



***IREAD-3***

**Indiana Reading Evaluation  
and Determination**

**2021–2022**

**Volume 1  
Annual Technical Report**

## **ACKNOWLEDGMENTS**

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

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

## TABLE OF CONTENTS

1. INTRODUCTION.....	1
1.1 BACKGROUND AND HISTORICAL CONTEXT .....	1
1.2 PURPOSE AND INTENDED USES OF THE IREAD-3 ASSESSMENT.....	2
1.3 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF IREAD-3 .....	3
1.4 AVAILABLE TEST FORMATS AND SPECIAL VERSIONS .....	3
1.5 STUDENT PARTICIPATION .....	4
2. SUMMARY OF OPERATIONAL PROCEDURES .....	6
2.1 ADMINISTRATION PROCEDURES.....	6
2.2 DESIGNATED SUPPORTS AND ACCOMMODATIONS .....	6
3. ITEM BANK AND TEST CONSTRUCTION.....	8
3.1 OVERVIEW OF ITEM DEVELOPMENT .....	8
3.2 OPERATIONAL FORM CONSTRUCTION.....	8
4. CLASSICAL ANALYSES OVERVIEW .....	9
4.1 CLASSICAL ITEM ANALYSES.....	9
4.1.1 <i>Item Discrimination</i> .....	9
4.1.2 <i>Distractor Analysis</i> .....	10
4.1.3 <i>Item Difficulty</i> .....	10
4.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS .....	10
4.3 CLASSICAL ANALYSES RESULTS.....	13
5. ITEM RESPONSE THEORY, ITEM CALIBRATION, AND EQUATING.....	14
5.1 IRT MODELS .....	14
5.2 IRT SUMMARIES .....	15
6. SCORING AND REPORTING .....	16
6.1 MAXIMUM LIKELIHOOD ESTIMATION .....	16
6.1.1 <i>Likelihood Function</i> .....	16
6.1.2 <i>Derivatives</i> .....	16
6.1.3 <i>Extreme Case Handling</i> .....	17
6.1.4 <i>Standard Errors of Estimates</i> .....	18
6.2 TRANSFORMING THETA SCORES TO REPORTING SCALE SCORES .....	18
6.3 OVERALL PERFORMANCE CLASSIFICATION .....	19
6.4 REPORTING CATEGORY SCORES.....	19
6.5 LEXILE® SCORES .....	20
6.6 COMPARISON OF SCORES TO PREVIOUS YEAR .....	20

7. QUALITY ASSURANCE PROCEDURES ..... 21

7.1 QUALITY ASSURANCE IN TEST CONFIGURATION ..... 21

7.2 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION ..... 21

7.2.1 *Production of Content*..... 21

7.2.2 *Web Approval of Content During Development*..... 22

7.2.3 *Platform Review*..... 22

7.2.4 *User Acceptance Testing and Final Review*..... 23

7.2.5 *Functionality and Configuration*..... 24

7.3 QUALITY ASSURANCE IN DATA PREPARATION ..... 25

7.4 QUALITY ASSURANCE IN ITEM ANALYSIS AND EQUATING ..... 26

7.5 QUALITY ASSURANCE IN SCORING AND REPORTING ..... 26

7.5.1 *Quality Assurance in Test Scoring*..... 26

7.5.2 *Quality Assurance in Reporting*..... 28

8. REFERENCES..... 29

## LIST OF TABLES

Table 1: Required Uses and Citations of IREAD-3 .....	2
Table 2: Number of Students Participating in IREAD-3 2021–2022 .....	5
Table 3: Distribution of Demographic Characteristics of Tested Population .....	5
Table 4: IREAD-3 Items by Type .....	8
Table 5: Thresholds for Flagging Items in Classical Item Analysis .....	9
Table 6: DIF Classification Rules.....	13
Table 7: Operational Item p-Value Five-Point Summary and Range .....	13
Table 8: Operational Item Parameter Five-Point Summary and Range, Spring 2022 ...	15
Table 9: Operational Item Parameter Five-Point Summary and Range, Summer 2022	15
Table 10: Theta and Scaled-Score Limits for Extreme Ability Estimates .....	18
Table 11: Scaling Constants on the Reporting Metric .....	19
Table 12: Proficiency Levels .....	19
Table 13: Proficiency Levels for Grade 2 Students.....	19
Table 14: Overview of Quality Assurance Reports .....	27

## LIST OF APPENDICES

Appendix A: Operational Item Statistics
Appendix B: Test Characteristic Curves
Appendix C: Distribution of Scale Scores and Standard Deviations
Appendix D: Distribution of Reporting Category Scores

## 1. INTRODUCTION

The Indiana Reading Evaluation and Determination (IREAD-3) 2021–2022 Technical Report is provided to document and make transparent all methods used in item development, test construction, psychometric methods, standard setting, score reporting methods, summarizing student assessment results, and providing supporting evidence for intended uses and interpretations of the test scores. The technical report is presented as five separate, self-contained volumes that cover the following topics:

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

The Indiana Department of Education (IDOE) communicates the quality of the IREAD-3 assessments by making these technical reports accessible to the public. Not all volumes are produced annually, and some volumes have only minor updates between years.

### 1.1 BACKGROUND AND HISTORICAL CONTEXT

IREAD-3 was first administered to students during the spring of 2012 in accordance with House Enrolled Act 1367. The IREAD-3 assessment was constructed to measure foundational reading standards through grade 3. In 2014, the new Indiana Academic Standards (IAS) in English/Language Arts (ELA) were adopted for IREAD-3. IREAD-3 assessments do not measure all the IAS for ELA, but rather the standards most relevant to foundational reading proficiency.

In June 2017, IDOE commissioned an independent alignment evaluation of the 2017 forms through edCount, Indiana’s vendor for the IREAD-3 assessment study. The purpose of the study was to review supporting documentation for the assessment, including an analysis of the relationship between the content assessed by the test and the underlying construct it is supposed to measure. The study’s outcome determined that the items aligned to the standards and the forms aligned to the blueprint.

Starting Spring 2022, IDOE provides Indiana schools the option to administer the IREAD-3 assessment to students in grade 2. This allows students and educators to receive information earlier for students who need additional support to learn to read. IDOE, with the guidance and support from the Technical Advisory Committee (TAC), analyzed the current IREAD-3 assessment to ensure it would serve this new purpose. To provide schools with an early indicator regarding students’ foundational reading skillset grade 2, a new cut score was developed to indicate when a grade 2 student is “on track” to reach

proficiency in grade 3. Refer to Volume 6 of this technical report for a summary of the Grade 2 Policy Content Setting for IREAD-3 that took place in July 2022. This new “On Track” cut was announced to schools on August 1, 2022. Schools who opted in to administer IREAD-3 to grade 2 students communicated individual student results and instructional responses based on those results to families. It is IDOE’s plan to conduct validation research for the “On Track” cut using Spring 2023 scores. Upon completion, the “On Track” cut, along with the existing “Pass” cut, will be incorporated into the score reports starting Spring 2023. Grade 2 students who achieved the “Pass” cut score will be exempted from the IREAD-3 assessment in grade 3. The grade 2 opt-in option is not available in Summer retest administrations.

## 1.2 PURPOSE AND INTENDED USES OF THE IREAD-3 ASSESSMENT

IREAD-3 is a standards-referenced assessment that applies principles of evidence-centered design to yield overall and reporting-category-level test scores at the student level and other levels of aggregation that reflect student proficiency in foundational reading skills as defined in the IAS. IREAD-3 supports instruction and student learning by providing immediate feedback to educators and parents that can be used to inform instructional strategies that remediate or enrich instruction. IREAD-3 also supports instruction by providing aggregate data on a larger scale to consider effectiveness of curriculum, instructional programming, and current educational strategies. An array of reporting metrics allows achievement to be monitored at both the student and aggregate levels.

The IREAD-3 assessment draws items from an existing item bank (see Volume 2). IREAD-3 items measure knowledge and skills to ensure students can read proficiently before moving on to grade 4. Items on the test forms were constructed to uniquely measure students’ reading skills on the IAS in ELA. Cambium Assessment, Inc. (CAI) inherited the IREAD-3 item bank from Indiana’s previous testing vendor and did not perform any new item development.

Table 1: Required Uses and Citations of IREAD-3 outlines the required uses and citations of IREAD-3.

*Table 1: Required Uses and Citations of IREAD-3*

Required Use	Required Use Citation
House Enrolled Act (HEA) 1367, also known as Public Law 109 in 2010, requires the evaluation of reading skills for students who are in third grade beginning in the spring of 2012. This legislation was created to ensure that all students can read proficiently at the end of grade three. In response to HEA 1367, educators from across the state worked with the Indiana Department of Education to develop a test blueprint and to review test questions that have now become the Indiana Reading Evaluation and Determination (IREAD-3) Assessment. The intent of HEA 1367 is to ensure every student has the opportunity for future success through literacy. The results will have a positive effect on our entire state as the need for remedial education in middle and high school is reduced and dropout rates and juvenile delinquency are lowered.	House Enrolled Act 1367, Public Law 109

### **1.3 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF IREAD-3**

IDOE manages the IREAD-3 assessment program with the assistance of Indiana educators, the Indiana State Board of Education Technical Advisory Committee (TAC), and several vendors (listed below). IDOE fulfills the diverse requirements of implementing IREAD-3 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 IREAD-3 program, including coordination with other IDOE offices, accredited Indiana public and non-public schools, and vendors.

#### **Indiana Educators**

Indiana educators participated in most aspects of the conceptualization and development of IREAD-3. Educators participated in the development of the IAS, clarification of how these standards will be assessed, creation of the blueprint and test design, standard setting to determine the IREAD-3 cut score, and committee reviews of test items and passages.

#### **Technical Advisory Committee**

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

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

Cambium Assessment, Inc. (CAI) is the current vendor selected through the state-mandated competitive procurement process. In the winter of 2017, CAI became the primary party responsible for building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting assessment results for IREAD-3 as described in this report.

#### **Human Resources Research Organization**

For the 2021–2022 IREAD-3 assessment, the Human Resources Research Organization (HumRRO) conducted independent verifications of scoring activities.

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

IREAD-3 was administered as an online, fixed-form assessment. Students unable to participate in the online administration had the option to use a paper-and-pencil form. Students participating in the computer-based IREAD-3 could use standard online testing

features in the Test Delivery System (TDS), which included a selection of font colors and sizes and the ability to zoom in and out and highlight text.

Students with disabilities could take the IREAD-3 with or without accommodations. Additionally, a separate form was administered to hard-of-hearing students, and a braille test form was available for students with visual impairments. More details about accommodations can be found in Volume 3.

## **1.5 STUDENT PARTICIPATION**

In Spring 2022, all accredited Indiana public and non-public school students in grade 3, students from schools who opted to participate in grade 2 testing, as well as students in grades 4 and 5 who had not passed the assessment previously, took the IREAD-3 assessment. Students who did not pass the assessment during the previous administration could retest in Summer 2022, except grade 2 students. Students granted a Good Cause Exemption (GCE) did not participate in the retest administration.

The purpose of granting a GCE is to exempt a student who does not pass IREAD-3 from having to participate in future IREAD-3 testing. A GCE does not impact a student's IREAD-3 score or passing status, nor does it remove the student's score from a school's percentage passing calculation. Students who have previously been retained two times prior to promotion to grade 4, students with Individualized Education Programs (IEPs), and English learners with Individual Learning Plans (ILPs) are eligible for GCEs. Table 2 shows the number of students assessed and the number of students reported for IREAD-3 by administration. The number of tested and reported students are not equal because of cases where tests were not complete or were invalidated for issues such as a student having to switch accommodation type after a testing session had begun. A maximum of only one score is reported for a student even if more than one test session was attempted. Table 3 presents the distribution of students by counts and percentages by administration. The subgroup categories reported include gender, ethnicity, students classified as special education (SPED), English learners, and Section 504 Plan status.

*Table 2: Number of Students Participating in IREAD-3 2021–2022*

Admin	Grade	Number Tested	Number Reported
Spring 2022	2	20,392	20,199
Spring 2022	3	85,485	85,212
Summer 2022	3	16,395	16,265

*Table 3: Distribution of Demographic Characteristics of Tested Population*

Admin	Grade	Group	All Students	Male	Female	White	Black/ African American	Asian	Hispanic	American Indian/ Alaska Native	Native Hawaiian/ Other Pacific Islander	Multiracial/ Two or More Races	Special Education	English Learner	Section 504 Plan
Spring 2022	2	N	20,392	10,458	9,934	13,892	2,426	234	2,718	33	16	1,073	3,402	1,734	170
		%	100	51.28	48.72	68.12	11.9	1.15	13.33	0.16	0.08	5.26	16.68	8.5	0.83
Spring 2022	3	N	85,485	43,652	41,833	53,942	12,208	2,533	11,735	132	79	4,856	14,348	8,365	1,688
		%	100	51.06	48.94	63.1	14.28	2.96	13.73	0.15	0.09	5.68	16.78	9.79	1.97
Summer 2022	3	N	16,395	8,719	7,676	7,702	4,177	281	3,152	31	21	1,031	4,563	2,360	375
		%	100	53.18	46.82	46.98	25.48	1.71	19.23	0.19	0.13	6.29	27.83	14.39	2.29

## **2. SUMMARY OF OPERATIONAL PROCEDURES**

### **2.1 ADMINISTRATION PROCEDURES**

The Indiana Reading Evaluation and Determination (IREAD-3) assessment for Spring 2022 was administered to eligible students from March 7 through 18, 2022. The Summer 2022 assessment was available May 23 through July 15, 2022, to students who did not pass the Spring 2022 administration.

The key personnel involved with the IREAD-3 administration included the Corporation Test Coordinators (CTCs), Co-Op role (Co-Op), Non-Public School Test Coordinators (NPSTCs), School Test Coordinators (STCs), and Test Administrators (TAs). Test administration manuals (TAMs) were provided so that personnel involved with statewide assessment administrations could maintain both standardized administration conditions and test security.

A secure browser developed by Cambium Assessment, Inc. (CAI) was required to access the online IREAD-3 assessments. The online browser provided a secure environment for student testing by disabling the hot keys, copy, and screen-capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines).

### **2.2 DESIGNATED SUPPORTS AND ACCOMMODATIONS**

Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., Text-to-Speech) are provided digitally through instructional or assessment technology, while non-embedded designated features (e.g., scribe) are non-digital or provided through online supports not contained within the Test Delivery System (TDS). Standard accommodations are made available to all students to use as needed. Non-standard accommodations are generally available for students for whom there is a documented need in an Individualized Education Program (IEP), Section 504 Plan, or Individual Learning Plan (ILP).

State-approved non-standard accommodations do not compromise learning expectations, constructs, or grade-level standards. Such accommodations help generate valid outcomes of the assessments so that students receiving an accommodation can fully demonstrate what they know and are able to do. Psychometrically, 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).

Accessibility supports, both standard and non-standard, discussed in this document include embedded and non-embedded features. Standard accommodations were supports that were universally available to all students as they accessed instructional or assessment content. There were also designated features that were available to students for whom an informed educator or team of educators had identified the need for non-

standard accommodations. Educators making those decisions were trained on accessibility guidelines and understood the range of designated supports available for students with an IEP, Section 504 Plan, or ILP. Indiana STCs and TAs were responsible for ensuring that arrangements for accommodations were made before the test administration dates. The available accommodation options for eligible students included braille, streamline, assistive technology (e.g., adaptive keyboards, touch screen, switches), and scribe. Accommodations were assigned to students through the Test Information Distribution Engine (TIDE) and, in the case of non-standard accommodations, were approved by the Indiana Department of Education (IDOE) before being applied to the TDS testing interface.

### 3. ITEM BANK AND TEST CONSTRUCTION

#### 3.1 OVERVIEW OF ITEM DEVELOPMENT

Operational items used on the Indiana Reading Evaluation and Determination (IREAD-3) test forms were drawn from the previously established IREAD-3 item bank. Volume 2 of this technical report contains details on the IREAD-3 item bank.

#### 3.2 OPERATIONAL FORM CONSTRUCTION

Operational test forms (refer to Volume 2) include multiple-choice (MC) item types to measure the Indiana Academic Standards (IAS). Table 4 briefly describes the item types used and the number of items by item type. A more detailed description and examples for each of the item types are also provided in Appendix B of Volume 2 of this technical report.

Previously developed fixed forms built by Indiana’s prior vendor were used for all test administrations. Tests were pre-equated using previously established item parameters.

*Table 4: IREAD-3 Items by Type*

<b>Response Type</b>	<b>Description</b>	<b>Spring 2022</b>	<b>Summer 2022</b>
MC	Student selects one correct answer from a number of options.	38	38

## 4. CLASSICAL ANALYSES OVERVIEW

### 4.1 CLASSICAL ITEM ANALYSES

Cambium Assessment, Inc. (CAI) psychometricians monitor the behavior of items while test forms are administered in a live environment. This is accomplished using CAI's Quality Monitor (QM) system, which yields an item-analysis report on the performance of test items throughout the testing window. During the administration of the 2021–2022 Indiana Reading Evaluation and Determination (IREAD-3) assessment, this system served as a key check for the early detection of potential problems with item scoring, including the incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that could 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 can 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 5, and a description of the statistics is provided in the following paragraphs.

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

Analysis Type	Flagging Criteria
Item Discrimination	Adjusted biserial/polyserial correlation statistic is less than 0.25 for multiple-choice (MC) items.
Distractor Analysis	Adjusted biserial correlation statistic is greater than 0.00 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 0.25 or greater than 0.95 for MC items.

#### 4.1.1 ITEM DISCRIMINATION

The item discrimination index indicates the extent to which each item differentiates between test takers who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for MC items was calculated as the correlation between the item score and the ability estimate for students. Point biserial correlations and the number of flagged items for operational items can be found in Appendix A, Operational Item Statistics. All operational items had a higher point biserial correlation than the flagging criteria. No IREAD-3 operational items were flagged for item discrimination.

### 4.1.2 DISTRACTOR ANALYSIS

Distractor analysis for MC items is used to identify items that may have marginal distractors, ambiguous correct responses, an incorrect key, or more than one correct answer that attracts high-scoring students. For MC items, the correct response should have been the most frequently selected option by high-scoring students. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative. No IREAD-3 operational items were flagged for distractor analysis.

### 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. Conventional item  $p$ -values are summarized in Section 4.3, Classical Analyses Results. The  $p$ -values and number of flagged items for operational items can be found in Appendix A, Operational Item Statistics. Operational items had  $p$ -values within the expected range.

## 4.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

Note that differential item functioning (DIF) summaries are provided only when field-test analyses occur. No items were field tested during the 2021–2022 school year, and thus no DIF summaries appear in this year’s technical report.

