



IREAD-3

**Indiana Reading Evaluation
and Determination**

2018–2019

**Volume 1
Annual Technical Report**

ACKNOWLEDGMENTS

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

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

Table of Contents

1.	INTRODUCTION	1
1.1	Background and Historical Context.....	1
1.2	Purpose and Intended Uses of the IREAD-3 Assessment.....	1
1.3	Participants in the Development and Analysis of IREAD-3	2
1.4	Available Test Formats and Special Versions.....	3
1.5	Student Participation.....	3
2.	SUMMARY OF OPERATIONAL PROCEDURES.....	5
2.1	Administration Procedures.....	5
2.2	Designated Supports and Accommodations	5
3.	ITEM BANK AND TEST CONSTRUCTION	7
3.1	Overview of Item Development.....	7
3.2	Operational Form Construction.....	7
4.	CLASSICAL ANALYSES OVERVIEW	8
4.1	Classical Item Analyses	8
4.1.1	<i>Item Discrimination</i>	8
4.1.2	<i>Distractor Analysis</i>	9
4.1.3	<i>Item Difficulty</i>	9
4.1.4	<i>Mean Total Score</i>	9
4.2	Differential Item Functioning Analysis.....	9
4.3	Classical Analyses Results	12
5.	ITEM RESPONSE THEORY, ITEM CALIBRATION, AND EQUATING	14
5.1	Item Response Theory Models	14
5.2	IRT Summaries.....	15
6.	SCORING AND REPORTING	16
6.1	Maximum Likelihood Estimation	16

6.1.1	Likelihood Function.....	16
6.1.2	Derivatives.....	16
6.1.3	Extreme Case Handling.....	17
6.1.4	Standard Errors of Estimates.....	18
6.2	Transforming Theta Scores to Reporting Scale Scores.....	18
6.3	Overall Performance Classification.....	19
6.4	Reporting Category Scores.....	19
7.	QUALITY CONTROL PROCEDURES.....	20
7.1	Scoring Quality Check.....	20
8.	REFERENCES.....	21

List of Appendices

Appendix A: Operational Item Statistics

Appendix B: Test Characteristic Curves

Appendix C: Distribution of Scale Scores and Standard Errors

Appendix D: Distribution of Reporting Category Scores

LIST OF TABLES

Table 1: Required Uses and Citations of IREAD-3	2
Table 2: Number of Students Participating in IREAD-3 2018–2019	4
Table 3: Distribution of Demographic Characteristics of Tested Population	4
Table 4: IREAD-3 Items by Type	7
Table 5: Thresholds for Flagging Items in Classical Item Analysis	8
Table 6: DIF Classification Rules.....	12
Table 7: Operational Item p-Value Five-Point Summary and Range, Spring 2019....	13
Table 8: Operational Item Parameter Five-Point Summary and Range, Spring 2019.....	15
Table 9: Operational Item Parameter Five-Point Summary and Range, Summer 2019.....	15
Table 10: Theta and Scaled-Score Limits for Extreme Ability Estimates	18
Table 11: Scaling Constants on the Reporting Metric.....	18
Table 12: Proficiency Levels.....	19

1. INTRODUCTION

The Indiana Reading Evaluation and Determination (IREAD-3) 2018–2019 technical report is provided to document and make transparent all methods used in item development, test construction, psychometric methods, standard setting, score reporting methods, summaries of student assessment results, and supporting evidence for intended uses and interpretations of the test scores. The technical report is presented as 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.

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) IREAD-3 were adopted. IREAD-3 assessments do not measure all the IAS in 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 the vendor edCount for the IREAD-3 assessment. The purpose of the study was to review the 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 outcome of the study determined that items aligned to the standards, and the forms aligned to the blueprint.

1.2 PURPOSE AND INTENDED USES OF THE IREAD-3 ASSESSMENT

IREAD-3 is a criterion-referenced assessment that applies principles of evidence-centered design to yield overall and reporting-category-level test scores at the student level and at other levels of aggregation that reflect student performance of 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. An array of reporting metrics allows achievement to be monitored at both student and aggregate levels.

The IREAD-3 assessment draws items from an existing item bank (see Volume 2). AIR inherited the IREAD-3 item bank from Indiana’s previous testing contractor, and no new development was performed. IREAD-3 content standards are aligned with knowledge and skills that 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.

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 Indiana state 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, Indiana 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 academic standards, clarification of how these standards will be assessed, blueprint and test design, and committee reviews of test items and passages.

Technical Advisory Committee

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

American Institutes for Research

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

Human Resources Research Organization

For the 2018–2019 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. More details about accommodations can be found in Volume 3. In addition to the resources available to all students, students with visual impairments could take the braille form. Students with disabilities could take the IREAD-3 with or without accommodations. In addition, a separate form was administered to hard-of-hearing students.

1.5 STUDENT PARTICIPATION

All Indiana public school students in grade 3 were required to participate in the state assessment in spring 2019 as well as grade 4 and 5 students who did not pass that assessment during 2018 or 2017. Students who did not pass the assessment during the previous administration could also retest in summer 2019 unless the student obtains a Good Cause Exemption (GCE). A GCE is an exemption from IREAD-3 for students who did not pass the initial administration and either 1) have previously been retained two times prior to promotion to grade four; 2) have the case conference committee determine that a student with disability promotion is more appropriate; or 3) have a committee determine that an English Learners (ELs) whose Individual Learning Plan (ILP) promotion is more appropriate. Table 2 shows the number of students assessed and the number of students reported for IREAD-3 by administration. Table 3 presents the distribution of students by counts and percentages by administration. The subgroup categories reported here are gender, ethnicity, students with special education (SPED), English Learners, and Section 504.

Table 2: Number of Students Participating in IREAD-3 2018–2019

Admin	Number Tested	Number Reported
Spring 2019	86,006	85,881
Summer 2019 Retest	12,682	12,613

Table 3: Distribution of Demographic Characteristics of Tested Population

Admin	Group	All Students	Male	Female	White	Black/ African American	Asian	Hispanic	American Indian/ Alaska Native	Native Hawaiian/ Other Pacific Islander	Multiracial/ Two or More Races	Special Education	English Learner	Section 504
Spring 2019	N	86,006	44,185	41,821	55,729	11,923	2,283	11,134	134	75	4,728	13,678	7,899	1,711
	%	100.00	51.37	48.63	64.80	13.86	2.65	12.95	0.16	0.09	5.50	15.90	9.18	1.99
Summer 2019 Retest	N	12,682	6,998	5,684	5,614	3,569	197	2,552	21	8	721	3,908	2,026	426
	%	100.00	55.18	44.82	44.27	28.14	1.55	20.12	0.17	0.06	5.69	30.82	15.98	3.36

2. SUMMARY OF OPERATIONAL PROCEDURES

2.1 ADMINISTRATION PROCEDURES

The IREAD-3 assessment for spring 2019 was administered to eligible students during an early window (authorization by IDOE was required) from March 11–15, 2019, and during the official window from March 18–29, 2019. The summer 2019 retest was available May 28 to July 19, 2019, to students who did not pass the spring 2019 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) who proctored the assessment. Test administration manuals were provided so that personnel involved with statewide assessment administrations could maintain both standardized administration conditions and test security.

A secure browser developed by AIR was required to access the online 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

Accessibility supports discussed in this document included both embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features that were available to all students as they accessed instructional or assessment content, designated features that were available to students for whom an informed educator or team of educators had identified the need, and accommodations that were available for students for whom there was documentation on an Individualized Education Program (IEP), Section 504 Plan, or Individual Language Plan (ILP).

Educators making these decisions were trained on the process and understood the range of designated supports available.

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

disability-related, construct-irrelevant impediments to performance” (Koretz & Hamilton, 2006, p. 562).

The TAs and STCs in Indiana 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. Detailed descriptions for each of these accommodations can be found in Appendix J of Volume 5.

3. ITEM BANK AND TEST CONSTRUCTION

3.1 OVERVIEW OF ITEM DEVELOPMENT

Operational items used on the IREAD-3 test forms were drawn from the previously established IREAD-3 item bank. Volume 2 is a separate, stand-alone report containing details on the IREAD-3 item bank.

3.2 OPERATIONAL FORM CONSTRUCTION

Operational test forms (see Volume 2) include multiple-choice (MC) and multi-select (MS) item types to measure the 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.

