Amira ISIP Assess

TECHNICAL GUIDE

ABSTRACT

The *Amira ISIP Assess Technical Guide* provides a comprehensive overview of the design, structure, and psychometric validation of the Amira ISIP Assessment for the 2026–2027 school year. It details how the tool functions as a universal screener, benchmark assessment, and progress monitoring system—leveraging AI-powered speech recognition and a 3rd-generation approach that extends traditional methods for assessing early literacy skills, detecting dyslexia risk, and supporting differentiated instruction. Grounded in leading research frameworks, the guide explains the constructs measured, the scoring methodology, and the technical standards that ensure validity, reliability, and equity across diverse student populations.

DATE

2026-2027 School Year

Table of Contents

1. INTRODUCTION....4
1.1 Theoretical Framework....4
1.2 Purpose and Use....5
1.3 Statement of Standards Alignment....6
2. CONSTRUCTS MEASURED....8
2.1 Phonological and Phonemic Awareness....8
2.2 Alphabetic Knowledge....14
2.3 Phonics/Decoding....15
2.4 Oral Reading Fluency....18
2.5 Vocabulary....20
2.6 Spelling/Encoding....22
2.7 Reading Comprehension....23
2.8 Oral Language....24
2.9 Rapid Automatized Naming....29
2.10 Visual Attention....30
3. TEST DESIGN....32
3.1 Assessment Blueprint and Design....32
3.2 Item Development and Expert Review....32
3.3 Field-Testing and Psychometric Validation....33
3.4 Content Management....35
3.5 Accommodations....35
3.6 UX Studies....38
3.7 Amira's Real-Time Calibration Design....39
3.8 Administration....43
4. MEASUREMENT MODEL....45
4.1 IRT Model....45
4.2 Item Calibration....45
4.4 Vertical Scaling....47
4.5 Calibration Studies....48
4.6 Differential Item Functioning....51
5. SCORING & REPORTS....60
5.1 Norm-Referenced vs. Criterion-Referenced Scores: ARM and MAST....60
5.2 ARM....61
5.3 MAST....62
5.4 Estimated Mastery....64
5.5 DRI....65
6. LINKING & EQUATING....67

6.1 Linking to Legacy ISIP....67
6.2 WCPM Equating....68
7. DEVELOPMENT OF NATIONAL NORMS....70
7.1 Stratification....70
7.2 Weighting....71
8. CLASSIFICATION ACCURACY....75
8.1 Student Sample....75
8.2 Candidate Amira ISIP Screener Cut Scores....77
8.3 Methodology....79
8.4 Results....80
8.5 Classification Accuracy Study by Subgroup....84
8.6 Classification Accuracy Study of Amira ISIP Subscores....87
9. RELIABILITY & VALIDITY....90
9.1 Reliability of Forms: Universal Screener, Benchmark, and Progress Monitoring....90
9.2 Standard Error of Measurement....101
9.3 Validity....107
10: PRIVACY & SECURITY....121
10.1 Amira's Commitment to Privacy and Security....121
10.2 Privacy and Security Compliance at a Glance....121
10.3 Compliance with Governing Standards: FERPA, COPPA, and GDPR....123
10.4 Amira's Privacy Policy....124
10.5 Student Audio: De-Identified Storage on AWS, and Nowhere Else....125
10.6 Security Provisions for Audio, Including Encryption....126
10.7 Policy: No Audio Sharing with Any Subprocessor Except AWS....126
10.8 Data Sharing Agreement (DSA) Policy and Model DSA....127
10.9 Consents Provided by Districts to Amira....132
10.10 Challenges and Judicial Precedent....134
10.11 Subprocessor Policy....137
10.12 State Legislation: Landscape and Amira's Conformance....138
10.13 SOC 2 as a Compliance Process, and Amira's Adherence....140
10.14 Conclusion: All Reasonable Steps to Protect Student Privacy....141
APPENDIX A: Advisor Information....142
APPENDIX B: Criteria for Item Quality Evaluation....143
APPENDIX C: Amira ISIP Tasks & Expected Times....145
APPENDIX D: SODA Progress Monitoring....146

1. INTRODUCTION

1.1 Theoretical Framework

Amira ISIP's theoretical framework for identifying reading risk is built upon three synergistic and distinct pillars of research: pragmatic guidelines from the International Dyslexia Association (IDA), Dr. Nell Duke's Active View Framework, and the Multiple Deficit Model (MDM) from Dr. David Francis and Jack Fletcher. The assessment leverages the IDA's recommended approaches and constructs to identify signals and markers of reading struggle, aligning with grade-specific screening recommendations. Amira Learning collaborates actively with these leading researchers and organizations to ensure its assessment reflects the latest, most evidence-based frameworks for predicting and explaining reading difficulties.

Dr. Nell Duke's Active View (AV) Framework serves as Amira ISIP's primary theoretical foundation for ensuring comprehensive coverage of reading mastery. This framework is designed to assess and improve reading comprehension by emphasizing five key components: Activation, Connecting, Thinking, Imaging, and Evaluating. The AV Framework offers a detailed and structured approach to understanding the multifaceted nature of reading, moving beyond models that focus solely on decoding or phonological processing to encompass cognitive, linguistic, and metacognitive processes. This holistic approach allows educators to identify not only *where* a reader struggles but also *why*, facilitating targeted interventions. Complementing this, the MDM provides a neuroscience-based framework for understanding the origins and causes of reading struggle. Guided by Dr. David Francis, the MDM posits that multiple cognitive, genetic, and environmental factors interact to produce reading difficulties associated with dyslexia. It suggests evaluating reading risk across seven constructs: phonological processing tasks, rapid automatized naming, orthographic processing tasks, working memory tasks, processing speed tasks, language skills assessment, and reading fluency, all of which Amira ISIP is organized to assess at each grade level.

The integration of these three frameworks — the IDA's pragmatic guidelines, Duke's Active View, and the MDM — allows Amira ISIP to offer a comprehensive, holistic, and developmentally appropriate assessment of reading difficulties. This ensures that basic reading skills, comprehension abilities, and underlying cognitive processes are all evaluated. Furthermore, Amira ISIP enhances this theoretical foundation with cutting-edge artificial intelligence (AI) and speech recognition technology, including Reading Error Detection (RED) Models, Multimodal Learning Analytics, Natural Language Processing (NLP), and Machine Learning. This technological integration uniquely combines the benefits of observational and digital tests, providing accurate, efficient, and comprehensive insights for early identification and targeted interventions for students at risk for reading difficulties. See the Amira ISIP Theoretical Framework for a more thorough description.

1.2 Purpose and Use

Amira ISIP serves three distinct and crucial purposes in supporting early literacy development: as a universal screener, a benchmark assessment, and a progress monitoring tool.

Universal Screener
Amira ISIP's universal screener is an online, computer-based solution primarily for Kindergarten to Grade 8 students to assess early literacy skills and identify students at risk of reading difficulties, including dyslexia. The universal screener is typically configured as the beginning-of-year assessment and is fully configurable at the LEA (Local Educational Agency) or SEA (State Educational Agency) level. This initial assessment differs from other assessments in that the inclusion of the Rapid Automatized Naming (RAN) task is strongly encouraged through grade 6 to yield the most valid and accurate classification outcomes. This screener integrates early literacy assessment with dyslexia screening, aligning with state requirements.

Benchmark Assessment
Amira ISIP's benchmark assessment can be administered at the beginning-of-year (BOY), middle-of-year (MOY), and end-of-year (EOY) to provide crucial data at key points throughout the school year, offering insights into student growth and detailed data aligned with state standards and school curricula. It is specifically designed for grades PreK-8. The benchmark assessment includes tasks that cover all components of the "reading rope" model, such as Phonological Awareness, Phonics/Decoding, High Frequency Word Recognition, Oral Reading Fluency, Receptive and Expressive Vocabulary, and Oral and Reading Comprehension.

The assessment leverages advanced AI algorithms for functions like RED, multimodal learning analytics (audio analysis, eye-tracking, keystroke data), Natural Language Processing (NLP) for comprehension, and Machine Learning. This AI-driven approach ensures high accuracy, efficiency, and fairness across diverse student populations, including those with varying accents and dialects, English learners, and students with disabilities. Amira ISIP dynamically adjusts task difficulty to achieve a median administration time of 15-18 minutes, minimizing test-taker fatigue and maximizing accuracy. It provides real-time, actionable data to teachers, empowering them with immediate feedback and resources tailored to identified skill gaps, and supports various accommodations to ensure equitable access.

Progress Monitoring
Amira ISIP's progress-monitoring tools are integrated into its suite to assess students' academic performance on an ongoing basis, evaluate their rate of improvement, and continuously signal which skills require supportive instruction. Progress monitoring

assessments are equivalent to the tasks configured on the benchmark assessment.

Districts can implement progress monitoring through three distinct approaches, each offering different levels of administrative control and data collection. The first method involves district-level screening windows, in which administrators establish multiple assessment periods designated as either benchmark or progress-monitor evaluations. When students log in during these windows, they automatically receive the assigned assessment.

The second approach empowers teachers through on-demand assessments (ODAs), allowing them to assign evaluations at any time, regardless of district scheduling. These ODAs automatically function as progress monitors when administered outside district windows, but inherit the window's designation when used within established periods. The third method leverages Amira Tutor, in which teachers simply assign students to the automated tutoring system. In all cases, Amira generates comprehensive data following each administration, including overall scores and subscores commensurate with the assessment configuration.

1.3 Statement of Standards Alignment

The development, technical documentation, and ongoing evaluation of this assessment system are designed to conform to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). In particular:

1. Validity
We have sought to provide evidence of validity consistent with the Standards' requirements (Standards, Chapter 1). This includes rationale for intended score uses and interpretations, support from multiple sources of evidence (e.g. content relevance, criterion relationships, internal structure, and consequential validity), and clear articulation of relevant construct definitions and score meaning.

2. Reliability / Precision
Reliability and precision analyses follow the Standards' guidance (Chapter 2), including assessment of internal consistency, measurement error, decision consistency, and, where applicable, generalizability or alternate-form reliability. Reliability coefficients are reported and interpreted in light of the tool's intended use.

3. Test Design, Scoring, and Interpretation
Procedures of item development, scoring, administration, and reporting are grounded in best practices as recommended in the Standards (Chapters 4–6). They are designed to minimize construct-irrelevant variance, ensure fairness, and support valid score interpretation for the intended examinee populations.

4. Fairness, Bias, and Generalizability

We attend to the Standards' principles regarding fairness and equity (Chapter 3), including efforts to evaluate differential item functioning, examine test performance across subgroups, and mitigate construct-irrelevant barriers. Representativeness of samples and generalizability of findings are also considered.

5. Supporting Documentation and Transparency
In alignment with the Standards (Chapter 7), we provide extensive technical documentation, including item specifications, calibration processes, validation studies, administration protocols, and interpretation guidance. This transparency fosters accountability and enables independent review.

By explicitly aligning with the AERA/APA/NCME Standards, we affirm our commitment to rigorous measurement practices, ethical test use, and defensible score interpretations.

2. CONSTRUCTS MEASURED

The following sections describe the Amira ISIP Assessment design, along with links and screenshots illustrating Amira ISIP in action.

2.1 Phonological and Phonemic Awareness

Amira ISIP uses several different tasks to assess Phonological and Phonemic awareness, carefully sequenced from foundational to advanced. The tasks reflect the developmental continuum of early literacy and were selected based on research evidence of efficacy in predicting literacy and reading difficulties such as dyslexia, as well as the success of task administration and scoring within the Amira ISIP screening context.

The assessment addresses multiple levels of phonological awareness, including:

- **Syllable Awareness**: Tasks include segmenting and blending syllables, such as in nonword repetition and phonological working memory tasks.
- **Rhyme Recognition**: Embedded within auditory listening and discrimination tasks to evaluate rhyme detection.
- **Onset-Rime Manipulation**: Items that require students to isolate and blend onset and rime units support early decoding readiness.
- **Individual Phoneme-Level Awareness**: Dedicated tasks assess phoneme segmentation, blending, isolation, deletion, and substitution.
- **Advanced Manipulation**: Deletion and substitution tasks become more complex across grades, supporting accurate classification and instruction.

Each task is supported by automatic scoring models and embedded within a computer-adaptive assessment experience to reduce cognitive load while maximizing diagnostic precision. Specific examples of tasks are listed below.