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

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

- Male/Female
- White/African American
- White/Hispanic

DIF refers to items that appear to function differently across identifiable groups, typically different demographic groups. Identifying DIF is important because it provides a statistical

indicator that an item may contain cultural or other 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 on other types of items. In this example, it is the instruction, not the item, that exhibits bias. However, DIF can indicate bias, so all items were evaluated for DIF. Items flagged for DIF 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.

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

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

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

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

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

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

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

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

The MH conditional odds ratio is calculated as

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

Table 6: DIF Classification Rules

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

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

### 4.3 CLASSICAL ANALYSES RESULTS

This section presents a summary of results from the classical item analysis for the 2021–2022 IREAD-3 operational items. The summaries here are aggregates; item-specific details can be found in Appendix A, Operational Item Statistics.

Table 7 provides summaries of the  $p$ -values by percentile and range by administration for operational items. Indiana students' performance indicates the desired variability across the scale. The variability informs us that the constructed operational forms had a good discrimination for Indiana students.

Table 7: Operational Item  $p$ -Value Five-Point Summary and Range

Administration	Grade	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Spring 2022	2	0.29	0.33	0.51	0.59	0.73	0.84	0.90
Spring 2022	3	0.40	0.54	0.71	0.76	0.87	0.93	0.96
Summer 2022	3	0.35	0.36	0.5	0.59	0.67	0.79	0.81

## 5. ITEM RESPONSE THEORY, ITEM CALIBRATION, AND EQUATING

Item Response Theory (IRT) (van der Linden & Hambleton, 1997) is used to calibrate all items and derive scores for all Indiana Reading Evaluation and Determination (IREAD-3) items and assessments. IRT is a general framework that models test responses resulting from an interaction between students and test items. IRT encompasses many related measurement models that allow for varied assumptions about the nature of the data. Simple unidimensional models are the most common models used in K–12 operational assessment programs. In some instances, item dependencies exist, and more complex models are employed.

Cambium Assessment, Inc. (CAI) used previously established item parameters to score the IREAD-3 assessments in Spring 2022 and Summer 2022.

### 5.1 IRT MODELS

IREAD-3 employed IRT models for item calibration and student ability estimation. The IREAD-3 assessment is made up of multiple-choice (MC) items and two-point composite items. All MC items will use the three-parameter logistic (3PL) model. All polytomous items will use the generalized partial credit model.

#### **Three-Parameter Logistic Model**

In the case of the 3PL, we have:

$$p_{ij}(z_{ij}|\theta_j, a_i, b_i, \dots, b_i, c_i, \dots, c_i) = \left\{ \begin{array}{l} c_i + (1 - c_i) \frac{\exp(1.7 * a_i(\theta_j - b_i))}{1 + \exp(1.7 * a_i(\theta_j - b_i))} = p_{ij}, \text{ if } z_{ij} = 1 \\ \frac{1 - c_i}{1 + \exp(1.7 * a_i(\theta_j - b_i))} = 1 - p_{ij}, \text{ if } z_{ij} = 0 \end{array} \right\},$$

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

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

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

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

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

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

## 5.2 IRT SUMMARIES

The statistical summaries of the pre-equated operational item parameters used to score the Spring and Summer administrations can be found in Table 8 and Table 9.

*Table 8: Operational Item Parameter Five-Point Summary and Range, Spring 2022*

Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
a	0.39	0.43	0.88	1.07	1.46	1.70	1.92
b	-3.89	-3.08	-1.70	-1.26	-0.86	-0.15	0.51
c	-2.77	0.15	0.07	0.15	0.20	0.35	0.42

*Table 9: Operational Item Parameter Five-Point Summary and Range, Summer 2022*

Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
a	0.39	0.49	0.78	1.03	1.37	1.59	1.76
b	-3.02	-2.45	-1.61	-1.24	-0.81	-0.24	-0.18
c	0.01	0.02	0.10	0.15	0.21	0.28	0.28

Another way to view the technical properties of IREAD-3 test forms is via test characteristic curves (TCCs). These plots are displayed in Appendix B, Test Characteristic Curves.

## 6. SCORING AND REPORTING

### 6.1 MAXIMUM LIKELIHOOD ESTIMATION

Cambium Assessment, Inc. (CAI) generated ability estimates using pattern scoring, a method that scores students depending on how they answer individual items. Scoring details are provided in the following paragraphs.

#### 6.1.1 LIKELIHOOD FUNCTION

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

$$L(\theta) = L(\theta)^{3PL}L(\theta)^{CR},$$

where

$$L(\theta)^{3PL} = \prod_{i=1}^{N_{3PL}} P_i^{z_i} Q_i^{1-z_i}$$

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

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

$$q_i = 1 - p_i,$$

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

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

#### 6.1.2 DERIVATIVES

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

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

where

$$\begin{aligned}\frac{\partial \ln L(\theta)}{\partial \theta} &= \frac{\partial \ln L(\theta)^{MC}}{\partial \theta} + \frac{\partial \ln L(\theta)^{CR}}{\partial \theta} \\ \frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} &= \frac{\partial^2 \ln L(\theta)^{MC}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} \\ \frac{\partial \ln L(\theta)^{MC}}{\partial \theta} &= \sum_{i=1}^{N_{MC}} D a_i \frac{(P_i - c_i) Q_i}{1 - c_i} \left( \frac{z_i}{P_i} - \frac{1 - z_i}{Q_i} \right) \\ \frac{\partial^2 \ln L(\theta)^{MC}}{\partial^2 \theta} &= - \sum_{i=1}^{N_{MC}} D^2 a_i^2 \frac{(P_i - c_i) Q_i}{(1 - c_i)^2} \left( 1 - \frac{z_i c_i}{P_i^2} \right) \\ \frac{\partial \ln L(\theta)^{CR}}{\partial \theta} &= \sum_{i=1}^{N_{CR}} D a_i \left( \exp \left( \sum_{k=1}^{z_i} D a_i (\theta - \delta_{ki}) \right) \right) \left( \frac{z_i}{1 + \sum_{j=1}^{m_i} \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki}))} \right. \\ &\quad \left. - \frac{\sum_{j=1}^{m_i} j \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki}))}{\left( 1 + \sum_{j=1}^{m_i} \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki})) \right)^2} \right) \\ \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} &= \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left( \left( \frac{\sum_{j=1}^{m_i} j \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki}))}{1 + \sum_{j=1}^{m_i} \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki}))} \right)^2 \right. \\ &\quad \left. - \frac{\sum_{j=1}^{m_i} j^2 \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki}))}{1 + \sum_{j=1}^{m_i} \exp(\sum_{k=1}^j D a_i (\theta - \delta_{ki}))} \right)\end{aligned}$$

where  $\theta_t$  denotes the estimated  $\theta$  at iteration  $t$ .  $N_{CR}$  is the number of items that are scored using the generalized partial credit model (GPCM), and  $N_{3PL}$  is the number of items scored using the 3PL model.

### 6.1.3 EXTREME CASE HANDLING

Extreme unreliable student ability estimates are truncated to the lowest observable scores (LOT/LOSS) or the highest observable scores (HOT/HOSS). Note that

- LOT = lowest observable theta score;
- LOSS = lowest observable scale score;
- HOT = highest observable theta score; and
- HOSS = highest observable scale score.

Estimated theta values lower than the LOT or higher than the HOT will be truncated to the LOT and HOT values and will be assigned the LOSS and HOSS associated with the LOT and HOT.

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded and an MLE cannot be generated. All incorrect and all correct cases will be scored by assigning the lowest observable and highest observable scale score, respectively.

Table 10 gives the LOT/LOSS and HOT/HOSS for the IREAD-3 assessment.

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

Lowest Observable Theta (LOT)	Highest Observable Theta (HOT)	Lowest Observable Scale Score (LOSS)	Highest Observable Scale Score (HOSS)
-4.22992	1.785323	200	650

### 6.1.4 STANDARD ERRORS OF ESTIMATES

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

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

where

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left( \left( \frac{\sum_{j=1}^{m_i} j \text{Exp}(\sum_{k=1}^j D a_i (\hat{\theta} - b_{ik}))}{1 + \sum_{j=1}^{m_i} \text{Exp}(\sum_{k=1}^j D a_i (\hat{\theta} - b_{ik}))} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 \text{Exp}(\sum_{k=1}^j D a_i (\hat{\theta} - b_{ik}))}{1 + \sum_{j=1}^{m_i} \text{Exp}(\sum_{k=1}^j D a_i (\hat{\theta} - b_{ik}))} \right) - \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i) Q_i}{(1 - c_i)^2} \left( 1 - \frac{z_i c_i}{P_i^2} \right)$$

and where  $m_i$  is the maximum possible score point (starting from 0) for the  $i$ th item,  $D$  is the scale factor, 1.7,  $N_{GPCM}$  is the number of items that are scored using GPCM items, and  $N_{3PL}$  is the number of items scored using 3PL model.

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

## 6.2 TRANSFORMING THETA SCORES TO REPORTING SCALE SCORES

Scale scores were reported for each student who took the IREAD-3 assessments. The scale scores were based on the operational items presented to the student and did not

include the filler item. The scale score is the linear transformation of the item response theory (IRT) ability estimate:

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

The summary of IREAD-3 scale scores for each administration is provided in Appendix C, Distribution of Scale Scores and Standard Deviations.

*Table 11: Scaling Constants on the Reporting Metric*

Slope (a)	Intercept (b)
74.81	516.44

### 6.3 OVERALL PERFORMANCE CLASSIFICATION

Each student was assigned an overall performance category in accordance with his or her overall scale score. Table 12 and Table 13 provide the scale score range for performance standards for IREAD-3 for grade 3 and grade 2, respectively. For grade 3, the lower bound of Level 2, Pass, marks the minimum cut score for proficiency in the foundational reading skills required by the end of grade 3. For grade 2, the Level 3 Pass cut is the same Level 2 Pass cut for grade 3, whereas the lower bound of Level 2, On Track, marks the minimum cut score required for students to be considered on track for proficiency by the end of grade 3.

*Table 12: Proficiency Levels*

Level 1 Did Not Pass	Level 2 Pass
200–445	446–650

*Table 13: Proficiency Levels for Grade 2 Students*

Level 1 At Risk	Level 2 On Track	Level 3 Pass
200–404	405–445	446–650

### 6.4 REPORTING CATEGORY SCORES

Reporting category scores are reported as raw score percentage correct, based on the operational items contained in a reporting category on the given form. Scores are reported for

- Reading: Foundations and Vocabulary
- Reading: Nonfiction

- Reading: Literature

## **6.5 LEXILE® SCORES**

In Spring and Summer 2022, IREAD-3 reported Lexile®<sup>1</sup> measures for students in the *Did Not Pass* performance level. MetaMetrics provided conversion tables between IREAD-3 scale scores and Lexile measures.

## **6.6 COMPARISON OF SCORES TO PREVIOUS YEAR**

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

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

---

<sup>1</sup> Lexile® measures are the intellectual property of Metametrics, Inc.

## 7. QUALITY ASSURANCE PROCEDURES

Quality assurance (QA) procedures are enforced throughout all stages of the Indiana Reading Evaluation and Determination (IREAD-3) test development, administration, and scoring and reporting. This chapter describes QA procedures associated with the following:

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

As QA procedures pervade all aspects of test development, the discussion of QA procedures is not limited to this chapter and is also discussed in chapters describing all phases of test development and implementation.

### 7.1 QUALITY ASSURANCE IN TEST CONFIGURATION

The Cambium Assessment, Inc. (CAI) scoring engine and the accuracy of data files are checked prior to their use in an operational test administration by using a simulated student response data file to check whether student responses entered in the Test Delivery System (TDS) were captured accurately and the scoring specifications were applied accurately. The simulated data file is scored independently by two programmers, following the scoring rules.

In addition to checking the scoring accuracy, CAI also thoroughly checks the test configuration file. For the operational administration, the test configuration file is the key file that contains all specifications for the item selection algorithm, and eventually the scoring algorithm, such as the test blueprint specification, slopes, and intercepts for theta-to-scale score transformation, cut scores, and the item information (cut scores, answer keys, item attributes, item parameters, passage information, etc.). The accuracy of the information in the configuration file is checked and confirmed numerous times independently by multiple staff members prior to the testing window.

### 7.2 QUALITY ASSURANCE IN COMPUTER-DELIVERED TEST PRODUCTION

#### 7.2.1 PRODUCTION OF CONTENT

The production of computer-based tests includes four key steps:

1. Final content is previewed and approved in a process called web approval. During web approval, items are “packaged” and presented to reviewers exactly as they will be displayed to the student.

2. A complete test configuration is approved. The final test configuration gathers content, form information, display information, and relevant scoring and psychometric information from the item bank and packages it for deployment.
3. Tests are initially deployed to a test site where they undergo platform review, a process during which reviewers ensure that each item displays properly on a large number of platforms representative of those used in the state for testing purposes.
4. The final system is deployed to a staging environment accessible to the Indiana Department of Education (IDOE) for user acceptance testing (UAT) and final review.

### **7.2.2 WEB APPROVAL OF CONTENT DURING DEVELOPMENT**

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

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

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

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

### **7.2.3 PLATFORM REVIEW**

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

Platform review is conducted by a team. The team leader projects an item as it was web approved in ITS, and team members, each behind a different platform, look at the same item to ensure that it renders as expected.

## 7.2.4 USER ACCEPTANCE TESTING AND FINAL REVIEW

Each release of every CAI system goes through a complete testing cycle, including regression testing. With each release, and every time a test is published, the system must undergo a period of UAT. During UAT, the client is provided with login information to an identical (though smaller scale) testing environment to which the system has been deployed. CAI provides recommended testing scenarios and constant support during the UAT period. CAI resolves identified issues before the opening of the testing window. Issues that cannot be resolved are noted for future review and resolution if a current resolution is not feasible within the timeline. IDOE provides signoff for administration go-live at the conclusion of the UAT period.

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

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

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

The PCB approves a maintenance plan that includes every scheduled change to the system. Deviations from the maintenance plan must be approved by the PCB, including server or driver patches that differ from those approved in the maintenance plan. Every bug fix, enhancement, data correction, or new feature must be presented with the results of a quality assurance plan and approved by the PCB.

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

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

### ***Cutover and Parallel Processing***

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

The production environment runs independently of the other environments and is changed only with the approval of the PCB. When developing enhancements, they are developed and tested initially on the development and test environments in Washington, D.C. before being deployed to the staging environment in Rackspace.

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

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

### **7.2.5 FUNCTIONALITY AND CONFIGURATION**

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

One area of quality unique to online delivery is the quality of the delivery system. Three activities provide for the predictable, reliable, quality performance of the delivery system:

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

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

In addition, latency data are captured for each assessed student—data about how long it takes to load, view, or respond to an item. All this information is logged, as well, enabling

CAI to automatically identify schools or districts experiencing unusual slowdowns, often before they are able to notice any lags.

### **7.3 Quality Assurance in Data Preparation**

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

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

When the two data files match exactly, they are then passed over to two independent psychometricians, who each perform classical and IRT analyses. Any discrepancies are identified and resolved.

When all results from the independent analysts match, the final results are uploaded to CAI's ITS.

CAI's TDS has a real-time quality-monitoring component built in. As students test, data flow through the Quality Monitor (QM) software. QM conducts a series of data integrity checks, ensuring, for example, that the record for each test contains information for each item that was supposed to be on the test, and that the test record contains no data from items that have been invalidated. QM scores the test, recalculates performance-level designations, calculates subscores, compares item parameters to the reference item parameters in the bank, and conducts a host of other checks.

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

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

## **7.4 QUALITY ASSURANCE IN ITEM ANALYSIS AND EQUATING**

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

- Verify accuracy of program code and procedures; and
- Evaluate the communication and work flow among participants. If necessary, the team will reconcile differences and correct production or verification programs.

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

## **7.5 QUALITY ASSURANCE IN SCORING AND REPORTING**

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

### **7.5.1 QUALITY ASSURANCE IN TEST SCORING**

CAI verifies the accuracy of the scoring engine using simulated test administrations. The simulator generates a sample of students with an ability distribution that matches that of the state. The ability of each simulated student is used to generate a sequence of item responses consistent with the underlying ability. Although the simulations were designed to provide a rigorous test of the adaptive algorithm for adaptively administered tests, they also provide a check of the full range of item responses and test scores in fixed-form tests. Additionally, these simulations ensure that students at all performance levels are exposed to the full range of test item content as dictated by the IREAD-3 test blueprints. Simulations are always generated using the production item selection and scoring engine to ensure that verification of the scoring engine is based on a very wide range of student response patterns.

To verify the accuracy of the Centralized Reporting System (CRS), item response data is merged with the demographic information taken either from previous year assessment data. If current year enrollment data are available by the time simulated data files are created, online reporting can be verified using the current year's testing information. By populating the simulated data files with real school information, it is possible to verify that special school types and special districts are being handled properly in the CRS.

Specifications for generating simulated data files are included in the analysis output student data file specifications document submitted to IDOE each year. Review of all simulated data is scheduled to be completed prior to the opening of the test

administration, so that the integrity of item administration, data capture, and item and test scoring and reporting can be verified before the system goes live.

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

QA reports are generated on a regular schedule, and item analysis reports are evaluated frequently at the opening of the testing window to ensure that items are performing as anticipated. Each time the reports are generated, the lead psychometrician reviews the results. If any unexpected results are identified, the lead psychometrician alerts the project manager immediately to resolve any issues. Table 14 presents an overview of the quality assurance (QA) reports.

*Table 14: Overview of Quality Assurance Reports*

QA Reports	Purpose	Rationale
Item Statistics	To confirm whether items work as expected	Early detection of key errors

### ***Item Analysis Report***

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

***Item p-Value.*** For multiple-choice items, the proportion of students selecting each response option is computed, and if the keyed response is not the modal response, the item is flagged. Although the correct response is not always the modal response, keyed response options flagged for both low biserial correlations and non-modal response are indicative of miskeyed items.

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

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

### **7.5.2 QUALITY ASSURANCE IN REPORTING**

Scores for the IREAD-3 online assessments are assigned by automated systems in real time. The machine rubrics are created and reviewed along with the items, then validated and finalized during rubric validation following field testing. The review process “locks down” the item and rubric when the item is approved for web display (Web Approval). During operational testing, actual item responses are compared to expected item responses (given the IRT parameters), which can detect miskeyed items, item drift, or other scoring problems. Potential issues are automatically flagged in reports available to psychometricians.

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

During the operational testing window and before scores are reported, IDOE and CAI collaborate to perform a final quality assurance for scoring called “test deck.” IDOE completes test events for demo students using specific response patterns that should result in expected scores. CAI independently scores these test events and provides the results to IDOE prior to reporting student results. IDOE checks that scores are reported as expected as a final confirmation of scoring and reporting.

## 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.
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.
- Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.
- Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40:106–108.
- van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997) *Handbook of modern item response theory*. New York: Springer-Verlag.
- Zwick, R. (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement (ETS Research Report No. 12-08). Princeton, NJ: Educational Testing Service.