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

Table 4: IREAD-3 Items by Type

Response Type	Description	Spring 2019	Summer 2019 Retest
MC	Student selects one correct answer from a number of options.	37	37
MS	Student selects all correct answers from a number of options.	2	2

4. CLASSICAL ANALYSES OVERVIEW

4.1 CLASSICAL ITEM ANALYSES

AIR psychometricians monitor the behavior of items while test forms are administered in a live environment. This is accomplished using AIR's Quality Monitor (QM) system, which yields an item-analysis report on the performance of test items throughout the testing window. During administration of the 2018–2019 IREAD-3, 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 could be produced to flag only items with statistics falling outside a specified range or to generate reports based on all items in the pool. The criteria for flagging and reviewing items are 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) or multi-point 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.
Item Difficulty (non-MC items)	Proportion of students receiving any single score point is greater than 0.95 for non-MC items.
Inverted Mean Total Score	Mean total score for a lower score point exceeds the mean total score for a higher score point for multi-point items.

4.1.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiates between those examinees who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for 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. 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 was used to identify items that may have had marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracted high-scoring students. For MC items, the correct response should have been the most frequently selected option by high-scoring students. The discrimination value of the correct response should have been substantial and positive, and the discrimination values for distractors should have been lower and, generally, negative. All operational items had a negative distractor. 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. For constructed-response items, item difficulty was calculated using the item's relative mean score and the average proportion correct (analogous to p -value and indicating the ratio of the item's mean score divided by the maximum possible score points). Conventional item p -values are summarized in Section 4.3. The p -values and number of flagged items for operational items can be found in Appendix A. Most of the operational items had p -values within the expected range. One spring 2019 IREAD-3 operational item was flagged. The flagged item was verified by AIR content experts and psychometricians reported that the item behaved as expected.

4.1.4 Mean Total Score

For multi-point items, mean total score was calculated using the item's relative mean score and the average proportion correct (analogous to p -value and indicating the ratio of the item's mean score divided by the maximum possible score points). Items were flagged when the proportion of students in any score point category was greater than 0.95. In addition, multi-point items are flagged if the average ability estimate of students in a score-point category is lower than the average ability estimate of students in the next lower score-point category. For example, if students who receive three points on a multi-point item score lower, on average, on the total test than students who received only two points on the item, the item will be flagged for review. The p -values and number of flagged items for operational items can be found in Appendix A. All of the multi-point operational items had p -values following the expected mean total score. No IREAD-3 operational items were flagged for mean total score.

4.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014) provides a guideline for when sample sizes

permitting subgroup differences in performance should be examined and appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors.

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

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

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF was important because it provided a statistical indicator that an item may contain cultural or other bias. DIF -flagged items were further examined by content experts, who were asked to re-examine each flagged item to decide whether the item should have been excluded from the pool due to bias. Not all items that exhibit DIF are biased; characteristics of the education system may also lead to DIF. For example, if schools in certain areas 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, the instruction, not the item, exhibits bias. However, DIF can indicate bias, so all items were evaluated for DIF.

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

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

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

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

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

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

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

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

The MH conditional odds ratio is calculated as

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

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

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

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

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

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

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

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

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, or female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., White or male). If the DIF statistics fell into the “C” category for any group, the item showed significant DIF and was reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts reviewed all items flagged based on DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance.

Table 6: DIF Classification Rules

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

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.

4.3 CLASSICAL ANALYSES RESULTS

This section presents a summary of results from the classical item analysis for the 2019 IREAD-3 spring operational items. The summaries here are aggregates; item-specific details are found in Appendix A.

Table 7 provides summaries of the p -values by percentile and range by 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	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Spring 2019	0.46	0.60	0.76	0.82	0.90	0.96	0.98
Summer 2019	0.35	0.37	0.53	0.62	0.72	0.82	0.83

DIF summary tables based on Indiana students can be found in Appendix A. Across all operational items and DIF comparison groups, less than 16% of spring 2019 IREAD-3 items were classified as C DIF. AIR content specialists and psychometricians reviewed flagged items to ensure that they were free of bias. The review of the flagged items did not produce any serious issues with items.

5. ITEM RESPONSE THEORY, ITEM CALIBRATION, AND EQUATING

IRT (van der Linden & Hambleton, 1997) is used to calibrate all items and derive scores for all 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.

AIR used previously established item parameters to score the IREAD-3 assessments in spring and summer 2019.

5.1 ITEM RESPONSE THEORY MODELS

IREAD-3 employed IRT models for item calibration and student ability estimation. The IREAD-3 assessment is made up of 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:

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

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

Generalized Partial Credit Model

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

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

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

score for the person j , k indexes step of the item i , and $b_{i,k}$ is the k^{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 2019

Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
a	0.39	0.42	0.84	1.06	1.46	1.71	1.92
b	-3.89	-3.12	-1.74	-1.27	-0.86	-0.12	0.51
c	0.02	0.03	0.08	0.15	0.20	0.36	0.42

Table 9: Operational Item Parameter Five-Point Summary and Range, Summer 2019

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 the test characteristic curves (TCCs). These plots are displayed in Appendix B.

6. SCORING AND REPORTING

6.1 MAXIMUM LIKELIHOOD ESTIMATION

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

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

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

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

$$q_i = 1 - p_i$$

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., 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

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

and where θ_t denotes the estimated θ at iteration t . N_{CR} is the number of items that are scored using the 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 the 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 of Theta (LOT)	Highest of Theta (HOT)	Lowest of Scale Score (LOSS)	Highest of 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

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

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

For spring 2019, 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 IRT ability estimate:

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

The summary of IREAD-3 scale scores for each administration is provided in Appendix D.

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 provides the scale score range for performance standards for IREAD-3. The lower bound of the level 2, Pass, marks the minimum cut score for proficiency.

Table 12: Proficiency Levels

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

6.4 REPORTING CATEGORY SCORES

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

- Reading Foundations and Vocabulary
- Nonfiction
- Literature

7. QUALITY CONTROL PROCEDURES

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

7.1 SCORING QUALITY CHECK

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

Additionally, HumRRO provided replication of the psychometric scoring process for IREAD-3. IDOE approved and published scores only when all three independent systems matched.

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**

2018–2019

**Volume 2
Test Development**

ACKNOWLEDGMENTS

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

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

TABLE OF CONTENTS

1.	INTRODUCTION	1
1.1	Claim Structure	1
1.2	Underlying Principles Guiding Development	1
1.3	Organization of this Volume	2
2.	IREAD-3 ITEM BANK SUMMARY	3
3.	ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS	4
3.1	Overview	4
3.2	Item Specifications	4
4.	IREAD-3 BLUEPRINTS AND TEST CONSTRUCTION	5
4.1	IREAD-3 Blueprints	5
4.2	Test Form Construction.....	5
5.	REFERENCES	7

LIST OF TABLES

Table 1: Item Types and Descriptions	3
Table 2: Blueprint Percentage of Test Items Assessing Each Reporting Category	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 basic reading skills and reading comprehension based on the Indiana Academic Standards (IAS). The Indiana State Board of Education approved the IAS in April 2014 for English/Language Arts (ELA). The IAS are intended to implement more rigorous standards with the goal of challenging and motivating Indiana’s students to acquire stronger critical thinking, problem solving, and communications skills promoting college-and-career readiness.

1.1 CLAIM STRUCTURE

The IREAD-3 assessment was designed to measure foundational reading standards through grade 3. 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 identify and comprehend most new variations of word meaning and new text-based vocabulary. Examples of specific knowledge, skills, and abilities for grade 3 students scoring at the Pass level may include the following:

- Identify 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 among 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

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, described the interaction types that could be used, provided guidelines for targeting the appropriate cognitive engagement, offered suggestions for controlling item difficulty, and offered sample items.

Items were written with the goal that virtually every item would be accessible to all students, either by itself or in conjunction with accessibility tools, such as text-to-speech 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. The details of these processes follow.

1.3 ORGANIZATION OF THIS VOLUME

This volume is organized in three sections:

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

2. IREAD-3 ITEM BANK SUMMARY

As described above, all items used on the IREAD-3 assessment are aligned to the IAS. AIR inherited the IREAD-3 item bank from Indiana’s previous testing contractor, and no new development was performed.