2.1.1 Blending

In this task, spoken words are presented as sequences of individual sound units (ranging from syllables to onsets and rimes to individual phonemes). The student must blend the provided sounds together into the full word. The task begins with Amira explaining that the student will hear the individual sounds that make up a word. The student is then prompted to blend these sounds and select the image that matches the blended word.

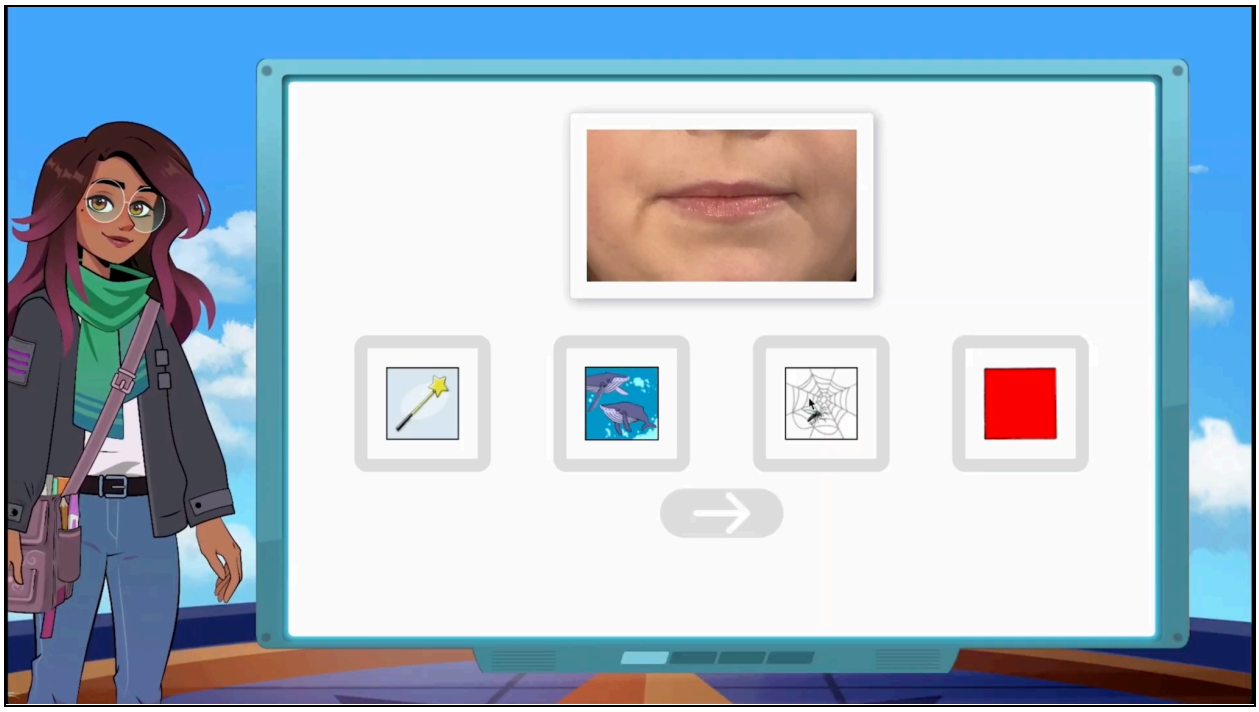


Figure 2.1 Screenshot of the Kindergarten Blending Task

The blending task is structured as follows. This is an example of syllable-level blending, but the item bank also contains many onset-rime and phoneme blending items.

1. Amira: "You are going to see some pictures. I will say their names. After I say the names of the pictures, my friend in the video will say the sounds in one of those names. You put the sounds together and decide what picture my friend is saying."
2. Four images are displayed on the screen. Amira names each image aloud.
3. Video of teacher articulating the segmented sounds (in this item, syllables): "/baɪ/ /sɪ/ /kəl/."
4. The student blends the sounds to form the word "bicycle" and selects the corresponding image.
5. The responses are scored dichotomously (correct/incorrect) and used towards the student's ability calculation.

See an example video of the blending task here.

2.1.2 Phoneme segmentation task

This task requires students to listen to one-syllable words and segment them into their constituent phonemes. The full articulation of the word is provided, and students are then asked to segment the word. The student is not presented with any text associated with the word to be segmented.

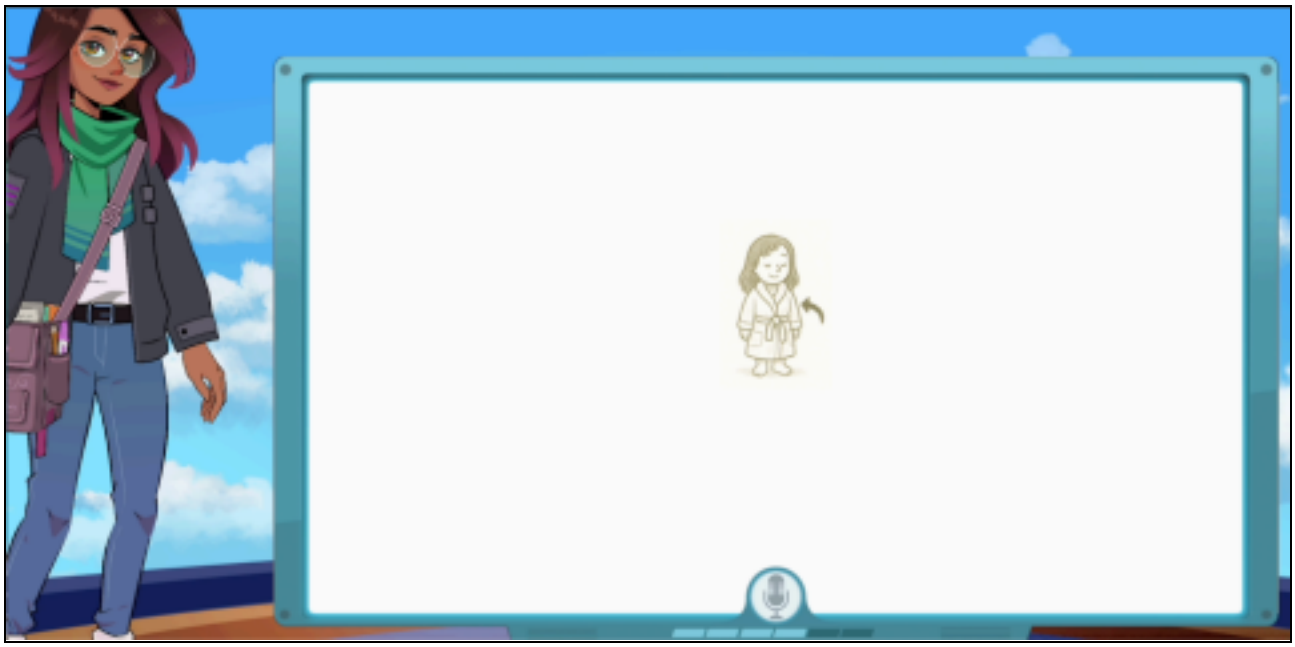


Figure 2.2 Screenshot of the Phoneme Segmentation Task

The single phoneme segmentation task is structured as follows:

1. Amira provides the student with directions by modeling how to segment a word. At the end of Amira's example, she blends the word so the student can hear the word both ways.
2. Amira then tells the student that it's their turn to segment each word.

For each assessment item:

3. Amira provides a picture of the word, then says, "the word is [word]" and then prompts the student to segment it.
4. The responses are scored using Amira's machine learning models.

2.1.3 Phonological elision task

In the phonological elision task, students are asked to say the sounds that remain after deleting a specific phoneme or word part. For half of the words, the deletion occurs at the beginning of the word (either syllable or phoneme/onset-rime manipulation), and for the other half of the words, the deletion occurs at the end of the word (either syllable or phoneme manipulation).

Amira delivers this task under the cover story of figuring out *mystery words* to say to Spot, a dog that students become acquainted with when they are first introduced to Amira's software.

The phonological elision task is structured as follows:

1. Amira provides directions to the student: "We're going to say some mystery words to Spot. I'm going to say a word, and then give you a part of that word you should not

say."

2. A warm-up item is presented:
    a. "For example, can you say the word **cup**?" → [student says the word]
    b. "Now, can you say **cup** without the /k/ sound?" **-->**[student responds and Amira provides feedback]

For each assessment item:

3. Amira says the word and asks the student to repeat it.
4. After Amira's models detect that the student has said the word back, Amira says, "Now tell me what word would be left if I said [word] without the [phoneme or word-part] sound."
5. The student responds.
6. The responses are scored using Amira's machine learning models.

See an example video of the phonological elision task here.

2.1.4 Phonological working memory task

Working memory, particularly working memory for language-related tasks, is an important cognitive skill that supports reading and spelling development (Swanson & Sachse-Lee, 2001). Children with reading difficulties, such as dyslexia, often exhibit deficits in working memory, which can affect their ability to hold and manipulate phonological and orthographic information during reading and writing tasks (Jeffries & Everatt, 2004).

A meta-analysis by Swanson et al. (2009) found that individuals with dyslexia consistently perform worse than typical readers on measures of verbal working memory, such as digit span and nonword repetition tasks. Additionally, longitudinal studies have shown that poor working memory skills in early childhood predict later reading difficulties.
Amira ISIP uses a Pseudoword (Nonword) Repetition task to assess phonological working memory. In this task, a video is shown of an adult vocalizing a sequence of syllables that combine to form a pseudo-word. The student is then prompted to repeat this pseudo-word. This task is supported for students in kindergarten and grade 1.

The sequences of syllables are carefully developed according to varying degrees of difficulty (e.g., varying syllable counts), to ensure they don't form words in any commonly spoken language, and to be age-appropriate (e.g., utilize phonemes and syllables that are appropriate to the speech capabilities of children at each age level).

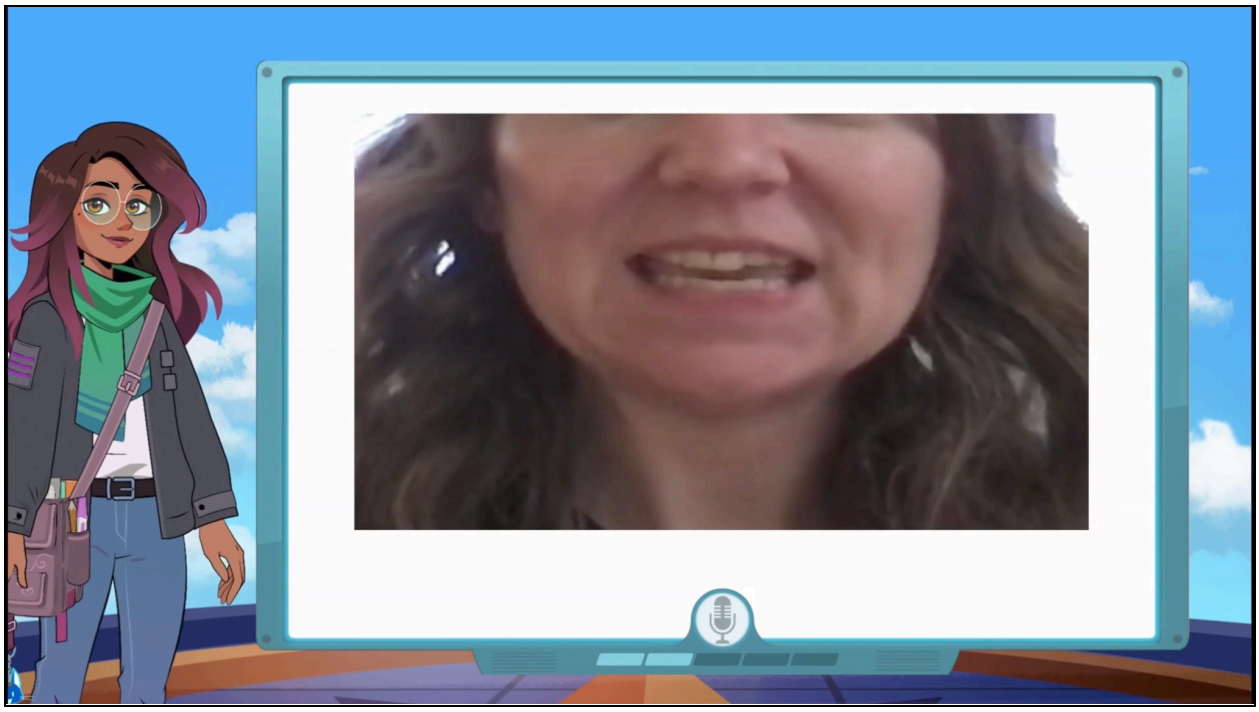


Figure 2.3 Screenshot of the Phonological Working Memory Task

The phonological working memory task is structured as follows:

1. Amira provides directions to the student: "Here is my friend. My friend is going to say some words that aren't real words, like **zevy**."
2. A warm-up item is presented: [a woman on the screen sounds out *zeh-vy*] "Please say **zevy** to my friend" → [student responds and Amira provides feedback].