***IREAD-3***

**Indiana Reading Evaluation and  
Determination**

**2021–2022**

**Volume 2  
Test Development**

## **ACKNOWLEDGMENTS**

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

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

## TABLE OF CONTENTS

1.	INTRODUCTION.....	1
1.1	Claim Structure .....	1
1.2	Underlying Principles Guiding Development .....	2
1.3	Organization of This Volume .....	2
2.	IREAD-3 BLUEPRINT.....	3
2.1	IREAD-3 Blueprint.....	3
3.	ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS.....	4
3.1	Overview of Item Development .....	4
3.2	Use of Item Specifications in Item Development .....	4
3.3	IREAD-3 Item Bank Summary .....	5
4.	IREAD-3 TEST CONSTRUCTION .....	6
4.1	Test Form Construction .....	6
5.	REFERENCES.....	7

## LIST OF TABLES

Table 1: Blueprint Percentage of Test Items Assessing Each Reporting Category .....	3
Table 2: Item Types and Descriptions .....	5

## LIST OF APPENDICES

Appendix A: IREAD-3 Blueprints  
Appendix B: Example Item Types  
Appendix C: Item Specifications

## 1. INTRODUCTION

The IREAD-3 assessment was designed to measure foundational reading skills and reading comprehension based on the Indiana Academic Standards (IAS). The Indiana State Board of Education (SBOE) approved the IAS in April 2014 for English/Language Arts (ELA). The IAS are intended to implement more rigorous standards, with the goal of promoting college-and-career readiness by challenging and motivating Indiana’s students to acquire stronger critical thinking, problem solving, and communications skills.

### 1.1 CLAIM STRUCTURE

The IREAD-3 assessment was designed to measure foundational reading standards at the end of grade 3. For school year 2021–2022, schools had the option of administering IREAD-3 to grade 2 students. Students who score at the Pass level on the IREAD-3 assessment demonstrate proficient understanding when reading and responding to grade-level literary and informational texts; students can also identify and comprehend most new variations of word meaning and new text-based vocabulary. Examples of specific knowledge, skills, and abilities for students scoring at the Pass level may include the following:

- Identify the main idea and supporting details in text
- Use information from the text to comprehend basic story plots
- Connect prior knowledge with literal information from nonfiction text
- Recall major points and make predictions about what is read
- Determine what characters are like by what they say or do in the story
- Determine the theme or author’s message in fiction and nonfiction text
- Distinguish basic text elements (e.g., problem and solution, fact and opinion, cause and effect)
- Distinguish beginning, middle, and ending sounds made by different letter patterns
- Identify simple, multiple-meaning words
- Use sentence clues to find meanings of unknown words
- Determine the meanings of words using knowledge of synonyms and antonyms
- Recognize common genres
- Read words with several syllables

Grade 2 students may score at the Pass, On-Track, or At-Risk level. Grade 2 students who score at the Pass level do not need to participate in IREAD-3 during grade 3; students who score at the On-Track level are not flagged for specific remediation but will still participate in grade 3; and students who score at the At-Risk level require remediation

efforts and targeted instruction in foundational reading skills to ensure they can achieve proficiency by the end of grade 3.

## **1.2 UNDERLYING PRINCIPLES GUIDING DEVELOPMENT**

The IREAD-3 item bank was established using a structured, evidence-centered design. The process for development began with detailed item specifications. The specifications, discussed in a later section, describe the interaction types that can be used, provide guidelines for targeting the appropriate cognitive engagement, and offer sample items and suggestions for controlling item difficulty.

Items for IREAD-3 were written with the goal that virtually every item would be accessible to all students, either by itself or in conjunction with accessibility tools, such as Text-to-Speech (TTS) or assistive technologies.

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

## **1.3 ORGANIZATION OF THIS VOLUME**

This volume is organized into the following four topics:

1. An explanation of the IREAD-3 test blueprint
2. An overview of the item development process that supports the validity of the claims the IREAD-3 assessment was designed to support
3. An overview of the IREAD-3 item pool
4. A description of test construction for the IREAD-3 assessment

## 2. IREAD-3 BLUEPRINT

Indiana educator committees, in collaboration with content experts, created the blueprints for IREAD-3.

### 2.1 IREAD-3 BLUEPRINT

Test specifications or blueprints provide the following guidelines:

- Length of the assessment
- Content areas to be covered and the acceptable number of items across standards within each content area or reporting category

*Table 1: Blueprint Percentage of Test Items Assessing Each Reporting Category*

Reporting Category	Reading Foundations and Vocabulary	Reading: Nonfiction	Reading: Literature	Total
Points	10–14	12–16	12–16	36–40
Percent	25–35%	30–40%	30–40%	100%

The IREAD-3 blueprint is provided in Appendix A. The blueprint is organized by reporting category and specifies the number of items required for each category to elicit the needed information from the student to justify strand-level scores.

The blueprint also defines the standards within each reporting category. The standards have assigned point ranges to ensure that the material is represented on a test form with the proper emphasis relative to other standards in that reporting category. The ranges in the blueprint allow each student to experience a wide range of content while still providing flexibility during form construction.

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

#### 3.1 OVERVIEW OF ITEM DEVELOPMENT

A previous Indiana vendor developed the IREAD-3 item bank using a rigorous, structured process that engaged stakeholders at critical junctures. The vendor used item writers with extensive experience developing items for standardized assessments, with most being teachers who had substantial knowledge of grade 3 curriculum and instruction. Educators reviewed items for content, bias, and sensitivity.

The item development process begins by defining passage and item specifications, and continues with

- selecting and training item writers;
- writing and review of items internally; and
- having state personnel and stakeholder committees conduct reviews.

Each step in this process helps ensure that the items can support the claims on which they are based. More information about the item development process can be found in the IREAD-3 *Spring 2018 Technical Report*.

#### 3.2 USE OF ITEM SPECIFICATIONS IN ITEM DEVELOPMENT

The IREAD-3 item specifications, given in Appendix C, were created by Indiana in summer 2015. Item specifications guided the item development process for all IREAD-3 items. Like the items themselves, item specifications go through item development and committee review.

The IREAD-3 item specifications include the following:

- **Content Standard.** This section identifies the standard being assessed.
- **Evidence Statement.** This section provides a statement that describes the knowledge and skills that an assessment item should elicit from students.
- **Content Limits/Constraints.** This section provides the limits/constraints that delineate the specific content that the standard measures, as well as the parameters in which items must be developed to assess the standard accurately, including the lower and upper complexity limits of items.
- **Depth of Knowledge Demands.** This section provides the demands that all IREAD-3 item specifications have; a Depth of Knowledge (DOK) value is based on Webb's DOK categories.
- **Item Type.** This section identifies which of two possible item types (multiple-choice and multi-part multiple-choice) is to be used.

- **Sample Items.** In this section, sample items present a range of response mechanisms. Each sample item contains detailed notes delineating the cognitive demands of the item and an explanation of its difficulty level.

### 3.3 IREAD-3 ITEM BANK SUMMARY

As described in Section 1, Introduction, all items used on the IREAD-3 assessment are aligned to the Indiana Academic Standards (IAS). Cambium Assessment, Inc. (CAI) inherited the IREAD-3 item bank from Indiana’s previous testing contractor, and no new development was performed.

Table 2 lists the item types used on IREAD-3 assessments and provides a brief description of each. Examples of various item types can be found in Appendix B.

*Table 2: Item Types and Descriptions*

<b>Response Type</b>	<b>Description</b>
Multiple-Choice (MC)	Student selects one correct answer from a number of options.
Multi-Part Multiple-Choice	Student selects one correct answer from a number of options for each part of the item.

## 4. IREAD-3 TEST CONSTRUCTION

Indiana assessment forms were constructed using the IREAD-3 blueprint and item pool. The construction of test forms is a process that requires judgment from content experts. It is based on psychometric criteria to ensure that certain technical characteristics of the test forms meet industry expected standards. The processes used for blueprint development and test form construction are described to support the claim that they are technically sound and consistent with expectations of current professional standards.

IREAD-3 is designed to support the claims described at the outset of this volume.

### 4.1 TEST FORM CONSTRUCTION

At the start of the IREAD-3 contract, Cambium Assessment, Inc. (CAI) was provided with a set of pre-built fixed forms to be delivered for the Spring 2019 and Summer 2019 administrations. Subsequent test administrations have also used pre-built forms, as specified in the IREAD-3 form re-use plan. More information about the test construction process can be found in the *IREAD-3 Spring 2018 Technical Report*.

The first segment of the forms includes four items and a sample item that the test administrator reads aloud to students, as well as stand-alone multiple-choice and multi-part multiple-choice items. Segments two and three comprise multiple-choice and multi-part multiple-choice items that are linked to reading passages.

As noted previously, segment one on the IREAD-3 assessment contains items that are read aloud to students. Students with a hard-of-hearing accommodation will not be able to access these items as intended in the construct being measured; thus, these four items are not administered to these students. As a result, students with the hard-of-hearing accommodation will have a lower total achievable raw score; however, their scale scores will be adjusted such that they are comparable to those of other students, irrespective of those four items.

## 5. REFERENCES

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



***IREAD-3***

**Indiana Reading Evaluation  
and Determination**

**2021-2022**

**Volume 3  
Test Administration**

## TABLE OF CONTENTS

1. INTRODUCTION.....	1
2. TESTING PROCEDURES AND TESTING WINDOWS.....	3
2.1 Grade 2 ILEARN-3 Testing .....	<b>Error! Bookmark not defined.</b>
2.2 Eligible Students .....	5
2.3 Testing Accommodations .....	6
2.3.1 Available Accommodations .....	8
3. ADMINISTRATOR TRAINING .....	10
3.1 Online Administration .....	10
3.1.1 Roles and Responsibilities in the Online Testing Systems .....	11
3.2 Test Administration Resources .....	12
4. TEST SECURITY PROCEDURES .....	15
4.1 Security of Test Materials.....	15
4.2 Investigating Test Irregularities .....	16
4.3 Tracking and Resolving Test Irregularities .....	17
4.4 CAI's System Security .....	18
REFERENCES.....	19

## LIST OF TABLES

Table 1: Designated Features and Accommodations Available in Spring 2022 .....	7
Table 2: User Guides and Manuals .....	13
Table 3: Examples of Test Irregularities and Test Security Violations .....	17

## LIST OF APPENDICES

Appendix A: <i>IREAD-3 Test Administrator’s Manual (TAM)</i>	
Appendix B: <i>Online Test Delivery System (TDS) User Guide</i>	
Appendix C: <i>Indiana Accessibility and Accommodations Guidance Manual</i>	
Appendix D: <i>Test Information Distribution Engine (TIDE) User Guide</i>	
Appendix E: <i>Indiana Assessments Policy Manual</i>	
Appendix F: <i>Technology Setup for Online Testing Quick Guide</i>	
Appendix G: <i>Accessibility and Accommodations Implementation and Setup Module</i>	
Appendix H: <i>Test Delivery System (TDS) Training Webinar Module</i>	
Appendix I: <i>Test Administration Overview Webinar Module</i>	
Appendix J: <i>Test Information Distribution Engine (TIDE) Webinar Module</i>	
Appendix K: <i>Test Delivery System (TDS) Webinar Module</i>	
Appendix L: <i>Online Reporting System (ORS) Webinar Module</i>	
Appendix M: <i>Technology Requirements for Online Testing Webinar Module</i>	
Appendix N: <i>Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux</i>	
Appendix O: <i>Indiana Online Practice Test User Guide</i>	
Appendix P: <i>Online Reporting System (ORS) User Guide</i>	
Appendix Q: <i>IREAD-3 ISR Interpretive Guide</i>	

## 1. INTRODUCTION

In Spring 2022, pursuant to IC 20-32-8.5-2, the Indiana Reading Evaluation and Determination (IREAD-3) test was administered to Indiana students in grade 3. For the first time in Spring 2022, IREAD-3 was also made available to a selected population of grade 2 students as part of a pilot study. Students in grades 4 and 5 who had not previously passed the IREAD-3 assessment were given the opportunity to retest during this administration. Students who did not pass the assessment during the previous administration could also retest in Summer 2022 unless the student obtained a Good Cause Exemption (GCE). A GCE serves to exempt a student from future IREAD-3 testing. Students who have previously been retained two times prior to promotion to grade four; students with Individualized Education Programs (IEPs), and English learners (ELs) with Individual Learning Plans (ILPs) are eligible for GCEs.

In Spring and Summer 2022, IREAD-3 was administered in CAI's Test Delivery System (TDS) as one test ID with three segments. Students were instructed to log out between each section and test administrator (TA) approval was required for a student to advance to each segment. A paper-pencil test was provided to students who could not take the test online due to their individual education plan (IEP).

The first four items on the IREAD-3 assessment are phonetics items that require a student to listen to item content. Students who are deaf or hard of hearing are not able to access this content. A separate hard of hearing test form was deployed for students who were designated as hearing impaired to ensure that their performance on the assessment was not impacted. The hard of hearing test form was available online in the TDS. Students testing with a paper-pencil accommodation skipped the first four items on the assessment.

Assessment instruments should have established test administration procedures that support useful interpretations of score results, as specified in Standard 6.0 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). This volume provides details on testing procedures, accommodations, test administrator training and resources, and test security procedures implemented for IREAD-3. Specifically, it provides the following evidence related to test administration for the validity of the assessment results:

- A description of the population of students who take IREAD-3.
- A description of the training and documentation provided to TAs in order for them to follow the standardized procedures for administration.
- A description of offered test accommodations that are 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 reproduction of any kind.

- A description of CAI's quality monitoring (QM) system and the test irregularity investigation process to detect cheating, monitor real-time item quality, and evaluate test integrity.

## **2. TESTING PROCEDURES AND TESTING WINDOWS**

Administering the 2021-2022 IREAD-3 assessments required coordination, detailed specifications, and proper training. Several groups of people were involved in the administration process, from those setting up secure testing environments to those administering the tests. Without the proper training and coordination of these individuals, the standardization of the test administration could have been compromised. IDOE worked with CAI to develop and provide the training and documentation necessary for the administration of IREAD-3 under standardized conditions within all testing environments, for both online and on paper-pencil tests.

All students were required to take a practice test at their school using the TDS interface prior to taking the IREAD-3 assessment. The practice test sessions helped students become familiar with TDS functionality and item types by providing students with sample test items similar to those they would encounter on the IREAD-3 assessment. Indiana students also had the opportunity to interact with released, non-secure items on a public facing Released Items Repository (RIR) assessment, available on the Indiana Assessment Portal. The IREAD-3 RIR was deployed in October 2018.

The Spring IREAD-3 and Summer IREAD-3 assessments were administered as one test with three segments. Schools had the flexibility to test at any time within the testing window, but it was recommended that schools administer no more than one segment per testing day. Schools were instructed to administer the three segments in chronological order.

The IREAD-3 assessment is a timed assessment, with each of the three segments lasting between 30-35 minutes. The Spring IREAD-3 testing window was March 7 – 18, 2022. The Summer IREAD-3 testing window was May 11 – July 15, 2022.

### **2.1 GRADE 2 IREAD-3 TESTING**

[Placeholder] Student literacy is most impacted through strong instruction at the early grade levels. To support instruction and early intervention, Indiana encouraged schools to opt-in to administer IREAD-3 at grade 2 beginning school year 2022-2023.

The Indiana Department of Education (IDOE) conducted analyses to confirm the degree of alignment between IREAD-3 and Indiana Academic Standards at grade 2 to ensure that grade 2 students had the opportunity to learn the foundational reading skills assessed. IDOE presented the results of the study to the Technical Advisory Committee (TAC), which confirmed that the content was appropriate for assessment of grade 2 students.

IDOE created a new cut score for IREAD-3 (on the current scale) to indicate “on track for reading proficiency at grade two.” IDOE then implemented a model where schools may elect to administer IREAD-3 to grade 2 students to obtain an earlier indicator of reading proficiency.

- Students who achieve the “Pass” proficiency level (446) in grade 2 pass the reading proficiency assessment and do not participate in IREAD-3 during grade three.
- Students who achieve the “On Track” proficiency level in grade 2 are not required to receive any specific remediation, but must participate in IREAD-3 during their grade 3 school year.
- Students who achieve the “At Risk” proficiency level in grade 2 must receive intervention during grade three and participate in IREAD-3 during their grade 3 school year.

## **2.2 ELIGIBLE STUDENTS**

Students in grade 3 were required to take IREAD-3 in Spring 2022 with or without accommodations if provided by an Individual Education Plan, Section 504 Plan, or ILP, including students who have been retained twice. Students who did not pass IREAD-3 in Spring 2022 had the option to participate in the assessment in either the summer retest window or in grade 5, if needed.

For the purpose of a pilot study in Spring 2022, a representative sample of grade 2 schools were recruited for testing. These students tested under the same guidelines as grade 3 students. Grade 2 students who achieved the IREAD-3 passing score do not have to test again as grade 3 students. Those who did not achieve the passing score will test along with all other grade 3 students in Spring 2023. Grade 2 students from this pilot test population were not eligible to take the summer retest in 2022.

The IREAD-3 assessment measures foundational reading standards for grade 3 students. Based on the Indiana Academic Standards, IREAD-3 is a summative assessment that was developed in accordance with IC 20-32-8.5-2. All grade 3 students are required to participate in IREAD-3 unless they have secured a valid exemption (i.e., a Good Cause Exemption) or are taking an alternate assessment (i.e., I AM):

- **Public and Accredited Non-Public School Students:** Indiana accredited public and non-public school students enrolled in grade 3 were required to participate in the IREAD-3.
- **Home Education Program Students:** Students who received instruction at home and were registered appropriately with their corporation office as Home Education Program students were eligible to participate in statewide assessments. If parents or guardians identified an IREAD-3 assessment as a selected measure of their child's annual progress, students could participate in an IREAD-3 administration, as directed by the Corporation Test Coordinator (CTC).
- **English Learners (ELs):** All ELs are required to participated in statewide assessments, including IREAD-3.

**Students with Disabilities:** Indiana has established the procedures to ensure the inclusion in IREAD-3 testing of all grade 3 students with disabilities. Federal and state law require that all students participate in the state testing system, including IREAD-3. In Indiana, a student with an IEP will participate in IREAD-3 without accommodations or with approved accommodations. Students who participate in Indiana's Alternate Measure (I AM) will not participate in IREAD-3. Per the Individuals with Disabilities Education Improvement Act (IDEA) and Title 511 Article 7-Special Education, published December 2014 by the Indiana State Board of Education, decisions regarding which assessment option a student will participate in are made annually by the student's IEP team and are based on the student's curriculum, present levels of academic achievement, functional performance, and learning characteristics. Decisions cannot be based on program setting, category of disability, percentage of time in a particular placement or classroom, or any considerations regarding a school's Adequate Yearly Progress (AYP) designation.

IDOE instructed schools to maintain documentation locally for any student who needed a passing score on IREAD-3 but was unable to test in spring or summer 2022 to administer IREAD-3 to the student during the next available test window.

