standard 7.5

workshops include a review of the OIB, IDOE determined what item information was provided to the panelists, including item metadata, item difficulty, associated performance level in benchmark assessments, and even impact data; as well as when such information is made available.

3. STANDARD SETTING MATERIALS AND PROCEDURES

3.1 PERFORMANCE-LEVEL DESCRIPTORS

PLDs define the content area knowledge and skills that students at each performance level are expected to demonstrate. In particular, Policy PLDs articulate the overall claims about a student's performance in each performance level. Range PLDs are key elements in standard setting processes. Range PLDs define the content area knowledge, skills, and processes that test takers at a particular performance level are expected to possess. The standard setting panelists based their judgments about the location of the performance standards on the PLDs, as well as the IAS.

Indiana's policy group is made up of a member from the SBOE, a member from higher education, administrators at the high school and grade 3–8 levels, special education administrators and leaders, and the IDOE leadership. The policy group created the Policy PLDs in May 2018. The Range PLDs were drafted by educators in a meeting held June 18–21, 2018. Policy and Range PLDs are presented in Appendix D, Range Performance-Level Descriptors (PLDs).

Central to their training in the Bookmark method, panelists used the PLDs to develop a representation of students who are "just barely" described by each of the PLDs. During this training task, panelists learned that while PLDs are written to characterize typical members of each performance level, their bookmark placements would be directed toward characterizing and identifying the most minimally qualified members of each performance level. Characterizing a student as "just barely" meeting the performance standard is not an intuitive judgment, and panelists worked to identify the minimum characteristics of student achievement for entry into each performance level. Each panel produced a "just barely" PLD to help guide their discussions and bookmark placements. To develop a common understanding among panelists, each panel was asked to

- Review and parse PLDs
- Discuss characteristics of students classified near thresholds of performance standards
- Identify the characteristics that distinguish students "just above" the performance standard from those "just below"
- Determine what evidence was necessary to conclude that a student possessed the minimum knowledge and skills needed to meet the performance standard
- Summarize knowledge and skills of students who "just barely" meet each performance standard, or are "just barely" described by each PLD

These discussions yielded common descriptions of students "just barely" characterized by each PLD within each room.

3.2 ORDERED-ITEM BOOKLETS

Following review of PLDs and development of "just barely" PLDs, panelists reviewed the OIBs. An OIB is a collection of assessment items ordered from easiest to most difficult. Each page in the OIB corresponds to a level of achievement on the ILEARN assessments, and panelists use the OIB to recommend the minimum level of achievement required to enter each performance level.

Items were ordered according to their response probability (RP) level based on their Item Response Theory (IRT) parameters. In IRT, the item characteristic curve for each item indicates the likelihood of responding correctly for each point along the student achievement scale. The response probability criterion refers to the location on the achievement scale that corresponds to a given probability of success. In context of the standard setting workshop, this criterion is used to develop a common understanding of what constitutes mastery when evaluating whether a student can respond successfully to an item. An RP value of 0.67 was used as the mastery criterion for all of the standard setting workshops. Panelists were asked to consider whether, for example, a "just barely" proficient student had a 0.67 likelihood of answering the item correctly. They were also encouraged to ask this question in other related ways, including whether two-thirds of "just barely" proficient students would answer the item correctly, or whether a "just barely" proficient student would respond correctly to item two of three times.

Dichotomously scored (e.g., incorrect vs. correct) ILEARN ELA, Mathematics, and grade 5 Social Studies items were calibrated using the two-parameter logistic model (2PL). Multi-point, partial credit items were calibrated using the generalized partial credit model (GPC) with ordering of score point pages in the OIB based on step-level difficulties. Science stand-alone items were calibrated using the Rasch model and performance tasks were calibrated using the Rasch Testlet Model.

The OIBs were augmented with items to minimize any gaps in the test information in critical regions. Increasing the number of items across the range of item difficulties provides panelists with greater context to identify important shifts in the knowledge and skill requirements of assessment items. Often panelists become focused on the cognitive demands of a single item when deliberating on the location of a performance standard. This propensity is exacerbated when there are relatively few items in a given location, which can cause judgment about one item to take on too much importance. Even when there are sufficient items to establish reliable performance standards for a central proficient performance standard, there are typically fewer items available in locations associated with performance standards categorizing achievement Below and Above Proficient; thus, movement of the bookmark by even a page or two may result in very large increases or decreases in the percentage of students meeting the standard. Augmenting the OIB moderates the impact associated with each OIB page, especially for performance standards in the tails of the ability distribution (Cizek & Bunch, 2006).

3.2.1 Composition of OIBs

Within each ELA and Mathematics assessment, online test takers were administered a unique test form meeting the test blueprint and determined by the CAT algorithm. A fixed-form operational test form was also administered both online and on paper, for accommodated populations. Within each Science and Social Studies assessment, all online test takers were administered a test form with a common set of items used for operational scoring, as well as a set of embedded items used for linking or field testing. The operational test form was also administered on paper with item substitutions for a few technology-enhanced items that could not be represented on paper. For all subjects, the operational items administered online served as the basis for the OIBs.

For ELA and Mathematics OIBs, the set of operational items administered to a student at the mean of the student population was pulled and this served as the baseline OIB. To minimize gaps in test information, the ELA and Mathematics OIBs were augmented by additional operational items not on the baseline OIB. Each ELA OIB was augmented with 10–15 operational items and each Mathematics OIB was augmented with 12–18 operational items.

For Science and Social Studies, the online fixed form was used as the baseline OIB. To minimize gaps in test information, the Science and Social Studies OIBs were augmented by additional operational not on the online fixed form (e.g., items operational only on paper forms) and field-test items. Each Science OIB was initially constructed with 47–55 operational items and was augmented with an additional 4–19 field-test items, and the Social Studies grade 5 OIB was augmented with 17 field-test items.

All items selected for inclusion in the OIB were reviewed for statistical integrity. It is important to note that each OIB was initially constructed with respect to the assessment blueprint, which specifies the composition of each assessment with respect to the range of content assessed by each operational form. The augmented OIBs were as proportional to the operational test blueprints as possible; the blueprints are presented in Appendix E[6] Test Blueprints. In some instances, it was necessary to over-emphasize one or more standards in order to appropriately minimize gaps in the OIB.

The OIBs were presented online, allowing panelists to view items in the same context as student test takers. The composition of the OIBs by assessment and grade are

[6] Responsive to Standards for Education and Psychological Testing: Standards 7.1 and 12.4

summarized in Table 3 below. A technical summary of the OIBs are presented in Appendix F, Ordered Item Booklets, including for each page in the OIB, the item score point associated with the presented item, the difficulty represented by the page, and the standard error of the difficulty. In addition, the appendix indicates the overall percent of students who would score at or above the standard associated with each OIB page, and the location of external benchmarks within the booklet.

Table 3: The Composition of the Ordered-Item Booklets

Grade	Number of Items in OIB			Pages in OIB (Total Points)
	Operational	Field Test	Total	
ELA 3	42	--	42	53
ELA 4	46	--	46	54
ELA 5	43	--	43	53
ELA 6	43	--	43	55
ELA 7	48	--	48	60
ELA 8	46	--	46	55
Mathematics 3	59	--	59	63
Mathematics 4	59	--	59	62
Mathematics 5	54	--	54	60
Mathematics 6	59	--	59	63
Mathematics 7	57	--	57	60
Mathematics 8	58	--	58	66
Science 4	29	17	46	53
Science 6	41	4	45	57
Biology	34	19	53	62
Social Studies 5	39	17	56	60

3.2.2 Review of OIBs

For each item in the OIB, panelists were instructed to ask what a student must know and be able to do to answer each question and what makes each item in the OIB more difficult than the preceding item. This review of the OIB allowed panelists to gain new perspectives on the knowledge and skill requirements of items and to share information regarding their thoughts on the location of the threshold region. During this discussion, the workshop leader circulated through the room to monitor progress, assisted panelists who might have had trouble with the task, and answered any questions.