Table 1 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 1: Item Types and Descriptions

Response Type	Description
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.

3. ITEM DEVELOPMENT PROCESS THAT SUPPORTS VALIDITY OF CLAIMS

3.1 OVERVIEW

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

The process begins with the definition of passage and item specifications, and continues with

- selection and training of item writers;
- writing and internal review of items; and
- review by state personnel and stakeholder committees.

Each of these steps has a role in ensuring that the items can support the claims on which they will be based. More information about the item development process can be found in the IREAD-3 Spring 2018 technical report.

3.2 ITEM SPECIFICATIONS

IREAD-3 item specifications, given in Appendix C, were created by a previous Indiana vendor in summer 2015. The item specifications also went through item development and committee review. Item specifications guided the item development process for all IREAD-3 items.

The IREAD-3 item specifications include the following:

- **Content Standard.** This identifies the standard being assessed.
- **Evidence Statement.** Statements that describe the knowledge and skills that an assessment item should elicit from students.
- **Content Limits/Constraints.** This section delineates the specific content that the standard measures and the parameters in which items must be developed to assess the standard accurately, including the lower and upper complexity limits of items.
- **Depth of Knowledge Demands.** All IREAD-3 item specifications have a Depth of Knowledge (DOK) value 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. Notes delineating the cognitive demands of the item and an explanation of its difficulty level are detailed for each sample item.

4. IREAD-3 BLUEPRINTS AND TEST CONSTRUCTION

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

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

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

4.1 IREAD-3 BLUEPRINTS

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 2: Blueprint Percentage of Test Items Assessing Each Reporting Category

Reporting Category	Reading Foundations and Vocabulary	Nonfiction	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, ensuring that the form contains enough items at that 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.

4.2 TEST FORM CONSTRUCTION

At the start of the IREAD-3 contract, AIR was provided with a set of pre-built fixed forms to be delivered for the Spring 2019 and Summer 2019 administrations. 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 items that are read aloud by the test administrator to students and stand-alone multiple-choice and multi-part multiple-choice items. Segments two and three consist of multiple-choice and multi-part multiple-choice items that are linked to reading passages.

As noted above, segment one on the IREAD-3 assessment contains four items and a sample item that are read aloud to students. For students with a hard-of-hearing designation, any accommodation will invalidate the construct measured by these items. Thus, these four items are not administered to students with the hard-of-hearing accommodation. That means that the students with the accommodation will have four fewer operational items and four fewer score points than all other students.

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**

2018-2019

**Volume 3
Test Administration**

ACKNOWLEDGMENTS

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

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

TABLE OF CONTENTS

1. INTRODUCTION.....	1
2. TESTING PROCEDURES AND TESTING WINDOWS	3
2.1 Eligible Students	4
2.2 Testing Accommodations	5
3. ADMINISTRATOR TRAINING	9
3.1 Online Administration	9
3.2 Test Administration Resources.....	11
4. TEST SECURITY PROCEDURES	14
4.1 Security of Test Materials.....	14
4.2 Investigating Test Irregularities.....	15
4.3 Tracking and Resolving Test Irregularities	16
4.4 AIR’s System Security	17
REFERENCES	19

LIST OF TABLES

Table 1: Designated Features and Accommodations Available in Spring 2019	6
Table 2: User Guides and Manuals.....	12
Table 3: Examples of Test Irregularities and Test Security Violations	16

LIST OF APPENDICES

- Appendix A: *Online Test Delivery System (TDS) User Guide*
- Appendix B: *Technology Setup for Online Testing Quick Guide*
- Appendix C: *2018-2019 Additional Configurations and Troubleshooting Guide for Windows, Mac, Chrome OS, and Linux*
- Appendix D: *Indiana Online Practice Test User Guide*
- Appendix E: *Test Information Distribution Engine (TIDE) User Guide*
- Appendix F: *Braille Requirements Manual for Online Testing*
- Appendix G: *Online Reporting System User Guide*

Appendix H: *IREAD-3 Spring 2019 Test Administrators Manual (TAM)*
Appendix I: *IREAD-3 Summer 2019 Test Administrators Manual (TAM)*
Appendix J: *Accessibility and Accommodations Implementation and Setup Module*
Appendix K: *Indiana Assessments Policy Manual*
Appendix L: *Released Item Repository Quick Guide*
Appendix M: *Why it is Important to Assess Webinar Module*
Appendix N: *Test Administrator Training Webinar Module*
Appendix O: *Test Administration Overview Webinar Module*
Appendix P: *Test Information Distribution Engine (TIDE) Webinar Module*
Appendix Q: *Test Delivery System (TDS) Webinar Module*
Appendix R: *Online Reporting System (ORS) Webinar Module*
Appendix S: *Technology Requirements for Online Testing Webinar Module*
Appendix T: *Indiana Accessibility and Accommodations Guidance Manual*
Appendix U: *IREAD-3 ISR Interpretive Guide*

1. INTRODUCTION

In spring 2019, pursuant to House Enrolled Act 1367 (also known as Public Law 109) the Indiana Reading Evaluation and Determination (IREAD-3) test was administered to Indiana students in Grade 3. Students in grades 4 and 5 who had not previously passed the IREAD-3 assessment were given the opportunity to retest. Students who did not pass the assessment during the previous administration could also retest in summer 2019 unless the student obtained a Good Cause Exemption (GCE). A GCE is an exemption from IREAD-3 for students who did not pass the initial administration and either 1) have previously been retained two times prior to promotion to grade four; 2) have the case conference committee determine that a student with disability promotion is more appropriate; or 3) have a committee determine that an English Learners (ELs) whose Individual Learning Plan (ILP) promotion is more appropriate.

In spring 2019, IREAD-3 was administered in AIR's Test Delivery System (TDS) under one test ID with three segments. Test Administrator approval was required for a student to advance to each segment.

In summer 2019, IREAD-3 was administered in TDS as three segments with each segment under a separate test ID. TAs assigned each segment separately to students.

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 the item 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 test administrators 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 AIR'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 2018-2019 IREAD-3 assessments required coordination, detailed specifications, and proper training. In addition to these efforts, several individuals were involved in the administration process, from those setting up secure testing environments to those administering the tests. Without the proper training and coordination of these individuals, the standardization of the test administration could have been compromised. IDOE worked with AIR to develop and provide the training and documentation necessary for the administration of IREAD-3 under standardized conditions within all testing environments, both online and on paper-pencil tests.

All students were required to take a practice test conducted at their school prior to taking the IREAD-3 assessments. The practice test sessions contained sample test items that reflected similar test items that the student encountered on the IREAD-3 assessments and helped students become familiar with TDS functionality and item types. 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 [IREAD-3 portal](#). The IREAD-3 RIR was deployed in October 2018 which resulted in students having access to the items online five months prior to the opening of the assessment window in March 2019.

The spring IREAD-3 assessment was administered as one test with three segments. The Summer IREAD-3 test was also comprised of three segments, but each segment was administered under a separate test ID within TDS and combined into a single test for scoring. Schools had the flexibility to test over the span of 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 was a timed assessment with each of the three segments lasting between 30-35 minutes. The spring IREAD-3 testing window was March 18-29, 2019. The summer IREAD-3 testing window was May 28-July 19, 2019.

2.1 ELIGIBLE STUDENTS

Students in grade 3 were required to take IREAD-3 in spring 2019 with or without accommodations if provided by their Individual Education Plan, Section 504 Plan, or ILP and also including students who have been retained twice. Students who did not pass IREAD-3 in spring 2019 could take the Summer IREAD-3 re-test or could take the IREAD-3 assessment in grades 4 and 5 if a passing score was not achieved.

The IREAD-3 assessment measures foundational reading standards to grade 3 students each spring. Based on the Indiana Academic Standards, IREAD-3 is a summative assessment that was developed in accordance with House Enrolled Act 1367 (also known as Public Law 109 in 2010)

- **Public and Private School Students:** Indiana public and private 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 district 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. ELs can receive a Good Cause Exemption from IREAD-3 if it has been determined that promotion is appropriate.
- **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. In Indiana, a student on an IEP participates under one of these three general options:
 1. IREAD-3 without accommodations
 2. IREAD-3 with approved accommodations
 3. Indiana Alternate Measure (I AM) Alternate Assessment

Per the Individuals with Disabilities Education Improvement Act (IDEA) and Title 511 Article 7-Special Education, published December 2014 by the Indiana State Board of Education, decisions regarding which assessment option a student will participate in are made annually by the student's IEP team and are based on the student's curriculum, present levels of academic achievement, functional performance, and learning characteristics. Decisions cannot be based on program setting, category of disability, percentage of time in a particular placement or classroom, or any considerations regarding a school's Adequate Yearly Progress (AYP) designation.