## **2.3 TESTING ACCOMMODATIONS**

Students participating in the online, fixed-form IREAD-3 assessment are able to use the standard online testing features in TDS. These features include the ability to select an alternate background and font color, mouse pointer size and color, and font size before testing. During the test, students can zoom in and out to increase or decrease the size of text and images, highlight items and passages (or sections of items and passages), cross out response options by using the strikethrough function, use a notepad to make notes, and flag a question for review using the mark for review function.

All Indiana State Assessments have appropriate accommodations available to allow these options accessible to students with disabilities and ELs, including ELs with disabilities. Accommodations are provided to students with current IEPs or Section 504 Plans, as well as to students identified as English Learners (ELs). Accommodations available for eligible students participating in the IREAD-3 assessments are described in the *IREAD-3 Test Administrator's Manual (TAM)* (Appendix A), which were accessible before and during testing from the IREAD-3 portal.

IREAD-3 assessments provide two categories of assessment support to students: designated features and accommodations, both embedded (delivered through TDS) and non-embedded. Designated features for IREAD-3 are those supports that are available for use by any student for whom the need has been indicated by an educator (or team of educators with parent/guardian and student). The *Online Test Delivery System (TDS) User Guide* published on the IREAD-3 portal (Appendix B) provides instructions on how to access and use these features.

Volume 1 of this technical report lists the allowed accommodations and the number of students who were provided with accommodations during the Spring 2022 IREAD-3 test administration. Table 1 provides a list of designed features and accommodations that were offered during the Spring 2022 administration.

Table 1: Designated Features and Accommodations Available in Spring 2022

	Designated Features	Accommodations
Embedded	Color Contrast (Computer) Language (English or Braille) Masking Mouse Pointer Print Size	Hard of Hearing Test Form Streamline Text to Speech
Non-embedded	Access to Sound Amplification System Assistive Technology to Magnify/Enlarge Special Furniture or Equipment for Viewing Tests Time of Day for Testing Altered Special Lighting Conditions Color Acetate Film for Paper Assessments	Read Aloud to Self Large Print Booklet Braille Booklet Print Booklet Interpreter for Sign Language Read Aloud Script for Paper Booklet Human Reader Tested Individually Alternate Indication of a Response Braille Transcript for Audio Items Student Provided with Additional Breaks Bi-Lingual Word to Word Dictionary Color Acetate Film for Paper Test Student Provide with Extended Testing Time for Testing Sessions (e.g., 50% additional time)

IDOE also collected information about non-standard accommodation requests under a Special Requests section in TIDE below the designated features and accommodations. These special requests required IDOE approval.

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

If an EL or student with an IEP or Section 504 Plan used any accommodations during the test administration, this information was recorded by the test administrator (TA) in his or her required administration information.

Guidelines recommended for making accommodation decisions included the following:

1. Accommodations should facilitate an accurate demonstration of what the student knows or can do.
2. Accommodations should not provide the student with an unfair advantage or negate the validity of a test; accommodations must not change the underlying skills being measured by the test.
3. Accommodations must be the same or nearly the same as those used by the student while completing daily classroom instruction and routine assessment activities.
4. Accommodations must be necessary for 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:

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

### **2.3.1 AVAILABLE ACCOMMODATIONS**

The TA and the school test coordinator (STC) were responsible for ensuring that arrangements for accommodations were made before the test administration dates. IDOE provided a separate accessibility manual, the *Indiana Accessibility and Accommodations Guidance Manual* (Appendix C), as a supplement to the test administration manuals for individuals involved in administering tests to students with accommodations.

For eligible students with IEPs, Section 504 Plans, or ILPs participating in computer-based assessments, a full comprehensive list of accommodations is listed in the *TIDE User Guide* (Appendix D).

The Accommodation Guidelines provide information about the available tools, supports, and accommodations that are available to students taking the IREAD-3 assessments. For further information, please refer to the *Indiana Assessments Policy Manual* (Appendix E).

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

### 3. ADMINISTRATOR TRAINING

IDOE established and communicated to educators and key personnel involved with IREAD-3 administrations a clear, standardized procedure for the administration of IREAD-3 that was to be followed in all administrations, including those with accommodations. Key personnel included Corporation Test Coordinators (CTCs), Corporation Information Technology Coordinators (CITCs), Non-Public School Test Coordinators (NPSTCs), STCs, and TAs. The roles and responsibilities of staff involved in testing are further detailed in the next section.

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

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

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

The “*User Support*” section of both the *Online Test Delivery System (TDS) User Guide* (Appendix B) and the *Test Information Distribution Engine (TIDE) User Guide* (Appendix D) provides instructions to address possible technology challenges during test administration. The CAI Indiana Help Desk collaborated with IDOE to provide support to Indiana schools as they administered the state assessment.

#### 3.1 ONLINE ADMINISTRATION

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

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

### 3.1.1 ROLES AND RESPONSIBILITIES IN THE ONLINE TESTING SYSTEMS

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

#### *CTCs*

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

#### *CITCs*

CITCs were responsible for ensuring that testing devices were properly configured to support testing and coordinating participation in the 2021-2022 system readiness test (SRT). IDOE recommended that all schools complete an SRT prior to their first test administration. The SRT is a simulation of online testing at the state level that ensures student testing devices and local school networks were correctly configured to support online testing.

#### *NPSTCs*

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

#### *STCs*

Before each administration, STCs and CTCs were required to verify that student eligibility was correct in TIDE and that any accommodations or test settings were correct. To participate in a computer-based online test, students were required to have been listed as eligible for that test in TIDE. See the *Test Information Distribution Engine User Guide* (Appendix D) for more information.

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

#### *Test Administrators*

TAs administered both a practice test session prior to student's administration of the IREAD-3 assessment, and the operational IREAD-3 assessment.

TAs were responsible for reviewing necessary user manuals and user guides to prepare the testing environment and ensure that students did not have books, notes, scratch

paper, or electronic devices. They were required to administer IREAD-3 following the directions found in the guide. 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 being used during administration of IREAD-3.

### **3.2 TEST ADMINISTRATION RESOURCES**

The list of webinars and training resources for the spring 2022 IREAD-3 administration is provided below. Training materials were available online at <https://iread3.portal.cambiumast.com/teachers.html> and are included as appendices to this report volume.

- **Test Administrator (TA) Certification Course:** All educators who administer the IREAD-3 assessment are required to complete an online TA Certification Course.
- **Accessibility and Accommodations Implementation and Setup Module:** This online module provides information on the accessibility and accommodations in Indiana for the IREAD-3 tests (Appendix G).
- **Student Interface Training Webinar Module:** This online module provides information and a step by step guide through the student interface in the test delivery system.
- **Test Delivery System (TDS) Training Webinar Module:** This online module provides information and a step by step guide through the test administrator interface in the test delivery system (Appendix H).
- **Test Administration Overview Webinar Module:** This module provides a general overview of the TA's role in the test administration process, including key responsibilities before, during, and after the testing window (Appendix I).
- **Test Information Distribution Engine (TIDE) Webinar Module:** This module provides a general overview of TIDE and the features applicable to educators and administrators before, during, and after testing (Appendix J).
- **Test Delivery System (TDS) Webinar Module:** This module provides a general overview of TDS and the features available for both the TA and the student interface within TDS (Appendix K).
- **Online Reporting System (ORS) Webinar Module:** This module provides a general overview of ORS where student scores, including individual scores and aggregate scores, displayed after students completed the IREAD-3 assessments (Appendix L).
- **Technology Requirements for Online Testing Webinar Module:** This module provides technology requirements for corporation and school technology coordinators to ensure that their testing devices are set up properly before testing (Appendix M).

The administration resources comprising various tutorials and user guides (user manuals, quick guides, etc.) were available for Indiana personnel on the IREAD-3 Portal at <https://iread3.portal.cambiumast.com/teachers.html>.

Table 2 presents the list of available user guides and manuals related to the IREAD-3 administration. The table also includes a short description of each resource and its intended use.

*Table 2: User Guides and Manuals*

<b>Resource</b>	<b>Description</b>
<i>Online Test Delivery System (TDS) User Guide</i>	This user guide supports TAs who manage testing for students participating in the IREAD-3 practice tests and operational tests (Appendix B).
<i>Technology Setup for Online Testing Quick Guide</i>	This document explains in four steps how to set up technology in Indiana corporations and schools. (Appendix F).
<i>Additional Configurations and Troubleshooting Guide for Windows, Mac, Android, Chrome OS, and Linux</i>	This manual provides information about hardware, software, and network configurations for running various testing applications provided by CAI (Appendix N).
<i>Indiana Online Practice Test User Guide</i>	This user guide provided an overview of the IREAD-3 Practice Test (Appendix O).
<i>Test Information Distribution Engine (TIDE)</i>	This user guide described the tasks performed in the Test Information Distribution Engine (TIDE) for IREAD-3 assessments (Appendix D).
<i>Online Reporting System (ORS) User Guide</i>	This user guide provides an overview of the different features available to educators to support viewing student scores for the IREAD-3 assessment (Appendix P).
<i>Indiana Accessibility and Accommodations Guidance</i>	The accessibility manual establishes the guidelines for the selection, administration, and evaluation of accessibility supports for instruction and assessment of all students, including students with disabilities, English learners (ELs), ELs with disabilities, and students without an identified disability or EL status (Appendix C).
<i>IREAD-3 ISR Interpretive Guide</i>	This user guide is an annotated Individual Student Report (ISR) that provides information on how to read and interpret a student's IREAD-3 test results (Appendix Q).

## Department Resources and Support

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

- Weekly newsletter distributed via email from the IDOE Office of Assessment to all officially designated CTCs in IDOE's database. The newsletter was titled "IREAD-3 Assessment Update" and included new announcements relevant to the IREAD-3 assessment, reminders of upcoming milestones, and a 'planning ahead' section with important dates in the IREAD-3 program. The IDOE Office of Assessment contact information was also available at the end of each weekly newsletter so that corporations and schools could contact the IDOE directly if there were any questions.

- On an “as needed” basis, communications were sent out via email memos. These messages generally addressed specific issues that needed to be transmitted quickly to administrators and teachers in the field or important information that the IDOE wanted to ensure was clearly outlined due to its importance to the IREAD-3 program. The distribution was to superintendents, principals, and school leaders.
- General information about the assessments was posted on the IDOE Office of Assessment website at <https://www.in.gov/doe/students/assessment/>. This *Accessibility and Accommodations Guidance* in the IREAD-3 Policy and Guidance section of the website was often referenced to address questions pertaining to accommodations and overall accessibility.

### **IREAD-3 Practice Tests**

The purpose of practice tests is to familiarize students with the system, functionality, and item types that will appear on the IREAD-3 examination. Users could also watch tutorials on each item to familiarize themselves with the different features and response instructions for each item type. Practice tests are not intended to guide classroom instruction.

The IREAD-3 practice tests were deployed on October 13, 2021, and remained available throughout the testing window. Online practice tests were designed for use with the CAI Secure Browser, and CAI’s TDS delivered the practice tests in secure mode using the same test delivery engine as the operational test. The Indiana portal provided a list of supported web browsers that could be used to administer the practice tests.

The design of the secure mode ensured that students, teachers, and educators were familiar with the online testing system before operational testing began. Both practice and operational tests were delivered through the same system, and IDOE required all students to take a practice test prior to taking the operational IREAD-3 test.

Students taking the IREAD-3 assessment on paper were also required to take a test prior to taking the operational IREAD-3 assessment. Paper testers took a paper-based practice test, located at the beginning of the paper-and-pencil assessment booklets. The TA script provided specific instructions to ensure the students completed the paper practice test items before starting the operational IREAD-3 assessment. A practice test answer key was included within the TA script and provided educators the opportunity to ensure that their students understood how to respond to the different question types represented on the IREAD-3 assessment.

## **4. TEST SECURITY PROCEDURES**

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

The test security procedures for IREAD-3 included the following:

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

To support these policies and procedures, IDOE leveraged security measures within CAI systems. For example, students taking the IREAD-3 assessments were required to acknowledge a security statement confirming their identity and acknowledging that they would not share or discuss test information with others. Additionally, students taking the online assessments were logged out of a test within the CAI Secure Browser after 20 minutes of inactivity.

In developing the IREAD-3 TAM (Appendix A), IDOE and CAI ensured that all test security procedures were available to everyone involved in test administration. Each manual included protocols for reporting any deviations in test administration.

If IDOE determined that an irregularity in test administration or security had occurred, it acted following their approved procedures, including invalidating student scores if appropriate.

### **4.1 SECURITY OF TEST MATERIALS**

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 testing session, secure materials (e.g., scratch paper) were required to be returned immediately to the STC and placed in locked storage. Secure materials were never to be left unsecured and were not permitted to remain in classrooms or be removed from the school's campus overnight.

Secure materials that did not need to be returned to the print vendor for scanning and scoring could be destroyed securely following outlined security guidelines, but were not allowed to be discarded in the trash. In addition, any monitoring software that might have allowed test content on student workstations to be viewed or recorded on another computer or device during testing had to be disabled.

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

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

All accommodated test materials (regular print, large print, and braille) were treated as secure documents, and processes were in place to protect them from loss, theft, and reproduction.

To access the online IREAD-3 tests, a Secure Browser was required. The CAI Secure Browser provided a secure environment for student testing by disabling hot keys, copy, and screen capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines). Users could not access other applications from within the secure browser, even if they knew the keystroke sequences. Students were not able to print from the secure browser. During testing, the desktop was locked down. The Secure Browser was designed to ensure test security by prohibiting access to external applications or navigation away from the test. See the *Online Test Delivery System (TDS) User Guide* (Appendix B) for further details.

## **4.2 INVESTIGATING TEST IRREGULARITIES**

CAI's quality monitoring (QM) system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QM system and any anomalies (such as tests not meeting blueprint, unexpected test lengths, or other unlikely issues) are flagged. CAI psychometricians ran quality assurance reports and alerted the program team of any issues. 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.

Item statistics and blueprint reports were run and reviewed weekly during the Spring and Summer 2022 testing windows. Analyses relying on student ability were run after the summer administration when all items were calibrated and placed on the same scale.

CAI psychometricians monitored testing anomalies throughout the testing window. Evidence was collected for evaluation, including blueprint match, test times that were much longer than the state average, and item response patterns using the person-fit index. The flagging criteria used for these analyses are configurable and were confirmed by IDOE. While analyses used to detect the testing anomalies could be run anytime within the testing window, analyses relying on state averages typically were held until the close of the testing window to ensure final data were used.

No unexpected results were identified during the 2021-2022 IREAD-3 test windows. Had any unexpected results been identified, the lead psychometrician would have alerted the program team leads immediately to resolve any issues.

### 4.3 TRACKING AND RESOLVING TEST IRREGULARITIES

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

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

CTCs, NPSTCs, STCs, and TAs had to discuss the details of a test irregularity to determine whether test invalidation was appropriate. CTCs, NPSTCs, and STCs had to submit to IDOE a *Testing Concerns and Security Violations Report* when invalidating any student test in response to a test security breach or interaction that compromised the integrity of the student's test administration.

During the testing window, TAs were also required to immediately report any test incidents (e.g., disruptive students, loss of Internet connectivity, student improprieties) to the STC. A test incident could include testing that was interrupted for an extended period due to a local technical malfunction or severe weather. STCs notified CTCs or NPSTCs of any test irregularities that were reported. CTCs or NPSTCs were responsible for submitting requests for test invalidations to the IDOE via TIDE. IDOE made the final decision on whether to approve the requested test invalidation and the decision was recorded and processed through TIDE. CTCs or NPSTCs could track the status and final decisions of requested test invalidations and irregularities in TIDE. This information was stored in TIDE for the school year and remained available until TIDE was updated for the 2021-2022 school year.

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

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

Description
Student(s) making distracting gestures/sounds or talking during the test session that creates a disruption in the test session for other students.
Student(s) leaving the test room without authorization.
TA or Test Coordinator leaving related instructional materials on the walls in the testing room.
Student(s) cheating or providing answers to each other, including passing notes, giving help to other students during testing, or using handheld electronic devices to exchange information.
Student(s) accessing or using unauthorized electronic equipment (e.g., cell phones, smart watches, iPods, or electronic translators) during testing.
Disruptions to a test session such as a fire drill, school-wide power outage, earthquake, or other acts.

---

TA or Test Coordinator failing to ensure administration and supervision of the assessments by qualified, trained personnel.

---

TA giving incorrect instructions.

---

TA or Test Coordinator giving out his or her username/password (via email or otherwise), including to other authorized users.

---

TA allowing students to continue testing beyond the close of the testing window.

---

TA or teacher coaching or providing any other type of assistance to students that may affect their responses. This includes both verbal cues (e.g., interpreting, explaining, or paraphrasing the test items or prompts) and nonverbal cues (e.g., voice inflection, pointing, or nodding head) to the correct answer. This also includes leading students through instructional strategies such as think-aloud, asking students to point to the correct answer or otherwise identify the source of their answer, requiring students to show their work to the TA, or reminding students of a recent lesson on a topic.

---

TA providing students with unallowable materials or devices during test administration or allowing inappropriate designated features and/or accommodations during test administration.

---

TA providing a student access to another student's work/responses.

---

TA allowing students to continue testing beyond the close of the testing window.

---

TA or Test Coordinator modifying student responses or records at any time.

---

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

---

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

---

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

---

#### **4.4 CAI'S SYSTEM SECURITY**

CAI has built-in security controls for all of its data stores and transmissions. Unique user identification is a requirement for all systems and interfaces. All of CAI's systems encrypt data at rest and in transit. IREAD-3 data resides on servers at Rackspace, CAI's online hosting provider. Rackspace maintains 24-hour surveillance of both the interior and exterior of its facilities. Staff at both CAI and Rackspace receive formal training in security procedures to ensure that they know the procedures and implement them properly.

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

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

## **REFERENCES**

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*.