On each page in the OIB, panelists viewed the content of the item, the associated passage, content alignment, and the scoring key or rubric. In addition, for each page that presented a writing item, ELA panelists were provided a sample student essay response that scored at the particular score point.

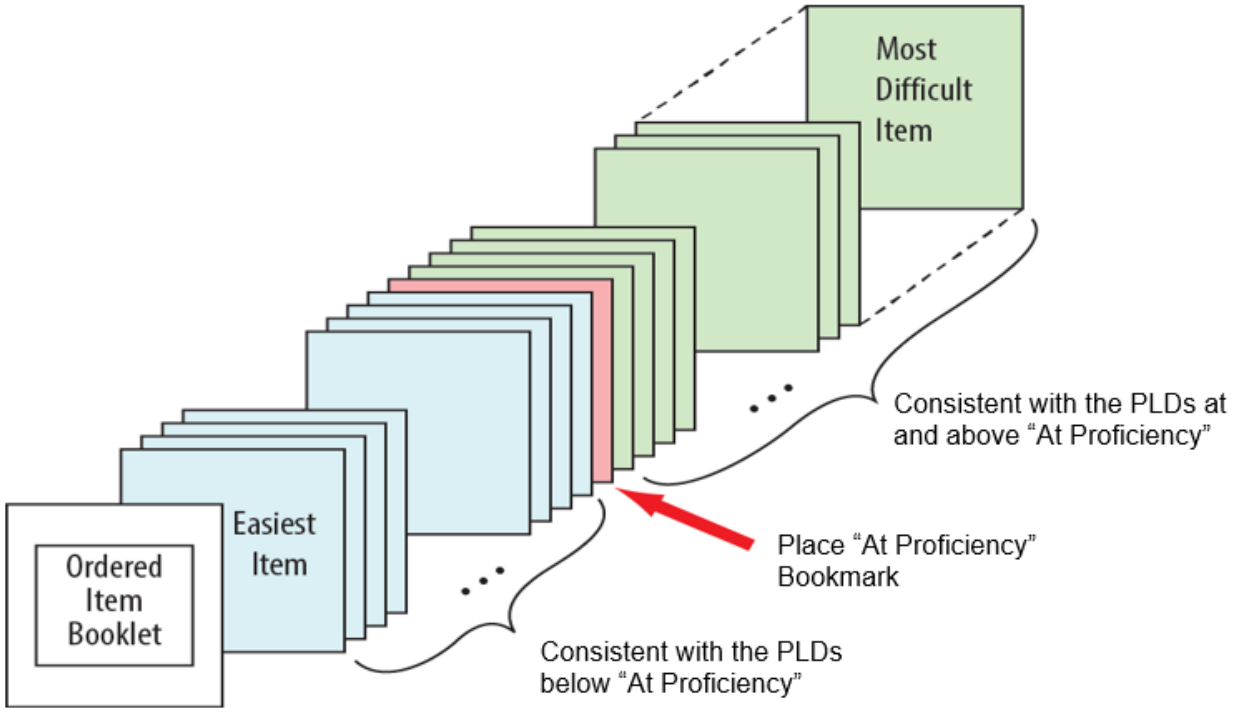
3.3 ILEARN BOOKMARK PLACEMENT

Prior to making their Round 1 bookmark placements, panelists were provided training in the identification of performance standards in the OIBs. As part of this training, panelists learned to identify a location in the OIB that best delineates two performance levels (i.e., between pages on which students must demonstrate mastery to meet the minimum requirements for inclusion in the Approaching Proficiency level from those items on which demonstration of mastery is not necessary).

Using their "just barely" PLDs as a guide, the panelists were then instructed to set a bookmark on the item that best delineated each of the performance levels. Panelists were reminded how to set bookmarks, and prior to making initial placements, facilitators led a group activity that reviewed the key concepts of the bookmark procedure, allowing facilitators to provide additional training if necessary. Prior to placing recommended performance standards in each round, panelists were asked to complete a readiness form to indicate their preparedness to recommend performance standards. This form asked panelists to assert their understanding of the tools used to recommend performance standards in each round. If a panelist indicated that they do not feel prepared to recommend performance standards, the workshop leader provided additional training and opportunities for discussion. All panelists had to indicate that they felt prepared to move forward before they recommended a cut. All ILEARN standard setting panelists indicated they understood the task at hand and felt ready to recommend performance standards. Samples of readiness forms used for completing the bookmark task are presented in Appendix G, Readiness Forms.

Bookmark placement was conducted in three rounds, allowing panelists to make independent judgments while still benefiting from discussion with their fellow panelists. Panelists were instructed to identify their recommended cut scores for At Proficiency, Approaching Proficiency, and Above Proficiency in each round. The placement of the bookmark is illustrated in Figure 1. Each panelist used their "just barely" PLDs to identify which item represented the lower bound of each performance level. In the example, a panelist concluded that students who were "just barely" at the At Proficiency level would demonstrate mastery on the item on the page indicated by the arrow, while students below the At Proficiency level would not. Therefore, the panelist decided that the At Proficiency performance level would begin on the page indicated by an arrow. The panelist believed that students below the At Proficiency performance level would not be able to demonstrate mastery of items beyond the indicated page in the OIB.

Figure 1: Example of Bookmark Placement



3.4 Panelist Feedback

Prior to Round 2, panelists were provided feedback about the bookmark placements made by fellow panelists. After making their Round 1 bookmark placements, panelists reconvened and began with a discussion of panelist feedback about the bookmark locations recommended by each panelist, beginning with table-level feedback and discussion, and progressing to room-level discussion. Each table spent time reviewing and discussing cut-score placements, focusing on the lowest and highest recommended performance standards both at the table and across the panel. Panelists were asked to review the items between the lowest and highest performance standards at their table, discussing the standards and the "just barely" PLDs. Discussion was then expanded to the room level, with each table reviewing the basis for their own recommendations for the group at large.

3.5 Benchmark Information

Following discussion of panelist feedback, panelists were presented with benchmark data, performance standards comparable to other important assessment systems, including national and international benchmarks such as NAEP and Smarter. To facilitate comparisons of Indiana performance standards with other national and international benchmarks, panelists were provided with the locations of performance standards from these other assessment systems in their OIBs before beginning Round 2. In particular,

performance standard locations for the following assessments were provided as part of panelists' OIB review:

- Smarter ELA and Mathematics performance standards in grades 3–8; Social Studies grade 5 used the performance standard cut from ELA grade 5
- NAEP performance standards in ELA and Mathematics in grades 3–8 and Science in grades 4, 6, and Biology

3.6 IMPACT DATA

Prior to Round 3, panelists were again provided feedback about the bookmark placements made by fellow panelists. After making their Round 2 bookmark placements, panelists reconvened and began with a discussion of panelist feedback about the bookmark locations recommended by each panelist, beginning with table-level feedback and discussion, then progressing to room-level discussion.