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

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

2.2 TESTING ACCOMMODATIONS

Students participating in the online, fixed test form IREAD-3 assessments were able to use the standard online testing features in TDS. These features included the ability to select an alternate background and font color, mouse pointer size and color, and font size before testing. During the tests, students could zoom in and zoom out to increase or decrease the size of text and images, highlight items and passages (or sections of items and passages), cross out response options by using the strikethrough function, use a notepad to make notes, and 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 were provided to students with disabilities enrolled in public schools with current IEPs or Section 504 Plans, as well as to students identified as English Learners (ELs).

The accommodations available for eligible students participating in the IREAD-3 assessments are described in the *Test Administrator Manual (TAMs)* (Appendices H and I of this report volume), which were accessible before and during testing from the [IREAD-3 portal](#).

The IREAD-3 assessments provided two categories of assessment supports to students. These included designated features and accommodations, both embedded (delivered through TDS) and non-embedded. Volume 1 of this technical report lists the allowed accommodations and the number of students who were provided with accommodations during the spring 2019 IREAD-3 test administration.

Table 1 provides a list of designed features and accommodations that were offered during the spring 2019 administration. Designated features for the 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 A of this report volume) provides instructions on how to access and use these features.

Table 1: Designated Features and Accommodations Available in Spring 2019

	Designated Features	Accommodations
Embedded	Color Contrast (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 appropriate allowable accommodations. Computer-based test settings and accommodations were required to be identified in the Test Information Distribution Engine (TIDE) before starting a test session. Some settings and accommodations could not be changed after a student started the test.

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

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 that are being measured by the test
3. Accommodations must be the same or nearly the same as those needed and used by the student in completing daily classroom instruction and routine assessment activities
4. Accommodations must be necessary for enabling the student to demonstrate knowledge, ability, skill, or mastery

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

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

Available Accommodations

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

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

- Contracted UEB braille

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

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 K in this report volume).

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 its educators and key personnel involved with IREAD-3 administrations a clear, standardized procedure for the administration of IREAD-3, including administrations with accommodations. Key personnel involved with IREAD-3 administrations included Corporation Test Coordinators (CTCs), Corporation Information Technology Coordinators (CITC), 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 AIR TA Certification Course before administering the test. There were also several training modules developed by AIR in collaboration with IDOE to assist with test administration. The modules included topics on AIR systems, test administration, and accessibility and accommodations. The modules 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 A in this report volume) 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
- Information on secure browser features and keyboard shortcuts

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

3.1 ONLINE ADMINISTRATION

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

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

Roles and Responsibilities in the Online Testing Systems

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

CTCs

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

CITCs

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

NPSTCs

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

STCs

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

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

Test Administrators

TAs administered IREAD-3 and administered a practice test session prior to student's administration of the 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 ensure 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 2019 IREAD-3 administration is provided below. Training materials were all available online at <https://iread3.portal.airast.org/resources/test-administrators-and-educators/> and are included as appendices to this report volume.

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

The administration resources comprising various tutorials and user guides (user manuals, quick guides, etc.) were available at the IREAD-3 Portal at <https://iread3.portal.airast.org/resources/test-administrators-and-educators/>

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

Resource	Description
<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 (see Appendix A).
<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. (see Appendix B).
<i>2019-2020 Additional Configurations and Troubleshooting Guide for Windows, Mac, Chrome OS, and Linux</i>	This manual provides information about hardware, software, and network configurations for running various testing applications provided by American Institutes for Research (AIR) (see Appendix C).
<i>Indiana Online Practice Test User Guide</i>	This user guide provided an overview of the IREAD-3 Practice Test (see Appendix D).
<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 (see Appendix E).
<i>Braille Requirements Manual for Online Testing</i>	This manual provided an overview of how to ensure your computer devices are set up properly to successfully administer the online Braille assessments for IREAD-3 (see Appendix F).
<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 (see Appendix G).
<i>2018-2019 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 (see Appendix U).

Department Resources and Support

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

- 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 information on 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.

- Communications via email memos took place on an “as needed” basis. These messages generally addressed specific issues that needed to be transmitted quickly to administrators and teachers in the field or important information that the IDOE wanted to ensure was clearly outlined due to its importance to the 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 <https://www.doe.in.gov/assessment>, such as testing windows for all state-administered assessments. The *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 the practice tests was to familiarize students with the system, functionality, and item types that appeared on the IREAD-3 tests. The practice tests were not intended to guide classroom instruction. Users could also use the tutorials on each item and familiarize themselves with the different features and response instructions for each item type.

The IREAD-3 practice tests were deployed on October 1, 2018, and remained available throughout the testing window. Online practice tests were designed for use with the AIR Secure Browser. The portal provided a list of supported web browsers to administer the practice tests. AIR’s TDS delivered the practice tests in secure mode and used the same test delivery engine as the operational test.

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

Students taking the IREAD-3 assessment on paper were also required to take a paper-based practice test prior to taking the operational IREAD-3 assessment. The practice test items were delivered to students at the beginning of the paper-and-pencil assessment booklets. The TA script provided specific instructions to ensure that the students completed the paper practice test items prior to 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
- Guidelines to determine if test invalidation was appropriate/necessary

To support these policies and procedures, IDOE leveraged security measures within AIR systems. For example, students taking the 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 AIR Secure Browser after 20 minutes of inactivity.

In developing the IREAD-3 TAMs (Appendices H and I of this report volume), IDOE and AIR ensured that all test security procedures were available to everyone involved in test administration. Each manual included protocols for reporting any deviations in test administration.

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

- Invalidation of student scores

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 administration, initial or make-up test session, secure materials (e.g., scratch paper) were required to be returned immediately to the STC and placed in locked storage. Secure materials were never to be left unsecured and were not permitted to remain in classrooms or be removed from the school's campus overnight. Secure materials that did not need to be returned to the print vendor for scanning and scoring were allowed to be destroyed securely following outlined security guidelines, but were not allowed to be discarded in the trash. In addition, any monitoring software that might have allowed test

content on student workstations to be viewed or recorded on another computer or device during testing had to be disabled.

It 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
- 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 of any kind.

To access the online IREAD-3 tests, a Secure Browser was required. The AIR Secure Browser provided a secure environment for student testing by disabling hot keys, copy, and screen capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines). Users could not access other applications from within the secure browser, even if they knew the keystroke sequences. Students were not able to print from the secure browsers. 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* in Appendix A for further details.

4.2 INVESTIGATING TEST IRREGULARITIES

AIR's quality monitoring (QM) system gathers data used to detect cheating, monitors real-time item function, and evaluates test integrity. Every completed test runs through the QM system, and any anomalies (such as tests not meeting blueprint, unexpected test lengths, or other unlikely issues) are flagged. AIR psychometricians 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 2019 testing windows. Analyses relying on student ability were not able to be run until the summer after all items were calibrated and placed on the same scale.

AIR psychometricians monitored testing anomalies throughout the testing window. A variety of evidence was collected for the evaluation. These include blueprint match, unusual much longer test times as compared to the state average, and item response patterns using the person-fit index. The flagging criteria used for these analyses are

configurable and were set by IDOE. While analyses used to detect the testing anomalies could be run anytime within the testing window, analyses relying on state averages are typically held until the close of the testing window to ensure final data is being used.

No unexpected results were identified during the 2019 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 2019-2020 school year.

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

Table 3: Examples of Test Irregularities and Test Security Violations

Description
Student(s) making distracting gestures/sounds or talking during the test session that creates a disruption in the test session for other students.
Student(s) leaving the test room without authorization.
TA or Test Coordinator leaving related instructional materials on the walls in the testing room.
Student(s) cheating or providing answers to each other, including passing notes, giving help to other students during testing, or using handheld electronic devices to exchange information.