***IREAD-3***

**Indiana Reading Evaluation  
and Determination**

**2021–2022**

**Volume 4  
Evidence of Validity and  
Reliability**

## **ACKNOWLEDGMENTS**

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

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

## TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE .....	4
1.1. Reliability.....	5
1.2. Validity .....	7
2. PURPOSE OF IREAD-3 .....	10
3. EVIDENCE OF CONTENT VALIDITY .....	11
3.1. Content Standards .....	11
4. RELIABILITY.....	12
4.1. Marginal Reliability .....	12
4.2. Test Information Curves and Standard Error of Measurement.....	13
4.3. Reliability of Performance Classification .....	17
4.3.1 Classification Accuracy .....	18
4.3.2 Classification Consistency .....	19
4.4. Precision at Cut Scores .....	21
5. EVIDENCE ON INTERNAL–EXTERNAL STRUCTURE.....	23
5.1. Correlations Among Reporting Category Scores .....	23
5.2. Confirmatory Factor Analysis .....	24
5.2.1 Factor Analytic Methods .....	25
5.2.2 Results.....	27
5.2.3 Discussion .....	28
5.3. Local Independence .....	28
5.4. Convergent and Discriminant Validity.....	29
6. FAIRNESS IN CONTENT .....	31
6.1. Statistical Fairness in Item Statistics .....	31
7. SUMMARY .....	33
8. REFERENCES .....	34

## LIST OF TABLES

Table 1: Number of Items for Each Reporting Category by Administration .....	11
Table 2: Marginal Reliability Coefficients by Administration .....	13
Table 3: Descriptive Statistics .....	18
Table 4: Classification Accuracy Index (Grade 2) .....	19
Table 5: Classification Accuracy Index (Grade 3) .....	19
Table 6: False Classification Rates (Grade 2).....	20
Table 7: False Classification Rates (Grade 3).....	21
Table 8: Classification Accuracy and Consistency (Grade 2).....	21
Table 9: Classification Accuracy and Consistency (Grade 3).....	21
Table 10: Performance Levels and Associated Conditional Standard Error of Measurement.....	21
Table 11: Observed Correlation Matrix Among Reporting Categories .....	23
Table 12: Goodness-of-Fit Second-Order CFA.....	27
Table 13: Correlations Among Factors .....	28
Table 14: Q <sub>3</sub> Statistics .....	29
Table 15: Observed Score Correlations Spring.....	30

## LIST OF FIGURES

Figure 1: Sample Test Information Function .....	15
Figure 2: Conditional Standard Error of Measurement (Spring, Grade 2) .....	16
Figure 3: Conditional Standard Error of Measurement (Spring, Grade 3) .....	16
Figure 4: Conditional Standard Error of Measurement (Summer) .....	17
Figure 5: Second-Order Factor Model (IREAD-3) .....	27

## LIST OF APPENDICES

Appendix A: Reliability Coefficients

Appendix B: Conditional Standard Error of Measurement

## 1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

IREAD-3 was constructed to measure foundational reading standards in grade 3. The Indiana Academic Standards (IAS) in English Language Arts (ELA) are the foundation of IREAD-3, which was first administered to students during the spring of 2012 in accordance with the State of Indiana’s House Enrolled Act 1367. During the 2021–2022 school year there were two administrations: Spring 2022 and Summer 2022. The main test administration was online with braille and hard of hearing accommodations available. A paper-and-pencil version of the assessment was also available. Full descriptions of available accommodations are given in Volume 3, Section 1.2, of this technical report.

Since the Spring 2022 administration, the Indiana Department of Education (IDOE) has provided Indiana schools with option to administer the IREAD-3 assessment to students in grade 2. These students are administered the same IREAD-3 test form and scored with the same item response theory (IRT) parameters as grade 3 students. The only difference is that the On-Track cut score for grade 2 was developed to indicate when a grade 2 student is “on track” to reach proficiency in grade 3.

With the implementation of the IREAD-3 assessment, both reliability evidence and validity evidence were necessary to support appropriate inferences of student academic performance made on the basis of IREAD-3 scores. This volume of the technical report presents empirical evidence about the reliability and validity of the 2021–2022 IREAD-3 assessment.

The purpose of this volume is to provide empirical evidence that supports a validity argument regarding the uses and inferences for the IREAD-3 assessment. To that end, this volume addresses the following topics:

- *Reliability*. Estimates of marginal reliability for each administration are reported in this volume; these estimates are presented by administration in the main body and by demographic subgroups in Appendix A. This section of the report also includes conditional standard errors of measurement (CSEMs) and classification accuracy and consistency results by administration.
- *Content Validity*. Evidence is provided to show that test forms were constructed to measure foundational reading skills represented in the IAS, with blueprints that contained a sufficient number of items targeting each reporting category.
- *Internal Structure Validity*. Evidence regarding the internal relationships among subscale scores is shown in order to justify the IRT measurement model; this includes observed evidence and evidence gathered from correlations among reporting categories per administration. Confirmatory factor analysis (CFA) has also been performed using the second-order factor model. Additionally, local item independence, an assumption of unidimensional IRT, was tested using Yen’s  $Q_3$  fit statistic.
- *Test Fairness*. Fairness is statistically analyzed using differential item functioning (DIF) in tandem with content alignment reviews by specialists.

## 1.1 RELIABILITY

The term *reliability* refers to consistency in test scores and 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:

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

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

- The *test-retest* method measures stability over time. With this method, the same test is administered twice to the same group at two different points in time. If test scores from the two administrations are highly correlated, then the test scores are deemed to have a high level of stability. For example, if the result is highly stable, those who scored high on the first test administration tend to obtain a high score on the second administration. The critical factor, however, is the time interval. The time interval should not be too long, which could allow for changes in the test takers' true scores. Likewise, it should not be too short, or memory and practice may confound the results. The test-retest method is most effective for measuring constructs that are stable over time, such as intelligence or personality traits.
- The *parallel-forms* method is used for measuring equivalence. With this design, two parallel forms of the test are administered to the same group. This method requires two similar forms of a test. However, it is difficult to create two strictly parallel forms. When this method is applied, the effects of memory or practice can be eliminated or reduced, since the tests are not purely identical as is the case with the test-retest method. The reliability coefficient from this method indicates the degree to which the two tests are measuring the same construct. While there are many possible items to administer to measure any particular construct, it is feasible to administer only a sample of items on any given test. If there is a high correlation between the scores of the two tests, then inferences regarding high reliability of scores can be substantiated. This method is commonly used to estimate the reliability of performance or aptitude tests.
- The *split-half* method uses one test divided into two halves within a single test administration. It is crucial to make the two half-tests as parallel as possible, as the correlation between the two half-tests is used to estimate the reliability of the whole test. In general, this method produces a coefficient that underestimates the reliability of the full test. To correct the estimate, the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910) can be applied. While this method is convenient, varying splits of the items may yield different reliability estimates.
- The *internal consistency* method can be employed when it is not possible to conduct repeated test administrations. Whereas other methods often compute the

correlation between two separate tests, this method considers each item within a test to be a one-item test. There are several other statistical methods based on this idea: coefficient *alpha* (Cronbach, 1951), Kuder-Richardson Formula 20 (Kuder & Richardson, 1937), Kuder-Richardson Formula 21 (Kuder & Richardson, 1937), stratified coefficient *alpha* (Qualls, 1995), and the Feldt-Raju coefficient (Feldt & Brennan, 1989; Feldt & Qualls, 1996).

- *Inter-rater reliability* is the extent to which two or more individuals (coders or raters) agree. Inter-rater reliability addresses the consistency of the implementation of a rating system.

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

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

Returning to the definition of reliability as the ratio of true score variance to observed score variance, the following equation results:

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

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

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

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

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

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

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

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

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

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

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

## 1.2 VALIDITY

The term *validity* refers to the degree to which “evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Messick (1989) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment.” Both of these definitions emphasize evidence and theory to support inferences and interpretations of test scores. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) 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 (refer to Section 4.2, Alignment of IREAD-3 Test Forms to the Content Standards and Benchmarks). In order for test score inferences to support a validity claim, the items should be representative of the content domain and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies, in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct (refer to Volume 2 of this technical report for details). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance.

For example, a mathematics item targeting a specific mathematics skill that requires advanced reading proficiency and vocabulary has a high level of construct-irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores. Statistical analyses, such as CFA or multidimensional scaling, are also used to evaluate content relevance. Results from CFA for the IREAD-3

assessment are presented in Section 5.2, Confirmatory Factor Analysis. Evidence based on test content is a crucial component of validity, because construct underrepresentation or irrelevancy could result in unfair advantages or disadvantages to one or more groups of test takers.

The second source of validity evidence is based on “the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA, APA, & NCME, 2014). This evidence is collected by surveying test takers about their performance strategies or responses to particular items. Because items are developed to measure 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: analyzing the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests (refer to Volume 1, Section 5.2). Other possible analyses that can be used to examine internal structure are a dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (refer to Section 3, Reliability, and Section 5, Evidence of Internal-External Structure, for details).

A fourth source of evidence for validity is the relationship of test scores to external variables. The *Standards* (AERA, APA, & NCME, 2014) divide this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization.

- Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence delineates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multitrait-multimethod matrix can be used (refer to Section 5.4, Convergent and Discriminant Validity, for details).
- Test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends upon the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another.
- Validity generalization is related to whether the evidence is situation specific or can be generalized across different settings and times. For example, sampling errors or range restrictions may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

The fifth source of evidence for validity is that the intended and unintended consequences of test use should be included in the test-validation process. Determining the validity of the test should depend upon evidence directly related to the test; this process should not

be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is in fact due to an unintended, confounding aspect of the test, this would interfere with the test's validity. As described in this volume and 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 whether sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

## 2. PURPOSE OF IREAD-3

IREAD-3 is a standards-referenced test constructed to measure student proficiency on foundational reading skills found in the IAS. The test was developed using principles of evidence-centered design and adheres to the principles of universal design to ensure that all students have access to test content. IREAD-3 is a grade three reading assessment developed in accordance with state legislation. IREAD-3 is designed to measure foundational reading skills based on Indiana Academic Standards through grade three. The Indiana State Board of Education set forth guidance schools must use when making decisions about grade-level promotion, instructional plans, and Good Cause Exemption eligibility for individual students. The intent is to ensure each student receives the appropriate reading remediation based on IREAD-3 test data and their individual learning needs. Test results from these assessments can be employed to evaluate students' reading progress and help teachers improve their instruction and provide targeted reading instruction to students, which will have a positive effect on student literacy over time. This volume provides evidence of content validity in Section 3, Evidence of Content Validity. Volume 2, Test Development, describes the IAS and test blueprints in more detail.

IREAD-3 test scores are useful indicators for understanding individual students' academic performance of the IAS. The overall scale score and reporting category percent-correct scores were provided for each student to indicate student strengths and weaknesses in different content areas of the test relative to the other areas and to the district and state. These scores help teachers identify and respond to student needs, provided that they are viewed with the usual caution that accompanies use of reporting category scores. Thus, the reliability coefficients for these test scores and the validity of the test scores must be examined to support practical use of these tests across the state. This volume discusses how individual test scores can be used to measure test reliability in Section 4, Reliability. Volume 5 of this technical report is the score interpretation guide and provides details on all generated scores and their appropriate uses and limitations.

### 3. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the IREAD-3 assessment are representative of the content standards of the larger knowledge domain. It describes the content standards for IREAD-3 and discusses the test development process, mapping IREAD-3 tests to the standards. A complete description of the test development process can be found in Volume 2, Test Development.

#### 3.1 CONTENT STANDARDS

The IREAD-3 assessment measures foundational reading standards. It is designed to measure basic reading skills and reading comprehension based on the IAS. Blueprints were developed to ensure that the test and the items were aligned to the prioritized standards that they were intended to measure. The IREAD-3 blueprint is available in Volume 2, Appendix A. Table 1 presents the number of items measuring each reporting category by administration.

*Table 1: Number of Items for Each Reporting Category by Administration*

Reporting Category	Administration	
	Spring	Summer
Reading: Foundations and Vocabulary	12	12
Reading: Nonfiction	12	14
Reading: Literature	14	12

## 4. RELIABILITY

### 4.1 MARGINAL RELIABILITY

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

Within the item response theory (IRT) framework, measurement error varies across the range of ability. The amount of precision is indicated by the test information at any given point of a distribution. The inverse of the TIF represents the SEM. SEM is equal to the inverse square root of information. The larger the measurement error, the less test information is being provided. The amount of test information provided is at its maximum for students toward the center of the distribution, as opposed to students with more-extreme scores. Conversely, measurement error is minimal for the portion 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 is defined as:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 2 presents the marginal reliability coefficients by administration.

*Table 2: Marginal Reliability Coefficients by Administration*

Administration	Grade	Marginal Reliability
Spring	2	0.899
Spring	3	0.841
Summer	3	0.867

## 4.2 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

Within the IRT framework, measurement error varies across the range of ability as a result of the test, providing varied information across the range of ability as displayed by the TIF. The TIF describes the amount of information provided by the test at each score point

along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. This means, for instance, that if the measurement error is large, less information is being provided by the assessment at the specific ability level.

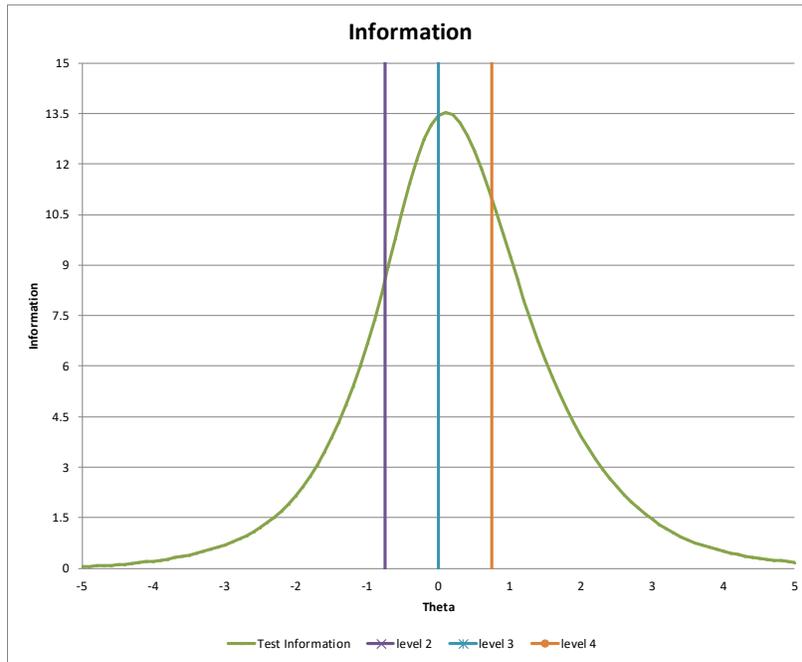
Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most-precise scores in this range. Where the curve is lower at the tails indicates that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

Computing these TIFs is useful to evaluate where the test is maximally informative. In IRT, the TIF is based on the estimates of the item parameters in the test, and the formula used for the IREAD-3 assessment is calculated as

$$TIF(\theta_i) = \sum_{j=1}^{N_{GPCM}} D^2 a_j^2 \left( \frac{\sum_{s=1}^{m_j} s^2 \text{Exp}(\sum_{h=1}^s D a_j (\theta_i - b_{jh}))}{1 + \sum_{s=1}^{m_j} \text{Exp}(\sum_{h=1}^s D a_j (\theta_i - b_{jh}))} - \left( \frac{\sum_{s=1}^{m_j} s \text{Exp}(\sum_{h=1}^s D a_j (\theta_i - b_{jh}))}{1 + \sum_{s=1}^{m_j} \text{Exp}(\sum_{h=1}^s D a_j (\theta_i - b_{jh}))} \right)^2 \right) + \sum_{j=1}^{N_{3PL}} D^2 a_j^2 \left( \frac{Q_j [P_j - c_j]^2}{P_j [1 - c_j]} \right),$$

where  $N_{GPCM}$  is the number of items that are scored using generalized partial credit model items,  $N_{3PL}$  is the number of items scored using the 3PL model,  $i$  indicates item  $i$  ( $i \in \{1, 2, \dots, N\}$ ),  $m_i$  is the maximum possible score of the item,  $s$  indicates student  $s$ , and  $\theta_s$  is the ability of student  $s$ .

Figure 1: Sample Test Information Function



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

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

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented instead of the TIFs for the Spring 2022 and Summer 2022 administrations in Figure 2 through Figure 4. These plots are based on the scaled scores reported during the 2021–2022 School Year. The vertical line represents the performance category cut score.

Figure 2: Conditional Standard Error of Measurement (Spring, Grade 2)

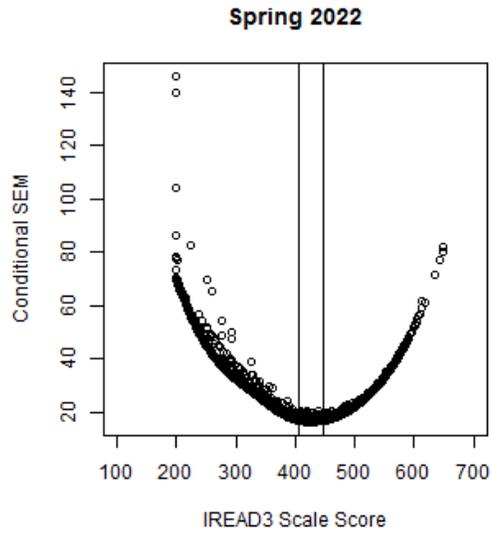


Figure 3: Conditional Standard Error of Measurement (Spring, Grade 3)

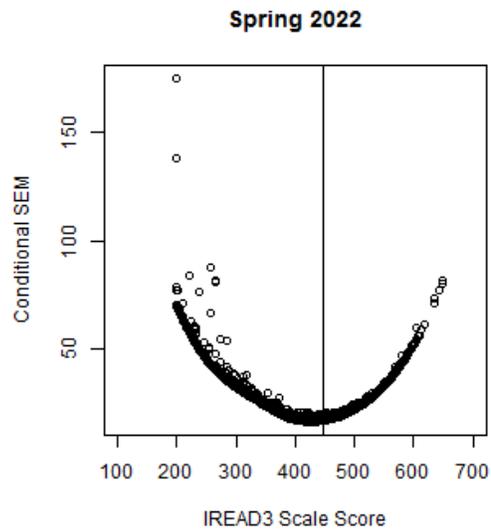
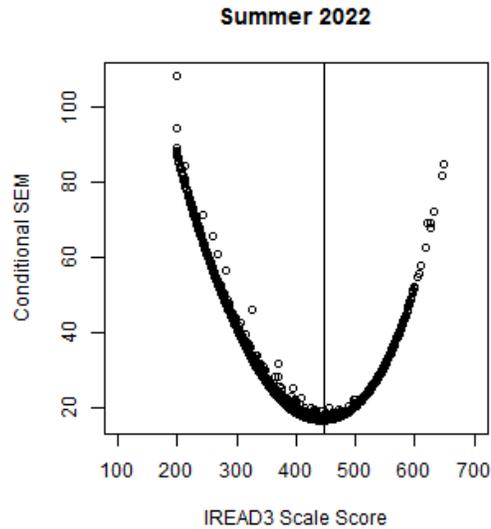


Figure 4: Conditional Standard Error of Measurement (Summer)



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

Reporting category summaries presented in Appendix A and Appendix B include the average CSEM by scale score and corresponding performance levels for each scale score.

### 4.3 RELIABILITY OF PERFORMANCE CLASSIFICATION

When students complete IREAD-3 assessments, they are placed into performance levels given their observed scaled score. The cut score for student classification into the different performance levels were previously determined.