Following discussion of panelist feedback, panelists were presented with impact data, which detailed the percentage of students expected to score at or above the recommended Round 2 performance standards. Panelists discussed any implications of the impact data, both at their tables and across the panel more generally, focusing on whether the impact was in line with their expectations. Following the presentation of impact data, panelists were provided, for each item in the OIB, the percentage of students expected to achieve the ability level indexed by that page.

After completing their discussions, panelists again worked through the OIB, placing their Round 3 bookmarks for all three performance levels, beginning with At Proficiency and followed by Approaching Proficiency and Above Proficiency.

3.6.1 Estimating Student Performance Data

While the ILEARN OIBs were constructed based on calibration of the online testing population, the percentage of students within the state who meet or exceed each potential performance standard (i.e., each page in the OIB) was estimated based on all students participating in the first operational administration of the assessment, including students who tested online and students who tested on paper. For Biology, only first time testers from the spring administration were included.

Prior to Round 3 of the Bookmark procedure, the percentage of students meeting the standards, based on the Round 2 median cut score, was presented to panelists.

3.7 VERTICAL ARTICULATION

Performance standards should ideally be well-articulated across grades. Unless there are systemic differences in the quality of instruction across grades, the expectation is that students who meet the standards and are prepared for instruction in the subsequent grade will likely continue to meet standards as they progress through their school years, and therefore we would not expect to see large changes in the proficiency rates from

grade to grade. While this vertical articulation is incorporated into the development of the IAS, as well as the test specifications for each of the ILEARN assessments, maintaining and reinforcing the cross-grade articulation in the setting of meaningful performance standards is important, especially for ELA and Mathematics, where students are assessed annually. Lack of articulation in these subjects can result in confusion, such as when there are unreasonably large shifts in student performance-level classifications from grade to grade.

Articulation was considered from two perspectives: (1) the percentage of students meeting standards across grades and courses, and (2) the location of the performance standards on the vertically-linked ILEARN scale, which allowed panelists to evaluate their recommended performance standards with respect to expected student growth from grade to grade.

To help foster consistency in the identification of performance standards across grades, after performance standards were recommended for the initial grade level in each grade band, ELA and Mathematics table leaders convened to participate in a vertical moderation session. Table leaders were shown the percentage of students scoring at or above each of the performance standards, and the percentage of students classified at each performance level across tests. Where the percentage of students expected to meet standards varied greatly between grade- or course-based assessments, table leaders were asked to consider modifications to the recommended standards that would achieve a more articulated system. In these instances, table leaders reviewed the OIBs and considered whether their content supported the adjustment. Thus, while table leaders worked to articulate standards across grades, they also ensured that any changes resulting from the moderation meeting would be consistent with the knowledge and skills described in the PLDs.

With anchor grade performance standards in hand, the AIR evaluated both the impact data from each grade-level assessment, as well as student ability estimates from the vertically-linked ILEARN scale, to interpolate the likely location of each performance standard for each of the remaining grade-level assessments.

To recommend performance standards in these adjacent grade assessments, the standard bookmark procedures were modified so that panelists were instructed to determine whether the “just barely” PLDs supported the placement of a specific bookmark on the interpolated page. If the PLDs did not support the placement of the bookmark on the interpolated page, then panelists were asked whether they could identify a bookmark placement near the interpolated page that would be supported by the PLDs. Panelists were instructed that their bookmark placements must be guided by content considerations, which may recommend the bookmark be placed on the interpolated page in the OIB or a different location. Otherwise, bookmark placements proceeded as with the anchor grade rounds. Following Round 1 bookmark placements, panelists received feedback about the bookmark placements of panelists at their table and for the room as a whole, and were presented with benchmark data. Following Round 2 bookmark placements, panelists again received feedback about the bookmark placements of

panelists at their table and for the room as a whole, and were then presented with impact data.

A final moderation session was conducted following the completion of workshop activities for the interpolated grades. This final moderation activity ensured that table leaders had an opportunity to review the entire system of recommended standards and to make any desired adjustments prior to completion of the workshop. As with the initial moderation session, in those instances where table leaders chose to adjust a performance standard during the final moderation session, they reviewed their OIBs to ensure that the adjustments had a basis in test content.

The advantage of this approach is that it results in a system of performance standards that are more consistent across grade levels. At the most basic level, it ensures that there are no wide fluctuations in the proportion of students meeting each performance standard across grades. Cross-grade articulation informed by the vertical scale also ensures that there are no reversals in recommended performance standards across grades.

3.8 WORKSHOP EVALUATION

Throughout the process, panelists were encouraged to provide feedback concerning the standard setting workshop procedures and outcomes via group discussions, practice activities, and completion of readiness forms prior to placing their bookmarks.

At the end of each day, panelists were asked to complete a workshop evaluation form designed to elicit feedback on all aspects of the workshop, including clarity of training and tasks, appropriateness of the time spent on activities, and satisfaction with the outcome of the workshop. This feedback can be found in Appendix K, Summary of Panelist Evaluations.

4. RECOMMENDED PERFORMANCE STANDARDS AND IMPACT DATA

For the ILEARN in ELA, Mathematics, Science, and Social Studies, Appendix H, Recommended Cut Scores by Round, presents the minimum, maximum, and median bookmark placement for each round of bookmark placements, as well as any bookmarks placed during Moderation sessions, and the resulting final recommendations following the standard setting workshops. As panelists discussed the reasons for their bookmark placements in the context of feedback from other panelists and impact data, variability across tables often decreased across rounds. The figures in Appendix I, Workshop Agendas, illustrate variability in median table bookmark placements for the three performance standards over the three rounds. These figures illustrate how variability in bookmark decisions changed from the first round to the second round, and from the second round to the third round. In general, there was considerable consistency in the placement of performance standards across rounds.

For each assessment, the final recommended performance standard is the outcome from the final moderation, or in the absence of moderation, the median bookmark page following Round 3.

The final recommended performance standards for each assessment, grade, and performance standard are presented in Table 4, along with the projected impact each performance standard would have on Indiana public school students tested in 2019. The final recommended OIB page numbers are the median bookmarks of each panel following Round 3 bookmark placement, and subsequent moderation.

Table 4: Final Recommended Performance Standards

Grade	Performance Level	OIB Page	RP67	Estimated Percentage of Students At or Above Performance Standard
ELA 3	Approaching Proficiency	9	-1.12	69%
	At Proficiency	25	-0.54	46%
	Above Proficiency	43	0.20	18%
ELA 4	Approaching Proficiency	8	-0.75	69%
	At Proficiency	24	-0.10	45%
	Above Proficiency	45	0.63	19%
ELA 5	Approaching Proficiency	9	-0.37	71%
	At Proficiency	26	0.32	47%
	Above Proficiency	44	1.26	15%
ELA 6	Approaching Proficiency	7	-0.11	73%
	At Proficiency	21	0.59	47%
	Above Proficiency	41	1.38	17%