For each assessment item:

3. Amira says the pseudo-word syllable sequence and asks the student to repeat it.
4. After Amira ISIP's models detect that the student has attempted a response, Amira says "Got it!". If Amira detects that the student is making no attempt, she will give the student up to one opportunity to replay the video of the pseudo-word.
5. The student responds.
6. The responses are scored using Amira's machine learning models.

See an example video of the phonological working memory task here.

2.1.5 Phoneme manipulation (Substitution)

Amira's phoneme manipulation sequence includes tasks that implicitly engage students in word chaining—a skill involving successive phoneme substitutions across a set of related items. While students are prompted on a single substitution per item, the sequencing of assessment items reflects chaining logic by progressively altering onset, medial, or coda phonemes in a structured series.

Specifically, the task below requires students to perform a phoneme substitution on either

the first or last phoneme in consonant-vowel-consonant (CVC) words. The articulation of the starting word is presented, followed by an indication, both visually and verbally, of which phoneme should be changed and to which phoneme it should be changed. Then, the student is asked to say the resulting word after the phoneme manipulation.

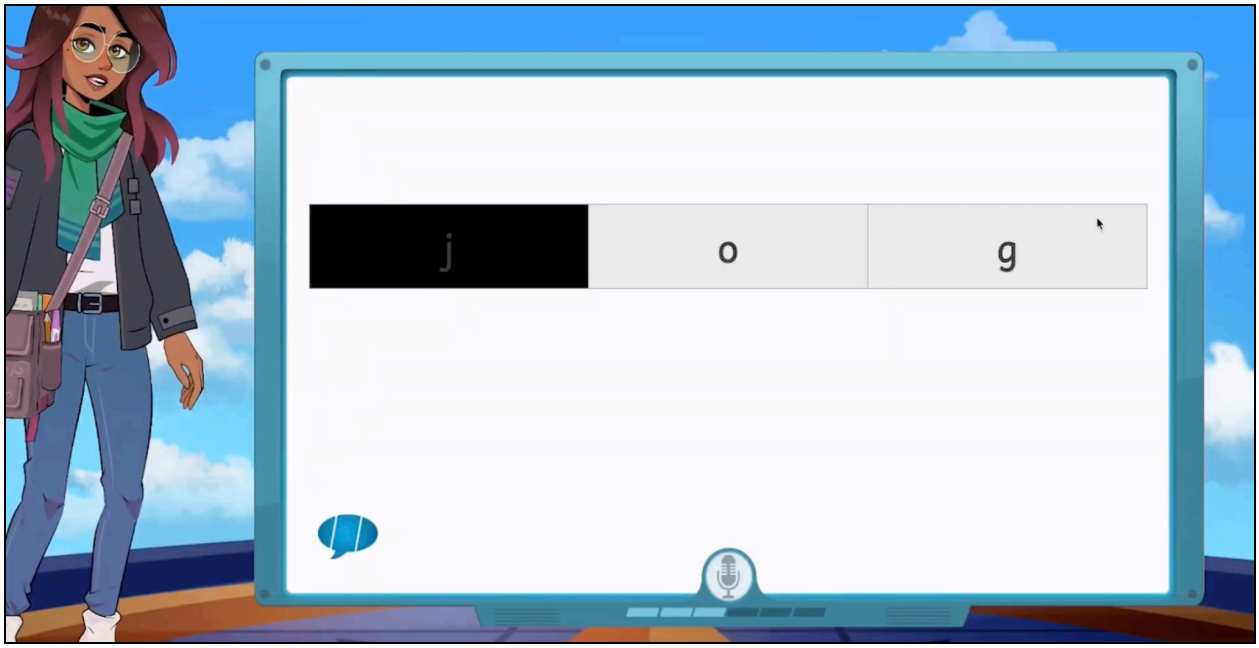


Figure 2.4 Screenshot of the Phoneme Manipulation

The phoneme manipulation is structured as follows:

1. Amira provides directions to the student.
2. A warm-up item is presented.

For each assessment item:

3. A number of boxes appear on the screen, representing the number of phonemes in the word.
4. Amira says, "Let's play with the word **DUG**". Here, **DUG** represents the starting word. Amira says, "Now let's change the /d/ sound to the /j/ sound. What's the word?" While Amira is saying the latter directions, the box corresponding to the phoneme to be changed will flash to indicate to the student whether it is the first or the last sound. In the **DUG**–to–**JUG** example, the first box will flash, as shown in the screenshot above.
5. The student is then given an opportunity to say the resulting word (in this case, **JUG**).
6. The response is scored using Amira's machine learning models.

See an example video of the phoneme manipulation task here.

2.2 Alphabetic Knowledge

2.2.1 Letter Name Fluency

Amira ISIP's Letter Name Fluency task displays the letters of the alphabet in text on the screen, one at a time, and requires students to verbally name each letter within a specified time window.

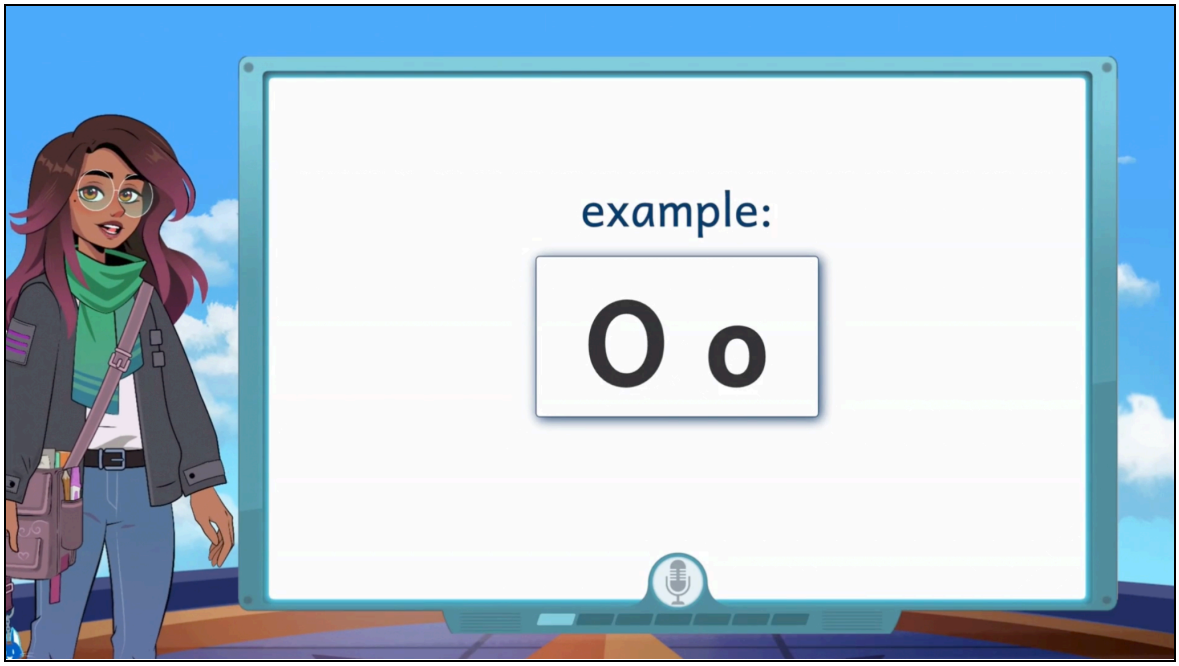


Figure 2.5 Screenshot of the Warm-up Example for the Letter Name Fluency

This task leverages Amira's ability to listen to speech, enabling the software to emulate the typical approaches teachers use to assess mastery of alphabetic knowledge. A student is typically presented with ten items in this task.

The Letter Name Fluency is structured as follows:

1. Amira provides directions to the student.
2. A warm-up example is presented.
3. A letter is presented to the student on screen as text.
4. The student is asked to name the letter shown on the screen.
5. The student has a configured interval of time to articulate the correct letter name.
6. Amira scores the item dichotomously.

See an example video of the letter naming task here.

2.2.2 Letter Sound Fluency

Amira's Letter Sound Fluency task displays the letters of the alphabet in text on the screen,

one at a time, and requires students to produce each letter's sound within a specified time window. A student is typically presented with six to ten items in this task.
The Letter Sound Fluency task is structured as follows:

1. Amira provides directions to the student.
2. A warm-up example is presented.
3. Amira displays the uppercase and lowercase instantiations of one letter.
4. The student is asked to say the sound that the letter shown on the screen makes.
5. The student has a configured interval of time in which they are given to articulate the correct phoneme.
6. Amira scores the item dichotomously. If there are multiple correct responses (as with vowels), Amira accepts any version as correct.

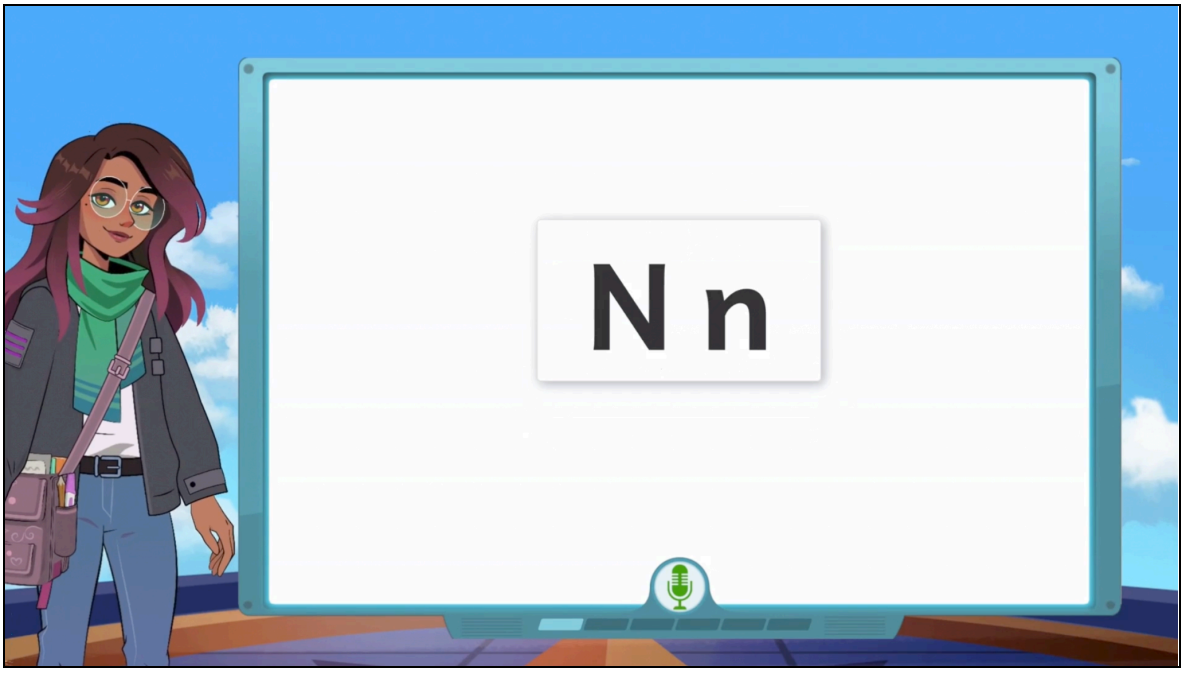


Figure 2.6 Screenshot of the Letter Sound Task

See an example video of the letter sound fluency task here.

2.3 Phonics/Decoding

2.3.1 Pseudo-word Decoding

The goal of this task is to measure a student's capacity to decode, converting printed text into a sequence of sounds and then blending those sounds into complete pseudo-words.