Student(s) accessing or using unauthorized electronic equipment (e.g., cell phones, smart watches, iPods, or electronic translators) during testing.

Disruptions to a test session such as a fire drill, school-wide power outage, earthquake, or other acts.

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

TA giving incorrect instructions.

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

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

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

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

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

TA 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 AIR'S SYSTEM SECURITY

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

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

AIR's systems implement sophisticated, configurable privacy rules that can limit access to data to only appropriately authorized personnel. AIR maintains logs of key activities

and indicators, including data backup, server response time, user accounts, system events and security, and load test results.

REFERENCES

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



IREAD-3

**Indiana Reading Evaluation
and Determination**

2018–2019

**Volume 4
Evidence of Validity and
Reliability**

TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE	5
1.1 Reliability	5
1.2 Validity.....	8
2. PURPOSE OF IREAD-3.....	11
3. EVIDENCE OF CONTENT VALIDITY	12
3.1 Content Standards	12
4. RELIABILITY	13
4.1 Marginal Reliability	13
4.2 Test Information Curves and Standard Error of Measurement.....	13
4.3 Reliability of Performance Classification	16
4.3.1 Classification Accuracy	17
4.3.2 Classification Consistency.....	18
4.4 Precision at Cut Scores.....	20
5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE	21
5.1 Correlations Among Reporting Category Scores.....	21
5.2 Confirmatory Factor Analysis	22
5.2.1 Factor Analytic Methods.....	22
5.2.2 Results.....	24
5.2.3 Discussion.....	25
5.3 Local Independence	26
5.4 Convergent and Discriminant Validity.....	27
6. FAIRNESS IN CONTENT.....	29
6.1 Statistical Fairness in Item Statistics.....	29
7. SUMMARY	31
8. REFERENCES	32

LIST OF TABLES

Table 1: Number of Items for Each Reporting Category by Administration	12
Table 2: Marginal Reliability Coefficients by Administration.....	13
Table 3: Descriptive Statistics.....	17
Table 4: Classification Accuracy Index	18
Table 5: False Classification Rates	19
Table 6. Classification Accuracy and Consistency	20
Table 7: Performance Levels and Associated Conditional Standard Error of Measurement.....	20
Table 8: Observed Correlation Matrix Among Reporting Categories.....	21
Table 9: Goodness-of-Fit Second-Order CFA	25
Table 10: Correlations Among Factors	25
Table 11: Q ₃ Statistics	27
Table 12: Observed Score Correlations Spring	28
Table 13: Observed Score Correlations Summer	28

LIST OF FIGURES

Figure 1: Sample Test Information Function.....	14
Figure 2: Conditional Standard Error of Measurement (Spring).....	15
Figure 3: Conditional Standard Error of Measurement (Summer)	16
Figure 4: Second-Order Factor Model (ELA)	24

LIST OF APPENDICES

Appendix A: *Reliability Coefficients*

Appendix B: *Conditional Standard Error of Measurement*

ACKNOWLEDGMENTS

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

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

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

IREAD-3 was constructed to measure foundational reading standards grade 3. The Indiana Academic Standards (IAS) in English Language Arts (ELA) are the foundation of IREAD-3. IREAD-3 was first administered to students during the spring of 2012 in accordance with House Enrolled Act 1367. During the 2018-2019 school year there two administrations, a spring administration and a summer retest administration. The main test administration was online with braille and hard of hearing accommodations available. A paper-and-pencil version was also available. Full descriptions of available accommodations are listed in Volume 3, Section 1.2. The number of students who were provided with accommodations is presented in Volume 1, Appendix E.

With the implementation of the IREAD-3 assessment, both reliability evidence and validity evidence were necessary to support appropriate inferences of student academic performance from IREAD-3 scores. This volume provides empirical evidence about the reliability and validity of the 2018–2019 IREAD-3 assessment, given its intended uses.

The purpose of this volume is to provide empirical evidence to support a validity argument regarding the uses and inferences for the IREAD-3 assessment. This volume addresses the following:

- *Reliability.* Marginal reliability estimates for each administration are reported in this volume, the reliability estimates are presented by administration in the main body, and by demographic subgroups in Appendix A. This section 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 the IAS with a sufficient number of items targeting each area of the blueprint.
- *Internal Structure Validity.* Evidence is provided regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and correlations among reporting categories per administration. Confirmatory factor analysis has also been performed using the second-order factor model. Additionally, local item independence, an assumption of unidimensional IRT, was tested using the Q_3 statistic.
- *Test Fairness.* Fairness is statistically analyzed using differential item functioning (DIF) in tandem with content alignment reviews by specialists.

1.1 RELIABILITY

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

consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

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

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

- The *test-retest* method measures stability over time. With this method, the same test is administered twice to the same group at two different points in time. If test scores from the two administrations are highly correlated, then the test scores are deemed to have a high level of stability. For example, if the result is highly stable, those who scored high on the first test administration tend to obtain a high score on the second administration. The critical factor, however, is the time interval. The time interval should not be too long, which could allow for changes in the test takers' true scores. Likewise, it should not be too short, or memory and practice may confound the results. The test-retest method is most effective for measuring constructs that are stable over time, such as intelligence or personality traits.
- 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 the standard errors of measurement (SEMs)—the smaller the standard error, the higher the precision of the test scores. For example, classical test theory assumes that an observed score (X) of 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, we arrive at

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

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

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

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

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

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

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

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

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

In contrast, the SEMs in IRT vary over the ability continuum. These heterogeneous errors are a function of a 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

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 (see 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 (see Volume 2 of this technical report for details). Test scores can be used to support an intended validity claim when they contain minimal construct-irrelevant variance.

For example, a Mathematics item targeting a specific mathematics skill that requires advanced reading proficiency and vocabulary has a high level of construct-irrelevant variance. Thus, the intended construct of measurement is confounded, which impedes the validity of the test scores. Statistical analyses, such as factor analysis or multidimensional scaling, are also used to evaluate content relevance. Results from factor analysis for the 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: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests (see Volume 1, Section 5.2). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (see Section 3, Reliability, and Section 5, Evidence of Internal-External Structure, for details).

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

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

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

test first requires an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

2. PURPOSE OF IREAD-3

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

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

IREAD-3 assessments are criterion-referenced tests designed to measure student performance on the IAS in ELA. As a comparison, norm-referenced tests are designed to compare or rank all students to one another.

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 tailor their instruction, provided that they are viewed with the usual caution that accompanies use of reporting category scores. Thus, we must examine the reliability coefficients for these test scores and the validity of the test scores to support practical use of these tests across the state. Volume 5 of this technical report is the score interpretation guide and provides details on all generated scores and their appropriate uses and limitations.

3. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the IREAD-3 were representative of the content standards of the larger knowledge domain. We describe the content standards for IREAD-3 and discuss 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. The IREAD-3 blueprint is available in Volume 2, Appendix A. Blueprints were developed to ensure that the test and the items were aligned to the prioritized standards that they were intended to measure. 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 the test based on the average conditional standard errors, estimated at different points on the performance scale, for all students. The marginal reliability coefficients are nearly identical or close to the coefficient *alpha*. For our analysis, the marginal reliability coefficients were computed using operational items.

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

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

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

where σ_{θ}^2 is the true score variance of θ and

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

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

Table 2 presents the marginal reliability coefficients by administration.