Grade	Performance Level	OIB Page	RP67	Estimated Percentage of Students At or Above Performance Standard
ELA 7	Approaching Proficiency	5	0.09	75%
	At Proficiency	24	0.90	49%
	Above Proficiency	43	1.72	20%
ELA 8	Approaching Proficiency	6	0.15	79%
	At Proficiency	21	1.03	50%
	Above Proficiency	44	1.85	21%
Mathematics 3	Approaching Proficiency	7	-1.57	76%
	At Proficiency	17	-0.99	58%
	Above Proficiency	47	-0.16	25%
Mathematics 4	Approaching Proficiency	9	-0.95	74%
	At Proficiency	22	-0.35	53%
	Above Proficiency	49	0.54	21%
Mathematics 5	Approaching Proficiency	7	-0.62	72%
	At Proficiency	23	0.14	47%
	Above Proficiency	47	0.88	22%
Mathematics 6	Approaching Proficiency	8	-0.16	70%
	At Proficiency	23	0.59	46%
	Above Proficiency	47	1.39	20%
Mathematics 7	Approaching Proficiency	10	-0.10	68%
	At Proficiency	28	0.83	41%
	Above Proficiency	43	1.67	18%
Mathematics 8	Approaching Proficiency	12	0.13	65%
	At Proficiency	29	1.20	37%
	Above Proficiency	48	2.01	18%
Science 4	Approaching Proficiency	12	-0.36	65%
	At Proficiency	24	0.12	46%
	Above Proficiency	40	0.69	24%
Science 6	Approaching Proficiency	12	-0.68	73%
	At Proficiency	26	0.08	47%
	Above Proficiency	46	0.89	19%
Biology	Approaching Proficiency	12	-0.43	63%
	At Proficiency	28	0.18	39%
	Above Proficiency	47	0.93	17%

Grade	Performance Level	OIB Page	RP67	Estimated Percentage of Students At or Above Performance Standard
Social Studies 5	Approaching Proficiency	8	-0.46	63%
	At Proficiency	18	0.04	45%
	Above Proficiency	42	0.87	21%

Table 5 shows the estimated percentage of student classified at each performance level based on final panelist-recommended standards for the overall student population across grade levels and courses. The results of Table 5 are represented graphically in Figure 2 through Figure 5. Appendix J, Recommended Cut Scores by Round, presents the estimated percentage of students classified at each performance level disaggregated by gender and ethnicity.

Table 5: Percentage of Students at Each Performance Level Based on Final Recommended Performance Standards

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
ELA 3	31	23	28	18
ELA 4	31	24	26	19
ELA 5	29	24	31	15
ELA 6	27	26	29	17
ELA 7	25	26	29	20
ELA 8	21	29	29	21
Mathematics 3	24	19	32	25
Mathematics 4	26	21	33	21
Mathematics 5	28	25	25	22
Mathematics 6	30	24	26	20
Mathematics 7	32	27	23	18
Mathematics 8	35	28	19	18
Science 4	35	19	22	24
Science 6	27	25	28	19
Biology	37	24	22	17
Social Studies 5	37	18	24	21

Figure 2: Percentage of Students at Each Performance Level Based on Final Recommended Performance Standards — ELA

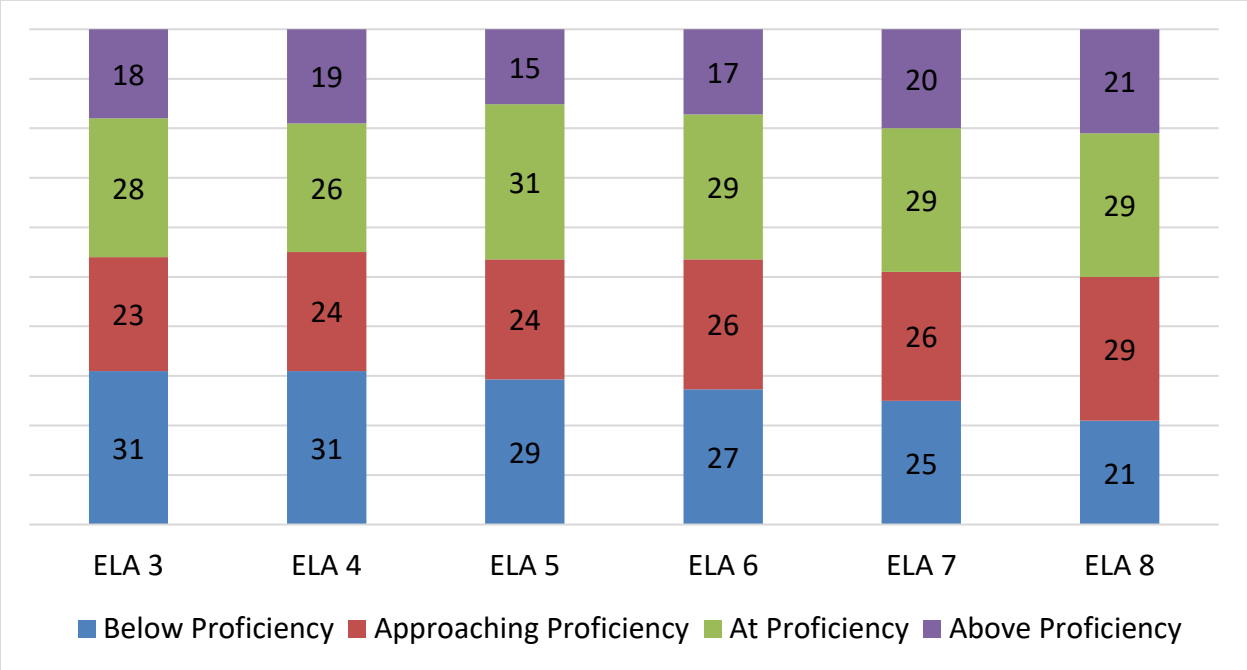


Figure 3: Percentage of Students at Each Performance Level Based on Final Recommended Performance Standards — Mathematics

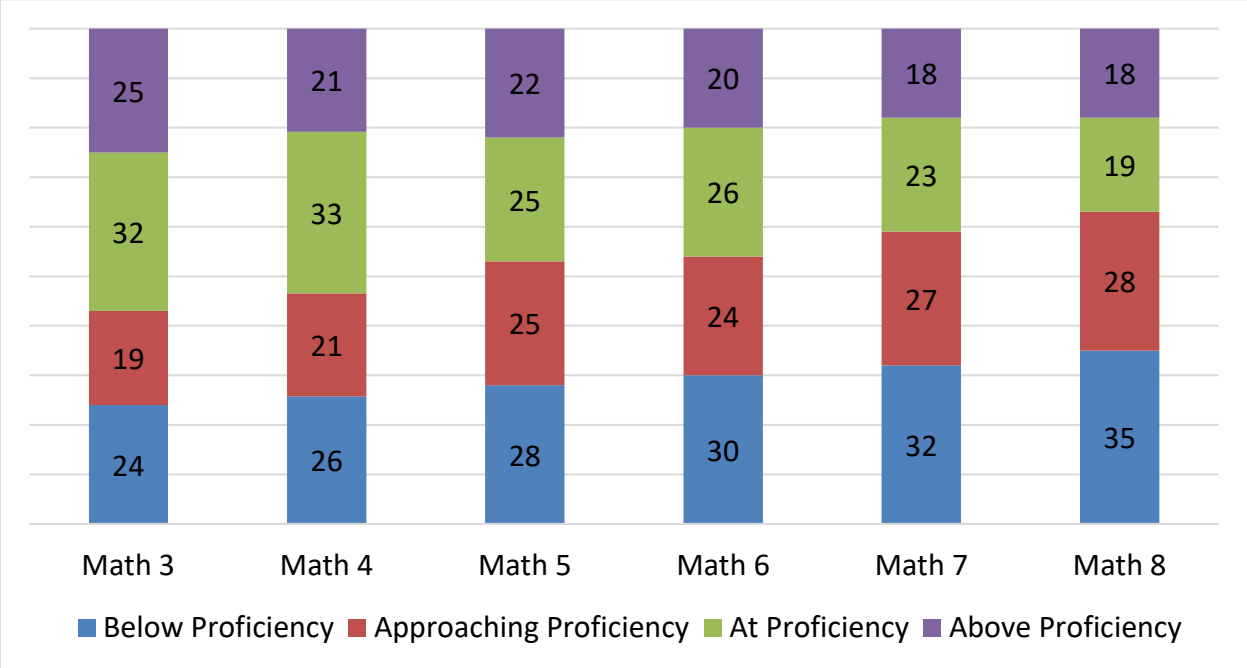


Figure 4: Percentage of Students at Each Performance Level Based on Final Recommended Performance Standards — Science

