Indiana's Alternate Measure

2018–2019

Volume 1
Annual Technical Report

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Hyesuk Jang, Yuan Hong, Myung Hee Im, Katherine Krehbiel, Hashim Evans, and Celine Bryan. 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

1. INTRODUCTION

The Indiana's Alternate Measure (I AM) 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, 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 I AM 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 I AM standard setting process.
7. *Special Studies*. This volume compiles any special studies conducted for the I AM; it is updated annually to reflect studies relevant to the respective test administration. No special studies occurred in 2018-2019.

The Indiana Department of Education (IDOE) communicates the quality of the I AM by making these technical reports accessible to the public.

1.1 BACKGROUND AND HISTORICAL CONTEXT OF TEST

I AM was constructed to measure student achievement in English/Language Arts (ELA), Mathematics, Science, and Social Studies relative to the Indiana Alternate Academic Standards, or Content Connectors. I AM was first administered to students in Spring 2019, replacing the Indiana Standards Tool for Alternate Reporting (ISTAR).

1.2 PURPOSE AND INTENDED USES OF THE INDIANA'S ALTERNATE MEASURE

I AM is a criterion-referenced test that applies 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 achievement of Indiana's Alternate Standards, or Content Connectors. I AM 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.

I AM draws all items from the I AM item bank that includes the ISTAR legacy items (see Volume 2). I AM content standards in ELA, Mathematics, Science, and Social Studies are aligned with knowledge and skills that are essential for competitive employment and post-secondary education. The American Institutes for Research (AIR) and the IDOE worked together to ensure that the items in the test forms constructed for all grades uniquely measure students' mastery of the Indiana Alternate Academic Standards (Content Connectors) in ELA, Mathematics, Science, and Social Studies.

Table 1 outlines the required participation criteria of I AM. The purpose and intended use of I AM is found on the IDOE web page.

Table 1: Participation Criteria for I AM

Participation Criteria
Review of student record indicates a disability that significantly impacts intellectual functioning and adaptive behavior. Adaptive behavior is defined as essential for someone to live independently and to function safely in daily life.
The student requires extensive, repeated, individualized instruction and support that is not of a temporary nature.
The student uses substantially adapted materials and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate, and transfer skills across multiple settings.
Goals listed in the Individual Education Plan (IEP) for this student are linked to the enrolled grade level Alternate Achievement Standards (Indiana Content Connectors).

1.2.1 Participants in the Development and Analysis of the Indiana's Alternate Measure

The IDOE manages the Indiana state assessment program with the assistance of Indiana educators, a Technical Advisory Committee (TAC), and several vendors (listed in the following paragraphs). The IDOE fulfills the diverse requirements of implementing the I AM 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 alternate assessment 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 I AM. Educators participate in the development of the academic standards, clarification of how these standards will be assessed, test design, and committee reviews of test items and passages.

Technical Advisory Committee

The State Board of Education convenes a TAC 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 independent and higher education institutions.

Corporation-Based Advisory Group

Indiana also utilizes a corporation-based advisory group, Assessment Implementation Advisory Group, that meets monthly to advise on implementation practices. This group consists of educators, technology coordinators, and test coordinators.

American Institutes for Research

AIR is the vendor selected through the state-mandated competitive procurement process. In Spring 2018, 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 the I AM described in this report. Additionally, AIR is responsible for developing and maintaining the I AM item bank, which is used for the I AM test construction.

Assessment Systems Corporation

The Assessment Systems Corporation conducts independent verifications of scoring activities in Spring 2019.

Caveon Test Security

Caveon Test Security monitored web pages and social media during the Spring 2019 test administration to ensure that any secure testing materials, such as items and prompts, were not leaked to the public or unauthorized personnel.

1.3 Available Test Formats and Special Versions

I AM is administered as an online, stage-adaptive assessment using multiple-choice (MC) item types. Students who are unable to participate in the online administration are administered the test in a paper-and-pencil format as an accommodation. This format is available in regular print, large print, and braille. The paper-and-pencil format includes the same operational items as the online format of the assessment. Students participating in the computer-based I AM use text-to-speech to hear the item stimulus, stem, and answer choices. Similarly, test administrators use a script to read the item stimulus, stem, and

answer choices to students who participate in the paper-pencil format. Students participating in the computer-based I AM can use standard online testing features in the test delivery system (TDS), which includes a selection of font colors and sizes and the ability to zoom in, zoom out, and highlight text. Students can take I AM with or without accommodations. English Learners (ELs) can take the Spanish language version of the I AM Mathematics, Science, and Social Studies assessments. No students took the Spanish forms for 2019. The Spanish forms were the same tests as English but just translated in Spanish. The items were translated by third-party vendor that provides professional translation services. During test development, it was ensured that scores obtained using the Spanish language version or other alternative modes of administration were comparable to those received on the standard online test adhering to the same blueprints. The test summary comparison between the standard online form and the paper-and-pencil form, which matches the Spanish language form, is provided in Volume 2.

1.4 STUDENT PARTICIPATION

All Indiana public school students in grades 3–8 and 10 in ELA and Mathematics, grades 4 and 6 and high school in Science, and grade 5 in Social Studies can participate in the state assessments. Table 2 shows the number of students tested and the number of students reported in the Spring 2019 I AM administration by grade and subject area. Table 3 through Table 6 present the distribution of students, in counts and percentages, by subgroups for ELA, Mathematics, Science, and Social Studies, respectively. The subgroup categories reported here are gender, primary disability, and race/ethnicity.

Table 2: Number of Students Participating in the I AM 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	766	713	3	765	709						
4	841	772	4	840	766	4	838	766			
5	877	818	5	873	809				5	867	802
6	1016	961	6	1009	953	6	1001	943			
7	1042	986	7	1045	988						
8	1157	1103	8	1158	1101						
10	1141	1078	10	1140	1078	Biology	1067	1013			

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

Grade	Group	All Students	Female	Male	Autism	Non-Autism	Moderate and Severe Intellectual Disability	Non-Moderate and Severe Intellectual Disability	African American	Hispanic	White
3	*N*	766	260	506	281	483	185	579	124	96	478
	%	100	33.94	66.06	36.68	63.05	24.15	75.59	16.19	12.53	62.4
4	*N*	841	289	552	251	588	215	624	122	123	518
	%	100	34.36	65.64	29.85	69.92	25.56	74.2	14.51	14.63	61.59
5	*N*	877	280	597	256	618	216	658	155	113	544
	%	100	31.93	68.07	29.19	70.47	24.63	75.03	17.67	12.88	62.03
6	*N*	1,016	358	658	307	696	267	736	164	128	629
	%	100	35.24	64.76	30.22	68.5	26.28	72.44	16.14	12.6	61.91
7	*N*	1,042	357	685	304	731	260	775	186	141	644
	%	100	34.26	65.74	29.17	70.15	24.95	74.38	17.85	13.53	61.8
8	*N*	1,157	421	736	324	831	310	845	178	136	770
	%	100	36.39	63.61	28	71.82	26.79	73.03	15.38	11.75	66.55
10	*N*	1,141	428	713	291	843	290	844	198	106	761
	%	100	37.51	62.49	25.5	73.88	25.42	73.97	17.35	9.29	66.7

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

Grade	Group	All Students	Female	Male	Autism	Non-Autism	Moderate and Severe Intellectual Disability	Non-Moderate and Severe Intellectual Disability	African American	Hispanic	White
3	*N*	765	261	504	280	483	184	579	124	96	477
	%	100	34.12	65.88	36.6	63.14	24.05	75.69	16.21	12.55	62.35
4	*N*	840	289	551	251	587	215	623	121	123	519
	%	100	34.4	65.6	29.88	69.88	25.6	74.17	14.4	14.64	61.79
5	*N*	873	277	596	254	616	215	655	153	113	543
	%	100	31.73	68.27	29.1	70.56	24.63	75.03	17.53	12.94	62.2
6	*N*	1,009	357	652	303	693	266	730	162	126	626
	%	100	35.38	64.62	30.03	68.68	26.36	72.35	16.06	12.49	62.04
7	*N*	1,045	360	685	307	732	260	779	184	143	647
	%	100	34.45	65.55	29.38	70.05	24.88	74.55	17.61	13.68	61.91
8	*N*	1,158	421	737	323	833	311	845	180	136	769
	%	100	36.36	63.64	27.89	71.93	26.86	72.97	15.54	11.74	66.41
10	*N*	1,140	429	711	288	845	291	842	199	105	760
	%	100	37.63	62.37	25.26	74.12	25.53	73.86	17.46	9.21	66.67

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

Grade	Group	All Students	Female	Male	Autism	Non-Autism	Moderate and Severe Intellectual Disability	Non-Moderate and Severe Intellectual Disability	African American	Hispanic	White
4	*N*	838	289	549	249	587	215	621	121	123	518
	%	100	34.49	65.51	29.71	70.05	25.66	74.11	14.44	14.68	61.81
6	*N*	1,001	352	649	301	687	265	723	160	125	622
	%	100	35.16	64.84	30.07	68.63	26.47	72.23	15.98	12.49	62.14
Biology	*N*	1,067	384	683	265	801	281	785	213	104	675
	%	100	35.99	64.01	24.84	75.07	26.34	73.57	19.96	9.75	63.26

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

Grade	Group	All Students	Female	Male	Autism	Non-Autism	Moderate and Severe Intellectual Disability	Non-Moderate and Severe Intellectual Disability	African American	Hispanic	White
5	*N*	867	276	591	254	610	214	650	152	114	538
	%	100	31.83	68.17	29.3	70.36	24.68	74.97	17.53	13.15	62.05

2. SUMMARY OF OPERATIONAL PROCEDURES

2.1 ADMINISTRATION PROCEDURES

The 2018–2019 I AM administration window opened on April 8, 2019, and closed on May 17, 2019, across all subjects. The key personnel involved with the I AM administration included the Corporation Test Coordinators (CTCs), School Test Coordinators (STCs), and Test Administrators (TAs) who proctored the test. A Test Administrator's Manual and Test Coordinator's Manual, 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 I AM tests. The online browser provides 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. If the test was paused for more than 14 days, the test opportunity would expire. The STC would be required to submit a test irregularity request to reopen the assessment.

2.2 DESIGNATED FEATURES AND ACCOMMODATIONS

Accessibility supports discussed within this document include

1. 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;
2. designated features that are available to students for whom the need has been identified by an informed educator or team of educators; and
3. accommodations that are available for students for whom there is documentation on an IEP or Individual Learning Plan (ILP).

Scores achieved by students using designated features are included for federal accountability purposes. All educators making decisions on use of these features are trained in the process and understand the range of designated features 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., color contrast, print size) are provided digitally through instructional or assessment technology; non-embedded designated features (e.g., assistive technology to magnify/enlarge) are non-digital. Students who require assistive technology must have permissive mode turned on to allow for the assistive technology to function in conjunction with the secure environment. These accommodations are generally available for students for whom there is a documented need on an IEP 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 or ILP generate valid outcomes on the assessments so that students can fully demonstrate what they know and are able to do. From the psychometric perspective, the purpose of providing accommodations is to "increase the validity of inferences about students with disabilities by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

The TAs and STCs in Indiana are responsible for ensuring that arrangements for accommodations are made before the test administration dates. The available accommodation options for eligible students include braille, interpreter for sign language, streamline, assistive technology (e.g., adaptive keyboards, touch screen, switches), calculation devices, and multiplication tables.

Table 7 through Table 14 list the number of students who are marked in the Test Information Distribution Engine (TIDE) as receiving each accommodation during the Spring 2019 test administration.

Table 7: ELA Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	10
Embedded Accommodations							
Permissive Mode	129	180	222	218	186	238	185
Streamlined Mode	17	14	12	17	17	20	12
Non-Embedded Accommodations*							
Alternate Indicator of Response	625	650	670	747	744	758	495
Paper Test Booklet	61	66	56	58	59	55	41
Large Print Booklet	28	39	40	33	31	34	22
Read Aloud Script for Paper Booklet				1			
Braille booklet**			1		1		
Read Aloud to Self							
Special Request							
Non-Standard Accommodation (approved by the IDOE)							

* No data was collected on the number of students who used the "Student Provided Access to Own Resources" accommodation.

** When the IDOE followed up with corporations as to whether they required braille in TIDE, no corporations needed a braille test booklet. As a result, no braille booklets were used.

Table 8: ELA Total Students with Allowed Embedded and Non-Embedded Designated Features

Designated Supports	Grade						
	3	4	5	6	7	8	10
Embedded Designated Features							
Translations							
Masking	766	840	876	1,016	1,040	1,156	1,139
Mouse Pointer	4	6	6	14	9	6	3
Print Size	7	10	10	12	8	11	8
Color Contrast	2	9	9	2	1	3	9
Non-Embedded Designated Features							
Color acetate film for paper assessment	2	1	2		1		
Assistive technology to magnify/enlarge	28	34	32	35	29	41	20
Access to sound amplification system	10	7	17	11	5	9	6
Special furniture or equipment for viewing test	41	35	34	27	28	22	23
Special lighting conditions	24	22	13	21	19	18	8
Time of day for testing altered	200	190	165	185	178	187	154

Table 9: Mathematics Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade						
	3	4	5	6	7	8	10
Embedded Accommodations							
Permissive Mode	129	180	224	211	184	236	186
Streamlined Mode	17	14	12	17	16	20	11
Non-Embedded Accommodations*							
Alternate Indicator of Response	625	651	668	738	747	756	495
Handheld/Adaptive Calculator	118	124	144	299	307	322	333
Multiplication Table		1	2	165	134	155	119
Paper Test Booklet	62	64	56	54	61	53	42
Large Print Booklet	27	41	41	31	32	34	23
Read Aloud Script for Paper Booklet				1			
Hundreds Chart	164	176	164	165	140	153	98
Braille booklet**			1		1		
Read Aloud to Self							
Special Request							
Non-Standard Accommodation (approved by the IDOE)							

* No data was collected on the number of students who used the "Student Provided Access to Own Resources" accommodation.

** When the IDOE followed up with corporations as to whether they required braille in TIDE, no corporations needed a braille test booklet. As a result, no braille booklets were used.

Table 10: Mathematics Total Students with Allowed Embedded and Non-Embedded Designated Features

Designated Supports	Grade						
	3	4	5	6	7	8	10
Embedded Designated Features							
Translations							
Masking	765	839	872	1,009	1,042	1,157	1,138
Mouse Pointer	1	5	5	11	8	7	4
Print Size	5	5	7	12	9	13	9
Color Contrast	1	5	12	2	1	3	8
Non-Embedded Designated Features							
Color acetate film for paper assessment	2	1	2		1		
Assistive technology to magnify/enlarge	27	34	33	35	28	41	22
Access to sound amplification system	10	7	17	11	5	9	5
Special furniture or equipment for viewing test	43	35	34	26	28	22	23
Special lighting conditions	24	21	13	20	19	18	10
Time of day for testing altered	202	190	165	181	179	187	155

Table 11: Science Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade		
	4	6	Biology
Embedded Accommodations			
Permissive Mode	181	210	162
Streamlined Mode	14	17	10
Non-Embedded Accommodations*			
Alternate Indicator of Response	647	734	468
Paper Test Booklet	66	55	42
Large Print Booklet	39	30	19
Read Aloud Script for Paper Booklet Braille booklet		1	
Read Aloud to Self			
Special Request			
Non-Standard Accommodation (approved by the IDOE)			

* No data was collected on the number of students who used the Student Provided Access to Own Resources" accommodation.

** When the IDOE followed up with corporations as to whether they required braille in TIDE, no corporations needed a braille test booklet. As a result, no braille booklets were used.

Table 12: Science Total Students with Allowed Embedded and Non-Embedded Designated Features

Designated Supports	Grade		
	4	6	Biology
Embedded Designated Features			
Translations			
Masking	837	1001	1065
Mouse Pointer	3	8	1
Print Size	6	10	2
Color Contrast	4	2	18
Non-Embedded Designated Features			
Color acetate film for paper assessment	1		
Assistive technology to magnify/enlarge	34	34	15
Access to sound amplification system	7	11	8
Special furniture or equipment for viewing test	35	26	18
Special lighting conditions	22	19	7
Time of day for testing altered	191	181	131

Table 13: Social Studies Total Students with Allowed Embedded and Non-Embedded Accommodations

Accommodations	Grade
	5
Embedded Accommodations	
Permissive Mode	221
Streamlined Mode	12
Non-Embedded Accommodations*	
Alternate Indicator of Response	664
Paper Test Booklet	55
Large Print Booklet	40
Read Aloud Script for Paper Booklet	
Braille booklet**	1
Read Aloud to Self	
Special Request	
Non-Standard Accommodation (approved by the IDOE)	

* No data was collected on the number of students who used the Student Provided Access to Own Resources" accommodation.

** When the IDOE followed up with corporations required braille in TIDE, no corporations as to whether they needed a braille test booklet. As a result, no braille booklets were used.

Table 14: Social Studies Total Students with Allowed Embedded and Non-Embedded Designated Features

Designated Supports	Grade
	5
Embedded Designated Features	
Translations	
Masking	866
Mouse Pointer	3
Print Size	6
Color Contrast	12
Non-Embedded Designated Features	
Color acetate film for paper assessment	2
Assistive technology to magnify/enlarge	33
Access to sound amplification system	17
Special furniture or equipment for viewing test	33
Special lighting conditions	12
Time of day for testing altered	163

3. ITEM BANK AND TEST CONSTRUCTION

3.1 OVERVIEW OF ITEM DEVELOPMENT

All operational items used on the I AM test forms are drawn from the I AM item bank. Volume 2 is a separate, stand-alone report containing complete details on the I AM item bank. The I AM bank includes the newly developed I AM items and ISTAR legacy items. The parameter estimation of I AM items is described in Section 5 of this volume.

3.2 TEST DESIGN

The I AM assessments were designed to be stage adaptive, with all students administered common test segments in Part 1, with subsequent routing to Part 2 tiered segments based on Part 1 ability estimates.

Part 1 consisted of the following three sections:

- Segment 1: Two practice items confirming the student's participation in the practice test and 3 operational or operational field-test (OFT) items of varying complexity;
- Segment 2: 17 operational or OFT items of varying complexity; and,
- Segment 3: Embedded field-test (EFT) items that matched the same blueprint as a subset of the OFT items. These EFT items were only used for scoring if they were needed to meet the blueprint following item data review (IDR). The items included in this EFT section were fixed, so every student participating in the assessment saw the same items.

In Part 2, more targeted items were administered to the student based on his or her performance in Part 1. There were three stage adaptive segments of Part 2: Tier 1 (low complexity), Tier 2 (moderate complexity), and Tier 3 (high complexity). Each form in Part 2 included two sections, one Tier segment and EFT, as follows:

- Segments 4–6: 12 operational items and,
- Segment 7: Randomly selected EFT items that were not used for scoring

Each I AM test included 32 operational or operational field test items that were used for scoring as well as 2 practice items and 15 EFT items. The EFT items were administered in both Part 1 and Part 2. Table 15 shows the number of EFT items in Part 1 and Part 2.

Table 15: Number of EFT Items

Test	Number of EFT Items in Part 1	Number of EFT Items in Part 2
ELA 3	5	10
ELA 4	6	9
ELA 5	7	8
ELA 6	2	13
ELA 7	6	9
ELA 8	4	11
ELA 10	6	9
Mathematics 3	9	6
Mathematics 4	5	10
Mathematics 5	7	8
Mathematics 6	7	8
Mathematics 7	7	8
Mathematics 8	8	7
Mathematics 10	5	10
Science 4	5	10
Science 6	6	9
Biology	9	6
Social Studies 5	5	10

3.3 FIELD TESTING

The 2019 I AM test forms contained new field-test items and a collection of items eligible for field testing from ISTAR. The EFT slots (in paper-and-pencil tests) or segments (in online tests) are located with fixed positions across all subjects, such that item location and motivation effects, if they exist, will 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.

The field-test engine randomly samples field-test items for each individual test administration, essentially creating thousands of unique EFT forms. This sampling approach to embedding field-test items results in the following important outcomes:

- Reduction in the number of EFT items that each student must respond to and more efficient “spiraling” of items, which reduces clustering of item responses, resulting in more-precise parameter estimates
- More-generalizable item statistics, because they are not based on items appearing in a single position
- A truly representative sample of respondents for each item

The EFT algorithm consists of two different algorithms—one for identifying which field-test items will be administered to which student (the distribution algorithm) and one for selecting the position on the test for each item administered to the student (the positioning algorithm). When a student starts a test, the system randomly selects a pre-determined number of item groups, stopping when it has selected item groups containing at least the minimum number of field-test items designated for administration to each student. This randomization ensures that (1) each item is seen by a representative sample of Indiana students, and (2) every item is as likely as every other item to appear in a class or school, minimizing clustering effects.

3.4 OPERATIONAL FORM CONSTRUCTION

Operational test form development (see Volume 2) includes an array of item types used to measure the Content Connectors. Table 16 describes the item type in the I AM pool, and Table 17 shows the number of items by item type used in the operational forms. In the operational forms, only multiple-choice items were administered. The description and examples of each of the item types are also provided in Volume 2.

Table 16: Item Types and Descriptions

Response Type*	Description
Multiple-Choice (MC)	Student selects one correct answer from three options.
Multiple-Select (MS) (Science only)	Student selects all correct answers from a number of options.
Table-Match (MI) (Science only)	Student checks a box to indicate if information from a column header matches information from a row.

* Note that the abbreviations MC, MS, and MI correlate to the attributes used in AIR's Item Tracking System (ITS).

Table 17: Operational Items by Item Type and Grade

Subject	Item Type	3	4	5	6	7	8	10	Biology
ELA	MC	44	44	44	44	44	44	44	
Mathematics	MC	43	44	44	44	44	44	43	
Science	MC		41		41				44
Social Studies	MC			42					

4. CLASSICAL ANALYSES OVERVIEW

4.1 CLASSICAL ITEM ANALYSES

AIR psychometricians collectively monitor the behavior of items while test forms are administered in a live environment. This is accomplished using AIR's quality monitoring system, which yields an item analysis report on the performance of test items throughout the test window. During administration of the 2018–2019 I AM 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 might 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 could be produced to flag only items with statistics falling outside a specified range or to generate reports based on all items in the pool. The criteria for flagging and reviewing items is provided in Table 18, and a description of the statistics is provided in the following subsections. In the 2018–2019 I AM, the MC items were only used for scoring but the non-MC items such as MS and MI items were also administered as the field test items and included in the item calibration. Classical item statistics, the number of flagged items, and differential item functioning (DIF) categories for operational items and field-test items can be found in Appendices A and B, respectively. The flagged items from the new I AM items were reviewed in the IDR meeting, and the items that survived IDR were included in the pool.

Table 18: Thresholds for Flagging Items in Classical Item Analysis

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

* The IDOE made the decision to forego committee review for any items with an adjusted biserial/polyserial correlation less than 0.05 that are not needed to meet the blueprint. AIR shared these items with the IDOE to make final determinations.

4.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those test takers who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between

high- and low-achieving students. The discrimination index for MC items was calculated as the correlation between the item score and the ability estimate for students.

4.1.2 Distractor Analysis

Distractor analysis for MC 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 option most frequently selected 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.

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 non-MC 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.

4.1.4 Mean Total Score

For multi-point non-MC items, the average score of respondents in each score point is checked to ensure logical ordering of the average scores. For example, the mean score for score point zero should be lower than the mean score for score point one. There was no item flagged due to the inverted mean total score.

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. To identify such potential problems, all I AM items were evaluated in terms of DIF statistics based on the analyses made before the bank was established and also after the I AM was administered in Spring 2019 in Indiana.

DIF analyses were performed for the following groups:

- Male/Female
- White/African American

- White/Hispanic
- Autism/Other
- Moderate and Severe Intellectual Disability/Other

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 could contain cultural or another 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 are 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 in 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 MHχ^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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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 19. 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 19: DIF Classification Rules

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

Because of the unreliability of the DIF statistics when calculated on small samples, caution must be used in evaluating DIF classifications for items where focal or reference groups are fewer than 200 students (Mazor, Clauser, & Hambleton, 1992; Camilli & Shepard, 1994; Muniz, Hambleton, & Xing, 2001; Sireci & Rios, 2013). Because those sample sizes are not tenable for the alternate assessment program, AIR used a much smaller threshold (n = 50), which, although it may not have the power to detect real differences between subgroups, provides at least some opportunity to flag and evaluate items for possible bias. Data review participants were cautioned that DIF results based on small samples are less reliable.

4.3 CLASSICAL ANALYSES RESULTS

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

Table 20 through Table 23 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. 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 20: 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	44	0.26	0.29	0.36	0.41	0.47	0.59	0.72
4	44	0.19	0.29	0.43	0.49	0.56	0.74	0.87
5	44	0.31	0.43	0.47	0.53	0.60	0.70	0.91
6	44	0.29	0.31	0.43	0.48	0.59	0.69	0.79
7	44	0.27	0.39	0.46	0.53	0.62	0.75	0.82
8	44	0.29	0.35	0.41	0.52	0.60	0.81	0.85
10	44	0.24	0.33	0.43	0.54	0.67	0.80	0.87

Table 21: 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	43	0.23	0.27	0.38	0.44	0.53	0.59	0.75
4	44	0.19	0.24	0.37	0.46	0.49	0.58	0.80
5	44	0.24	0.28	0.33	0.41	0.47	0.55	0.70
6	44	0.20	0.28	0.34	0.43	0.54	0.64	0.92
7	44	0.20	0.24	0.30	0.40	0.47	0.55	0.69
8	44	0.19	0.24	0.34	0.41	0.47	0.51	0.54
10	43	0.19	0.22	0.28	0.37	0.49	0.57	0.67

Table 22: 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	41	0.22	0.24	0.36	0.46	0.54	0.68	0.77
6	41	0.25	0.29	0.37	0.43	0.52	0.68	0.69
Biology	44	0.25	0.29	0.43	0.48	0.59	0.78	0.79

Table 23: 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	42	0.29	0.34	0.41	0.48	0.56	0.69	0.81

DIF summary tables based on Indiana students can be found in Appendix A. The items that meet the minimum counts (n = 50) for both focal and reference groups were included only in the DIF analysis. Across all operational items and DIF comparison groups, one to three items were classified 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 produce any serious issues with items.

5. ITEM CALIBRATION AND ESTABLISHING I AM SCALE

Item response theory (IRT; van der Linden & Hambleton, 1997) is used to calibrate all items and derive scores for all the Indiana's I AM items. 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, and items are often calibrated using a sample of students from within a state population.

5.1 ITEM RESPONSE THEORY MODELS

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

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

where Z represents the vector of item responses, and θ represents a student's true ability.

Traditional item response models differ only in the form of the function $P(Z)$. The one-parameter model (also known as the Rasch model) is used to calibrate dichotomously scored I AM items and takes the form

$$P(x_j = 1|\theta_k, b_j) = \frac{1}{1+e^{(\theta_k - b_j)}} = P_{j1}(\theta_k).$$

The b parameter is often called the *location* or *difficulty* parameter; the greater the value of b, the greater the difficulty of the item. The one-parameter model assumes that the probability of a correct response approaches zero as proficiency (θ_k-b_j) decreases toward negative infinity. In other words, the one-parameter model assumes that no guessing occurs. In addition, the one-parameter model assumes that all items are equally discriminating.

For items that have multiple, ordered response categories (i.e., partial credit items), I AM items are calibrated using the Rasch family Masters' (1982) partial credit model. Under Masters' model, the probability of a response in category i for an item with m_j categories can be written as

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

5.2 ESTABLISHING THE I AM BANK

Item banks for I AM assessments were calibrated twice following the close of the Spring 2019 test window. The item parameters used for the IDR meeting were produced from the initial calibration. The final item parameters used to score the tests were produced from the second calibration. The procedures of each calibrations are briefly described as follows.

Pre-IDR calibration:

- Step 1: Freely calibrated all the items in each test excluding (1) ISTAR items that have legacy parameters and (2) the ISTAR items that were administered in the tiers (segments 4–6) only and have no legacy parameters. The calibration was centered on the mean item difficulty of all the items being included in the calibration.
- Step 2: Anchored on the item parameters obtained in Step 1 and calibrated the ISTAR items that were administered in the tiers (segments 4–6) only and have no legacy parameters.

Post-IDR calibration and linking:

To establish the I AM scale for these assessments, all the items that survived the IDR were re-calibrated. In this run, the ISTAR items that have legacy parameters and were administered in segments 4–6 only were excluded. The calibration was centered on the mean item difficulty of all the items being included in the calibration. The mean-mean equating method was used to put the legacy ISTAR items administered in Part 1, segment 1 and 2, onto the I AM scale. Applying the same linking constants, the legacy ISTAR items administered in the tiers were also put onto the I AM scale. The post-IDR calibrated and linked parameters were used for the final scoring.

5.3 IRT ANALYSES RESULTS

Table 24 displays the number of students contributing to the I AM IRT model.

Table 24: N Students Used in I AM Calibration

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10	Biology
ELA	641	695	746	898	920	1038	1024	
Mathematics	640	694	748	890	921	1047	1019	
Science		696		880				960
Social Studies			740					

5.3.1 IRT Summaries

The IRT statistical properties of the final operational test forms used for the I AM are summarized in Table 25 through Table 28.

Table 25: Operational Item Parameter Five-Point Summary and Range, ELA

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	b	-1.07	-0.87	-0.25	0.09	0.36	0.90	1.60
4	b	-1.57	-1.06	-0.59	-0.02	0.25	0.65	1.60
5	b	-1.77	-1.29	-0.75	-0.17	0.15	0.53	1.30
6	b	-1.55	-1.16	-0.49	0.00	0.36	0.80	1.00
7	b	-1.53	-1.01	-0.35	-0.04	0.38	0.57	1.59
8	b	-1.57	-1.19	-0.67	-0.21	0.10	0.84	1.43
10	b	-1.54	-1.39	-0.73	-0.19	0.17	0.88	1.09

Table 26: Operational Item Parameter Five-Point Summary and Range, Mathematics

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	b	-1.37	-1.13	-0.53	-0.05	0.29	0.95	1.38
4	b	-2.37	-0.74	-0.37	-0.13	0.30	1.17	2.00
5	b	-1.31	-0.80	-0.48	-0.10	0.36	0.74	0.80
6	b	-1.66	-1.09	-0.62	-0.05	0.38	0.85	1.66
7	b	-1.64	-0.68	-0.37	0.03	0.44	0.70	0.99
8	b	-1.17	-0.84	-0.46	-0.17	0.12	0.94	1.06
10	b	-0.97	-0.71	-0.30	0.11	0.49	0.97	1.04

Table 27: 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	-1.51	-1.17	-0.49	0.00	0.47	0.96	1.85
6	b	-2.00	-1.12	-0.46	0.02	0.42	0.93	1.23
Biology	b	-1.62	-1.34	-0.69	0.10	0.33	1.26	1.38

Table 28: 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	b	-1.75	-0.94	-0.36	-0.01	0.33	0.59	0.71

5.3.2 I AM Test Characteristic Curves (2019)

Another way to view the technical properties of the I AM test forms is via the test characteristic curves (TCCs). These plots are displayed in Appendix C with three-tier TCCs for each test.

6. SCORING AND REPORTING

6.1 MAXIMUM LIKELIHOOD ESTIMATION

The abilities were estimated using maximum likelihood estimation (MLE). Parameter estimates were calibrated using the Rasch model for dichotomously scored items and Masters' partial credit model for polytomous items.

6.1.1 Likelihood Function

The likelihood function for generating the MLEs is based on a mixture of item types, including MC (typically worth one point) and non-MC (often worth more than one point but scored for integer partial credit), and can therefore be expressed as

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N} \left[\frac{1}{1 + exp[-D(\theta - b_i)]}\right]^{x_i} \left[1 + \frac{1}{1 + exp[-D(\theta - b_i)]}\right]^{1-x_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N} \frac{exp \sum_{k=1}^{X_i} D(\theta - \delta_{ki})}{\sum_{j=1}^{m_i} exp \sum_{k=1}^{j} D(\theta - \delta_{ki})}$$

where b_i is the location (i.e., difficulty) parameter, x_i is the observed response to the item, i indexes item, and δ_{ki} is the k^{th} step for item i with m total categories.

We subsequently find the optimal point to maximize the log-likelihood as the student's theta (i.e., MLE) given the set of items administered to the student.

6.1.2 Derivatives

Finding the MLE requires an iterative method, such as Newton-Raphson iterations. Because the log-likelihood is a monotonic function of the likelihood, the following derivatives based on the log-likelihood function (with Rasch constraints) are used:

$$\frac{\partial ln L(\theta)^{MC}}{\partial \theta} = \sum_{i=1}^{N} \left\{x_i - \left[\frac{1}{1 + exp[-(\theta - b_i)]}\right]\right\}$$

$$\frac{\partial ln L(\theta)^{CR}}{\partial \theta} = \sum_{i=1}^{N} \left\{x_i - \left[\frac{\sum_{j=1}^{m_i} jexp \sum_{k=1}^{x_i} (\theta - \delta_{ki})}{1 + \sum_{j=1}^{m_i} exp \sum_{k=1}^{x_i} (\theta - \delta_{ki})}\right]\right\}$$

$$\frac{\partial^2 lnL(\theta)^{MC}}{\partial\theta^2} = -\sum_{i=1}^{N} \left(1 - \left[\frac{1}{1+exp[-(\theta - b_i)]}\right]\right)\left[\frac{1}{1+exp[-(\theta - b_i)]}\right]$$

$$\frac{\partial^2 lnL(\theta)^{CR}}{\partial\theta^2} = \sum_{i=1}^{N} \left[\left(\frac{\sum_{j=1}^{m_i} jexp\sum_{k=1}^{x_i}(\theta - \delta_{ki})}{1+\sum_{j=1}^{m_i} exp\sum_{k=1}^{x_i}(\theta - \delta_{ki})}\right)^2 - \left(\frac{\sum_{j=1}^{m_i} j^2 exp\sum_{k=1}^{x_i}(\theta - \delta_{ki})}{1+\sum_{j=1}^{m_i} exp\sum_{k=1}^{x_i}(\theta - \delta_{ki})}\right)\right]$$

Hence, the estimated MLE is found via the following maximization routine:

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

where

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

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

and where θ_t denotes the estimated θ at iteration t.

6.1.3 Standard Errors of Estimate

The standard error of the MLE is estimated by

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

where $I(\theta_j) = -\left(\frac{\partial^2 lnL(\theta)}{\partial^2\theta}\middle|\theta = \theta_j\right)$ is the Fisher information at the MLE and is calculated as follows:

$$\frac{\partial^2 log\left([p(\theta)]^{Z_i}[q(\theta)]^{1-Z_i}\right)}{\partial^2\theta} = -D^2 p_i(\theta) q_i(\theta)$$

where

$$q_i = 1 - p_i.$$

In general, the second derivate for the i^{th} Masters' partial credit model item is

$$\frac{\partial^2 log\left(P(\theta)\right)}{\partial\theta^2} = D^2 \frac{\left[\sum_{j=1}^{m_i} \quad jexp\left(\sum_{k=1}^{j} \quad D(\theta - b_{ki})\right)\right]^2}{\left[1 + \sum_{j=1}^{m_i} \quad exp \sum_{k=1}^{j} \quad D(\theta - b_{ki})\right]^2} - D^2 \frac{\left[\sum_{j=1}^{m_i} \quad j^2 exp\left(\sum_{k=1}^{j} \quad D(\theta - b_{ki})\right)\right]}{\left[1 + \sum_{j=1}^{m_i} \quad exp \sum_{k=1}^{j} \quad D(\theta - b_{ki})\right]}.$$

6.1.4 Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded and an MLE cannot be generated. For the I AM scoring, the extreme cases were handled according to the following guidelines:

i. Score all incorrect and all correct cases by either adding or subtracting 0.3 to/from an item score.
ii. Generate MLE for every other case and apply the following rule:
    a. If MLE is lower than –4, assign theta to –4.
    b. If MLE is higher than 4, assign theta to 4.

These truncated lowest observable theta (LOT) and highest observable theta (HOT) and the associated scale scores for each grade and subject are provided in Table 29.

Table 29: Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates

Subject	Grade	Lowest Observable Theta	Highest Observable Theta	Lowest Observable Scale Score (LOSS)	Highest Observable Scale Score (HOSS)
ELA	3–8 & 10	–4	4	1300	1700
Mathematics	3–8 & 10	–4	4	2300	2700
Science	4, 6, & Biology	–4	4	3300	3700
Social Studies	5	–4	4	4300	4700

6.1.5 Standard Error of LOT/HOT Scores

The standard error for LOT and HOT was computed using the LOT and HOT ability estimates given the administered items. For example, in the formula in Standard Error of Measurement (see Section 6.1.3), $\hat{\theta}$ = LOT or HOT, and difficulties (*b*) are for the administered items.

6.2 Transforming Theta Scores to Reporting Scale Scores

For Spring 2019, the I AM scale scores were reported for each student who took the ELA, Mathematics, Science, and Social Studies assessments. The scale scores were based on the operational items presented to the student and did not include any field-test items.

The scale score is the linear transformation of the IRT ability estimate using the scaling constants a and b shown in Table 30:

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

Scale scores are reported and compared as integers, with their decimal digits rounded down.

Table 30: Scaling Constants on the Reporting Metric

Subject	Grade	Slope (a)	Intercept (b)
ELA	3–8 & 10	50	1500
Mathematics	3–8 & 10	50	2500
Science	4, 6, & Biology	50	3500
Social Studies	5	50	4500

The summary of the I AM 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.

6.3 Overall Performance Classification

Each student was assigned an overall performance category in accordance with his or her overall scale score. Table 31 through Table 34 provide the scale score range for performance standards for ELA, Mathematics, Science, and Social Studies. The lower bound of the level 3, Proficient, marks the minimum cut score for proficiency.

Table 31: Proficiency Levels for ELA by Grade

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
3	1300 – 1463	1464 – 1481	1482 – 1700
4	1300 – 1478	1479 – 1497	1498 – 1700
5	1300 – 1474	1475 – 1488	1489 – 1700
6	1300 – 1466	1467 – 1486	1487 – 1700
7	1300 – 1485	1486 – 1497	1498 – 1700
8	1300 – 1464	1465 – 1490	1491 – 1700
10	1300 – 1467	1468 – 1505	1506 – 1700

Table 32: Proficiency Levels for Mathematics by Grade

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
3	2300 – 2462	2463 – 2473	2474 – 2700
4	2300 – 2461	2462 – 2478	2479 – 2700
5	2300 – 2459	2460 – 2470	2471 – 2700
6	2300 – 2461	2462 – 2477	2478 – 2700
7	2300 – 2466	2467 – 2477	2478 – 2700
8	2300 – 2463	2464 – 2474	2475 – 2700
10	2300 – 2470	2471 – 2484	2485 – 2700

Table 33: Proficiency Levels for Science by Grade

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
4	3300 – 3475	3476 – 3496	3497 – 3700
6	3300 – 3465	3466 – 3488	3489 – 3700
Biology	3300 – 3471	3472 – 3502	3503 – 3700

Table 34: Proficiency Levels for Social Studies by Grade

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
5	4300 – 4488	4489 – 4499	4500 – 4700

6.4 REPORTING CATEGORY SCORES

Reporting category scores were computed using all items for scoring within each reporting category for categories that have at least a minimum of seven (7) items in the blueprint. There were 58 items rejected in the IDR; due to the loss of these items, Mathematics grade 10 Tier 3 students did not receive the subscore of Equations and Inequalities (EI), Science grade 4 Tier 2 students did not receive the subscore of Investigating (I), and braille Science grade 6 Tier 3 students did not receive the subscore of Analyzing, Interpreting, and Computational Thinking (AICT), because there are only six items available in the reporting categories after IDR. The reporting category scores were computed as a percent-correct score for each student and an average percent-correct score for aggregate reporting.

Table 35: Reporting Category Scores Not Reported

Test	Tier	Subscore
Online/Paper Mathematics 10	Tier 3	Equations and Inequalities (Linear and Systems)
Online/Paper Science 4	Tier 2	Investigating
Paper Science 6	Tier 3	Analyzing, Interpreting, and Computational Thinking

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. Procedures that cannot be reasonably automated are replicated by two independent analysts at AIR.

7.1 SCORING QUALITY CHECK

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

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

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.

Camilli, G., & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.

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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–173.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small Sample Studies to Detect Flaws in Item Translations, *International Journal of Testing, 1*(2), 115-135.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning, *Educational Research and Evaluation, 19*(2-3), 170-187.

Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, *40*(2):106–108.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Indiana's Alternate Measure

2018–2019

Volume 2
Test Development

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Hyesuk Jang, Yuan Hong, Katherine Krehbiel, Hashim Evans, and Celine Bryan. 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

Indiana's Alternate Measure (I AM) is the summative accountability assessment for students with significant cognitive disabilities in grades 3–8 and high school. The I AM assessment is a stage-adaptive assessment. Performance on Part 1 of the assessment determines placement into one of three Part 2 forms: Tier 1 (low complexity); Tier 2 (moderate complexity); or Tier 3 (high complexity). The assessment measures student achievement and growth according to Indiana's Alternate Standards, called Content Connectors, which are aligned to and derived from the Indiana Academic Standards (IAS). I AM assesses:

- English/Language Arts: Grades 3–8 and 10
- Mathematics: Grades 3–8 and 10
- Science: Grades 4 and 6, and Biology End-of-Course Assessment (ECA) (high school)
- Social Studies: Grade 5

In June of 2018, the Indiana State Board of Education approved the adoption of new Content Connectors for English/Language Arts (ELA), Mathematics, Science, and Social Studies. Various stakeholders planned, designed, and facilitated the review, revision, and development of the Content Connectors. These alternate academic standards are designed to measure the knowledge and skills of students with significant cognitive disabilities and who are assessed with an alternate assessment. A systematic process was followed to ensure they are appropriately aligned to IAS and readily available to teachers, parents, and students across the state. Alternate standards are necessary to ensure all students have access to grade-level-aligned content and to achieve educational accountability.

1.1 CLAIM STRUCTURE

The I AM assessments are designed to support the following claims about proficiency:

Level 1 – Below Proficiency

Indiana students below proficiency have not met current grade-level Content Connectors. Students may require significant support to develop the knowledge, application, and skills to be on track for post-secondary education or competitive integrated employment.

Level 2 – Approaching Proficiency

Indiana students approaching proficiency have nearly met current grade-level Content Connectors by demonstrating some basic knowledge, application, and skills. Students may require support to be on track for post-secondary education or competitive integrated employment.

Level 3 – At Proficiency

Indiana students at proficiency have met current grade-level Content Connectors by demonstrating essential knowledge, application, and skills to be on track for postsecondary education or competitive integrated employment.

1.2 Organization of This Volume

This volume is organized into three subsequent sections:

- An overview of the I AM blueprint development process that supports the validity of the claims that I AM assessments are designed to support
- An overview of the item development process that supports the validity of the claims that I AM assessments are designed to support
- An overview of the test construction process for the I AM assessments that supports the validity of the claims that I AM assessments are designed to support

2. I AM Assessment Blueprints

The I AM assessments are designed to measure student achievement of the Indiana Content Connectors. The Indiana Content Connectors were designed as an extension to the IAS and adopted by the Indiana State Board of Education to measure the knowledge and skills of students with significant cognitive disabilities. To ensure that the I AM assessments appropriately measure the knowledge and skills of the I AM student population, assessment blueprints were constructed to represent the range of content defined in the Indiana Content Connectors. This ensures that the assessments result in an accurate classification of student achievement. The I AM assessments are designed to support the claims about proficiency described in this volume in Section 1.1, Claim Structure.

This section describes the development of I AM assessment blueprints that yield valid and reliable assessment scores and proficiency-level classifications to indicate whether students have demonstrated the knowledge and skills associated with the Indiana Content Connectors. The details in this section support the claim that the blueprints are technically sound and consistent with expectations of current professional standards.

2.1 I AM Blueprint Development

The American Institutes for Research (AIR) worked closely with the Indiana Department of Education (IDOE) to create blueprints that guide the development process for the I AM assessments. Blueprints are the assessment design specifications that ensure assessment scores support the performance-level descriptors (PLDs) described in Section 1.1, Claim Structure. Blueprints specify the proportionality of how I AM assesses the Indiana Content Connectors, including the relative range of each Content Connector on the assessment as represented in the minimum and maximum number of items to be administered to each student.

AIR and IDOE recruited Indiana educators to inform I AM blueprint development in June 2018. These educators represented different regions of the state, diverse student populations, and content and accessibility expertise. Panels of content and special education educators serving students with significant cognitive disabilities were convened at each grade level where they recommended priorities and associated item ranges noted within the blueprints. Educators also considered the vertical articulation of content across grades 3–10.

To ensure that the I AM assessments provide valid assessment of the Content Connectors and align with the Indiana Learning Evaluation Assessment Readiness Network (ILEARN) assessments and expectations, the I AM assessment blueprints were constructed to include the range of content defined in the IAS represented on ILEARN and the aligned Content Connectors as appropriate for the I AM student population to achieve the result of the accurate classification of student achievement.

The workshop began with a large group session to orient participants to the workshop objectives and review the agenda activities to meet those objectives. IDOE oriented and standardized the participants in IDOE expectations.

During the large group session, discussion emphasized that blueprints that reflect the breadth of the subject area content domains, cognitive complexity, and vertical articulation across grades must be developed to ensure assessments align to the IAS (Content Connectors for the I AM population).

Participants then disseminated into grade-level groups.

In order to design blueprints that yield valid and reliable assessment scores and proficiency-level classifications that indicate whether students have demonstrated the knowledge and skills associated with the Content Connectors, the blueprint meeting participants began by reviewing the Content Connectors and identifying key evidence that demonstrated proficiency in each Content Connector.

Next, using the ILEARN reporting categories that were created by Indiana educators during the ILEARN workshops in February 2018, AIR and IDOE presented two documents for each content area to the participants, including:

1) A completed ILEARN blueprint for the content area and grade with the percentages and item minimums/maximums for the reporting categories and IAS for reference

2) A “blank” draft I AM blueprint for the content area and grade with all percentages and item minimums/maximums for the reporting categories and Content Connectors left blank for the participants to assign prioritization of each standard and determine critical importance of each standard for the I AM student population

Because grade 10 blueprints for ELA and Mathematics were not constructed by the ILEARN committees, participants used the ILEARN blueprints developed for grades 7 and 8 ELA and Mathematics as a reference point for the I AM grade 10 discussions. However, the grade 10 workshop participants were given wide latitude to change the

blueprint based on their discussions during the workshop sessions. Grade 10 ELA and Mathematics workshop participants received the following:

1) A completed ILEARN blueprint for the content area for grades 7 and 8 with the percentages and item minimums/maximums for the reporting categories and IAS for reference

2) A list of all Content Connectors in general blueprint form without reporting categories, prioritization, percentages, or item minimums/maximums listed so that participants could determine reporting categories, assign Content Connector priority, and determine critical importance for the I AM student population at grade 10.

Within each subject-area and grade-level panel, panelists worked independently to classify each reporting category as critically important (3), important (2), or less important (1) to demonstrate mastery of the Content Connectors at grade level. Panelists discussed and rationalized their priorities and came to a consensus about the weights of each reporting category. Once weights were determined, percentages were assigned by reporting category.

Next, subject-area panels convened to review the system of weighted reporting categories across the grade-level panels. The goal of the subject-area panel meeting was to ensure any shifts across grades were thoughtful and intended.

The next step was to classify the Content Connectors according to the relevance of the content being assessed within each of the reporting categories. Panelists worked in subject-area and grade-level groups to indicate Content Connectors that best inform the reporting category to those that provide less information for the reporting category.

Panelists first worked independently in Google Polls to classify each Content Connector as: (3) a standard that best informs the reporting category, (2) a standard that provides some information for the reporting category, or (1) a standard that provides little information for the reporting category to demonstrate mastery of the reporting category. After making individual, initial classifications, AIR staff tabulated the scores using Google Polls to show areas of consensus and areas of disagreement in real time. When a majority agreed (e.g., 4 out of 6 panelists) on a Content Connector's classification, that classification (3, 2, or 1) was assigned to the Content Connector. When there was disagreement about the priority of a standard, panelists discussed and rationalized their prioritization/classification and came to a consensus. The panel came to a majority decision about each of the classifications in a draft blueprint.

Next, all grade-level panels convened as one subject-area group to review the prioritized Content Connectors that emerged from the grade-level panels. The purpose of the entire subject-area group meeting was to ensure that any shifts in importance of Content Connectors across grade levels was thoughtful and intended.

Panels re-evaluated as a large group the previous proportions based on the review of individual Content Connectors, working toward the end goal of final blueprint percentages and determination of reporting category weights.

Following the close of the workshop, AIR worked to incorporate panelists' feedback in the development of public-facing blueprints for the I AM assessments. These were presented for IDOE review prior to a follow-up webinar with workshop participants.

Subject-area panels were reconvened via a follow-up webinar during the week of June 25, 2018. A separate webinar was held for each subject area to review the final draft blueprints to ensure that they met the intent of the individual committees. A guided review of the draft blueprints illustrated how each of the blueprint elements was generated from the panelist feedback based on requirements of the assessment system, reporting framework, and their rating of the Content Connectors and reporting category weights.

Subject-area panels evaluated whether revisions should be made to the proposed grade-level blueprints that would better meet IDOE's assessment goals. At the conclusion of each webinar, participants confirmed that the recommended blueprints satisfied the requirements for I AM and that the I AM blueprints developed during the June 2018 meetings met the following objectives:

- Measure the breadth and depth of Indiana Content Connectors, aligned to and derived from the IAS
- Provide weight to the Content Connectors and reporting categories as identified by educators
- Produce accurate and precise test scores and performance-level classifications
- Meet required item count limits
- Remain consistent related to measurable content across test administrations

2.1.1 ELA Blueprints

The I AM blueprints developed for ELA grades 3–8 and 10 are provided in Appendix A: I AM English/Language Arts Blueprints.

The key features of the I AM ELA blueprints include reporting categories, reporting category allocations, Content Connectors, Content Connector allocations (number of minimum and maximum items per Content Connector), and the total number of operational items.

Reporting Categories

The I AM ELA blueprints are organized by reporting category and specify the number of items required for each reporting category ensuring that the form contains enough items from that category to elicit enough information from the student to justify reporting category-level scores. The I AM ELA grade 3 blueprint includes an additional reporting category for Reading Foundations.

Reporting categories comprise a broad domain or segment of the subject area identified by educators as containing meaningful sets of interrelated Content Connectors. Reporting categories are broad to allow for individual-level reporting of student performance. In

many cases, the reporting category combines two or more related domains, as indicated by educators.

The I AM ELA blueprints in grades 6–8 and 10 also include Speaking and Listening Content Connectors that contribute to the student score as a whole.

Reporting Category Allocations

The I AM ELA blueprints include the overall percentage of the assessment characterized by each reporting category. For ELA, educators placed an emphasis on Reading Foundations and literary text in grade 3, continued to emphasize literary text in grades 4 and 5, and then transitioned to more emphasis on nonfiction text in grades 6–8 and 10. On the I AM ELA assessment, the focus of reading is centered on comprehending text. To meet the varied needs of this population, reading takes several forms, including listening to text read aloud.

Content Connectors

The I AM ELA blueprints list the code for each Content Connector in each reporting category.

Content Connector Allocations

The I AM ELA blueprints also specify the number of minimum and maximum items per Content Connector. A Content Connector with a range that starts at 0 indicates that the Content Connector may not be assessed each year. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction.

Total Number of Operational Items

The total number of operational items on each I AM ELA assessment is 32.

2.1.2 Mathematics Blueprints

The I AM blueprints developed for Mathematics grades 3–8 and 10 are provided in Appendix B: I AM Mathematics Blueprints.

The key features of the I AM Mathematics blueprints include reporting categories, reporting category allocations, Content Connectors, Content Connector allocations (number of minimum and maximum items per Content Connector), and the total number of operational items.

Reporting Categories

The I AM Mathematics blueprints are organized by reporting category and specify the number of items required for each reporting category ensuring that the form contains enough items from that category to elicit enough information from the student to justify reporting category- level scores.

Reporting categories comprise a broad domain or segment of the subject area identified by educators as containing meaningful sets of interrelated Content Connectors. Reporting categories are broad to allow for individual-level reporting of student performance. In many cases, the reporting category combines two or more related domains, as indicated by educators.

The I AM Mathematics blueprints also include Content Connectors in a category that is reported as an aggregate score. The items assessing those Content Connectors will contribute to the student score as a whole.

Reporting Category Allocations

The I AM Mathematics blueprints include the overall percentage of the assessment characterized by each reporting category. For Mathematics, educators determined that all reporting categories should have equal emphasis in grades 3 and 4. For grades 5 and 6, educators placed an emphasis on Number Sense and transitioned to more emphasis on Algebra and Functions in grades 7–8. In grade 10, the educators determined that all reporting categories should have equal emphasis.

Content Connectors

The I AM Mathematics blueprints list the code of each Content Connector in each reporting category.

Content Connector Allocations

The I AM Mathematics blueprints also specify the number of minimum and maximum items per Content Connector. A Content Connector with a range that starts at 0 indicates that the Content Connector may not be assessed each year. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction.

Total Number of Operational Items

The total number of operational items on each on each I AM Mathematics assessment is 32.

2.1.3 Science Blueprints

The I AM blueprints developed for Science grades 4 and 6 and Biology are provided in Appendix C: I AM Science Blueprints.

The key features of the I AM Science blueprints include reporting categories, reporting category allocations, Content Connectors, Content Connector allocations (number of minimum and maximum items per Content Connector), and the total number of operational items.

Reporting Categories

The I AM Science blueprints are organized by reporting category and specify the number of items required for each reporting category ensuring that the form contains enough

items from that category to elicit enough information from the student to justify reporting category-level scores.