Table 2: Marginal Reliability Coefficients by Administration

Administration	Marginal Reliability
Spring	0.788
Summer	0.866

4.2 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

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

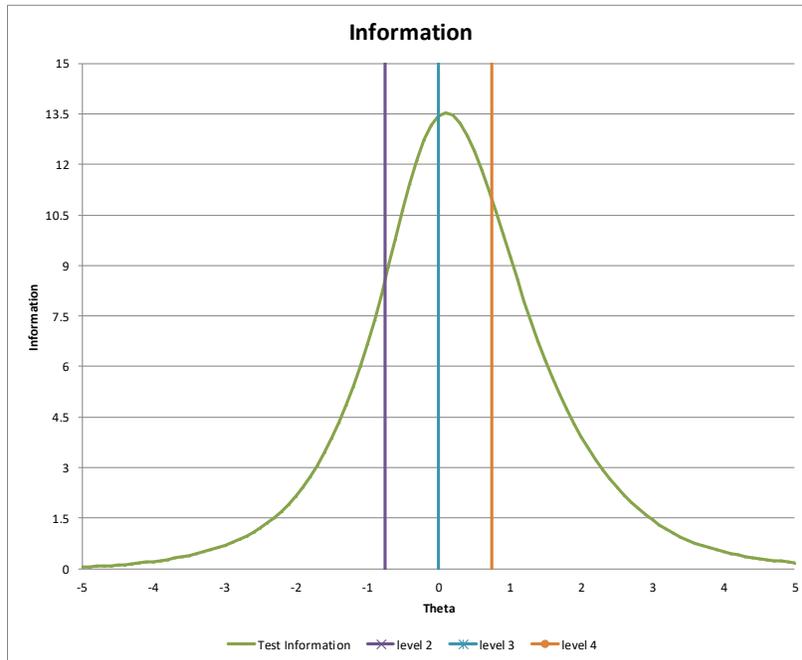
Figure 1 displays a sample TIF with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning it provides the most-precise scores in this range. Where the curve is lower at the tails indicates that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

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

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

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 θ_s is the ability of student s .

Figure 1: Sample Test Information Function



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

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

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented in Figure 2 and Figure 3, instead of the TIFs for the spring and summer administrations. These plots are based on the scaled scores reported in 2019. The vertical line represents the performance category cut score.

Figure 2: Conditional Standard Error of Measurement (Spring)

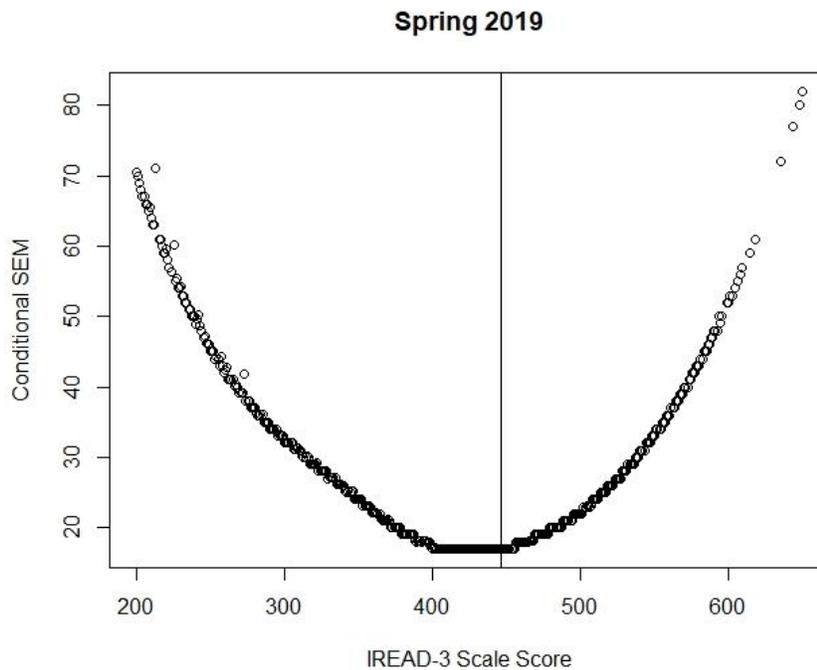
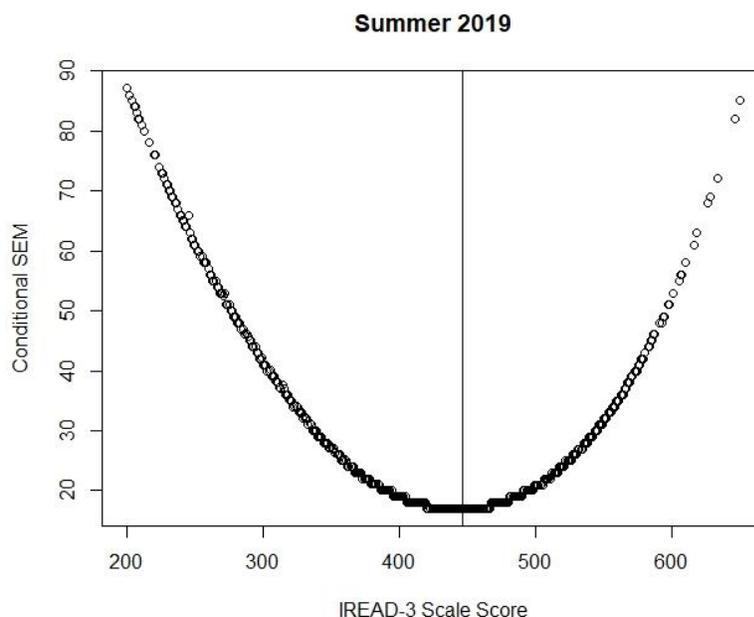


Figure 3: Conditional Standard Error of Measurement (Summer)



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 includes 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 to computing misclassification probabilities and is designed to explore the following research questions:

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

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

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

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

For these analyses, we used students from the spring 2019 IREAD-3 population data files that had an overall score reported. 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 AIR’s scoring engine.

Table 3: Descriptive Statistics

Administration	Sample Size	Mean Theta	Standard Deviation of Theta	Mean Scale Score	Standard Deviation of Scale Scores
Spring	86,006	-0.101	1.085	508.857	81.150
Summer	12,682	-1.092	0.936	434.767	70.042

4.3.1 Classification Accuracy

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

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

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

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

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

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

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 4×4 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 provides the overall CAI based on the observed score approach. There is no industry standard, but these numbers suggest that misclassification would not be frequent in the population data.

Table 4: Classification Accuracy Index

Administration	Overall Accuracy Index
Spring	0.998
Summer	0.907

4.3.2 Classification Consistency

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

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

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

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

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

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

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

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

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

Table 5 provides the FPR and FNR for the IREAD-3 administrations.

Table 5: False Classification Rates

Administration	FPR	FNR	Accuracy
Spring	0.118	0.024	0.858
Summer	0.101	0.079	0.820

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

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

Classification consistency with classification accuracy results are presented in Table 6. All accuracy values are higher than 0.90 and classification rates are higher than 0.87.

Classification accuracy is slightly higher than classification consistency. Classification consistency rates can be lower than classification accuracy because the consistency is based on two tests with measurement errors, while the accuracy is based on one test with a measurement error and the true score.

Table 6. Classification Accuracy and Consistency

Administration	Accuracy	Consistency
Spring	0.998	0.941
Summer	0.907	0.875

4.4 PRECISION AT CUT SCORES

Table 7 presents mean CSEM at each performance level by administration. These tables also include performance-level cut scores and associated CSEM.

Table 7: Performance Levels and Associated Conditional Standard Error of Measurement

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
Spring	1	20.699	--	--
	2	35.311	446	17.014
Summer	1	24.074	--	--
	2	21.747	446	17.024

5. EVIDENCE ON INTERNAL-EXTERNAL STRUCTURE

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

On IREAD-3 assessments, there are three reporting categories: Reading Foundations and Vocabulary, Nonfiction, and 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 for student performance.

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

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

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score.

5.1 CORRELATIONS AMONG REPORTING CATEGORY SCORES

Table 8 presents the observed correlation matrix of the reporting category percent correct scores for both administrations. The average correlation was 0.73 for the spring administration and 0.67 for the summer administration.

Table 8: Observed Correlation Matrix Among Reporting Categories

Administration	Reporting Category	Number of Items	RFV	NF	L
Spring	Reading: Foundations and Vocabulary (RFV)	12	1.000		
	Reading: Nonfiction (NF)	12	0.684	1.000	
	Reading: Literature (L)	14	0.710	0.795	1.000
Summer	Reading: Foundations and Vocabulary (RFV)	12	1.000		
	Reading: Nonfiction (NF)	14	0.652	1.000	
	Reading: Literature (L)	12	0.632	0.713	1.000

5.2 CONFIRMATORY FACTOR ANALYSIS

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

The results in this section are intended to provide evidence that the methods for reporting 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, 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, 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.

5.2.1 Factor Analytic Methods

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

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

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

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

The factor analysis models are based on the matrix 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 vech vectorizes a symmetric matrix. That is, vech stacks each column of the matrix to form a vector. Note that the WLSMV approach (Muthén, du Toit, & Spisic, 1997) employs a weight matrix of asymptotic variances (i.e., the diagonal of the weight matrix) instead of the full asymptotic covariances.