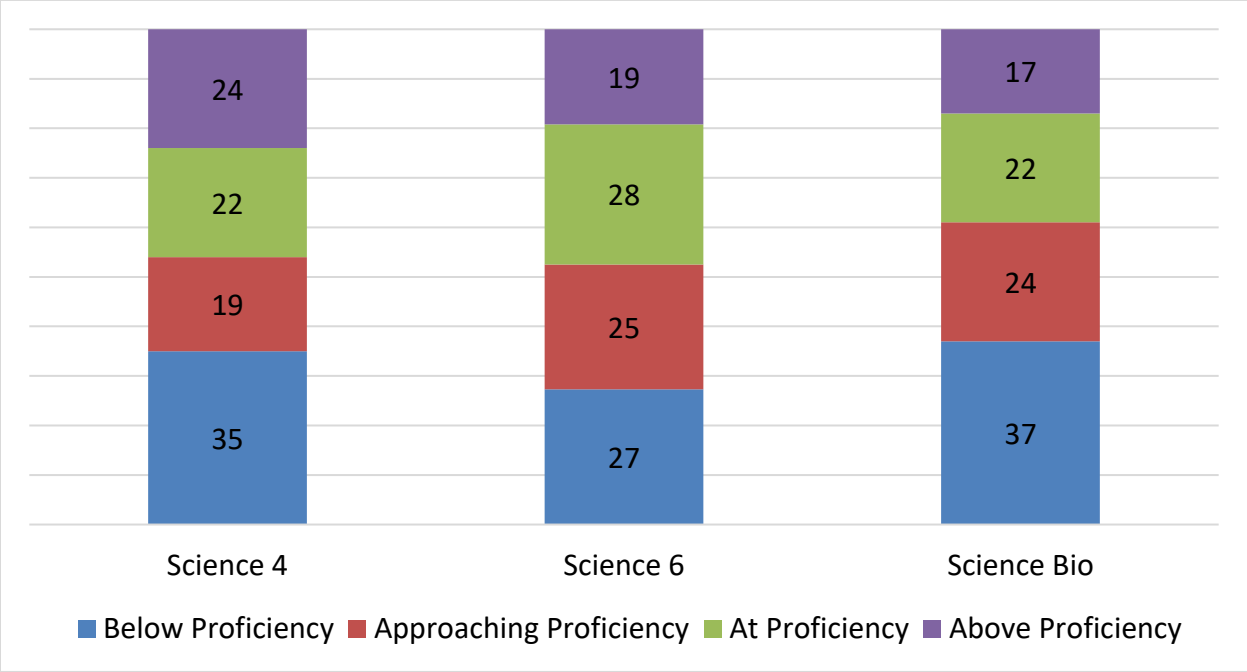
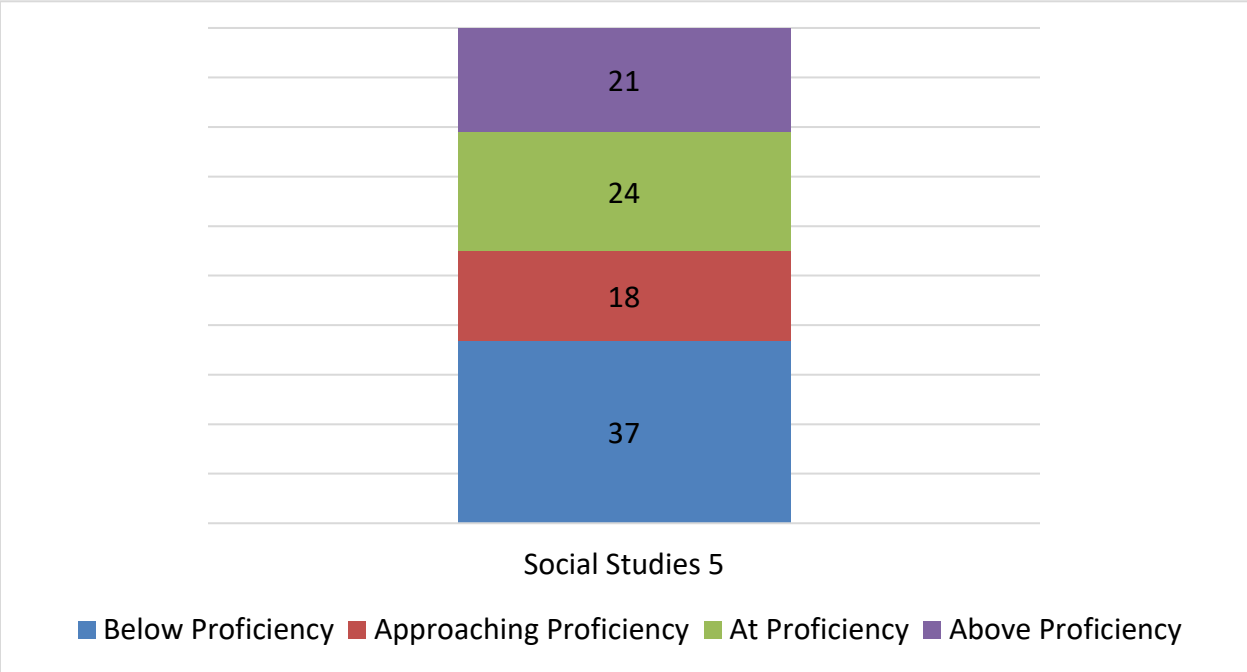


Figure 5: Percentage of Students at Each Performance Level Based on Final Recommended Performance Standards — Social Studies



The IDOE reported ELA and Mathematics student performance on the vertically linked scale established by Smarter. The IRT vertical scale was formed by linking across grades using common items in adjacent grades. Grade 6 was used as the baseline, and each grade was successively linked onto the scale. More details about the vertical scaling methods can be found in Chapter 9 of the 2013–2014 Technical Report (Smarter Balanced, 2016). Each Science and Social Studies assessment was reported on a separate within-test scale. The scale score is the linear transformation of the IRT ability estimate, θ:

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

Table 6 lists the scaling constants *a* and *b* for all grades and subjects.

Table 6: Scaling Constants

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

Applying the ILEARN scale score transformations to the performance standards recommended by the workshop panels results in the system of scale score ranges for each of the ILEARN performance-level classifications identified in Table 7.