Using pseudo-words requires students to rely on their decoding skills rather than recognizing words from memory and familiarity. The Pseudo-word/Non-word Decoding task is presented as a series of made-up words, with Amira listening for the proper sound-outs based on common letter-sound correspondences and for successfully blending the sounds

into the full pseudo-word unit. Amira ISIP's pseudo-word items are carefully constructed to reflect the expected decoding skills of students at the target grade level, to be phonotactically valid, and to avoid biases that may be present for bilingual/EL (English Learner) and other special populations (i.e., pseudo-words that are real words in other languages, especially if the decoding patterns differ from English, are excluded). Kindergarten and Grade 1 items are short and mostly monosyllabic. The words used in the task conform to standard and typical patterns in the English lexicon.



Figure 2.7 Screenshot of the Pseudo-word/Non-word Decoding Task

The Pseudo-word Decoding task is structured as follows:

1. Amira provides directions to the student.
2. A warm-up item is presented.
3. A pseudo-word is presented in text form.
4. The student is asked to decode and pronounce the full pseudo-word.
5. The student has a configured interval for articulating the pseudo-word. 6. Amira scores the item.

See an example video of the pseudo-word decoding task here.

2.3.2 Word Identification Fluency

Amira ISIP measures the word identification fluency construct using a Word Identification Fluency task. In this activity, the student is presented with decodable words of varying difficulty and is asked to read the word aloud.

The Word Identification Fluency tests the basic ability to read words in isolation. The words

presented are mostly at the student's grade level but vary in difficulty. Words are chosen to test a student's mastery of all letter-sound correspondences and basic decoding skills that are expected at the student's level.

Grade	Sample Item
Kindergarten	Cup
Grade 1	Home
Grade 2	Spring
Grade 3	Quickly

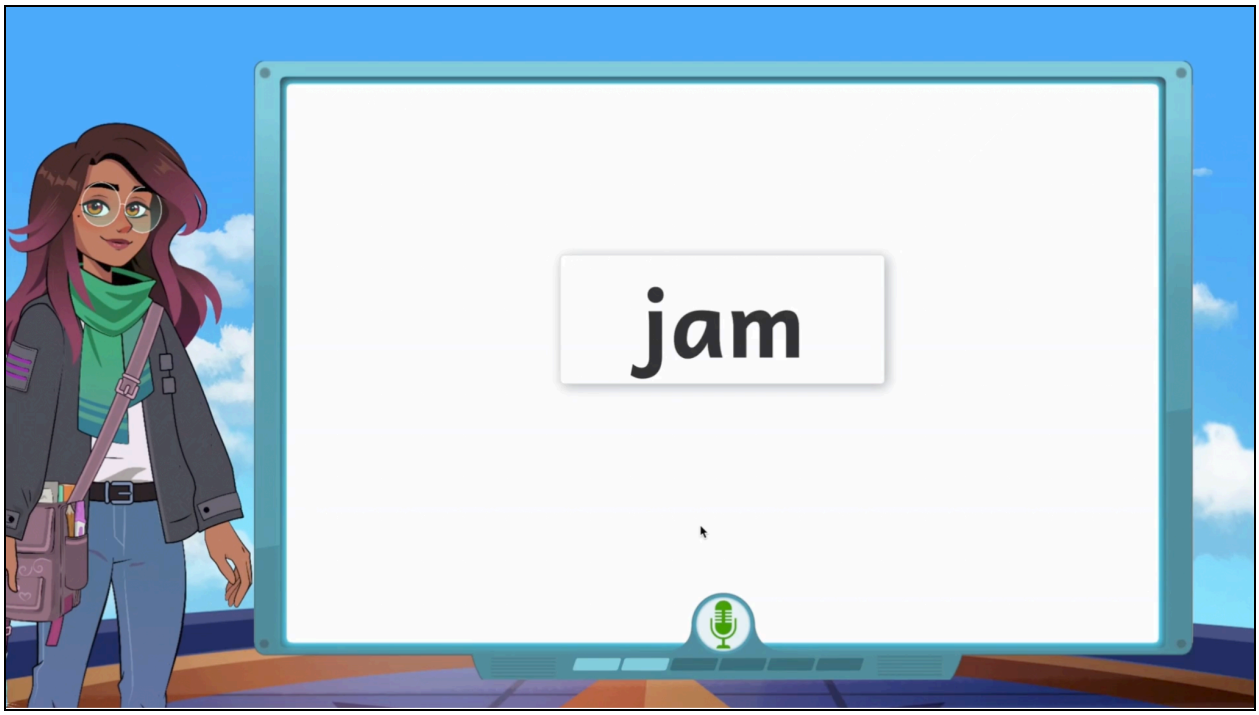


Figure 2.8 Screenshot of Word Identification Fluency task for Grade K

The Word Identification Fluency task is structured as follows:

1. Amira provides directions to the student.
2. A warm-up item is presented.
3. An isolated word is presented to the student in text form.
4. The student is asked to read the word.
5. The student has a configured interval for reading the complete word.
6. Amira scores the item.

The Word Identification Fluency task is supported for Grades K-6. The number of items presented varies from 4 to 20, depending on the grade level, with the number of words increasing with higher grade levels.

See an example video of the word identification task here.

2.4 Oral Reading Fluency

Amira ISIP administers an Oral Reading Fluency (ORF) task to assess students' ability to read words in the context of a passage, employing accuracy, prosody, and speed metrics, including Words Correct Per Minute (WCPM).

Amira ISIP presents a grade-level passage without images, divided into sections. The student reads one section at a time and then proceeds to the next. If the student struggles to read fluently, the text is adjusted to a lower level. Once the student finishes a section, Amira allows them to move on to the next block of text.

Typically, Amira provides enough text for the student to read for 90 seconds to 4 minutes. The ORF task is timed to assess fluency, not to encourage speed. That is, the goal is to evaluate the rate at which students can read with natural ease, not to read as quickly as possible. All ORF reading passages are designed so that the 4-minute limit will provide ample time for students to read all the text.

Speed, accuracy, prosody, and detailed reading miscue and timing information collected during the ORF task help to richly inform Amira ISIP's assessment of the student's abilities across the different threads of the reading rope. Amira ISIP's ORF task is not only comprehensive in its assessment of fluency, accuracy, and miscues, but it is also highly predictive of a student's overall reading abilities and future reading success.

Amira ISIP's ORF task is designed to be a strong predictor of students' performance on broader reading assessments, including state standardized tests and other widely recognized literacy measures. The predictive accuracy of this task stems from its detailed assessment of critical reading constructs, as outlined in the "reading rope" model, which includes word recognition, decoding, and comprehension.

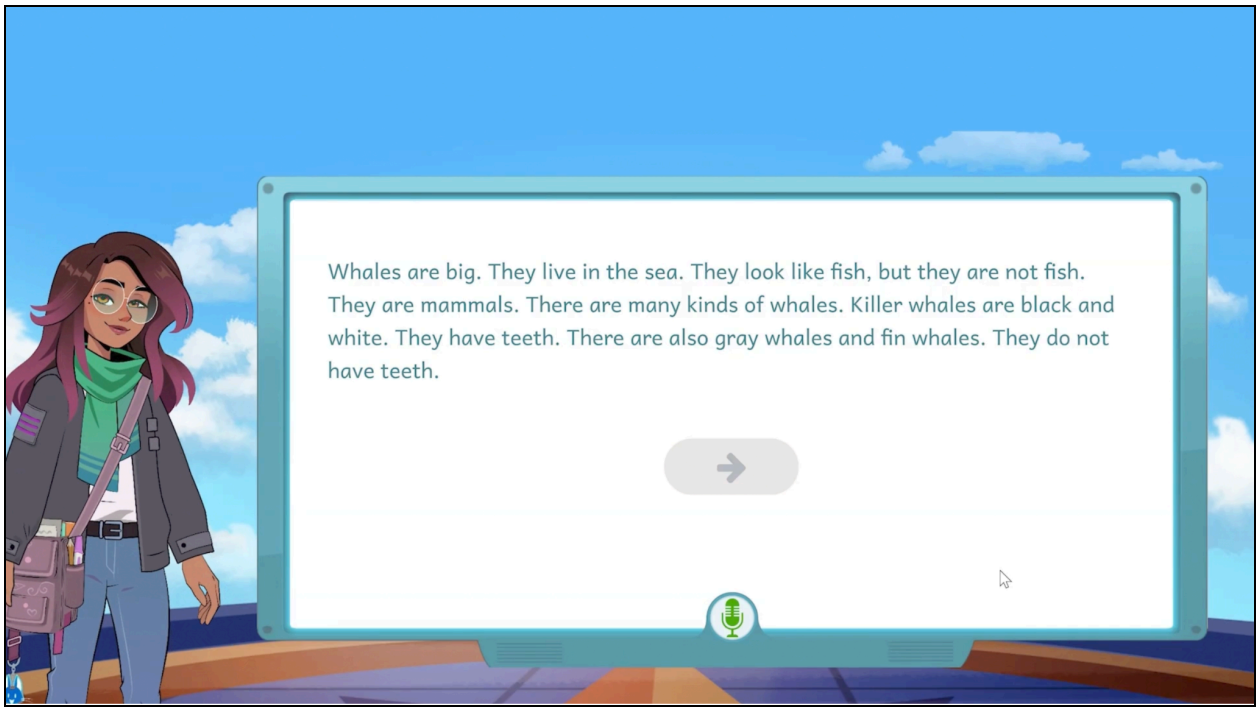


Figure 2.9 Screenshot of the Oral Reading Fluency Task

When a student completes a block in the passage, Amira moves them to the next block.

The ORF task is structured as follows:

1. Amira provides directions to the student.
2. Amira presents a short passage broken into blocks.
3. The student reads the passage, one block at a time.
4. If necessary, Amira adjusts text complexity based on the student's observed ability.
5. Timing information is kept at the word level.
6. On passage completion, Amira scores the ORF passage, identifying which words were correctly read and which words were not.
7. Amira uses each word as an item and additionally uses overall metrics like WCPM and error rate to compute final scores.

See an example video of the ORF task here. Content from Amira's ORF passages has been carefully constructed to align to different grade levels, including early readers in Kindergarten.

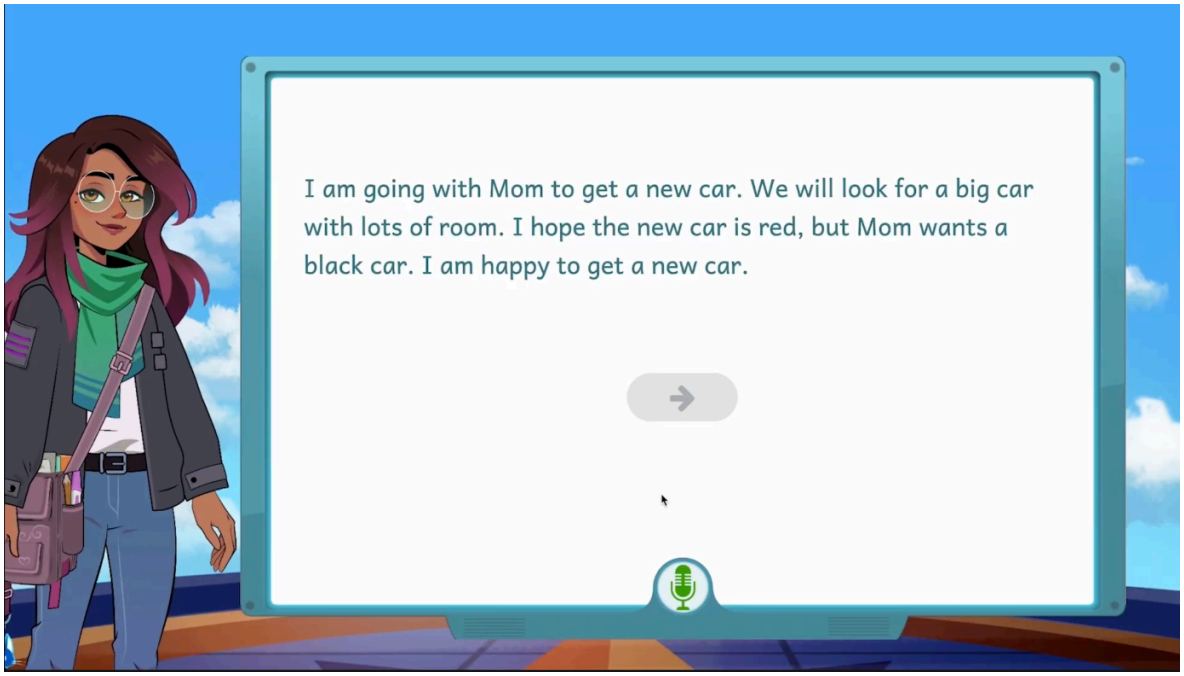


Figure 2.10 Screenshot of the Oral Reading Fluency Task at Kindergarten.