Misclassification probabilities are computed for the *Pass* and *Do Not Pass* 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 (Rudner, 2001) that computes misclassification probabilities and is designed to explore the following research questions:

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

The second approach (Lee, Hanson, & Brennan, 2002; Guo, 2006) computes misclassification probabilities using an IRT-based method for students scoring at each score point. This approach is designed to explore the following research questions:

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

Both approaches yield student-specific classification probabilities that can be aggregated to form overall misclassification rates for the test.

These analyses used scores reported in the IREAD-3 state student data files. Table 3 provides the sample size, mean, and standard deviation of the observed theta data. The theta scores are based on the maximum likelihood estimates (MLEs) obtained from CAI’s scoring engine.

Table 3: Descriptive Statistics

Administration	Grade	Sample Size	Mean Theta	Standard Deviation of Theta	Mean Scale Score	Standard Deviation of Scale Scores
Spring	2	20,199	-1.19	1.16	427.14	86.83
Spring	3	85,212	-0.34	1.17	490.77	87.46
Summer	3	16,265	-1.23	0.98	424.43	73.61

### 4.3.1 CLASSIFICATION ACCURACY

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

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

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

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

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

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

where  $pl_j$  is the  $j$ th student’s performance level and the values of  $E_{vl}$  are the elements used to populate the matrix  $\mathbf{E}$ , a  $4 \times 4$  matrix of conditionally expected numbers of

students to score within each performance-level bin based on their true scores. The overall CAI of the test can then be estimated from the diagonal elements of the matrix

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

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

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

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

Table 4 and Table 5 provide the overall CAIs based on the observed score approach (Rudner, 2001). There is no industry standard, but these numbers suggest that misclassification would not be frequent in the population data.

*Table 4: Classification Accuracy Index (Grade 2)*

Administration	Overall Accuracy Index	Cut Accuracy Index	
		Cut 1 and Cut 2	Cut 2 and Cut 3
Spring	0.878	0.940	0.937

*Table 5: Classification Accuracy Index (Grade 3)*

Administration	Overall Accuracy Index
Spring	0.947
Summer	0.915

### 4.3.2 CLASSIFICATION CONSISTENCY

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

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

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

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

Similarly, the probability that a true score is below the cut can be estimated as

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

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

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

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

Table 6 and Table 7 provide the FPR, FNR, and accuracy measures derived with the IRT-based method (Lee, Hanson, & Brennan, 2002; Guo, 2006) for the IREAD-3 administrations.

*Table 6: False Classification Rates (Grade 2)*

Administration	1/2 Cut			2/3 Cut		
	FPR	FNR	Accuracy	FPR	FNR	Accuracy
Spring	0.032	0.024	0.944	0.032	0.027	0.941

*Table 7: False Classification Rates (Grade 3)*

Administration	FPR	FNR	Accuracy
Spring	0.029	0.02	0.951
Summer	0.045	0.037	0.918

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

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

Classification consistency with classification accuracy results derived using the IRT-based method (Lee, Hanson, & Brennan, 2002) are presented in Table 8 and Table 9. All accuracy values are higher than 0.91 and classification rates are higher than 0.88. Classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy rates because consistency is based on two tests with measurement errors, while accuracy is based on one test with a measurement error and the true score.

*Table 8: Classification Accuracy and Consistency (Grade 2)*

Administration	1/2 Cut		2/3 Cut	
	Accuracy	Consistency	Accuracy	Consistency
Spring	0.944	0.922	0.941	0.916

*Table 9: Classification Accuracy and Consistency (Grade 3)*

Administration	Accuracy	Consistency
Spring	0.951	0.930
Summer	0.918	0.884

#### 4.4 PRECISION AT CUT SCORES

Table 10 presents the mean CSEM at each performance level by administration. These tables also include performance-level cut scores and associated CSEM.

*Table 10: Performance Levels and Associated Conditional Standard Error of Measurement*

Administration	Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Spring	2	1	27.174	--	--

<b>Administration</b>	<b>Grade</b>	<b>Performance Level</b>	<b>Mean CSEM</b>	<b>Cut Score (Scale Score)</b>	<b>CSEM at Cut Score</b>
		2	17.045	405	17.038
		3	26.138	446	17.066
Spring	3	1	22.147	--	--
		2	33.400	446	17.020
Summer	3	1	25.774	--	--
		2	21.325	446	17.019

## 5. EVIDENCE ON INTERNAL–EXTERNAL STRUCTURE

This section explores the internal structure of the assessment using the scores provided at the reporting category level. The relationship of the subscores is just one indicator of the test dimensionality.

IREAD-3 assessments have three reporting categories: (1) Reading: Foundations and Vocabulary; (2) Reading: Nonfiction; and (3) Reading: Literature.

Overall scale scores and reporting-category percent correct were provided to students. Evidence is needed to verify that scale scores and percent correct for each reporting category provide both different and useful information about student performance.

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

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

Another approach 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.

### 5.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 11 presents the observed correlation matrix of the reporting category percent-correct scores for both administrations. The average correlation was 0.75 for the spring administration, and 0.68 for the summer administration.

*Table 11: Observed Correlation Matrix Among Reporting Categories*

Administration	Grade	Reporting Category	Number of Items	RFV	NF	L
Spring	2	Reading: Foundations and Vocabulary (RFV)	12	1.00		
		Reading: Nonfiction (NF)	12	0.70	1.00	
		Reading: Literature (L)	14	0.73	0.80	1.00
Spring	3	Reading: Foundations and Vocabulary (RFV)	12	1.00		
		Reading: Nonfiction (NF)	12	0.72	1.00	

Administration	Grade	Reporting Category	Number of Items	RFV	NF	L
		Reading: Literature (L)	14	0.74	0.82	1.00
Summer	3	Reading: Foundations and Vocabulary (RFV)	12	1.00		
		Reading: Nonfiction (NF)	14	0.67	1.00	
		Reading: Literature (L)	12	0.65	0.72	1.00

## 5.2 CONFIRMATORY FACTOR ANALYSIS

IREAD-3 test items were designed to measure different standards and higher-level reporting categories. Test scores were reported as an overall performance measure. Additionally, scores in the various reporting categories were also provided as indices of strand-specific performance. The strand scores were reported in a fashion that aligned with the theoretical structure of the test derived from the test blueprint.

The results in this section are intended to provide evidence that the methods for reporting IREAD-3 strand scores align with the underlying structure of the test and also provide evidence for appropriateness of the selected IRT models. This section is based on a second-order confirmatory factor analysis (CFA), in which the first-order factors load onto a common underlying factor. The first-order factors represent the dimensions of the test blueprint, and items load onto factors they are intended to measure.

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

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

The results in this section were based on the data collected from the initial administration of the IREAD-3 assessments, which was the Spring 2019 administration. The purpose is to provide validity evidence regarding the dimensionality of the assessments. Given there is no major change in test design, this analysis does not need to be conducted in subsequent administrations.

### 5.2.1 FACTOR ANALYTIC METHODS

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

As previously stated, the method of reporting scores used for the IREAD-3 assessments implies separate factors for each reporting category that are connected by a single underlying factor. This model is subsequently referred to as the implied model. In factor analytic terms, this suggests that test items load onto separate first-order factors, with the first-order factors connected to a single underlying second-order factor. The use of the CFA in this section establishes some validity evidence for the degree to which the implied model is reasonable.

A chi-square difference test is often applied to assess model fit. However, it is sensitive to sample size, almost always rejecting the null hypothesis when the sample size is large. Therefore, instead of conducting a chi-square difference test, other goodness-of-fit indices were used to evaluate the implied model for IREAD-3.

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

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

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

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

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

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

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

For the implied model a second-order factor analysis is posited, in which test items are coerced to load onto the reporting categories they are designed to target and all reporting categories share a common underlying factor. The second-order factor analysis can be mathematically represented as:

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

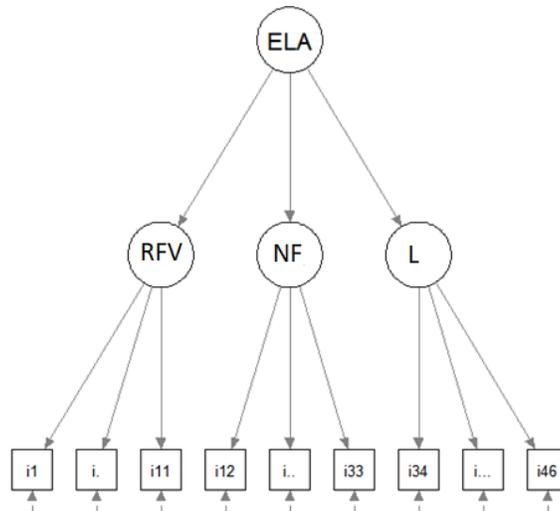
where  $\hat{\Sigma}$  is the implied correlation matrix among test items,  $\Lambda$  is the  $p \times k$  matrix of first-order factor loadings relating item scores to first-order factors,  $\Gamma$  is the  $k \times 1$  matrix of second-order factor loadings relating the first-order factors to the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

The second-order factor model can also be represented graphically, and a sample of the generalized approaches is provided on the following page. The general structure of the second-order factor analysis for ELA is illustrated in Figure 5, where Reading Foundations and Vocabulary (RFV), Nonfiction (NF), and Literature (L) represent the three reporting categories. This figure is generally representative of the factor analyses performed for all grades and subjects, with the understanding that the number of items within each reporting category could vary across the grades.

The purpose of conducting CFA for IREAD-3 is to provide evidence that each individual assessment in IREAD-3 implies a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

Figure 5: Second-Order Factor Model (IREAD-3)

## Generalized Second Order Factor Structure



## 5.2.2 RESULTS

Several goodness-of-fit statistics from each of the analyses are presented in Table 12, which shows the summary results obtained from CFA. Three goodness-of-fit indices were used to evaluate model fit of the item parameters to the manner in which students actually responded to the items. The root mean square error of approximation (RMSEA) is referred to as a badness-of-fit index, such that a value closer to 0 implies better fit and a value of 0 implies best fit. In general, an RMSEA below 0.05 is considered a good fit, and an RMSEA over 0.1 suggests a poor fit (Browne & Cudeck, 1993).

The Tucker-Lewis index (TLI) and the comparative fit index (CFI) are incremental goodness-of-fit indices that compare the implied model to the baseline model where no observed variables are correlated (i.e., there are no factors). Values greater than 0.9 are recognized as acceptable, and values over 0.95 are considered a good fit (Hu & Bentler, 1999). As Hu and Bentler (1999) suggest, the selected cut-off values of the fit index should not be overgeneralized and should be interpreted with caution.

Based on the fit indices, the model showed good fit across content domains. RMSEA was 0.022, and CFI and TLI were equal to or greater than 0.984.

Table 12: Goodness-of-Fit Second-Order Confirmatory Factor Analysis

Administration	df	RMSEA	CFI	TLI	Convergence
Spring	663	0.022	0.985	0.984	YES

Table 13 provides the estimated correlations between the reporting categories from the second-order factor model by administration. In all cases, these correlations are very

high. However, the results provide empirical evidence that there is some detectable dimensionality among reporting categories.

*Table 13: Correlations Among Factors*

Administration	Reporting Category	Number of Items	RFV	NF	L
Spring	Reading: Foundations and Vocabulary (RFV)	12	1.00		
	Reading: Nonfiction (NF)	12	0.953	1.00	
	Reading: Literature (L)	14	0.928	0.975	1.00

### 5.2.3 DISCUSSION

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

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

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

## 5.3 LOCAL INDEPENDENCE

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

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

When local independence is not met, there are issues of multidimensionality that are unaccounted for in the modeling of the data (Bejar, 1980). In fact, Lord (1980) noted that “local independence follows automatically from unidimensionality” (as cited in Bejar, 1980, p.5). From a dimensionality perspective, there may be nuisance factors that are influencing relationships among certain items, after accounting for the intended construct of interest. These nuisance factors can be influenced by a number of testing features, such as 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 test taker for item  $i$ ,  $T_i(\hat{\theta}_j)$  is the estimated true score for item  $i$  of test taker  $j$ , which is defined as

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

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

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

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

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

When there are  $n$  items,  $n(n-1)/2$ ,  $Q_3$  statistics will be produced. The  $Q_3$  values are expected to be small. Table 14 presents summaries of the distributions of  $Q_3$  statistics — minimum, 5th percentile, median, 95th percentile, and maximum — by administration. Overall, only four items had a  $Q_3$  value greater than the critical value of 0.2 for  $|Q_3|$  (Chen & Thissen, 1997).

Table 14:  $Q_3$  Statistics

Administration	Grade	$Q_3$ Distribution				
		Minimum	5th Percentile	Median	95th Percentile	Maximum
Spring	2	-0.121	-0.055	-0.021	0.035	0.292
Spring	3	-0.100	-0.054	-0.024	0.015	0.304
Summer	3	-0.082	-0.058	-0.025	0.014	0.160

## 5.4 CONVERGENT AND DISCRIMINANT VALIDITY

According to Standard 1.14 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), it is necessary to provide evidence of convergent and discriminant validity evidence. The a priori expectation is that subscores within the same subject (e.g., ELA and ELA) will correlate more positively than subscore correlations across subjects (e.g., ELA and Mathematics). These correlations are based on a small number of items, typically around 8–18; as a consequence, the observed score correlations will be smaller in magnitude because of the very large measurement error at the subscore level.

Part of demonstrating validity evidence is showing that assessment scores are related as expected with criteria and other variables for all student groups. However, a second

independent test measuring the same constructs as ELA and Mathematics in Indiana, which would allow for a cross-test set of correlations, was not available. Therefore, the correlations between subscores within and across tests were examined alternatively.

Grade 3 students provide an opportunity for such comparisons to be performed alternatively, as students in grade 3 take both ILEARN ELA and Mathematics assessments in addition to the IREAD-3 assessment. Table 15 shows the observed correlations between *ILEARN* Grade 3 ELA and Mathematics subscores and the IREAD-3 subscores. In general, the pattern is consistent with the prior expectation that subscores within a test correlate more highly than correlations between tests measuring a different construct, with a few small notes on the writing dimensions.

*Table 15: Observed Score Correlations Spring*

Subject	Reporting Category	IREAD-3		
		Cat1	Cat2	Cat3
ILEARN ELA Grade 3	Key Ideas and Textual Support/Vocabulary	0.55	0.66	0.65
	Structural Elements and Organization/Connection of Ideas/Media Literacy	0.50	0.62	0.60
	Writing	0.56	0.62	0.61
ILEARN Mathematics Grade 3	Algebraic Thinking and Data Analysis	0.61	0.65	0.64
	Computation	0.58	0.63	0.61
	Geometry and Measurement	0.60	0.63	0.61
	Number Sense	0.58	0.63	0.61

\*Cat1 = Reading Foundations and Vocabulary, Cat2 = Nonfiction, Cat3 = Literature

## 6. FAIRNESS IN CONTENT

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

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

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

### 6.1 STATISTICAL FAIRNESS IN ITEM STATISTICS

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

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

- Male/Female;

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

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 4.2, of the *2021–2022 IREAD-3 Annual Technical Report*.

## 7. SUMMARY

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

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

## 8. REFERENCES

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

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

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



**Indiana Reading Evaluation  
and Determination**

**2021–2022**

**Volume 5:  
Score Interpretation Guide**

## **ACKNOWLEDGMENTS**

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

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

## TABLE OF CONTENTS

1. INDIANA SCORE REPORTS.....	1
1.1 Overview of Indiana’s Score Reports .....	1
1.2 Overall Scores and Reporting Categories .....	2
1.3 Online Reporting System .....	3
1.3.1 Individual Student Report.....	3
1.3.2 Interpretive Guide.....	5
1.3.3 Data File .....	7
2. INTERPRETATION OF REPORTED SCORES .....	8
2.1 Scale Score .....	8
2.2 Performance Level .....	8
2.3 Performance Category for Reporting Categories .....	8
2.4 Cut Scores .....	9
2.5 Lexiles .....	9
2.6 Appropriate Uses for Scores and Reports.....	9
3. SUMMARY .....	11

## LIST OF TABLES

Table 1: Reporting Categories for IREAD-3 .....	2
Table 2: IREAD-3 Assessment Proficiency Cut Scores (Grade 2).....	9
Table 3: IREAD-3 Assessment Proficiency Cut Scores .....	9

## LIST OF FIGURES

Figure 1: Individual Student Report.....	4
Figure 2: Supplemental Interpretive Guide .....	6
Figure 3: Data File .....	7

## **1. INDIANA SCORE REPORTS**

Pursuant to IC 20-32-8.5-2, the *Indiana Reading Evaluation and Determination, Grade 3* (IREAD-3) assessment was administered to Indiana students in grade 3 in Spring 2022. In Spring 2022, IREAD-3 was also administered to grade 2 students from schools that opted to administer IREAD-3 to grade 2 students. Students in grades 4 and 5 who had not previously passed IREAD-3 were given the opportunity to retest during Spring 2022 and Summer 2022.

The purpose of this volume is to describe the information available from the scores reported for the 2021–2022 IREAD-3 assessments, and to define appropriate uses and inferences that can be drawn from them. This volume also documents the features of the score reports provided through the Indiana Online Reporting System (ORS), which is designed to assist stakeholders in reviewing, downloading, and appropriately interpreting test results.

### **1.1 OVERVIEW OF INDIANA’S SCORE REPORTS**

IREAD-3 was administered in Spring 2022 and Summer 2022. Test scores from each assessment were provided to IDOE corporations and schools through ORS beginning on March 21, 2022, for spring, and May 24, 2022, for summer.

The ORS is a web-based application that provides IREAD-3 results at various, privileged levels (<https://in.reports.cambiumast.com>). Assessment results are available to users according to their roles and the access they are given based on the authentication granted to them. There are four levels of user roles: corporation, school, teacher, and roster. Each user is given drill-down access to reports in the system based on his or her assigned role. This means that teachers can access data only for rosters of their own students; school administrators can access data only for the students in their own schools; and corporation administrators can access data for all schools and students in their corporation.

Users have the following types of access to the system:

- State: access to all state, corporation, school, teacher, and student test data.
- Co-Op Role (Co-Op) and Corporation Test Coordinator (CTC): access to all test data for their corporation and for the schools and students in their corporation.
- Non-Public School Test Coordinator (NPSTC), School Test Coordinator (STC), and Principal (PR): access to all test data for their school and the students in their school.
- Test Administrator (TA): access to the individual student report (ISR) and test data in a data file for students within his or her rosters.

Access to reports is password protected, and users can access data at their assigned level and below. For example, an STC can access the test data for students in his or her own school but not for students in another school.

## 1.2 OVERALL SCORES AND REPORTING CATEGORIES

Each student receives a single scale score if there is a valid score to report. The validity of a score is determined using invalidation rules, which define a set of parameters under which a student’s test may be counted. A student’s score will be automatically invalidated if he or she fails to respond to at least one item in each test segment. Normally, a student takes a test in the Test Delivery System (TDS) and then submits it. TDS then forwards the test for scoring before ORS reports the scores. However, tests may also be manually invalidated before reaching ORS if testing irregularities occur (e.g., cheating, unscheduled interruptions, loss of power or Internet connection).