Table 7: ILEARN Scale Score Ranges Based on Final Performance Standards

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
ELA 3	5060–5415	5416–5459	5460–5514	5515–5760
ELA 4	5090–5443	5444–5492	5493–5546	5547–5810
ELA 5	5110–5471	5472–5523	5524–5594	5595–5850
ELA 6	5130–5491	5492–5543	5544–5603	5604–5870
ELA 7	5130–5506	5507–5567	5568–5628	5629–5890
ELA 8	5150–5510	5511–5576	5577–5637	5638–5920
Mathematics 3	6080–6381	6382–6424	6425–6487	6488–6730
Mathematics 4	6100–6428	6429–6473	6474–6540	6541–6800
Mathematics 5	6110–6452	6453–6509	6510–6565	6566–6850
Mathematics 6	6110–6487	6488–6544	6545–6604	6605–6870
Mathematics 7	6120–6492	6493–6561	6562–6624	6625–6920
Mathematics 8	6120–6508	6509–6589	6590–6650	6651–6950
Science 4	7350–7481	7482–7505	7506–7534	7535–7650
Science 6	7350–7465	7466–7503	7504–7544	7545–7650

Grade	Below Proficiency	Approaching Proficiency	At Proficiency	Above Proficiency
Biology	7350–7477	7478–7508	7509–7546	7547–7650
Social Studies 5	8350–8476	8477–8501	8502–8542	8543–8650

5. EVALUATION OF THE STANDARD SETTING WORKSHOP

5.1 PANELIST EVALUATION OF STANDARD SETTING WORKSHOP

Following the completion of standard setting tasks, panelists were asked to evaluate different aspects of the workshop and the resulting recommendations. At the end of the workshop, all but two panelists indicated that training on the main components and tools of the bookmark procedure was adequate, and that they understood how to use each component.

Generally, panelists indicated that the amount of time allotted for different activities within the standard setting workshop was “about right.” Overall, panelists expressed general satisfaction with the workshop and offered suggestions for improving the experience in future meetings.

Across all panels, most participants indicated they agreed that students classified at each performance level are fairly placed into each of the performance-level classifications based on the knowledge and skills described in the Indiana Academic Standards (IAS), as summarized in Table 8. Appendix K, Summary of Panelist Evaluations, shows panelists’ responses to the evaluation forms.

Table 8: Summary of Panelist Evaluation of Recommended Performance Standards

Workshop Evaluation Question	Strongly Disagree	Disagree	Agree	Strongly Agree
I am confident that students classified as At Proficiency are proficient in knowledge and skills described in the Indiana Academic Standards.	0	2	43	59
I am confident that students classified as Approaching Proficiency are fairly classified in knowledge and skills described in the Indiana Academic Standards.	0	5	41	58
I am confident that students classified as Above Proficiency exceed proficiency in knowledge and skills described in the Indiana Academic Standards.	0	1	41	62

5.2 INDEPENDENT OBSERVER REVIEW OF STANDARD SETTING WORKSHOP

IDOE invited members of the SBOE TAC to attend and observe the standard setting workshop. One observer attended and submitted a report to the SBOE describing their experience at the workshop; the report was produced independently without input or review from IDOE.

6. ADOPTION OF FINAL PERFORMANCE STANDARDS

On July 25, the SBOE adopted the panelist-recommended performance standards.

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.

Cizek, G.J. & Bunch, M.B. (2006). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousands Oaks, CA: SAGE Publications.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlba.

Smarter Balanced Assessment Consortium (2016). 2013–2014 Technical Report. Retrieved from: https://portal.smarterbalanced.org/library/en/2013-14-technical-report.pdf.

Indiana Learning Evaluation Readiness Network (ILEARN)

February 11–12, 2019

Volume 7
U.S. Government
Standard Setting

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, Kevin Clayton, and Kyra Bilenki. Major contributors from the Indiana Department of Education include the Director of Assessment, Assistant Directors of Assessment, and Program Leads.

Table of Contents

LIST OF APPENDICES

List of Tables

List of Figures

Background

The U.S. Government assessment is an optional end-of-course assessment (ECA) pursuant to Indiana Code 20-32-5. Although the assessment was administered for the first time in spring 2019, educators were advised that the assessment could be used as a final exam for the spring 2019 semester and that a proficiency indicator would be available in the Online Reporting System (ORS) beginning on May 6, 2019. Thus, while standard setting workshops for most ILEARN assessments were scheduled for July, after item calibration and equating and ability estimation activities were finished, standard setting for the U.S. Government assessment had to be completed without reference to information about item difficulty, student ability, or impact data.

The purpose of this document is to provide a technical summary of the process used to recommend a proficient performance standard for the ILEARN U.S. Government ECA using the Angoff method of standard setting (Angoff, 1971). Because the standard setting workshop was to be conducted prior to the administration of any assessment items, a standard setting method that did not rely on item difficulty or student ability was required. The Angoff method provided an assessment-centered, research-based procedure that could be implemented without item statistics. Following this approach, panelists first developed range performance level descriptors (PLDs), followed by "Just Barely," or "Threshold," PLDs. Panelists then worked through the assessment items to identify the knowledge and skills necessary for students to successfully respond to each item. Panelists used their understanding of just barely proficient students and the knowledge and skill requirements of the assessment items to assign the probability that a just barely proficient student would respond correctly to each assessment item. Panelists made the probability judgment in two rounds and had the opportunity to review their Round 1 judgments with their fellow panelists. AIR used its online standard setting tool to employ the Angoff method.

Overview of Standard Setting Activities

The IDOE implemented a standard setting workshop to recommend a proficient performance standard to demarcate student performance as At Proficiency or Below Proficiency with respect to the Indiana Academic Standards (IAS) in U.S. Government. The standard setting workshop was conducted February 11 and 12, 2019. Educators from around Indiana identified and recommended an assessment score on the Spring 2019 U.S. Government ECA to IDOE associated with a proficient level of performance.

Standard setting refers to methods of identifying performance standards that indicate whether a student has performed to an established level of achievement. Standard setting involves expert judgment that is typically informed by student performance data. A vast amount of literature describes a wide range of standard setting techniques. Some of these techniques are normative and identify performance standards that yield a desired percentage of test takers placed in two or more categories. Other techniques focus on what students know and are able to do.

Based on conversations between IDOE and AIR, and input from the Indiana State Board of Education Technical Advisory Committee (SBOE TAC), the Angoff method of standard setting was chosen to avoid relying on item response theory (IRT) parameters or impact data.

Staff from AIR used the Angoff method to set achievement standards. The performance standards recommended from the process were:

- Content referenced, because they were based on a rigorous application of the IAS
- Reasonable, because they were based on the expert, informed judgments of the standard setting panels; and
- Credible, because a diverse group of panelists followed a rigorous and well-supported standard setting procedure.

Agenda

AIR designed a schedule that accounted for the range and just barely PLD creation and the assignment of probabilities which allowed work to be completed in two days. The agenda can be found in Appendix A.

Orientation and Training

Training is an essential element of a standard setting workshop. The first day of the workshop began with orienting panelists on the workshop activities. Major workshop activities included development of range PLDs and, threshold or just barely PLDs, review of assessment items, and the assignment of the probability that a just barely proficient student would respond correctly to each assessment item. Panelists received training before each workshop activity. For both the range and just barely PLDs, panelists reviewed sample PLDs to gain understanding of the language and rigor typically used in each type of PLD. Panelists then worked as a group, led by the AIR facilitator, to develop initial PLDs to ensure that all panelists

understood how to construct the PLDs. As part of the training for the assignment of probability estimates that just barely proficient students would respond correctly to assessment items, panelists worked through sample items as a group and then engaged in a practice round and reviewed their probability judgments as a group. Panelists were administered a survey to ensure that they understood the main concepts of the Angoff method. Before each judgment round, panelists completed a readiness form indicating that they felt prepared to make the probability judgments.