Reporting categories comprise a broad domain or segment of the subject area identified by educators as containing meaningful sets of interrelated Content Connectors. Reporting categories are broad to allow for individual-level reporting of student performance. In many cases, the reporting category combines two or more related domains, as indicated by educators.

Reporting Category Allocations

The I AM Science blueprints include the overall percentage of the assessment characterized by each reporting category. For grade 4 Science, educators determined that Questioning and Modeling was of greatest priority. For grade 6 Science, educators placed an emphasis on Investigating. In the Biology ECA, educators determined that Analyzing Data and Mathematical Thinking should receive the greatest emphasis.

Content Connectors

The I AM Science blueprints list the code of each Content Connector in each reporting category.

Content Connector Allocations

The I AM Science blueprints also specify the number of minimum and maximum items per Content Connector. A Content Connector with a range that starts at 0 indicates that the Content Connector may not be assessed each year. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction.

Total Number of Operational Items

The total number of operational items on each on each I AM Science assessment is 32.

2.1.4 Social Studies Blueprint

The I AM blueprint developed for Social Studies grade 5 is provided in Appendix D: I AM Social Studies Blueprint.

The key features of the I AM Social Studies blueprint include reporting categories, reporting category allocations, Content Connectors, Content Connector allocations (number of minimum and maximum items per Content Connector), and the total number of operational items.

Reporting Categories

The I AM Social Studies blueprint is organized by reporting category and specifies the number of items required for each reporting category, ensuring that the form contains enough items from that category to elicit enough information from the student to justify reporting category-level scores.

Reporting Categories comprise a broad domain or segment of the subject area identified by educators as containing meaningful sets of interrelated Content Connectors. Reporting categories are broad to allow for individual-level reporting of student performance. In many cases, the reporting category combines two or more related domains, as indicated by educators.

Reporting Category Allocations

The I AM Social Studies blueprint includes the overall percentage of the assessment characterized by each reporting category. For grade 5 Social Studies, educators placed an emphasis on Civics and Government/History.

Content Connectors

The I AM Social Studies blueprint lists the code of each Content Connector in each reporting category.

Content Connector Allocations

The blueprint also specifies the number of minimum and maximum items per Content Connector. A Content Connector with a range that starts at 0 indicates that the Content Connector may not be assessed each year. The item ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction.

Total Number of Operational Items

The total number of operational items on each on the I AM Social Studies assessment is 32.

2.1.5 Test Length

As indicated, the I AM assessments include 32 operational items. The 2019 I AM test design also includes 15 embedded field-test (EFT) items placed into fixed positions at the end of Part 1 (segment 3) and the end of Part 2 (segment 7). The EFT items in segment 7 differed based on the Part 2 tier. All test forms contain fixed operational items but vary with respect to the EFT items. Table 1 shows the number of operational items, the number of EFT items, and the total number of items administered on each 2019 assessment.

Table 1: Observed Spring 2019 Test Length by Grade and Subject

Subject	Grades	Number of Operational or Operational Field-Test Items	Number of EFT Items Per Test	Total Items Per Test
ELA	3–8, 10	32	15	47
Mathematics	3–8, 10	32	15	47
Science	4 & 6 & Biology ECA	32	15	47
Social Studies	5	32	15	47

2.1.6 Reporting Category Percentages

The blueprint is designed to support reporting at the Reporting Category level in addition to the overall test score. Individual scores for each Reporting Category provide information to help identify areas in which a student may have had difficulty.

Tables 2–5 provide the percentage of operational items required in the blueprints by Reporting Category for each grade level by subject. The percentages represent an acceptable range of item counts.

Table 2: Blueprint Percentage of Items Assessing Each Reporting Category in ELA

Grade	Reporting Category			
	Key Ideas and Textual Support/Vocabulary	***Structural Elements and Organization/Connection of Ideas/Media Literacy***	***Writing***	***Reading Foundations***
3	22–31%	22–25%	22–25%	22–31%
4	34–41%	31–38%	22–25%	N/A
5	34–44%	28–38%	22–28%	N/A
	Key Ideas and Textual Support/Vocabulary	***Structural Elements and Organization/Connection of Ideas/Media Literacy***	***Writing***	***Speaking and Listening (Aggregate Only)***
6	28–38%	25–34%	22–25%	3–6%
7	28–44%	25–34%	22–25%	3–6%
8	28–44%	25–34%	22–25%	3–6%
10	28–38%	25–34%	22–25%	3–6%

Table 3: Blueprint Percentage of Items Assessing Each Reporting Category in Mathematics

Grade	Reporting Category				
	Algebraic Thinking and Data Analysis	***Computation***	***Geometry and Measurement***	***Number Sense***	***Process Standards (Aggregate Only)***
3	22–25%	22–25%	22–25%	22–25%	6–12%
4	22–25%	22–25%	22–25%	22–25%	6–12%
	Algebraic Thinking	***Computation***	***Geometry and Measurement, Data Analysis, and Statistics***	***Number Sense***	***Process Standards (Aggregate Only)***
5	22–25%	22–25%	22–25%	25–28%	3–12%
	Algebra and Functions	***Computation***	***Geometry and Measurement, Data Analysis, and Statistics***	***Number Sense***	***Process Standards (Aggregate Only)***
6	25–28%	22–25%	22–25%	25–28%	3–12%
	Algebra and Functions	***Data Analysis, Statistics, and Probability***	***Geometry and Measurement***	***Number Sense and Computation***	***Process Standards (Aggregate Only)***
7	25–28%	22–25%	22–25%	22–25%	3–6%
8	28–31%	22–25%	22–25%	22–25%	3–6%
	Equations and Inequalities (Linear and Systems)	***Functions (Linear and Non-linear)***	***Geometry and Measurement***	***Number Sense and Data Analysis***	***Process Standards (Aggregate Only)***
10	22–25%	22–25%	22–25%	22–25%	3–12%

Table 4: Blueprint Percentage of Items Assessing Each Reporting Category in Science

Grade	Reporting Category			
	Analyzing, Interpreting, and Computational Thinking	***Explaining Solutions, Reasoning, and Communicating***	***Investigating***	***Questioning and Modeling***
4	22–25%	22–25%	22–25%	25–34%
6	22–25%	22–25%	25–34%	22–25%
	Analyzing Data and Mathematical Thinking	***Communicating Explanations and Evaluating Claims Using Evidence***	***Developing and Using Modeling to Describe Structure and Function***	***N/A***
Biology	40–50%	22–25%	28–37%	N/A

Table 5: Blueprint Percentage of Items Assessing Each Reporting Category in Social Studies

Grade	Reporting Category		
	Civics and Government/History	***Economics***	***Geography***
5	50–56%	22–25%	22–25%

3. I AM ITEM POOL AND DEVELOPMENT

In order for the I AM assessments to yield valid and reliable assessment scores and proficiency-level classifications, the needs formalized by the I AM assessment blueprints guide the I AM item pool development. The I AM item pool consists of two source types: legacy operational items from the Indiana Standards Tool for Alternate Reporting (ISTAR) and newly-developed items for field testing.

3.1 LEGACY OPERATIONAL ITEMS

In order to support blueprint and test design requirements as new items for the I AM pool were developed and field tested, legacy operational items that aligned to the new Indiana Content Connectors and that met I AM blueprint needs were retained for operational use on the 2018–2019 I AM assessments. Items were also evaluated and selected for alignment to the new 2018 I AM item specifications when possible. However, because there is variance between the item specifications in use when the legacy operational items were developed and the new I AM item specifications, full alignment of the legacy operational items to the new I AM item specifications was not possible. For future administrations, it is the intention to replace legacy operational items with new I AM items as they become available for operational use so that full alignment of the I AM assessments with the new item specifications is eventually possible.

3.2 FIELD-TEST ITEMS

In order to begin growing the I AM operational pool, AIR and IDOE developed new items for field testing based on blueprint needs and that fully aligned to the new Content Connectors and item specifications.

AIR completed a preliminary legacy operational pool analysis in June 2018 based on metadata indicating alignment to the IAS. Based on this preliminary analysis, AIR created I AM item development plans and developed new items that targeted the depth and breadth of coverage required by the test blueprints, with the intent to grow the pool over time.

I AM field-test item development was guided by a rigorous, structured process that engaged stakeholders at critical junctures. 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 item change and comment. When either the internal

AIR reviewers or stakeholders (during committee meetings) review items in ITS, they can see the items as they will appear to the student, with all accessibility features and tools available.

3.2.1 Item Types

The majority of the I AM items are Multiple-Choice (MC). Six I AM Science items are of a different type: Five are Multiple-Select (MS) items and one is a Table Match item (MI).

Table 6 lists the I AM item types in the I AM item bank and provides a brief description of each. Examples of each item type can be found in Appendix E: Item Type Examples.

Table 6: I AM Item Types and Descriptions

Response Type	Description
Multiple-Choice (MC*)	Student selects one correct answer from a number of options.
Multiple-Select (MS) (Science only)	Student selects all correct answers from a number of options.
Table Match (MI) (Science only)	Student checks a box to indicate if information from a column header matches information from a row.

*Note that the abbreviations MC, MS, and MI correlate to the attributes used in AIR's Item Tracking System (ITS).

3.2.2 Underlying Principles Guiding Item Development

I AM item development is based on the needs formalized by the I AM assessment blueprints and is guided by detailed item specifications. The specifications, discussed in Section 3.2.4, describe the interaction types that can be used, provide guidelines for targeting the appropriate cognitive engagement, offer suggestions for controlling item difficulty, and offer sample items.

Items are written with the goal that virtually every item be accessible to all students within the designated population, 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 System (TDS), which has received Web Content Accessibility Guidelines (WCAG) 2.0 AA certification, offers a wide array of accessibility tools, and is compatible with most assistive technologies.

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

We seek to ensure that the items measure the standards in a fair and meaningful way by engaging educators and other stakeholders at each step of the process. Educators evaluate the alignment of items to the standards and item specifications and offer guidance and suggestions for improvement. They also participate in the review of items for accessibility and fairness.

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

The process is guided by passage and item specifications, and includes

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

Each of these steps has a role in ensuring that the items can support the claims that will be based on them. Table 7 describes how each step contributes to these goals. Each step in the process is discussed in more detail below the table.

Table 7: Summary of How Each Step of Development Supports the Validity of Claims

Item Development Step	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, passage topics, content limits, Depth of Knowledge (DOK), and guidelines for meeting tier requirements.	Avoids the use of any item types with accessibility constraints, provides language guidelines.	
Selection and training of item writers	Ensures that item writers have the background to understand the unique needs of the alternate student population, as well as specific details related to standards and specifications.	Training in language accessibility and fairness prevents the introduction of unnecessary barriers.	
Writing and internal review of items	Checks content and tier alignment; evaluates and improves overall quality.	Eliminates editorial issues; flags and removes bias and accessibility issues.	
Markup for translation and accessibility features		Adds text-to-speech to reduce barriers.	Adds text-to-speech and Spanish translations.
Review by state personnel and stakeholder committees	Checks content and tier alignment; evaluates and improves overall quality.	Flags sensitivity issues.	
Field testing	Provides statistical check on quality; flags issues.	Flags items that appear to function differently for subsequent review.	May reveal usability or implementation issues with markup.
Post field-test reviews	Final, more focused check on flagged items.	Final, more focused review on items flagged for differential item function.	

3.2.3 I AM Passage Specifications

I AM English/Language Arts (ELA) development begins with passage specifications. Detailed passage specifications ensure that all passages align to the correct grade level and provide sufficient complexity and appropriate subject matter.

Passage specifications for ISTAR were developed by educators in the summer of 2017. These passage specifications were used to review passages for the I AM assessment by educator stakeholders in collaboration with IDOE content experts and AIR content experts during a Passage Review workshop in August 2018. At the end of this workshop, participants affirmed through an end-of-workshop survey that the ISTAR passage specifications elicited passages that are appropriate for the I AM student population and were therefore appropriate for continued use as I AM passage specifications.

Using the following tools and resources, passages for the I AM ELA assessments are evaluated quantitatively for content and vocabulary:

- Lexile® Framework for Reading
- ATOS® Readability Formula
- Flesch-Kincaid Grade Level
- EDL Core Vocabularies

The Lexile® Framework for Reading was developed by MetaMetrics, Inc., and employs a scientific formula to calculate the Lexile level of a text based on the semantic and syntactic elements of a text.

The ATOS® readability formula takes into account the most important predictors of text complexity, which are average sentence length, average word length, and word difficulty level. The results are provided in a grade-level scale.

The Flesch-Kincaid Grade Level measures sentence length by the average number of words in a sentence and word length by the average number of syllables in a word to provide a U.S. grade level of education that an average student would require to be able to understand the text.

The EDL Core Vocabularies resource is used for all grades to determine the readability of vocabulary words. The EDL is composed of words introduced in reading instruction and found on frequency lists. This resource is used to determine the grade level when selecting vocabulary to assess.

Table 8 provides the quantitative specifications for I AM passages by grade for word count, Lexile range, Flesch-Kincaid range, and ATOS range.

Table 8: I AM Quantitative Passage Specifications

I AM Grade(s)	Max Word Count	Lexile Range	Flesch-Kincaid Range	ATOS Range
3	250	300–740	1.5–2.0	1.5–2.8
4–5	280	300–820	1.5–5.7	2.0–4.8
6–8	300	300–925	2.0–6.5	2.5–6.0
10	350	400–1050	2.3–7.0	2.8–6.5

Each I AM passage is also evaluated qualitatively. The complexity of the passages is reduced through the three tiers, from most complex to least complex (Tier 3 being the most complex and Tier 1 being the least complex). It is assumed that students have experience with text in their grade spans or those of earlier grade spans.

Table 9 provides the qualitative specifications for passages by tier.

Table 9: I AM Qualitative Passage Specifications

Tier 1	Tier 2	Tier 3
● **Passage topic** is grade and age appropriate. ● **Sentences** are short and use primarily simple structure, with concrete language and clearly connected pronouns. ● **Passage** is comprised of high-frequency, commonly used vocabulary. ● **Topic** is directly stated and supported with concrete details. ● **Dialogue** is either not used or limited, with no more than one or two people speaking in brief interactions. ● **Illustrations** are used to support the concepts in the passage (typically, 2–3 throughout text, appearing before any associated text). ● **Text features** have simple information with limited detail. ● **Figurative language**, if assessed, is simple. ● **Assessed vocabulary** is two or more grades below the assessed grade.	● **Passage topic** is grade and age appropriate. ● **Sentences** may include compound subjects and predicates and introductory phrases. ● **Passage** is comprised of mostly high frequency, commonly used vocabulary and some basic subject-specific vocabulary. ● **Topic** may be directly stated or require simple inferences. ● **Dialogue** is limited, with two people speaking in brief interactions. ● **Images** are sometimes used to support the concepts in the passage (typically one right below title). ● **Text features** have information with few details. ● **Figurative language**, if assessed, is simple. ● **Assessed vocabulary** is two or more grades below the assessed grade.	● **Passage topic** is grade and age appropriate. ● **Sentences** may be a mix of simple and compound structures, as well as some complex constructions. ● **Passage** includes some common expressions, controlled vocabulary, and some subject-specific language. ● **Topic** may include more inferential concepts and themes with multiple characters. ● **Dialogue** may include two or more people speaking. ● **Images** are sometimes used to support the concepts in the passage (typically one right below title). ● **Text features** have information with complex ideas. ● **Figurative language**, if assessed, is simple. ● **Assessed vocabulary** is two or more grades below the assessed grade.

These quantitative and qualitative specifications help test developers create passages that will support appropriate difficulty. The specifications are used in subsequent reviews by IDOE and panelists during committee reviews.

3.2.4 I AM Item Specifications

Item specifications guide the I AM item development process. In July 2018, Indiana educators met to develop item specifications for the new 2018 Content Connectors for ELA, Mathematics, Science, and Social Studies.

The I AM item specifications were designed to provide guidance on how to construct valid and reliable items aligned to the Content Connectors.

The item specifications were developed specifically for the I AM student population to ensure that the I AM assessments provide a valid assessment of the Content Connectors

and align with the I AM assessment blueprints. This allows the I AM assessments to provide an accurate classification of student achievement.

Using evidence statements, educators analyzed the Content Connectors for various dimensions outlined on the Item Specification templates.

The workshop began with a large group session to orient participants to the workshop objectives and review the agenda activities to meet those objectives. IDOE oriented and standardized the participants in IDOE expectations.

The large group session focused on helping panelists understand that to ensure assessments align to the Content Connectors, Item Specifications must be developed that reflect the breadth of the subject area content domains, cognitive complexity, and vertical articulation across grades.

Next, subject-area panels convened. Each subject-area group completed two item specification templates as preparation and training for the grade-level work that followed. Discussion was guided by AIR facilitators and IDOE.

In grade-level groups, the participants worked in smaller 3-member groups to develop the item specifications for all Content Connectors assessed on the I AM Blueprints for their grade and subject area. Item specifications were completed based on educator discussions by AIR facilitators and IDOE. The small groups will be given a designated number of item specifications to complete before reconvening with the larger group.

At designated checkpoints, participants completed peer reviews of the sections they had developed to that point. This was critical to ensure that grade-level expectations were met, each grade/grade-band working group was consistent in their approach to writing item specifications, and that grade-level specific content limits were respected.

Following the initial completion of item specifications by grade-level panels, the entire subject area reconvened to review the work done in the grade-level panels. Each break-out group presented their work for the full subject-area panel to review for consistency across the subject area. Modifications were made by the note-takers to match the panelists' discussions. An AIR/IDOE content matter expert facilitated.

Following close of the workshop, AIR reviewed the teacher-crafted item specifications to ensure completeness, rigor, and accuracy. As part of that process, AIR developed any missing sample items as necessary, and these were included in the final item specification drafts that were reviewed and approved by IDOE.

Specifications for all assessed grades and subjects include the following:

- **Reporting Category.** This is the blueprint reporting category that the Content Connector is a part of for the I AM assessments.
- **Content Connector.** This includes the language and the coding used for the Content Connector (Indiana's alternate standards, aligned to and derived from IAS).
- **Indiana Academic Standard.** This includes the language and coding used for the IAS that the Content Connector is aligned to and derived from.

- **Content Limits.** This section denotes grade-level limitations for assessment. Content limits delineate what terms, concepts, or procedures are acceptable at a particular grade level for a particular standard—and in some cases what is not acceptable.
- **Recommended Response Mechanisms.** This section identifies the ways in which students may respond to a prompt.
- **Construct-Relevant Vocabulary.** This section lists any key vocabulary that can be used in the item.
- **Cognitive Complexity (DOK).** This section indicates a number between 1 and 6. The number corresponds to the Links for Academic Learning (LAL) DOK model, which has six cognitive complexity levels to account for the differentiated needs and abilities of the special education population. DOK represents cognitive complexity and is defined for each Content Connector. Items are to match the recommended DOK of the Content Connector to which it is aligned.
- **Evidence Statements.** Because students with significant cognitive disabilities are a diverse population with a variety of needs, I AM items are classified into one of three tiers. Generally, Tier 1 items are less complex than Tier 2 items, and Tier 2 items are less complex than Tier 3 items. The I AM item specifications include an evidence statement for each tier. Evidence statements describe the knowledge and skills that an assessment item elicits from students.
    - **Tier 1** questions and answer choices include low structural-level items with a range of item difficulty and complexity. Graphics are provided for most answer choices, along with text, which give students a visual support to answer the questions.
    - **Tier 2** questions and answer choices include medium structural-level items with a range of item difficulty and complexity. They may include more introductory phrases in the questions and fewer graphics in answer choices than in Tier 1. They also include a greater level of complexity in how students respond to the questions than in Tier 1.
    - **Tier 3** questions and answer choices include high structural-level items with a range of item difficulty and complexity. There is more text and few to no graphics in the answer choices. There may be more abstract ideas and inferencing. There is more complexity in how students respond to the questions than in Tier 2.
- **Accessibility and Accommodation Considerations.** This section provides guidance regarding graphics, as well as auditory and visual considerations.
- **Sample Item.** In this section, a sample item is provided along with its corresponding tier.

Table 10 presents a sample ELA specification for one grade 3 Content Connector.

Table 10: Sample ELA Specifications for Grade 3

Reporting Category	Key Ideas & Textual Support/Vocabulary
Content Connector	**3.RN.2.2.a.1:** Determine the main idea of a text.
IAS Standard	**3.RN.2.2:** Determine the main idea of a text; recount the key details and explain how they support the main idea.
Content Limits	Items must be passage based. Tier 1 and 2 items should avoid the word "best" in the stem. Tier 1 items should contain picture support in answer choices when possible to aid comprehension. Tier 2 items can contain picture support in answer choices. Tier 3 items should not contain picture support. Tier 1 distractors should demonstrate clearly incorrect understanding of events or details in the passage. Tier 2 distractors should be possible misunderstanding of events or details in the passage or unrelated details or events in the passage. Text complexity will increase with Tiers.
Recommended Response Mechanisms	Multiple Choice (MC) Table Match (TM) Multi-Select (MS)
Construct-Relevant Vocabulary	main idea
Cognitive Complexity	4
Evidence Statements	
Evidence Statements	**Tier 1** Students can identify a key detail in the text.
	Tier 2 Students can identify an explicitly stated main idea of the text.
	Tier 3 Students can determine the main idea of a text.
Accessibility and Accommodation Considerations	
Stimulus Graphic Limitations	Stimulus graphics will be limited to clear photos, illustrations, diagrams, tables, and charts that directly relate to the passage topic. Information contained within stimulus graphics is ineligible for assessment unless specifically prescribed by Content Connector and/or evidence statements.
Visual and Auditory Considerations	Graphics will be provided in formats that are accessible to students to understand or process information. Graphics that do not contribute to the student's understanding should not be included.

Sample Item	
Tier 3	[Stimulus: Passage about the history of telephones] Which sentence tells the main idea? A. No one uses telephones anymore. B. Telephones are a lot bigger than they used to be. **C. Telephones have changed a lot over the years.**

At the time of item specification development, available item types for the Recommended Response Mechanisms section of the I AM item specifications included two-, three-, or four-option MC; five-option MS; and table match. For Mathematics only, numeric/equation response was also considered an available item type. IDOE and AIR conducted a cognitive laboratory study in the fall of 2018 to learn more about how students taking I AM interact with different item types. For the I AM student population, three-option MC was recommended as the most appropriate response mechanism. Based on the results of this study, I AM item specifications were edited to remove references to item types no longer being considered for I AM from evidence statements and sample items. The edits to the evidence statements and sample items were approved by educator committees. Please note, however, that additional item types were retained in the Recommended Response Mechanisms section for further consideration based on future studies that may occur.

All newly developed I AM items align to the 2018 I AM item specifications. Legacy operational items on the 2018–2019 I AM assessments were selected for "best fit" to the new 2018 I AM Content Connectors and item specifications. However, because legacy operational items were developed prior to the creation of I AM item specifications, not all legacy operational items align fully to the I AM item specifications. Alignment of operational legacy items to the 2018 I AM Content Connectors was deemed sufficient when alignment to the new 2018 I AM item specifications was not possible. Future I AM administrations will continue to replace legacy operational items with new I AM items as the I AM pool depth and breadth increases, with ongoing efforts being made to align I AM administrations solely to the 2018 I AM item specifications.

3.2.5 Training of Item Writers

All AIR item writers developing I AM items have at least a bachelor's degree, and many have teaching experience. All item writers are trained in

- the principles of universal design;
- the avoidance of bias and sensitivity issues;
- language accessibility guidelines; and
- the I AM Passage and Item Specifications.

Key material is included as Appendix G: Language, Accessibility, Bias, and Sensitivity Guidelines and Checklist.

3.3 INTERNAL REVIEW

AIR's I AM assessment development structure utilizes highly effective units of test developers 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.

AIR 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 used is included here as Appendix F: Item Review Checklist. The I AM Internal Review cycle includes five levels:

- Preliminary Review
- Content Review 1
- Special Education Review
- Edit Review 1
- Senior Review 1

3.3.1 Preliminary Review

First, items are written independently by test developers. For the 2018–2019 I AM administration, ELA items associated with literary or nonfiction passages were written after the passages had been reviewed and approved by educators (Passage Review is described in more detail in Section 3.4.2). After items are written by test developers, the items go through Preliminary Review. The Preliminary Review is conducted by team leads or senior content staff. Sometimes, the Preliminary Review is conducted in a group setting, led by a senior test developer. During the Preliminary Review process, test developers, either individually or as a group, analyze items to ensure the following:

- The item aligns with the academic standard.
- The item matches the item specifications 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 LAL DOK level.
- The vocabulary used in the item is appropriate for the grade and subject matter.
- The item considers language accessibility and is fair to all students.
- 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 make clear what is being asked, it is stated positively, and it does not rely on negatives—such as no, not, none, never—unless absolutely necessary).

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

3.3.2 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. He or she 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 his or her own experience in test development. If substantive changes are deemed to be necessary, this reviewer rejects the item or sends the item back to a test developer with requested changes and then reviews the item again.

3.3.3 Accessibility Review

The accessibility reviewer reviews and revises items to make sure they not only meet the content standards but are also as accessible as possible to students across a wide spectrum of cognitive and physical disabilities. If the accessibility reviewer has concerns about the accessibility of an item, the item gets sent back to the Content Review 1 review level for revision.

3.3.4 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 items 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 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.

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 MC items, editors check that options are parallel to the extent possible in structure and fit logically

and grammatically with the stem; they also confirm that the key accurately and correctly answers the question as posed, is not inappropriately obvious, and is the only correct answer to an item among the distractors.

3.3.5 Senior Review 1

By the time an I AM item arrives at Senior Review 1, it has been thoroughly vetted by both content reviewers and editors. Senior reviewers (i.e., 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, alignment to the standard, and consistency with the expectations for the highest quality.

3.4 Review by State Personnel and Stakeholder Committees

All I AM items have been through an exhaustive external review process. I AM items in the item bank are reviewed by IDOE content experts, and then reviewed and approved by a stakeholder committee that evaluates content, accessibility, bias/fairness, and sensitivity.

3.4.1 State (Client) Review

After items have been developed in the I AM item bank, state content experts review all items prior to committee review. At this stage in the review process, clients can request edits, such as wording edits, scoring edits, or alignment/DOK updates. An AIR content lead reviews and implements client-requested edits that are aligned to I AM Content Connectors and item specifications. At this stage, items are ready for committee review.

3.4.2 Passage Review

For the 2018–2019 I AM administration, there was a separate review and acceptance process for passages that preceded item development. During the 2018 ELA Passage Review, passages were reviewed against the I AM Passage Specifications, which include criteria for passage quality, quantitative metrics for readability and grade-level appropriateness, accessibility, fairness, sensitivity, and bias.

Committee members were designed to include two subject matter experts, two administrators or instructional coaches, and two special education teachers or accessibility specialists.

Committee members accepted passages as they appeared, or they recommended revisions based on a quality criteria checklist.

A summary of the 2018 I AM Passage Review meeting appears in Appendix H: 2018 I AM Passage Review Committee Summary.

Future I AM administrations will forgo passage review as a separate step preceding item development. Passage Review is important with long passages with numerous

associated items to make sure the passage is acceptable before beginning work on developing associated items. With alternate assessments, however, passages are short with typically only 3–5 associated items. It is therefore more conducive to develop the passage while developing the items, which leads to the need for simultaneous passage/item development and review.

3.4.3 Content and Fairness Committee Review

During the Content and Fairness Committee Review, items are reviewed for content validity, grade-level appropriateness, and alignment to the content standards and item specifications. Committee members are typically grade-level and subject-matter experts, or they may include accessibility specialists or corporation/school-level administrators. During this review, committee members also review the items for bias, fairness, sensitivity, and accessibility.

Committee members accept items as they appear or they recommend revisions based on a quality criteria checklist.

A summary of the 2018 I AM Content and Fairness Committee Review meeting appears in Appendix I: 2018 I AM Content and Fairness Committee Review Summary.

3.5 Markup for Translation and Accessibility Features

After all approved state- and committee-recommended edits have been applied, the items are considered “locked” and ready for accessibility markup. 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 markups, whether for translations or text-to-speech, follow similar processes. A 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.

Currently, all I AM items are tagged with text-to-speech. I AM Mathematics, Science, and Social Studies are also tagged with Spanish translations.

3.6 Field Testing

Newly developed I AM items are field tested as embedded field-test items in the I AM assessment. Field testing is described in detail in Volume 1, Section 3.3, of this technical report.

3.7 Post-Field-Test Review

Following field testing, items are subject to additional reviews. These include:

- Key verification, for items that are key-scored
- Data review, for items that failed standard flagging criteria

We discuss each of these processes in the following sections.

3.7.1 Key Verification

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

3.7.2 Item Data Review

Volume 1**,** Sections 4.1 and 4.2, of this technical report describe in detail the statistical flags that send items to item data review. The flags are designed to highlight potential content weaknesses, miskeys, or possible bias issues.

I AM items that are field tested are flagged for review in the following areas:

- Item Quality and Performance
- Item Difficulty
- Differential Item Functioning

Item Quality and Performance

I AM MC items are flagged for item quality and performance if the correlation for the key is less than 25% and/or if the correlation for the distractor(s) is greater than 0.

I AM MS and table match items are flagged for item quality and performance if the correlation with test score is less than 25%.

Item Difficulty

I AM MC items are flagged for item difficulty if the percentage of students selecting the key is less than 25% or greater than 95% and/or if students select an incorrect option more often than they select the key.

I AM MS and table match items are flagged for item difficulty if the percentage of students achieving ANY score point is greater than 95%.

Differential Item Functioning (DIF)

To evaluate DIF, AIR evaluates the likelihood of correct responses between students in different groups who were matched on ability. With fair items, students of the same ability should have the same likelihood of responding correctly, regardless of group membership. When items are flagged for DIF, groups matched on ability have different likelihoods of responding correctly based on group membership only.

Committee members are 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: Item Data Review Training Material.

A summary of the 2019 I AM Item Data Review Committee meeting appears in Appendix K: 2019 I AM Item Data Review Committee Summary.

3.8 Strategy for Pool Evaluation and Replenishment

AIR seeks to release approximately six items per grade and subject from the pool each year for use in Indiana's Released Items Repository (RIR). To grow the operational pool each year, AIR intends to develop items to be included in six field-test slots on each content-area form. The total number of items on the field-test forms on each year's assessments from which these six items will be randomly selected for any one student is based on what the anticipated student population can support in order to ensure that each field-test item is administered to at least 200 students. The current I AM student population in grades 3–6 supports the development and testing of 18 field-test items per year (six items each in three forms). The current I AM student population in grades 7–10 supports the development and testing of 24 field-test items per year (six items each in four forms).

Our general strategy for item development planning gathers information from three sources, including:

1. Characteristics of released items to be replaced
2. Characteristics of legacy items to be replaced
3. Tabulations of content coverage to identify gaps in the pool

4. I AM Assessment Construction

4.1 Assessment Form Construction

During Fall 2018, AIR psychometricians and content experts worked with IDOE content specialists and leadership to build assessments for the Spring 2019 administration. The I AM assessments are designed to support the claims about proficiency described in this volume in Section 1.1. I AM assessments are built according to the assessment blueprints that were designed to yield valid and reliable test scores and proficiency-level classifications. I AM assessments are also built according to a test design that provides guidance for the stage-adaptive structure of the assessments.

In this section, the processes used for assessment construction are described to support the claim that they are technically sound and consistent with expectations of current professional standards. These processes include the utilization of a structured test design plan and collaborative participation from all parties.

4.1.1 Test Design

I AM is a stage-adaptive assessment administered in segments. In Part 1, all students take the same assessment form (20 operational items) across a range of complexities. Performance on this first set of items determines the next set of items received in one of three Part 2 forms (each contains 12 operational items): Tier 1 (low complexity); Tier 2 (moderate complexity); or Tier 3 (high complexity). Each Part 2 form (Tier 1, Tier 2, or Tier 3) contains unique items along with items from adjacent tiers. Performance on items from both parts is combined for the final summative scale scores. The overall scale scores for Indiana students align with three proficiency levels (Below Proficiency, Approaching Proficiency, and At Proficiency).

Table 11 illustrates the I AM test design for forms in each grade and subject.

Table 11: I AM Test Design 2018–2019

I AM Test Design 2018–2019			
Part 1	**Part 2**		
	Tier 1	**Tier 2**	**Tier 3**
item 1	item 21	item 21	item 30
item 2	item 22	item 22	item 31
item 3	item 23	item 23	item 32
item 4	item 24	item 30	item 36
item 5	item 25	item 31	item 37
item 6	item 26	item 32	item 38
item 7	item 27	item 33	item 39
item 8	item 28	item 34	item 40
item 9	item 29	item 35	item 41
item 10	item 30	item 36	item 42
item 11	item 31	item 37	item 43
item 12	item 32	item 38	item 44
item 13			
item 14			
item 15			
item 16	**Key**		
item 17	**Tier 1 item**		
item 18	**Tier 2 item**		
item 19	**Tier 3 item**		
item 20			

Part 1 is administered to all students. Performance in Part 1 determines placement into one of three Part 2 forms: Part 2 Tier 1, Part 2 Tier 2, or Part 2 Tier 3.

Parts 1 and 2 have a total of 32 operational items.

As shown in Table 11, 44 unique operational items are needed for form building. This is due to the cross-tier linking pattern in the Part 2 forms. Each Part 2 form (Part 2 Tier 1, Part 2 Tier 2, Part 2 Tier 3) contains unique items along with items from adjacent tiers.

4.1.2 Operational Field-Test Items

Because of the new 2018 Content Connectors, blueprints, and item specifications, not all of the Indiana legacy operational items were retained for operational use on the new I AM assessments. For the 2018–2019 I AM assessments, newly developed items were used in an operational field-test role where needed to meet the depth and breadth of the blueprint requirements. To the extent possible, the 2019 EFT items included items that

matched the same blueprint coverage as the operational field-test (OFT) items. This provided redundancy to help ensure against the potential loss of operational field-test items during Item Data Review.

Future I AM assessment administrations will not include operational field-test items.

Tables 12–19 show how many operational items (called "OP Legacy Items" in the tables) and how many OFT items (called "OFT New I AM Items" in the tables) were on each form by grade and subject. Two tables are provided for each subject. The first table is for the assessments as constructed. Because OFT items were included in the 2018–2019 assessments, Item Data Review (IDR) needed to be conducted before a final determination on the usability of the OFT items for scoring could be made. Therefore, post-IDR tables are also included to show the number of OFT items used for scoring. Despite including EFT items that provided redundant coverage, full coverage was not possible due to pool limitations. Therefore, in the cases highlighted below, the number of OFT items post-IDR was slightly lower than the numbers as constructed.

Table 12: Number of OP Legacy Items and OFT New I AM Items in ELA (as constructed)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
3	22	10	22	10	22	10
4	24	8	24	8	23	9
5	24	8	24	8	24	8
6	23	9	24	8	24	8
7	23	9	23	9	22	10
8	23	9	23	9	23	9
10	25	7	25	7	23	9

Table 13: Number of OP Legacy Items and OFT New I AM Items in ELA (post-IDR)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
3	22	10	22	10	22	10
4	24	8	24	8	23	9
5	24	8	24	8	24	8
6	23	9	24	8	24	8
7	23	9	23	9	22	10
8	23	9	23	9	23	9
10	25	7	25	7	23	9

Table 14: Number of OP Legacy Items and OFT New I AM Items in Mathematics (as constructed)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
3	22	10	24	8	23	9
4	21	11	23	9	22	10
5	21	11	22	10	20	12
6	23	9	23	9	23	9
7	21	11	21	11	20	12
8	23	9	23	9	22	10
10	19	13	21	11	18	14

Table 15: Number of OP Legacy Items and OFT New I AM Items in Mathematics (post-IDR)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
3	22	10	24	8	23	8
4	21	11	23	9	22	10
5	21	11	22	10	20	12
6	23	9	23	9	23	9
7	21	11	21	11	20	12
8	23	9	23	9	22	10
10	19	13	21	11	18	13

Table 16: Number of OP Legacy Items and OFT New I AM Items in Science (as constructed)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
4	25	7	24	8	22	10
6	20	12	21	11	18	14
Biology	25	7	25	7	24	8

Table 17: Number of OP Legacy Items and OFT New I AM Items in Science (post-IDR)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
4	25	6	24	5	22	7
6	20	11	21	9	18	12
Biology	25	7	25	7	24	8

Table 18: Number of OP Legacy Items and OFT New I AM Items in Social Studies (as constructed)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
5	20	12	21	11	20	12

Table 19: Number of OP Legacy Items and OFT New I AM Items in Social Studies (post-IDR)

Grade	Tier 1		Tier 2		Tier 3	
	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items	OP Legacy Items	OFT New I AM Items
5	20	10	21	9	20	10

4.2 Test Form Assembly

As already discussed, the I AM assessment blueprints describe the content to be covered and the allocations for Reporting Categories and Content Connectors. To assemble the 2019 test forms, AIR content specialists selected operational items to represent the blueprint for each grade and subject. Where Indiana legacy operational items were not available, newly-developed I AM items were selected as OFT items.

4.2.1 Role of the AIR Content Team

AIR ELA, Mathematics, Science, and Social Studies content teams were responsible for the initial form construction and subsequent revisions. AIR content teams performed the following tasks:

- Selection of the operational items
- Selection of the OFT 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
- Revision of the forms to incorporate feedback from IDOE

4.2.2 Role of the AIR Technical Team

The AIR technical team, which included psychometricians and statistical support associates, prepared the item bank by updating the ITS with current item statistics and provided test construction training to the internal content team. During test construction, at least one psychometrician was facilitating with each content area. The technical team performed the following tasks:

- Preparing item bank statistics and updating of AIR's ITS
- Creating 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
- Creating statistical summary and materials for IDOE review

4.2.3 Role of the IDOE

The IDOE assessment and content specialists reviewed and approved selected items and forms provided by AIR. Feedback provided by IDOE was addressed in subsequent rounds by AIR until all I AM forms were approved by IDOE.

4.3 Blueprint Match

The I AM assessment blueprints are designed to support reporting at the Reporting Category level in addition to the overall test score. Individual scores for each Reporting Category provide information to help identify areas in which a student may have had difficulty.

Table 20, Table 22, Table 24, and Table 26 provide the percentage of test items assessing each reporting category that appeared on the Spring 2019 forms as constructed (pre-IDR).

Table 21, Table 23, Table 25, and Table 27 provide the percentage of test items assessing each reporting category after IDR.

Table 20: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in ELA (as constructed)

Grade	Reporting Category			
	Key Ideas and Textual Support/Vocabulary	***Structural Elements and Organization/Connection of Ideas/Media Literacy***	***Writing***	***Reading Foundations***
3	30%	20%	21%	29%
4	35%	22%	43%	N/A
5	48%	21%	31%	N/A
	Key Ideas and Textual Support/Vocabulary	***Structural Elements and Organization/Connection of Ideas/Media Literacy***	***Writing***	***Speaking and Listening (Aggregate Only)***
6	31%	20%	39%	9%
7	32%	21%	36%	11%
8	35%	22%	35%	8%
10	31%	28%	33%	8%

Table 21: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in ELA (post-IDR)

Grade	Reporting Category			
	Key Ideas and Textual Support/Vocabulary	***Structural Elements and Organization/Connection of Ideas/Media Literacy***	***Writing***	***Reading Foundations***
3	32%	20%	22%	27%
4	35%	22%	43%	N/A
5	49%	21%	30%	N/A
	Key Ideas and Textual Support/Vocabulary	***Structural Elements and Organization/Connection of Ideas/Media Literacy***	***Writing***	***Speaking and Listening (Aggregate Only)***
6	31%	20%	39%	9%
7	32%	21%	36%	11%
8	36%	23%	34%	7%
10	31%	28%	33%	8%

Table 22: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in Mathematics (as constructed)

Grade	Reporting Category				
	Algebraic Thinking and Data Analysis	***Computation***	***Geometry and Measurement***	***Number Sense***	***Process Standards (Aggregate Only)***
3	26%	35%	15%	18%	6%
4	36%	13%	28%	16%	8%
	Algebraic Thinking	***Computation***	***Geometry and Measurement, Data Analysis, and Statistics***	***Number Sense***	***Process Standards (Aggregate Only)***
5	33%	20%	29%	14%	4%
	Algebra and Functions	***Computation***	***Geometry and Measurement, Data Analysis, and Statistics***	***Number Sense***	***Process Standards (Aggregate Only)***
6	16%	33%	15%	33%	3%
	Algebra and Functions	***Data Analysis, Statistics, and Probability***	***Geometry and Measurement***	***Number Sense and Computation***	***Process Standards (Aggregate Only)***
7	34%	26%	13%	21%	6%
8	26%	24%	25%	18%	6%
	Equations and Inequalities (Linear and Systems)	***Functions (Linear and Non-linear)***	***Geometry and Measurement***	***Number Sense and Data Analysis***	***Process Standards (Aggregate Only)***
10	19%	30%	19%	29%	3%

Table 23: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in Mathematics (post-IDR)

Grade	Reporting Category				
	Algebraic Thinking and Data Analysis	***Computation***	***Geometry and Measurement***	***Number Sense***	***Process Standards (Aggregate Only)***
3	25%	34%	16%	19%	6%
4	35%	14%	26%	17%	9%
	Algebraic Thinking	***Computation***	***Geometry and Measurement, Data Analysis, and Statistics***	***Number Sense***	***Process Standards (Aggregate Only)***
5	33%	20%	28%	15%	4%
	Algebra and Functions	***Computation***	***Geometry and Measurement, Data Analysis, and Statistics***	***Number Sense***	***Process Standards (Aggregate Only)***
6	18%	32%	15%	32%	3%
	Algebra and Functions	***Data Analysis, Statistics, and Probability***	***Geometry and Measurement***	***Number Sense and Computation***	***Process Standards (Aggregate Only)***
7	33%	27%	13%	22%	6%
8	25%	25%	25%	19%	6%
	Equations and Inequalities (Linear and Systems)	***Functions (Linear and Non-linear)***	***Geometry and Measurement***	***Number Sense and Data Analysis***	***Process Standards (Aggregate Only)***
10	19%	31%	19%	28%	3%

Table 24: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in Science (as constructed)

Grade	Reporting Category			
	Analyzing, Interpreting, and Computational Thinking	***Explaining Solutions, Reasoning, and Communicating***	***Investigating***	***Questioning and Modeling***
4	20%	23%	25%	32%
6	23%	20%	26%	32%
	Analyzing Data and Mathematical Thinking	***Communicating Explanations and Evaluating Claims Using Evidence***	***Developing and Using Modeling to Describe Structure and Function***	***N/A***
Biology	33%	25%	43%	N/A

Table 25: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in Science (post-IDR)

Grade	Reporting Category			
	Analyzing, Interpreting, and Computational Thinking	***Explaining Solutions, Reasoning, and Communicating***	***Investigating***	***Questioning and Modeling***
4	22%	24%	22%	32%
6	24%	19%	24%	32%
	Analyzing Data and Mathematical Thinking	***Communicating Explanations and Evaluating Claims Using Evidence***	***Developing and Using Modeling to Describe Structure and Function***	***N/A***
Biology	34%	25%	41%	N/A

Table 26: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in Social Studies (as constructed)

Grade	Reporting Category		
	Civics and Government/History	***Economics***	***Geography***
5	54%	18%	28%

Table 27: Observed Spring 2019 Percentage of Items Assessing Each Reporting Category in Social Studies (post-IDR)

Grade	Reporting Category		
	Civics and Government/History	***Economics***	***Geography***
5	51%	20%	30%

In almost every case, the percentages across reporting categories in the Spring 2019 forms met the required blueprint range. The grades and subjects where blueprint requirements were not met are listed in Table 28.

Table 28: I AM 2019 Blueprint Discrepancies

Status	***Subject***	***Grade***	***Reporting Category***	***Blueprint***		***Tier 1***	***Tier 2***	***Tier 3***
				Minimum	***Maximum***			
Pre-IDR	ELA	10	KITS	9	12	12	13 (+1)	12
Post-IDR	ELA	10	KITS	9	12	12	13 (+1)	12
Post-IDR	Mathematics	10	EI	7	8	8	7	6 (–1)
Post-IDR	Science	4	I	7	8	7	6 (–1)	7

It is the intention for future I AM assessments to fully meet the I AM blueprint requirements. To ensure that the item pool can support the blueprint needs, annual item development plans will be developed based on a pool analysis against blueprint needs. The Item Development Plans for ELA, Mathematics, Science, and Social Studies for the 2019–2020 I AM administration are provided in Appendices L-O.

Developing and maintaining a robust operational pool aligned to I AM blueprint requirements will allow for future I AM assessment administrations to continue to yield valid and reliable test scores and proficiency-level classifications that indicate whether students taking the I AM assessment have demonstrated the knowledge and skills associated with the Indiana Content Connectors.

5. 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.

Fisher, D., Frey, N., & Lapp, D. (2012). *Text complexity: Raising rigor in reading*. Newark, DE: International Reading Association.

Flowers, C., Wakeman, S., Browder, D.M., & Karvonen, M. (2009). Links for academic learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice*, 28(1), 25–37.

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: 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), 21–22.

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.

Taylor, S.E. (1989). *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies.* Orlando, FL: Steck-Vaughn.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved 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's Alternate Measure

2018–2019

Volume 3
Test Administration

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Hyesuk Jang, Yuan Hong, Katherine Krehbiel, Hashim Evans, and Celine Bryan. 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 students with significant cognitive disabilities for operational use beginning with the 2018–2019 school year. This new assessment program, referred to as Indiana's Alternate Measure (I AM), replaced the Indiana Standards Tool for Alternate Reporting (ISTAR) in English/Language Arts (ELA), Mathematics, Science, and Social Studies. I AM is a two-stage adaptive assessment that comprises ELA and Mathematics assessments for grades 3–8 and 10, Science assessments for grades 4 and 6, a Social Studies assessment for grade 5, and the Biology End-of-Course assessment. In 2018–2019 both stages of all tests were administered online. Print and large print accommodations were available for students who could not access the assessment online. Braille was offered; however, no students taking I AM in 2018–2019 required the braille accommodation.

As specified in Standard 6.0 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), assessment instruments are required to have established test administration procedures to support useful interpretations of score results. This volume of the I AM Technical Report provides details on the testing procedures, accommodations, Test Administrator (TA) training and resources, and test security procedures implemented for I AM. Specifically, it provides the following test administration-related evidence for the validity of the assessment results:

- A description of the student population that takes the I AM
- A description of the training and documentation provided to TAs in order for them to follow standardized administration procedures
- A description of available 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 to mitigate loss, theft, and test content reproduction of any kind
- A description of the American Institute for Research's (AIR's) Quality Monitor (QM) system and the test irregularity investigation process to detect cheating, monitor real-time item quality, and evaluate test integrity

2. TESTING PROCEDURES AND TEST WINDOWS

Administering the 2018–2019 I AM required coordination, detailed specifications, and proper training. In addition to these efforts, several individuals were involved in the administration process, from those setting up testing environments to those administering the tests. Without the proper training and coordination of these individuals, the standardization of test administration could have been compromised. The Indiana Department of Education (IDOE) worked with AIR to develop and provide the training and documentation necessary for the successful administration of I AM under standardized conditions within all testing environments. The I AM test window was April 8 through May 17, 2019.

The accommodations available for eligible students participating in the I AM assessments are described in both the *I AM Test Administrator's Manual* (TAM) (Appendix M), located on the Indiana Assessment Portal, and the *2018–2019 Indiana Accessibility and Accommodations Guidance* (Appendix Y), located on the IDOE website. Both documents were available before and during testing.

For eligible students participating in the computer-based I AM ELA, Mathematics, Social Studies, Science, and Biology assessments, the accommodations made available are described in the *I AM Online Test Delivery System (TDS) User Guide* (Appendix T), which was accessible before and during testing through the Indiana Assessment Portal.

All students were required to take a subject-specific practice test within the operational test environment prior to taking the operational Spring 2019 I AM. Students who were administered the paper I AM form completed the practice test items in the paper-and-pencil test booklet. The practice tests contained sample test items and helped students become familiar with the test system's functionality, if applicable, and item types. Indiana alternate assessment students and TAs also had the opportunity to interact with released, non-secure items on a public-facing Released Item Repository (RIR) assessment that is available on the Indiana Assessment Portal. The I AM RIR was deployed on October 1, 2018, allowing students access to the items online for seven months prior to the opening of the test window.

Before a student could enter the operational test environment to participate in the subject-specific practice test and operational test, TAs had to complete the Learner Characteristics Inventory (LCI) for each student (see Appendix U).

I AM is a stage-adaptive assessment administered in two segments, where a student's answers in Segment 1 determined the next group of items presented to the student in Segment 2. Each I AM assessment included 32 operational or operational field-test (OFT) items that were used for scoring, as well as 15 embedded field-test (EFT) items. The student's total score is based on performance from both segments of the assessment.

The I AM assessments were untimed and were delivered to students individually. Students could start and finish one segment of an assessment in a single day or over the course of multiple days, if needed. TAs were advised, however, that students could not complete Segment 1 and begin Segment 2 on the same day. Both segments of an assessment had to be completed within two weeks of beginning the assessment.

2.1 ELIGIBLE STUDENTS

Students with significant cognitive disabilities who met the criteria to participate in the alternate assessments, as defined by Title 20 of the Indiana Code and federal law, participated in I AM.

Students eligible to participate in I AM were required to take the assessments appropriate for the grade level/subject in which they were receiving instruction. These students represented the following groups:

- **Public School Students, including Charter School Students.** Indiana public school and charter school students who met the participation criteria to participate in the alternate assessment and were enrolled in tested grade levels/subjects were required to participate in I AM.
- **Private School Students.** Indiana private school students who met the participation criteria to participate in the alternate assessment and were enrolled in tested grade levels/subjects were required to participate in I AM.
- **Accredited Nonpublic School Students.** Indiana students who attended accredited nonpublic schools and who met the participation criteria to participate in the alternate assessment and were enrolled in tested grade levels/subjects were required to participate in I AM.
- **Choice School Students.** Indiana Choice school students who met the participation criteria to participate in the alternate assessment and were enrolled in tested grade levels/subjects were required to participate in I AM.
- **Home Education Program Students.** Students who met the participation criteria to participate in the alternate assessment and 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 I AM assessment as a selected measure of their child's annual progress, students could participate in an I AM administration, as directed by the Corporation Test Coordinator (CTC).
- **English Learners (ELs).** All ELs participated in statewide assessments. ELs who were enrolled in school in the United States for less than one year could be exempt for one administration from the I AM ELA assessments if a student's EL team agreed that exemption was appropriate.
- **Students with Disabilities.** Indiana has established procedures to ensure the inclusion for testing of all public elementary and secondary school students with disabilities. Federal and state law 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 four general options:

    1. Indiana Learning Evaluation Readiness Network (ILEARN) without accommodations

2. ILEARN with approved accommodations
3. I AM without accommodations
4. I AM with approved accommodations.

A student's Case Conference Committee (CCC) determined, based on the criteria provided and the student's individual and unique needs, whether a student with disabilities participated in general education assessments with or without testing accommodations, or in the alternate assessment with or without accommodations. A student was eligible to participate in I AM in lieu of ILEARN if the CCC determined the student met the following criteria:

- Review of student record indicates a disability that significantly impacts intellectual functioning and adaptive behavior. Adaptive behavior is defined as essential for a person to live independently and function safely in daily life.
- The student requires extensive, repeated, individualized instruction and support that is not of a temporary nature.
- The student uses substantially adapted materials and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate, and transfer skills across multiple settings.
- Goals listed in the IEP for this student are linked to the enrolled grade-level Alternate Academic Standards (Indiana Content Connectors).

2.2 TESTING ACCOMMODATIONS

Students participating in the computer-based I AM were able to use the standard online testing features in the Test Delivery System (TDS). Before testing, TAs were able to select an alternate background and font color, mouse pointer size and color, and font size. During the assessments, 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 or masking function, or use the online basic Desmos calculator.

All I AM assessments had appropriate accommodations available to make these options accessible to students with significant cognitive disabilities who required additional accommodations, per the student's IEP. Online accommodations included permissive mode (to use assistive technology) and streamline mode. As an accommodation, students could also participate in I AM by using a paper-and-pencil test booklet, a large print test booklet, or a braille test booklet.

The I AM assessments provided three categories of assessment features to students. These included universal tools, designated features, and accommodations. Section 3.2 in Volume 1 of this technical report lists the allowed accommodations and the number of students who were provided with accommodations during the Spring 2019 test administration.

Table 1 provides a list of universal tools, designated features, and accommodations that were offered in the Spring 2019 administration. Universal tools are accessibility features of the TDS that are delivered either digitally (i.e., embedded) or separately (i.e., non-embedded). Designated features for I AM 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 *I AM Online Test Delivery System (TDS) User Guide*, available through the Indiana Assessment Portal (and included as Appendix T), provides instructions on how to access and use these features.

Table 1: Universal Tools, Designated Features, and Accommodations Available in Spring 2019

Universal Tools (for all students)	Designated Features	Accommodations (available per IEP)
Embedded/Online		
Online calculator Expandable passages Highlighter Masking Strikethrough Text-to-Speech (required) Zoom in and zoom out for text and graphics Line Reader	Color contrast Masking Mouse pointer (size and color) Print size (zoom in and zoom out) Translations (stacked)	Permissive mode to use assistive technology devices Streamlined mode
Non-Embedded		
Headphones or noise buffers to block out distractions Low-tech assistive writing instrument Preferential seating Scratch paper, including lined or graph paper Handheld calculator for paper assessment Student tested individually	Color acetate film for paper assessment Assistive technology to magnify/ enlarge text and images Access to sound amplification Special furniture or equipment for viewing test Special lighting conditions Time of day for testing altered	Alternate indicator of response Calculator Multiplication table Paper test booklet Large print test booklet Read-aloud script for paper-and-pencil test booklet Hundreds chart Interpreter for American Sign Language Braille test booklet Read aloud to self Student provided access to own resources

IDOE also collected information about non-standard accommodation requests under a Special Requests section in the Test Information Distribution Engine (TIDE). These special requests required IDOE approval.