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

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

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

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

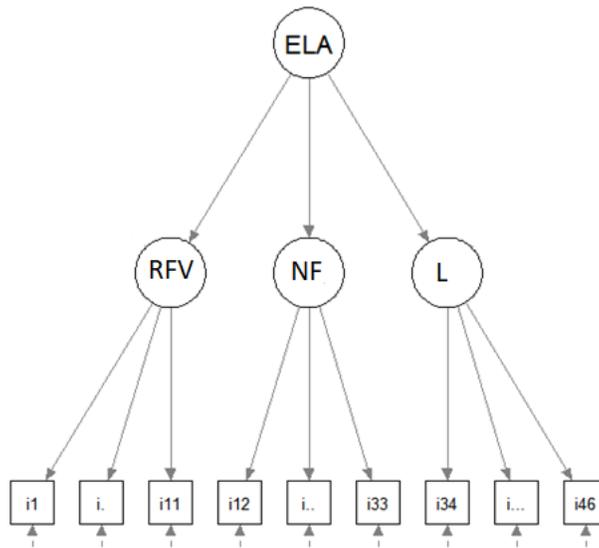
$$\hat{\Sigma} = \Lambda(\Gamma\Phi\Gamma' + \Psi)\Lambda' + \Theta,$$

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

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 4, 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 confirmatory factor analysis for IREAD-3 was to provide evidence that each individual assessment in IREAD-3 implied a second-order factor model: a single underlying second-order factor with the first-order factors defining each of the reporting categories.

Figure 4: Second-Order Factor Model (ELA)
Generalized Second Order Factor Structure



5.2.2 Results

Several goodness-of-fit statistics from each of the analyses are presented in Table 9, which shows the summary results obtained from confirmatory factor analysis. Three

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

Based on the fit indices, the model showed good fit across content domains. RMSEA was below 0.025, and CFI and TLI were equal to or greater than 0.973.

Table 9: Goodness-of-Fit Second-Order CFA

IREAD-3					
Administration	df	RMSEA	CFI	TLI	Convergence
Spring	663	0.022	0.985	0.984	YES
Summer	662	0.025	0.974	0.973	YES

Table 10 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 10: 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
Summer	Reading: Foundations and Vocabulary (RFV)	12	1.00		
	Reading: Nonfiction (NF)	14	0.876	1.00	
	Reading: Literature (L)	12	0.862	0.938	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 that there are detectable distinctions among reporting categories.

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

Overall, these results provide empirical evidence and justification for the use of our 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 11 presents summaries of the distributions of Q_3 statistics—minimum, 5th percentile, median, 95th percentile, and maximum by administration. Overall, only two items had a Q_3 value greater than the critical value of 0.2 for $|Q_3|$ (Chen & Thissen, 1997).

Table 11: Q_3 Statistics

Administration	Q_3 Distribution				
	Minimum	5th Percentile	Median	95th Percentile	Maximum
Spring	-0.101	-0.055	-0.025	0.004	0.292
Summer	-0.085	-0.061	-0.025	0.022	0.169

5.4 CONVERGENT AND DISCRIMINANT VALIDITY

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

Table 12 and Table 13 show the observed correlations between ILEARN Grade 3 ELA and Mathematics subscores and the IREAD-3 subscores, where students took both assessments. In general, the pattern is consistent with the a priori expectation that subscores within a test correlate more highly than correlations between tests measuring a different construct, with a few small notes on the writing dimensions.

Table 12: Observed Score Correlations Spring

Subject	Reporting Category	IREAD-3		
		Cat1	Cat2	Cat3
ILEARN ELA Grade 3	Key Ideas and Textual Support/Vocabulary	0.54	0.65	0.64
	Structural Elements and Organization/Connection of Ideas/Media Literacy	0.47	0.59	0.57
	Writing	0.48	0.56	0.55
ILEARN Mathematics Grade 3	Algebraic Thinking and Data Analysis	0.56	0.62	0.60
	Computation	0.52	0.59	0.57
	Geometry and Measurement	0.51	0.57	0.55
	Number Sense	0.50	0.56	0.54

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

Table 13: Observed Score Correlations Summer

Subject	Reporting Category	IREAD-3		
		Cat1	Cat2	Cat3
ILEARN ELA Grade 3	Key Ideas and Textual Support/Vocabulary	0.31	0.35	0.34
	Structural Elements and Organization/Connection of Ideas/Media Literacy	0.23	0.25	0.25
	Writing	0.29	0.30	0.29
ILEARN Mathematics Grade 3	Algebraic Thinking and Data Analysis	0.41	0.41	0.37
	Computation	0.39	0.39	0.35
	Geometry and Measurement	0.37	0.35	0.31
	Number Sense	0.36	0.35	0.32

*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. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002), including:

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

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

6.1 STATISTICAL FAIRNESS IN ITEM STATISTICS

Analysis of the content alone is 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
- English Learners (ELs)/Not ELs

A detailed description of the DIF analysis that was performed is presented in Volume 1, Section 4.2, of the *2018–2019 IREAD-3 Annual Technical Report*. The DIF statistics for each operational test item are presented in the appendix A of Volume 1 of the *2018–2019 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.



IREAD-3

**Indiana Reading Evaluation
and Determination**

2018–2019

**Volume 5
Score Interpretation Guide**

TABLE OF CONTENTS

1. INDIANA SCORE REPORTS	1
1.1 Overview of Indiana's Score Reports.....	1
1.2 Overall Scores and Reporting Categories.....	1
1.3 Online Reporting System	3
1.3.1 Individual Student Report.....	Error! Bookmark not defined.
1.3.2 Interpretive Guide.....	5
1.3.3 Data File.....	6
2. INTERPRETATION OF REPORTED SCORES	7
2.1 Scale Score	7
2.2 Standard Error of Measurement	7
2.3 Performance Level.....	8
2.4 Performance Category for Reporting Categories.....	8
2.5 Cut Scores	8
2.6 Appropriate Uses for Scores and Reports	8
3. SUMMARY.....	10

LIST OF TABLES

Table 1: Reporting Categories for IREAD-3..... 2
Table 2: IREAD-3 Assessment Proficiency Cut Scores 8

LIST OF FIGURES

Figure 1: Individual Student Report 4
Figure 2: Supplemental Interpretive Guide 5
Figure 3: Data File 6

ACKNOWLEDGMENTS

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

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

1. INDIANA SCORE REPORTS

In Spring 2019, pursuant to House Enrolled Act 1367, also known as Public Law 109, the Indiana Reading Evaluation and Determination (IREAD-3) assessment was administered to Indiana students in grade 3. Students in grades 4 and 5 who had not previously passed the IREAD-3 were given the opportunity to retest.

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

1.1 OVERVIEW OF INDIANA’S SCORE REPORTS

The IREAD-3 assessment was administered in spring and summer 2019. Test scores from each spring 2019 assessment were provided to IDOE corporations and schools through the ORS beginning on March 18, 2019, for spring and June 10, 2019, for summer.

The ORS (<https://in.reports.airast.org>) is a web-based application that provides IREAD-3 results at various, privileged levels. 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 users: 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 all aggregated test data for their rosters and the students within their rosters

Access to reports is password protected, and users can access data at their assigned level and below. For example, an STC 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 the ORS reports the scores. However, tests may also be manually invalidated before reaching the ORS if testing irregularities occur (e.g., cheating, unscheduled interruptions, loss of power or Internet 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 that the student's knowledge and skills fall below proficiency as measured by the assessment. Conversely, higher scale scores indicate that 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.

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-level descriptors set out content-area knowledge and skills that students at each performance level are expected to possess, and they are determined by comparing a student's scale score against carefully established cut scores, unique to each grade and subject. Cut points are listed in Section 2.5, Cut Scores. Performance levels can be interpreted on the basis of performance-level descriptors, which represent a descriptive analysis of a student's abilities based on his or her performance level.

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 Nonfiction Literature

1.3 ONLINE REPORTING SYSTEM

The 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 the ORS. In 2019, score data was published to ORS one day after the test window opened for the spring 2019 administration and one week after the test window opened for summer 2019 administration. Quality control verification was conducted on the score data for both test windows before data was released in ORS.