IDOE reviewed and approved all training materials used in the standard setting meeting. The training PowerPoint presentation slides can be found in Appendix I.

Performance Level Descriptors

PLDs for the U.S. Government ECA were not included as part of the Spring 2018 PLD workshop. Thus, U.S. Government standard setting workshop panelists began with a review of the policy PLDs adopted in Spring 2018 and then worked to construct range PLDs for the U.S. Government academic content standards. U.S. Government has only one performance standard, which will be used to classify students as At Proficiency or Below Proficiency with respect to the standards.

Policy PLDs

Policy PLDs articulate the overall claims about a student's performance in each performance level. Prior to the U.S. Government standard setting meeting, AIR and IDOE used the policy PLDs created in May 2018 to draft the policy PLDs for U.S. Government for educator review. During the meeting, the facilitator first walked panelists through the policy PLDs for grade 5 Social Studies, outlining the key descriptors at each performance level. The training was meant to help panelists internalize the sense of rigor conveyed by the descriptors at each performance level (proficient and not proficient).

Following this discussion, panelists reviewed the U.S. Government policy PLDs. The facilitator engaged the panelists in a room-level discussion and discussed the expectations of the policy PLDs and how they should inform the range PLDs. The goal of the discussion was for the panelists to develop a shared sense of the kind of student described by each proficiency level.

Range PLDs

Range PLDs define the content area knowledge, skills, and processes that test takers at a particular performance level are expected to possess. They describe the prototypical members of the given performance level. For the U.S. Government assessment, only a proficient PLD was required to differentiate students who were proficient with respect to the IAS from those who were not.

The facilitator described the process for creating range PLDs and shared the tools used for creating them, including the IAS and example range PLDs.

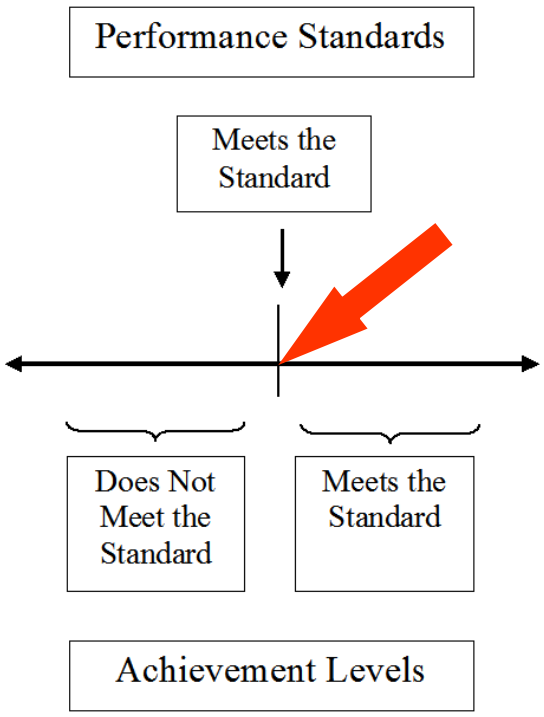
Per discussions between IDOE and AIR, 15 range PLDs (27% of the 56 total standards) were drafted prior to the meeting by AIR content specialists. IDOE reviewed these draft range PLDs before they were shared with panelists during the meeting. From the two larger reporting categories (i.e., Functions of Government, Institutions and Processes of Government), six standards were selected. From the smaller reporting category (i.e., Historical Foundations of American Government), three standards were selected. These draft range PLDs were used for training and discussion purposes. Panelists worked as a group to discuss and refine the draft range PLDs. The remaining standards were distributed among the three tables, and panelists at those tables crafted the range PLDs for their assigned standards. The facilitators spent approximately 30 minutes training panelists to construct range PLDs, and then panelists had two and a half hours to write the assigned range PLDs.

"Just Barely" PLDs

Just barely PLDs (also called threshold PLDs) define what students who just barely qualify for entry into a performance level know and are able to do. As a result, these descriptors represent the beginning level of the associated performance level. The just barely descriptors are represented in Figure 1. The red arrow in Figure 1 indicates that the focus on the lowest point of the range PLDs. Just barely descriptors help narrow the focus of panelists to the most basic and essential knowledge and skills required to differentiate students who fall at the border between two performance levels. The just barely proficient descriptors were developed by the panelists as part of the standard setting activity. AIR developed, with IDOE review, a just barely template and three illustrative examples prior to the meeting. The panelists used the template to develop their own just barely PLDs.

Panelists wrote just barely PLDs only for the 42 standards represented by items on the spring 2019 operational assessment form. Three standards were used for training, and the remaining standards were divided across the three tables.

Figure 1: Just Barely Descriptors



Reviewing the Assessment Form

Central to the Angoff method is the base year assessment form. The assessment form contained the operational assessment items on which panelists were setting standards. Because the operational assessment form was composed entirely of items for which no item statistics were yet available, seven embedded field-test items measuring standards that are required in order to meet the blueprint were added to the operational assessment items, in case some operational assessment items were rejected in the item data review process. The assessment form was presented to panelists electronically. Panelists could interact with items the same way students would during testing. In their review of assessment items, panelists were instructed to think about what students need to know and are able to do to respond successfully to each item.

Training in Assigning the Probability of Correct Responding

Prior to assigning the likelihood of students responding to items correctly, panelists received training in the Angoff method and use of the online standard setting tool to enter their values.

The workshop facilitator trained panelists to assign probability estimates to assessment items using a set of practice U.S. Government assessment items. The facilitator worked through each practice assessment item with the panel. The facilitator and panelists discussed the

knowledge and skill requirements of the practice item, and panelists shared their judgments about how likely a just barely proficient student would be to respond correctly to the item and their rationale for these judgments.

Discussions focused on the performance of students just barely meeting the proficient standard to ensure that panelists were basing their judgments on this special group of students. The facilitator then showed panelists how to assign the likelihood that a just barely proficient student would answer the item correctly in the standard setting tool.

The key to training was moderating educator judgment. People pervasively demonstrate unrealistic optimism when judging the probability of a positive outcome (Sharot, Korn, & Dolan, 2011). Because of this cognitive bias, panelists are prone to overestimate the probability that students will respond correctly to assessment items. AIR and IDOE worked to convey to panelists that when the difficulty of an assessment item matches the ability of just barely proficient students, those students would have a 0.5 probability of responding correctly. In addition, panelists were instructed that because most of the assessment items were constructed to measure proficiency of the IAS, most assessment items would be distributed near the proficient level performance standard.

In the online standard setting tool, the probabilities of correct response from which panelists could select ranged from 0.25 to 0.75 in increments of 0.05. AIR reviewed a large bank of U.S. Government assessment items, which showed a 0.47 mean probability of a correct response with a standard deviation of 0.20, indicating that most items would fall in the 0.25 to 0.75 range. A 0.25 probability of a correct response indicates a chance rate of responding, and most items should have a higher probability. Conversely, 0.75 is a high probability of a correct response, and few items written to grade level should have a probability this high.

After completing the training, panelists took a short survey (see Appendix B) regarding what it means to be just barely proficient and the relationship between the knowledge and skills needed to respond correctly to an item and the likelihood that a just barely proficient student would respond correctly.

After all panelists finished the survey, the facilitator led a discussion about the questions and answers. AIR staff collected the surveys and confirmed that panelists completed the survey and that they had indicated the correct answers.