Students participating in I AM who required computer-based accommodations (e.g., permissive mode) were provided the opportunity to participate in practice activities for the statewide assessments with appropriate allowable accommodations. Computer-based

test settings and accommodations were required to be identified in TIDE before starting a test session. Some settings and accommodations could not be changed after a student started the test.

If a student used any accommodations during the test administration, this information was recorded by the TA in his or her required administration information.

Guidelines recommended for making accommodation decisions included the following:

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

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

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

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. IDOE provided a separate accessibility policy manual (*2018–2019 Indiana Assessments Policies, Administration, and Security Manual*, included as Appendix X of this technical report; the current manual is available online at https://www.doe.in.gov/sites/default/files/assessment/2019-2020-indiana-assessments-policy-manual.pdf, as a supplement to the TAMs, for individuals involved in administering assessments to students with accommodations.

For eligible students with IEPs who participated in I AM paper-based assessments, the following accommodations were available:

- Paper test booklet
- Large print test booklet

As noted earlier, braille test booklets were not created for the 2018–2019 administration due to the fact that no students taking I AM needed a braille test booklet accommodation.

For eligible students with IEPs who participate in computer-based I AM assessments, a comprehensive list of accommodations is included in the *Test Information Distribution Engine (TIDE) User Guide* (Appendix H of this report).

The Accessibility and Accommodations Guidance provides information about the available tools, supports, and accommodations that were available to students taking the I AM assessments. For further information, please refer to both the *2018–2019 Indiana Assessments Policies, Administration, and Security Manual* (Appendix X) and the *2018–2019 Indiana Accessibility and Accommodations Guidance* (Appendix Y).

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

3. ADMINISTRATOR TRAINING

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

3.1 IN-PERSON TEST ADMINISTRATOR TRAINING

TAs were required to attend one half-day training session in Indiana before administering the I AM. Before the Spring 2019 assessment administration, over a period of three weeks, AIR conducted 20 training sessions statewide on the 2018–2019 test administration. TAs who could not attend an in-person training session, and were excused by IDOE, had access to either a Moodle course training session or a live-stream of the February 25 training session conducted this year. These training sessions provided an overview of the alternate assessment and the online systems used during test administration. These online systems included the I AM Portal, the TDS, TIDE, and the Online Reporting System (ORS). During the training session, AIR used video vignettes, which included Indiana educators and students, to illustrate important concepts. Appendix S includes the PowerPoint presentation used during each training session.

All test administration personnel were required to attend the training when a session became available in their area of the state. If approved by IDOE, TAs could attend the live-stream session or complete the Moodle course in lieu of attending an in-person training. Appendix K: I AM Educator Brochure provides the dates of each training and the number of members trained in each training session or workshop. The trainings occurred at various dates from February 25 through March 22 of this year. The number of participants in these trainings varied from 61 to 186 per session, thereby training a total of 1,609 participants.

3.2 COMPUTER-BASED ADMINISTRATION

TAMs and guides were available online for school and corporation staff. The *I AM Online Test Delivery System (TDS) User Guide* (Appendix T) was designed to familiarize TAs with TDS and contains tips and screen captures throughout the text. The user guide contained

- 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 Indiana Secure Browser features and keyboard shortcuts.

The *User Support* sections of both the *I AM Online TDS User Guide* (Appendix T) and the *TIDE User Guide* (Appendix H) provide instructions to address possible technology challenges during test administration. The AIR Help Desk collaborated with IDOE to provide support to Indiana schools as they administered the state assessment.

The *I AM Online TDS User Guide* (Appendix T) provides 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 A) as well as the *Additional Configurations and Troubleshooting Guides* (Appendices B–F) provide 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 were summarized in the following sections.

Roles and Responsibilities in the Online Testing Systems

CTCs, STCs, and TAs each had specific roles and responsibilities in the online testing systems. See the *I AM Test Administrator's Manual (TAM)* (Appendix M) and the *I AM Test Coordinator's Manual (TCM)* (Appendix N) for their specific responsibilities before, during, and after testing.

Corporation Test Coordinators

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.

School Test Coordinators

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 were required to appear as eligible for that test in TIDE. See the *TIDE User Guide* (Appendix H) for more information.

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

Test Administrators

TAs administered the I AM assessments to students. Prior to administration of I AM, TAs completed the LCI (see Appendix U) for each student and administered a practice test session.

TAs were responsible for attending an in-person training, reviewing necessary user manuals and user guides to prepare the testing environment, and ensuring that students did not have access to books, notes, or electronic devices. They were required to administer the I AM assessment following the directions found in the *I AM Test Administrator's Manual (TAM)* (Appendix M). Any deviation in test administration was required to be reported by TAs to the STC, who was to report it to the CTC. Then, if necessary, the CTC was to report it to IDOE. TAs also ensured that only the resources allowed for specific tests were available and no additional resources were used during administration of the I AM assessments.

3.3 TEST ADMINISTRATION RESOURCES

The list of training resources for the Spring 2019 I AM test administration is provided in this section. These materials were all available online on the Indiana Assessment Portal. (PDFs of these six resources have also been included in this technical report as Appendices J, I, G, W, K, and R, respectively.)

1. **Understanding Indiana's Alternate Measure (I AM) Webinar Module.** This online module walks Indiana educators through the new I AM assessments to prepare educators for the Spring 2019 assessment.
2. **Test Information Distribution Engine (TIDE) Webinar Module.** This module provides a general overview of the AIR system, TIDE, and the features applicable to educators before, during, and after testing.
3. **Technology Requirements for Online Testing Webinar Module.** This module provides technology requirements for CTCs and STCs to ensure their testing devices are set up properly before testing.
4. **Online Reporting System (ORS) Webinar Module.** This module provides a general overview of the ORS, where student scores (including individual scores and aggregate scores) are displayed after students complete the I AM assessments.
5. **I AM Educator Brochure.** This brochure provides an overview of the new I AM assessment to prepare educators for the Spring 2019 assessment.
6. **I AM Training FAQs.** This document includes frequently asked questions from the I AM in-person training.

Table 2 presents the list of available user guides and manuals related to the I AM administration. These materials were all available on the Indiana Portal. (PDFs of these six publications have also been included in this technical report as Appendices A–F, H, M–Q, T, and V, respectively.)

Table 2: User Guides and Manuals

Resource	Description
Technology Setup for Online Testing Quick Guide	This document explains in four steps how to set up technology in Indiana corporations and schools.
Additional Configurations and Troubleshooting Guide for Android	This document contains additional configurations and troubleshooting for a school or corporation's network and Android workstations.
Additional Configurations and Troubleshooting Guide for Chrome OS	This document contains additional configurations and troubleshooting for a school or corporation's network and Chrome OS workstations.
Additional Configurations and Troubleshooting Guide for Linux	This document contains additional configurations and troubleshooting for a school or corporation's network and Linux workstations.
Additional Configurations and Troubleshooting Guide for Mac	This document contains additional configurations and troubleshooting for a school or corporation's network and Mac workstations.
Additional Configurations and Troubleshooting Guide for Windows	This document contains additional configurations and troubleshooting for a school or corporation's network and Windows workstations.
Test Information Distribution Engine (TIDE) User Guide	This user guide describes the tasks performed in TIDE for I AM assessments.
I AM Test Administrator's Manual (TAM)	This manual provides information on the policies and procedures surrounding the I AM assessments, as well as an overview of the specific roles and responsibilities required before, during, and after testing.
I AM Test Coordinator's Manual (TCM)	This provides an overview of I AM test administration activities intended for Test Coordinators.
Released Item Repository Quick Guide	This quick guide provides an overview of how to administer the I AM RIR.
Released Item Repository Answer Keys	These answer keys provide information on the items included in the RIR for each tested grade and content area.
I AM Practice Test User Guide	This user guide provides an overview of and supports TAs for students participating in the I AM Practice Test.
I AM Online Test Delivery System (TDS) User Guide	This user guide supports TAs who manage testing for students participating in the LCI and the I AM practice and operational tests.
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 I AM assessments.

Department Resources and Support

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

- A weekly newsletter was distributed via email to CTCs from the IDOE Office of Assessment every Monday. The newsletter was titled, "I AM Assessment Update," and included information on new announcements relevant to the I AM assessment, reminders of upcoming milestones, and a planning-ahead section that included important dates in the I AM program. The IDOE Office of Assessment contact

information was also available at the end of each weekly newsletter so that corporations could contact IDOE directly with any questions.

- A weekly newsletter was distributed via email to educators from the IDOE Office of the State Superintendent of Public Instruction every Friday. The newsletter was titled, "IN the Know," and included information on new announcements relevant to the I AM assessment, as well as updates from other offices in the IDOE. Access to various social media platforms, as well as information on accessing previous weekly updates, was also available at the end of each weekly newsletter.
- Communications via newsletter from either the Office of Assessment or the Office of the State Superintendent of Public Instruction took place on an "as needed" basis. These messages generally addressed specific issues that needed to be communicated quickly to administrators and teachers in the field or information that the IDOE wanted to ensure was clearly outlined due to its importance to the I AM program. An example of this would be a memo that IDOE sent to corporations in Fall 2018 containing extensive information on scheduling guidance for I AM to help guide schools in planning the I AM testing schedule. The distribution was to superintendents, principals, and school leaders.
- General information about the assessments (such as dates of test windows for all state-administered assessments) was posted on the IDOE Office of Assessment website. The Accessibility and Accommodations Guidance in the I AM Policy and Guidance section of the IDOE website was designed to address questions pertaining to accommodations and overall accessibility.
- The *2018–2019 Indiana Assessments Policies, Administration, and Security Manual* (Appendix X) was also posted on the IDOE Office of Assessment website. This manual discussed CTC and STC responsibilities regarding IDOE communication and monitoring of test administration. The manual provided guidance on students opting out of an assessment and specific categories of students; descriptions on the various roles of personnel involved in test administration; and what needs to be done before, during, and after test administration. The manual also discussed formal security and integrity training for school and corporation personnel as well as the different aspects surrounding test security.
- The *2018–2019 Indiana Accessibility and Accommodations Guidance* (Appendix Y) was also posted on the IDOE Office of Assessment website. This manual includes the guidelines for the selection, administration, and evaluation of accessibility supports for instruction and assessment of all students, including students with disabilities, ELs, ELs with disabilities, and students without an identified disability or EL status.

I AM Released Item Repository

The I AM RIR is a collection of non-secure items that are available to the public via the Indiana Assessment Portal and are intended to allow students, parents, and educators access to content that will be similar to what the student will encounter when taking the I

AM assessments. The I AM RIR was deployed on October 1, 2018 and remained available all year. An answer key for each grade and content area (Appendix P) accompanied the RIR, which provided educators the opportunity to see how their students were performing on the assessment and where educators might focus efforts to improve student performance before the administration of the I AM assessment.

I AM Practice Tests

The purpose of the practice tests was to familiarize students with the system, functionality, and item types that will be on the I AM tests. These training tests were not intended to guide classroom instruction. Users could also use the tutorials on each item and familiarize themselves with the different features and response instructions for each item type.

In order to administer the I AM practice tests, I AM TAs had to be certified. As a result, I AM TAs could administer the practice tests to students only after attending the in-person TA training and completing the LCI for each student. Computer-based practice tests were designed for use with the Indiana Secure Browser or a supported web browser. The Indiana Assessment Portal provided a list of supported web browsers and their versions for the practice tests. AIR's TDS delivered the training tests in secure mode and used the same test delivery engine as with the operational assessment.

The design of the Indiana Secure Browser ensures that students, teachers, and educators are familiar with the online testing system before operational testing begins. The same set of tools, accommodations, and embedded supports used for the operational assessments were available in the I AM practice test because both training tests and operational assessments were delivered through the same system. IDOE required all students to take the practice test prior to the operational I AM assessment.

Students taking the I AM on paper were also required to take a paper-and-pencil practice test prior to taking the operational I AM assessment. The required 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 students completed the paper practice test items before starting the operational I AM 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. If non-embedded accessibility supports are used, assessment security can become an issue when other test formats are used (e.g., large print) or when someone other than the student is allowed to see the test (e.g., interpreter, reader, scribe). To ensure test security and confidentiality, TAs need to keep testing materials in a secure place to prevent unauthorized access, keep all test content confidential and refrain from sharing information or revealing test content, and return all materials as instructed.

Some of the same considerations for test security apply to embedded accessibility supports. For example, ensuring that only authorized personnel have access to the test and that test materials are kept confidential is critical in technology-based assessments. In addition, it is important to guarantee that students cannot access any unauthorized programs, the Internet, saved data, or computer shortcuts while they are taking the assessment. In most cases, any specially required hardware devices and appropriate applications, such as switches, should be compatible with computer-delivered assessments. Prior to testing, educators should check device compatibility and make appropriate adjustments if necessary.

The test security procedures for I AM included the following:

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

TAs were trained on test security procedures via an in-person training as well as training webinars hosted by both AIR and IDOE. Before testing, educators were required to sign the Test Security and Integrity Agreement, acknowledging that they would adhere to test security procedures. Both AIR and IDOE are committed to ensuring that, going forward, the test security policies and procedures are clearly presented in the user guides and TAMs.

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 training for everyone involved in test administration, outlining appropriate incident-reporting procedures, detecting test irregularities, and planning for investigation and handling of test security violations.

Furthermore, every school corporation or other test administration location that administers tests under the Indiana assessment system must have a locally developed, written test security policy that is shared with staff. While IDOE does not require school board approval of this policy, corporations should follow local level practices to determine if this policy needs to be approved by the local school board. The corporation’s test security policy must

- specify that secure test materials should not be delivered to school buildings more than one week (preferably less) in advance of test administration;
- specify that teachers and other school staff members are not allowed access to secure materials (except for the TAM) more than four hours in advance of the test administration; and
- describe the entity's plan for ensuring the security of assessment materials during testing and storage of all secure assessment materials before, during, and after testing. All test materials should be stored at a central location under lock and key.

If a TA has reason to believe that a violation in test security has occurred, he or she should notify the STC immediately. The STC should then notify the CTC.

If he or she has reason to believe that a lapse in test security has occurred, the CTC must do the following:

- Submit a Testing Irregularity Report to the IDOE Office of Student Assessment
- Submit a Test Irregularity request in TIDE
- Securely transmit relevant evidence of irregularities via secure file transfer protocol after the incident occurred
- Maintain the confidentiality of all evidence and documentation related to test security investigations

If IDOE has reason to believe that a violation in testing security has occurred, it has an obligation to investigate the incident as soon as possible. Additionally, IDOE receives data forensic information from AIR after testing has concluded. Following a review of the data forensic analysis, IDOE contacts corporations where there may be a test security issue. Corporations and schools are required to comply with IDOE's requests for documentation and information relevant to their initial investigation. IDOE may involve the school corporation or conduct a separate investigation.

If the IDOE determined that an irregularity in test administration or security had occurred, IDOE notified the CTC in writing, indicating the status of the case. Depending on the severity of the incident and its potential impact to the IDOE program, actions may have included but were not limited to the following:

- Invalidation of student scores
- A requirement for corporations to complete documentation and conduct interviews to gather more details regarding any test sessions identified as concerns
- IDOE's requirement of additional action steps taken by the corporations

4.1 Security of Test Materials

Before test materials were finalized, items went through multiple reviews, including review by various committees. It was critical to maintain the security of test items during these

committee meetings. Items were accessed directly from AIR's secure Item Tracking System (ITS) for online committee meetings; thus, no printed copies needed to be transported to meetings. For any materials that had to be printed, the materials were printed on light green paper with each page marked as secure in the header and/or footer. Such printed materials were distributed to participants only after these individuals signed the AIR and state nondisclosure forms. AIR staff reviewed the security procedures with the committee members.

All test items, test materials, and student-level testing information were deemed secure and were required to be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of 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. Under no circumstances were students permitted to assist in either preparing secure materials before testing or in organizing and returning materials after testing. After any administration, initial or make-up, secure materials (e.g., test booklets and TA scripts) 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 to be removed from the school's campus overnight. 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.

Printed test booklets were shipped to each Indiana school corporation one week prior to the start of the test window. Corporations were required to return printed test materials to the vendor one week after the end of the test window. Due to the fact that these materials were in corporations for six weeks, the security of the test booklets was a critical responsibility.

CTCs were therefore required to develop, implement, and assess procedures for the secure storage, administration, and delivery of standardized test materials back to testing vendors by established deadlines. Failure by a school corporation or its employees to securely store, administer, and return all secure test materials by established deadlines was considered an integrity breach under 511 IAC 5-5-3, which may have resulted in an action under Indiana Code 20-28-5-7. While student responses would not be scored, schools were required to immediately return student answer booklets found more than one week after the pick-up date.

It is considered a testing security violation for an individual to fail to follow security procedures set forth by IDOE, and no individual was permitted to

- use another staff member's username and/or password to access vendor systems or administer tests;
- use a student's login information to access practice tests or operational assessments;
- review test questions prior to, during, or after test administration;

- give test takers access to test questions prior to testing;
- copy, reproduce, or use in any manner any portion of any secure assessment, for any reason;
- alter student answer documents (paper-and-pencil or online) prior to, during, or after testing;
- share or post actual or paraphrased test items/content or student responses in a public forum, social media, text, or email;
- comment on test content in a public forum, social media, text, or email;
- take pictures, snapshots, or videos of assessment materials;
- deviate from the prescribed administration procedures specified in the TAM;
- make answer keys available to test takers;
- score student responses on the test locally before submitting the assessment for scoring to the test contractor, as designated by IDOE; or
- participate in, direct, aid, counsel, assist, encourage, or fail to report any of the acts prohibited in this section.

All schools and corporations were expected to treat all special document test materials (print, large print, braille) as secure documents, and ensure processes were in place to protect them from loss, theft, and reproduction of any kind.

To access the I AM tests, an Indiana Secure Browser was required. The Indiana 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 Indiana Secure Browser, even if they knew the keystroke sequences. Students were also unable to print from the Indiana Secure Browser. During testing, the desktop was locked down. The Indiana Secure Browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. See the *I AM Online TDS User Guide* in Appendix T for further details.

4.2 INVESTIGATING TEST IRREGULARITIES

Throughout the test window, TAs were required to report breaches of protocol and testing irregularities to the appropriate STC, who was responsible for relaying the report to IDOE. Online test invalidation requests were submitted, as appropriate, through the *Test Irregularities* module under *Administering Tests* in AIR's TIDE.

AIR's QM system gathered data used to detect irregularities, monitored real-time item function, and evaluated test integrity. Every completed test ran through the QM system, and any anomalies (such as unscored or missing items, unexpected test lengths, or other unlikely issues) were flagged, and immediate notification went to AIR psychometricians and the project team through quality assurance (QA) reports. The forensic analysis report

from the QM system flagged unlikely patterns of behavior in testing administrations aggregated at the following levels: test administration, TA, and school.

AIR psychometricians could monitor testing anomalies throughout the test window. A variety of evidence was collected for the evaluation. These included unusual changes in test scores across administrations, much shorter or longer item response times as compared to the state average, and item response patterns using the person-fit index. The flagging criteria used for these analyses were configurable and could be changed by the user. The analyses used to detect the testing anomalies could be run anytime within the test window.

If any unexpected results were identified, the lead psychometrician alerted the project manager immediately to resolve any issues.

4.3 GUIDELINES FOR TEST INVALIDATION

During the test window, TAs were required to immediately report any test irregularities (e.g., disruptive students, loss of Internet connectivity, student improprieties) to the STC. A test irregularity could include testing that was interrupted for an extended period due to a local technical malfunction or severe weather. STCs notified CTCs of any test irregularities that were reported. CTCs were responsible for submitting requests for test invalidations to IDOE via TIDE. STCs notified CTCs of any test irregularities that were reported. Schools invalidated tests at the local level. Once a test was invalidated, IDOE was informed via a Testing Irregularity Report.

4.4 AIR'S SYSTEM SECURITY

AIR has built-in security controls in all of 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. IDOE data resides on servers at Rackspace, AIR's 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 our 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.

5. 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's Alternate Measure

2018–2019

Volume 4
Evidence of Reliability and Validity

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Hyesuk Jang, Yuan Hong, Myung Hee Im, Katherine Krehbiel, Hashim Evans, and Celine Bryan. 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. 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. This new program, named Indiana's Alternate Measure (I AM), replaced the Indiana Standards Tool for Alternate Reporting (ISTAR). It is delivered as an online assessment. Online accommodated and paper-and-pencil versions were available to students whose Individualized Education Program (IEP) or Individual Learning Plan (ILP) 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/Paper-and-pencil tests	3–8, and 10
Mathematics	Online/Paper-and-pencil tests	3–8, and 10
Science	Online/Paper-and-pencil tests	4, 6, and high school
Social Studies	Online/ Paper-and-pencil tests	5

* Stacked Spanish is available for the mathematics, science, and social studies online assessments. Printed braille forms are also available. Full descriptions of available accommodations are listed in Volume 3, Section 1.2, Testing Accommodations. The number of students who were provided with accommodations is presented in Volume 1, Section 2.2, Designated Supports and Accommodations.

With the implementation of these tests, both validity and reliability evidence is necessary to support appropriate inferences of student academic performance from the I AM scores. This volume provides empirical evidence about the validity and reliability of the 2018–2019 I AM, given its intended uses.

The I AM assessments for students with significant cognitive disabilities are designed for measuring student achievement according to Indiana's Alternate Academic Standards (Indiana Content Connectors). Measuring the wide variation of intellectual ability is one of the challenges in developing alternate assessments. To consider the variability of performance within this population, the I AM assessments utilize the attemptedness status of students' engagement with and the stage-adaptive tests. To determine whether a student is sufficiently engaged to produce a valid test score, students are categorized as according to a defined proficiency level, No Mode of Communication (NMC) or Undetermined (UND), according to the number of times they do not respond to the first five items and also the total number of responded items in Part 1. The tests assigned as NMC or UND are not scored. The stage-adaptive tests route students to one of three complexity levels in Part 2 based on their performance in Part 1. Part 2 is also constructed to target different levels of cognitive processing based on results from Part 1 to yield greater measurement precision across the distribution of student achievement levels observed among the I AM student population. Details of I AM test design are described in Volume 1, Section 3.2 Test Design.

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

- **Content Validity.** Evidence is provided in Section 3, Evidence of Content Validity, to show that test forms were constructed to measure the Indiana Content Connectors with a sufficient number of items targeting each area of the blueprint.
- **Internal Structure Validity.** Evidence is provided in this volume 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. Additionally, local item independence, an assumption of unidimensional IRT, was tested using the Q_3 statistic.
- **Test Fairness.** Fairness is statistically analyzed in Section 5, Fairness in Content, using differential item functioning (DIF) in tandem with content alignment reviews by specialists.
- **Reliability.** Section 6, Reliability, addresses the marginal reliability estimates for each test. The reliability estimates are presented by grade and subject, as well as by demographic subgroups. This section also includes conditional standard errors of measurement (CSEMs) and classification accuracy and consistency results by grade and subject.

1.1 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, p. 11). 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 the inferences and interpretations of test scores. *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggests five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be carefully considered.

The first source of evidence for validity is the relationship between the test content and the intended test construct (see Section 3.2, Alignment of I AM Test Forms to the Content Connectors 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 (for details, see Volume

2 of this technical report, Test Development). 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 multi-dimensional scaling, are also used to evaluate content relevance. 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 group of test takers.

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, p. 15). This evidence is collected by surveying test takers about their performance strategies or responses to particular items. Because items are developed to measure particular 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, which is 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 used for analyzing the internal structure of tests (see Volume 1, Section 4.2, Differential Item Functioning Analysis). Other possible analyses used to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (for details, see Section 4, Evidence on Internal-External Structure, and Section 6, Reliability).

A fourth source of evidence for validity is the relationship of test scores to external variables. *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: (1) convergent and discriminant evidence, (2) test-criterion relationships, and (3) 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. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy depends mainly 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, one may need to consider sampling errors or range restrictions to determine whether the conclusions of a test can be assumed for the larger population.

Fifth, 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 affect the test's validity. As described in this volume, as well as in Volume 1, 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 if 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.

1.2 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 repeatedly takes the same or parallel tests, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

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

Test score reliability is traditionally estimated using both classical and IRT approaches. Classical estimates of test reliability, such as Cronbach's alpha, provide an index of the internal consistency reliability of the test, or the likelihood that a student would achieve the same score in an equivalently constructed test form. While classical indicators provide a single estimate of the reliability of test forms, the precision of test scores varies with respect to the information value of the test at each location along the append. 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 stage-adaptive design targets test information near student's ability level in each tier, 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.

2. PURPOSE OF INDIANA'S ALTERNATE MEASURE ASSESSMENTS

The primary purpose of I AM is to yield test scores at the student level and other levels of aggregation that reflect student performance relative to the Indiana Content Connectors. I AM supports instruction and student learning by measuring student performance and providing feedback to educators and parents. Assessments can be used as an indicator to determine whether the students with significant cognitive disabilities in Indiana are ready with the knowledge and skills that are essential for post-secondary education or competitive integrated employment.

I AM assessments also provide evidence in the requirements for state and federal accountability systems. Test scores can be employed to evaluate students' learning progress and help teachers to 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 Indiana Content Connectors in English/Language Arts (ELA), Mathematics, Science, and Social Studies. The test was developed using principles of evidence-centered design and adherence to the principles of universal design to ensure all students have access to the test content. Volume 2, Test Development, describes the Indiana Content Connectors and test blueprints in more detail. The I AM test scores are useful indicators for understanding individual students' academic performance on the Indiana Content Connectors and whether students are progressing in their performance over time. Additionally, individual test scores can be used for measuring reliability of the test which can be found in Section 3, Reliability.

I AM is a criterion-referenced test designed to measure student performance on the Indiana Content Connectors in ELA, Mathematics, Science, and Social Studies. As a comparison, norm-referenced tests are designed to compare or rank all students to one another.

With the overall scale score, percent-correct scores at the reporting category (domain) level were provided for each student to indicate student performance in different content areas of the test relative to the other areas and to the district and state. These scores serve as useful feedback for teachers to tailor their instruction, provided that they are viewed with the usual caution that accompanies 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 across the state. Volume 6 of this technical report series is the score interpretation guide and provides details on all scores generated and their appropriate uses and limitations.

3. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the I AM were representative of the Content Connectors of the larger knowledge domain. We describe the Content Connectors for I AM and discuss the assessment development process, mapping I AM assessments to the Content Connectors. A complete description of the test development process can be found in Volume 2, *Test Development*.

3.1 CONTENT STANDARDS

I AM was aligned to the ELA, Mathematics, Science, and Social Studies Content Connectors adopted in June 2018. I AM Content Connectors are available for review on the Content Connectors page of the Indiana Department of Education (IDOE) website. Blueprints were developed to ensure that the assessment and items were aligned to the prioritized Content Connectors that they were intended to measure.

Table 2–Table 5 present the reporting categories by grade and test, as well as the number of items measuring each category used for the reporting category scores. For ELA (grades 6, 7, 8, 10) and Mathematics, there are the items included in the overall score, but not any reporting category score. A complete description of the blueprint and test form construction process can be found in Volume 2, Section 4 (I AM Blueprints and Assessment Construction), of the I AM technical report.

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

Grade	Reporting Category	Number of Items
3	Key Ideas and Textual Support/Vocabulary (KITS)	10
3	Reading Foundations (RF)	7
3	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	8
3	Writing (W)	7
4	Key Ideas and Textual Support/Vocabulary (KITS)	13
4	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	11
4	Writing (W)	8
5	Key Ideas and Textual Support/Vocabulary (KITS)	14
5	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	9
5	Writing (W)	9
6	Key Ideas and Textual Support/Vocabulary (KITS)	12
6	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	11
6	Writing (W)	8

Grade	Reporting Category	Number of Items
7	Key Ideas and Textual Support/Vocabulary (KITS)	14
7	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	9
7	Writing (W)	8
8	Key Ideas and Textual Support/Vocabulary (KITS)	14
8	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	9
8	Writing (W)	8
10	Key Ideas and Textual Support/Vocabulary (KITS)	12–13
10	Structural Elements and Organization/Connection of Ideas/Media Literacy (SECM)	10–11
10	Writing (W)	8

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

Grade	Reporting Category	Number of Items
3	Algebraic Thinking and Data Analysis (ATDA)	7
3	Computation (C)	7
3	Geometry and Measurement (GM)	7–8
3	Number Sense (NS)	8
4	Algebraic Thinking and Data Analysis (ATDA)	7
4	Computation (C)	7–8
4	Geometry and Measurement (GM)	7
4	Number Sense (NS)	7
5	Algebraic Thinking (AT)	7–8
5	Computation (C)	7
5	Geometry and Measurement, Data Analysis, and Statistics (GMDAS)	8
5	Number Sense (NS)	8–9
6	Algebra and Functions (AF)	8
6	Computation (C)	7
6	Geometry and Measurement, Data Analysis, and Statistics (GMDAS)	7
6	Number Sense (NS)	9

Grade	Reporting Category	Number of Items
7	Algebra and Functions (AF)	8
7	Data Analysis, Statistics, and Probability (DASP)	7
7	Geometry and Measurement (GM)	7
7	Number Sense and Computation (NSC)	8
8	Algebra and Functions (AF)	10
8	Data Analysis, Statistics, and Probability (DASP)	7
8	Geometry and Measurement (GM)	7
8	Number Sense and Computation (NSC)	7
10	Equations and Inequalities (Linear and Systems) (EI)	6–8
10	Functions (Linear and Non-linear) (F)	8
10	Geometry and Measurement (GM)	7
10	Number Sense and Data Analysis (NSDA)	7–8

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

Grade	Reporting Category	Number of Items
4	Analyzing, Interpreting, and Computational Thinking (AICT)	7
4	Explaining Solutions, Reasoning, and Communicating (ESRC)	7
4	Investigating (I)	6–7
4	Questioning and Modeling (QM)	8–10
6	Analyzing, Interpreting, and Computational Thinking (AICT)	7–8
6	Explaining Solutions, Reasoning, and Communicating (ESRC)	7
6	Investigating (I)	8
6	Questioning and Modeling (QM)	8
10	Analyzing Data and Mathematical Thinking (ADMT)	13–14
10	Communicating Explanations and Evaluating Claims Using Evidence (CEEC)	8
10	Developing and Using Modeling to Describe Structure and Function (UM)	10–11

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

Grade	Reporting Category	Number of Items
5	Civics and Government/History (CGH)	16
5	Economics (ECON)	7
5	Geography (GEO)	7

3.2 Alignment of I AM Test Forms to the Content Connectors and Benchmarks

The independent alignment study was conducted November 6 through 8 of this year by the third-party vendor, edCount.

4. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE

In this section, we explore the internal structure of Indiana's Alternate Measure (I AM) 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, Mathematics, Science, and Social Studies, there are three to four reporting categories that differ in some cases by grade (see Table 2 through Table 5 for reporting category information). Students were provided with their percentage of correct answers based on each reporting category. Evidence is needed to verify that scores for each category provide useful information on 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 the 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 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 and disattenuated correlations between the subscores for test or at grade level are provided in the following sections. The theta estimates of each subscore were used for the correlations.

4.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 6–Table 9 present the observed correlation matrix of the reporting category scores for each subject area. In ELA, the correlations among the reporting categories ranged from 0.52–0.72. For Mathematics, the correlations were between 0.08–0.55. In Science, the correlations among reporting categories ranged from 0.31–0.62. In Social Studies, the correlations among reporting categories ranged from 0.48–0.62.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a larger standard error of measurement (SEM) 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 10–Table 13 display disattenuated correlations. Disattenuated values greater than 1 are reported as 1.00*. The overall average disattenuated correlation was 0.95 for ELA, 0.93 for Mathematics, 0.96 for Science, and 1.00 for Social Studies. The values suggest that validity evidence of internal structure is supported.

Table 6: Observed Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Key Ideas and Textual Support/Vocabulary (Cat1)	10	1.00	0.33	0.36	0.37
	Reading Foundations (Cat2)	7		1.00	0.35	0.22
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat3)	8			1.00	0.27
	Writing (Cat4)	7				1.00
4	Key Ideas and Textual Support/Vocabulary (Cat1)	13	1.00	0.51	0.68	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11		1.00	0.47	
	Writing (Cat3)	8			1.00	
5	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00	0.61	0.62	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	9		1.00	0.56	
	Writing (Cat3)	9			1.00	
6	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1.00	0.58	0.55	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11		1.00	0.51	
	Writing (Cat3)	8			1.00	
7	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00	0.64	0.57	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	9		1.00	0.63	
	Writing (Cat3)	8			1.00	
8	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00	0.61	0.59	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	9		1.00	0.43	
	Writing (Cat3)	8			1.00	
10	Key Ideas and Textual Support/Vocabulary (Cat1)	12–13	1.00	0.67	0.72	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–11		1.00	0.61	
	Writing (Cat3)	8			1.00	

Table 7: Observed Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	7	1.00	0.49	0.36	0.55
	Computation (Cat2)	7		1.00	0.26	0.45
	Geometry and Measurement (Cat3)	7-8			1.00	0.34
	Number Sense (Cat4)	8				1.00
4	Algebraic Thinking and Data Analysis (Cat1)	7	1.00	0.42	0.42	0.35
	Computation (Cat2)	7–8		1.00	0.44	0.49
	Geometry and Measurement (Cat3)	7			1.00	0.42
	Number Sense (Cat4)	7				1.00
5	Algebraic Thinking (Cat1)	7–8	1.00	0.20	0.31	0.10
	Computation (Cat2)	7		1.00	0.40	0.32
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	8			1.00	0.30
	Number Sense (Cat4)	8–9				1.00
6	Algebra and Functions (Cat1)	8	1.00	0.35	0.35	0.40
	Computation (Cat2)	7		1.00	0.29	0.29
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	7			1.00	0.39
	Number Sense (Cat4)	9				1.00
7	Algebra and Functions (Cat1)	8	1.00	0.20	0.18	0.15
	Data Analysis, Statistics, and Probability (Cat2)	7		1.00	0.25	0.21
	Geometry and Measurement (Cat3)	7			1.00	0.17
	Number Sense and Computation (Cat4)	8				1.00
8	Algebra and Functions (Cat1)	10	1.00	0.22	0.16	0.24
	Data Analysis, Statistics, and Probability (Cat2)	7		1.00	0.24	0.21
	Geometry and Measurement (Cat3)	7			1.00	0.24
	Number Sense and Computation (Cat4)	7				1.00
10	Equations and Inequalities (Linear and Systems) (Cat1)	6–8	1.00	0.21	0.08	0.25
	Functions (Linear and Non-linear) (Cat2)	8		1.00	0.14	0.37
	Geometry and Measurement (Cat3)	7			1.00	0.26
	Number Sense and Data Analysis (Cat4)	7–8				1.00

Table 8: Observed Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
4	Analyzing, Interpreting, and Computational Thinking (Cat1)	7	1.00	0.41	0.53	0.45
	Explaining Solutions, Reasoning, and Communicating (Cat2)	7		1.00	0.53	0.57
	Investigating (Cat3)	6–7			1.00	0.61
	Questioning and Modeling (Cat4)	8–10				1.00
6	Analyzing, Interpreting, and Computational Thinking (Cat1)	7–8	1.00	0.31	0.48	0.43
	Explaining Solutions, Reasoning, and Communicating (Cat2)	7		1.00	0.35	0.36
	Investigating (Cat3)	8			1.00	0.40
	Questioning and Modeling (Cat4)	8				1.00
Biology	Analyzing Data and Mathematical Thinking (Cat1)	13–14	1.00	0.62	0.62	
	Communicating Explanations and Evaluating Claims Using Evidence (Cat2)	8		1.00	0.58	
	Developing and Using Modeling to Describe Structure and Function (Cat3)	10–11			1.00	

Table 9: Observed Correlation Matrix Among Reporting Categories (Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government/History (Cat1)	16	1.00	0.55	0.62
	Economics (Cat2)	7		1.00	0.48
	Geography (Cat3)	7			1.00

Table 10: Disattenuated Correlation Matrix Among Reporting Categories (ELA)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Key Ideas and Textual Support/Vocabulary (Cat1)	10	1.00	0.96	0.70	1.00*
	Reading Foundations (Cat2)	7		1.00	1.00*	1.00*
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat3)	8			1.00	0.90
	Writing (Cat4)	7				1.00
4	Key Ideas and Textual Support/Vocabulary (Cat1)	13	1.00	0.80	1.00*	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11		1.00	0.83	
	Writing (Cat3)	8			1.00	
5	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00	0.97	0.95	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	9		1.00	0.96	
	Writing (Cat3)	9			1.00	
6	Key Ideas and Textual Support/Vocabulary (Cat1)	12	1.00	0.96	0.93	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	11		1.00	0.99	
	Writing (Cat3)	8			1.00	
7	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00	1.00*	1.00*	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	9		1.00	1.00*	
	Writing (Cat3)	8			1.00	
8	Key Ideas and Textual Support/Vocabulary (Cat1)	14	1.00	1.00*	0.93	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	9		1.00	0.82	
	Writing (Cat3)	8			1.00	
10	Key Ideas and Textual Support/Vocabulary (Cat1)	12–13	1.00	1.00*	1.00*	
	Structural Elements and Organization/Connection of Ideas/Media Literacy (Cat2)	10–11		1.00	1.00*	
	Writing (Cat3)	8			1.00	

Table 11: Disattenuated Correlation Matrix Among Reporting Categories (Mathematics)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
3	Algebraic Thinking and Data Analysis (Cat1)	7	1.00	1.00*	1.00*	1.00*
	Computation (Cat2)	7		1.00	0.84	1.00*
	Geometry and Measurement (Cat3)	7-8			1.00	0.98
	Number Sense (Cat4)	8				1.00
4	Algebraic Thinking and Data Analysis (Cat1)	7	1.00	1.00*	1.00*	0.96
	Computation (Cat2)	7–8		1.00	1.00*	1.00*
	Geometry and Measurement (Cat3)	7			1.00	1.00*
	Number Sense (Cat4)	7				1.00
5	Algebraic Thinking (Cat1)	7–8	1.00	0.73	0.84	0.33
	Computation (Cat2)	7		1.00	1.00*	1.00*
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	8			1.00	0.91
	Number Sense (Cat4)	8–9				1.00
6	Algebra and Functions (Cat1)	8	1.00	1.00*	1.00*	1.00*
	Computation (Cat2)	7		1.00	0.92	0.85
	Geometry and Measurement, Data Analysis, and Statistics (Cat3)	7			1.00	1.00*
	Number Sense (Cat4)	9				1.00
7	Algebra and Functions (Cat1)	8	1.00	1.00*	0.82	1.00*
	Data Analysis, Statistics, and Probability (Cat2)	7		1.00	1.00*	1.00*
	Geometry and Measurement (Cat3)	7			1.00	0.77
	Number Sense and Computation (Cat4)	8				1.00
8	Algebra and Functions (Cat1)	10	1.00	0.70	0.68	0.88
	Data Analysis, Statistics, and Probability (Cat2)	7		1.00	0.99	0.76
	Geometry and Measurement (Cat3)	7			1.00	1.00*
	Number Sense and Computation (Cat4)	7				1.00
10	Equations and Inequalities (Linear and Systems) (Cat1)	6–8	1.00	1.00*	1.00*	1.00*
	Functions (Linear and Non-linear) (Cat2)	8		1.00	1.00*	1.00*
	Geometry and Measurement (Cat3)	7			1.00	1.00*
	Number Sense and Data Analysis (Cat4)	7–8				1.00

Table 12: Disattenuated Correlation Matrix Among Reporting Categories (Science)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3	Cat4
4	Analyzing, Interpreting, and Computational Thinking (Cat1)	7	1.00	1.00*	1.00*	1.00*
	Explaining Solutions, Reasoning, and Communicating (Cat2)	7		1.00	1.00*	1.00*
	Investigating (Cat3)	6–7			1.00	1.00*
	Questioning and Modeling (Cat4)	8–10				1.00
6	Analyzing, Interpreting, and Computational Thinking (Cat1)	7–8	1.00	0.69	1.00*	0.98
	Explaining Solutions, Reasoning, and Communicating (Cat2)	7		1.00	0.91	0.87
	Investigating (Cat3)	8			1.00	1.00*
	Questioning and Modeling (Cat4)	8				1.00
Biology	Analyzing Data and Mathematical Thinking (Cat1)	13–14	1.00	1.00*	0.98	
	Communicating Explanations and Evaluating Claims Using Evidence (Cat2)	8		1.00	1.00*	
	Developing and Using Modeling to Describe Structure and Function (Cat3)	7			1.00	

Table 13: Disattenuated Correlation Matrix Among Reporting Categories (Social Studies)

Grade	Reporting Category	Number of Items	Cat1	Cat2	Cat3
5	Civics and Government/History (Cat1)	16	1.00	1.00*	1.00*
	Economics (Cat2)	7		1.00	0.99
	Geography (Cat3)	7			1.00

4.2 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 the probability of correct responses is the product of independent probabilities over all items (Chen & Thissen, 1997):

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

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that "local independence follows automatically from unidimensionality" (as cited in Bejar

[1980], p. 5). From a dimensionality perspective, there may be nuisance factors that are influencing relationships among certain items after accounting for the intended construct of interest. These nuisance factors can be influenced by a number of testing features, such as speediness, 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 examinee for item *i*, $T_i(\hat{\theta}_j)$ is the estimated true score for item *i* of examinee *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 examinee *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 14–Table 17 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 about 90% of the items, between the 5th and 95th percentiles for most of grades and subjects, were equal to or smaller than a critical value of 0.2 for $|Q_3|$ (Chen & Thissen, 1997), except for two grades in Mathematics which have the value of 0.21 and 0.22 for $|Q_3|$.

Table 14: ELA Q3 Statistic

Grade	Q_3 Distribution					Within Passage Q_3*	
	Minimum	5th Percentile	Median	95th Percentile	Maximum	Minimum	Maximum
3	−0.481	−0.192	−0.036	0.155	0.626	−0.254	0.362
4	−0.325	−0.173	−0.040	0.133	0.341	−0.325	0.341
5	−0.341	−0.173	−0.042	0.140	0.409	−0.152	0.409
6	−0.299	−0.181	−0.036	0.129	0.365	−0.145	0.167
7	−0.303	−0.170	−0.044	0.129	0.261	−0.177	0.149
8	−0.256	−0.162	−0.042	0.123	0.276	−0.180	0.174
10	−0.320	−0.154	−0.037	0.095	0.287	−0.152	0.238

* Within Passage Q3, values are computed for each item pair within a passage.

Table 15: Mathematics Q3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
3	−0.363	−0.193	−0.044	0.149	0.302
4	−0.342	−0.190	−0.040	0.150	0.304
5	−0.496	−0.224	−0.036	0.184	0.362
6	−0.298	−0.165	−0.046	0.148	0.324
7	−0.518	−0.214	−0.041	0.171	0.602
8	−0.413	−0.181	−0.047	0.146	0.413
10	−0.379	−0.166	−0.035	0.118	0.399

Table 16: Science Q3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
4	−0.381	−0.203	−0.052	0.189	0.339
6	−0.314	−0.201	−0.048	0.179	0.365
Biology	−0.342	−0.167	−0.042	0.126	0.366

Table 17: Social Studies Q3 Statistic

Grade	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
5	−0.322	−0.169	−0.045	0.128	0.350

4.3 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 validity evidence demonstrating that assessment scores are related as expected with criterion and other variables for all student groups. However, a second, independent test measuring the same constructs as ELA, Mathematics, Science, and Social Studies in Indiana, which could easily permit for a cross-test set of correlations, was not available. Therefore, the correlations between subscores within and across assessments 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 (e.g., typically around 7 to 11); as a consequence, 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 content area and across subjects and grades. Each correlation table shows the observed or disattenuated subscore correlations among two or three subjects: tables of grades 3, 7, and 8 include ELA and Mathematics; tables of grades 4, 6, and 10 include ELA, Mathematics and Science; and tables of grade 5 include ELA, Mathematics, and Social Studies. In general, the pattern is consistent with the a-priori expectation that subscores within an assessment correlate more highly than correlations between assessments measuring a different construct.

Table 18: Grade 3 Observed Score Correlations

Subject	Reporting Category	ELA				Mathematics			
		Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.33	0.36	0.37	0.44	0.39	0.30	0.43
	RF (Cat2)		1.00	0.35	0.22	0.22	0.22	0.21	0.27
	SECM (Cat3)			1.00	0.27	0.23	0.21	0.20	0.22
	W (Cat4)				1.00	0.28	0.22	0.22	0.28
Mathematics	ATDA (Cat1)					1.00	0.49	0.36	0.55
	C (Cat2)						1.00	0.26	0.45
	GM (Cat3)							1.00	0.34
	NS (Cat4)								1.00

Table 19: Grade 4 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.51	0.68	0.46	0.54	0.49	0.47	0.48	0.57	0.63	0.67
	SECM (Cat2)		1.00	0.47	0.35	0.38	0.34	0.20	0.38	0.37	0.51	0.40
	W (Cat3)			1.00	0.44	0.51	0.46	0.43	0.41	0.55	0.57	0.59
Mathematics	ATDA (Cat1)				1.00	0.42	0.42	0.35	0.35	0.34	0.49	0.44
	C (Cat2)					1.00	0.44	0.49	0.32	0.43	0.46	0.54
	GM (Cat3)						1.00	0.42	0.35	0.35	0.44	0.50
	NS (Cat4)							1.00	0.26	0.35	0.45	0.45
Science	AICT (Cat1)								1.00	0.41	0.53	0.45
	ESRC (Cat2)									1.00	0.53	0.57
	I (Cat3)										1.00	0.61
	QM (Cat4)											1.00

Table 20: Grade 5 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Social Studies		
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3
ELA	KITS (Cat1)	1.00	0.61	0.62	0.32	0.33	0.50	0.26	0.63	0.54	0.60
	SECM (Cat2)		1.00	0.56	0.18	0.37	0.44	0.27	0.53	0.53	0.52
	W (Cat3)			1.00	0.26	0.36	0.50	0.26	0.54	0.46	0.51
Mathematics	AT (Cat1)				1.00	0.20	0.31	0.10	0.37	0.16	0.26
	C (Cat2)					1.00	0.40	0.32	0.33	0.36	0.31
	GMDAS (Cat3)						1.00	0.30	0.53	0.46	0.45
	NS (Cat4)							1.00	0.30	0.28	0.25
Social Studies	CGH (Cat1)								1.00	0.55	0.62
	ECON (Cat2)									1.00	0.48
	GEO (Cat3)										1.00

Table 21: Grade 6 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.58	0.55	0.36	0.37	0.38	0.41	0.55	0.39	0.42	0.50
	SECM (Cat2)		1.00	0.51	0.33	0.33	0.41	0.38	0.43	0.37	0.44	0.42
	W (Cat3)			1.00	0.37	0.28	0.42	0.39	0.43	0.41	0.36	0.45
Mathematics	AF (Cat1)				1.00	0.35	0.35	0.40	0.28	0.25	0.34	0.32
	C (Cat2)					1.00	0.29	0.29	0.29	0.22	0.29	0.29
	GMDAS (Cat3)						1.00	0.39	0.35	0.33	0.29	0.38
	NS (Cat4)							1.00	0.36	0.32	0.32	0.38
Science	AICT (Cat1)								1.00	0.31	0.48	0.43
	ESRC (Cat2)									1.00	0.35	0.36
	I (Cat3)										1.00	0.40
	QM (Cat4)											1.00

Table 22: Grade 7 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.64	0.57	0.22	0.27	0.43	0.20
	SECM (Cat2)		1.00	0.63	0.16	0.26	0.42	0.21
	W (Cat3)			1.00	0.22	0.27	0.42	0.20
Mathematics	AF (Cat1)				1.00	0.20	0.18	0.15
	DASP (Cat2)					1.00	0.25	0.21
	GM (Cat3)						1.00	0.17
	NSC (Cat4)							1.00

Table 23: Grade 8 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.61	0.59	0.28	0.35	0.30	0.27
	SECM (Cat2)		1.00	0.43	0.23	0.25	0.21	0.21
	W (Cat3)			1.00	0.23	0.36	0.22	0.21
Mathematics	AF (Cat1)				1.00	0.22	0.16	0.24
	DASP (Cat2)					1.00	0.24	0.21
	GM (Cat3)						1.00	0.24
	NSC (Cat4)							1.00

Table 24: Grade 10 Observed Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science		
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3
ELA	KITS (Cat1)	1.00	0.67	0.72	0.30	0.34	0.16	0.41	0.70	0.62	0.61
	SECM (Cat2)		1.00	0.61	0.28	0.29	0.17	0.36	0.58	0.51	0.50
	W (Cat3)			1.00	0.30	0.38	0.17	0.38	0.64	0.57	0.57
Mathematics	EI (Cat1)				1.00	0.21	0.08	0.25	0.30	0.30	0.30
	F (Cat2)					1.00	0.14	0.37	0.38	0.32	0.29
	GM (Cat3)						1.00	0.26	0.15	0.11	0.08
	NSDA (Cat4)							1.00	0.41	0.35	0.35
Science	ADMT (Cat1)								1.00	0.62	0.62
	CEEC (Cat2)									1.00	0.58
	UM (Cat3)										1.00

Table 25: Grade 3 Disattenuated Score Correlations

Subject	Reporting Category	ELA				Mathematics			
		Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.96	0.70	1.00*	0.87	0.80	0.78	0.80
	RF (Cat2)		1.00	1.00*	1.00*	0.77	0.80	0.97	0.90
	SECM (Cat3)			1.00	0.90	0.52	0.51	0.59	0.48
	W (Cat4)				1.00	0.96	0.82	1.00*	0.92
Mathematics	ATDA (Cat1)					1.00	1.00*	1.00*	1.00*
	C (Cat2)						1.00	0.84	1.00*
	GM (Cat3)							1.00	0.98
	NS (Cat4)								1.00

Table 26: Grade 4 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.80	1.00*	1.00*	0.89	0.95	0.81	1.00*	1.00*	1.00*	1.00*
	SECM (Cat2)		1.00	0.83	0.88	0.71	0.76	0.39	0.99	0.77	0.98	0.68
	W (Cat3)			1.00	1.00*	0.93	1.00*	0.84	1.00*	1.00*	1.00*	1.00*
Mathematics	ATDA (Cat1)				1.00	1.00*	1.00*	0.96	1.00*	0.99	1.00*	1.00*
	C (Cat2)					1.00	1.00*	1.00*	0.87	0.94	0.91	0.97
	GM (Cat3)						1.00	1.00*	1.00*	0.90	1.00*	1.00*
	NS (Cat4)							1.00	0.73	0.79	0.93	0.84
Science	AICT (Cat1)								1.00	1.00*	1.00*	1.00*
	ESRC (Cat2)									1.00	1.00*	1.00*
	I (Cat3)										1.00	1.00*
	QM (Cat4)											1.00

Table 27: Grade 5 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Social Studies		
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3
ELA	KITS (Cat1)	1.00	0.97	0.95	0.65	0.83	0.92	0.60	0.91	0.93	0.99
	SECM (Cat2)		1.00	0.96	0.42	1.00*	0.92	0.71	0.89	1.00*	0.99
	W (Cat3)			1.00	0.58	0.98	1.00*	0.66	0.86	0.86	0.92
Mathematics	AT (Cat1)				1.00	0.73	0.84	0.33	0.80	0.41	0.63
	C (Cat2)					1.00	1.00*	1.00*	0.87	1.00*	0.93
	GMDAS (Cat3)						1.00	0.91	1.00*	1.00*	0.98
	NS (Cat4)							1.00	0.72	0.81	0.67
Social Studies	CGH (Cat1)								1.00	1.00*	1.00*
	ECON (Cat2)									1.00	0.99
	GEO (Cat3)										1.00

Table 28: Grade 6 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	0.96	0.93	0.79	0.81	0.78	0.78	0.95	0.73	0.86	0.94
	SECM (Cat2)		1.00	0.99	0.85	0.85	0.96	0.81	0.86	0.78	1.00*	0.91
	W (Cat3)			1.00	0.98	0.74	1.00*	0.86	0.88	0.90	0.87	1.00
Mathematics	AF (Cat1)				1.00	1.00*	1.00*	1.00*	0.75	0.71	1.00*	0.93
	C (Cat2)					1.00	0.92	0.85	0.79	0.62	0.92	0.84
	GMDAS (Cat3)						1.00	1.00*	0.87	0.86	0.84	1.00*
	NS (Cat4)							1.00	0.82	0.77	0.86	0.93
Science	AICT (Cat1)								1.00	0.69	1.00*	0.98
	ESRC (Cat2)									1.00	0.91	0.87
	I (Cat3)										1.00	1.00*
	QM (Cat4)											1.00

Table 29: Grade 7 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	1.00*	1.00*	0.79	0.86	0.95	0.66
	SECM (Cat2)		1.00	1.00*	0.58	0.88	0.98	0.74
	W (Cat3)			1.00	0.87	0.95	1.00*	0.74
Mathematics	AF (Cat1)				1.00	1.00*	0.82	1.00*
	DASP (Cat2)					1.00	1.00*	1.00*
	GM (Cat3)						1.00	0.77
	NSC (Cat4)							1.00

Table 30: Grade 8 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics			
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4
ELA	KITS (Cat1)	1.00	1.00*	0.93	0.59	0.73	0.84	0.68
	SECM (Cat2)		1.00	0.82	0.58	0.63	0.72	0.63
	W (Cat3)			1.00	0.54	0.84	0.68	0.59
Mathematics	AF (Cat1)				1.00	0.70	0.68	0.88
	DASP (Cat2)					1.00	0.99	0.76
	GM (Cat3)						1.00	1.00*
	NSC (Cat4)							1.00

Table 31: Grade 10 Disattenuated Score Correlations

Subject	Reporting Category	ELA			Mathematics				Science		
		Cat1	Cat2	Cat3	Cat1	Cat2	Cat3	Cat4	Cat1	Cat2	Cat3
ELA	KITS (Cat1)	1.00	1.00*	1.00*	0.90	1.00*	1.00*	0.88	1.00*	1.00*	0.98
	SECM (Cat2)		1.00	1.00*	0.88	0.95	1.00*	0.80	0.91	0.93	0.84
	W (Cat3)			1.00	0.99	1.00*	1.00*	0.90	1.00*	1.00*	1.00
Mathematics	EI (Cat1)				1.00	1.00*	1.00*	1.00*	0.89	0.99	0.95
	F (Cat2)					1.00	1.00*	1.00*	1.00*	1.00*	0.94
	GM (Cat3)						1.00	1.00*	1.00*	1.00*	0.73
	NSDA (Cat4)							1.00	0.85	0.84	0.78
Science	ADMT (Cat1)								1.00	1.00*	0.98
	CEEC (Cat2)									1.00	1.00*
	UM (Cat3)										1.00

5. 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). They include:

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

AIR content experts received extensive training on the principles of universal design and applied these principles in the development of all test materials. In the review process, adherence to the principles of universal design was verified by Indiana leadership.