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

- Scale score
- Overall subject performance level

The top of the report includes the following:

- Student's name
- Scale score
- performance level

The middle section includes the following:

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

The bottom of the report includes the following:

- 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?

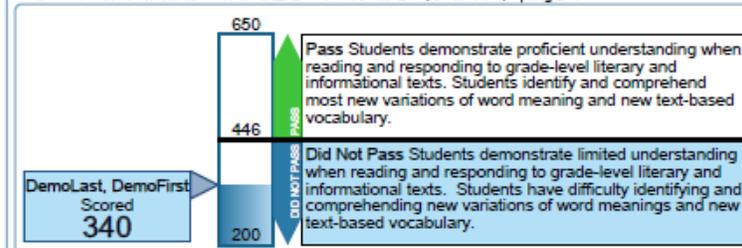
Test: IREAD-3

Year: Spring 2019

Name: DemoLast, DemoFirst

Overall Performance on the IREAD-3 Test: DemoLast, DemoFirst, Spring 2019			
Name	STN	Scale Score	Passing Status
DemoLast, DemoFirst	199999231	340	Did Not Pass

Scale Score and Performance on the IREAD-3 Test: DemoLast, DemoFirst, Spring 2019



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: DemoLast, DemoFirst, Spring 2019

Strand	Percent Correct
Reading: Foundations and Vocabulary	50
Reading: Nonfiction	17
Reading: Literature	21

Based on data from the IREAD-3, Spring 2019 administration.
 Report Generated: 6/4/2019 2:05:39 PM EDT
 Data presented in this system are considered preliminary. Official data will be released from the Department following the Spring and Summer 2019 administrations.
 For help in understanding your student's scores and this report, contact your student's teacher or school principal.

1.3.1 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 also available in five different languages: Arabic, Chinese, Burmese, Spanish, and Vietnamese.

Figure 2: Supplemental Interpretive Guide



Indiana Reading Evaluation and Determination IREAD-3 Assessment Results

Working Together for Student Success

Dear Parent/Guardian,

This report provides information about your child’s performance on the Indiana’s 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 the results with your child and his/her teacher. Thank you for supporting your child’s education.



Dr. Jennifer McCormick
State Superintendent of Public Instruction

INFORMATION ON INDIANA’S IREAD-3 ASSESSMENT

IREAD-3 measures foundational reading standards through grade 3. Overall student results in 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 just 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 and not just at 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.	999999001	516	Pass

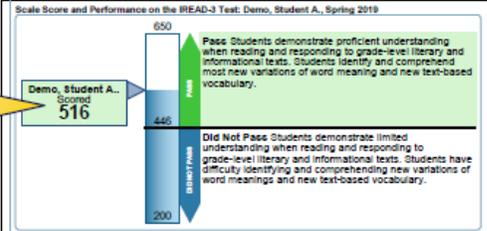
We encourage you to review these results with your child and his/her teacher. If you have questions about the contents of this report, contact your local school or district.

Things 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.

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



Demo, Student A. Scored **516**

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.

Did Not Pass: Students demonstrate limited understanding when reading and responding to grade-level literary and informational texts. Students have difficulty identifying and comprehending new variations of word meanings and new text-based vocabulary.

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.

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 understand more about your child’s proficiency level, go to www.doe.in.gov/assessment/iread-3-families
- To practice questions similar to what your child has seen on IREAD-3, go to <https://inpt.tds.airast.org/student>

For more information about this assessment, go to www.doe.in.gov/assessment/iread-3

Indiana Department of Education

1.3.2 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 administration of the test. 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												
1	Student F	Student L	STN	Student D	Gender	Ethnicity	Special Ed	Identified	Section 5C	Enrolled	C	Enrolled	S	Enrolled	S	Enrolled	C	Enrolled	C	Reading	C	Reading	S	Reading	P	Reading	f	Reading	f	Reading	Literature	Pe	
2	Demo	Student1	123456789	26/07/200	M	White	Y	N	N		3	Demo Sch	9999_9995	Demo Dis	9999		1	650	Pass		100	100	100										
3	Demo	Student2	123456789	30/05/200	F	White	N	N	N		3	Demo Sch	9999_1000	Demo Dis	10000		1	403	Did Not P		57	29	67										
4	Demo	Student3	123456789	19/05/201	F	White	N	N	N		3	Demo Sch	9999_1000	Demo Dis	10001		1	493	Pass		86	79	75										
5	Demo	Student4	123456789	21/08/200	M	White	N	N	N		3	Demo Sch	9999_1000	Demo Dis	10002		1	411	Did Not P		71	43	50										
6	Demo	Student5	123456789	05/01/201	M	White	Y	N	N		3	Demo Sch	9999_1000	Demo Dis	10003		1	472	Pass		86	93	50										
7																																	

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. Students' scores and performance levels are summarized at aggregate levels. 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 can be used to illustrate students' current level of performance and are most powerful when used to measure their growth over time. Lower scale scores can indicate that the student'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. 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 STANDARD ERROR OF MEASUREMENT

A student's score is best interpreted when recognizing that his or her knowledge and skills fall within a score range and are not just precise numbers. A scale score (the observed score on any test) is an estimate of the true score. If a student takes a similar test several times, the resulting scale scores would vary across administrations, sometimes being a little higher, a little lower, or the same. The SEM represents the precision of the scale score, or the range in which the student would likely score if a similar test were administered several times. The SEM can be interpreted as the degree of uncertainty of a student's score based on a statistical analysis of the student's answers on a test. When interpreting scale scores, it is recommended to always consider the range of scale scores incorporating the SEM of the scale score.

The \pm next to the student's scale score provides information about the certainty, or confidence, of the score's interpretation. The boundaries of the score band are one SEM above and below the student's observed scale score, representing a range of score values that is likely to contain the true score. For example, 445 ± 15 indicates that if a student were tested again, it is likely that he or she would receive a score between 430 and 460.

2.3 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 2.5, Cut Scores). Performance-level descriptors are descriptions of content-area knowledge and skills that students at each performance level are expected to possess. Thus, performance levels can be interpreted in relation to performance-level descriptors.

Performance-level descriptors are available on the IDOE web page at www.doe.in.gov/sites/default/files/assessment/iread-3-cut-score-performance-level-descriptors_0.pdf.

2.4 PERFORMANCE CATEGORY FOR REPORTING CATEGORIES

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

2.5 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

Test	Level 1 Did Not Pass	Level 2 Pass
IREAD-3	200–445	446–650

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 are able to do in certain subject 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, performance level indicators for reporting categories can be used to identify an individual student's relative strengths and weaknesses in reporting categories for a content area.

Assessment results on student test performance can be used to help teachers or schools make decisions about how to support students' learning. Aggregate score reports on the teacher and school levels provide information about students' strengths and weaknesses and can be used to improve teaching and student learning. For example, a group of students may have performed well overall but not as well in several reporting categories. In

this case, teachers or schools can identify their students' strengths and weaknesses through the group's performance by reporting category and can then promote instruction in specific areas where student performance is below overall performance. Furthermore, by narrowing the student performance result by subgroup, teachers and schools can determine what strategies may need to be implemented to improve teaching and student learning, particularly for students from disadvantaged subgroups. For example, teachers might see student assessment results by gender and observe that a particular group of students is struggling with literary response and analysis in reading. Teachers can then provide additional instruction for these students that will enhance their performance and enable them to achieve the benchmarks for literary response and analysis.

In addition, assessment results can be used to compare students' performance among different students and groups. Teachers can evaluate how their students perform compared with students in other schools and corporations by overall scores and reporting category scores. Furthermore, scale scores can be used to measure the growth of individual students over time if data are available. The scale score in IREAD-3 is reported on within-test scales.

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 take into account 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 ORS. The results were released in real-time during the test window beginning one day (spring 2019) and one week (summer 2019) after the start of the respective test windows.

The reporting system is interactive. When educators or administrators log in, they see a summary of data about students for whom they are responsible (a principal would see the students in his or her school; a teacher would see students in his or her class). They can then drill down through various levels of aggregation all the way to ISRs. The system allows them to tailor the content more precisely, moving from subject area through reporting categories and even to standards-level reports for aggregates. Aggregate reports are available at every user level, and authorized users can print these or download them (or the data on which they are based). 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.