Practice in Assigning the Likelihood

Practice Round

Panelists then logged in to the standard setting tool to practice performing the Angoff judgment task using a six-item practice assessment form. Panelists worked to evaluate the knowledge and skill requirements of the practice items with respect to their just barely PLDs to assign a probability that just barely proficient students would respond correctly to each item and to practice assigning the probability in the standard setting tool. The six items covered a range

of standards and included both multiple-choice and multi-select item types represented in the U.S. Government assessment. The six items were used for discussion purposes after the panelists completed the practice task. Similar to the training described previously, discussions focused on the performance of students just barely meeting the proficient standard.

Once all panelists completed the practice task, they were given feedback. The feedback included item position, item ID, median panelist likelihood value, minimum panelist likelihood value, maximum panelist likelihood value, the 25th percentile of panelist likelihood values, the 75th percentile of panelist likelihood value, and the interquartile range (IQR) of panelist likelihood values. The facilitator led a discussion, focusing on items in which panelist likelihood values varied the most. Example feedback is given in Appendix E.

Round 1

Panelists then signed the Round 1 Readiness Form (Appendix C) indicating that they understood the task at hand and were ready to make their recommendations.

Panelists were instructed to keep in mind the characteristics of students who just barely qualified for the proficient performance level and made independent judgments about the probability that a just barely proficient student would answer each item in the assessment form correctly. The recommendations were recorded in the online standard setting tool.

Panelists received and discussed feedback from their Round 1 ratings for tables and the entire room. The feedback was in the form of statistics that described the central tendency and variability of the panelists' ratings. The facilitator worked with the room as a whole to discuss items with the greatest variation among panelists. Panelists then completed review of items within their tables. An example of feedback from Round 1 is given in Appendix E.

Round 2

Panelists were instructed to keep in mind feedback from Round 1 and what the characteristics are of a student who just barely qualifies for the proficient performance level. After signing the Round 2 Readiness Form (Appendix C), each participant made an independent Round 2 judgment about the probability that a just barely proficient student would respond correctly to each item in the assessment form.

Workshop Evaluation

At the completion of the workshop, panelists completed an evaluation form. The summary of results can be found in Appendix F. The evaluation form was designed to elicit feedback on all aspects of the workshop, including clarity of training and tasks, appropriateness of the time spent on activities, and satisfaction with the workshop's outcome.

Meeting Logistics

Educator Panel

The Educator Panel consisted of 12 panelists recruited by IDOE from across the state. IDOE focused on schools that were participating in the U.S. Government ECA, but some panelists were drawn from schools that were not participating. The recruiting plan for obtaining panelists for the standard setting meetings was intended to result in a representative group of panelists who would render informed recommendations to the state on the placement of the performance standards. Diverse groups of panelists bring a wide range of perspectives and experience to the standard setting effort, ensuring that the recommendations are thoughtful and representative of broad education constituencies.

The members of the Educator Panel were asked to fill out an Educator Panel Demographic Information sheet. The results from the demographic sheet are summarized in Appendix D and include breakdowns by gender, race/ethnicity, years of experience, and summary of current role (e.g., teacher, administrator).

Meeting Staff

The panelist room included AIR facilitators, content support, and IDOE staff. The facilitators conducted training and practice, led discussions for two rounds of standard setting, decided when to begin and end each meeting phase, fielded panelist questions, and ensured that timely recommendations were provided. AIR staff greeted panelists when they arrived, registered them, provided assigned materials, and ensured the security of assessment materials at all times.

Table 1: IDOE and AIR Educator Panel Attendees

	Attendee	Affiliation	Role
1	Dr. Charity Flores	IDOE	Introductory Remarks
2	Dr. Kristine David	IDOE	Introductory Remarks
3	Tim Martin	IDOE	Social Studies Observer
4	Mary Williams	IDOE	Observer
5	Kelly Connelly	IDOE	Observer
6	Justin Mocas	IDOE	Observer
7	Tracie Morris	AIR	Program Management
8	Susan Sherwood	AIR	Program Management
9	Stephan Ahadi	AIR	Psychometrics
10	Elizabeth Ayers-Wright	AIR	Psychometrics
11	Kevin Clayton	AIR	Psychometric Support

	Attendee	Affiliation	Role
12	Kevin Dwyer	AIR	Facilitator
13	Mike Flynn	AIR	Facilitator
14	Scott Koenig	AIR	Content Support
15	Drew Azar	AIR	Technical Support

Meeting Materials

The following materials were required for the standard setting meeting:

- One LCD projector and screen for the Educator Panel meeting room
- One hard-wired laptop computer per panelist
- Pens, pencils, notepads
- Travel and other expense reimbursement forms for panelists to complete
- Non-disclosure agreements
- Training materials

Security Considerations

The fundamental purpose of the security plan was to ensure that item and data security was not compromised.

Panelists reviewed assessment items and assigned probability judgments within AIR's online standard setting tool. The tool can be accessed only through a secure website. Panelists were assigned usernames and passwords to access the site. At the end of each day, the site was locked so that panelists could not access the material outside of the meeting room. In addition, after the workshop, panelist passwords were reset so that panelists could no longer access the secure site. Ten days after the meeting, items and item content expired, automatically removing all workshop material from the secure website. AIR has saved extracts of all panelist notes and ratings internally on secure drives, so all panelist records will be preserved.

AIR worked to keep the physical workshop environment secure. All workrooms were kept locked and/or monitored by AIR staff at all times. Panelists were prohibited from using their phones or other electronic devices while in the meeting room. Any required printing of secure or confidential material was done on green paper and collected at the end of the workshop.

Meeting Results

The Round 1 results for all items are given in Appendix G. The first 54 item positions represent the 54 operational items used to determine a student's score. The median probability of correct responding across all panelists was used to compute the passing score on the first operational assessment form. The passing score was defined as the sum of the median probabilities across the operational assessment items. The passing score based on Round 1 was 26 of 54 items correct.

The Round 2 results are given in Appendix H for all items. Again, the first 54 item positions represent the operational items. After Round 2, the passing score remained at 26 items.

Although median probability judgments were stable between the two rounds, we were also interested in determining whether there was evidence of convergence in panelists' judgments following feedback and discussion in Round 2. In Round 1, the average IQR of the probability judgments was 0.08, indicating substantial agreement among panelists even in Round 1. In Round 2, the average IQR was 0.07, indicating greater convergence among panelists' probability judgments, although the increase was small.

Post Meeting Updates

Since the meeting was held prior to the testing window and without any student or item data, the raw score was the only available metric to be used as a passing score. However, to equate test forms year to year, it is preferable to use item response theory methods that take into account item difficulty. After the close of the testing window and following item calibrations, AIR used the student data from the Spring 2019 test form to identify the scale score (on the theta metric) corresponding to the raw scoring passing value determined by the U.S. Government standard setting. The calibration sample included only those students who answered all items. The linear transformation from the theta scale to the U.S. Government scale score will be used to report student achievement results for future administrations. Table 2 shows the slope and intercept used to transform the theta score to the scale score. Table 3 shows the resulting scale score ranges for the two performance levels.

Table 2. Linear Transformation Scaling Constants

Subject	Grade	Slope (*a*)	Intercept (*b*)
Social Studies	U.S. Government	50	8500

Table 3: ILEARN U.S. Government Assessment Proficiency Cut Scores

Level 1 Below Proficiency	Level 2 At Proficiency
8350–8496	8497–8650

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature neuroscience, 14*(11), 1475.