5.1 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is not sufficient to determine the fairness of an assessment. 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 utilized was 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, Section 4.2 Differential Item Functioning Analysis. 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 or the reference group. The details surrounding this review of items for bias is further described in Volume 2 of this technical report, *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
- Autism/Other
- Moderate and severe intellectual disability/Other

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 4.2, Differential Item Functioning Analysis. The DIF statistics for each operational test item are presented in Volume 1, Appendix A, Operational Item Statistics.

6. RELIABILITY

6.1 MARGINAL RELIABILITY

Marginal reliability is a measure of the overall reliability of the test based on the average CSEMs, 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 test information function (TIF) represents the SEM, which is equal to the inverse square root of information. The larger the SEM, 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:

$$\rho = \frac{\sigma_\theta^2 - \underline{\sigma_e^2}}{\sigma_\theta^2}$$

where σ_θ^2 is the square of the standard error of student ability estimate (θ). The marginal measurement error variance ($\underline{\sigma_e^2}$) can be estimated as the average of squared standard error of θ across all test takers. The marginal measurement error variance ($\underline{\sigma_e^2}$) is computed as:

$$\underline{\sigma_e^2} = \int_{-\infty}^{\infty} \frac{1}{I(\theta)} g(\theta) d\theta = \frac{\sum \sigma_\theta^2}{N},$$

where $g(\theta)$ is a density function and *N* is the number of students.

Table 32 presents the marginal reliability coefficients for all students by test. The marginal reliability coefficients for all grades of ELA, Science, and Social Studies range from 0.74–0.86, which is similar to other statewide standardized tests. The marginal reliability coefficients of lower grades in Mathematics had a similar level to other subjects, ranging from 0.67–0.79, while the higher grades had the marginal reliability coefficients of 0.54–0.64, which is expected due to the small standard deviations of theta scores from 0.57–0.68. The standard deviations of theta scores are provided in Table 33. The marginal reliability by each demographic subgroup are presented in Appendix A, Reliability Coefficients. In Appendix A, demographic subgroups include Female, Male, Autism, Non-Autism, Moderate and Severe Intellectual Disability, Non-Moderate and Severe Intellectual Disability, African American, Hispanic, and White.

Table 32: Marginal Reliability Coefficients

Test	Number of Students	Number of Operational Items*	Marginal Reliability
ELA 3	713	32	0.74
ELA 4	772	32	0.84
ELA 5	818	32	0.85
ELA 6	961	32	0.82
ELA 7	986	32	0.82
ELA 8	1103	32	0.83
ELA 10	1078	32	0.86
Mathematics 3	709	31–32	0.76
Mathematics 4	766	32	0.79
Mathematics 5	809	32	0.67
Mathematics 6	953	32	0.70
Mathematics 7	988	32	0.54
Mathematics 8	1101	32	0.59
Mathematics 10	1078	31–32	0.64
Science 4	766	29–31	0.78
Science 6	943	29–31	0.75
Biology	1013	32	0.83
Social Studies 5	802	30	0.81

* While 32 operational items were administered on all tests, for some tests, the number of operational items used for scoring in each tiered test varied because of loss of items in the Item Data Review.

6.2 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 assessment, 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 CSEM at each score point. For instance, if the SEM is large, then less information is being provided by the assessment at the specific ability level.

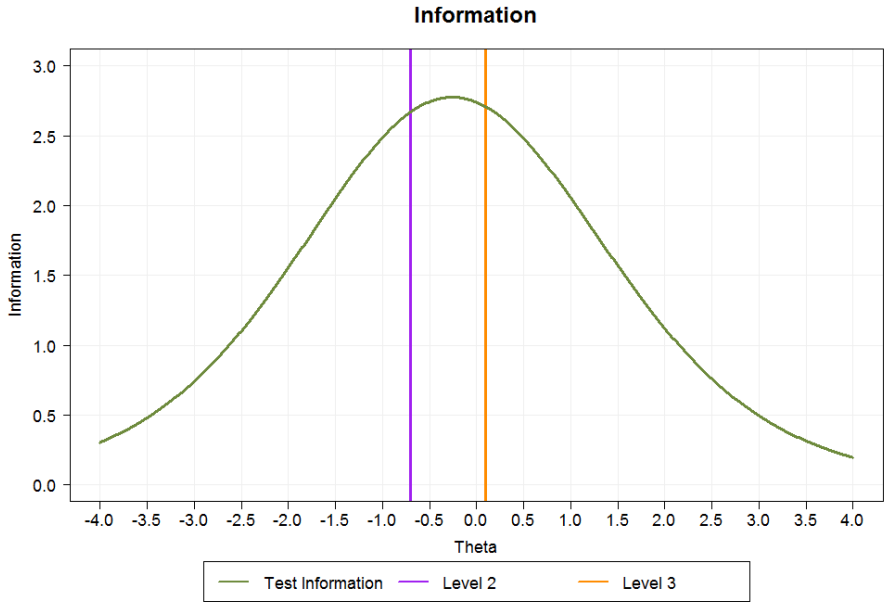
Figure 1 displays a sample TIF with two vertical lines indicating the performance cut scores. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most precise scores in this range. The test provides less information about test takers at the tails, where the curve is lower, 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 I AM is calculated as:

$$TIF(\theta_s) = \sum_{i=1}^{N_{PCM}} \left(\frac{\sum_{h=1}^{m_i} h^2 exp\left(\sum_{l=1}^{h} (\theta_s - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} (\theta_s - b_{il})\right)} - \left(\frac{\sum_{h=1}^{m_i} h exp\left(\sum_{l=1}^{h} (\theta_s - b_{il})\right)}{1 + \sum_{h=1}^{m_i} exp\left(\sum_{l=1}^{h} (\theta_s - b_{il})\right)} \right)^2 \right),$$

where N_{PCM} is the number of items that are scored using partial credit model (PCM) items, 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 SEM for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

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

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the SEMs are more useful for score interpretation. For this reason, SEM plots are presented in Figure 2 through Figure 5 for ELA, Mathematics, Science, and Social Studies respectively, instead of the TIFs. These plots are based on the scaled scores reported in 2019. Vertical lines represent two performance category cut scores.

When the maximum likelihood estimate (MLE) is used for score estimation, it is also common to compute the SEM from the numerically differentiated Hessian, which approximates as follows:

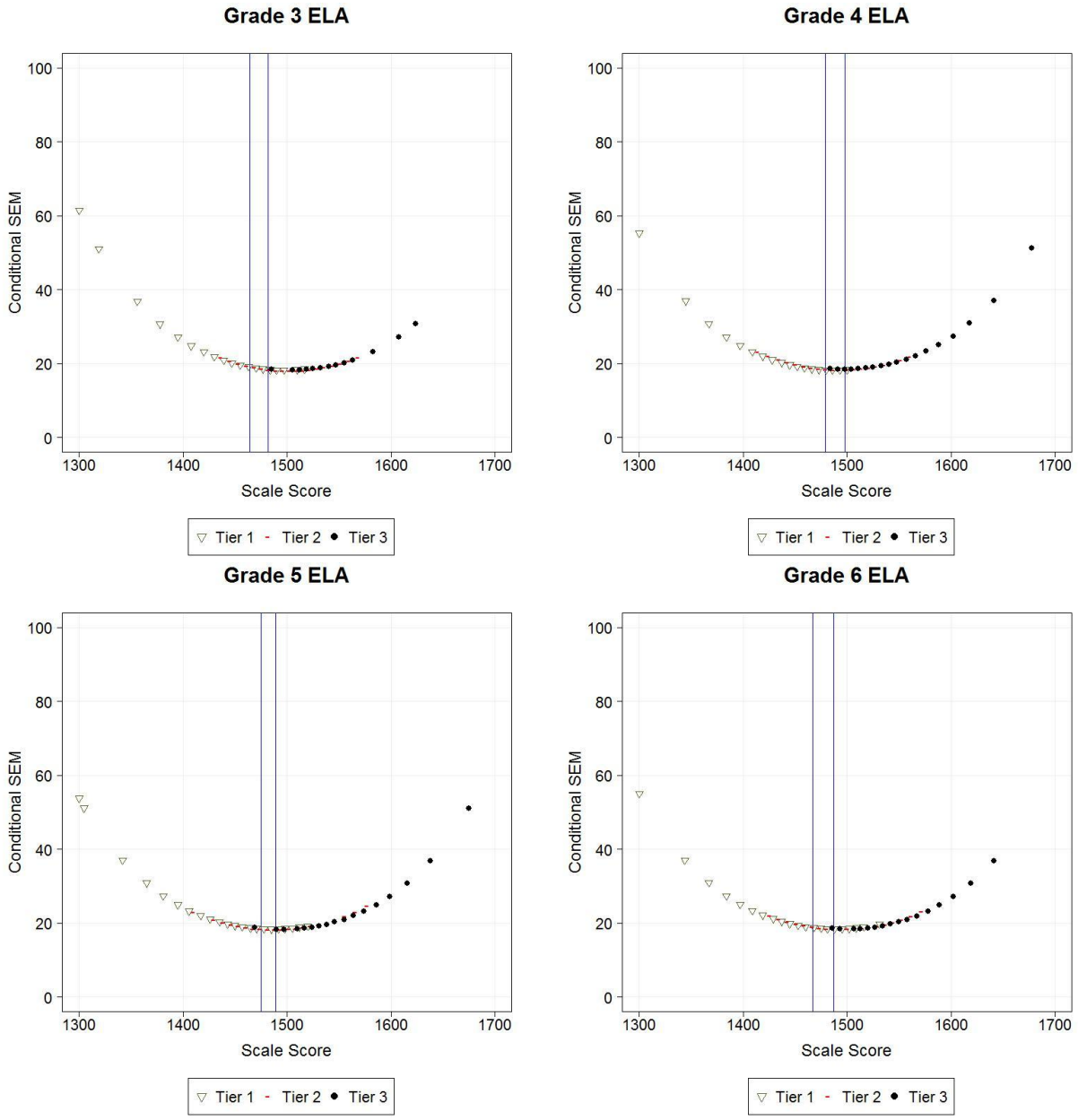
$$se(\hat{\theta}) = \frac{1}{\sqrt{-\left(\frac{\partial^2 lnL(\hat{\theta})}{\partial^2 \theta}\right)}},$$

where

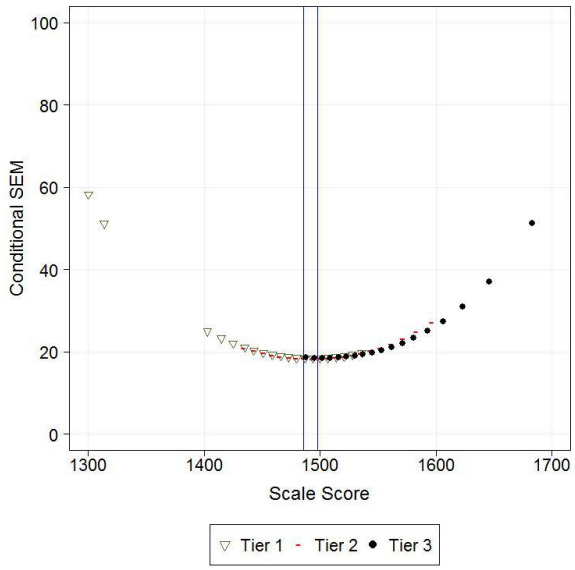
$$\frac{\partial^2 lnL(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{PCM}} \left(\left(\frac{\sum_{j=1}^{m_i} jexp\left(\sum_{k=1}^{j} (\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} (\hat{\theta} - b_{ik})\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} (\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} (\hat{\theta} - b_{ik})\right)} \right),$$

where N_{PCM} is the number of items that are scored using PCM items. Figure 2–Figure 5 (also the CSEM in Appendix B, Conditional Standard Error of Measurement) are based on the Hessian estimates. Vertical lines in the plots represent the Approaching Proficiency and At Proficiency performance category cut scores respectively.

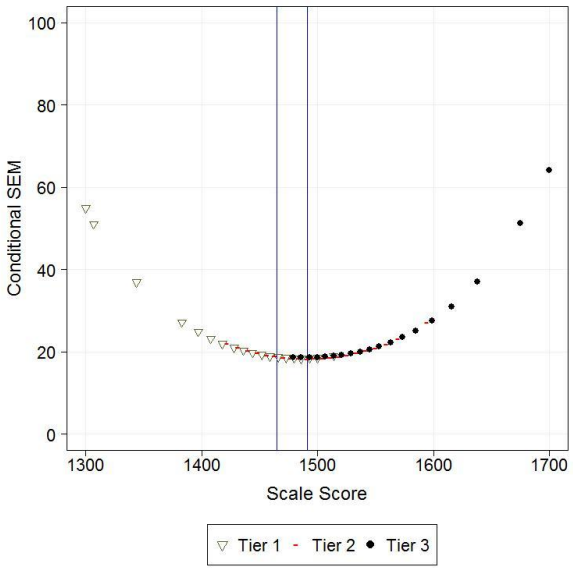
Figure 2: Conditional Standard Errors of Measurement (ELA)



Grade 7 ELA



Grade 8 ELA



Grade 10 ELA

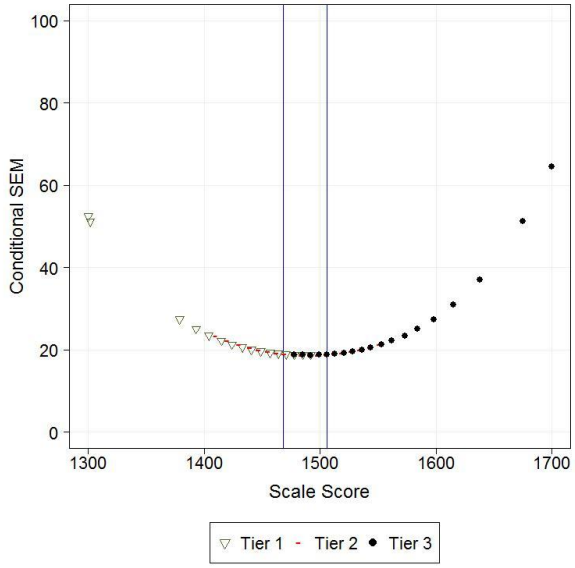
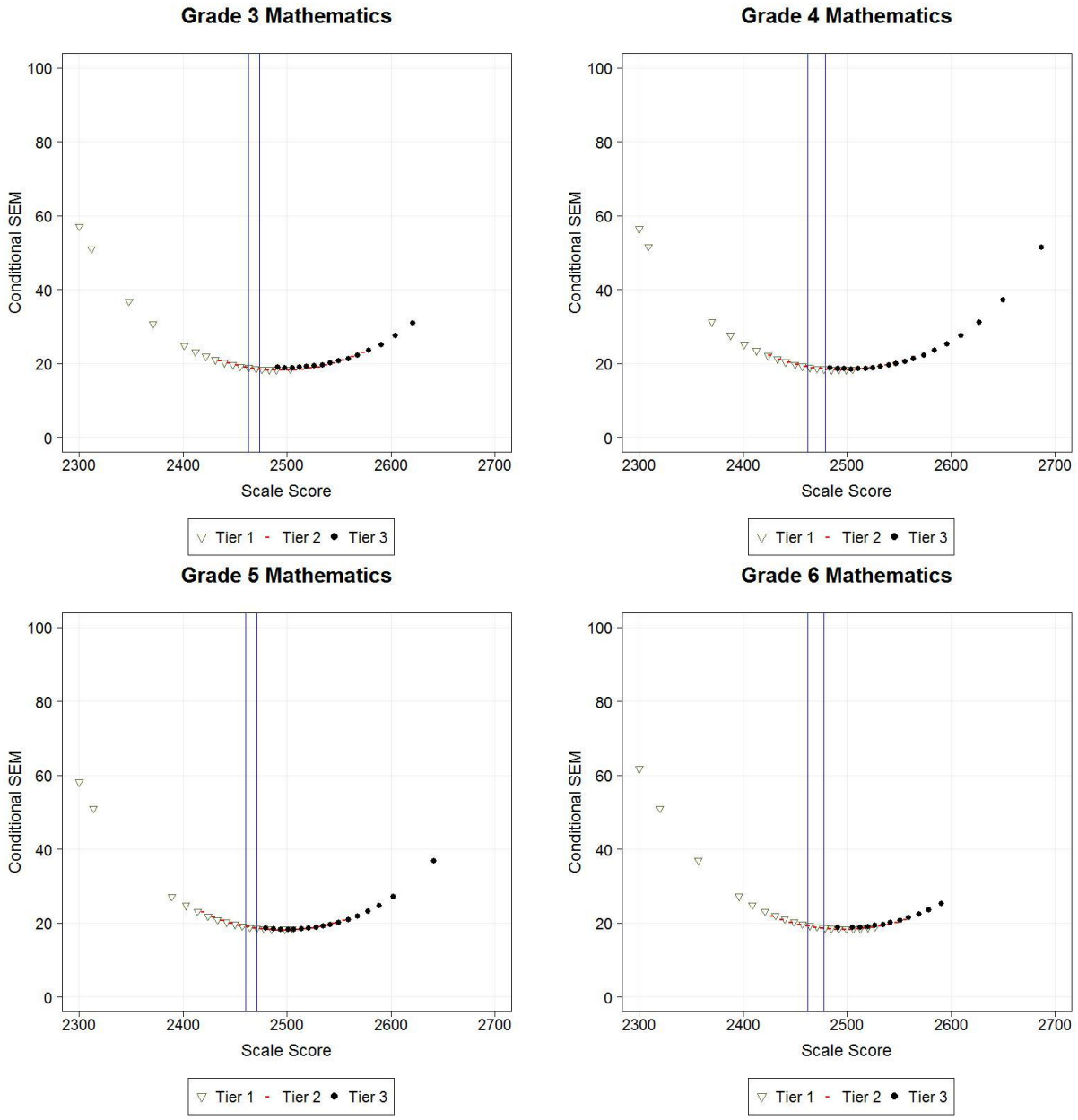
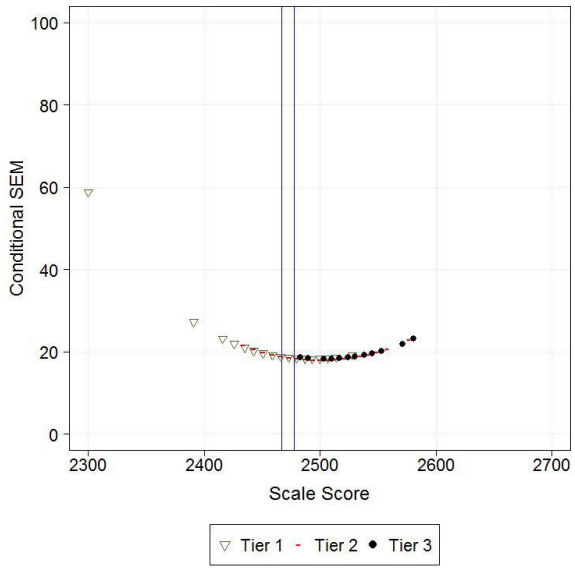


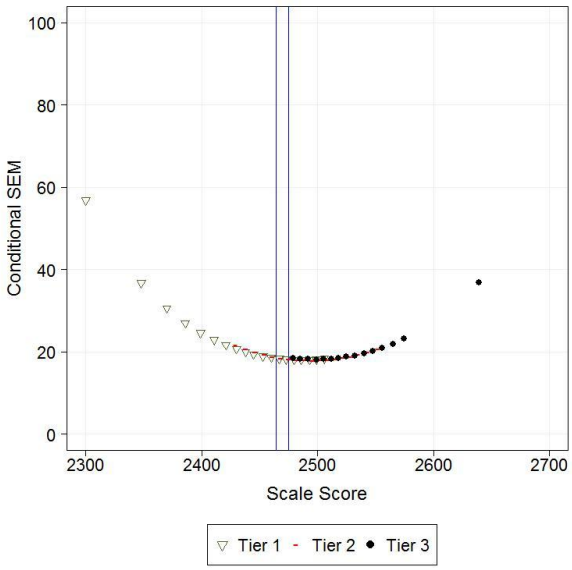
Figure 3: Conditional Standard Errors of Measurement (Mathematics)



Grade 7 Mathematics



Grade 8 Mathematics



Grade 10 Mathematics

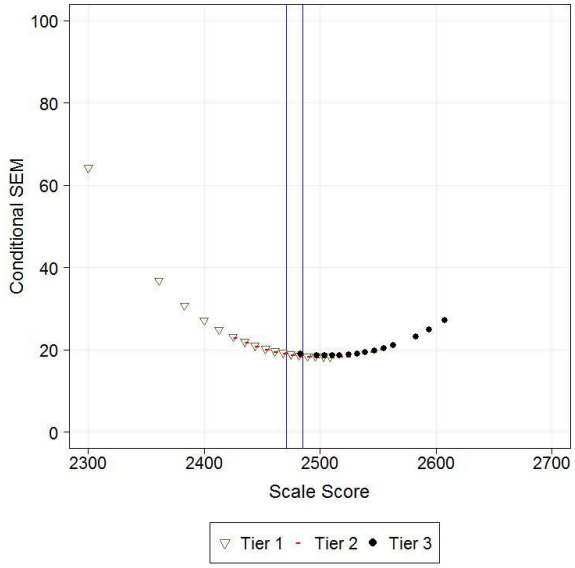


Figure 4: Conditional Standard Errors of Measurement (Science)

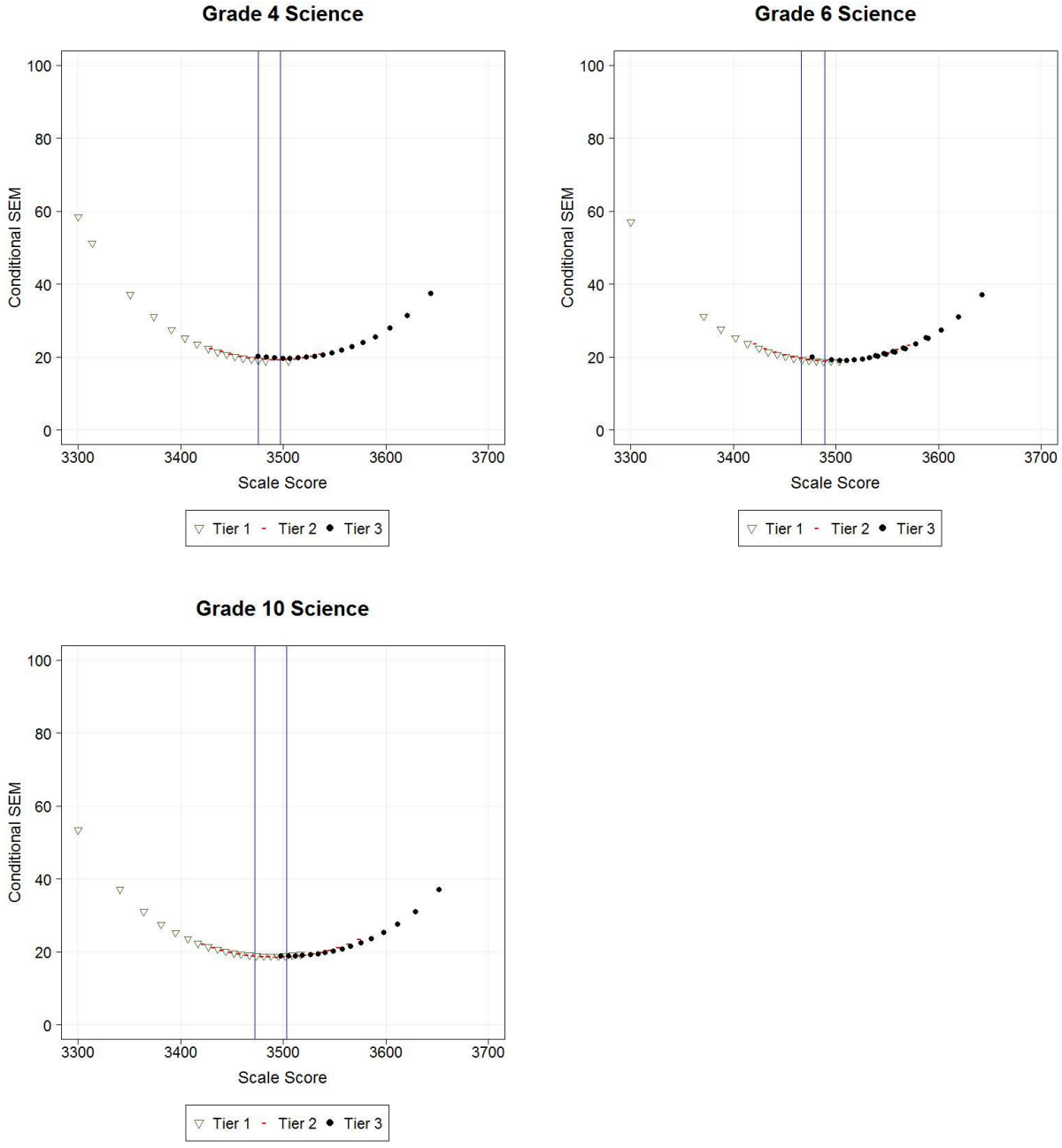
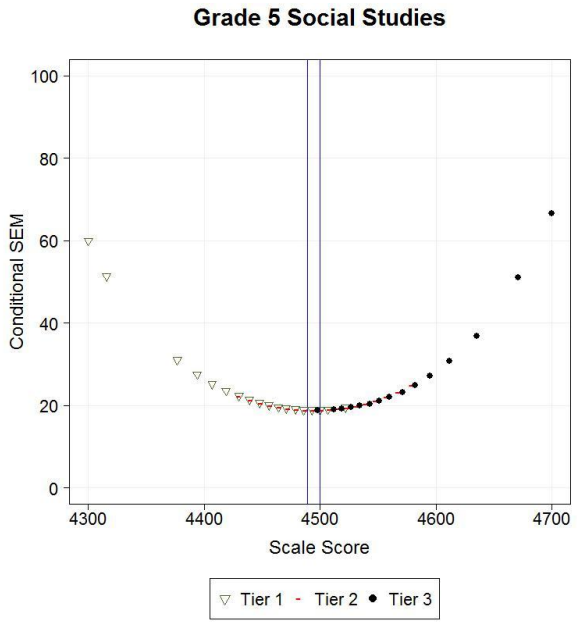


Figure 5: Conditional Standard Errors of Measurement (Social Studies)



Generally, the relationship between CSEM and scale score is U-shaped, with large CSEM values at the lower and upper ends of the scale score range and the smallest CSEMs in the middle, approximately at the Approaching Proficiency and At Proficiency cut scores. The middle section of the scale scores has most of the measurement information, and therefore, the CSEM remains low. The higher CSEMs at the lower and upper ends indicate a lack of easier or harder items compared to student ability.

Reliability coefficients by demographic subgroups are also presented in Appendix A, Reliability Coefficients. Appendix B, Conditional Standard Error of Measurement, includes scale-score-by-scale-score CSEMs and corresponding performance levels for each scale score.

6.3 Reliability of Performance Classification

When students complete an I AM assessment, they are placed into performance levels depending upon their observed scaled score. The cut scores for student classification into the different performance levels were determined after the I AM standard-setting process. A complete description of the standard-setting process can be found in Volume 6, *Standard Setting Report*.

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

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 score 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 score point?
2. What is the probability that the student's true score is above the cut score 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 I AM population data files that included the status of reported scores. Table 33 provides the sample size, mean, and standard deviation of the observed theta scores. The theta scores are based on the MLEs obtained from the American Institutes for Research's (AIR's) scoring engine.

Table 33: Descriptive Statistics by Test

Test	Sample Size	Average Theta	Standard Deviation of Theta	Average Scale Score	Standard Deviation of Scale Score
ELA 3	713	−0.51	0.80	1474.20	40.19
ELA 4	772	−0.22	1.05	1488.57	52.36
ELA 5	818	−0.16	1.09	1491.37	54.64
ELA 6	961	−0.25	0.97	1486.81	48.59
ELA 7	986	0.03	0.96	1501.21	48.12
ELA 8	1103	−0.18	1.04	1490.78	52.05
ELA 10	1078	0.18	1.18	1508.75	59.16
Mathematics 3	709	−0.41	0.83	2479.14	41.25
Mathematics 4	766	−0.41	0.88	2479.12	44.05
Mathematics 5	809	−0.58	0.70	2470.49	35.10
Mathematics 6	953	−0.44	0.73	2477.35	36.41

Test	Sample Size	Average Theta	Standard Deviation of Theta	Average Scale Score	Standard Deviation of Scale Score
Mathematics 7	988	−0.52	0.57	2473.45	28.76
Mathematics 8	1101	−0.65	0.61	2467.26	30.54
Mathematics 10	1078	−0.50	0.68	2474.42	34.03
Science 4	766	−0.25	0.91	3487.01	45.68
Science 6	943	−0.27	0.81	3485.80	40.63
Biology	1013	−0.06	1.01	3496.55	50.18
Social Studies 5	802	−0.24	1.00	4487.30	49.87

6.3.1 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 $l = 1,2,\cdots,L$. This probability can be estimated from evaluating the integral

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

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 a SEM of $\hat{\sigma}_j$, and using the asymptotic property of normality of the MLE, $\hat{\theta}_j$, we take $f(\cdot)$ as asymmetrically normal. Therefore, the above 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 (CDF). The expected number of students at level *l* based on students from observed level *v* can be expressed as

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

where pl_j is the *j*th student's performance level, the values of E_{vl} are the elements used to populate the matrix E, a 3×3 matrix of conditionally expected numbers of students to score within each performance-level bin based on their true scores. The overall CAI of the test can then be estimated from the diagonal elements of the matrix

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

where $N = \sum_{v=1}^{3} N_v, N_v$ is the observed number of students scoring in performance level v. The classification accuracy index for the individual cut score, *p*, ($CAIC_p$) is estimated by forming square partitioned blocks of the matrix E and taking the summation over all elements within the block as follows:

$$CAIC_p = \frac{\left(\sum_{v=1}^{p} \quad \sum_{l=1}^{p} \quad E_{vl} + \sum_{v=p+1}^{3} \quad \sum_{l=p+1}^{3} \quad E_{vl}\right)}{N}.$$

The *p*th cut score is $p(p = 1, 2)$.

Table 34–Table 37 provide the overall CAI and the classification accuracy index for the individual cut scores $(CAIC)$ based on the observed score approach. Here, the overall classification accuracy of the test ranges from 0.722–0.804 for ELA, from 0.679–0.747 for Mathematics, and from 0.728–0.779 for Science. The overall classification accuracy of Social Studies grade 5 is 0.816. There is no industry standard, but these numbers suggest that misclassification would not be frequent in the population data.

The overall cut-score accuracy rates are much higher, denoting that the degree to which we can reliably differentiate students between adjacent performance levels is typically from 0.837–0.903 for ELA, Science, and Social Studies, and from 0.785–0.851 for Mathematics.

Table 34: Classification Accuracy Index (ELA)

Grade	Overall Accuracy Index	Cut Score Accuracy Index	
		Cut 1 and Cut 2	Cut 2 and Cut 3
3	0.722	0.837	0.840
4	0.790	0.877	0.889
5	0.804	0.878	0.885
6	0.758	0.862	0.875
7	0.778	0.858	0.872
8	0.760	0.874	0.872
10	0.800	0.903	0.895

Table 35: Classification Accuracy Index (Mathematics)

Grade	Overall Accuracy Index	Cut Score Accuracy Index	
		Cut 1 and Cut 2	Cut 2 and Cut 3
3	0.747	0.841	0.831
4	0.727	0.851	0.840
5	0.690	0.807	0.804
6	0.706	0.828	0.836
7	0.682	0.790	0.785
8	0.700	0.800	0.809
10	0.679	0.804	0.817

Table 36: Classification Accuracy Index (Science)

Grade	Overall Accuracy Index	Cut Score Accuracy Index	
		Cut 1 and Cut 2	Cut 2 and Cut 3
4	0.750	0.849	0.876
6	0.728	0.855	0.852
Biology	0.779	0.878	0.893

Table 37: Classification Accuracy Index (Social Studies)

Grade	Overall Accuracy Index	Cut Score Accuracy Index	
		Cut 1 and Cut 2	Cut 2 and Cut 3
5	0.816	0.864	0.888

6.3.2 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 independently twice (Lee, Hanson, and 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 and 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 I AM assessment 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 score as

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

where c is the cut score required for passing in the same assigned metric, θ_j is true ability in the true-score metric, z_j is the item score, μ is the mean, and σ is the standard deviation of the population distribution. The function $p(\theta_j)$ is the probability of the 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 score as

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

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 score based on their observed score, but their true score would otherwise have classified them as below the cut score. The FNR is expressed as the proportion of individuals who scored below the cut score based on their observed score, but otherwise would have been classified as above the cut score based on their true scores. These rates are estimated as follows:

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

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

Table 38–Table 41 provide the FPR and FNR for the I AM assessments. In ELA, Science, and Social Studies, the FPR and FNR rates for the level 1 and level 2 cut scores are around 4–9%, and the rates for the level 2 and level 3 cut scores are around 4–8%. In Mathematics, the rates are around 6–12% at the cut scores between levels 1 and 2 and 5–13% at the cut scores between levels 2 and 3.

Table 38: False Classification Rates (ELA)

	1/2 cut		2/3 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**
3	0.075	0.088	0.083	0.077
4	0.066	0.056	0.062	0.049
5	0.048	0.074	0.048	0.067
6	0.070	0.068	0.045	0.080
7	0.081	0.061	0.060	0.068
8	0.066	0.060	0.071	0.057
10	0.058	0.039	0.046	0.059

Table 39: False Classification Rates (Mathematics)

	1/2 cut		2/3 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**
3	0.095	0.064	0.098	0.071
4	0.087	0.061	0.070	0.090
5	0.106	0.087	0.083	0.113
6	0.103	0.069	0.071	0.093
7	0.093	0.117	0.107	0.108
8	0.095	0.105	0.092	0.100
10	0.088	0.108	0.052	0.130

Table 40: False Classification Rates (Science)

	1/2 cut		2/3 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**
4	0.061	0.090	0.041	0.083
6	0.088	0.057	0.080	0.068
Biology	0.050	0.071	0.044	0.062

Table 41: False Classification Rates (Social Studies)

	1/2 cut		2/3 cut	
Grade	**FPR**	**FNR**	**FPR**	**FNR**
5	0.045	0.091	0.049	0.063

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

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

Classification consistency with classification accuracy results are presented in Table 42 and Table 43. In cut score 1/2 and cut score 2/3 results, all accuracy values are higher than 0.83, and consistency values are higher than 0.78 except for higher grades in Mathematics. 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 SEMs, while the accuracy is based on one test with an SEM and the true score. The accuracy and consistency rates for each performance level are higher for the levels with a smaller SEM.

Table 42: Classification Accuracy and Consistency (Cut Score 1 and Cut Score 2)

Grade	Accuracy	Consistency
ELA 3	0.837	0.776
ELA 4	0.877	0.828
ELA 5	0.878	0.830
ELA 6	0.862	0.811
ELA 7	0.858	0.806
ELA 8	0.874	0.824
ELA 10	0.903	0.865
Mathematics 3	0.841	0.782
Mathematics 4	0.851	0.792
Mathematics 5	0.807	0.739

Grade	Accuracy	Consistency
Mathematics 6	0.828	0.766
Mathematics 7	0.790	0.717
Mathematics 8	0.800	0.728
Mathematics 10	0.804	0.733
Science 4	0.849	0.789
Science 6	0.855	0.798
Biology	0.878	0.831
Social Studies 5	0.864	0.808

Table 43: Classification Accuracy and Consistency (Cut Score 2 and Cut Score 3)

Grade	Accuracy	Consistency
ELA 3	0.840	0.780
ELA 4	0.889	0.841
ELA 5	0.885	0.838
ELA 6	0.875	0.825
ELA 7	0.872	0.818
ELA 8	0.872	0.824
ELA 10	0.895	0.852
Mathematics 3	0.831	0.768
Mathematics 4	0.840	0.783
Mathematics 5	0.804	0.733
Mathematics 6	0.836	0.772
Mathematics 7	0.785	0.709
Mathematics 8	0.809	0.737
Mathematics 10	0.817	0.750
Science 4	0.876	0.825
Science 6	0.852	0.795
Biology	0.893	0.847
Social Studies 5	0.888	0.840

6.4 PRECISION AT CUT SCORES

Table 44–Table 47 present the mean CSEM at each performance level by grade and subject. These tables also include performance-level cut scores and associated CSEM. The different CSEMs at the same scale score can be estimated due to the different item sets that students took in Part 2 tests. The range of CSEMs are provided across tier tests. The I AM assessment scores are somewhat more precise for test scores near the middle of the scale, especially around the At Proficiency performance standard cut score. The tables below also show that test scores remain precise even for students in the lowest and highest performance levels.

Table 44: Performance Levels and Associated Conditional Standard Error of Measurement (ELA)

Grade	Performance Level	*N*	Mean CSEM	CSEM at Cut Scores*
3	1	281	22.15	
	2	120	18.73	19.09–19.53
	3	312	18.54	18.38–18.59
4	1	328	21.06	
	2	114	18.36	18.25–18.69
	3	330	20.41	18.27–18.40
5	1	340	21.04	
	2	84	18.31	18.39–18.63
	3	394	20.74	18.26–18.35
6	1	343	21.05	
	2	201	18.61	18.79–19.23
	3	417	20.11	18.42–18.49
7	1	375	20.29	
	2	133	18.39	18.41–18.66
	3	478	20.64	18.36–18.46
8	1	341	21.24	
	2	235	18.52	18.70–19.12
	3	527	20.96	18.37–18.55
10	1	241	21.81	
	2	320	18.77	18.96–19.11
	3	517	23.27	18.77–19.08

* CSEM values can vary depending on the item sets that students took.

Table 45: Performance Levels and Associated Conditional Standard Error of Measurement (Mathematics)

Grade	Performance Level	*N*	Mean CSEM	CSEM at Cut Score*
3	1	217	21.65	
	2	84	18.83	18.85–19.91
	3	408	19.10	18.49–19.38
4	1	246	21.85	
	2	164	18.87	19.05–19.78
	3	356	19.38	18.46–18.92
5	1	285	21.50	
	2	146	18.95	19.04–19.57
	3	378	18.68	18.59–18.95
6	1	305	21.56	
	2	214	19.08	19.36–19.65
	3	434	19.04	18.72–19.00
7	1	417	20.46	
	2	120	18.65	18.81–19.27
	3	451	18.37	18.44–18.73
8	1	506	20.19	
	2	149	18.37	18.41–19.03
	3	446	18.30	18.10–18.52
10	1	498	21.55	
	2	241	18.74	19.10–19.61
	3	339	18.58	18.56–18.96

* CSEM values can vary depending on the item sets that students took.

Table 46: Performance Levels and Associated Conditional Standard Error of Measurement (Science)

Grade	Performance Level	*N*	Mean CSEM	CSEM at Cut Score*
4	1	351	21.41	
	2	150	19.49	19.09–20.18
	3	265	21.08	18.83–19.63
6	1	277	21.47	
	2	222	19.31	19.33–20.13
	3	444	19.85	18.77–19.19
Biology	1	355	20.90	
	2	239	18.80	18.94–19.55
	3	419	20.78	18.78–18.88

* CSEM values can vary depending on the item sets that students took.

Table 47: Performance Levels and Associated Conditional Standard Error of Measurement (Social Studies)

Grade	Performance Level	*N*	Mean CSEM	CSEM at Cut Score*
4	1	484	20.88	
	2	46	18.89	18.88–18.93
	3	272	21.85	18.82–18.91

* CSEM values can vary depending on the item sets that students took.

SUMMARY

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

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

REFERENCES

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

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

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 and J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Chen, F., Bollen, K. A., Paxton, P., Curran, P., & Kirby, J. 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.) (pp. 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 ed.). 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: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

van Driel, Otto P. (1978). On various causes of improper solutions in maximum likelihood Factor analysis. *Psychometrika, 43*, 225–43.

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's Alternate Measure

2018–2019

Volume 5
Score Interpretation Guide

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Hyesuk Jang, Yuan Hong, Katherine Krehbiel, Hashim Evans, and Celine Bryan. 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. I AM SCORE REPORTS

In Spring 2019, pursuant to Chapter 5 of Indiana Code 20-32-5, "Indiana Statewide Testing for Educational Progress," the following Indiana's Alternate Measure (I AM) assessments were administered to Indiana students: English/Language Arts (ELA) and Mathematics for grades 3–8 and 10; Science for grades 4 and 6, and Biology for high school; and Social Studies for grade 5.

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

1.1 OVERVIEW OF I AM SCORE REPORTS

I AM assessments were administered in Spring 2019. Scores from each Spring 2019 assessment were provided to corporations and schools through the ORS on August 15, 2019, after the standard setting took place on July 22–24, 2019. Grade 3 ELA reports were not deployed on schedule due to resolving a reporting category issue. These were posted on August 20, 2019. The ORS provides information on student performance and aggregated summaries at several levels—state, corporation, school, and roster.

The ORS is a web-based application that provides I AM results to users at various levels. Assessment results are available to users on the basis of their roles and the privileges granted to each authenticated user. There are four types of access: state, corporation, school, and teacher. Users at each level are granted drill-down access to reports in the system in accordance with their assigned role. This means that teachers can access data only for their roster(s) of students, schools can access data only for the students in their school, and corporations can access data for all schools and students in their corporation.

Users have the following types of access to the ORS:

- State users can access all state, corporation, school, teacher, and student data.
- Co-Op Corporation Administrators (Co-Op) and Corporation Test Coordinators (CTCs) have access to all data for their corporations and for the schools and students in their corporations.
- School Test Coordinators (STCs) and Principals (PR) have access to all data for their school and for the students in their school.
- Test Administrators (TAs) can access all aggregate data for their roster(s) and the students within their roster(s).

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

1.2 Overall Scores and Reporting Categories

Students receive a single scale score for each subject assessed if there is a valid score to report. The validity of a score is determined using attempted rules, which define a set of parameters under which a student's attemptedness may be counted. All students begin the assessment with two practice items. Next, students are presented with three operational items. Educators document if the students are able to engage with these first five items. Students that are not able to engage are marked as no response. If the student has Mark as No Response (NR) for the first two practice items and first three operational items a student's score will be assigned as No Mode of Communication (NMC). For the students not assigned as NMC if the student fails to respond to at least five test items in segments 1 through 3, they will be assigned as Undetermined (UND). Students are counted as "attempted" if they are not assigned NMC or UND. Attempted students will be scored and provided a proficiency level designation. Normally, a student takes an assessment in the Test Delivery System (TDS) and then submits it. Once the assessment is completed in TDS, the assessment score is reported in ORS. However, assessments may also be manually invalidated before reaching the ORS if assessment irregularities occur (e.g., cheating, unscheduled interruptions).

A student's score is based only on the operational items on the assessment. A scale score is used to describe how well a student performed on an assessment and is an estimate of students' 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 Annual Technical Report. Lower scale scores indicate that the student does not possess sufficient knowledge and skills measured by the assessment. Conversely, higher scale scores indicate that the student has more proficient knowledge and skills measured by the assessment. Interpretation of scale scores is more meaningful when the scale scores are used along with performance levels and performance-level descriptors.

A student's scale score determines his or her overall performance level. Performance levels on an assessment correlate with proficiency categories into which students fall on the basis of their scale scores. For I AM, scale scores are mapped onto three performance levels:

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

Performance-level descriptors (PLDs) set out content-area knowledge and skills that students at each performance level are expected to possess; and cut scores, unique to each grade and subject, are determined by using Cut points are listed in Section 2.4, Cut Scores; additional details can be found in Volume 6: Standard-Setting Report.

Performance levels can be interpreted on the basis of PLDs, which represent a more descriptive analysis of a student's abilities. Generally, students performing on I AM at Level 3 have met current grade-level Content Connectors by demonstrating essential knowledge, application, and skills to be on track for post-secondary education or competitive integrated employment.

In addition to an overall score, students receive reporting category scores. Reporting categories represent distinct groups of knowledge within each grade and subject. For I AM, students' performance in each reporting category is reported using a percent-correct score for each student and an average percent-correct score for aggregate reporting. Tables 1 through 4 display the reporting categories by grade and subject.

Table 1: Reporting Categories for ELA

Grade	Reporting Category
3	1. Key Ideas and Textual Support/Vocabulary 2. Reading Foundations 3. Structural Elements and Organization/ Connection of Ideas/Media Literacy 4. Writing
4	1. Key Ideas and Textual Support/Vocabulary 2. Structural Elements and Organization/ Connection of Ideas/Media Literacy 3. Writing
5	1. Key Ideas and Textual Support/Vocabulary 2. Structural Elements and Organization/ Connection of Ideas/Media Literacy 3. Writing
6	1. Key Ideas and Textual Support/Vocabulary 2. Structural Elements and Organization/ Connection of Ideas/Media Literacy 3. Writing
7	1. Key Ideas and Textual Support/Vocabulary 2. Structural Elements and Organization/ Connection of Ideas/Media Literacy 3. Writing
8	1. Key Ideas and Textual Support/Vocabulary 2. Structural Elements and Organization/ Connection of Ideas/Media Literacy 3. Writing
10	1. Key Ideas and Textual Support/Vocabulary 2. Structural Elements and Organization/ Connection of Ideas/Media Literacy 3. Writing

Table 2: Reporting Categories for Mathematics

Grade	Reporting Category
3	1. Algebraic Thinking and Data Analysis 2. Computation 3. Geometry and Measurement 4. Number Sense
4	1. Algebraic Thinking and Data Analysis 2. Computation 3. Geometry and Measurement 4. Number Sense
5	1. Algebraic Thinking 2. Computation 3. Geometry and Measurement, Data Analysis, and Statistics 4. Number Sense
6	1. Algebra and Functions 2. Computation 3. Geometry and Measurement, Data Analysis, and Statistics 4. Number Sense
7	1. Algebra and Functions 2. Data Analysis, Statistics, and Probability 3. Geometry and Measurement 4. Number Sense and Computation
8	1. Algebra and Functions 2. Data Analysis, Statistics, and Probability 3. Geometry and Measurement 4. Number Sense and Computation
10	1. Equations and Inequalities (Linear and Systems) 2. Functions (Linear and Non-Linear) 3. Number Sense and Data Analysis 4. Geometry and Measurement

Table 3: Reporting Categories for Science

Grade	Reporting Category
4	1. Analyzing, Interpreting, and Computational Thinking 2. Explaining Solutions, Reasoning, and Communicating 3. Investigating 4. Questioning and Modeling
6	1. Analyzing, Interpreting, and Computational Thinking 2. Explaining Solutions, Reasoning, and Communicating 3. Investigating 4. Questioning and Modeling
Biology	1. Analyzing Data and Mathematical Thinking 2. Communicating Explanations and Evaluating Claims Using Evidence 3. Developing and Using Modeling to Describe Structure and Function

Table 4: Reporting Categories for Social Studies

Grade	Reporting Category
5	1. Civics and Government/History 2. Economics 3. Geography

1.3 ONLINE REPORTING SYSTEM

The 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 assessments are submitted by the students and processed into the ORS. In addition to each individual student's score report, the ORS produces aggregate score reports for teachers, schools, corporations, and states. The timely accessibility of aggregate score reports helps users monitor student 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 in the hierarchy. 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 so that the school performance can be compared with the corporation performance and the state 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, Available Reports on the Indiana Online Reporting System) lists the types of online reports: student, roster, teacher, school, and corporation.

1.4 AVAILABLE REPORTS ON THE I AM ONLINE REPORTING SYSTEM

The hierarchical structure of the Indiana ORS enables authorized users to view reports at their own level and at any lower level of aggregation. For example, an STC can view only the reports and data for his or her own school and for the students at the school. CTCs can view the reports and data for all schools and students in their corporations.

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. An *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
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 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 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 for 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. Table 6 presents the types of sub-groups and sub-group categories provided in the ORS.

Table 6: Indiana List of Sub-Groups by Category

Sub-Group Category	Sub-Group
Race/Ethnicity	White
	Black/African American
	Hispanic
	Asian
	American Indian/Alaska Native
	Native Hawaiian/Other Pacific Islander
	Multiracial/Two or More Races
Gender	Male
	Female
English Learner	English Learner
	Not English Learner
Special Education	Special Education
	Not Special Education
Section 504 Plan	Section 504 Plan
	Not Section 504 Plan
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

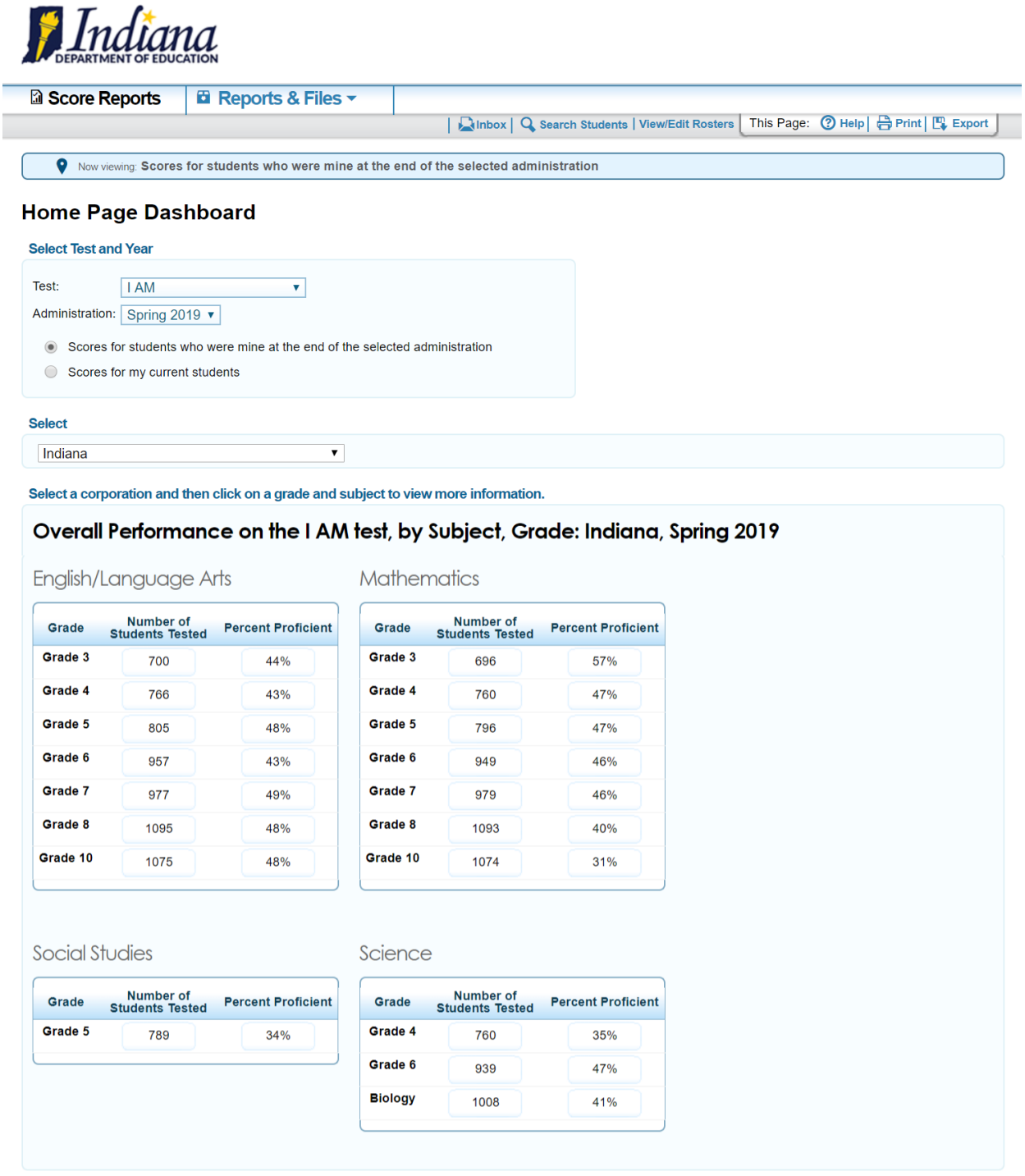
- displays summary data separated by grade and subject,
- bases the level of aggregation on a user's role, and
- reports the number of students assessed and percentage proficient.