A student’s score is based on the operational items on the assessment. A scale score describes how well a student performed on a test and is an estimate of students’ knowledge and skills as measured by the assessment. The scale score is transformed from a theta score, which is estimated on the basis of Item Response Theory (IRT) models as described in Volume 1. Lower scale scores indicate the student’s knowledge and skills fall below proficiency as measured by the assessment. Conversely, higher scale scores indicate the student has proficient knowledge and skills as measured by the assessment. Interpretation of scale scores is more meaningful when the scale scores are analyzed alongside performance levels and Performance-Level Descriptors (PLDs).

Based on the scale score, a student will receive an overall performance level. Performance levels are proficiency categories on an assessment, which students fall into based on their scale scores.

For IREAD-3, scale scores are mapped to two performance levels:

- Level 1: Did Not Pass
- Level 2: Pass

Performance levels can be interpreted through PLDs, which are a descriptive analysis of a student’s abilities based on performance. PLDs describe content-area knowledge and skills that students at each performance level are expected to possess and are determined by comparing a student’s scale score against carefully established cut scores that are unique to each grade and subject. Cut points are listed in Section 2.4, Cut Scores.

In addition to an overall score, students receive reporting-category scores. Reporting categories represent distinct areas of knowledge within each grade and subject. For IREAD-3, students’ performance in each reporting category is reported as a raw score percent correct.

Table 1 displays the IREAD-3 reporting categories.

*Table 1: Reporting Categories for IREAD-3*

Test	Reporting Category
IREAD-3	Reading: Foundations and Vocabulary Reading: Nonfiction Reading: Literature

## **1.3 ONLINE REPORTING SYSTEM**

ORS generates a set of online score reports that describe student performance for students, families, educators, and other stakeholders. The online score reports are produced after the tests are submitted by the students, machine-scored, and processed into ORS. Score data were published to ORS on March 21, 2022, for the Spring 2022 test administration, and on May 24, 2022, for the Summer 2022 test administration. Quality control verification was conducted on the score data for all test windows before data were released in ORS.

### **1.3.1 INDIVIDUAL STUDENT REPORT**

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

- Scale score
- Overall subject-performance level

The top of the report includes the student's

- name;
- scale score; and
- performance level.

The middle section includes:

- A barrel chart with the student's scale score
- PLDs with cut scores at each performance level

The bottom of the report includes information on student performance in each reporting category.

Figure 1 presents an example ISR for IREAD-3.

Figure 1: Individual Student Report



### Individual Student Report

How did my student perform on the test?

Test: IREAD-3

Year: Spring 2022

Name: Demo, Student J.

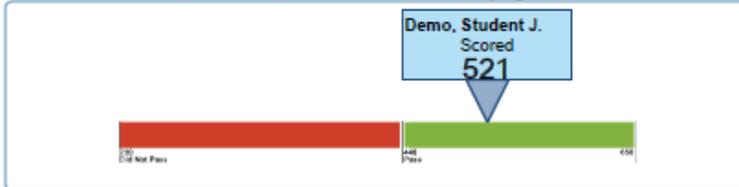
**Overall Performance on the IREAD-3 Test: Demo, Student J., Spring 2022**

Name	STN	Scale Score	Passing Status	Reported Lexile® Measure
Demo, Student J.	999999999	521	Pass	N/A

**Lexile® Information**

The Lexile® Framework for Reading is a scientific approach to reading and text measurement. A Lexile reader measure represents a person's reading ability on the Lexile scale.

**Scale Score and Performance on the IREAD-3 Test: Demo, Student J., Spring 2022**



**Passing Status Description**

**Pass**  
Students demonstrate proficient understanding when reading and responding to grade-level literary and informational texts. Students identify and comprehend most new variations of word meaning and new text-based vocabulary.

Your student's performance on the IREAD-3 assessment may be described in terms of percentage of total points earned for each of Indiana's grade 3 reading strands. Percentage of total points earned shows the total number of points your student earned on the test divided by the number of points the test was worth. These percentages are unique to this year's assessment items and may vary from year to year, so they should not be compared across years like scale scores may be. Please note these scores cannot be added together to equal the 3-digit scale score reported above.

**Performance on the IREAD-3 Test, by Strand: Demo, Student J., Spring 2022**

Strand	Percent Correct
Reading: Foundations and Vocabulary	100
Reading: Nonfiction	75
Reading: Literature	86

Based on data from the IREAD-3, Spring 2022 administration.  
Report Generated: 11/02/2022 1:20:28 PM EDT  
For help in understanding your student's score and this report, contact your student's teacher or school principal.

### **1.3.2 INTERPRETIVE GUIDE**

When printing ISRs, users have the option to print a supplemental “interpretive guide” (or “addendum” when printing a simple ISR), intended as a stand-alone document (see Figure 2), to help teachers, administrators, families, and students better understand the data presented in the ISR. The ISRs and the supplemental “interpretive guide” are available in five different languages: Arabic, Burmese, Chinese, Spanish, and Vietnamese.

Figure 2: Supplemental Interpretive Guide



**Indiana**  
DEPARTMENT OF EDUCATION  
*Working Together for Student Success*

## Indiana Reading Evaluation and Determination IREAD-3 Assessment Results

**Dear Parent/Guardian,**

This report provides information about your child's performance on the Indiana Reading Evaluation and Determination (IREAD-3) assessment. IREAD-3 is a summative assessment administered to all third graders enrolled in accredited Indiana schools to determine mastery of foundational reading skills.

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

Indiana Department of Education

**INFORMATION ON INDIANA'S IREAD-3 ASSESSMENT**

IREAD-3 measures foundational reading standards through grade three. Overall student results on IREAD-3 are reported as three-digit scale scores. These scale scores align with the two proficiency levels (Pass and Did Not Pass), based on the Indiana Academic Standards related to reading. IREAD-3 is a summative assessment, given at the end of instruction, to determine proficiency on a set of standards.

**UNDERSTANDING THE IREAD-3 ASSESSMENT**

**Individual Student Report**

How did my student perform on the test?  
Test: IREAD-3  
Year: Spring 2019  
Name: Demo, Student A.

**Basic test information**

A scale score is your child's overall numerical score placed on an alternative scale rather than using percent correct or a raw score.

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

Overall Performance on the IREAD-3 Test: Demo, Student A., Spring 2019			
Name	STN	Scale Score	Passing Status
Demo, Student A.	9999/99001	531	Pass

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

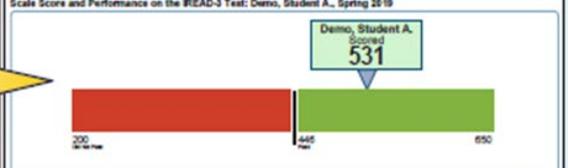
**Questions to consider with your child's teacher:**

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

For IREAD-3, the test scale is divided into two levels using one cut score: 446. The cut score is the score that separates the two levels. Based on your child's IREAD-3 scale score, he/she is placed into one of two proficiency levels: Pass or Did Not Pass.

Also included is a breakdown of performance across three domains within a content area, showing what percentage of the maximum points your child scored for each strand. These percentages cannot be added to achieve the scale score.

Scale Score and Performance on the IREAD-3 Test: Demo, Student A., Spring 2019



Performance on the IREAD-3 Test, by Strand: Demo, Student A., Spring 2019

Strand	Percent Correct
Reading: Foundations and Vocabulary	86
Reading: Nonfiction	86
Reading: Literature	86

**ADDITIONAL RESOURCES**

- To practice questions similar to what your child has seen on IREAD-3, go to <https://iread3.portal.cambiumast.com/>
- For more information about this assessment, go to <https://www.in.gov/doe/students/assessment/iread-3/>

Indiana Department of Education

Score Interpretation Guide

6

Indiana Department of Education

### 1.3.3 DATA FILE

ORS users have the option to quickly generate a comprehensive data file of their students' scores. Users can access data based upon their user role for current students or students that were theirs during the test administration. Data files (see Figure 3), which can be downloaded in Microsoft Excel or CSV format, contain a 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 as corporation, school, teacher, or roster reports.

Figure 3: Data File

A	B	C	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 FI	Student La	STN	Student D	Gender	Ethnicity	Special Ed	Identified	Section 5C	Socioecon	Enrolled C	Enrolled S	Enrolled S	Enrolled C	Enrolled C	Test name	Overall sc	Overall or	Reported I	College ar	Passing SI	Reporting									
Demofirst Demolast	19999923	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	Undetermined														
Demofirst Demolast	59999912	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	278	BR190L	Did Not Pass	21	33	29									
Demofirst Demolast	60000642	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	281	BR120L	Did Not Pass	21	0	29									
Demofirst Demolast	60000643	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	200	BR400L	Did Not Pass	7	0	14									
Demofirst Demolast	60000643	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	238	BR270L	Did Not Pass	14	25	0									
Demofirst Demolast	60000643	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	317	5L	Did Not Pass	29	17	21									
Demofirst Demolast	60000643	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	314	BR5L	Did Not Pass	43	17	14									
Demofirst Demolast	60000643	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	350	120L	Did Not Pass	50	42	36									
Demofirst Demolast	60000644	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	247	BR235L	Did Not Pass	20	8	7									
Demofirst Demolast	60000645	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	200	BR400L	Did Not Pass	10	8	14									
Demofirst Demolast	60000645	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	263	BR180L	Did Not Pass	20	17	7									
Demofirst Demolast	60000645	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	317	5L	Did Not Pass	30	8	21									
Demofirst Demolast	60000645	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	352	125L	Did Not Pass	40	25	36									
Demofirst Demolast	60000645	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	341	90L	Did Not Pass	20	33	29									
Demofirst Demolast	60000646	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	Undetermined														
Demofirst Demolast	60000647	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	Undetermined														
Demofirst Demolast	60000647	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	Undetermined														
Demofirst Demolast	60000648	01/03/19	M	White	N	N	N	N	N	5	Demo inst	9999_999	Demo Corj	9999	IREAD-3	394	270L	Did Not Pass	64	42	21									
Demofirst Demolast	60000648	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	Undetermined														
Demofirst Demolast	60000648	01/03/19	M	White	N	N	N	N	N	3	Demo inst	9999_999	Demo Corj	9999	IREAD-3	Undetermined														

## 2. INTERPRETATION OF REPORTED SCORES

A student's performance on a test is reported as a scale score and a performance level for the overall test and as a percentage of correct responses in each reporting category. This section describes how to interpret these scores.

### 2.1 SCALE SCORE

A scale score describes how well a student performed on a test, and can be interpreted as an estimate of a student's knowledge and skills as measured by his or her performance on the assessment. A scale score is the student's overall numeric score. IREAD-3 scale scores are reported on a within-test scale.

Scale scores are used to illustrate students' current level of performance. Lower scale scores can indicate that the student's knowledge and skills fall below proficiency as measured by the assessment. Conversely, higher scale scores can indicate that the student has proficient knowledge and skills as measured by the assessment. When combined across a student population, scale scores can not only describe school- and corporation-level changes in performance but can also 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 (PLDs). 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). Furthermore, the scale score of individual students should be interpreted cautiously when comparing two scale scores because small differences in scores may not reflect real differences in performance.

### 2.2 PERFORMANCE LEVEL

Performance levels are proficiency categories on an assessment that students fall into based on their scale scores. For IREAD-3, scale scores are mapped onto two performance levels (Level 1–Did Not Pass and Level 2–Pass) using performance standards (or cut scores: see Section 2.4, Cut Scores). PLDs are descriptions of content-area knowledge and skills that students at each performance level are expected to possess. Thus, performance levels can be interpreted in relation to PLDs.

PLDs are available on the IDOE web page at <https://www.in.gov/doe/files/IREAD-3-Performance-Level-Descriptors-v2.pdf>.

### 2.3 PERFORMANCE CATEGORY FOR REPORTING CATEGORIES

Students' performance in each reporting category is reported as a percent correct.

## 2.4 CUT SCORES

For all grades and subjects within IREAD-3, scale scores are mapped onto two performance levels (Level 1–Did Not Pass and Level 2–Pass). For each performance level, there is a minimum and maximum scale score that defines the range of scale scores students within each performance level have achieved. Collectively, these minimum and maximum scale scores are defined as cut scores and are the cutoff points for each performance level. Table 2 shows the cut scores for IREAD-3.

*Table 2: IREAD-3 Assessment Proficiency Cut Scores (Grade 2)*

Tested Grades	Level 1 At Risk	Level 2 On Track	Level 3 Pass
Grade 2	200–404	405–445	446–650

*Table 3: IREAD-3 Assessment Proficiency Cut Scores*

Tested Grades	Level 1 Did Not Pass	Level 2 Pass
Grades 3–5	200–445	446–650

## 2.5 LEXILES

The Lexile®<sup>1</sup> framework uses quantitative methods based on individual words and sentence lengths, rather than qualitative analysis of content, to produce scores. A Lexile® Measure for Reading is calculated on IREAD-3 assessments to provide an index of how well a student reads and understands the texts presented. This score can then be matched with a similar Lexile® Text Measure to help educators identify reading materials that best enhance instruction for each student. The Lexile® Text Measure is obtained by evaluating the readability of a piece of text, such as a book or an article.

For the Spring 2022 and Summer 2022 test administrations, Lexile® measures were only reported for IREAD-3 students who were classified into the *Did Not Pass* achievement level.

## 2.6 APPROPRIATE USES FOR SCORES AND REPORTS

Assessment results can provide information on individual students' performance on the test. Overall, assessment results demonstrate what students know and can do in certain subject

<sup>1</sup> Lexile® measures are the intellectual property of Metametrics, Inc.

areas and indicate whether students are on track to demonstrate the knowledge and skills necessary for college and career readiness. Additionally, assessment results can identify students' relative strengths and weaknesses in certain content areas. For example, reporting-category scores can be used to identify an individual student's relative strengths and weaknesses in reporting categories for a content area.

Although assessment results provide valuable information for understanding students' performance, these scores and reports should be used with caution. Scale scores are *estimates* of true scores and hence do not represent a precise measurement of student performance. A student's scale score is associated with measurement error; thus, users need to consider measurement error when using student scores to make decisions about student performance. Moreover, although student scores may be used to help make important decisions about students' placement and retention or teachers' instructional planning and implementation, assessment results should not be used as the only source of information for such judgments. Given that assessment results provide limited information, other sources on student performance—such as classroom assessment and teacher evaluation—should be considered when making decisions on student learning. Finally, when student performance is compared across groups, users need to consider the group size. The smaller the group, the larger the measurement error related to these aggregate data will be; thus, the data require interpretation with more caution.

### **3. SUMMARY**

The IREAD-3 results were reported online via the Online Reporting System (ORS). Reported scale scores were mapped onto two performance levels: Level 1–Did Not Pass (scale scores 200–445) and Level 2–Pass (scale scores 446–650). The results were released in real time during the test window beginning within the three weeks following the start of the respective test windows.

The reporting system is interactive. When educators or administrators log in, they can select “Reports & Files” to navigate to the page that allows them to generate a student data file or an individual student report (ISR) for students for whom they are responsible (e.g., a principal would see the students in his or her school, a teacher would see the students in his or her class). ISRs can be produced as individual PDF files or batched reports.

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



***IREAD-3***

**Indiana Reading Evaluation  
and Determination**

**2021–2022**

**Volume 6  
IREAD-3 Grade 2 Standard Setting**

## **ACKNOWLEDGMENTS**

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

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

## TABLE OF CONTENTS

1. EXECUTIVE SUMMARY.....	3
2. INTRODUCTION.....	6
2.1 PERFORMANCE STANDARDS AND VALIDITY OF TEST SCORE INTERPRETATIONS .....	6
3. OVERVIEW OF THE STANDARDS-CONFIRMATION APPROACH.....	9
3.1 WORKSHOP PANELISTS.....	9
3.2 WORKSHOP MATERIALS .....	9
4. WORKSHOP ACTIVITIES.....	12
4.1 WORKSHOP LOCATION .....	12
4.2 WORKSHOP STAFFING .....	12
4.3 ORIENTATION.....	12
4.4 REVIEW THE PLDS .....	13
4.5 REVIEW THE OIB .....	13
4.6 JUDGMENT TASK.....	13
5. OUTCOMES.....	15
5.1 EVALUATION .....	15
6. REFERENCES.....	16

## LIST OF TABLES

Table 1: Staffing Summary .....	12
---------------------------------	----

## LIST OF APPENDICES

Appendix A: IREAD-3 Grade 2 Cut Score Analysis
Appendix B: Composition of the Panel
Appendix C: Performance Level Descriptors
Appendix D: Ordered Item Booklet
Appendix E: Workshop Presentation Slides
Appendix F: Workshop Evaluation Results

## 1. EXECUTIVE SUMMARY

Beginning in the 2021–2022 school year, the Indiana Department of Education (IDOE) began administering Indiana Reading Evaluation and Determination (IREAD-3) to grade 2 students to provide an early indicator of reading proficiency. IDOE, with the guidance and support from the Technical Advisory Committee (TAC), analyzed the current IREAD-3 assessment to ensure it would serve this new purpose. To provide schools with information about student reading ability at grade 2, a new cut score would need to be developed to indicate when a grade 2 student is “On Track” to reach proficiency in grade 3. TAC recommended deriving the “On Track” cut statistically by using empirical data from a representative sample of grade 2 students taking IREAD-3 in spring 2022, followed by a standard confirmation meeting by content experts. The statistical analysis for deriving the cut was conducted following the conclusion of the first IREAD-3 test administration to a sample of grade 2 students in Spring 2022. The methodology and analysis of the statistical prediction of the cut score are detailed in Appendix A. The standard confirmation meeting included a policy standard-setting workshop in June 2022 and a content standard-setting workshop in July 2022 to recommend an “On Track” performance standard to the Indiana State Board of Education.

In the June 2022 policy workshop, panelists considered policy performance level descriptors (PLDs) for the current IREAD-3 assessment (delivered at grade 3) and created new policy performance level descriptors to describe different performance levels at grade 2. The committee kept the “Pass” policy PLD the same for grade 2 as for grade 3 but created two new descriptors: “On Track” and “At Risk”.

Pass	Students demonstrate proficient understanding when reading and comprehending literary and informational texts. Students identify and comprehend most new variations of word meaning and new text-based vocabulary. Students who Pass have demonstrated the foundational reading skills required by the end of grade three.
On Track	Students demonstrate expected understanding when reading and comprehending literary and informational texts. Students have a basic understanding of variations of word meanings and new text-based vocabulary. Students who are On Track require continued grade-level instruction in foundational reading skills, comprehension, and vocabulary in order to achieve the foundational reading skills required by the end of grade three.
At Risk	Students demonstrate limited understanding when reading and comprehending literary and informational texts. Students have a minimal understanding of word meaning and new text-based vocabulary. Students who are At Risk require additional remediation efforts and targeted instruction in foundational reading skills, comprehension, and vocabulary in order to achieve the foundational reading skills required by the end of grade three.