2.5 Vocabulary

In this task, Amira presents a word and asks the student to choose the word *that best matches* the target word from an array of three options. Amira reads the target word aloud and can read each of the multiple-choice options aloud on mouse-over, eliminating the need for students to read the words to complete the task. The goal of this task is to measure on-grade-level vocabulary skills, with each item chosen to represent a class of words that learners progressing at the state's expected pace should know. This task is supported in Grades K-6.

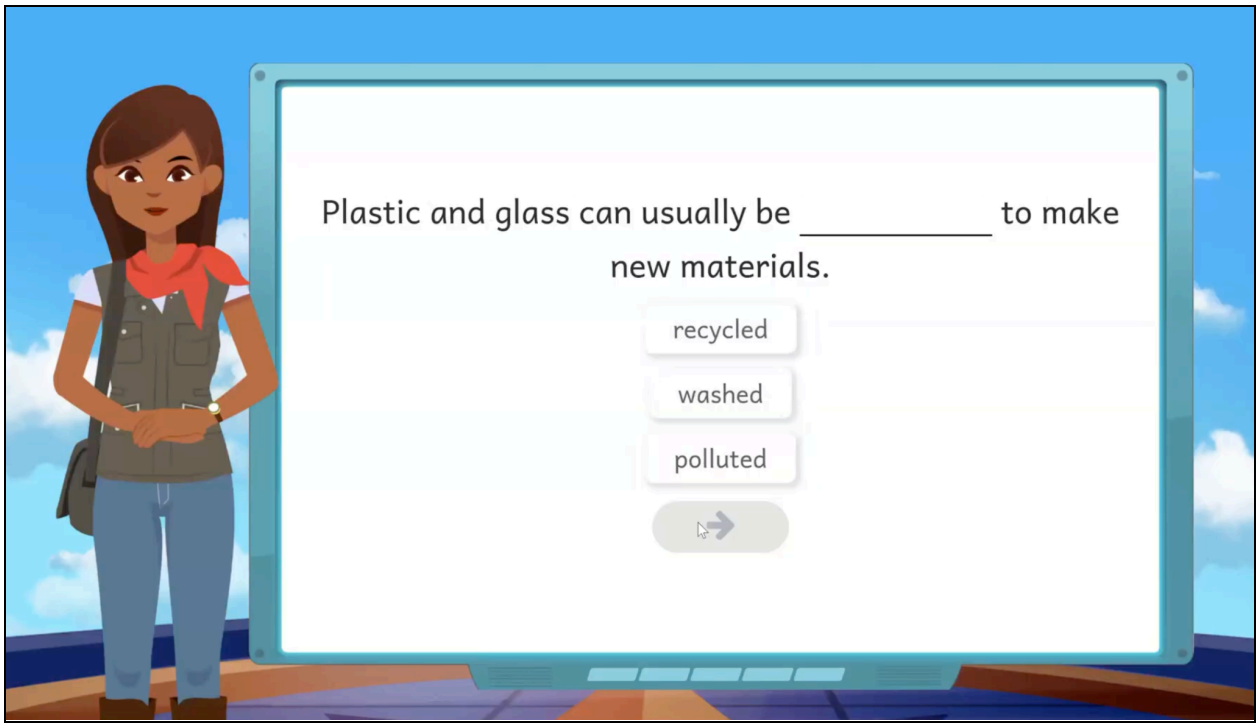


Figure 2.11 Screenshot of the Vocabulary Task - Words

See an example video of the vocabulary task here.

Amira also supports a configurable version of the vocabulary task, in which the choices are presented as pictures. In this version, she presents and reads the word aloud, and it is also shown in text. Amira asks the student to select which picture best shows that word. The correct picture is accompanied by two to three distractors. The screenshot below shows an item from the configurable picture version of the core vocabulary task.

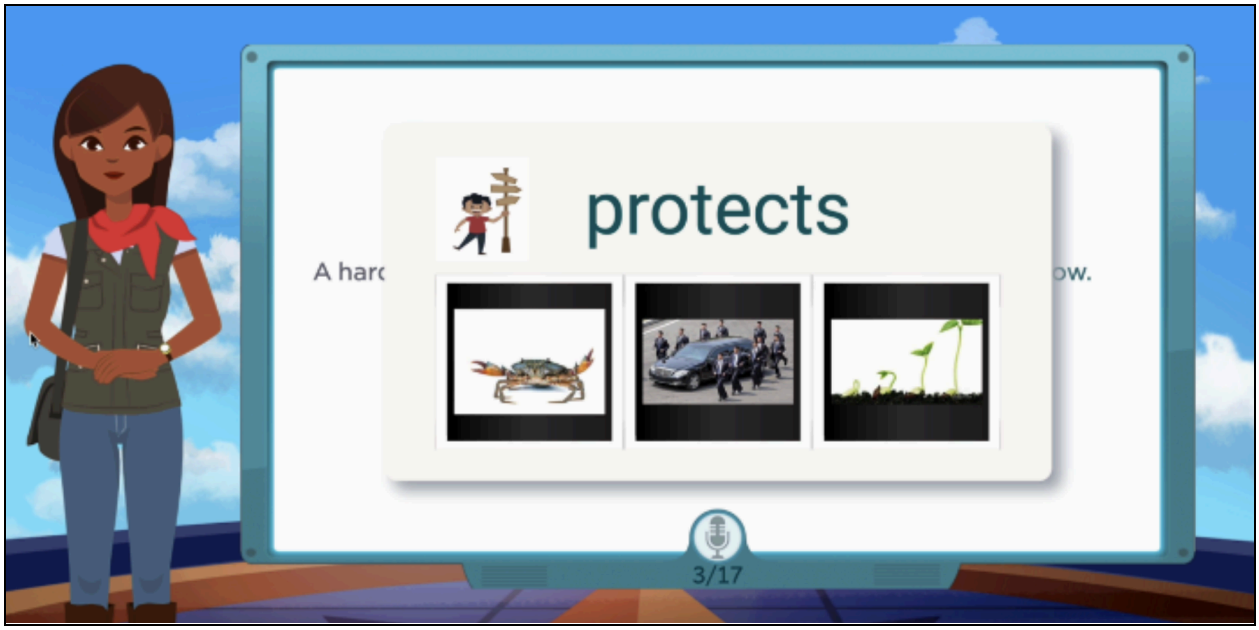


Figure 2.12 Screenshot of the Vocabulary Task - Pictures

Lastly, within Amira ISIP's Reading Comprehension task (described in section 2.8), specific items are designed to measure receptive vocabulary skills. Measuring vocabulary in context (i.e., embedded in reading passages) reflects how vocabulary knowledge is actually used during real reading. This captures both breadth (how many words you know) and depth (how well you know them) and supports authentic assessment practices aligned with natural reading processes. See an example screenshot of this reading comprehension task within the vocabulary construct below.

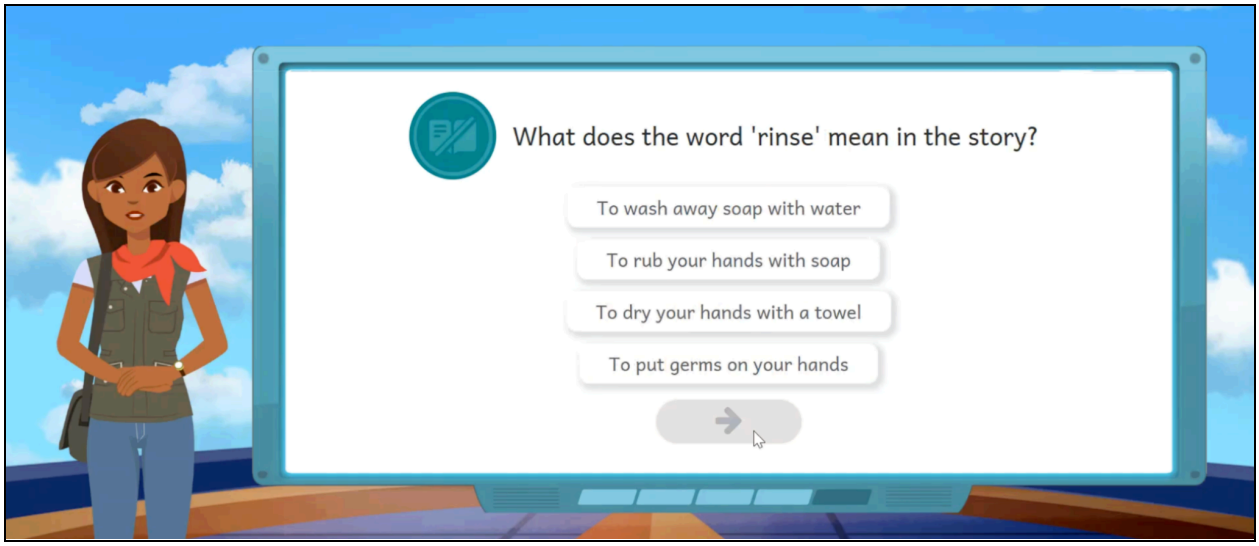


Figure 2.13 Screenshot of the Vocabulary Task - Comprehension

2.6 Spelling/Encoding

Amira ISIP utilizes a dedicated Spelling/Encoding task to assess this construct, which is supported in grades K through 8. In this task, Amira presents the student with a set of five to ten words. The student demonstrates their knowledge of encoding common spelling patterns to the best of their ability.

Item count is determined by grade and configuration. Amira articulates the words one by one, including using each word in a full sentence to give the student context. The student uses the keyboard to spell the word. Amira will repeat the word if needed. When the student has finished spelling, the green arrow activates, allowing them to move forward at their own pace. If too much time elapses, Amira will automatically move to the next item. Words vary in difficulty, and the time a student has to respond is calibrated in the software. Words are automatically scored by Amira as correct or incorrect. Additional error analysis by the teacher can help teachers understand specific spelling confusion. As with other items, the words included in the spelling task are specifically chosen for grade appropriateness and for letter-sound correspondence coverage.

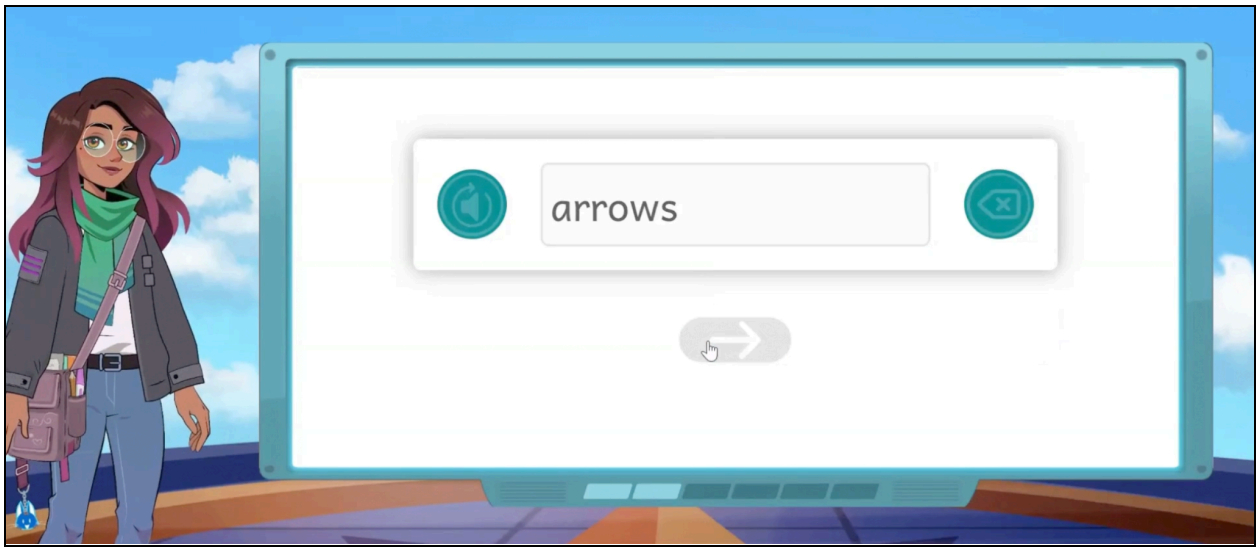


Figure 2.14 Screenshot of the Spelling/Encoding Task

The spelling/encoding task is structured as follows:

1. Amira provides directions to the student.
2. Amira reads the word, followed by an example of the word used in a sentence, followed by repeating the word again. For example, for the target word **rub**, Amira might say "Rub. I rub my eyes when they itch. Rub."
3. The student is then prompted to type the word into a text box, with an option to ask Amira to repeat the word if needed.
4. The responses are scored dichotomously based on whether the student spells the word correctly (1) or not (0).