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

- number of students assessed; and
- percentage proficient.

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

Figure 1: Sample State Summary Performance Report



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: I AM

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 I AM test, by Subject, Grade: Indiana, Spring 2019

English/Language Arts

Grade	Number of Students Tested	Percent Proficient
Grade 3	700	44%
Grade 4	766	43%
Grade 5	805	48%
Grade 6	957	43%
Grade 7	977	49%
Grade 8	1095	48%
Grade 10	1075	48%

Mathematics

Grade	Number of Students Tested	Percent Proficient
Grade 3	696	57%
Grade 4	760	47%
Grade 5	796	47%
Grade 6	949	46%
Grade 7	979	46%
Grade 8	1093	40%
Grade 10	1074	31%

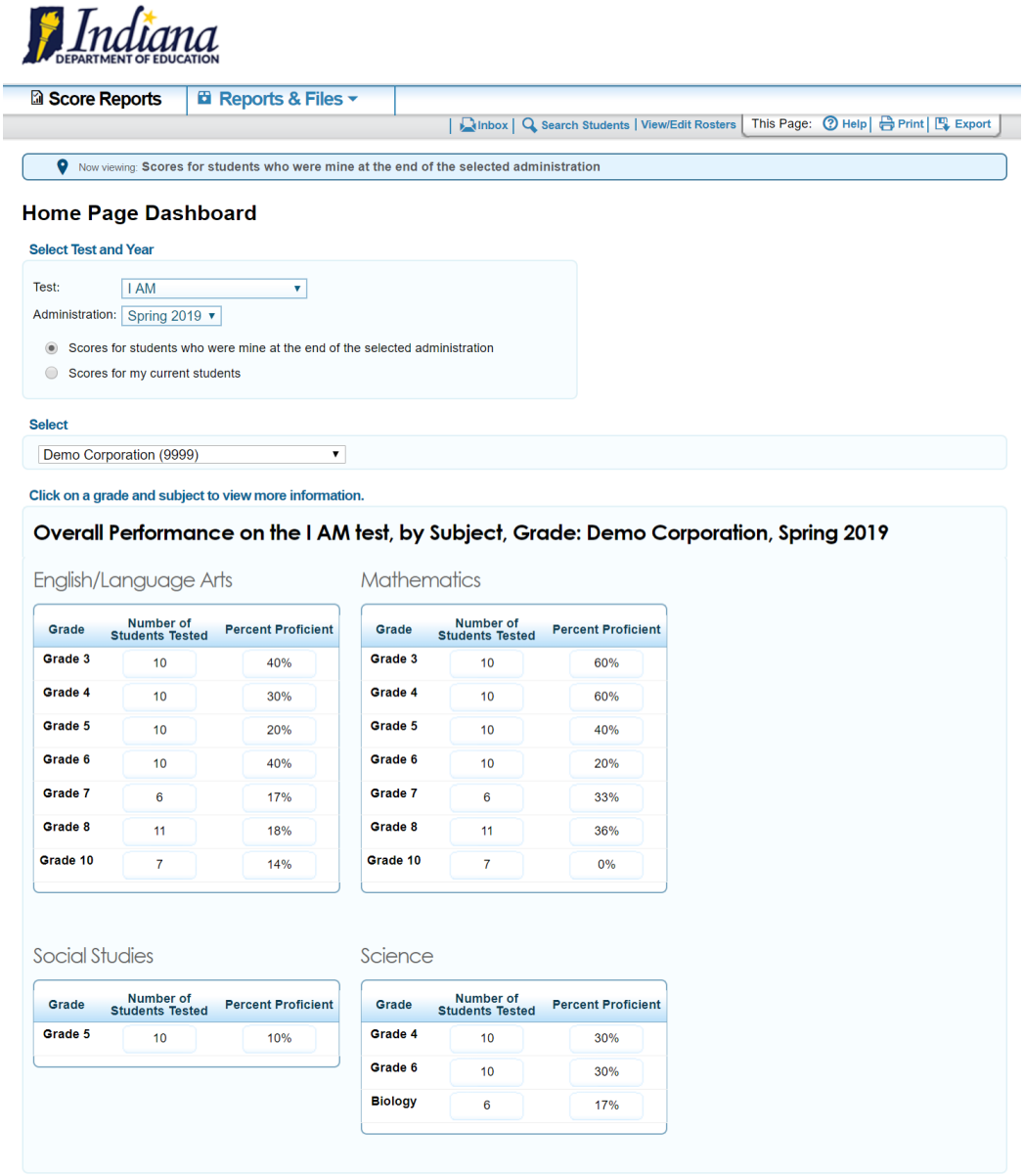
Social Studies

Grade	Number of Students Tested	Percent Proficient
Grade 5	789	34%

Science

Grade	Number of Students Tested	Percent Proficient
Grade 4	760	35%
Grade 6	939	47%
Biology	1008	41%

Figure 2: Corporation-Level Summary Performance Report



The Corporation Summary Report is similar to the State Summary, except that summary data are displayed for all students in the selected corporation who have completed the selected assessment 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 the:

- number of students;
- percentage proficient;
- number of students in each performance level; and
- percentage of students in each performance level.

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

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

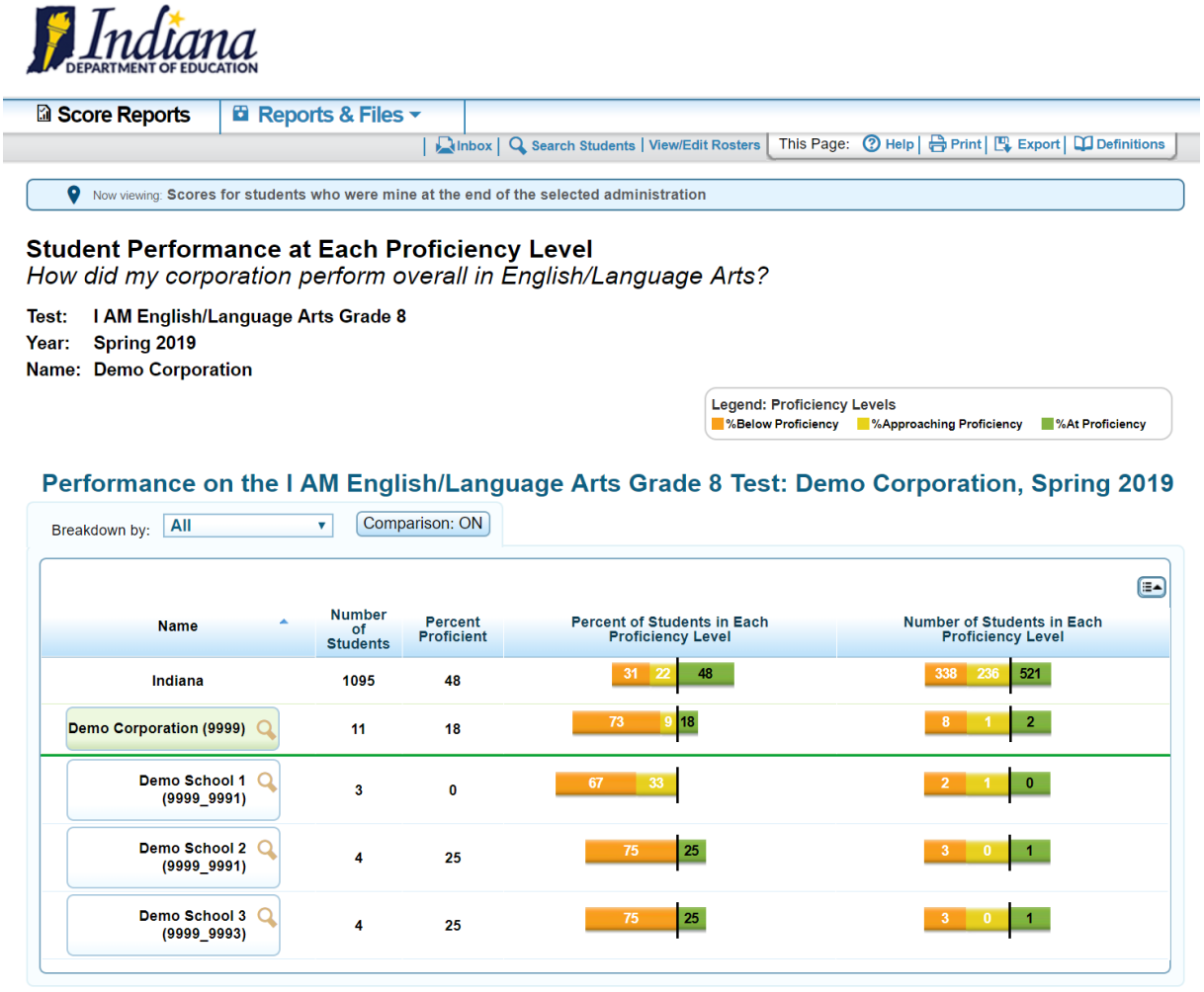
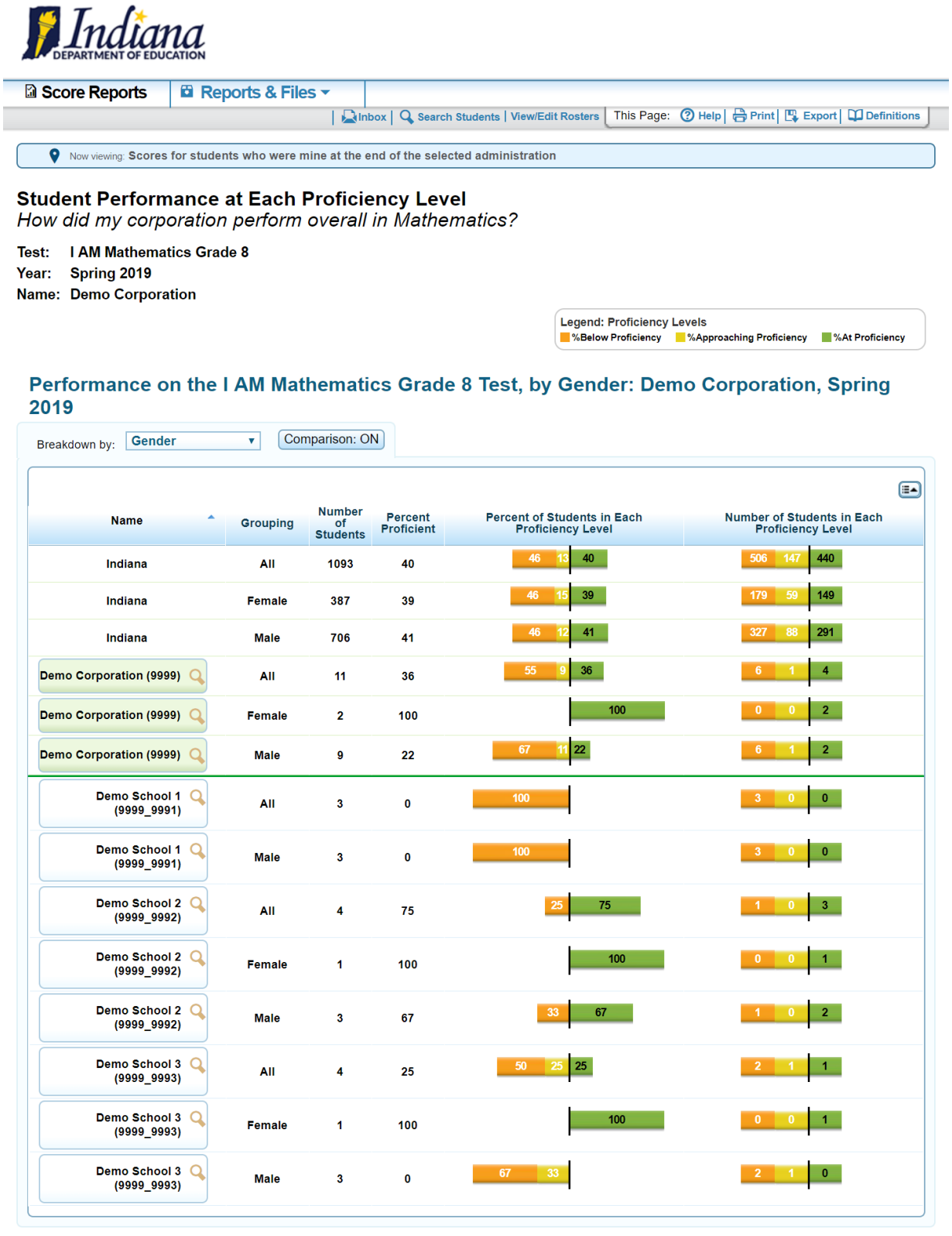


Figure 4: Corporation Aggregate-Level Subject Report, Grade 8 Mathematics by Gender



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 Mathematics?

Test: I AM Mathematics Grade 8
Year: Spring 2019
Name: Demo Corporation

Legend: Proficiency Levels
%Below Proficiency | %Approaching Proficiency | %At Proficiency

Performance on the I AM Mathematics Grade 8 Test, by Gender: Demo Corporation, Spring 2019

Breakdown by: Gender | Comparison: ON

Name	Grouping	Number of Students	Percent Proficient	Percent of Students in Each Proficiency Level	Number of Students in Each Proficiency Level
Indiana	All	1093	40	46 / 13 / 40	506 / 147 / 440
Indiana	Female	387	39	46 / 15 / 39	179 / 59 / 149
Indiana	Male	706	41	46 / 12 / 41	327 / 88 / 291
Demo Corporation (9999)	All	11	36	55 / 9 / 36	6 / 1 / 4
Demo Corporation (9999)	Female	2	100	100	0 / 0 / 2
Demo Corporation (9999)	Male	9	22	67 / 11 / 22	6 / 1 / 2
Demo School 1 (9999_9991)	All	3	0	100	3 / 0 / 0
Demo School 1 (9999_9991)	Male	3	0	100	3 / 0 / 0
Demo School 2 (9999_9992)	All	4	75	25 / 75	1 / 0 / 3
Demo School 2 (9999_9992)	Female	1	100	100	0 / 0 / 1
Demo School 2 (9999_9992)	Male	3	67	33 / 67	1 / 0 / 2
Demo School 3 (9999_9993)	All	4	25	50 / 25 / 25	2 / 1 / 1
Demo School 3 (9999_9993)	Female	1	100	100	0 / 0 / 1
Demo School 3 (9999_9993)	Male	3	0	67 / 33	2 / 1 / 0

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

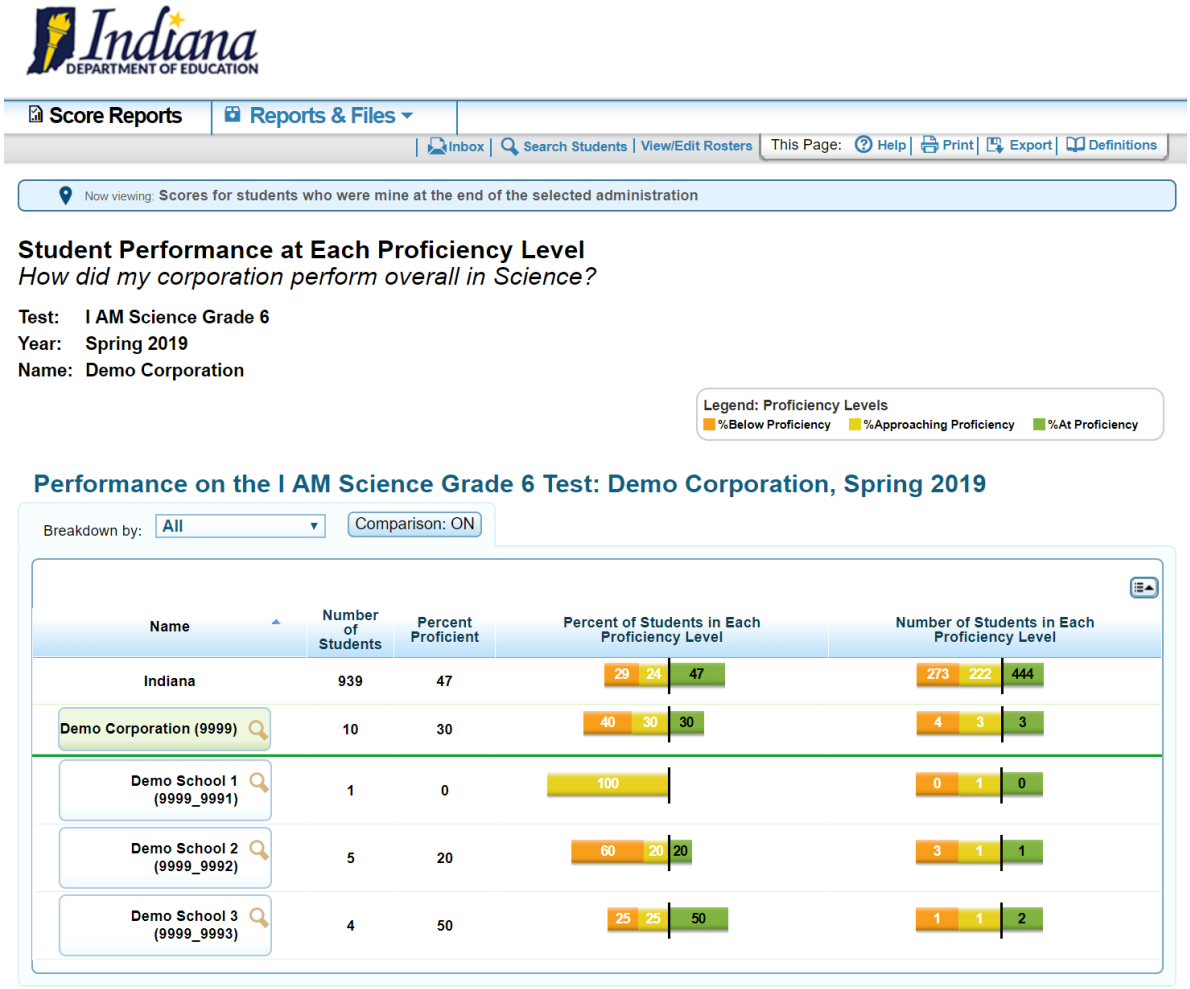
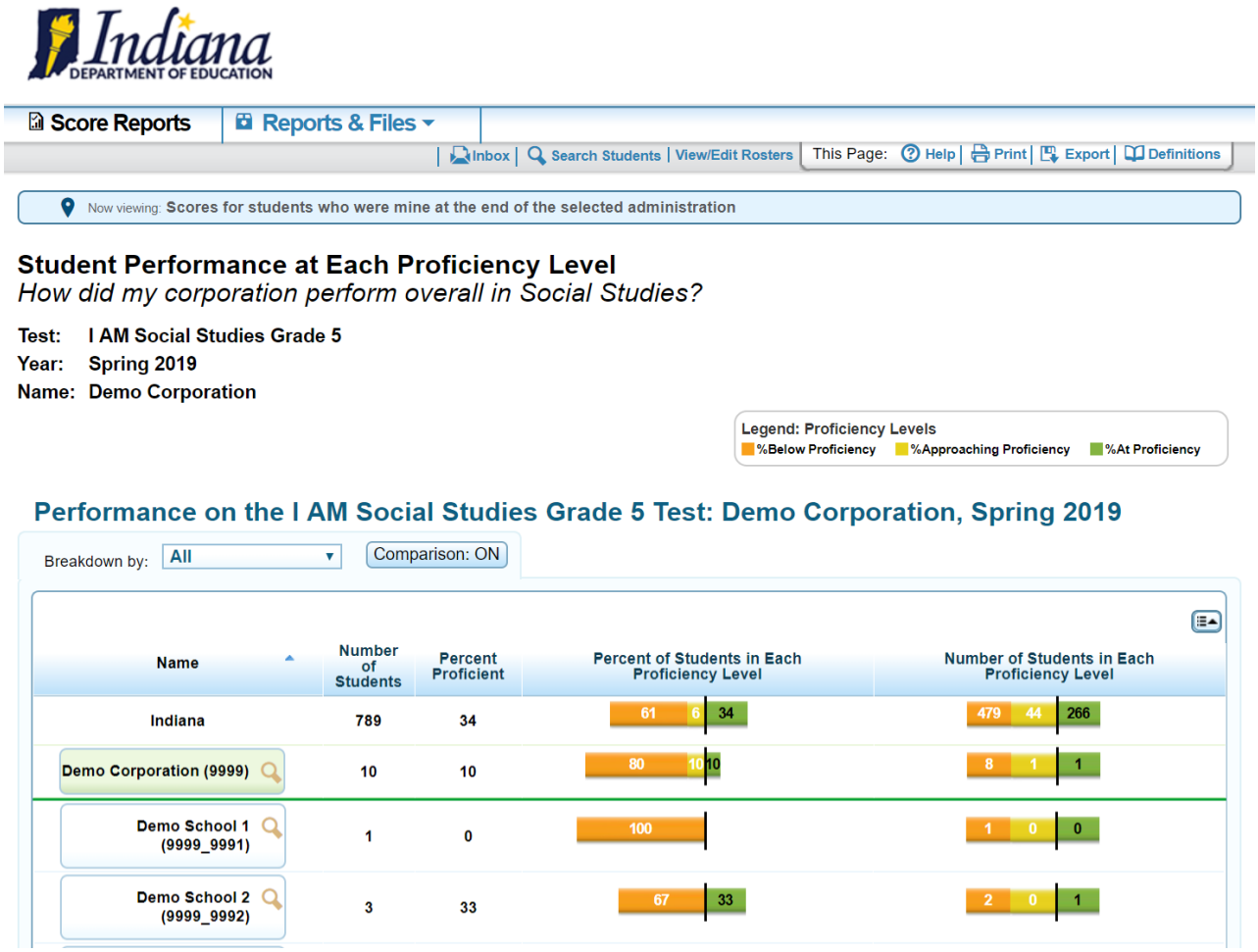


Figure 6: Corporation Aggregate-Level Subject Report, Grade 5 Social Studies



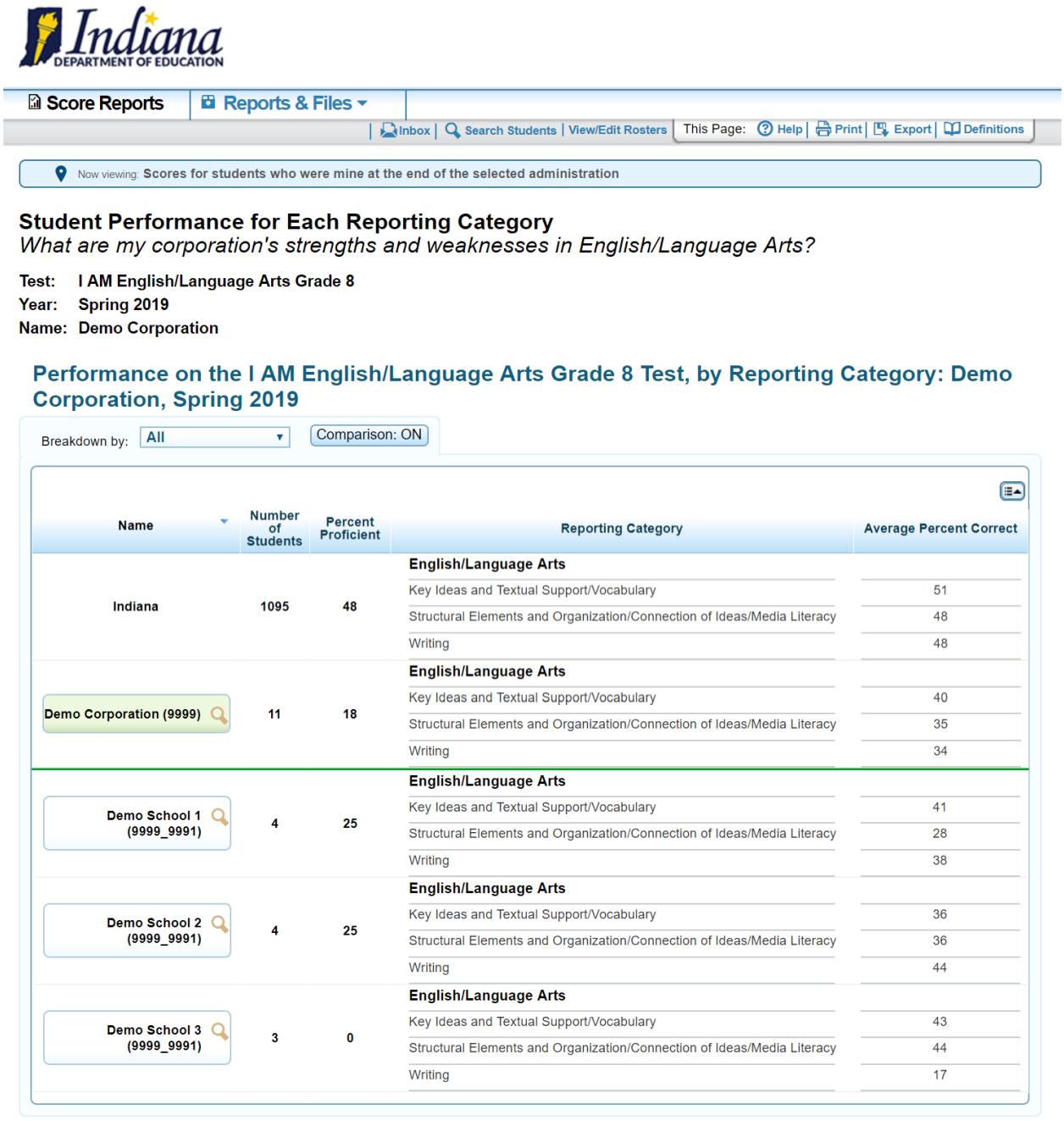
1.6.3 Aggregate-Level Reporting Category Report

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 the

- number of students,
- percent proficient, and
- average percent correct for each reporting 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 I AM.

Figure 7: Corporation Aggregate-Level Reporting Category Report for 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?

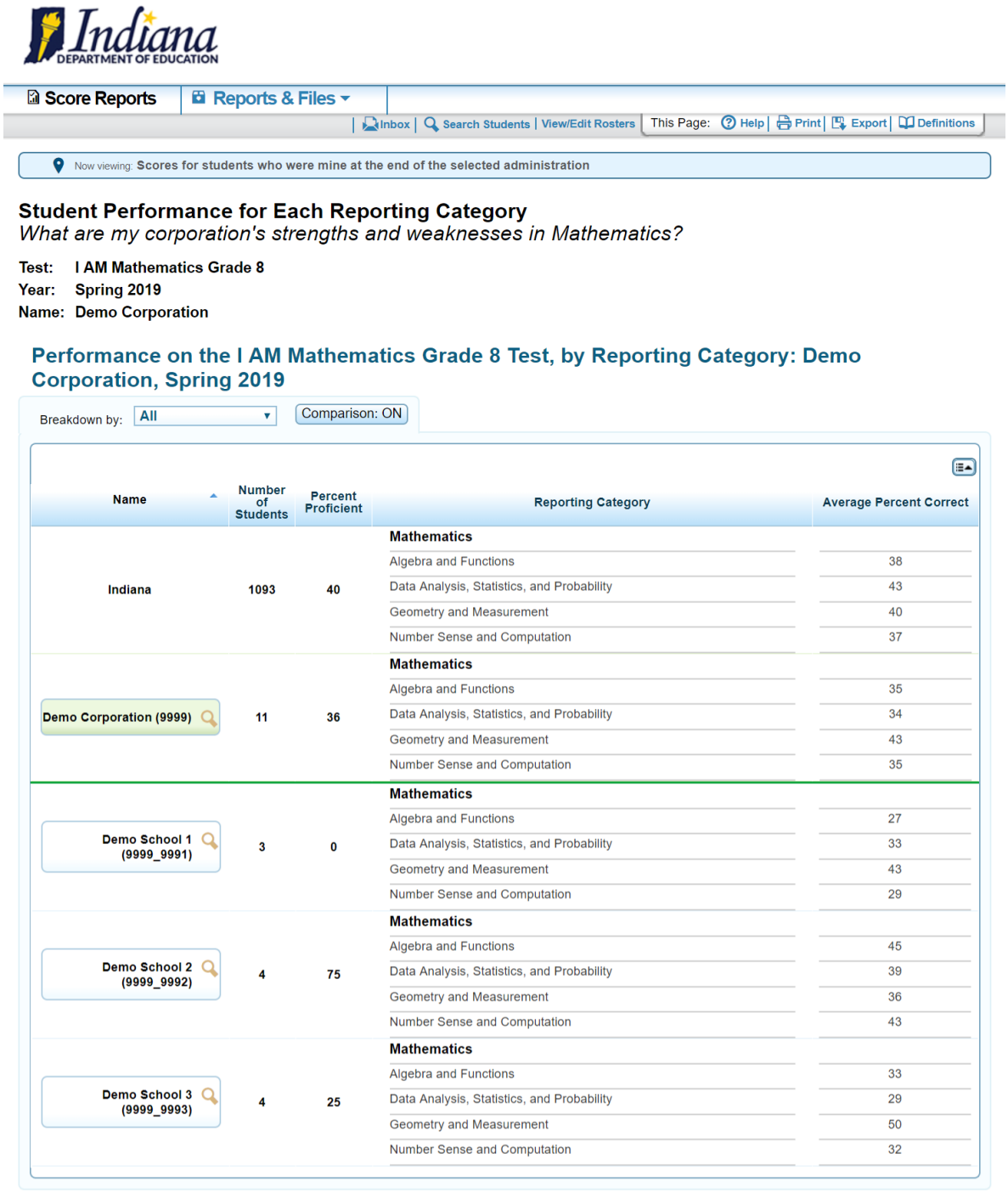
Test: I AM English/Language Arts Grade 8
Year: Spring 2019
Name: Demo Corporation

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

Breakdown by: All | Comparison: ON

Name	Number of Students	Percent Proficient	Reporting Category	Average Percent Correct
Indiana	1095	48	**English/Language Arts**	
			Key Ideas and Textual Support/Vocabulary	51
			Structural Elements and Organization/Connection of Ideas/Media Literacy	48
			Writing	48
Demo Corporation (9999)	11	18	**English/Language Arts**	
			Key Ideas and Textual Support/Vocabulary	40
			Structural Elements and Organization/Connection of Ideas/Media Literacy	35
			Writing	34
Demo School 1 (9999_9991)	4	25	**English/Language Arts**	
			Key Ideas and Textual Support/Vocabulary	41
			Structural Elements and Organization/Connection of Ideas/Media Literacy	28
			Writing	38
Demo School 2 (9999_9991)	4	25	**English/Language Arts**	
			Key Ideas and Textual Support/Vocabulary	36
			Structural Elements and Organization/Connection of Ideas/Media Literacy	36
			Writing	44
Demo School 3 (9999_9991)	3	0	**English/Language Arts**	
			Key Ideas and Textual Support/Vocabulary	43
			Structural Elements and Organization/Connection of Ideas/Media Literacy	44
			Writing	17

Figure 8: Corporation Aggregate-Level Reporting Category Report for 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?

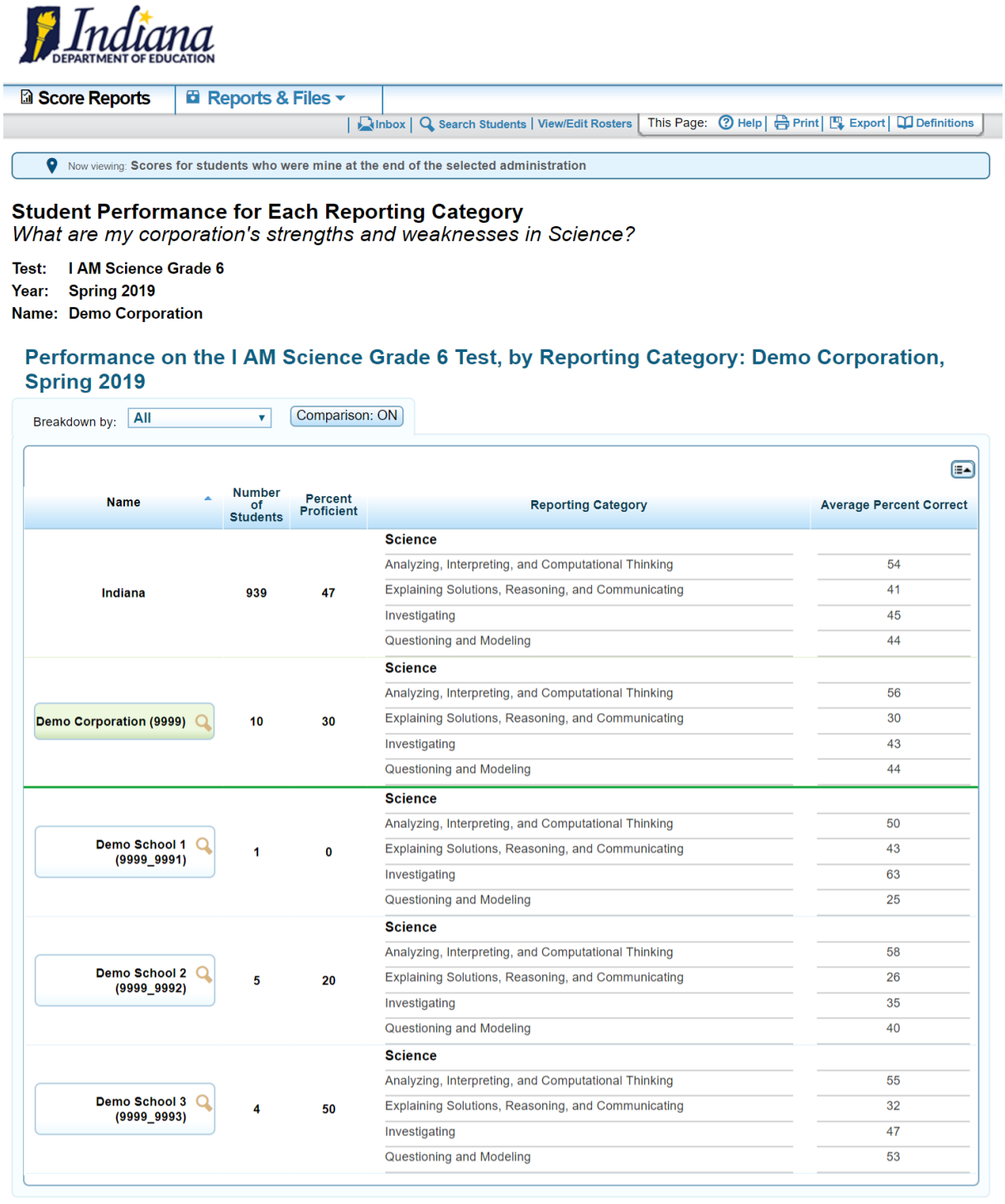
Test: I AM Mathematics Grade 8
Year: Spring 2019
Name: Demo Corporation

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

Breakdown by: All | Comparison: ON

Name	Number of Students	Percent Proficient	Reporting Category	Average Percent Correct
Indiana	1093	40	**Mathematics**	
			Algebra and Functions	38
			Data Analysis, Statistics, and Probability	43
			Geometry and Measurement	40
			Number Sense and Computation	37
Demo Corporation (9999)	11	36	**Mathematics**	
			Algebra and Functions	35
			Data Analysis, Statistics, and Probability	34
			Geometry and Measurement	43
			Number Sense and Computation	35
Demo School 1 (9999_9991)	3	0	**Mathematics**	
			Algebra and Functions	27
			Data Analysis, Statistics, and Probability	33
			Geometry and Measurement	43
			Number Sense and Computation	29
Demo School 2 (9999_9992)	4	75	**Mathematics**	
			Algebra and Functions	45
			Data Analysis, Statistics, and Probability	39
			Geometry and Measurement	36
			Number Sense and Computation	43
Demo School 3 (9999_9993)	4	25	**Mathematics**	
			Algebra and Functions	33
			Data Analysis, Statistics, and Probability	29
			Geometry and Measurement	50
			Number Sense and Computation	32

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



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?

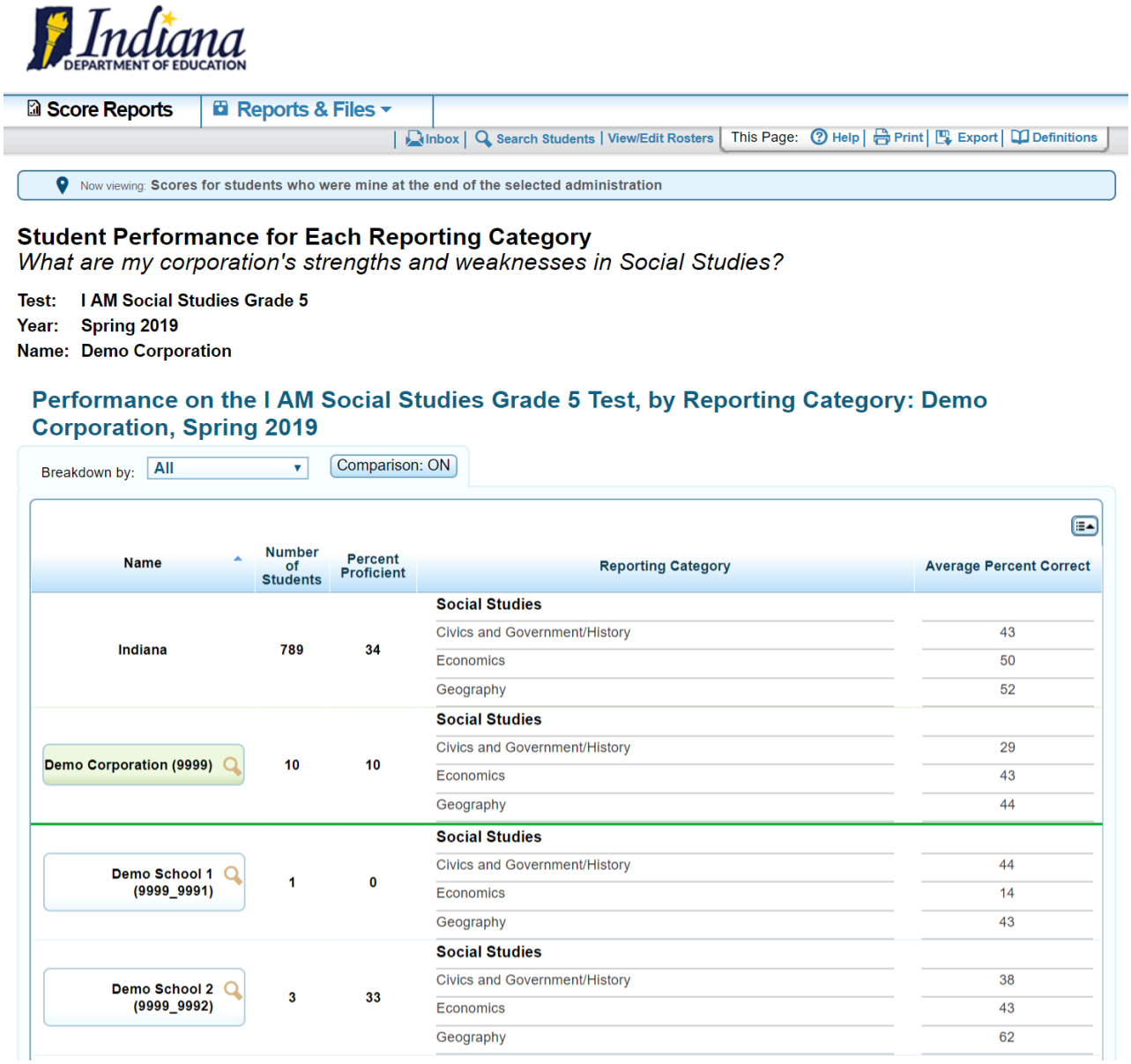
Test: I AM Science Grade 6
Year: Spring 2019
Name: Demo Corporation

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

Breakdown by: All | Comparison: ON

Name	Number of Students	Percent Proficient	Reporting Category	Average Percent Correct
Indiana	939	47	**Science**	
			Analyzing, Interpreting, and Computational Thinking	54
			Explaining Solutions, Reasoning, and Communicating	41
			Investigating	45
			Questioning and Modeling	44
Demo Corporation (9999)	10	30	**Science**	
			Analyzing, Interpreting, and Computational Thinking	56
			Explaining Solutions, Reasoning, and Communicating	30
			Investigating	43
			Questioning and Modeling	44
Demo School 1 (9999_9991)	1	0	**Science**	
			Analyzing, Interpreting, and Computational Thinking	50
			Explaining Solutions, Reasoning, and Communicating	43
			Investigating	63
			Questioning and Modeling	25
Demo School 2 (9999_9992)	5	20	**Science**	
			Analyzing, Interpreting, and Computational Thinking	58
			Explaining Solutions, Reasoning, and Communicating	26
			Investigating	35
			Questioning and Modeling	40
Demo School 3 (9999_9993)	4	50	**Science**	
			Analyzing, Interpreting, and Computational Thinking	55
			Explaining Solutions, Reasoning, and Communicating	32
			Investigating	47
			Questioning and Modeling	53

Figure 10: Corporation Aggregate-Level Reporting Category Report for 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: I AM Social Studies Grade 5
Year: Spring 2019
Name: Demo Corporation

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

Breakdown by: All | Comparison: ON

Name	Number of Students	Percent Proficient	Reporting Category	Average Percent Correct
Indiana	789	34	**Social Studies**	
			Civics and Government/History	43
			Economics	50
			Geography	52
Demo Corporation (9999)	10	10	**Social Studies**	
			Civics and Government/History	29
			Economics	43
			Geography	44
Demo School 1 (9999_9991)	1	0	**Social Studies**	
			Civics and Government/History	44
			Economics	14
			Geography	43
Demo School 2 (9999_9992)	3	33	**Social Studies**	
			Civics and Government/History	38
			Economics	43
			Geography	62

1.6.4 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:

- Overall subject scale score
- Overall subject performance level

Figure 11 through Figure 14 demonstrate examples of the Student-Level Subject Report for I AM.

Figure 11: Student-Level Subject Report for Grade 8 ELA

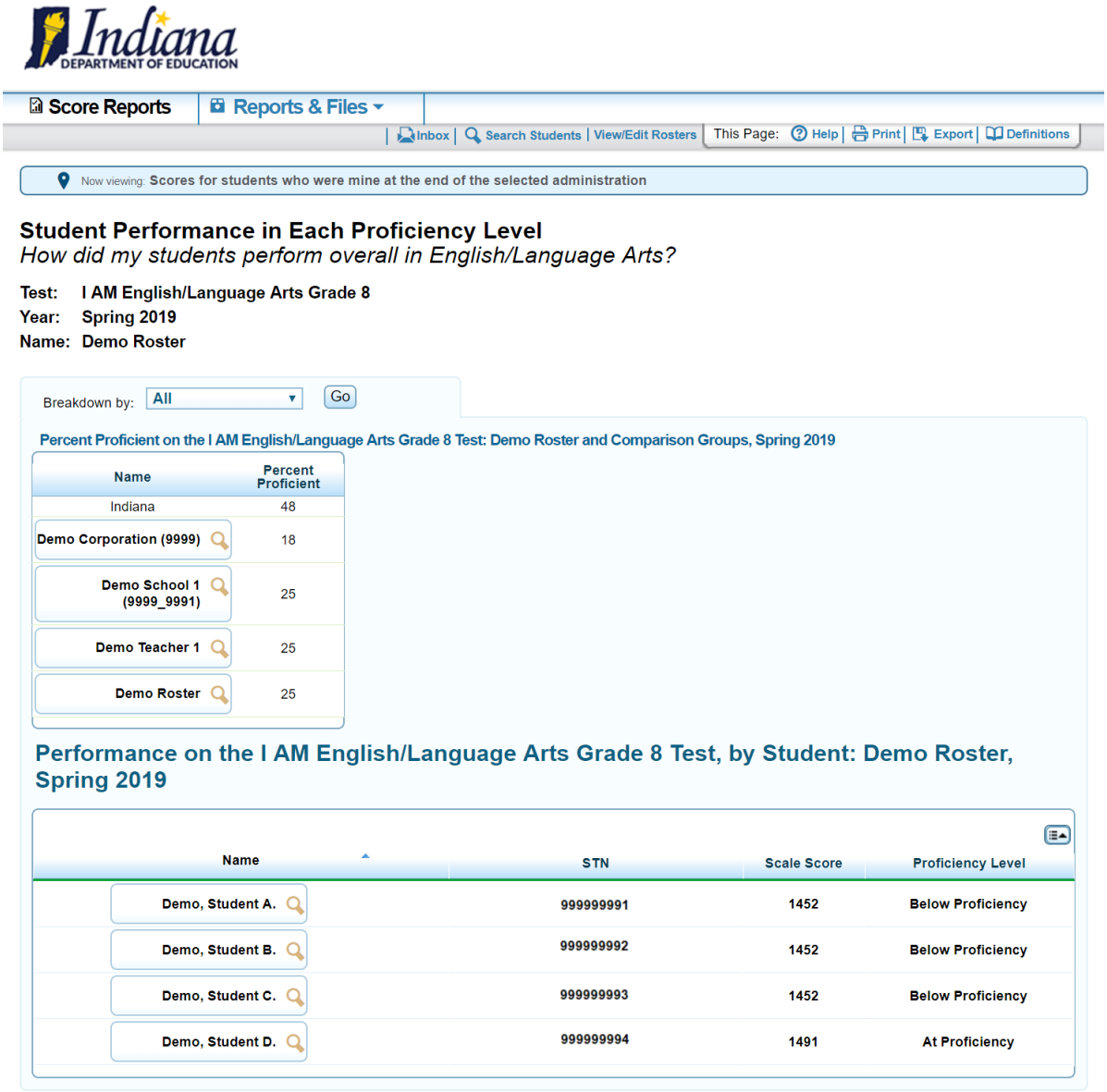


Figure 12: Student-Level Subject Report for Grade 8 Mathematics

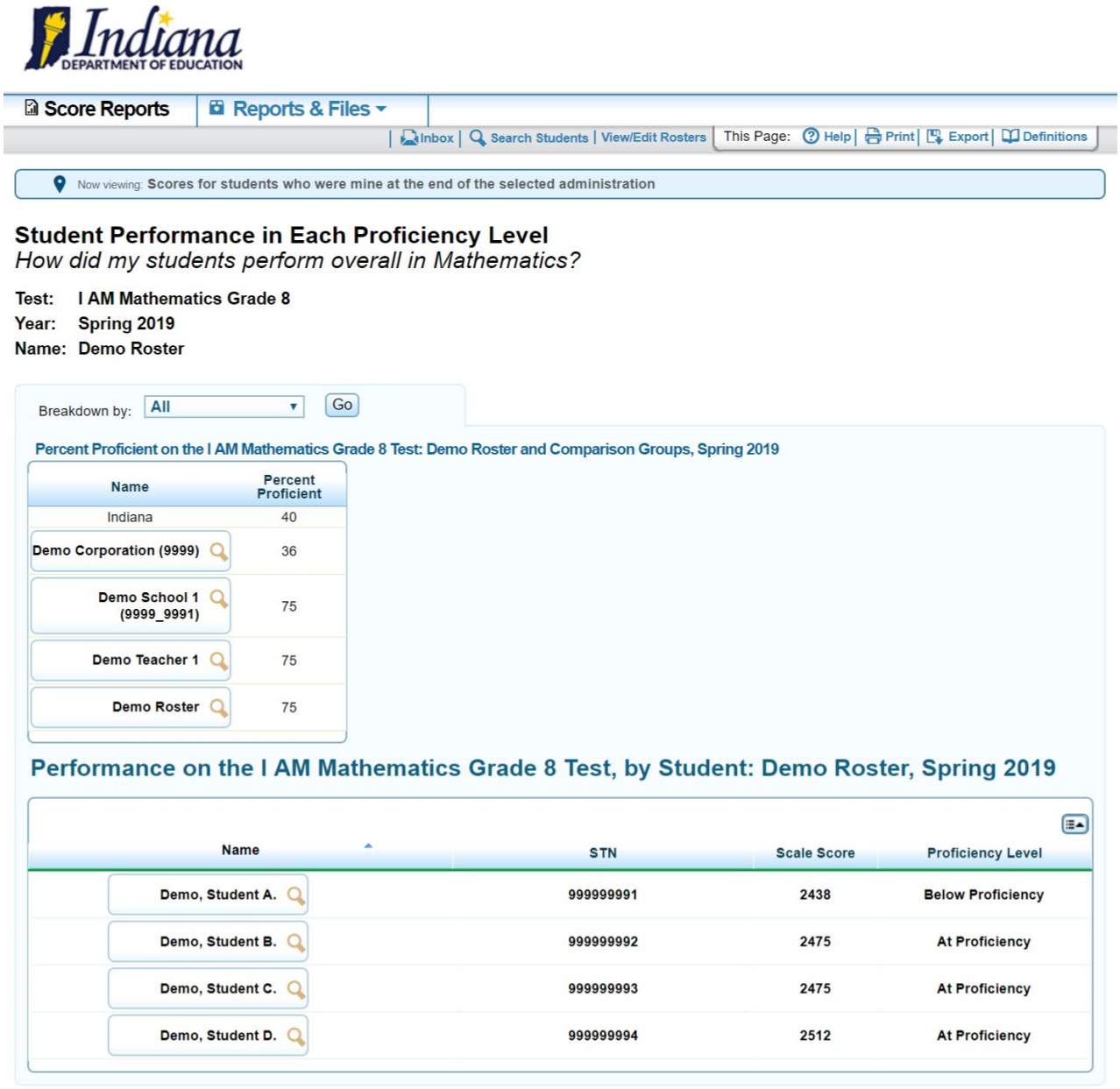


Figure 13: Student-Level Subject Report for Grade 6 Science

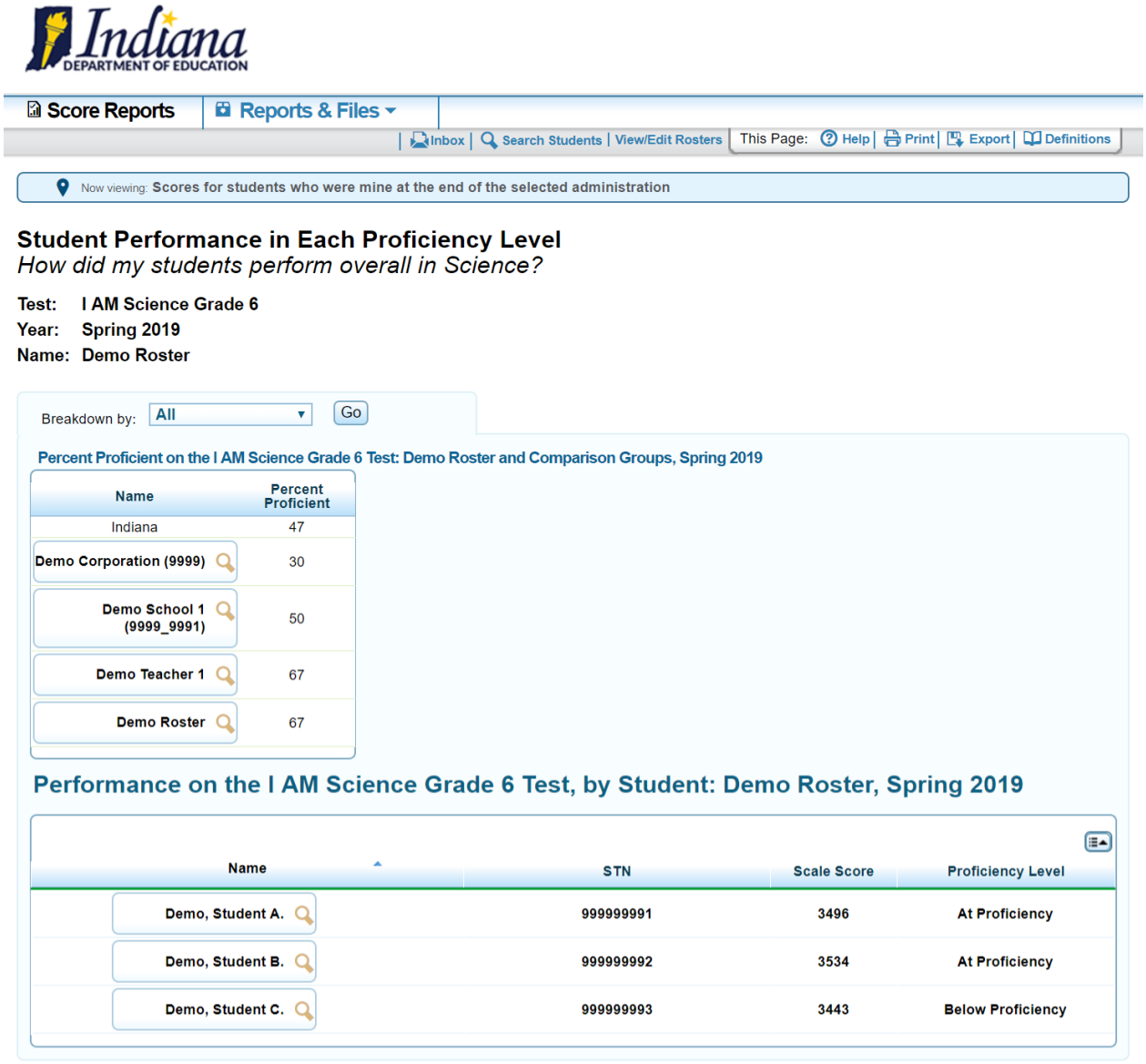
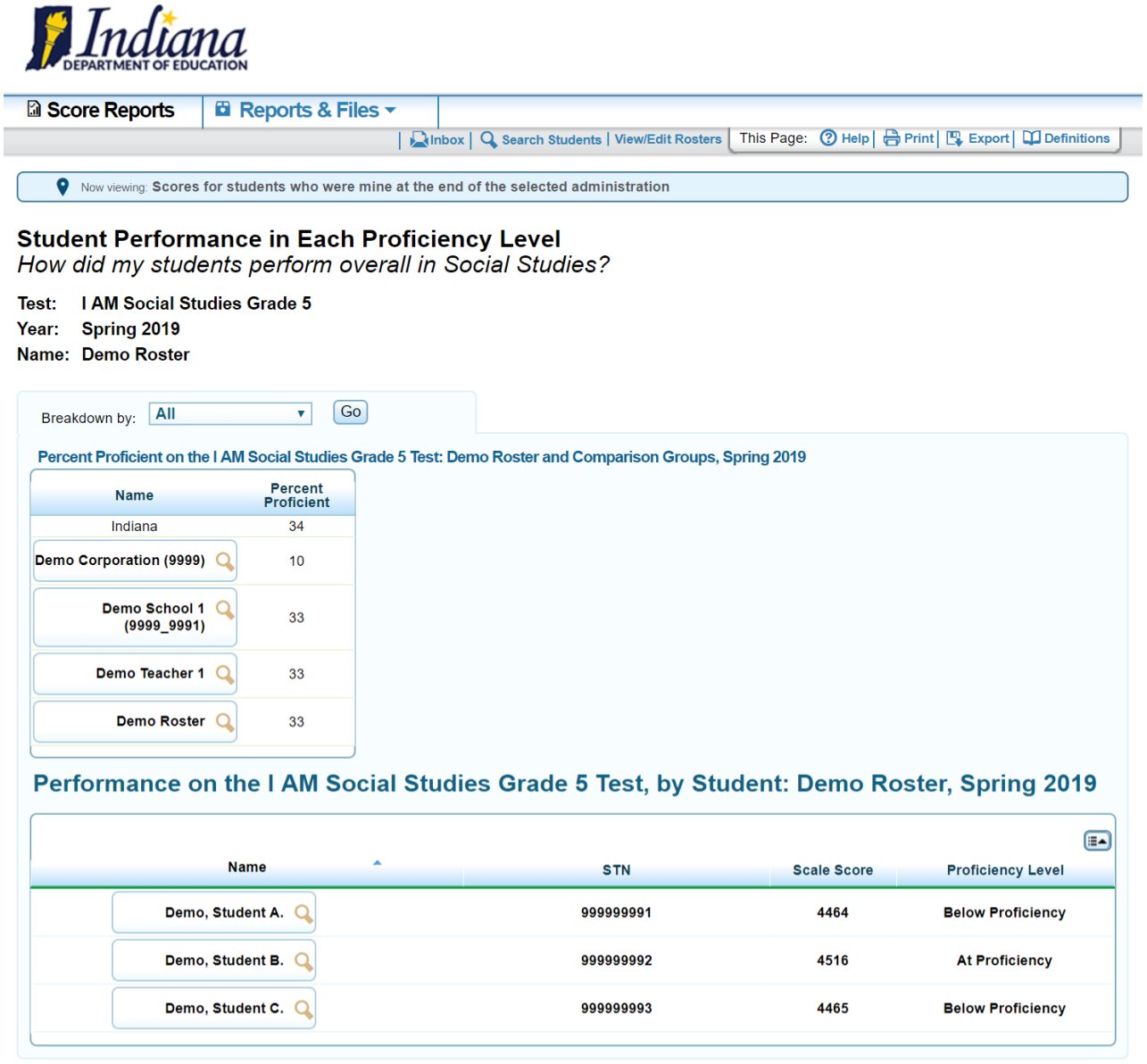


Figure 14: Student-Level Subject Report for Grade 5 Social Studies



1.6.5 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:

- Overall subject scale score
- Overall subject performance level
- Reporting category percent correct

Figure 15 through Figure 18 displays this information for I AM.