In the July 2022 workshop, panelists followed a modified bookmark method to recommend an “On Track” performance standard for grade 2 students. It was considered a modified bookmark method in a sense that panelists were provided a proposed cut along with a permissible range and were asked to affirm or move it within the permissible range. The reasoning for this is that the intended use of the grade 2 cut score is to inform educators about students' need for additional foundational reading instruction during grade 3. To avoid under-identifying students who may need support, a conservative approach was used by adding 1 or 2 standard errors (SEs) to the equipercentile indicator. This formed a basis for the statistically-derived cut along with a permissible range. Specifically, the proposed cut based on the statistical process was derived by adding 1 standard error to the equipercentile indicator, while the lower and upper limits of a permissible range were calculated by adding 1 and 2 standard errors to the equipercentile indicator, respectively. IDOE recruited 10 Indiana educators to serve on the panel. The workshop was conducted remotely. Panelists worked as a group to review Performance-Level Descriptors (PLDs) and an ordered-item booklet (OIB), composed primarily of items administered in the Spring 2022 test administration and ordered by item difficulty.

Each PLD is a summary of what students within each performance level are expected to know and be able to do. IDOE facilitators instructed panelists to use the PLDs to develop a mental representation of students at each level. After reviewing PLDs, panelists reviewed the OIB, a set of test items ordered from easiest to most difficult. The OIB for the standard-setting workshop was based primarily on the operational test form administered to Indiana students in Spring 2022. The OIB was augmented with other items in the IREAD-3 bank to minimize gaps in test information, especially for the difficulty range at or below the upper limit of the permissible range of the proposed “On Track” cut. All items were developed for the IREAD-3 assessment following Indiana item specifications and item review procedures.

The panelists reviewed the OIB within CAI’s standard-setting tool. Panelists were asked to consider the following two questions as they reviewed items in the OIB:

1. What do students need to know and be able to do to achieve this score point on this item?
2. Why is this page in the OIB more difficult than the previous pages?

After reviewing the PLDs and the OIB, panelists were prepared to perform the standard-setting task.

The basic question of the standard-setting task was whether the location of the proposed “On Track” performance standard accurately classifies students into each of the IREAD-3 grade 2 performance levels. Please refer to Appendix A for details of the cut score prediction study. Panelists were asked to judge whether the proposed “On Track” cut score reflected the content expectations delineated by the PLDs. There were two possible outcomes of the panelists’ deliberations:

1. Affirm: The proposed performance standard (page 12, scale score 390) accurately reflects the content expectations delineated by the PLDs and classifies students as belonging in the “On Track” performance level.
2. Move: A new location, within a permissible range (up to page 15, scale score 410), more accurately reflects the content expectations delineated by the PLDs and classifies students as belonging in the “On Track” performance level.

The workshop facilitators worked with the panelists with these two goals in mind: (1) Each panelist shares reasoning behind their recommendation (Affirm or Move) and (2) general consensus is expected after discussion, but majority decision may be accepted. At the end of the workshop, each panelist recommended a new location that was believed to more accurately reflect the content expectations delineated by the “On Track” PLD. All the recommendations fell within the permissible range, and the median recommended cut (page 14, scale score 405) was taken as the final recommendation.

## **2. INTRODUCTION**

IREAD-3 was first administered to students during Spring 2012 in accordance with House Enrolled Act 1367. The IREAD-3 assessment was constructed to measure foundational reading standards through grade 3. The new Indiana Academic Standards (IAS) in English/Language Arts (ELA) were adopted for IREAD-3 in 2014. The IAS are designed to help ensure that students are college and career ready by the end of high school. IREAD-3 assessments do not measure all IAS for ELA, but rather the standards most relevant to foundational reading proficiency.

Beginning in the 2021–2022 school year, IDOE began administering IREAD-3 to grade 2 students to provide an early indicator of reading proficiency. IDOE, with the guidance and support from the Technical Advisory Committee (TAC), analyzed the current IREAD-3 assessment to ensure that it would serve this new purpose. To provide schools with information about student reading ability at grade 2, a new cut score would need to be developed to indicate when a grade 2 student is “On Track” to reach proficiency in grade 3. TAC recommended deriving the “On Track” cut statistically by using empirical data from a representative sample of grade 2 students taking IREAD-3 in Spring 2022, followed by a standards-confirmation meeting by content experts. The statistical methodology and analysis for deriving the cut along with a permissible range are detailed in Appendix A. As a brief summary, the statistical process involved two steps. First, an equipercentile indicator for grade 2 was derived to represent the same percentile as the “Pass” cut for grade 3. Second, to avoid under-identifying students who may need support, a conservative approach was used by adding 1 or 2 standard errors (SEs) to the equipercentile indicator. This formed a basis for the statistically-derived cut along with a permissible range. Specifically, the proposed cut based on the statistical process was derived by adding 1 standard error to the equipercentile indicator, while the lower and upper limits of a permissible range were calculated by adding 1 and 2 standard errors to the equipercentile indicator, respectively.

This report describes the standard-setting workshop that was implemented to affirm or move the statistically derived cut within the permissible range.

### **2.1 PERFORMANCE STANDARDS AND VALIDITY OF TEST SCORE INTERPRETATIONS**

Validity refers to the degree to which test score interpretations are supported by evidence and speaks directly to the legitimate uses of test scores. Establishing the validity of test score interpretations is thus the most fundamental component of test design and evaluation. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) provide a framework for evaluating whether claims based on test score interpretations are supported by evidence. Within this

framework, the Standards describe the range of evidence that may be used to support the validity of test score interpretations.<sup>1</sup>

The kinds of evidence required to support the validity of test score interpretations depend centrally on the claims made for how test scores may be interpreted. Moreover, the Standards make explicit that validity is not an attribute of tests, but rather of test score interpretations. Some test score interpretations may be supported by validity evidence, while others are not. Thus, the test itself is not considered valid, but rather the validity of the intended interpretation and use of test scores is evaluated.

Central to evaluating the validity of test score interpretations is determining whether the test measures the intended construct. Such an evaluation in turn requires a clear definition of the measurement construct. For IREAD-3, the definition of the measurement construct is provided by the IAS. The IAS specify what students should know and be able to do by the end of each grade level in order for students to graduate ready for post-secondary education or entry into the workforce. IREAD-3 assessments do not measure all of the IAS for ELA, but rather the standards most relevant to foundational reading proficiency. Because directly measuring student achievement against each standard in the IAS would result in an impractically long test, each test administration is designed to measure a representative sample of the content domain defined by the IREAD-3 test blueprint. To ensure that each student is assessed on the intended breadth and depth of the standards, test form construction is guided by a set of test specifications, or blueprints, which indicate the number of items that should be sampled from each content strand, standard, and benchmark. Thus, the test blueprints represent a policy statement about the relative importance of content strands and standards in addition to meeting important measurement goals (e.g., sufficient items to report strand performance levels reliably). Because the test blueprint determines how student achievement of the IAS is evaluated, alignment of test blueprints with the content standards is critical. IDOE has published the IREAD-3 test blueprint that specifies the distribution of items across reporting strands.

Alignment of test content to the IAS<sup>2</sup> ensures that test scores can serve as valid indicators of the degree to which students have achieved the learning expectations detailed in the IAS. However, the interpretation of the IREAD-3 test scores rests fundamentally on how test scores relate to performance standards which define the extent to which students

---

<sup>1</sup> Responsive to Standards for Educational and Psychological Testing: Standard 9.13

<sup>2</sup> Responsive to Standards for Education and Psychological Testing: Standards 12.8 and 12.10

have achieved the expectations defined in the standards. IREAD-3 test scores are reported with respect to two proficiency levels in grade 3 and three proficiency levels in grade 2, demarcating the degree to which Indiana students have achieved the learning expectations defined by the IAS. The cut score establishing the “On Track” level of performance in grade 2 is critical, since it indicates that students demonstrate expected understanding when reading and comprehending literary and informational texts for grade 2 students and that they require continued grade-level instruction in foundational reading skills, comprehension, and vocabulary in order to achieve the foundational reading skills required by the end of grade 3. Procedures used to adopt the “On Track” performance standard for the IREAD-3 grade 2 are therefore central to the validity of test score interpretations.

This document describes the procedures that Indiana educators, serving as standard-setting panelists, followed to affirm or move the proposed IREAD-3 “On Track” performance standard for grade 2 students. The workshop employed a modified OIB procedure, similar to the Bookmark method used initially to recommend the IREAD-3 “Pass” performance standard, in which panelists used their expert knowledge of the academic content standards and student achievement to map the PLDs adopted by the Indiana State Board of Education onto an OIB based on the operational test forms administered to students in Spring 2022. It was considered a modified method in a sense that panelists were provided a proposed cut along with a permissible range and were asked to affirm or move it within the permissible range. The reasoning for this is that the intended use of the grade 2 cut score is to inform educators about students' need for additional foundational reading instruction during grade 3.

### **3. OVERVIEW OF THE STANDARDS-CONFIRMATION APPROACH**

#### **3.1 WORKSHOP PANELISTS**

IDOE worked to obtain a broadly representative panel for the grade 2 standard-setting workshop that reflected the teacher population in the state of Indiana. The diverse group of panelists brought a wide range of perspectives and experience to the standard-setting effort, ensuring that the evaluation of the proposed performance standard was thoughtful and representative of broad educational constituencies, and represented the range of expertise and experiences found in the educator population across the state.

Twelve panelists were recruited to inform policy performance level descriptors for the new grade 2 cut score. Panelists included school superintendents, principals, special education directors, and teachers serving a representative sample of Indiana’s diverse student population. The panel included educators from various geographic locations and ethnicities serving students in different socio-economic situations and of diverse demographics. The panel also included representatives from public schools, charter schools, and accredited nonpublic schools.

Ten panelists were recruited to evaluate the proposed “On Track” performance standard for grade 2 students. Although the workshop panel was relatively small, IDOE sought to identify panelists representing the range of educational contexts in the state of Indiana.

Appendix B presents the composition of the standard-setting panel. The table includes a record for each panelist and indicates the district they represented and their gender, ethnicity, and current position and main area of expertise. While it is critically important to include a range of stakeholders in the standards-confirmation process, experience has shown that it is essential for panelists to have direct knowledge of academic standards and student grade-level performance to participate meaningfully in the Bookmark procedure. For this reason, panel participation was restricted to classroom teachers, curriculum specialists, and special education teachers with expertise in ELA curriculum and instruction.

#### **3.2 WORKSHOP MATERIALS**

##### **3.2.1 Performance-Level Descriptors**

PLDs define the content-area knowledge and skills that students at each performance level are expected to demonstrate with respect to the IAS. The panelists based their judgments about the proposed “On Track” performance standard on the PLDs.

Prior to convening the workshop, Indiana educator committee drafted a PLD that described the range of achievement encompassed by the “On Track” performance level on the IREAD-3 test for grade 2 students. The range PLD was designed to be clear, concrete, and reflect Indiana’s expectations for an early indicator of proficiency based on the IAS. Committees of Indiana educators created the policy PLDs. PLDs that were used by panelists in the standards-confirmation workshop are presented in Appendix C.

### **3.2.2 Ordered-Item Booklet**

Following review of PLDs, panelists reviewed the OIB. An OIB is a collection of test items ordered from easiest to most difficult. Each page in the OIB corresponds to a level of achievement on IREAD-3, and panelists used the OIB to evaluate whether the proposed IREAD-3 “On Track” performance standard accurately identifies the minimum level of achievement required to qualify for entry into the “On Track” performance level.

#### **Composition of the OIB**

For IREAD-3 tests, all online test takers are administered a test form with a common set of items used for operational scoring. The operational items administered to Indiana students in Spring 2022, the last spring IREAD-3 summative test administration, served as the basis for the OIB. All items used to construct the Spring 2022 test forms were developed specifically for IREAD-3 following IDOE item specifications and item review procedures.

To minimize gaps in the OIB, the OIB was augmented with additional items eligible for operational use to represent the range of item difficulties more fully, especially for the difficulty range at or below the upper limit of the permissible range of the proposed “On Track” cut. Increasing the number of items for the targeted range of item difficulties provides panelists with greater context to identify important shifts in the knowledge and skill requirements of test items. Often panelists become focused on the knowledge and skill requirements of a single item when deliberating on the location of a performance standard. This propensity is exacerbated when there are relatively few items in a given location, which can cause judgments about one item to take on too much importance. Moreover, even though there are sufficient items in summative test forms to establish reliable performance standards for a central proficient performance standard, there are typically fewer items available in locations associated with performance standards categorizing achievement well below and above proficient; thus, in the context of a single operational test form, movement of the performance standard location by even a page or two may result in very large increases or decreases in the percentage of students meeting the standard. Augmenting the OIB moderates the impact associated with each OIB page, especially for performance standards measuring achievement near the tails of the ability distribution.

Items were ordered according to their response probability (RP) level based on their Item Response Theory (IRT) parameters. In IRT, the item characteristic curve for each item indicates the likelihood of responding correctly for each point along the student achievement dimension. The response probability criterion refers to the location on the achievement scale that corresponds to a given probability of success. In the context of the standards-confirmation workshop, this criterion is used to develop a common understanding of what constitutes mastery when evaluating whether a student can respond successfully to an item. An RP value of 0.67 was adopted as the mastery criterion to construct the OIB for the standard-setting workshop. Panelists were asked to consider whether, for example, an “On Track” student has a 0.67 likelihood of answering the item correctly.

The assessment is composed entirely of dichotomously scored (e.g., incorrect vs. correct) items. The items were calibrated using the three-parameter logistic (3PL) model.

The OIB was presented online, allowing panelists to view items in the same context as student test takers. A technical summary of the OIB is presented in Appendix D, 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. The appendix also indicates the existing IREAD-3 “Pass” performance standard associated with each OIB page and the overall percentage of students who would score at or above the standard.

## 4. WORKSHOP ACTIVITIES

### 4.1 WORKSHOP LOCATION

The content-setting workshop was conducted remotely. The workshop facilitator convened the panelists via the Webex online meeting software. All workshop activities employed Internet-based tools that panelists used to review range PLDs and the OIB, and evaluate the performance standards.

### 4.2 WORKSHOP STAFFING

A senior workshop coordinator was tasked with leading the introductory training and was responsible for working with each facilitator and monitoring the flow of activities across workshop. IDOE staff served as the workshop facilitators, leading the panel through training activities and execution of the content-setting process. Because test development staff served as workshop facilitators, they were highly qualified to serve as a subject-matter resource for panelists as they navigated the OIB. In addition, CAI project staff facilitated organization of meeting logistics and provided support to panelists as necessary. Table 1 summarizes workshop staffing.

IDOE staff monitored all standards-confirmation activities, and also addressed any policy or test development questions for panelists.

*Table 1: Staffing Summary*

	Staff	Role
IDOE	Lynn Schemel	Director of Assessment
	Mary Williams	Assistant Director of Assessment
	Alyson Traficante	Program Lead
	Kelly Connelly	Content Expert
CAI	Shuqin Tao	Senior Workshop Lead
	Elizabeth Wei	Senior Workshop Lead
	Maryam Pezeshki	Senior Workshop Lead
	Teresa Hall	Project Team
	Kevin Clayton	Data Analyst
	Christina Sneed	Data Analyst
	Jessica Singh	Data Analyst

### 4.3 ORIENTATION

Upon convening the workshop, the facilitator provided panelists with an introduction to the workshop goals and an overview of the workshop activities. Panelists understood that the overarching goal of the workshop was to begin with a statistically derived cut score

and either confirm that it was supported by the test content or move it to a new location within a permissible range. Presentation slides are provided in Appendix E.

#### **4.4 REVIEW THE PLDs**

Consistent with training in the bookmark method originally used to recommend the IREAD-3 “Pass” performance standard, content setting panelists used the PLD to develop a representation of students in the “On Track” performance level based on the “On Track” performance level descriptor.

#### **4.5 REVIEW THE OIB**

Panelists reviewed the OIB within CAI’s Standard Setting Tool. Panelists were asked to consider two questions as they reviewed item in the OIB.

1. What do students need to know and be able to do to achieve this score point on this item?
2. Why is this page in the OIB more difficult than the previous pages?

#### **4.6 JUDGMENT TASK**

After reviewing the PLDs, developing representations of students who just barely qualify for entry into each of the performance level classifications, and reviewing the OIB, panelists were prepared to perform the judgment task.

The basic question of the judgment task was whether the location of the proposed IREAD-3 “On Track” performance standard accurately classified students into the IREAD-3 “On Track” performance level. Based on the current location of the IREAD-3 “Pass” level performance standard, for example, panelists were asked to judge whether the proposed “On Track” performance standard validly differentiates grade 2 students who just barely qualify for entry into the “On Track” level performance classification from those not yet qualified for entry into the performance level. There were two possible outcomes of the panelists’ deliberations:

1. The proposed performance standard (page 12, scale score 391) accurately reflects the content expectations delineated by the PLDs and classifies students as belonging in the “On Track” performance level.
2. A new location, within a permissible range (up to page 15, scale score 408), more accurately reflects the content expectations delineated by the PLDs and classifies students as belonging in the “On Track” performance level.

If panelists concluded that the proposed performance standard did not validly differentiate students who qualify for entry into the “On Track” performance level based on the PLDs, they were asked to consider whether another location in the OIB would be supported. If the performance standard was seen to validly differentiate students qualifying for entry into the performance level by moving further up the OIB, panelists were asked to

recommend a different page in the OIB (up to page 15, scale score 408) that would support the “On Track” performance level.

## 5. OUTCOMES

The workshop facilitators worked with the panelists with these two goals in mind: (1) Each panelist shares reasoning behind their recommendation (Affirm or Move) and (2) general consensus is expected after discussion, but majority decision may be accepted. At the end of the workshop, each panelist recommended a new location that was believed to more accurately reflect the content expectations delineated by the “On Track” PLD. All the recommendations fell within the permissible range, and the median recommended cut (page 14, scale score 405) was taken as the final recommendation.

### 5.1 EVALUATION

Following the workshop activities, panelists completed a workshop evaluation form. The evaluation form was designed to elicit feedback on all aspects of the workshop, including clarity of training and tasks, appropriateness of the time spent on activities, and satisfaction with the outcome of the workshop. Workshop evaluation results are presented in Appendix F.

The evaluation results show that all panelists agreed or strongly agreed they understood the purpose of the workshop, and the training provided them with the information they needed to complete the tasks. They also indicated that the facilitator(s) guided the conversations appropriately and allowed committee participants to fully contribute. They also indicated that the workshop instructions and materials were clear, and the procedures used to complete the judgment task were fair and unbiased.

## 6. REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). Standards for educational and psychological testing.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Lawrence Erlbaum Associates.
- United Nations (2009). United Nations juridical yearbook 2005. <https://doi.org/10.18356/ae6d26f9-en>