See an example video of the encoding task here.

2.7 Reading Comprehension

Amira ISIP includes both a Listening Comprehension task (see Oral Language section) and a Reading Comprehension task, which follows the ORF passage that students read, to assess these skills. If both subtests are configured as part of the SEA or LEA setup, the student's performance on the Listening Comprehension task, compared with their performance on reading-based subtests, helps indicate the risk of dyslexia (or its absence). This allows Amira ISIP's Screener to effectively distinguish between low performance that may be linked to dyslexia and low performance caused by other factors, such as being an English learner (EL) or having developmental challenges, such as Specific Language Impairment (SLI).

The Reading Comprehension task extends the ORF task. After completing the ORF passage, the student answers three multiple-choice questions, each with one correct response and three distractors.

The Reading Comprehension task is structured as follows:

1. The student completes the ORF task and is prompted by Amira for the Reading Comprehension task.
2. Amira provides directions to the student.
3. Amira reads the question and answer choices aloud to the student. The student has access to the text for reference during each question.
4. For each assessment item, Amira poses several questions designed to test their understanding of the passage that they just read. For example, for the question: "When the story says, 'The man selling snacks is near,' what does the word near mean?" The students would choose the answer choice "close" to the correct answer to answer the question.
5. Amira scores the item automatically.

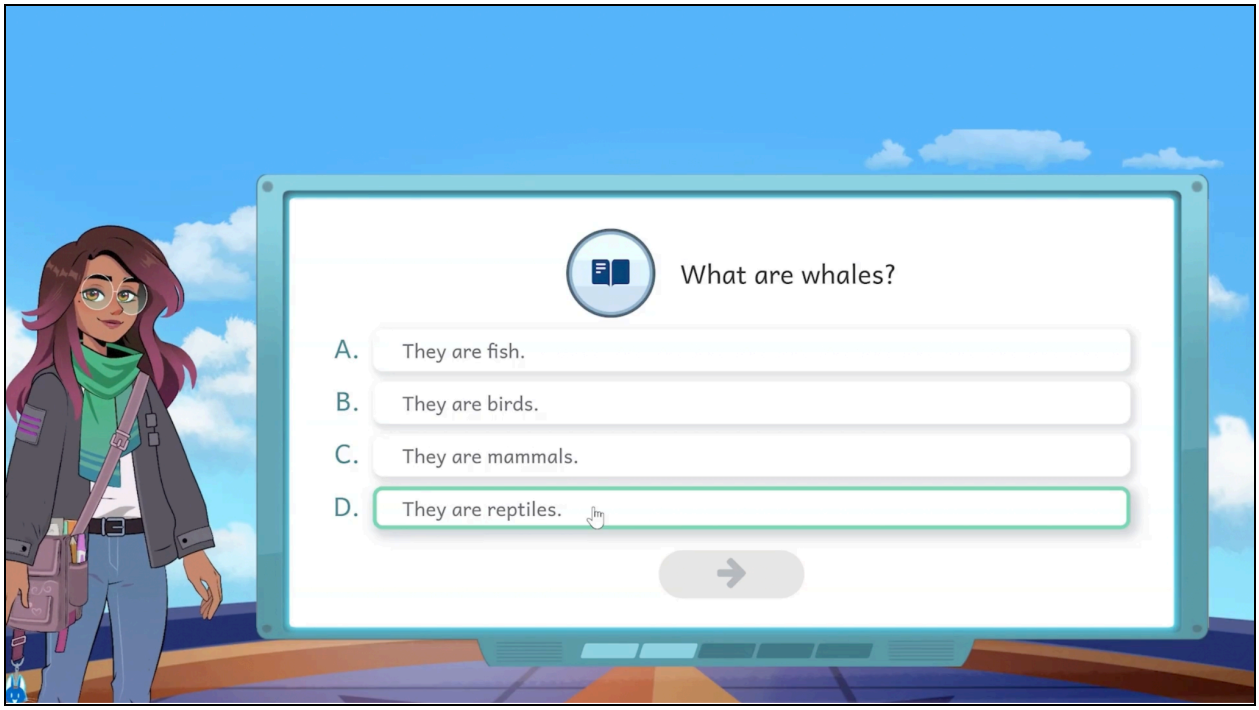


Figure 2.15 Screenshot of the Reading Comprehension task.

See an example video of the reading comprehension task here.

2.8 Oral Language

Amira ISIP uses up to three tasks to assess Oral Language ability in Kindergarten through grade 3. One of these tasks leverages Amira ISIP's unique AI-powered listening and comprehension capabilities, while the other two are traditional, well-established measures of oral language proficiency. The assessment can be customized to include any one or all three of these Oral Language tasks, depending on the preferences of the SEA or the district.

Task 1: Oral Language Vocabulary
Using a protocol similar to the standard PPVT (Peabody Picture Vocabulary Test), Amira ISIP

measures receptive vocabulary by having students identify pictures that represent the spoken word. In this task, the student is presented with a set of pictures, one of which "defines" the word provided by Amira, and the others serve as distractors. The student selects a picture employing a device-appropriate gesture. This approach, analogous to the Peabody, is widely used to measure receptive vocabulary, which is an essential component of oral language ability.

The task protocol is as follows:

1. A set of 4 picture options is shown on the screen
2. Amira speaks the word
3. The student selects a picture
4. If the selected picture corresponds to the spoken word, the item is marked correct.

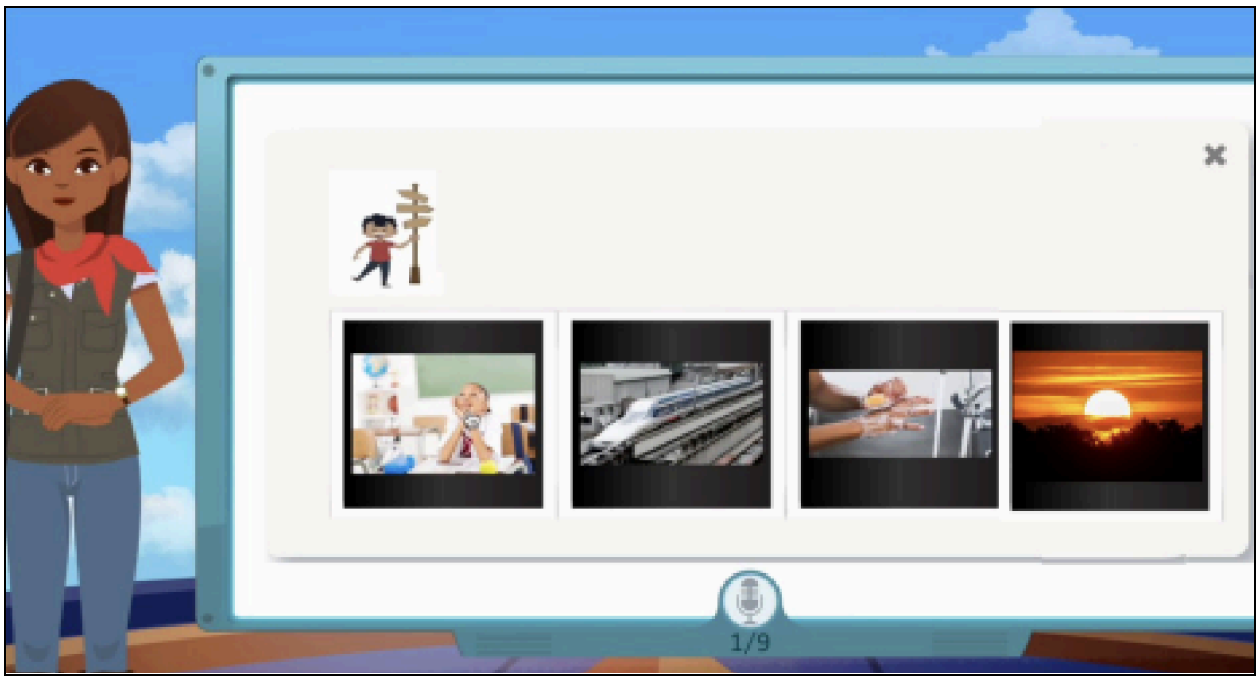


Figure 2.16 Screenshot of the Oral Language Vocabulary Task.

The Oral Language Vocabulary task requires minimal time to administer. A time-out feature is included, but it serves only to address non-responsive students, rather than acting as a timer. This task assesses receptive vocabulary and provides valuable insights into a child's language comprehension, a key predictor of future reading success.

Research has consistently demonstrated that picture vocabulary tests are robust measures of oral language ability. Studies have demonstrated that PPVT scores correlate strongly with other language assessments, including measures of expressive vocabulary and overall language competence (Dunn & Dunn, 2007). The task's strong psychometric properties, such as high test-retest reliability and content validity, have led to widespread acceptance as a measure of Oral Language. For example, Hoffman et al. (2011) found that PPVT scores predicted later reading comprehension, underscoring the test's importance in early

childhood literacy assessments. PVT has been widely used in large-scale studies, such as the Early Childhood Longitudinal Study, further validating its role as a key indicator of language development (Rathbun & Germino Hausken, 2001).

Task 2: Oral Language Comprehension Task

Amira ISIP features an Oral Language Comprehension task in which the student listens to a video where a teacher reads a short narrative passage aloud with prosody. After hearing the passage, the student answers up to five questions, which are also read aloud. This task does not involve any text presentation, reading, or the use of the alphanumeric keyboard. Answer choices are read aloud each time the student hovers over them with the mouse. No reading skills are required for this task.

The typical listening comprehension passage lasts between 60 and 75 seconds and tells a brief, character-driven story. Each passage is tailored to a specific grade level and calibrated for consistency across levels. The number and type of questions vary depending on the grade level of the passage and user preferences. After listening to the adult's reading, the student can choose to hear the passage again or proceed to the questions. The listening comprehension task is particularly recommended for kindergarten and grades 1 and 2.

The Listening Comprehension task is structured as follows:

1. Amira provides directions to the student.
2. A story is read aloud to the students without text, lasting 60 to 75 seconds. There is an option for the students to listen to the story again before answering the questions.
3. For each assessment item, Amira will read each question and answer choice out loud. The number and nature of the questions posed depend on the grade level of the passages. For example, for the question: “What does Tut do with the box on the bed?” The students would choose the answer choice “jumps in it” to correctly answer the question.
4. Amira scores the item automatically.