Figure 15: Student-Level Reporting Category Report for Grade 8 ELA

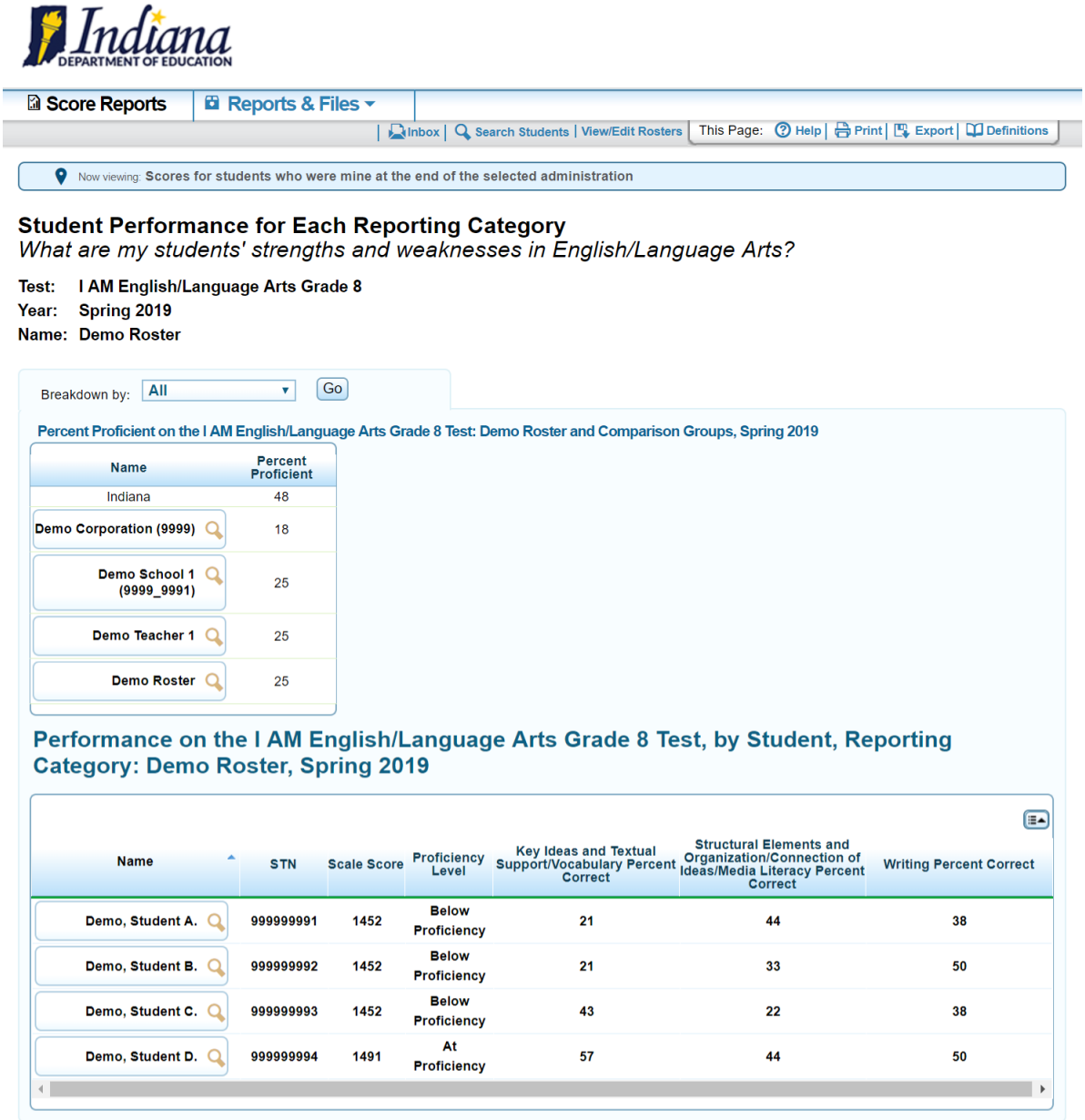
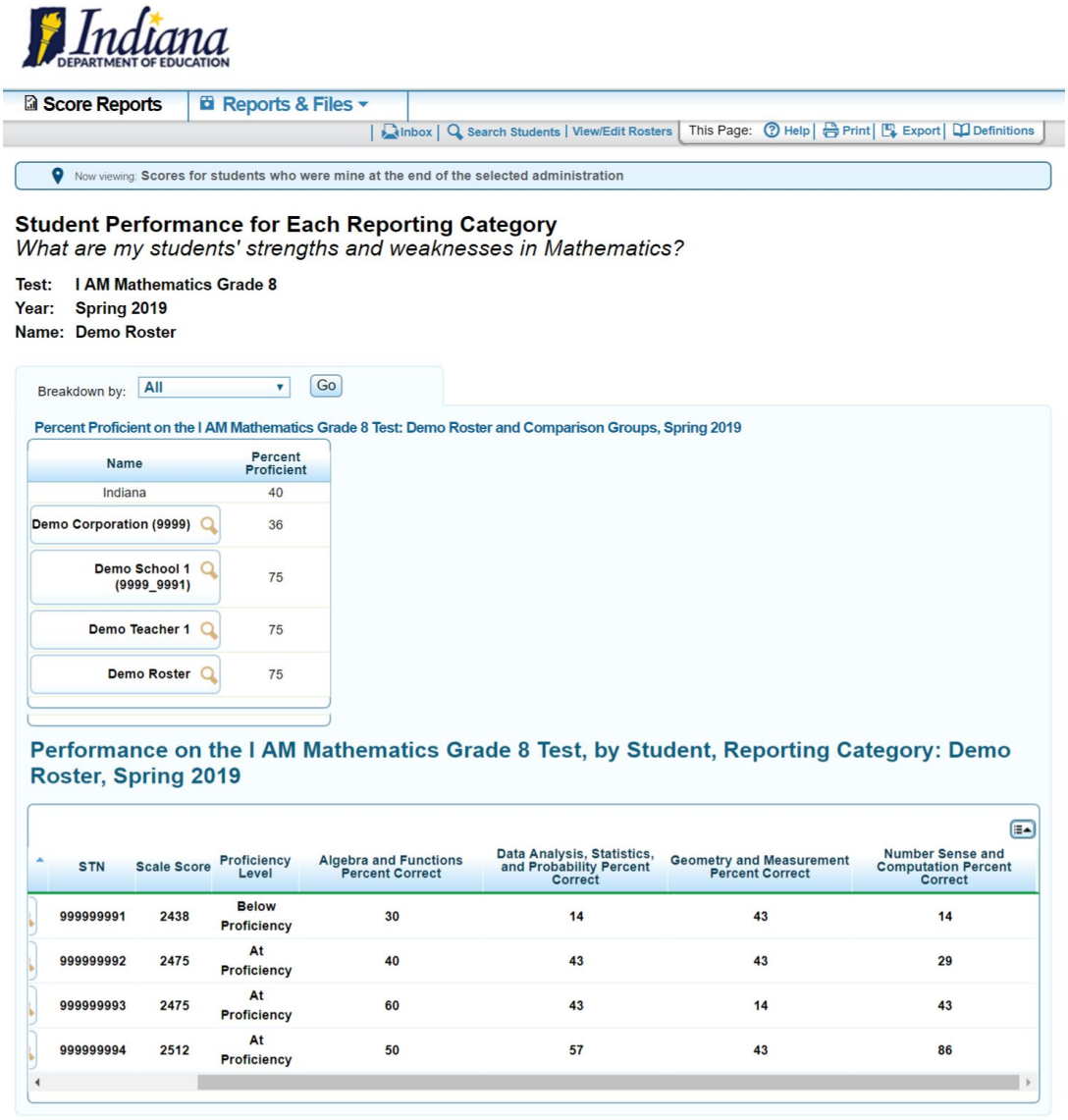


Figure 16: Student-Level Reporting Category Report for Grade 8 Mathematics



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 students' strengths and weaknesses in Mathematics?

Test: I AM Mathematics Grade 8
Year: Spring 2019
Name: Demo Roster

Breakdown by: All Go

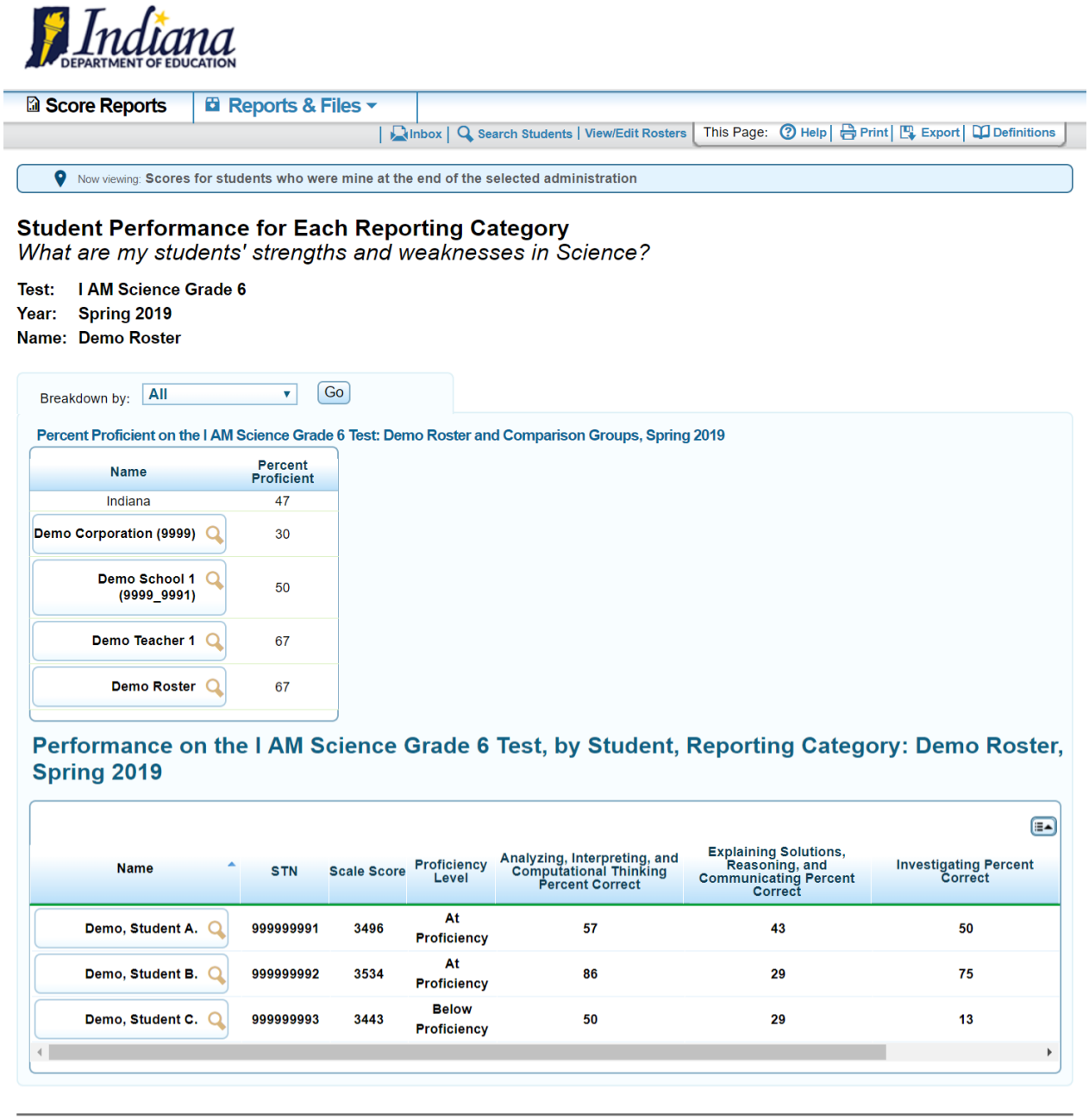
Percent Proficient on the I AM Mathematics Grade 8 Test: Demo Roster and Comparison Groups, Spring 2019

Name	Percent Proficient
Indiana	40
Demo Corporation (9999)	36
Demo School 1 (9999_9991)	75
Demo Teacher 1	75
Demo Roster	75

Performance on the I AM Mathematics Grade 8 Test, by Student, Reporting Category: Demo Roster, Spring 2019

STN	Scale Score	Proficiency Level	Algebra and Functions Percent Correct	Data Analysis, Statistics, and Probability Percent Correct	Geometry and Measurement Percent Correct	Number Sense and Computation Percent Correct
999999991	2438	Below Proficiency	30	14	43	14
999999992	2475	At Proficiency	40	43	43	29
999999993	2475	At Proficiency	60	43	14	43
999999994	2512	At Proficiency	50	57	43	86

Figure 17: Student-Level Reporting Category Report for 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 students' strengths and weaknesses in Science?

Test: I AM Science Grade 6
Year: Spring 2019
Name: Demo Roster

Breakdown by: All Go

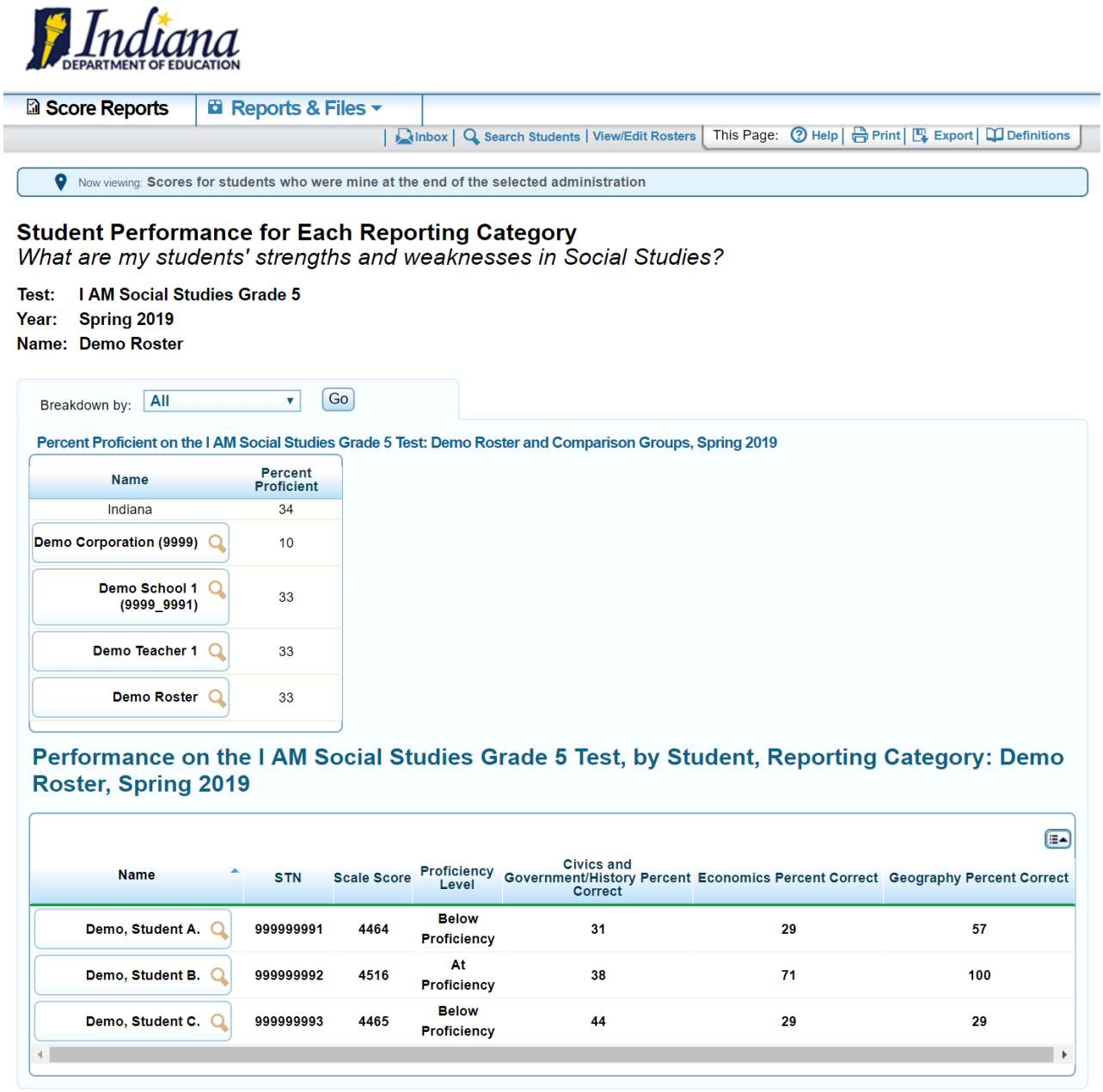
Percent Proficient on the I AM Science Grade 6 Test: Demo Roster and Comparison Groups, Spring 2019

Name	Percent Proficient
Indiana	47
Demo Corporation (9999)	30
Demo School 1 (9999_9991)	50
Demo Teacher 1	67
Demo Roster	67

Performance on the I AM Science Grade 6 Test, by Student, Reporting Category: Demo Roster, Spring 2019

Name	STN	Scale Score	Proficiency Level	Analyzing, Interpreting, and Computational Thinking Percent Correct	Explaining Solutions, Reasoning, and Communicating Percent Correct	Investigating Percent Correct
Demo, Student A.	999999991	3496	At Proficiency	57	43	50
Demo, Student B.	999999992	3534	At Proficiency	86	29	75
Demo, Student C.	999999993	3443	Below Proficiency	50	29	13

Figure 18: Student-Level Reporting Category Report for Grade 5 Social Studies



1.6.6 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 for that student:

- Overall subject scale score
- Overall subject performance level
- Percent proficiency for a student's state, corporation, and school
- Percent correct in each reporting category

The top section of the report includes key student information:

- Name
- Scale score
- Performance level

The middle section includes the following data:

- Bar chart display of the student's scale score
- Performance-level descriptors with cut scores at each performance level
- Average aggregated scale scores at the state, corporation, and school levels

The bottom section of the report contains detailed information on student performance in each reporting category.

Figure 19 through Figure 22 present examples of ISRs for an I AM assessment.

Figure 19: Individual Student Report for Grade 8 ELA

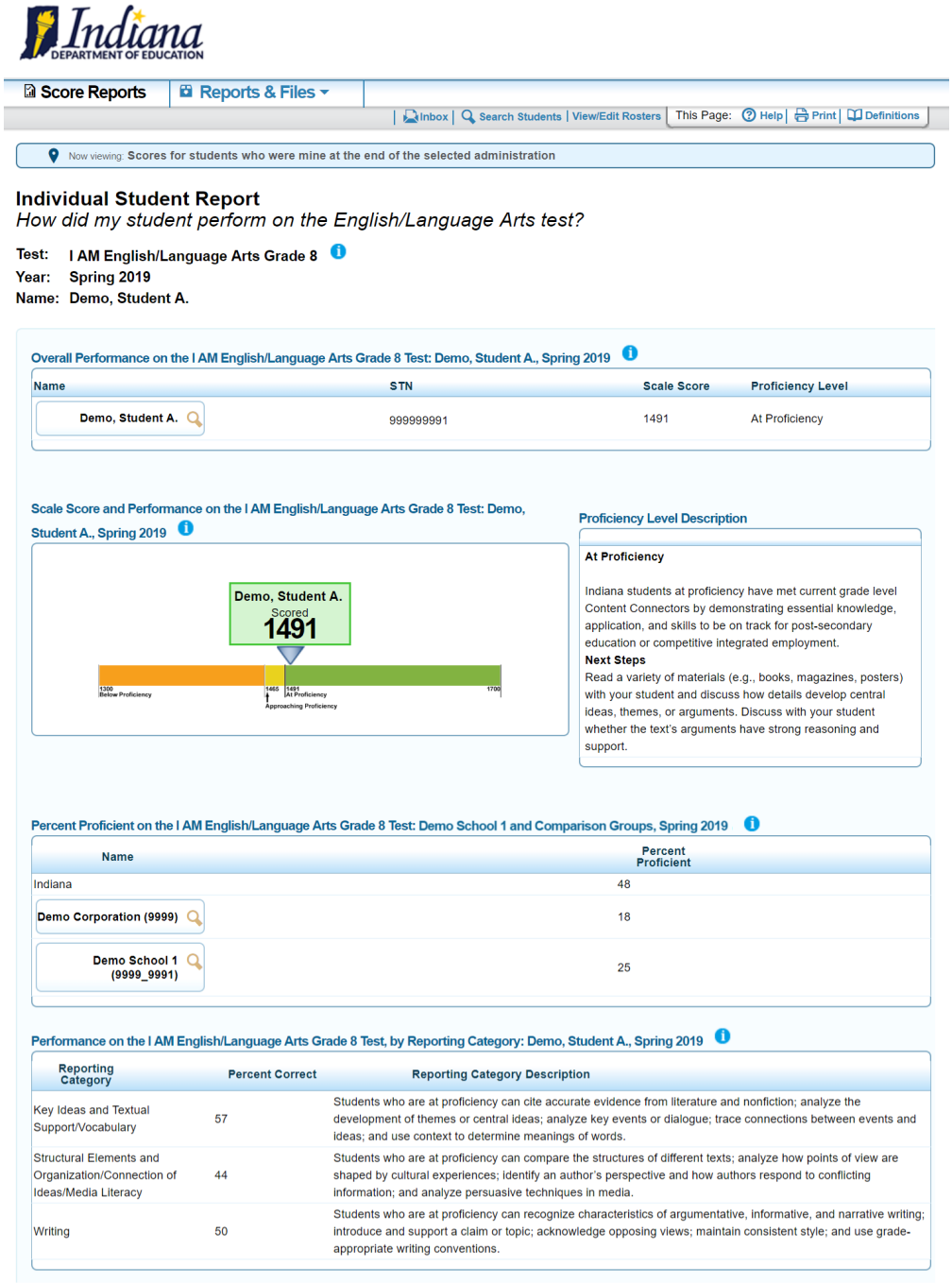


Figure 20: Individual Student Report for Grade 8 Mathematics

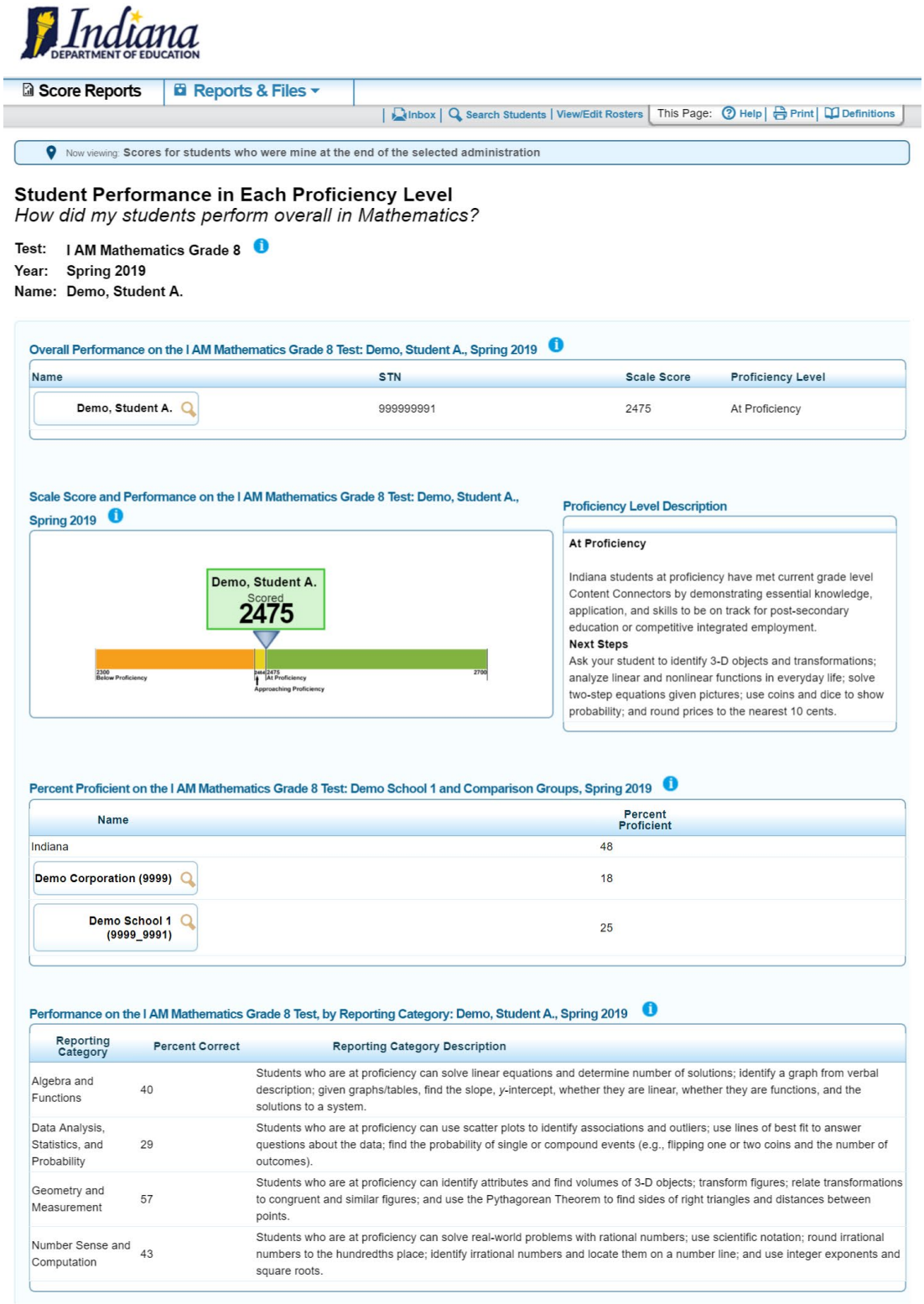
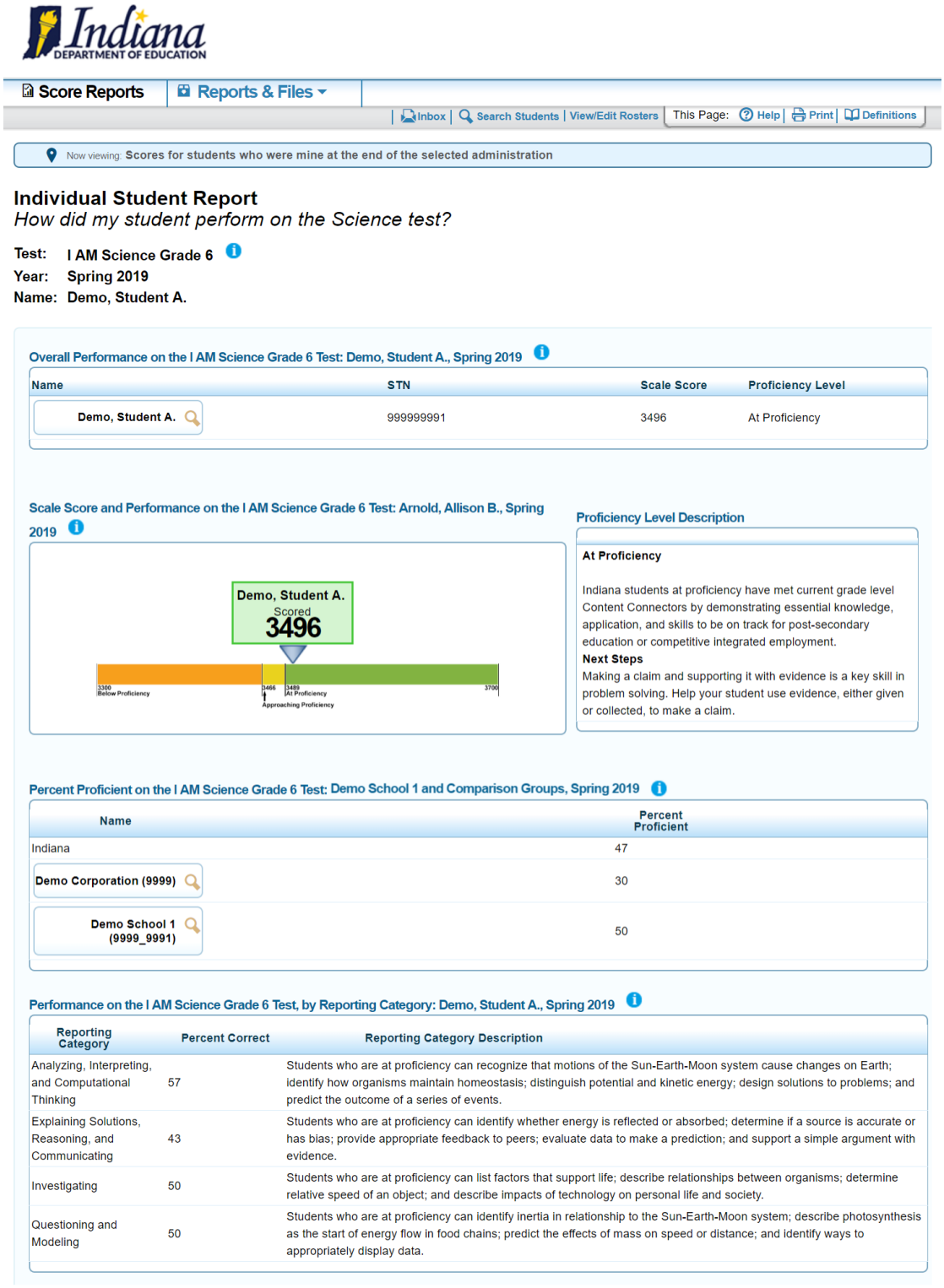


Figure 21: Individual Student Report for Grade 6 Science



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 Science test?

Test: I AM Science Grade 6
Year: Spring 2019
Name: Demo, Student A.

Overall Performance on the I AM Science Grade 6 Test: Demo, Student A., Spring 2019

Name	STN	Scale Score	Proficiency Level
Demo, Student A.	999999991	3496	At Proficiency

Scale Score and Performance on the I AM Science Grade 6 Test: Arnold, Allison B., Spring 2019

Proficiency Level Description

At Proficiency

Indiana students at proficiency have met current grade level Content Connectors by demonstrating essential knowledge, application, and skills to be on track for post-secondary education or competitive integrated employment.

Next Steps

Making a claim and supporting it with evidence is a key skill in problem solving. Help your student use evidence, either given or collected, to make a claim.

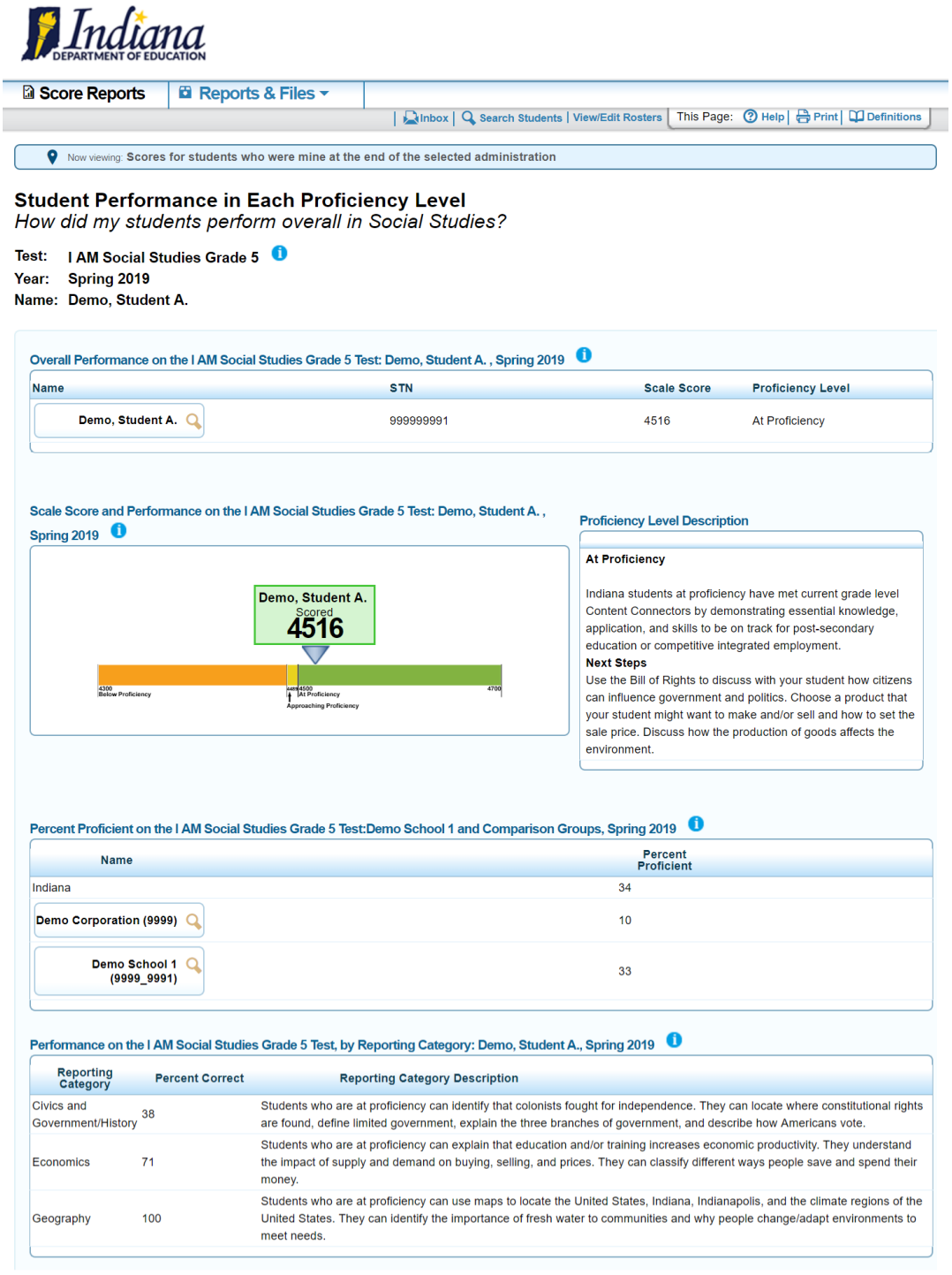
Percent Proficient on the I AM Science Grade 6 Test: Demo School 1 and Comparison Groups, Spring 2019

Name	Percent Proficient
Indiana	47
Demo Corporation (9999)	30
Demo School 1 (9999_9991)	50

Performance on the I AM Science Grade 6 Test, by Reporting Category: Demo, Student A., Spring 2019

Reporting Category	Percent Correct	Reporting Category Description
Analyzing, Interpreting, and Computational Thinking	57	Students who are at proficiency can recognize that motions of the Sun-Earth-Moon system cause changes on Earth; identify how organisms maintain homeostasis; distinguish potential and kinetic energy; design solutions to problems; and predict the outcome of a series of events.
Explaining Solutions, Reasoning, and Communicating	43	Students who are at proficiency can identify whether energy is reflected or absorbed; determine if a source is accurate or has bias; provide appropriate feedback to peers; evaluate data to make a prediction; and support a simple argument with evidence.
Investigating	50	Students who are at proficiency can list factors that support life; describe relationships between organisms; determine relative speed of an object; and describe impacts of technology on personal life and society.
Questioning and Modeling	50	Students who are at proficiency can identify inertia in relationship to the Sun-Earth-Moon system; describe photosynthesis as the start of energy flow in food chains; predict the effects of mass on speed or distance; and identify ways to appropriately display data.

Figure 22: Individual Student Report for 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

Student Performance in Each Proficiency Level

How did my students perform overall in Social Studies?

Test: I AM Social Studies Grade 5
Year: Spring 2019
Name: Demo, Student A.

Overall Performance on the I AM Social Studies Grade 5 Test: Demo, Student A. , Spring 2019

Name	STN	Scale Score	Proficiency Level
Demo, Student A.	999999991	4516	At Proficiency

Scale Score and Performance on the I AM Social Studies Grade 5 Test: Demo, Student A. , Spring 2019

Proficiency Level Description

At Proficiency

Indiana students at proficiency have met current grade level Content Connectors by demonstrating essential knowledge, application, and skills to be on track for post-secondary education or competitive integrated employment.

Next Steps

Use the Bill of Rights to discuss with your student how citizens can influence government and politics. Choose a product that your student might want to make and/or sell and how to set the sale price. Discuss how the production of goods affects the environment.

Percent Proficient on the I AM Social Studies Grade 5 Test:Demo School 1 and Comparison Groups, Spring 2019

Name	Percent Proficient
Indiana	34
Demo Corporation (9999)	10
Demo School 1 (9999_9991)	33

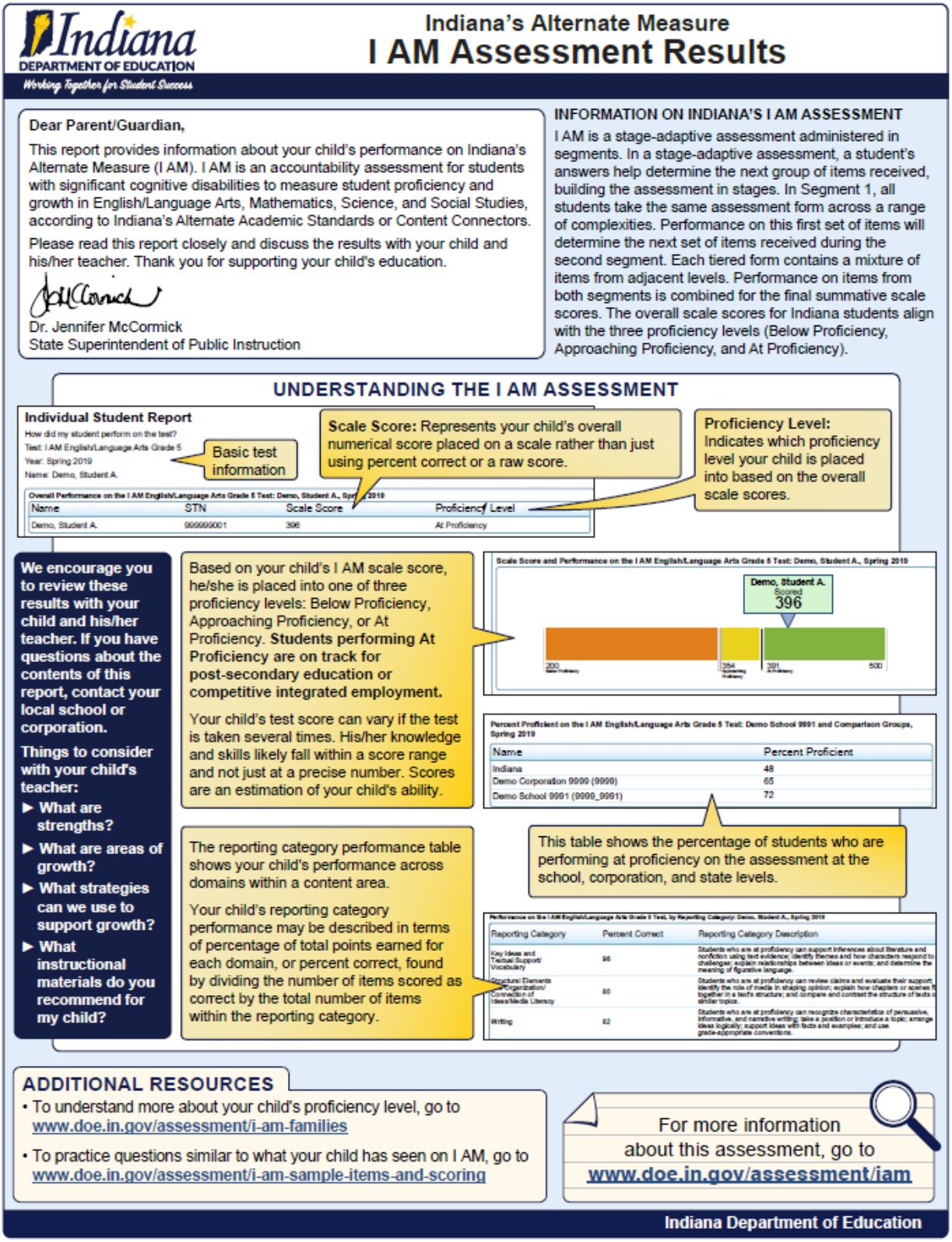
Performance on the I AM Social Studies Grade 5 Test, by Reporting Category: Demo, Student A., Spring 2019

Reporting Category	Percent Correct	Reporting Category Description
Civics and Government/History	38	Students who are at proficiency can identify that colonists fought for independence. They can locate where constitutional rights are found, define limited government, explain the three branches of government, and describe how Americans vote.
Economics	71	Students who are at proficiency can explain that education and/or training increases economic productivity. They understand the impact of supply and demand on buying, selling, and prices. They can classify different ways people save and spend their money.
Geography	100	Students who are at proficiency can use maps to locate the United States, Indiana, Indianapolis, and the climate regions of the United States. They can identify the importance of fresh water to communities and why people change/adapt environments to meet needs.

1.6.7 Interpretive Guide

When printing ISRs, users have the option to print a supplemental “interpretive guide” (called an “Addendum” when printing a Simple ISR) intended as a stand-alone document (see Figure 23) to help teachers, administrators, parents, and students better understand the data presented in the ISR.

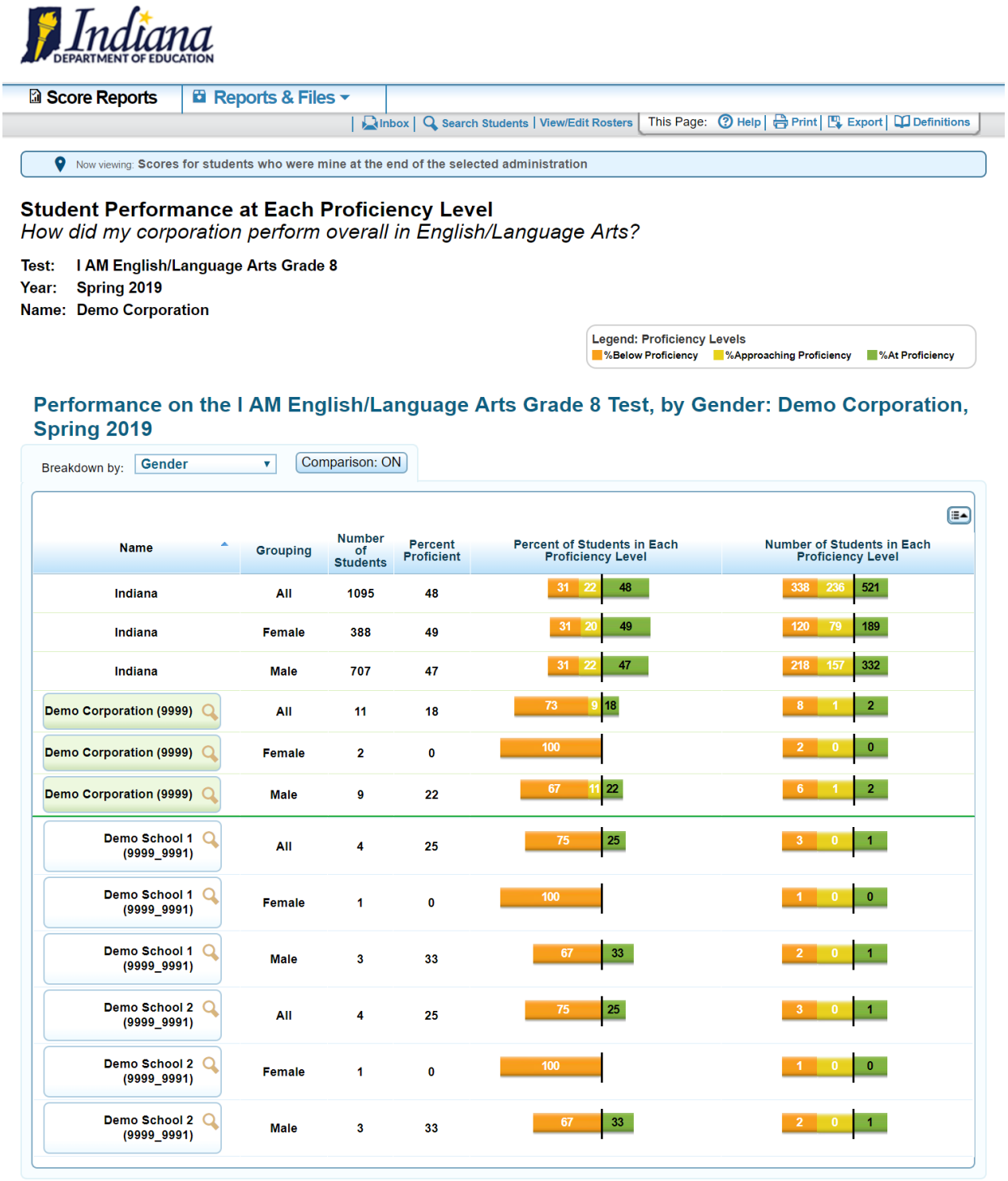
Figure 23: Supplemental Interpretive Guide



1.6.8 Reports by Sub-Group

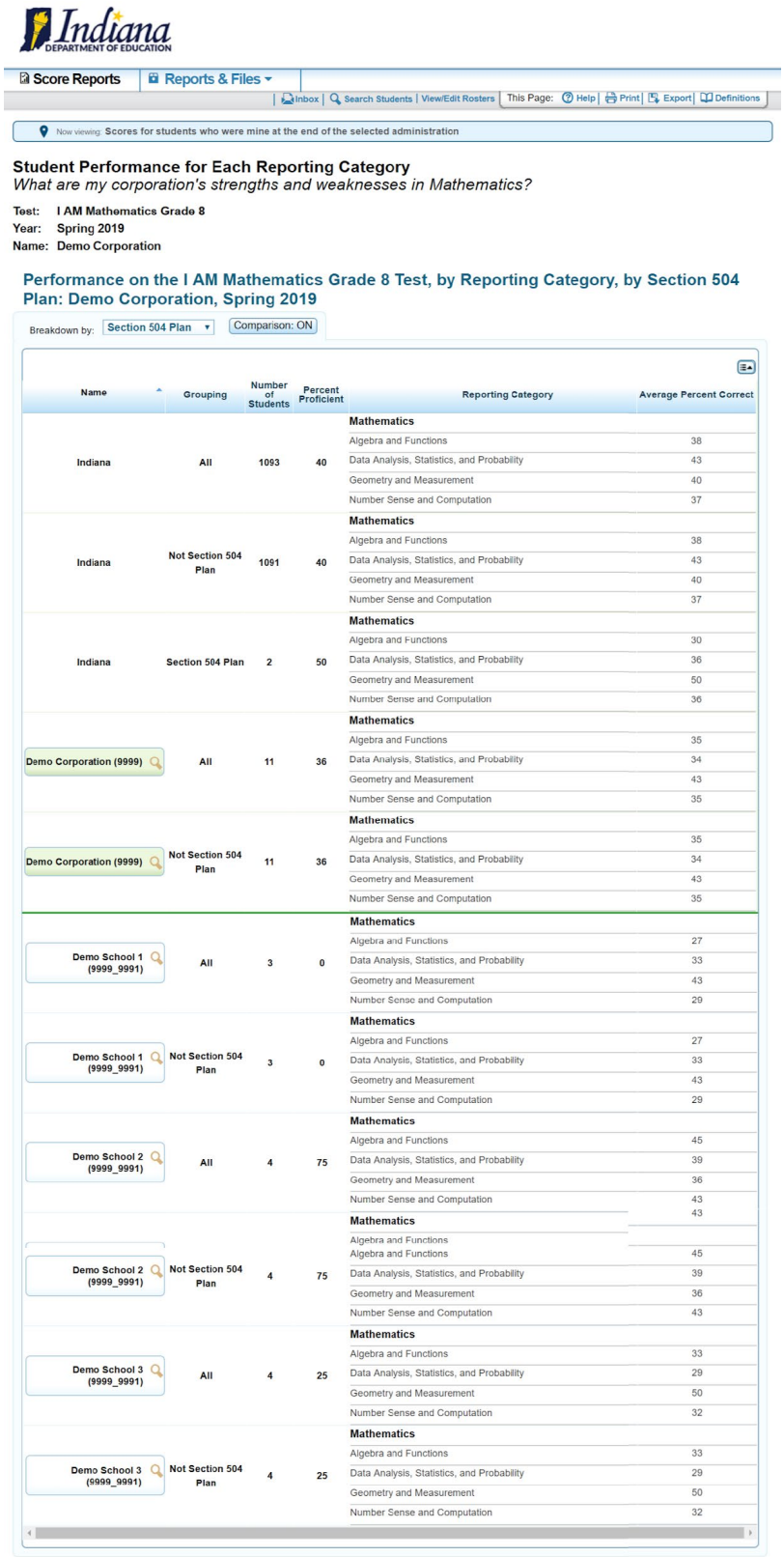
At the aggregate level, student performance can be broken down by demographic sub-group, such as gender (Figure 24) or English learner status (Figure 25).

Figure 24: Corporation Aggregate-Level Subject Report by Gender for Grade 8 ELA



Name	Grouping	Number of Students	Percent Proficient	Percent of Students in Each Proficiency Level	Number of Students in Each Proficiency Level
Indiana	All	1095	48	31 22 48	338 236 521
Indiana	Female	388	49	31 20 49	120 79 189
Indiana	Male	707	47	31 22 47	218 157 332
Demo Corporation (9999)	All	11	18	73 9 18	8 1 2
Demo Corporation (9999)	Female	2	0	100	2 0 0
Demo Corporation (9999)	Male	9	22	67 11 22	6 1 2
Demo School 1 (9999_9991)	All	4	25	75 25	3 0 1
Demo School 1 (9999_9991)	Female	1	0	100	1 0 0
Demo School 1 (9999_9991)	Male	3	33	67 33	2 0 1
Demo School 2 (9999_9991)	All	4	25	75 25	3 0 1
Demo School 2 (9999_9991)	Female	1	0	100	1 0 0
Demo School 2 (9999_9991)	Male	3	33	67 33	2 0 1

Figure 25: Corporation Aggregate-Level Reporting Category Report by Section 504 Plan Status for 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: I AM Mathematics Grade 8
Year: Spring 2019
Name: Demo Corporation

Performance on the I AM 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	Percent Proficient	Reporting Category	Average Percent Correct
Indiana	All	1093	40	**Mathematics**	
				Algebra and Functions	38
				Data Analysis, Statistics, and Probability	43
				Geometry and Measurement	40
				Number Sense and Computation	37
Indiana	Not Section 504 Plan	1091	40	**Mathematics**	
				Algebra and Functions	38
				Data Analysis, Statistics, and Probability	43
				Geometry and Measurement	40
				Number Sense and Computation	37
Indiana	Section 504 Plan	2	50	**Mathematics**	
				Algebra and Functions	30
				Data Analysis, Statistics, and Probability	36
				Geometry and Measurement	50
				Number Sense and Computation	36
Demo Corporation (9999)	All	11	36	**Mathematics**	
				Algebra and Functions	35
				Data Analysis, Statistics, and Probability	34
				Geometry and Measurement	43
				Number Sense and Computation	35
Demo Corporation (9999)	Not Section 504 Plan	11	36	**Mathematics**	
				Algebra and Functions	35
				Data Analysis, Statistics, and Probability	34
				Geometry and Measurement	43
				Number Sense and Computation	35
Demo School 1 (9999_9991)	All	3	0	**Mathematics**	
				Algebra and Functions	27
				Data Analysis, Statistics, and Probability	33
				Geometry and Measurement	43
				Number Sense and Computation	29
Demo School 1 (9999_9991)	Not Section 504 Plan	3	0	**Mathematics**	
				Algebra and Functions	27
				Data Analysis, Statistics, and Probability	33
				Geometry and Measurement	43
				Number Sense and Computation	29
Demo School 2 (9999_9991)	All	4	75	**Mathematics**	
				Algebra and Functions	45
				Data Analysis, Statistics, and Probability	39
				Geometry and Measurement	36
				Number Sense and Computation	43
Demo School 2 (9999_9991)	Not Section 504 Plan	4	75	**Mathematics**	
				Algebra and Functions	45
				Data Analysis, Statistics, and Probability	39
				Geometry and Measurement	36
				Number Sense and Computation	43
Demo School 3 (9999_9991)	All	4	25	**Mathematics**	
				Algebra and Functions	33
				Data Analysis, Statistics, and Probability	29
				Geometry and Measurement	50
				Number Sense and Computation	32
Demo School 3 (9999_9991)	Not Section 504 Plan	4	25	**Mathematics**	
				Algebra and Functions	33
				Data Analysis, Statistics, and Probability	29
				Geometry and Measurement	50
				Number Sense and Computation	32

1.6.9 Data File

ORS users have the option to quickly generate a comprehensive data file of their students' scores. Data files (an example of which is shown in Figure 26) 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.

Figure 26: Sample Data File

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
Student D	Gender	Ethnicity	Special Ed	Identified	Section 50	Enrolled G	Enrolled S	Enrolled S	Enrolled C	Enrolled C	English/La	English/La	Key Ideas	Structural	Writing Pr	Mathema	Mathema	Algebra ai	Computat	Geometry	Number S	Science Sc	Science Pr	Analyzing	Explaining	Investigat	Questioning
16/10/200	F	American	Y	Y	Y	6	Demo inst	9999_9991	Demo Cor	9999	1445	1	17	55	25	2483	3	63	57	43	22	3443	1	38	14	38	25
16/10/200	F	American	Y	Y	Y	6	Demo inst	9999_9991	Demo Cor	9999												N/A	N/A	N/A	N/A	N/A	N/A
16/10/200	F	American	Y	Y	Y	6	Demo inst	9999_9991	Demo Cor	9999												3443	1	25	29	13	50
16/10/200	F	American	Y	Y	Y	6	Demo inst	9999_9991	Demo Cor	9999	1460	1	17	45	50												
16/10/200	F	American	Y	Y	Y	6	Demo inst	9999_9991	Demo Cor	9999	1481	2	25	64	50	2471	2	25	86	29	11	N/A	N/A	N/A	N/A	N/A	N/A
26/10/200	F	White	N	N	N	6	Demo inst	9999_999(	Demo Cor	9999						2476	2	38	29	57	33						
26/10/200	F	Black/Afri	Y	Y	N	6	Demo inst	9999_999(	Demo Cor	9999	1486	2	50	55	38	2462	2	50	43	43	11	3450	1	29	43	13	38
26/10/200	F	White	N	N	N	6	Demo inst	9999_999(	Demo Cor	9999						2462	2	38	29	29	44						
26/10/200	F	Black/Afri	N	Y	N	6	Demo inst	9999_999(	Demo Cor	9999						2421	1	25	14	29	11	3489	3	43	43	38	63
01/01/201	M	Native Ha	Y	N	N	6	Demo inst	9999_9991	Demo Cor	9999						2490	3	25	57	71	33						
01/01/201	F	Black/Afri	Y	Y	Y	6	Demo inst	9999_9991	Demo Cor	9999	1473	2	42	36	50	Invalidate	Invalidate	Invalidate	Invalidate	Invalidate	Invalidated						
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1700	3	100	100	100	2700	3	100	100	100	100	3700	3	100	100	100	100
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2300	1	0	0	0	0	3300	1	0	0	0	0
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1700	3	100	100	100	2300	1	0	0	0	0	3300	1	0	0	0	0
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2683	3	88	100	100	100	3700	3	100	100	100	100
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1700	3	100	100	100	2700	3	100	100	100	100	3700	3	100	100	100	100
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2300	1	0	0	0	0	3300	1	0	0	0	0
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1700	3	100	100	100	2300	1	0	0	0	0	3300	1	0	0	0	0
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2700	3	100	100	100	100	3700	3	100	100	100	100
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1700	3	100	100	100	2700	3	100	100	100	100	3700	3	100	100	100	100
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2300	1	0	0	0	0	3300	1	0	0	0	0
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1678	3	92	100	100	2527	3	63	86	71	33	3700	3	N/A	100	100	100
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999						2700	3	100	100	100	100						
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2300	1	0	0	0	0	3300	1	0	0	0	0
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999						2300	1	0	0	0	0						
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1678	3	92	100	100	2357	1	0	0	14	11	3300	1	0	0	0	0
13/04/200	M	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1678	3	92	100	100	2683	3	100	86	100	100	3700	3	N/A	100	100	100
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999	1300	1	0	0	0	2700	3	100	100	100	100	3700	3	N/A	100	100	100
13/04/200	F	White	Y	Y	Y	6	Demo inst	9999_999(	Demo Cor	9999						2605	3	88	86	86	89						

2. Interpretation of Reported Scores

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

2.1 Scale Score

A scale score is used to describe how well a student performed on an assessment and can be interpreted as an estimate of a students' knowledge and skills as measured by their performance on the assessment. A scale score is the student's overall numeric score. Scale scores can be used to illustrate students' current level of performance and are most 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 assessment. Conversely, higher scale scores can indicate that the student has proficient knowledge and skills measured by the assessment. 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 Standard Error of Measurement (SEM). The details of SEM and the graphs of the conditional SEM of each test are provided in Volume 4 Evidence of Reliability and Validity. 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.

2.2 Performance Level

For I AM, scale scores are mapped onto three performance levels (Level 1–Below Proficiency, Level 2–Approaching Proficiency, and Level 3–At Proficiency) using performance standards (or cut scores—see Section 2.4, Cut Scores). 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 I AM at Level 3 are considered on track to demonstrate progress toward mastery of the knowledge and skills necessary for competitive employment and post-secondary education. Because performance levels are for the classification of the 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 skills for a given level.

Performance-level descriptors are available on the Indiana Department of Education web page.

2.3 PERCENT CORRECT FOR REPORTING CATEGORIES

Students' performance on each reporting category was reported using percent correct and was computed using all items for scoring in categories that have a minimum of seven items in the blueprint. Due to the loss of items after Item Data Review, leaving only six items available in the reporting categories, Mathematics grade 10 Tier 3 students did not receive the subscore for Equations and Inequalities (EI); Science grade 4 Tier 2 students did not receive the subscore for Investigating (I); and braille Science grade 6 Tier 3 students did not receive the subscore for Analyzing, Interpreting, and Computational Thinking (AICT). The reporting category scores were computed as a percent-correct score for each student and an average percent-correct score for aggregate reporting.

2.4 CUT SCORES

For all grades and subjects within I AM, scale scores are mapped onto three performance levels: Level 1–Below Proficiency, Level 2–Approaching Proficiency, and Level 3–At 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 they constitute the cutoff points for each performance level. Tables 7 through 10 show the cut scores for I AM.

Table 7: I AM ELA Assessment Proficiency Cut Scores

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
3	1300–1463	1464–1481	1482–1700
4	1300–1478	1479–1497	1498–1700
5	1300–1474	1475–1488	1489–1700
6	1300–1466	1467–1486	1487–1700
7	1300–1485	1486–1497	1498–1700
8	1300–1464	1465–1490	1491–1700
10	1300–1467	1468–1505	1506–1700

Table 8: I AM Mathematics Assessment Proficiency Cut Scores

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
3	2300–2462	2463–2473	2474–2700
4	2300–2461	2462–2478	2479–2700
5	2300–2459	2460–2470	2471–2700
6	2300–2461	2462–2477	2478–2700
7	2300–2466	2467–2477	2478–2700
8	2300–2463	2464–2474	2475–2700
10	2300–2470	2471–2484	2485–2700

Table 9: I AM Science Assessment Proficiency Cut Scores

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
4	3300-3475	3476-3496	3497-3700
6	3300-3465	3466-3488	3489-3700
Biology	3300-3471	3472-3502	3503-3700

Table 10: I AM Social Studies Grade 5 Assessment Proficiency Cut Scores

Grade	Level 1 Below Proficiency	Level 2 Approaching Proficiency	Level 3 At Proficiency
5	4300–4488	4489–4499	4500–4700

2.5 AGGREGATED SCORES

Percentage of students proficient, percentage of students in each proficiency level, and students' percent correct scores are aggregated at roster, teacher, school, corporation, and state levels to represent how well a group of students performs overall and by reporting category on an assessment. When students' scores are aggregated, these 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).

2.6 APPROPRIATE USES FOR SCORES AND REPORTS

Assessment results can be used to provide information on individual students' performance on the assessment. 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 competitive employment and post-secondary education. Additionally, assessment results can be used to identify students' 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.

Results on students' performance on the assessment 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 students' strengths and weaknesses and can be used to improve teaching and students' 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 students' learning, particularly for students from disadvantaged sub-groups. For example, teachers might see students' assessment results by gender and observe that a particular group of students is struggling with literary response and analysis in reading. 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.

Although assessment results provide valuable information to understand students' performance, these scores and reports should be used with caution. It is important to note that scale scores are estimates of true scores and hence do not represent a precise measure of student performance. A student's scale score is associated with measurement error; users must therefore 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 relied on as the only source of information. Given that assessment results provide limited information, other sources of data 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 must take into account the group size. The smaller the group, the larger the measurement error related to these aggregated data, thus requiring more cautious interpretation.

3. SUMMARY

I AM results are reported online via the ORS. The results were released after the testing window closed and standard setting had been completed. In the 2019–2020 school year, the ORS will report results on assessments as they are completed, beginning 12 business days after the first test is completed.

The reporting system 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 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 student 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 or download the reports as well as the data on which the reports are based. ISRs can be produced individually or batched as PDF files.

All authorized users can download reports, including data about students for whom they are responsible, at any time. The various reports may be used to inform stakeholders regarding student performance and instructional strategies.

Indiana's Alternate Measure

2018–2019

Volume 6
Standard Setting Report

ACKNOWLEDGMENTS

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

Major contributors to this technical report include the following staff from American Institutes for Research: Stephan Ahadi, Hyesuk Jang, Yuan Hong, Katherine Krehbiel, Hashim Evans, and Celine Bryan. 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. EXECUTIVE SUMMARY

1.1 BACKGROUND

This document provides an executive summary of the procedures used to recommend the Approaching Proficiency and At Proficiency performance standards for the Indiana's Alternate Measure (I AM) assessments. The I AM assessments are designed for students with significant cognitive disabilities who participate in a school curriculum that is consistent with the grade-level Indiana Alternate Academic Standards (IAAS) or Content Connectors. The assessments consist of English/Language Arts (ELA) and Mathematics assessments in grades 3–8 and 10; Science assessments in grade 4, grade 6, and Biology; and a Social Studies assessment in grade 5. The standard-setting workshop was conducted July 22 through 24 of this year. The workshop employed the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001), a widely used method in which standard-setting panelists use their expert knowledge of the academic content standards and student achievement to map the Performance Level Descriptors (PLDs) adopted by the Indiana Department of Education (IDOE) to an ordered-item booklet (OIB) based on the assessments administered to students in Spring 2019.

1.2 OVERVIEW OF STANDARD-SETTING ACTIVITIES

The IDOE implemented a standard-setting workshop to recommend two performance standards (Approaching Proficiency and At Proficiency) to distinguish students into three performance levels (Below Proficiency, Approaching Proficiency, and At Proficiency) with respect to the Indiana Content Connectors. The standard-setting workshop was conducted with educators from around Indiana to identify and recommend to the IDOE the performance standards on the Spring 2019 assessments.

The workshop began by orienting panelists to the workshop activities. Major workshop activities included the following: development of "Just Barely" PLDs; review of test items and the OIB; standard setting in three rounds; and anchor grade and final moderation for Mathematics and ELA. Panelists received training prior to beginning each workshop activity.

After the large group orientation, panelists broke into their grade- and subject-area panels. After receiving initial training on the use of the online standard-setting tool, panelists were administered the grade- and subject-area assessment for which they would be recommending performance standards. The assessment was administered to panelists electronically. Panelists interacted with items with the same functionality as was provided to students during the Spring 2019 test administration.

After completing the test administration, panelists reviewed the Range PLDs developed for each of the standards assessed in their grade and subject area assessment. Panelists then worked to develop threshold PLDs (i.e., "Just Barely" PLDs), identifying the characteristics of achievement differentiating those students who just barely qualify for entry into each of the performance levels from those students not yet qualified for entry into the performance level.

Following the development of threshold PLDs, panelists began reviewing their OIBs. The OIBs comprised a Spring 2019 test administration, minus some items with identical impact data, and augmented with additional items to minimize any gaps in assessment information between pages. In their review of their OIBs, panelists were instructed to answer two questions about each item: "What does a student need to know and be able to do to achieve the score point?" and "What makes this item more difficult than the preceding items in the OIB?"

Sixty-eight ELA and Mathematics panelists worked in grade-band panels (grades 3–4, 5–6, 7–8, and 10). Panels began by recommending performance standards in anchor grades (grades 4, 6, 8, and 10) followed by adjacent grades (grades 3, 5, and 7). Educators followed the methods outlined below for these grades. Following recommendation of performance standards for the anchor grades, grade-band panels convened to articulate recommended performance standards across the anchor grade levels. Following recommendation of performance standards for the adjacent grades, grade-band panels reconvened to do a final articulation of recommended performance standards across all grade levels.

Twenty-five panelists broken into three panels of educators who set standards for Science grade 4, grade 6, and Biology. Social Studies grade 5 also worked as a separate group of seven panelists. These recommendations were not articulated based on the content expectations for each grade.

Upon completion of their OIB review, panelists received training on the Bookmark procedure. Panelists then logged into the standard-setting tool developed by the American Institutes for Research (AIR) to practice performing the Bookmark method using a six-item practice OIB. Panelists worked to evaluate the knowledge and skill requirements of the practice items with respect to their "Just Barely" PLDs to practice assigning performance standards in the standard-setting tool.

Upon completion of training, panelists signed the Round 1 Readiness Form, indicating they understood the task and were ready to make their recommendations. Panelists worked through each page of the OIB and placed their performance standards recommendations on the last page in the OIB, where about 50% of students who just barely qualify for entry into the performance level would respond successfully.

After placing and confirming their bookmarks, panelists received and discussed feedback from their Round 1 ratings by table and by the entire room. The feedback was in the form of statistics that described the central tendency of the panelists' ratings. The facilitator worked with the room as a whole to discuss results and gave panelists the opportunity to discuss with their tables.

Following Round 1 feedback and panel discussions, facilitators introduced the idea of benchmark data to the panelists. In addition to having well-articulated performance standards across grades and subjects, the IDOE's policy committee recommended that the performance standards benchmarked against a multi-state assessment (created by the National Center and State Collaborative [NCSC]) of students with significant intellectual disabilities. The IDOE's policy committee also recommended that the performance standards for the alternate assessment be considered in relationship to the

performance standards for the general education assessment for the general population (the Indiana Learning Evaluation Assessment Readiness Network [ILEARN]). Benchmark data sources are given in Table 1.

Panelists signed the Round 2 Readiness Form and completed their recommendations. Panelists were instructed to consider feedback from Round 1, the benchmark data, and the characteristics of a student who Just Barely qualifies for entry into the performance levels.

Table 1: Benchmark Data Sources

Subject	Benchmark Data
ELA	ILEARN and NCSC
Mathematics	ILEARN and NCSC
Science	ILEARN
Social Studies	ILEARN

Panelists received and discussed feedback from their Round 2 ratings for the table and entire room. Recommended performance cuts were compared to the benchmark data given, and rooms discussed whether results were in line with the benchmarks. Following Round 2 feedback and discussions, facilitators introduced impact data to the panelists. Impact data shows the percentage of Indiana students who would reach the standard for each page in the OIB.

Panelists signed the Round 3 Readiness Form and completed their recommendations. Panelists were instructed to consider feedback from Rounds 1 and 2, the benchmark data, the impact data, and the characteristics of a student who just barely qualifies for entry into the performance levels.

After Round 3, panelists filled out their workshop evaluation forms. For the Science and Social Studies panelists, their workshops were completed.

Mathematics and ELA table leaders participated in articulation and moderation between anchor grades. Impact data across anchor grades was displayed, and AIR psychometric staff explained the results and led discussions. After the anchor grade moderation session, ELA grade 3 table leaders decided to adjust the OIB page of Approaching Proficiency from 9 to 7. ELA grade 7 table leaders decided to adjust the OIB page of Approaching Proficiency from 8 to 10. Mathematics grade 6 table leaders decided to adjust the OIB page of At Proficiency from 12 to 14. Mathematics grade 8 table leaders decided to adjust the OIB page of At Proficiency from 14 to 10. Mathematics grade 10 table leaders decided to adjust the OIB page for Approaching Proficiency from 10 to 8, and the At Proficiency OIB page from 14 to 16.

Following the moderation, the performance standards recommended for the anchor grades were used to interpolate/extrapolate the location of performance standards in the remaining adjacent grades (grades 3, 5, and 7). These pages were not set, but rather given to panelists as a reference point. Mathematics and ELA panelists then completed Rounds 1, 2, and 3 as outlined.

At the completion of Round 3 for the adjacent grades, table leaders participated in a final moderation session. After the final moderation, only two committees changed the recommendation. ELA grade 3 table leaders decided to change the Approaching Proficiency OIB page from 9 to 7, and Grade 7 ELA grade 7 table leaders changed the Approaching Proficiency OIB page from 8 to 10.

1.3 MEETING RESULTS

Table 2 shows the median and range of OIB pages chosen by panelists (from either Round 3 or moderation, where applicable) as well as the impact percentage (the percentage of students scoring at or above the score indicated by the bookmark) for that page.

Figures 1 and 3 show the ELA and Mathematics impact percentages graphically. Figures 2 and 4 show the ELA and Mathematics scale scores. All I AM assessments are within grade scales. Figures 5 and 6 show the Science and Social Studies impact percentages.

Table 2: Final Results

Subject	Grade	I AM OIB Page Approaching Proficiency	Range of Pages Approaching Proficiency	I AM Impact Data for Approaching Proficiency	I AM OIB Page At Proficiency	Range of Pages At Proficiency	I AM Impact Data for At Proficiency
ELA	3	7	3	60%	12	4	45%
	4	13	4	60%	20	2	45%
	5	11	3	65%	17	3	51%
	6	9	5	65%	16	7	50%
	7	10	1	63%	18	2	50%
	8	11	2	71%	19	5	49%
	10	13	0	79%	27	1	49%
Mathematics	3	6	0	71%	10	0	59%
	4	7	3	68%	12	5	48%
	5	6	2	66%	10	2	48%
	6	8	5	66%	14	2	47%
	7	8	4	59%	11	2	47%
	8	6	8	55%	10	13	42%
	10	8	8	55%	16	8	32%
Science	4	12	18	57%	19	19	41%
	6	11	2	71%	19	7	48%
	Biology	15	2	67%	22	5	43%
Social Studies	5	13	12	41%	17	12	35%

Table 3: Benchmark Comparisons

Subject*	Grade	I AM	NCSC	ILEARN
ELA	3	45	51	46
	4	45	56	45
	5	51	58	47
	6	50	63	47
	7	50	56	49
	8	49	64	50
	10	49	70	
Mathematics	3	59	73	58
	4	48	53	53
	5	48	57	47
	6	47	58	46
	7	47	68	41
	8	42	61	37
	10	32	57	

**Note: Science and Social Studies were not included because NCSC did not include those subjects.*

Figure 1: ELA Impact Results

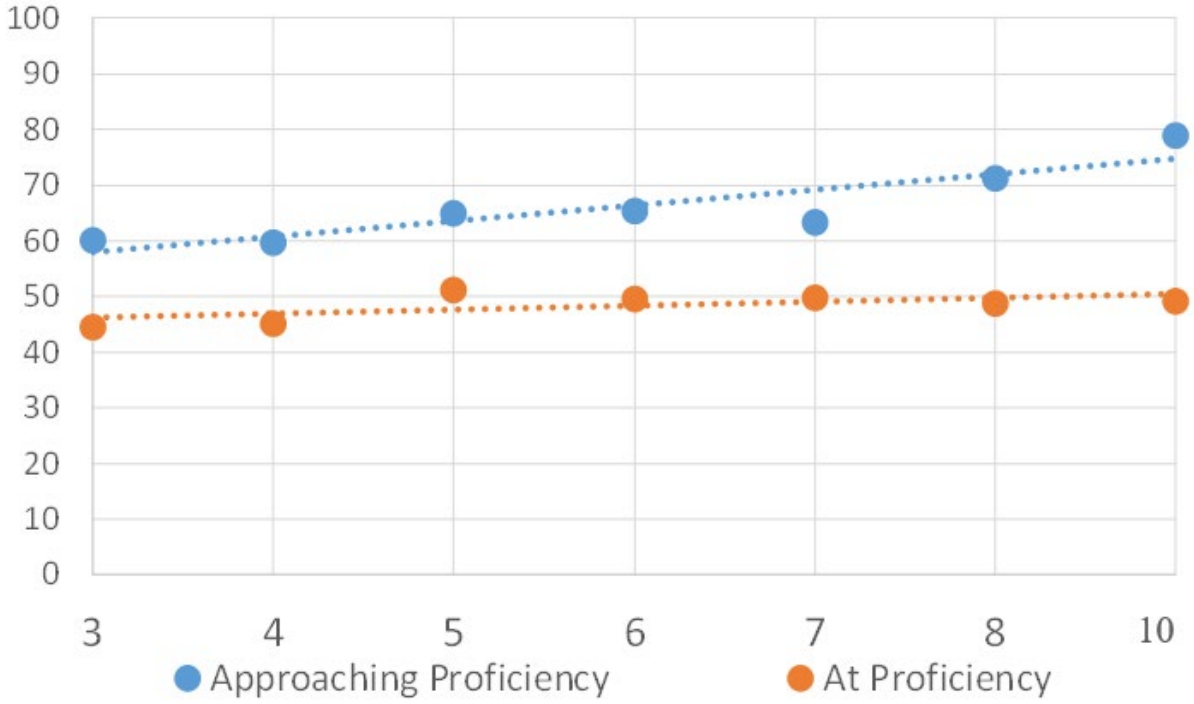


Figure 2: ELA Scale Score Results

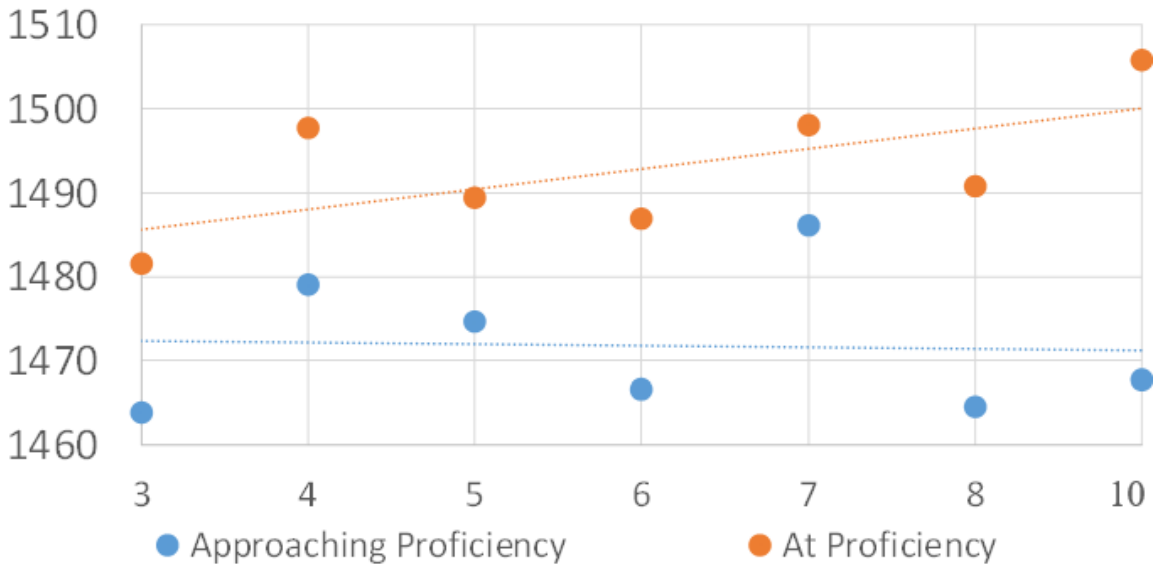


Figure 3: Mathematics Impact Results

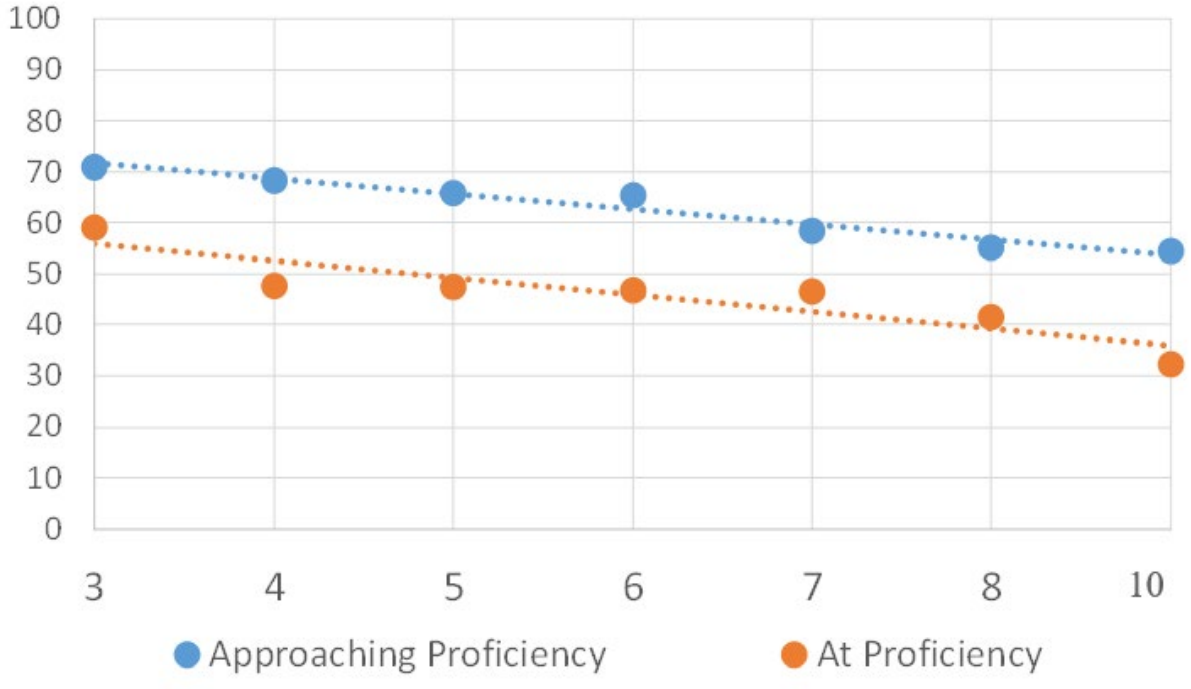


Figure 4: Mathematics Scale Score Results

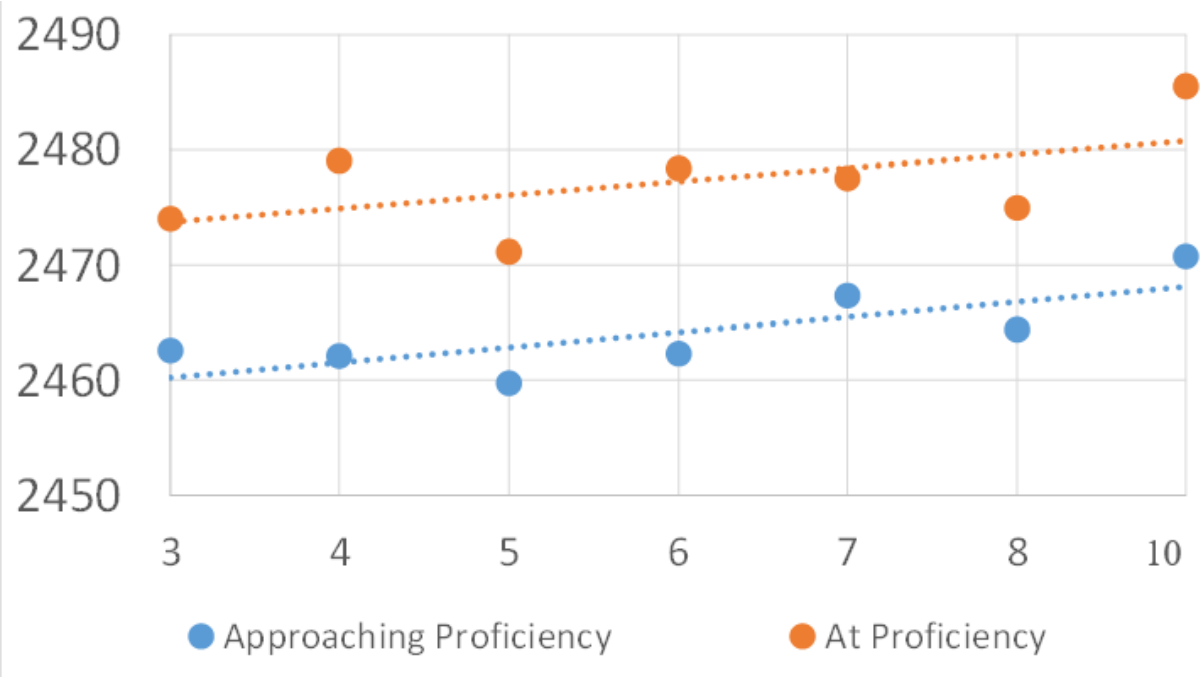
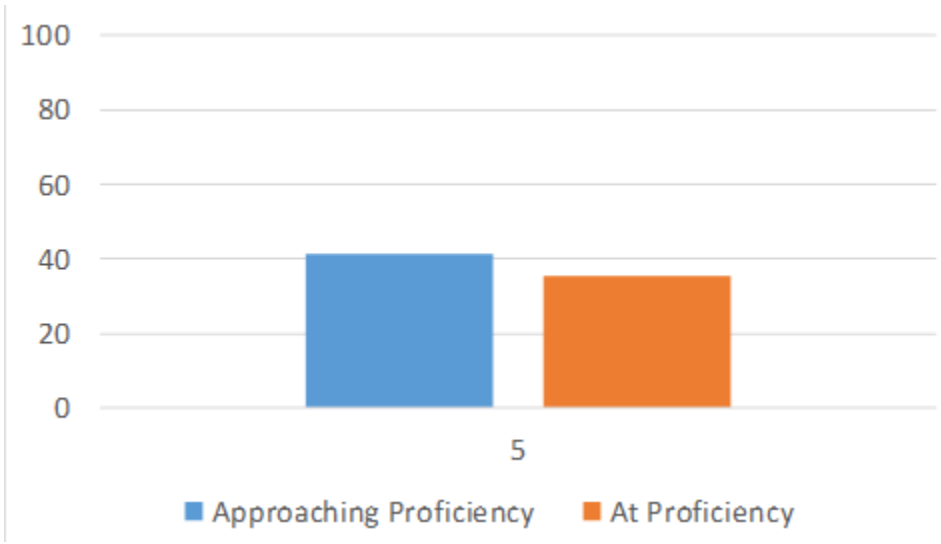


Figure 5: Science Impact Results

Figure 6: Social Studies Impact Results



2. INTRODUCTION

The I AM program for students with significant cognitive disabilities comprises assessments based on the Indiana Academic Standards (IAS) as expressed in the Indiana Content Connectors and the Range PLDs.

The I AM assessments require new performance standards (cut scores) to link performance on the assessments to the content standards. The IDOE contracted with AIR to establish cut scores for ELA and Mathematics in grades 3–8 and 10; Science in grade 4, grade 6, and Biology; and Social Studies in grade 5.

To fulfill this responsibility, AIR

1. implemented a defensible, valid, and technically sound method;
2. provided training on standard setting to all participants;
3. oversaw the process;
4. computed real-time data to inform the process; and
5. produced this technical report documenting the methods, approaches, processes, and outcomes of the standard-setting workshop.

The purpose of this report is to document the standard-setting process and resulting performance standard recommendations for the I AM in ELA, Mathematics, Science, Biology, and Social Studies.

3. STANDARD SETTING

One hundred educators from Indiana convened at the Sheraton Indianapolis Hotel at Keystone Crossing in Indianapolis, Indiana, from July 22 through 24 of this year, with the purpose of completing three rounds of standard setting to recommend two performance standards (cut scores) for the I AM assessments in each content area.

Standard setting is the process used to define achievement on I AM. Performance levels are defined by performance standards, or cut scores, that specify how many of the content standards students must know and be able to do to meet each performance level. Two cut scores (Approaching Proficiency and At Proficiency) are sufficient to define three performance levels.

A prerequisite to standard setting is to determine the nature of the categories, or performance levels, into which students are classified. The three performance level categories for the I AM are “Below Proficiency,” “Approaching Proficiency,” and “At Proficiency.” These categories, or performance levels, are associated with PLDs. PLDs link the assessment content to the IAS. There are multiple types of PLDs (Egan, Schneider, & Ferrara, 2012), including the following:

1. *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. The I AM Policy PLDs 2018–2019 can be found here.

2. *Range PLDs*: A description of what students should know and be able to do throughout the range of each performance level. For example, the Range PLD for Approaching Proficiency describes what students know and can do at that level all the way to just below the At Proficiency cut score. The Range PLDs for the I AM can also be found here.
3. *Target PLDs*: Sometimes called "Threshold" or "Just Barely" PLDs, these are created during the standard-setting workshop and are used only for standard setting. The Target PLDs describe what a student just barely scoring at the entry point of each performance level knows and can do.

On July 25, 2018, the IDOE worked with the seven-person Indiana stakeholder panel to make recommendations for I AM Policy PLDs. The IDOE led the I AM Policy PLD meeting, and AIR staff were present at the meeting in the role of note takers to document the process and the committee wording for the Policy PLDs. Policy PLDs define, at a broad policy level, the goals and rigor of the I AM assessment. The IDOE provided panelists with background on the I AM development process and on the purpose and role of PLDs within the assessment system. The IDOE discussed example PLDs from national and state alternate assessments, including NCSC, Dynamic Learning Maps (DLM), and several states. During the Policy PLD meeting, the panel drafted the following Policy PLDs: Below Proficiency, Approaching Proficiency, and At Proficiency.

On September 11–13, 2018 Indiana educators convened to develop the Range PLDs for each content area and grade level included in the I AM assessments. During the meeting, educators reviewed Policy PLDs and created Range PLDs. With the goal of reinforcing the alignment to ILEARN and ensuring a cohesive system of assessments, the IDOE invited the same policy panel that met on May 15, 2018 to develop ILEARN Policy PLDs to the extent possible. The goal of the I AM PLD meeting was to connect the content of the general assessment to the content of the alternate assessment for students with significant cognitive disabilities. The PLDs describe student performance at the following levels: Below Proficiency, Approaching Proficiency, and At Proficiency.

The participants in the standard-setting workshop primarily worked with the Range PLDs and Target PLDs.

3.1 The Bookmark Method

The Bookmark method of standard setting is well suited to support the establishment of cut scores on high-stakes assessments. It is appropriate for assessments like I AM that are scored using item response theory (IRT) and that use mixed-type items (e.g., multiple-choice with one key, multiple-select with two keys). The Bookmark method is appropriate for these types of assessments and simplifies the decision process for panelists by allowing them to perform the same judgment task for all items, regardless of item type. Because the Bookmark method directly relies on judgments made by experts, panelists and stakeholders report high confidence in the outcomes. It has proven to be technically

sound in litigation, and more than 30 states have selected and implemented this method, making it the most frequently used method of setting achievement standards on high-stakes state accountability assessments (Lewis et al., 2012; Karantonis & Sireci, 2006; Lewis & Lord-Bessen, 2017; Perie, 2005). For these reasons, the IDOE chose to use the Bookmark method to establish new performance standards.

The Bookmark method derives its name from the primary task required of panelists: the placement of a bookmark in an OIB to represent a cut score recommendation. Over the course of multiple rounds of judgments, panelists consider feedback and reference data provided for each round to recommended criterion-referenced cut scores using the Indiana Content Connectors, Range PLDs, Target PLDs, assessment content viewed in the OIBs, and panelist discussions.

3.2 WORKSHOP STRUCTURE

One large meeting room served as the all-participant training room. Twelve breakout rooms served as workspaces for the subject and grade-level panels, each with two tables. The overall workshop structure is illustrated in Table 4. Table 4 illustrates the number of participants, including the table leaders. The lowest number of participants in any room was seven, and the greatest number of participants in a room was nine. Appendix A, Standard Setting Panelists, contains the background information of the panelists.

Table 4: Overall Workshop Structure

Panel/ Rooms	Panelists	Table Leader Panelists	Subject	Grade
1	7	2	ELA	3, 4
2	6	2	ELA	5, 6
3	6	2	ELA	7, 8
4	7	2	ELA	10
5	6	2	Mathematics	3, 4
6	6	2	Mathematics	5, 6
7	7	2	Mathematics	7, 8
8	7	2	Mathematics	10
9	6	2	Science	4
10	6	2	Science	6
11	7	2	Science	Biology
12	5	2	Social Studies	5
Totals	76	24		

Table 5 summarizes the staff assignments for the workshop.

Table 5: Staff Assignments

Rooms	Subject	Grade	IDOE Staff	AIR Facilitator	AIR Assistant
Suite 9 (2nd Floor)	ELA	3, 4	Mark O'Malley	Katherine Krehbiel	Kelsey Conklin
Suite 10 (2nd Floor)	ELA	5, 6	Kelly Connelly (Niki Smithers on Monday morning)	Brett Craycraft	Hashim Evans
Plaza C (2nd Floor)	ELA	7, 8	Kristine David	Krista Bobbitt	Maureen Nalepa
Plaza D (2nd Floor)	ELA	10	Erin Thompson	Ann Harshbarger	Alexa McDorman
Suite 11 (2nd Floor)	Mathematics	3, 4	Justin Mocas	Peter Pluckebaum	Rachael Day
Suite 12 (2nd Floor)	Mathematics	5, 6	Sholonda Trice/Tobin Richardson	Eileen Heneghan	Lucas Melo
Suite 13 (2nd Floor)	Mathematics	7, 8	Charity Flores	Chris Kincheloe	Ronnie Pacini
Suite 14 (2nd Floor)	Mathematics	10	Andrew Jones	Paul Maxon	Maya Lewis
Plaza E (2nd Floor)	Science	4, 6	Karen Davies/ Niki Smithers (after Monday morning)	Cynthia Carr	Marjory Cohen
Suite 16 (2nd Floor)	Science	Biology	Tim Martin	Gabe Martinez	Kevin Cleary
Crosspointe Suite (Ground Floor)	Social Studies	5	Stephanie Thompson/ Felecia Jordan	Alex Linville	Scott Koenig

3.3 PARTICIPANTS AND ROLES

3.3.1 Indiana Department of Education Staff

IDOE staff were present throughout the process, and they provided overall policy context and answered any policy questions that arose. Staff represented the Office of Student Assessment (OSA), Test Development, and Special Education Services. They included the following:

- Dr. Charity Flores
- Kelly Connelly
- Dr. Kristine David
- Karen Davies
- Andrew Jones
- Felecia Jordan
- Tim Martin
- Justin Mocas
- Mark O'Malley
- Dr. Tobin Richardson
- Niki Smithers
- Erin Thompson
- Stephanie Thompson
- Sholonda Trice

3.3.2 Indiana Technical Advisory Committee

Indiana TAC member was also present and observed the process. Indiana TAC included the following:

- Dr. Chad Buckendahl

3.3.3 AIR Staff

AIR facilitated the workshop and each of the content-area rooms, provided psychometric and statistical support, and oversaw technical set-up and logistics. AIR team members included the following:

- Dr. Gary Phillips, AIR Vice President and Institute Fellow, facilitated and oversaw the workshop. He provided training to all participants, including the facilitators, the table facilitators, and all participants, and he supervised the psychometric analyses conducted during and after the workshop.
- Meg McMahon, AIR Vice President, Content and Test Development, oversaw the workshop and supervised the facilitation of each meeting room.

- Dr. Hyesuk Jang, Psychometric Support Manager, oversaw the set-up of analytics technology and psychometrics.
- Irene Hunting, Project Director, oversaw the workshop, supervised the program management team, and ensured that all logistics were accounted for.
- Jim Unger, Systems Support Analyst, and Mark Palomo, Systems Support Analyst, set up, tested, and performed troubleshooting on all technology during the workshop.

3.3.4 Room Facilitators

AIR provided a room facilitator and an assistant room facilitator for each room to guide the standard-setting process. Facilitators were content experts experienced in leading standard-setting processes and could answer any questions about the process, the items, and what the items were intended to measure. They also monitored time and motivated panelists to complete tasks within the scheduled time.

Before the workshop, it was necessary to ensure that each room facilitator was extensively knowledgeable about the constructs, processes, and technologies used in standard setting. Thorough training is essential to standardize the training and procedures across the grade/subject committees.

AIR facilitators were trained for their leadership role in the conduct of the standard-setting workshop. Before the workshop, all involved AIR staff participated in at least two comprehensive internal training sessions, in which the facilitators and their assistants were trained to use AIR's online standard-setting tool. This training covered six important functions:

- Operating and following the steps in the online standard-setting tool
- Taking the online assessment
- Placing bookmarks online
- Practicing leading discussions and getting feedback on information from Rounds 1 and 2
- Reviewing all workshop materials
- Conducting an online evaluation

3.3.5 Table Leaders

The IDOE pre-selected table leaders from the participant pool for their specialized knowledge or experience with the assessment, items, or standards. Table leaders also served as panelists and made their cut score recommendations along with the other table members.

As with room facilitators, it was necessary to ensure that each table leader was knowledgeable about the constructs, processes, and technologies used in standard setting and able to adhere to a standardized process across the grade/subject committees.

Table leaders were trained online as a group by IDOE staff before the standard-setting workshop. Training consisted of an overview of the table leaders' responsibilities and process guidance.

Table leaders fulfilled the following functions during the workshop:

- Helping panelists see the big picture
- Leading table discussions
- Supporting panelists with tasks
- Monitoring security of materials
- Reporting issues or misunderstandings to room facilitators
- Maintaining a supportive atmosphere of professionalism and respect

3.3.6 Educator Participants

To set the bookmarks, the IDOE recruited a diverse set of participants from across the state.

The IDOE selected a broadly representative group of 100 panelists based on teaching experience, qualifications, demographics, and specialized expertise, ensuring that a diverse range of perspectives informed the standard-setting process. The IDOE designated the most knowledgeable and experienced panelists at the workshop as table leaders.

Panelists included special education teachers, general education teachers, curriculum specialists, education administrators, and other stakeholders. The panel was also broadly representative of Indiana's special education teacher population in terms of gender, race/ethnicity, and regional composition. Table 6 is a numeric breakdown of participant characteristics.

Table 6: Panelist Characteristics

	Percentage of Panelists by Subject and Grade												**Overall Percent**
	ELA				**Mathematics**				**Science**			**Social Studies**	
Group	3/4	5/6	7/8	10	3/4	5/6	7/8	10	4	6	Bio	5	
Gender													
Male	0	0	0	11	25	25	11	22	13	13	33	14	14
Female	100	100	100	89	75	75	89	78	88	88	67	86	86
Race/Ethnicity													
Asian	11	0	0	0	0	0	0	0	0	0	0	0	1
Black	0	13	0	0	0	0	0	0	0	0	11	14	3
Hispanic	0	0	0	0	0	0	0	0	0	13	0	0	1
White	89	88	100	100	100	100	100	100	100	88	89	86	94

	Percentage of Panelists by Subject and Grade												Overall Percent
	ELA				Mathematics				Science			Social Studies	
Current Region of Occupation													
Central	67	50	38	44	38	50	33	22	63	63	22	29	43
North	11	38	25	33	38	50	33	44	38	25	67	29	36
South	22	13	38	22	25	0	33	33	0	13	11	43	21
Stakeholder Group*													
Special Education Teacher	56	25	50	56	63	50	44	56	63	50	0	43	46
General Education Teacher	44	75	75	56	38	38	56	67	38	50	78	43	55
ESL Teacher	0	0	0	0	0	0	0	11	0	0	0	14	2
Instructional Coach	0	0	0	0	0	0	0	11	13	0	0	0	2
Specialist	0	0	13	0	0	0	0	0	0	0	0	14	2
Higher Education Teacher	0	0	0	0	0	0	0	0	0	0	11	14	2
Administrator	11	0	0	22	13	13	0	0	0	13	22	14	9
School Level Taught Within Assigned Subject													
ES	67	50	13	0	38	38	0	0	38	25	11	43	27
MS	0	0	63	0	0	38	56	0	0	25	0	14	16
HS	0	13	13	33	0	13	0	56	25	13	78	0	20
ES and MS	22	25	13	0	38	13	11	11	38	25	0	14	17
ES and HS	0	0	0	0	13	0	0	0	0	0	0	0	1
MS and HS	0	0	0	11	13	0	22	0	0	0	0	14	5
ES, MS, and HS	0	0	0	22	0	0	0	11	0	0	0	0	2
N/A	11	13	0	33	0	0	11	22	0	13	11	14	11

**Note: Stakeholder group percentages do not add up to 100 because of overlap between positions.*

For results of the Bookmark method to be valid, the judgments must be made by individuals who are qualified to make them. Participants in the Indiana standard-setting workshop were highly qualified because of the variety of expertise and number of years of experience in instruction, curriculum, assessment, and special student populations. Most had taught for 11 years or more, and many held relevant certifications. They also represented a range of stakeholders, such as educators, administrators, and college faculty. Table 7 summarizes the qualifications of the panels.

Table 7: Panelist Qualifications

	Percentage of Panelists by Subject and Grade												Overall Percent
	ELA				Mathematics				Science			Social Studies	
Grade	3/4	5/6	7/8	10	3/4	5/6	7/8	10	4	6	Bio	5	
Years of Teaching Experience													
5 Years or Less	0	13	0	22	13	13	11	0	38	25	11	14	13
6 to 10 Years	11	25	13	11	38	13	22	33	13	13	22	14	19
11 Years or More	89	63	88	67	50	75	67	67	50	63	67	71	68
Years of Professional Experience in Education (other than classroom teaching)													
5 Years or Less	56	75	50	67	63	50	67	89	63	75	22	43	60
6 to 10 Years	11	13	13	11	13	0	0	11	13	13	0	43	11
11 years or More	33	13	38	22	25	50	33	0	25	13	78	14	29
Highest Degree Earned													
Bachelor	67	38	13	11	25	38	56	44	63	50	33	29	39
Master	33	63	88	78	75	63	44	56	25	50	56	71	58
Doctorate	0	0	0	0	0	0	0	0	13	0	11	0	2
Other	0	0	0	11	0	0	0	0	0	0	0	0	1

3.4 MATERIALS

3.4.1 Ordered-Item Booklets

The Bookmark method utilizes OIBs as the key tool for setting standards. All items contained in each OIB were accepted by the IDOE and its content specialists and advisors. For each grade and subject, a 40- or 41-item OIB, depending on the assessment, was assembled with operational and field test items selected from each tier of the stage-adaptive I AM assessment administered in Spring 2019. AIR's standard-setting tool provided panelists with online access to the OIBs.

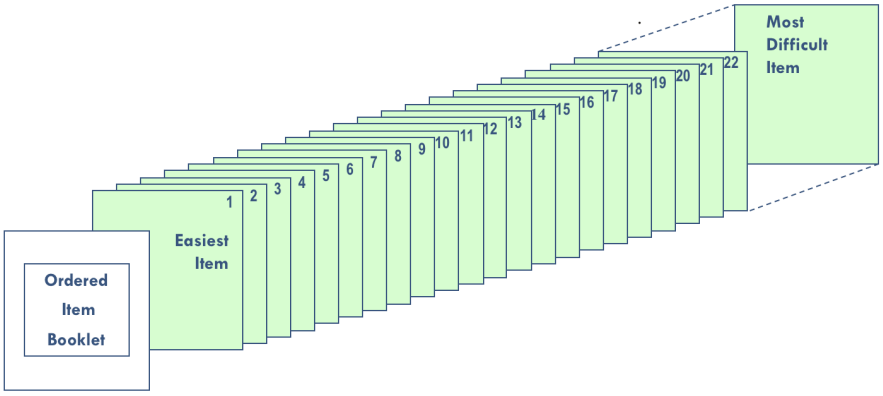
The items in each OIB were sorted in ascending order by IRT item difficulty, calculated by Winsteps and indicated by RP50. For one-point items, RP50 is the item difficulty point where a student has a 50% probability of answering the item correctly. Two-point items

appear on two pages, corresponding to each score point. Each page represents the item difficulty level where the students earned that score point with 50% probability.

While RP67 is the most used probability in standard setting (Huynh, 2006; Willians & Schulz, 2005), both RP50 and RP67 were considered in construction of OIB. After the evaluation of OIBs from both RP50 and RP67, RP50 that provided the better alignment of impacts was adopted. This is also consistent with previous practice in Indiana.

Each page of the online OIB presents a single item at a single score point, with the easier items located in the front of the OIB and the more difficult items in the back of the OIB. Panelists place the bookmark on the last page where a "Just Barely" student would respond to that item with at least 50% chance of answering it correctly. Each page of the OIB can correspond to a cut score; thus, when panelists place their "bookmark" for a performance level, they are in fact selecting the performance standard, indicated by the RP50 value of the item, for that performance level.

Figure 7: Ordered-Item Booklet



For the I AM assessments, the OIBs contained 40 or 41 assessment items. From the operational and field test items, there were insufficient items with the unique RP50 values that meet the blueprint for some reporting categories. Therefore, the strategy we used to select items for OIBs was to first include all the items with the unique RP50 values and add more items with the duplicate RP50 values for the OIBs that didn't meet the minimum of blueprints at the reporting categories. In some instances, it was necessary to over-emphasize one or more reporting categories in order to appropriately minimize gaps in the OIB.

The composition of the OIBs by assessment and grade are summarized in Table 8. The number of items for each reporting category is also provided in Table 9. 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.

Items in the 2019 item pool for each assessment were screened for use in the OIB. Selected items met the following three criteria:

- The items went through a rigorous item data review process that included teacher committee reviews.
- The OIB met the minimum of test blueprint.
- The sorted RP50 values produced acceptable fine gaps between adjacent OIB pages. This required removal of items with duplicate RP50 values.