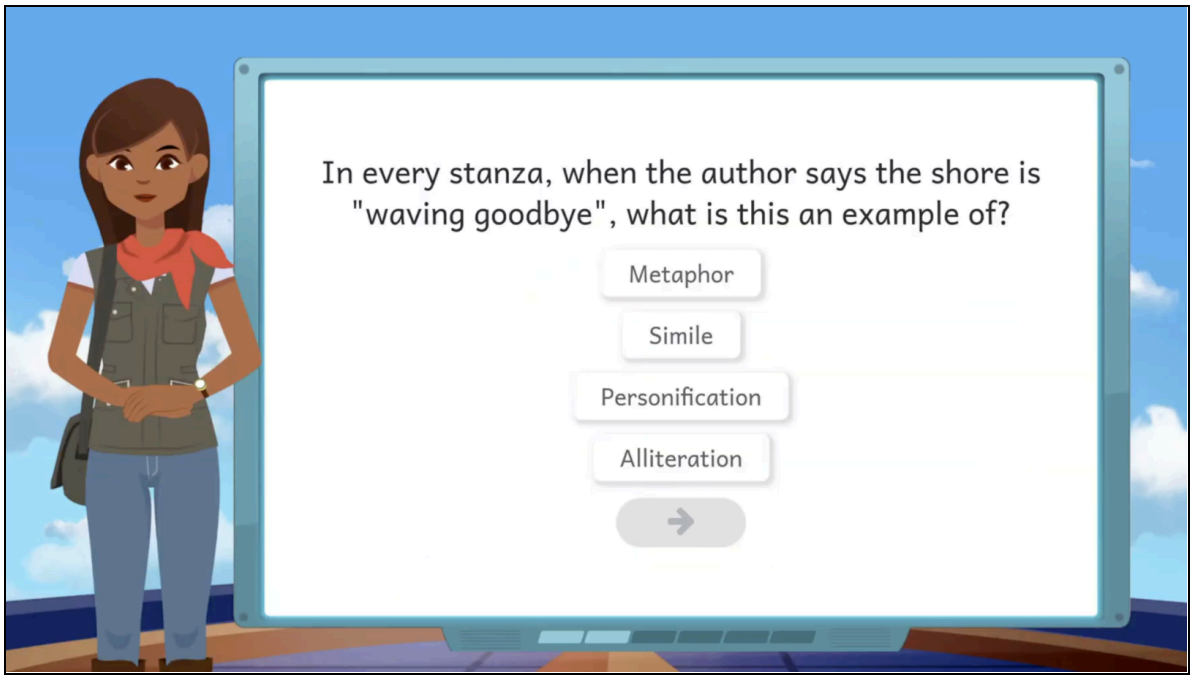


Figure 2.17 Screenshot of an Oral Multiple Choice Question
This task is shown after a spoken passage. All text is verbalized, so no reading is required.

Task 3: Oral Language Retell
Using Amira's capacity to listen and analyze spoken language, the Oral Language Retell task consists of the following protocol:

1. An actor conveys a very short narrative passage.
2. Amira asks the student to re-tell what they heard in their own words.
3. Amira collects the spoken language.
4. The Amira trained large language model then analyzes the retelling to determine which rubric elements the student conveyed

Five story rubric elements (basic setting, main character, major events, concrete details, conclusion), each scored independently and feeding the listening comprehension and comprehension scores. Syntax and grammar aren't scored, and a student isn't penalized for ungrammatical phrasing.

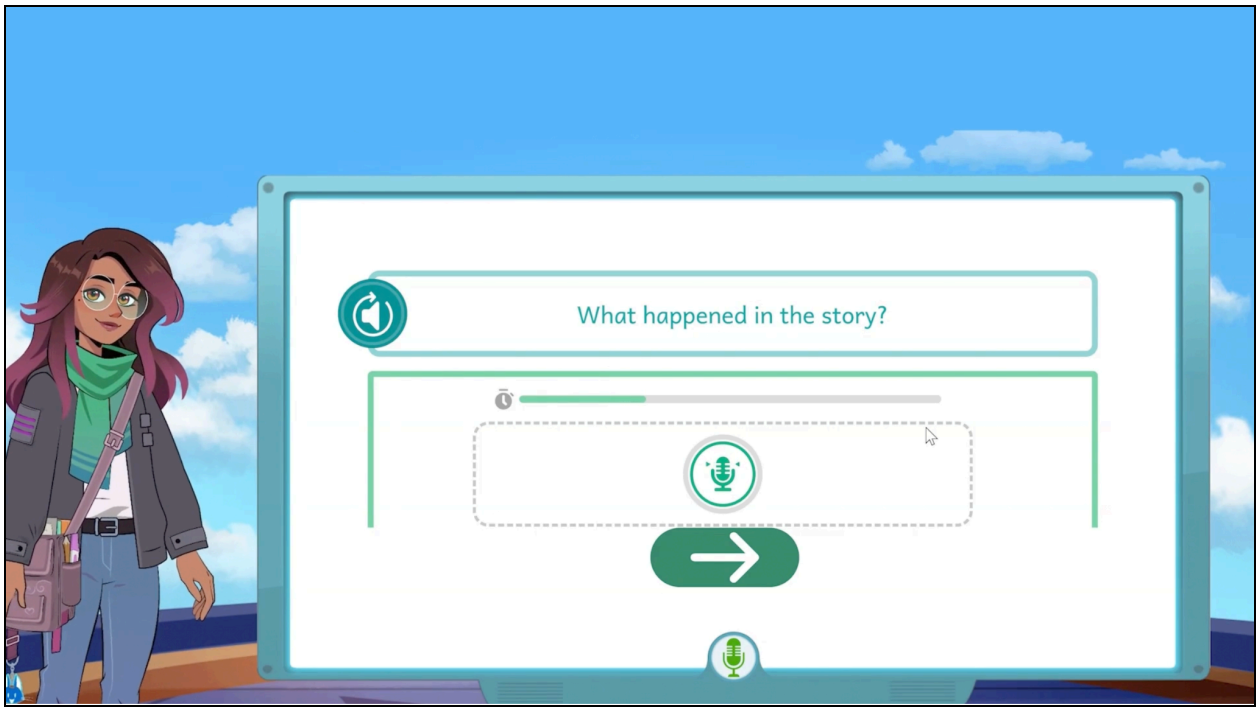


Figure 2.18 Screenshot of the Oral Language Retell Task.

In this task, the LLM analyzes a student's oral response to questions about a short story told aloud. The student re-tells the passage in their own spoken words.

See an example video of the oral retell task here.

Oral retell tasks are a powerful and effective way to assess a student's oral language proficiency. Amira ISIP's retell task asks students to listen to a story and then retell it in their own words. This allows the LLM models to evaluate key language skills with comprehension and recall of story content. The task provides a dynamic assessment of both expressive and receptive language abilities, offering insights into how well students can understand and organize language into coherent discourse.
Research has shown that oral retell tasks are particularly valuable for assessing the integration of listening comprehension with expressive language skills, making them an effective tool for identifying students at risk for language delays or reading difficulties (Morrow, 2005).

Studies consistently highlight the benefits of oral retell tasks in evaluating oral language proficiency. This approach not only assesses a child's ability to use narrative structure but also reflects their capacity to recall content. Strong oral retelling skills have been linked to better overall language and literacy outcomes, especially in reading comprehension and fluency (Snow, 2010). By capturing a comprehensive view of a student's language abilities, the oral retell task provides a robust measure of language mastery and helps guide targeted instructional interventions to support language development and improve reading achievement.

2.9 Rapid Automatized Naming

The Amira ISIP Benchmark administers a RAN task. The RAN task has been found to be a highly valid indicator of dyslexia risk (Denckla & Rudel, 1976; Wolf & Bowers, 1999) and highly predictive of the developmental trajectory of word-reading (word recognition) skills in Grades K, 1, and 2 (Boscardin et al., 2008).

Amira ISIP's RAN items were created by the University of Houston, and administration conforms to the methodology described in Jones, Branigan, and Kelly (2008) and Denckla and Rudell (1976). Amira ISIP can deliver three different forms of RAN—numbers, colors, and objects—ensuring the Benchmark relies on items within the general scope of a student's development and abilities. The purpose of the task is to assess speed and automaticity, not whether the students can identify the stimuli. That is, during a RAN task, children are not "reading" from the screen; they are recognizing images and naming them aloud. As such, they are expected to name the stimuli at a fairly rapid rate; children who struggle to do so may be at risk for dyslexia.

In all forms of the RAN task, the stimuli are likely to be known to children at very early ages. For example, in the number RAN task, the stimuli are numbers between 1 and 9, with zero always avoided. In the object RAN task, the stimuli used are line drawings of common objects (e.g., dog, book, hand, star). The Benchmark offers two dimensions of customizability: the RAN task type can be configured by the school district or by the software based on student ability. For example, if a student is unable to identify numbers, they can be administered the object- or color-RAN task.

The foundational output of the RAN task is the total processing time required to complete the task. Students are timed, and the system records the total time to completion, along with item-level accuracy. Amira ISIP also computes a RAN speed by dividing the number of items accurately named sequentially by the total processing time.



Figure 2.19 Screenshot of the object RAN Task

Note: The numbers 2, 3, 4, 5, 7, and 8 are used as distinct stimuli in this variant of the task.

RAN has strong predictive validity for differences in reading ability, and at least 104 studies published since 1990 have used it as a predictor of reading ability. Within each task type (numbers, colors, or objects), five or six different stimulus items are repeated in random order for a total of 36 stimuli, arranged in four horizontal rows of nine items each.

The RAN task is structured as follows:

1. Amira provides the student with directions to identify the stimuli from left to right, starting with the top line and moving line by line down to the bottom line, naming the items on each line from left to right.
2. Students are asked to go as fast and as accurately as they can.
3. Amira presents and demonstrates an abbreviated example consisting of six stimulus items per task type, repeated in random order, for a total of 18 stimuli, arrayed in two horizontal rows of nine items each.
4. The student then proceeds with the full task of 36 stimuli, arranged in four horizontal rows of nine items each.

In the abbreviated example Amira provides, she names the stimuli in the top row from left to right, then in the second row. Amira tells the students it's their turn and presents a screen with five or six stimulus items per task type, repeated in random order for a total of 36 stimuli, arranged in four horizontal rows of nine items each. The student reads the stimuli serially from left to right in each subsequent line, going from top to bottom. Total time to completion and item-level accuracy are recorded.
See an example video of the RAN task here.

2.10 Visual Attention

Amira ISIP uses a Visual Search task to assess Visual Attention, a crucial component of its screener, particularly for early grades. This task is designed to measure a student's ability to selectively focus on relevant visual information while disregarding distractions.

Visual attention is a cognitive ability that enables a student to selectively focus on relevant visual information while filtering out distractions. Effective visual attention is essential for key reading-related tasks such as efficiently scanning text, accurately identifying letters and words, and maintaining focus on the current word or letter while preparing for the next. Deficits in visual attention can significantly impact letter recognition, a key factor influencing reading skills and a predictor of reading difficulties in young children.

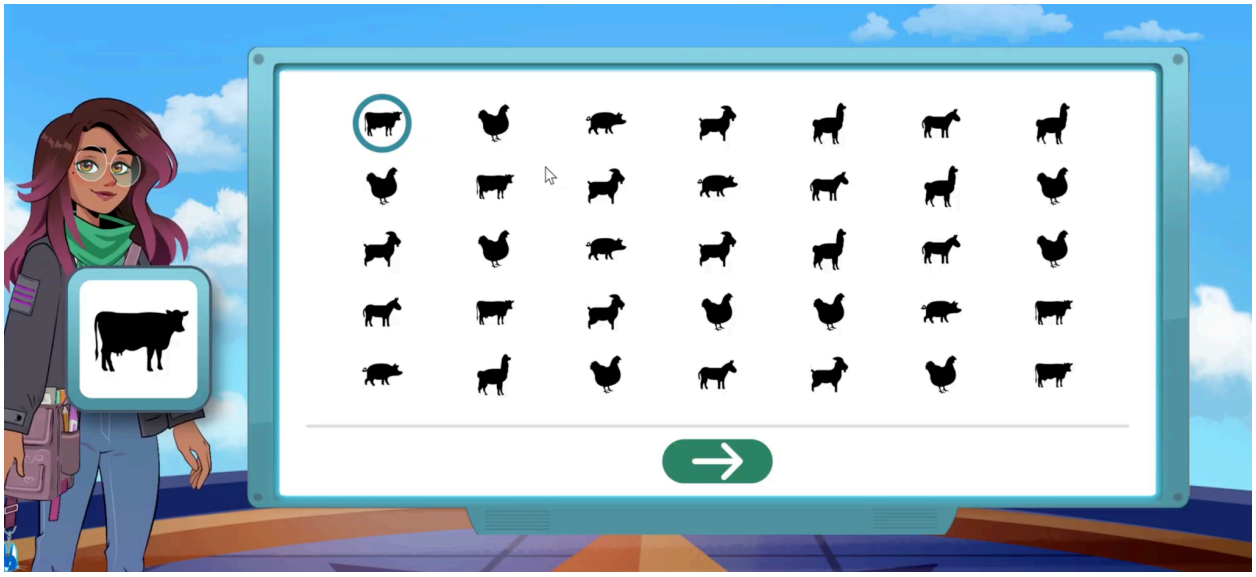


Figure 2.20 Screenshot of the Visual Attention Task

The visual attention task is structured as follows:

1. Students are presented with a visual display of numerous animal-like items arranged in a grid-like pattern across the screen.
2. Amira provides the student with instructions and a brief example, and the student can optionally request a repeat of the instructions.
3. The student's goal is to identify and mark specific target items within this display.
4. Students must systematically scan the display and click on the target items to identify them.
5. Students have 60 seconds to identify as many of the target items as possible. 6. This task is recommended for students in Kindergarten, Grade 1, and Grade 2. 7. Amira automatically scores all items and generates composite scores for time, accuracy, and completeness.