Table 8: The Composition of the Ordered-Item Booklets

Grade	Number of Items in OIB			Pages in OIB (Total Points)
	Operational	Field Test	Total	
ELA 3	32	8	40	40
ELA 4	32	8	40	40
ELA 5	31	10	41	41
ELA 6	33	7	40	40
ELA 7	32	8	40	40
ELA 8	35	6	41	41
ELA 10	31	9	40	40
Mathematics 3	34	6	40	40
Mathematics 4	32	8	40	40
Mathematics 5	35	5	40	40
Mathematics 6	36	4	40	40
Mathematics 7	29	11	40	40
Mathematics 8	37	3	40	40
Mathematics 10	29	11	40	40
Science 4	31	9	40	40
Science 6	33	6	39	41
Biology	32	8	40	40
Social Studies 5	30	10	40	40

Table 9:Number of Items for Each Reporting Category

Subject	Grade	Reporting Category	Blueprint Minimum	Blueprint Maximum	Number of Items in OIB
ELA	3	KITS	7	10	14
	3	RF	7	10	10
	3	SECM	7	8	8
	3	W	7	8	8
	4	KITS	11	13	18
	4	SECM	10	12	11
	4	W	7	8	11
	5	KITS	11	14	17
	5	SECM	9	12	12
	5	W	7	9	12
	6	KITS	9	12	12
	6	SECM	8	11	11
	6	W	7	8	15
	6	SL	1	2	2
	7	KITS	9	14	14
	7	SECM	8	11	11
	7	W	7	8	13
	7	SL	1	2	2
	8	KITS	9	14	21
	8	SECM	8	11	10
	8	W	7	8	8
	8	SL	1	2	2
	10	KITS	9	13	18
	10	SECM	8	11	12
	10	W	7	8	8
	10	SL	1	2	2
Mathematics	3	ATDA	7	8	8
	3	C	7	8	11
	3	GM	7	8	8
	3	NS	7	8	10
	3	PS	2	4	3
	4	ATDA	7	8	13
	4	C	7	8	8
	4	GM	7	8	8
	4	NS	7	8	7
	4	PS	2	4	4
	5	AT	7	8	8
	5	C	7	8	8
	5	GMDAS	7	8	13
	5	NS	8	9	9
	5	PS	1	4	2
Mathematics	6	AF	8	9	9

Subject	Grade	Reporting Category	Blueprint Minimum	Blueprint Maximum	Number of Items in OIB
	6	C	7	8	8
	6	GMDAS	7	8	8
	6	NS	8	9	14
	6	PS	1	4	1
	7	AF	8	9	9
	7	DASP	7	8	13
	7	GM	7	8	8
	7	NSC	7	8	8
	7	PS	1	2	2
	8	AF	9	10	10
	8	DASP	7	8	8
	8	GM	7	8	8
	8	NSC	7	8	12
	8	PS	1	2	2
	10	EI	7	8	8
	10	F	7	8	15
	10	NSDA	7	8	7
	10	GM	7	8	8
	10	PS	1	4	2
Science	4	AICT	7	8	8
	4	ESRC	7	8	13
	4	I	7	8	8
	4	QM	8	11	11
	6	AICT	7	8	12
	6	ESRC	7	8	9
	6	I	8	11	8*
	6	QM	7	8	8
	10	ADMT	13	16	16
	10	CEEC	7	8	8
	10	UM	9	11	16
Social Studies	5	CGH	16	18	23
	5	ECON	7	8	9
	5	GEO	7	8	8

**Note: 8 items for 10 pages including two 2-point items*

3.4.1.1 Background of 2019 Operational Field-Tested/Field-Tested Items

Indiana is updating and revising the item bank through new development and field-test activities including both operational field tests used for scoring and field tests not for scoring. The items that were field tested in Spring 2019 were reviewed and approved by the IDOE and added to the I AM item bank for operational use in Spring 2020.

3.4.1.2 A Brief Review of Item Development, Analysis, and Review

I AM items are written to the Indiana Content Connectors and I AM Item Specifications for each grade and subject. The PLDs, developed by educators across the state of Indiana, provide a full description of content to be targeted and tested for students with significant cognitive disabilities. Based on the IAS and the Indiana Content Connectors, the Range PLDs preserve the essence of the grade-level expectations, but they may modify the scope or complexity of the general education standards or take the form of introductory or prerequisite skills to the grade-level standards. The Range PLDs, directly linked to the Indiana Content Connectors, are expressed in terms of student content behaviors at three performance levels:

- Level 1 – Below Proficiency
- Level 2 – Approaching Proficiency
- Level 3 – At Proficiency

The IDOE provides the Range PLDs on its website.

Item analyses of the Spring 2019 administration included classical item analysis, differential item functioning (DIF) analysis, and IRT analysis. No vertical scale was considered.

In the IRT calibration, all items were concurrently calibrated (within grade) using Winsteps and Masters' partial credit model (Masters, 1982).

Flagged items went to IDOE item data review in mid-June 2019, and the standard-setting workshop used items that were approved by the IDOE.

3.5 Workshop Technology

Panelists used AIR's online tool for standard setting. Using this tool, panelists reviewed the content alignment and score points for each item, placed multiple rounds of bookmarks, and evaluated the impact that proposed cut scores will have on students. Panelists also saw their own bookmark placements, their table's bookmarks, the other tables' bookmarks, and the overall bookmarks for all grades. They were able to add notes and comments on the items as they reviewed them and examine reference and benchmark data on screen following each round.

Each panelist used an AIR laptop or Chromebook to take the assessment, review items and ancillary materials, and place bookmarks.

A full-time AIR IT specialist oversaw laptop setup and testing, answered questions, and ensured that technological processes ran smoothly and without interruption throughout the meeting.

3.6 EVENTS

The standard-setting workshop occurred over a period of three days. Table 10 summarizes each day's events, and this section describes each event listed in greater detail. Appendix B, Workshop Agenda, provides the complete agenda.

Table 10: Standard-Setting Agenda Summary

Day 1: Monday, July 22	Day 2: Tuesday, July 23	Day 3: Wednesday, July 24
● Orientation and introductions ● Breakout rooms ● Take the test ● Review content standards ● Review and confirm Range PLDs ● Write "Just Barely" Target PLDs ● Review OIBs ● Day 1 Evaluation Form	● Practice placing bookmarks ● Standard-setting quiz and readiness evaluation ● Place Round 1, 2, and 3 bookmarks for ELA and Mathematics grades 4, 6, 8, and 10; Science grade 4, grade 6, and Biology; and Social Studies grade 5 ● Review feedback, impact data, and benchmark data and discuss ● Write "Just Barely" Target PLDs for ELA and Mathematics grades 3, 5, and 7	● Review PLDs for ELA and Mathematics grades 3, 5, and 7 ● Place Round 1, 2, and 3 bookmarks for ELA and Mathematics grades 3, 5, and 7 ● Standard-setting workshop evaluations ● Final moderation

3.6.1 Orientation

Dr. Charity Flores from the IDOE and Dr. Gary Phillips from AIR welcomed panelists to the workshop in a large group setting. Dr. Flores gave a comprehensive overview of the importance of the standard-setting activity and emphasized the importance of keeping a content focus throughout the three-day workshop.

Dr. Phillips described the purpose and objectives of the meeting, explained the process designed to meet those objectives, and outlined the events that would happen each day. He outlined the responsibilities of the three groups of people at the workshop: panelists, AIR staff, and IDOE personnel. He explained that panelists were selected because they were experts, and he described how the process to be implemented over the three days was designed to elicit and apply their expertise to recommend new cut scores. He described how standard setting works and what would happen once the panelists had finalized their recommendations.

3.6.2 Confidentiality and Security

Standard setting uses live test items, requiring confidentiality to maintain their security. Participants were NOT allowed to do the following during and after the workshop:

- Discuss the test items outside of the meeting
- Remove any secure materials from the room on breaks or at the end of the day
- Discuss judgments or cut scores (theirs or others) with anyone outside of the meeting
- Discuss secure materials with non-participants
- Use cell phones in the meeting rooms (they were asked to turn off cell phone ringers)
- Take notes on anything other than provided materials
- Bring any other materials to the workshop

Participants were told that they could have general conversations about the process and days' events, but workshop leaders warned them against discussing details, particularly those involving items, cut scores, and any other confidential information.

Following the large-group orientation and discussion about security, the panelists moved to their individual work groups.

3.6.3 Take a Short Practice Test

Following the large-group training, panelists broke into their assigned rooms, where they took a short, fixed-form version test of the OIB. While testing, panelists were not allowed to discuss the items, hold any conversations, or access their phones.

Taking the items from the same pool and delivery system as the students provides the opportunity to interact with and become familiar with the look and feel of the student experience while testing.

3.6.4 Review the Indiana Alternate Academic Standards, Content Connectors, and Range PLDs

After finishing the practice assessment, panelists completed a thorough review of the IAAS, the Indiana Content Connectors, the Policy PLDs, and the Range PLDs for their grade and subject area. They identified key words describing the skills necessary for performance at each level and discussed the skills and knowledge that differentiated performance in each performance level.

Reviewing the standards and PLDs ensured that participants understood what students in Indiana are expected to know and be able to do, particularly the knowledge and skills students are expected to demonstrate at each performance level.

3.6.5 Write Target PLDs

As discussed in Section 3, Standard Setting, Target PLDs play a crucial role in the standard-setting task, as they describe the specific student behaviors that are of interest in placing a bookmark.

After reviewing and discussing the Indiana Content Connectors and PLDs, panelists worked in their table groups to draft Target PLDs that describe the skills that students scoring as Just Barely in one performance level have that students just below the performance standard do not have. Target PLDs describe students who are at the entry point of the range. At the "Just Barely" threshold, these students do reach the standard.

Extensive discussions took place within each of the panels so that each panelist had a clear vision of what a "Just Barely" performance meant. Clarity was based on the cognitive behaviors identified in the Range PLDs for each performance level.

After the Target PLDs for each performance level within each subject and grade were established, participants reviewed the online OIB for the assessment.

3.6.6 Bookmark Training and Placement

Sections 3.6.6.1 through 3.6.6.5 provide detailed information of the training that the participants received leading up to the actual Bookmark Placement Task (see Section 3.6.6.5: Round One Bookmark Placement).

3.6.6.1 Initial OIB Review

Armed with in-depth knowledge of the content structure of the assessment, panelists did an initial review of the OIB itself. The purpose of this review was to allow the panelists to obtain an overall impression of how the content standards were manifested in the form of test items. Panelists were encouraged to also consider what features of an item made it more difficult than the preceding items. No attempt was made to place a bookmark; rather, panelists discussed the items that seemed to naturally reflect one of the performance levels.

Part of the item discussions included an examination of the types of cognitive demands made on students and the types of behaviors, consistent with the eligible content, that students would have to demonstrate to succeed on the items.

3.6.6.2 How to Place a Bookmark

In preparation for the actual Bookmark Placement Task, extensive discussions took place within each room on the dynamics of bookmark placement. Multiple considerations in the judgment process were reviewed with the panelists by the room facilitators.

An important component of the Bookmark Placement Task is completely understanding the RP50 criterion.

The objective of standard setting is aspirational: Panelists think about the Target PLDs that describe students just barely meeting each performance level and identify what all students should know and be able to do, not what they actually know and can do as they review the OIB.

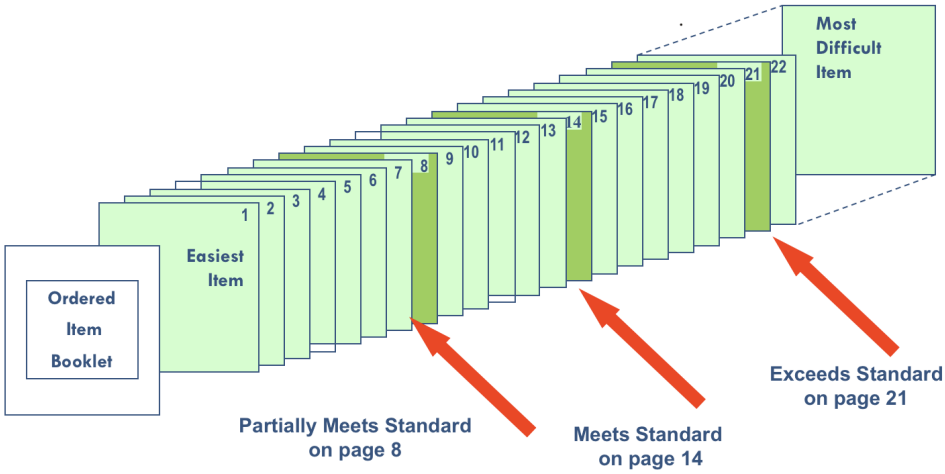
Panelists applied a 50% response probability rule when placing bookmarks. This rule requires panelists to identify the page in the OIB where 50% or more of students who just barely meet the standard (those described by the Target PLDs) would be able to answer the item on that page correctly.

The explanation of this rule provided to panelists was as follows:

> "Of 100 students who are 'Just Barely' at the standard, would at least 50% of students get this item correct?"

These "Just Barely" students are more likely to be able to correctly answer items at the beginning of the OIB and are less likely to be able to correctly answer items towards the end of the OIB. Items beyond that point in the OIB are items that less than 50% of the "Just Barely At Proficiency" students would correctly answer. Panelists place their bookmark on the first page in the OIB where they believe the "Just Barely At Proficiency" student would NOT have at least a 50% chance of answering correctly. Panelists repeated this process for the "Just Barely Approaching Proficiency" student.

Figure 8: Example Bookmark Placement



Workshop leaders from AIR advised panelists that, while some items may seem out of order, the item order is determined by item difficulty, which is computed from actual student performance on the items, not by content or cognitive process. The ordering of items in the OIB does not necessarily follow the sequence of instruction or the order of item presentation on the assessment.

Panelists were also informed that the placement of a bookmark may require panelists to evaluate a series of items for their demands on the student and that there may not be a perfectly clear location to place the bookmarks. Sometimes, sets of items that cluster near a cut point may need to be reviewed, and an expert judgment should be made about

where, in this item cluster, a cut score should be placed to best capture the “Just Barely” performance.

To keep panelists focused on the standard-setting task, and not on item critique, panelists referred item-related questions or comments to room facilitators and, in some cases, to IDOE staff to answer. Bookmarks were not to be placed on any item that panelists disagreed with or felt might be incorrect or unfair.

3.6.6.3 Bookmark Placement Practice Quiz

The purpose of the practice quiz (included as Appendix D, Bookmark Placement Practice Quiz) was to ensure that panelists were comfortable with the technology and the Bookmark Placement Task prior to setting any actual bookmarks. Each panelist took the practice quiz.

The quiz assessed panelists’ understanding in multiple ways. They must

- indicate on a diagram how performance standards and levels work together and where students just barely meeting each of the standards fall;
- answer questions about relative item difficulty in a hypothetical OIB; and
- demonstrate understanding by correctly applying the 50% rule to a hypothetical bookmark placement.

Following the administration of the practice quiz, an in-depth discussion took place in which panelists shared their understandings and any areas that may have been unclear. Panelists asked questions, and the room facilitators provided clarifications and further instructions until everyone felt comfortable with the task. At the end of the discussion, panelists expressed their view that the Bookmark Placement Task was clear.

3.6.6.4 Bookmark Placement Readiness Form

After completing the practice quiz, but before placing bookmarks, panelists completed a Bookmark Placement Readiness Form (Appendix E, Standard Setting Educator Panel – Readiness Form).

Every panelist must affirm their complete readiness to move to the Bookmark Placement Task on the Bookmark Placement Readiness Form before beginning the Bookmark Placement Task. Any panelist who is unable to affirm understanding will receive individual support until he or she can complete the form and move forward.

All panelists completed the form and were prepared and ready to place their bookmarks.

3.6.6.5 Round 1 Bookmark Placement

After completing the training activities, panelists were ready to set initial recommendations for cut scores. They began by identifying pages for the At Proficiency level first, followed by the Approaching Proficiency level for each grade and subject.

Panelists made their recommendation for the At Proficiency performance standard, followed by the recommendations for Approaching Proficiency. They spent the most time on the cut score for the At Proficiency standard, which helped them anchor the remaining level. The placement of the At Proficiency cut score in the OIB must leave enough room before it for the Approaching Proficiency cut score. A pop-up window in the online tool prompted panelists to verify their recommended cut scores before final submission.

After Round 1, but before Round 2, the room facilitators presented online feedback based on panelist recommendations to the panelists in their room. The feedback was in the form of statistics that described the central tendency and variability of the panelists' ratings. It included the median, minimum, and maximum ratings. The median is used because page numbers represent ordinal data, not interval data. To facilitate the discussion, the online system presented bar charts showing the comparison of a room median cut score, a table median cut score, and an individual panelist cut score. Further, before starting Round 2, the room facilitators introduced benchmark data and described how those data were to be used in reviewing the placement of the Round 2 bookmarks. Similar discussions were also held prior to the Round 3 bookmark placement, but impact data were added for panelist consideration in placing their Round 3 bookmarks.

Tables 11–14 present the results from Round 1 of the standard-setting activity. In addition to the median recommended page number for each subject, benchmark and impact data are presented in Tables 11–14 for the reader's convenience. The panelists did not see benchmark or impact data during their Round 1 bookmark placements. Feedback and benchmark data were reviewed prior to the Round 2 bookmark placement. Impact data were not presented until prior to the Round 3 Bookmark Placement Task.

Two types of benchmark data were available to the panelists. First, cut scores for the I AM were benchmarked against ILEARN, which measures general education student achievement and growth consistent with the IAS. ILEARN is the summative accountability assessment for Indiana students, and the data from the Spring 2019 administration were used for the benchmarking activity. Each page in the OIB indicated which performance level is compatible on the ILEARN assessment (ILEARN data were presented for panelist reference on the PowerPoint presentation).

The second source of benchmarking data for the panelists was the data from the most recent (2015) administration of the NCSC assessment. The NCSC is a multi-state, multi-organizational consortium funded through a General Supervision Enhancement Grant (GSEG) from the United States Education Department (USED) Office of Special Education Programs (OSEP). The NCSC assessments are the multistate comprehensive assessments for students with significant cognitive disabilities. For each page of the OIB, panelists were able to see which performance level was compatible on the NCSC assessment.

The impact data are from the Spring 2019 I AM assessment, and it was presented in advance of the Round 3 Bookmark Placement Task. The data were presented by grade and subject and indicated the percentage of students who would reach a performance standard if that page in the OIB were selected for the cut score for that standard.

Cell entries of "NA" indicate that no data were available.

Table 11: Round 1 Results: ELA, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included Only for Comparative Purposes

Subject and Grade	Median Round 1 Bookmark (Page #)		Range Round 1 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
ELA								
Grade 3	10	14	5	8	51.41	37.15	46	51
Table 1	10	14	5	8	51.41	37.15		
Table 2	10	14	1	5	51.41	37.15		
Grade 4	15	28	6	13	55.12	32.61	45	56
Table 1	16	29	6	11	50.36	29.44		
Table 2	15	24	5	13	55.12	39.83		
Grade 5	10	18	5	6	72.57	50.00	47	58
Table 1	10	17	2	2	72.57	51.22		
Table 2	10	19	5	6	72.57	45.14		
Grade 6	14	22	7	17	49.89	33.78	47	63
Table 1	14	25	7	17	49.89	29.95		
Table 2	14	21	7	4	49.89	33.78		
Grade 7	9	18	4	3	78.56	49.78	49	56
Table 1	9	19	4	2	78.56	44.53		
Table 2	9	17	2	1	78.56	49.78		
Grade 8	12	24	6	15	67.57	37.67	50	64
Table 1	12	26	5	5	67.57	32.91		
Table 2	12	22	6	11	67.57	42.33		
Grade 10*	13	27	8	12	79.09	49.31	50	70
Table 1	13	27	4	10	79.09	49.31		
Table 2	17	26	6	6	69.63	52.86		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

**Because ILEARN was not administered in grade 10, the grade 10 benchmarking activities used the data from the ILEARN grade 8.*

Table 12: Round 1 Results: Mathematics, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included Only for Comparative Purposes

Subject and Grade	Median Round 1 Bookmark (Page #)		Range Round 1 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Mathematics								
Grade 3	7	14	4	9	67.09	40.94	58	73
Table 1	8	14	4	6	67.09	40.94		
Table 2	7	14	4	5	67.09	40.94		
Grade 4	14	24	14	11	40.46	25.43	53	53
Table 1	16	24	14	11	38.73	25.43		
Table 2	14	25	4	4	40.46	24.86		
Grade 5	7	11	4	5	60.67	43.89	47	57
Table 1	7	12	2	3	60.67	43.89		
Table 2	8	11	4	5	53.15	43.89		
Grade 6	15	25	17	22	43.00	20.20	46	58
Table 1	15	24	17	12	43.00	20.20		
Table 2	16	28	4	16	38.71	15.12		
Grade 7	8	11	3	6	58.52	46.72	41	68
Table 1	8	12	3	6	58.52	34.61		
Table 2	6	10	2	4	69.76	46.72		
Grade 8	16	31	13	14	25.12	6.18	37	61
Table 1	16	28	3	9	25.12	8.12		
Table 2	10	31	10	7	41.64	6.18		
Grade 10*	14	25	16	13	42.14	18.30	37	57
Table 1	13	26	16	12	42.14	14.24		
Table 2	14	25	9	9	42.14	18.30		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

**Because ILEARN was not administered in grade 10, the grade 10 benchmarking activities used the data from the ILEARN grade 8.*

Table 13: Round 1 Results: Science, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included Only for Comparative Purposes

Subject and Grade	Median Round 1 Bookmark (Page #)		Range Round 1 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Science							NA	NA
Grade 4	22	31	16	5	35.40	19.86		
Table 1	16	28	11	5	48.63	24.17		
Table 2	27	33	10	5	26.91	19.57		
Grade 6	16	26	16	10	56.57	28.34		
Table 1	10	22	8	4	71.43	41.37		
Table 2	18	28	7	7	48.80	23.20		
Biology	10	17	12	12	79.05	60.42		
Table 1	15	22	12	12	66.84	42.53		
Table 2	10	17	3	3	79.05	60.42		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

Table 14: Round 1 Results: Social Studies, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included Only for Comparative Purposes

Subject and Grade	Median Round 1 Bookmark (Page #)		Range Round 1 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Social Studies							NA	NA
Grade 5	16	25	11	14	35.60	26.09		
Table 1	16	26	0	3	35.60	22.83		
Table 2	14	21	11	14	41.30	30.30		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

3.6.6.6 Round 2 Bookmark Placement

After placing the Round 1 bookmarks, workshop facilitators provided panelists with additional instructions for placing the Round 2 bookmarks. They described the goal of Round 2 as one of convergence, not consensus, on a common achievement standard.

Workshop facilitators reviewed the feedback data from Round 1, in which panelists could see how their individual recommended cut points compared with the other panelists in their room and for the room as a whole.

Following the discussion of the feedback data, the room facilitators introduced a discussion of benchmark data and how they would be used to inform the Bookmark Placement Task in Round 2. The facilitator explained how the panelists would see impact data for each cut score location they considered for their Round 2 placement. These data are automatically generated by AIR's bookmarking application.

Together, this information informed, but did not determine, panelist Round 2 decisions. Panelists discussed this information and the Round 1 cut scores before placing Round 2 bookmarks.

Round 2 results reflect panelist consideration of feedback and benchmark data. Tables 15–18 present the bookmarks and associated impact and benchmark data for each subject.

Table 15: Round 2 Results: ELA: Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 2 Bookmark (Page #)		Range Round 2 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
ELA								
Grade 3	9	12	3	4	57.21	44.51	46	51
Table 1	9	12	1	3	57.21	44.51		
Table 2	10	13	3	3	51.41	37.30		
Grade 4	13	20	6	8	59.74	45.31	45	56
Table 1	13	19	1	1	59.74	45.31		
Table 2	15	23	5	7	55.12	39.83		
Grade 5	10	16	3	4	72.57	55.14	47	58
Table 1	10	16	3	2	72.57	55.14		
Table 2	10	17	2	2	72.57	51.22		

Subject and Grade	Median Round 2 Bookmark (Page #)		Range Round 2 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Grade 6	11	18	8	10	57.09	43.58	47	63
Table 1	11	18	4	7	57.09	43.58		
Table 2	9	17	7	8	65.43	43.58		
Grade 7	9	18	1	2	78.56	49.78	49	56
Table 1	9	19	1	1	78.56	44.53		
Table 2	9	17	1	1	78.56	49.78		
Grade 8	11	21	6	12	71.26	42.82	50	64
Table 1	9	21	4	12	78.16	42.82		
Table 2	11	21	4	4	71.26	42.82		
Grade 10*	13	27	0	3	79.09	49.31	50	70
Table 1	13	27	0	3	79.09	49.31		
Table 2	13	26	0	1	79.09	52.86		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

**Because ILEARN was not administered in grade 10, the grade 10 benchmarking activities used the data from the ILEARN grade 8.*

Table 16: Round 2 Results: Mathematics: Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 2 Bookmark (Page #)		Range Round 2 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Mathematics								
Grade 3	6	10	1	1	71.02	59.21	58	73
Table 1	6	10	1	0	71.02	59.21		
Table 2	6	10	0	1	71.02	59.21		
Grade 4	8	13	7	16	63.73	47.69	53	53
Table 1	7	13	2	5	68.50	47.69		
Table 2	13	22	0	2	47.69	28.47		
Grade 5	7	11	3	1	60.67	43.89	47	57
Table 1	7	11	3	1	60.67	43.89		
Table 2	8	11	1	0	53.15	43.89		
Grade 6	8	13	7	6	65.58	46.95	46	58
Table 1	8	12	6	3	65.58	51.35		
Table 2	9	13	7	3	65.58	46.95		
Grade 7	6	10	3	4	69.76	46.72	41	68
Table 1	8	11	3	4	58.52	46.72		
Table 2	6	10	2	1	69.76	46.72		
Grade 8	7	17	11	15	44.54	18.55	37	61
Table 1	12	17	11	15	35.07	18.55		
Table 2	6	18	1	8	55.17	18.55		
Grade 10*	13	21	6	12	42.14	24.33	37	57
Table 1	13	22	4	12	42.14	24.33		
Table 2	12	19	6	4	42.14	25.02		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

**Because ILEARN was not administered in grade 10, the grade 10 benchmarking activities used the data from the ILEARN grade 8.*

Table 17: Round 2 Results: Science, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 2 Bookmark (Page #)		Range Round 2 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Science							NA	NA
Grade 4	13	20	10	19	55.54	37.12		
Table 1	15	28	7	12	48.63	24.17		
Table 2	12	17	3	5	56.69	43.17		
Grade 6	12	22	8	5	68.11	41.37		
Table 1	11	22	1	1	71.43	41.37		
Table 2	15	23	5	4	56.57	41.37		
Biology	10	17	8	8	79.05	60.42		
Table 1	8	16	8	8	84.95	66.32		
Table 2	10	17	0	0	79.05	60.42		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

Table 18: Round 2 Results: Social Studies, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 2 Bookmark (Page #)		Range Round 2 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Social Studies							NA	NA
Grade 5	13	19	13	13	41.30	30.71		
Table 1	13	19	3	6	41.30	30.71		
Table 2	10	18	13	13	48.23	30.71		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

3.6.6.7 Round 3 Bookmark Placement

Tables 19–22 contain the Round 3 panelist recommendations for the location of the cut scores. For Round 3, with the use of the standard-setting tool, the panelists had access to feedback data, benchmark data, and impact data before making their final cut score recommendations.

Table 19: Round 3 Results: ELA: Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 3 Bookmark (Page #)		Range Round 3 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
ELA								
Grade 3	9	12	3	4	57.21	44.51	46	51
Table 1	9	12	1	2	57.21	44.51		
Table 2	10	12	3	3	51.41	44.51		
Grade 4	13	20	4	2	59.74	45.31	45	56
Table 1	13	19	1	1	59.74	45.31		
Table 2	13	21	4	2	59.74	44.73		
Grade 5	11	17	3	3	65.14	51.22	47	58
Table 1	11	18	1	3	65.14	50.00		
Table 2	9	16	2	1	73.78	55.14		
Grade 6	9	16	5	7	65.43	49.77	47	63
Table 1	9	15	1	4	65.43	49.89		
Table 2	8	16	5	7	68.81	49.77		
Grade 7	8	18	1	2	78.56	49.78	49	56
Table 1	9	18	1	1	78.56	49.78		
Table 2	8	17	1	1	78.56	49.78		
Grade 8	11	19	2	5	71.26	48.93	50	64
Table 1	10	21	2	5	74.47	42.82		
Table 2	11	19	1	4	71.26	48.93		

Subject and Grade	Median Round 3 Bookmark (Page #)		Range Round 3 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Grade 10*	13	27	0	1	79.09	49.31	50	70
Table 1	13	27	0	0	79.09	49.31		
Table 2	13	27	0	1	79.09	49.31		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

**Because ILEARN was not administered in grade 10, the grade 10 benchmarking activities used the data from the ILEARN grade 8.*

Table 20: Round 3 Results: Mathematics: Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 3 Bookmark (Page #)		Range Round 3 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Mathematics								
Grade 3	6	10	0	0	71.02	59.21	58	73
Table 1	6	10	0	0	71.02	59.21		
Table 2	6	10	0	0	71.02	59.21		
Grade 4	7	13	3	5	68.50	47.69	53	53
Table 1	7	13	3	5	68.50	47.69		
Table 2	8	13	1	2	63.73	47.69		
Grade 5	6	10	2	2	66.04	47.52	47	57
Table 1	6	10	2	2	66.04	47.52		
Table 2	6	10	1	1	66.04	47.52		
Grade 6	8	12	5	2	65.58	51.35	46	58
Table 1	8	12	5	2	65.58	51.35		
Table 2	9	13	2	1	65.58	46.95		

Subject and Grade	Median Round 3 Bookmark (Page #)		Range Round 3 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Grade 7	8	11	4	2	58.52	46.72	41	68
Table 1	8	11	4	2	58.52	46.72		
Table 2	8	10	2	1	58.52	46.72		
Grade 8	6	14	8	13	55.17	26.67	37	61
Table 1	6	14	8	13	55.17	26.67		
Table 2	6	14	0	0	55.17	26.67		
Grade 10*	10	14	8	8	54.50	42.14	37	57
Table 1	7	14	5	3	58.56	42.14		
Table 2	11	18	4	1	48.76	25.02		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

**Because ILEARN was not administered in grade 10, the grade 10 benchmarking activities used the data from the ILEARN grade 8.*

Table 21: Round 3 Results: Science, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 3 Bookmark (Page #)		Range Round 3 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Science							NA	NA
Grade 4	12	19	18	19	56.69	41.29		
Table 1	15	24	18	16	48.63	31.80		
Table 2	11	17	3	12	56.69	43.17		
Grade 6	11	19	2	7	71.43	48.11		
Table 1	11	19	1	1	71.43	48.11		
Table 2	12	22	1	7	68.11	41.37		
Biology	15	22	2	5	66.84	42.53		
Table 1	15	22	2	2	66.84	42.53		
Table 2	15	22	2	5	66.84	42.53		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

Table 22: Round 3 Results: Social Studies, Median and Range of Recommended Bookmark Locations with Impact and Benchmark Data Included

Subject and Grade	Median Round 3 Bookmark (Page #)		Range Round 3 Bookmark (Page #)		Impact Data 2019 I AM (Percentage At or Above)		Benchmark Data ILEARN (Percentage At or Above)	Benchmark Data NCSC (Percentage At or Above)
	Appr Prof	At Prof	Appr Prof	At Prof	Appr Prof	At Prof	At Prof	At Prof
Social Studies							NA	NA
Grade 5	13	17	12	12	41.30	35.46		
Table 1	13	19	3	7	41.30	30.71		
Table 2	10	17	12	8	48.23	35.46		

Approaching Proficiency (Appr Prof), At Proficiency (At Prof).

3.6.7 Moderation

Performance standards for a statewide system must be coherent across grades and subjects. There should be no irregular peaks and valleys, and performance standards should be orderly across grades with no dramatic differences in expectation unless content rationale is sufficiently defined.

Before any moderation session was considered, staff from both AIR and the IDOE met to determine which performance standards, if any, should be submitted to table leaders for consideration of adjustment to improve articulation. AIR and IDOE staff also decided after which round the articulation activities should be conducted.

Following Round 3, the IDOE and AIR teams determined that moderation activities needed to be conducted for ELA and Mathematics. A short meeting among IDOE staff, AIR staff, and the table leaders was conducted, and it was determined that minor adjustments needed to be made to finalize the articulation of the cross-grade cut scores. Using the standard-setting tool, table leaders had access to feedback, benchmark, and impact data, as well as the OIB and their notes used during all three rounds of the standard setting. They were shown the final results of every other grade in the form of an articulated trend line. The goal was to show the panelists that the standards of the grades should rise along with the grade level at an even rate. The panelists were then given some suggested pages in the OIB where the cut would smooth the progression. The table leaders discussed among themselves, reviewed the data, and made changes only if they agreed and could justify the changes with the content of the item.

Table 23: Moderation Results

Form	Panelist	Mark	Position	Comment
Grade 3 ELA	Table 1 Panel 1	Level 2—Approaching Proficiency	7	We would recommend changing the Approaching Proficiency position to 7 for the following reasons: the answer choices included quotation marks, the answer choices did not have visual supports, the answer choices were in the passage with no obvious distractors.
Grade 7 ELA	Table 1 Panel 1	Level 2—Approaching Proficiency	10	After discussion, item 10 closely matches the content connector of the previous item and provides enough visual support to predict a just barely approaching proficiency student. Item 11 significantly challenges the just barely approaching proficiency student and is above our recommended cut score.
Grade 4 Mathematics	Table 1 Panel 1	Level 3—At Proficiency	12	Based on the content of the questions in numbers 12 and 13, we decided to adjust the recommendation from 13 to 12 because our room consensus was that we were confident that 50% or more students would correctly answer 12, but we were split on number 13. We felt there were distractors in number 13 (visual of the money in the question compared to the visual in the answer options) that would cause less than 50% of students to answer correctly.
Grade 6 Mathematics	Table 1 Panel 1	Level 3—At Proficiency	14	Students at this level should be able to identify the mode. Based on the just barely approaching PLD, students are able to identify the mode at this level. Evidence was provided by the group conversation in the committee earlier in the day when discussing this item. We feel comfortable making this change based on the feedback from the original bookmark placement.
Grade 8 Mathematics	Table 1 Panel 1	Level 3—At Proficiency	10	Due to the vocabulary used in #11, i.e., excessive words instead of a graphic or model, we felt that #10 was the last yes.
Grade 10 Mathematics	Table 2 Panel 1	Level 2—Approaching Proficiency	8	After further discussion, we feel that the content in questions 8 and 9 also align with our assessment of just barely approaching proficiency. For that reason, we believe the cut score should be 8.
Grade 10 Mathematics	Table 2 Panel 1	Level 3—At Proficiency	16	After further discussion, we believe that based on the PLD of this content connector, this question is more appropriate for the cut score. Our discussion about just barely at proficiency is represented by the content in this question. In addition, the previous two questions align more with the approaching proficiency.

3.6.8 Workshop Evaluations

Following completion of standard-setting activities, AIR staff and IDOE staff conducted a debriefing on the standard-setting process and the final cut scores. The debriefing session focused on the outcomes of the workshop. In addition, panelists were encouraged to discuss their satisfaction and level of comfort with the workshop process and with the performance standards they recommended. Panelists and table leaders then completed an Online Workshop Evaluation Form at the end of every day of the meeting.

Panelists independently completed the Online Workshop Evaluation Form, in which they described and assessed their experience taking part in the standard setting using the Bookmark method.

Participants answered questions with one of four different sets of response options. The first was a scale from 1 to 4, with 1 being “Strongly Disagree” and 4 being “Strongly Agree.” The second was a three-option scale with the options “Too Little,” “About Right,” and “Too Much.” The third was a three-option scale with the options “Somewhat Unclear,” “Somewhat Clear,” and “Very Clear.” The fourth was a three-option scale with the options “Not Important,” “Somewhat Important,” and “Very Important.” Workshop participants overwhelmingly indicated clarity in the instructions, materials, data, and process (see Tables 24–29).

Table 24: Day 1, All Rooms

"At the end of the workshop, please rate your agreement with the following statements."	**Percentage of the 3 and 4 Responses**
I understood the purpose of this standard-setting workshop.*	46%
I understand how the Bookmark Method will be used to recommend performance standards.	92%
I understand the role of Performance-Level Descriptors in the standard-setting process.	92%
I understand what is meant by students who are "just barely" described by the Performance-Level Descriptors.	94%
I understand how to review the ordered-item booklet.	94%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each item in the ordered-item booklet.	90%
I feel comfortable expressing my opinions in the workshop.	95%
I feel confident in my ability to recommend performance standards.	95%
"How appropriate was the amount of time spent working through different components of the standard-setting workshop?"	**Percentage of Combined "About Right" and "Too Much" Responses**
Large group introductory training	99%
Taking the online assessment	97%
Review of Performance-Level Descriptors	98%
Development of "Just Barely" Performance-Level Descriptors	91%
Review of ordered-item booklet	92%

Note: Number of responses = 100

**Due to an error on the evaluation form, Strongly Agree was indicated by a 1 and Strongly Disagree was indicated by a 2. Therefore, 46% of panelists rated "I understood the purpose of this standard-setting workshop" with a 1 or a 2.*

Table 25: Day 2, ELA and Mathematics Grades 4, 6, and 8 Rooms

"At the end of day 2,"	**Percentage of the 3 and 4 Responses**
I understand how to use the bookmark method to recommend performance standards.	96%
I understand the role of Performance-Level Descriptors in the standard-setting process.	98%
I understand how to review the ordered-item booklet.	98%
I understand what is meant by response probability (RP50).	94%
I understand how to place my bookmark in the OIB.	98%
I found my group's "Just Barely" PLD helpful in my deliberations.	88%
I found the panel feedback data helpful in my deliberations.	96%
I found the historical data helpful in my deliberations.	86%
I found the performance impact data helpful in my deliberations.	94%
I feel confident in my ability to recommend performance standards.	98%
"How appropriate was the amount of time spent working through different components of the standard-setting workshop?"	**Percentage of Combined "About Right" and "Too Much" Responses**
Review of bookmark procedures	100%
Review of response probability	94%
Review of the bookmark placement process	98%
Round 1 discussion	98%
Round 2 discussion	98%
Round 3 discussion	82%

Note: Number of responses = 50

Table 26: Day 2, ELA and Mathematics Grade 10 and Biology Rooms

"At the end of day 2,"	Percentage of the 3 and 4 Responses
I understood the purpose of this standard-setting workshop.	100%
I understood the concept of placing bookmarks in the OIB.	100%
I found the panel feedback helpful in my deliberations.	100%
I found the historical data helpful in my deliberations.	100%
I found the performance impact data helpful in my deliberations.	100%
I feel confident that my 'At Proficiency' bookmark represents the minimum level of performance of students who are at proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	100%
I feel confident that my 'Approaching Proficiency' bookmark represents the minimum level of performance of students who are approaching proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	100%
"At the end of the workshop,"	**Percentage of the 3 and 4 Responses**
The procedures used to recommend performance standards were fair and unbiased.	100%
The training provided me with the information I needed to recommend performance standards.	100%
The PLDs provided a clear picture of expectations for student performance at each level.	100%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each item.	96%
I found the panelist feedback data and discussions helpful in my decisions about where to place my bookmarks.	100%
I found the benchmark data and discussions helpful in my decisions about where to place my bookmarks.	100%
I found the impact data and discussions helpful in my decisions about where to place my bookmarks.	100%
I felt comfortable expressing my opinions throughout the workshop.	100%
Everyone was given the opportunity to express his or her opinions throughout the workshop.	100%
The procedures used to recommend performance standards were fair and unbiased.	100%

"How appropriate was the amount of time spent working through different components of the standard-setting workshop?"	**Percentage of Combined "About Right" and "Too Much" Responses**
Training on the Bookmark method	100%
Taking the online assessment	100%
Reviewing the ordered-item booklet	93%
Placement of your bookmarks in each round	100%
Review of the ordered-item booklet	89%
Development of "Just Barely" Performance-Level Descriptors	93%
Round 1 discussion	100%
Round 2 discussion	96%
Round 3 discussion	96%
"Please rate the clarity of the following components of the workshop."	**Percentage of the "Very Clear" and "Somewhat Clear"**
Instructions provided by the workshop leader	100%
Ordered-item booklet	100%
Performance-Level Descriptors	100%
"How important was each of the following factors in your placement of the bookmarks?"	**Percentage of the "Very Important" and "Somewhat Important"**
Performance-Level Descriptors	96%
Your perception of the difficulty of the items	96%
Your experiences with students	89%
Small group discussions	100%
Large group discussions	100%
Feedback data	100%
Impact data	96%

Standard Setting Outcome	Percentage of the 3 and 4 Responses
I am confident that students classified as At Proficiency are proficient in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	96%
I am confident that students classified as Approaching Proficiency the Standard are fairly classified as approaching proficiency in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	100%

Note: Number of responses = 27

Table 27: Day 2, Social Studies Grade 5

"At the end of day 2,"	Percentage of the 3 and 4 Responses
I understood the purpose of this standard-setting workshop.	100%
I understood the concept of placing bookmarks in the OIB.*	100%
I understood the concept of placing bookmarks in the OIB.*	100%
I found the historical data helpful in my deliberations.	83%
I found the performance impact data helpful in my deliberations.	83%
I feel confident that my 'At Proficiency' bookmark represents the minimum level of performance of students who are at proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	100%
I feel confident that my 'Approaching Proficiency' bookmark represents the minimum level of performance of students who are approaching proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	100%
"At the end of the workshop,"	**Percentage of the 3 and 4 Responses**
The procedures used to recommend performance standards were fair and unbiased.	100%
The training provided me with the information I needed to recommend performance standards.	100%
The PLDs provided a clear picture of expectations for student performance at each level.	100%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each item.	100%

I found the panelist feedback data and discussions helpful in my decisions about where to place my bookmarks.	100%
I found the benchmark data and discussions helpful in my decisions about where to place my bookmarks.	100%
I found the impact data and discussions helpful in my decisions about where to place my bookmarks.	100%
I felt comfortable expressing my opinions throughout the workshop.	100%
Everyone was given the opportunity to express his or her opinions throughout the workshop.	100%
The procedures used to recommend performance standards were fair and unbiased.	100%
"How appropriate was the amount of time spent working through different components of the standard-setting workshop?"	**Percentage of Combined "About Right" and "Too Much" Responses**
Training on the Bookmark method	100%
Taking the online assessment	100%
Reviewing the ordered-item booklet (OIB)**	100%
Placement of your bookmarks in each round	100%
Review of ordered-item booklet (OIB)**	100%
Development of "Just Barely" PLDs	100%
Round 1 discussion	100%
Round 2 discussion	100%
Round 3 discussion	100%
"Please rate the clarity of the following components of the workshop."	**Percentage of the "Very Clear" and "Somewhat Clear"**
Instructions provided by the workshop leader	100%
Ordered-item booklet	100%
Performance-Level Descriptors (PLDs)	100%

"How important was each of the following factors in your placement of the bookmarks?"	Percentage of the "Very Important" and "Somewhat Important"
Performance-Level Descriptors (PLDs)	100%
Your perception of the difficulty of the items	100%
Your experiences with students	100%
Small group discussions	100%
Large group discussions	100%
Feedback data	100%
Impact data	100%
Standard Setting Outcome	**Percentage of the 3 and 4 Responses**
I am confident that students classified as At Proficiency are proficient in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	100%
I am confident that students classified as Approaching Proficiency the Standard are fairly classified as approaching proficiency in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	100%

Note: Number of responses = 6

**Question was repeated on evaluation.*

***Question was repeated on evaluation.*

Table 28: Day 2, Science Grades 4 and 6

"At the end of day 2,"	**Percentage of the 3 and 4 Responses**
I understood the purpose of this standard-setting workshop.	100%
I understood the concept of placing bookmarks in the OIB.*	94%
I understood the concept of placing bookmarks in the OIB.*	94%
I found the historical data helpful in my deliberations.	75%
I found the performance impact data helpful in my deliberations.	94%
I feel confident that my 'At Proficiency' bookmark represents the minimum level of performance of students who are at proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	94%
I feel confident that my 'Approaching Proficiency' bookmark represents the minimum level of performance of students who are approaching proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	94%
"At the end of the workshop,"	**Percentage of the 3 and 4 Responses**
The procedures used to recommend performance standards were fair and unbiased.	94%
The training provided me with the information I needed to recommend performance standards.	100%
The PLDs provided a clear picture of expectations for student performance at each level.	81%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each item.	100%
I found the panelist feedback data and discussions helpful in my decisions about where to place my bookmarks.	94%
I found the benchmark data and discussions helpful in my decisions about where to place my bookmarks.	94%
I found the impact data and discussions helpful in my decisions about where to place my bookmarks.	94%
I felt comfortable expressing my opinions throughout the workshop.	94%
Everyone was given the opportunity to express his or her opinions throughout the workshop.	100%
The procedures used to recommend performance standards were fair and unbiased.	100%

"How appropriate was the amount of time spent working through different components of the standard-setting workshop?"	**Percentage of Combined "About Right" and "Too Much" Responses**
Training on the Bookmark method	100%
Taking the online assessment	100%
Reviewing the ordered-item booklet (OIB)**	88%
Placement of your bookmarks in each round	100%
Review of ordered-item booklet (OIB)**	94%
Development of "Just Barely" Performance-Level Descriptors	100%
Round 1 discussion	81%
Round 2 discussion	88%
Round 3 discussion	75%
"Please rate the clarity of the following components of the workshop."	**Percentage of the "Very Clear" and "Somewhat Clear"**
Instructions provided by the workshop leader	100%
Ordered-item booklet (OIB)	94%
Performance-Level Descriptors (PLDs)	100%
"How important was each of the following factors in your placement of the bookmarks?"	**Percentage of the "Very Important" and "Somewhat Important"**
Performance-Level Descriptors (PLDs)	100%
Your perception of the difficulty of the items	100%
Your experiences with students	100%
Small group discussions	100%
Large group discussions	100%
Impact data	100%

Standard Setting Outcome	Percentage of the 3 and 4 responses
I am confident that students classified as At Proficiency are proficient in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	88%
I am confident that students classified as Approaching proficiency the Standard are fairly classified as approaching proficiency in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	88%

Note: Number of responses = 16

**Question was repeated on evaluation.*

***Question was repeated on evaluation.*

Table 29: Day 3, ELA and Mathematics Adjacent Grades

"At the End of Day 3,"	Percentage of the 3 and 4 Responses
I understand how to review the ordered-item booklet (OIB).	98%
I understood the concept of vertical moderation.	98%
I understood the concept of extrapolating and interpolating standards for an adjacent grade.	96%
I understood the concept of placing bookmarks in the adjacent grade OIB.	98%
I found the panel feedback helpful in my deliberations.	98%
I found the historical data helpful in my deliberations.	94%
I found the performance impact data helpful in my deliberations.	98%
I feel confident that my 'At Proficiency' bookmark represents the minimum level of performance of students who are at proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	98%
I feel confident that my 'Approaching Proficiency' bookmark represents the minimum level of performance of students who are approaching proficiency in the knowledge and skills described in the Indiana Alternate Standards and Content Connectors.	98%
I feel confident in my ability to recommend performance standards.	98%

"At the end of the workshop,"	Percentage of the 3 and 4 Responses
I understood the purpose of this standard-setting workshop.	98%
The procedures used to recommend performance standards were fair and unbiased.	96%
The training provided me with the information I needed to recommend performance standards.	98%
The PLDs provided a clear picture of expectations for student performance at each level.	94%
Taking the online assessment helped me to better understand what students need to know and be able to do to answer each item.	98%
I found the panelist feedback data and discussion helpful in my decisions about where to place my bookmarks.	98%
I found the benchmark data and discussions helpful in my decisions about where to place my bookmarks.	96%
I found the impact data and discussions helpful in my decisions about where to place my bookmarks.	96%
I felt comfortable expressing my opinions throughout the workshop.	96%
Everyone was given the opportunity to express his or her opinions throughout the workshop.	96%
"How appropriate was the amount of time spent working through different components of the standard-setting workshop?"	**Percentage of Combined "About Right" and "Too Much" Responses**
Training on the Bookmark method	100%
Taking the online assessment	98%
Reviewing the ordered-item booklet (OIB)*	100%
Placement of your bookmarks in each round	100%
Review of ordered-item booklet (OIB)*	98%
Development of "Just Barely" Performance-Level Descriptors	100%
Round 1 discussion	100%
Round 2 discussion	98%
Round 3 discussion	100%

Please rate the clarity of the following components of the workshop.	Percentage of the "Very Clear" and "Somewhat Clear"
Instructions provided by the workshop leader	100%
Ordered-item booklet (OIB)	100%
Performance-Level Descriptors (PLDs)	100%
How important was each of the following factors in your placement of the bookmarks?	**Percentage of the "Very Important" and "Somewhat Important"**
Performance-Level Descriptors (PLDs)	100%
Your perception of the difficulty of the items	98%
Your experiences with students	96%
Small group discussions	100%
Large group discussions	100%
Feedback data	100%
Impact data	100%
Standard Setting Outcome	**Percentage of the 3 and 4 responses**
I am confident that students classified as At Proficiency are proficient in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	96%
I am confident that students classified as Approaching Proficiency the Standard are fairly classified as approaching proficiency in the knowledge and skills described the Indiana Alternate Standards and Content Connectors.	96%

Note: Number of responses = 50

**Question was repeated on evaluation.*

4. VALIDITY EVIDENCE

Validity evidence for standard setting is established in multiple ways. Validity evidence for standard setting is established in the professional standards developed by appropriate professional organizations and best practice in the literature and established validity criteria. The standards guide the evidence required of states to meet federal peer critical elements relevant to standard setting. In the following sections, we describe our adherence to these same professional standards

4.1 EVIDENCE OF ADHERENCE TO PROFESSIONAL STANDARDS AND BEST PRACTICES

The I AM standard-setting workshop was designed and implemented in a way consistent with the established practices and best-practice principles (Hambleton & Pitoniak, 2006; Hambleton, Pitoniak, & Copella, 2012; Kane, 2001). The workshop also adhered to the following professional standards recommended by the *Standards for educational and psychological testing* (American Educational Research Association

[AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) related to standard setting:

> Standard 5.21: When proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
>
> Standard 5.22: When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.
>
> Standard 5.23: When feasible and appropriate, cut scores defining categories and distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

The sections of this report documenting the rationale and procedures used in the standard-setting workshop address Standard 5.21. The Bookmark standard-setting procedure is appropriate for assessments with multiple item formats and scaled using IRT. Section 3.1 provides the justification for, and the additional benefits of, selecting the Bookmark method to establish the cut scores. Sections 3.6.1–3.6.8 document the process followed to implement the Bookmark method.

The design and implementation of the Bookmark procedure addresses Standard 5.22. The method directly leverages the subject matter expertise of the panelists placing the bookmarks and incorporates multiple, iterative rounds of ratings in which panelists modify their judgments based on feedback and discussion. Panelists apply their expertise in multiple ways throughout the process, including

- understanding the assessment and assessment items (from an educator and student perspective);

- describing the content measured by the assessment as described by the content standards;
- identifying the skills associated with each assessment item;
- describing the skills associated with "Just Barely" students for each performance level;
- selecting which assessment items students in each performance level should be able to answer correctly;
- evaluating and applying feedback and reference data to their Round 2 bookmarks; and
- considering the impact of the recommended cut scores on students.

Additionally, panelists' readiness evaluations provided evidence of a successful orientation to the process and understanding of the Bookmark procedure, while their workshop evaluations provide evidence of confidence in the process and resulting recommendations.

The recruitment process resulted in panels which were representative of important regional and demographic groups, and who were knowledgeable about the subject area and students' developmental level. Section 3.3.5, Educator Participants, and Tables 6 and 7 summarize details about the panel demographics and qualifications.

The provision of benchmark and impact data to panelists after Round 1 addresses Standard 5.23. This empirical data provides necessary and additional context describing student performance given the recommended standards.

4.2 Evidence in Terms of Peer Review Critical Elements

The United States Department of Education (USDOE) provides guidance for the peer review of state assessment systems. This guidance is intended to support states in meeting statutory and regulatory requirements under Title I of the Elementary and Secondary Education Act of 1965 (ESEA) (USDOE, 2015). The following two critical elements are relevant to standard setting, with evidence supporting each element immediately following:

> Critical Element 1.2: Substantive involvement and input of educators and subject-matter experts

Indiana educators played a critical role in establishing performance levels for the I AM. They reviewed and accepted the range PLDs, drafted and applied Target PLDs to delineate performance at each performance level, considered benchmark data and the impact of their recommendations, and formally recommended achievement standards.

Many subject-matter experts contributed to developing Indiana's performance standards. Contributing educators were subject-matter experts in their content area, the content standards, the curriculum that they teach, and in the developmental and cognitive capabilities of their students. AIR's facilitators were subject-matter experts in the subjects

tested, alternate assessments, and in facilitating effective standard-setting workshops. The psychometricians performing the analyses and calculations throughout the meeting were subject-matter experts in the measurement and statistics principles required of the standard-setting process. Finally, Dr. Phillips is a nationally known expert in assessment and measurement, including multiple methods of standard setting.

> Critical Element 6.2: Achievement standards setting. The State used a technically sound method and process that involved panelists with appropriate experience and expertise for setting its academic achievement standards and alternate academic achievement standards to ensure they are valid and reliable.

Evidence to support this critical element includes the following:

- The rationale for and technical sufficiency of the Bookmark method selected to establish performance standards (Section 3.1, The Bookmark Method)
- Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores (Section 3.6, Events and 4.1: Evidence of Adherence to Professional Standards and Best Practices)
- Panelists self-reported readiness to undertake the task (Section 3.6.6.4, Bookmark Placement Readiness Form) and confidence in the workshop process and outcomes (Section 3.6.8, Workshop Evaluations) supporting the validity of the process.
- The standard-setting panels consisted of panelists with appropriate experience and expertise, including content experts with experience teaching the Indiana Academic Content Standards in the tested grades and subjects, and individuals with experience and expertise teaching special and general education students in Indiana (Section 3.3.5, Educator Participants).

5. REFERENCES

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

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Cizek, G. J., & Koons, H. (2014). *Observation and report on Smarter Balanced standard setting: October 12–20, 2014.* Accessed from https://portal.smarterbalanced.org/library/en/standard-setting-observation-and-report.pdf.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp.79–106). New York, NY: Routledge.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). New York, NY: Routledge.

Huynh, H. (2006). A clarification on the response probability criterion RP50 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice*, 25: 19–20.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Karantonis, A. & Sireci, S. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*. 25. 4–12.

Lewis, D., & Lord-Bessen, J. (2017). Standard setting. *Handbook of Item Response Theory, Volume Three: Applications,* 229.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations (2nd Edition)* (pp. 225–253). New York, NY: Routledge.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

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: Erlbaum.

Perie, M. (2005, April). Angoff and Bookmark methods. Workshop presented at the annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

U. S. Department of Education, (2015). *Non-regulatory guidance for states for meeting requirements of the elementary and secondary education act of 1965, as amended.* Washington, DC. Accessed from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.

Williams, N. J., & Schulz, E. M. (2005, April). *An investigation of response probability (RP) values used in standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.