See an example of the visual attention task here.

3. TEST DESIGN

Amira Learning follows a rigorous, research-driven process for test development that aligns with best practices in educational assessment as outlined by the Council of Chief State School Officers (CCSSO) and the Standards for Educational and Psychological Testing (AERA, APA, NCME). This process ensures that all assessment items are valid, reliable, fair, and accessible, supporting accurate measurement of early literacy skills across diverse student populations. The development cycle engages educators and literacy experts at each phase, from initial item creation through review, field-testing, psychometric validation, acceptance, and continuous improvement to maintain assessment integrity and instructional relevance. With a dedicated records and version control system for item management, this process provides transparency, standardization, and detailed documentation at every stage of item development.

3.1 Assessment Blueprint and Design

The test development process begins with the creation of an assessment blueprint, which defines the content domains, cognitive complexity, and measurement objectives. Amira ISIP's assessment framework is grounded in research-based literacy models and national/state standards for grades pre-kindergarten through grade 8, ensuring alignment with instructional expectations. In addition, the blueprint was developed to reflect key theoretical frameworks associated with the science of reading and identification of reading difficulties, including the International Dyslexia Association, Multiple Deficits Model, and the Active View of Reading (see Amira's Theoretical Framework).

A research-based blueprint, grounded in national literacy standards, was developed to outline the required content domains and subdomains for a balanced assessment. The test design balances cognitive load and engagement, ensuring students can complete the Amira ISIP Assessment within reasonable time limits while remaining focused and motivated. During this phase, educators, literacy experts, and psychometricians collaborate to establish the assessment's construct map, identifying key skill areas such as phonemic awareness, decoding, fluency, vocabulary, and comprehension. A gap analysis is conducted to ensure full coverage of essential literacy competencies and identify any potential areas needing additional emphasis.

3.2 Item Development and Expert Review

Once the blueprint is established, Amira ISIP follows a structured item development process that incorporates principles of Universal Design for Learning (UDL), cognitive load theory, and bias/sensitivity considerations to ensure accessibility and fairness.

- **Recruitment and Training of Item Writers:** Item writers with expertise in early literacy, assessment design, bias and sensitivity guidelines, and Universal Design

(UD) principles, crafting items that meet diverse learner needs.
- **Guided Item Creation:** Items are developed to cover a range of difficulties and cognitive demands, ensuring appropriate alignment of national and state standards and avoiding bias through clear, accessible language. Item writers and reviewers are trained to apply UD principles, ensuring all items minimize construct-irrelevant barriers and are accessible to students with diverse needs.

Iterative Item Review Cycle

- **Preliminary Internal Review:** An initial internal review allows test developers to provide feedback on clarity, alignment, and developmental appropriateness, enhancing quality through collaborative review.
- **Subject Matter Expert Review:** Multiple rounds of internal expert review followed by evaluation for external experts to confirm content accuracy, alignment with standards, developmental appropriateness, and freedom from cultural, gender, or regional bias.
    - Each review committee typically comprised 8–12 members per grade band, including external educators and literacy researchers unaffiliated with Amira. External reviewers were state-nominated educators to ensure independence and objectivity.
    - Reviewers first rated items independently using standardized rubrics aligned to literacy constructs and Depth of Knowledge (DOK) levels, then met collectively to reconcile discrepancies and reach consensus.
    - Review summaries indicated ≥95% inter-rater agreement on content alignment, clarity, and developmental appropriateness.
- **Accessibility and Sensitivity Checks:** Items undergo thorough checks for compatibility with assistive technologies and for adherence to accessibility guidelines, ensuring readability and appropriateness.
- **External Bias and Sensitivity Review:** Partner states require a bias and sensitivity review for all items presented to students. Amira held workshops in August 2025 with educators. During these sessions, educators examined every item scheduled for the 2025–26 school year. Any item flagged during this process was excluded from test forms in the following school year. In total, less than 1% of items were removed as a result of this review.

3.3 Field-Testing and Psychometric Validation

Before becoming operational, test items undergo field-testing to collect empirical data on student performance. Amira ISIP follows a rigorous statistical validation process using Item Response Theory (IRT) and Classical Test Theory (CTT) to ensure items function as intended.

- Representative Sampling: Field tests include students from diverse backgrounds, language proficiency levels, and learning needs to ensure validity across populations.
- Classical Item Analysis: Classical item analysis identifies items that are functioning

as expected, including item difficulty and item discrimination. For multiple-choice items, distractor analysis is also conducted.

- Calibration with IRT: Items are calibrated using a 2-parameter model. IRT analysis identifies parameters for item difficulty, discrimination, and item fit. Items that do not fit the model, are too difficult or too easy, or have low discrimination parameters are flagged for revision or removal.
- Differential Item Functioning (DIF) Analysis: DIF testing ensures consistency in item performance across student subgroups, verifying that no item exhibits unintended bias toward any demographic group.
- Reliability and Consistency Checks: Items undergo tests of internal consistency and reliability, with results reviewed by psychometricians to confirm their accuracy.
- Validity Evidence Collection: Evidence is collected to confirm that each item measures the intended literacy skill accurately and is validated against comparable assessments.

Acceptance and Maintenance

- Stakeholder-Driven Acceptance Process: Final reviews involve subject-matter experts and stakeholders, who review each item's history, comments, and revisions. Committee feedback is used to adjust items, with flagged items refined or removed based on evidence. Approved items are documented, preserving a comprehensive record of all decisions.
- Norm Development and Maintenance: Normative data are regularly updated to reflect evolving student demographics, ensuring the Amira ISIP Assessment remains valid and equitable.
- Following successful field-testing, approved items are integrated into Amira ISIP's assessment system. The assessment engine dynamically adjusts item difficulty in real-time based on student responses, ensuring precise skill measurement. Item Bank Maintenance includes the following activities.
- Annual Item Development Plans (IDPs): Items are regularly reviewed and updated based on student performance data and educator feedback. Annual IDPs support the sustained growth of the item pool, addressing any identified gaps in content coverage and difficulty levels.
- Regular Item Pool Analysis and Refresh: The item pool is periodically refreshed based on comprehensive analyses.
- Longitudinal Validity Studies: Amira ISIP tracks student progress across multiple administrations to ensure that assessment results accurately reflect literacy growth.

Role of Educators and External Stakeholders in Review

1. Educator and Stakeholder Engagement
    - Item Review Committees: Subject matter experts serve on the Item Review, Bias, and Sensitivity Committees, providing expertise in content alignment, cognitive complexity, and developmental appropriateness.
    - Bias and Sensitivity Committees: Diverse representation on these

committees ensures that items are reviewed for fairness, accessibility, and cultural sensitivity, helping to avoid construct-irrelevant barriers in the ISIP assessment.

2. Training and Resources for Review Committees
    - Comprehensive Reviewer Training: Committee members receive detailed training on literacy standards, Depth of Knowledge (DOK) levels, item construction, national and state standards for literacy and content.
    - Ongoing Support for Reviewers: Facilitators support the committees, ensuring accurate documentation of decisions, which are logged for transparency and oversight.
3. Committee Review Process and Decision Logging
    - Outcome Documentation: Decisions to accept, revise, or reject items are logged within the records system, creating a permanent, auditable record.
    - Summary Reporting: Summary reports offer a comprehensive view of review outcomes, including committee recommendations and final approval status.

3.4 Content Management

A structured system of records and version control supports the development, review, storage, and publication of Amira ISIP Assessment items. This system incorporates review tracking to preserve the history and evolution of each item. This management solution records stakeholder feedback and maintains standardized reviews for each item. Each stage of development is documented, providing comprehensive records of all review actions and feedback for transparency and decision tracking.

Amira ISIP's test development process reflects the highest standards of educational assessment, ensuring validity, reliability, and fairness in measuring early literacy skills. By incorporating evidence-based design, psychometric validation, and continuous quality control, Amira ISIP provides a rigorous and equitable assessment system that supports data-driven instruction and student success.

3.5 Accommodations

Amira Learning is committed to providing equitable assessment materials and a program that is accessible to as many students as possible, regardless of their culture, background, learning needs, or disability. The platform is designed to make accommodations simple for teachers, ensuring every student has the opportunity to succeed. Many accommodations are built directly into the program and are universally available. For instance, all assessments and practice materials can be accessed in English only, Spanish only, or English with Spanish directions. Amira ISIP also allows additional time or breaks during assessments and can accommodate students taking assessments in both English and Spanish, either in separate administrations or all at once. Teachers can even configure whether English or Spanish is presented first. For ORF passages, Amira ISIP will automatically "downlevel" to an easier passage if the initial one is too difficult after 60 seconds.

Additionally, upon first login, Amira ISIP guides students through troubleshooting sound and voice issues, reducing the burden on teachers.

The general design of the Amira ISIP Assessment adheres to the principles of UDL, aiming to eliminate unnecessary hurdles and provide a flexible learning environment. This means information is presented in multiple ways, students can engage in learning in various ways, and they are provided options for demonstrating their learning. Amira ISIP Assessment met the Level AA standard under the Web Content Accessibility Guidelines (WCAG 2.0) in 2022, supporting best practices for user experience (UX) development, including features for visual or auditory impairments.

Beyond these built-in flexibilities, Amira ISIP offers several specific accommodations for students, especially when acknowledged by an IEP or 504 plan. These include:

- Spanish Proctoring for English Assessment: Allows students taking an English assessment to receive Spanish-language proctoring to ensure task understanding and provide an equitable opportunity.
- Spanish Screener: Enables students to be screened for dyslexia in Spanish to prevent disproportionate flagging of ELs and to identify reading mastery in Spanish.
- Paper-Based Administration: Provides a non-digital, teacher-administered test for students who cannot use the digital environment, offering an equivalent assessment.
- Braille test forms are available for visually impaired students, with both contracted and uncontracted formats offered in grades K–1 to accommodate individual readiness. Beginning in grade 2 through grade 8, only contracted Braille is provided..
- Paper-Based Spanish Administration: Similar to the paper-based option, but specifically for Spanish-speaking/reading students who cannot use the digital interface.
- Configure Time: Allows students to receive more time for the assessment.
- Retest: Enables re-assessment if an initial interaction is flawed due to environmental issues or if a student needs a trial run for readiness.
- Preparing Students: Ensures that students who need extra preparation are ready before the assessment.
- Removing Tasks: Reduces the time and complexity of the assessment process for students who require a less complex evaluation.
- Additionally, other allowable accommodations include providing a quiet setting for testing, small-group testing, breaks between tasks, assistive technology (e.g., hearing aids, glasses), enlarged materials (via screen magnification), colored overlays, filters, lighting adjustments, tracking devices, and whisper phones.

Review the Guidance for Assessment Accommodations for more information.

3.5.1 Attention to Linguistic Diversity

Amira ISIP equips teachers with a comprehensive toolkit designed to support students from diverse linguistic backgrounds and contexts. These resources come with clear, explicit

guidance to help teachers translate assessment results into actionable plans for differentiated, culturally sensitive instruction. The tools provided to address linguistic diversity include:

1. **Nationally Normed Measures**: These measures allow teachers to compare all students to the general population's mastery levels for fluency, comprehension, word recognition, and overall reading ability.
2. **Norms Specifically for EL**: Amira ISIP provides information on the same set of measures that are nationally normed for English Language Learners. This enables teachers to compare students against their peers and avoid over-response to the lagged mastery curve experienced by most English Learners.
3. **Spanish Language Diagnostic Screeners**: Amira ISIP offers Spanish language diagnostic screeners with their own national norms. This data enables teachers to differentiate between students who have foundational challenges with reading and those who simply lack exposure to English.
4. **Proficiency Scores**: Amira ISIP delivers proficiency scores analysis spanning the reading rope for:
    - Proficiency versus National Norms
    - Proficiency versus EL Norms
    - Proficiency in Spanish

By offering a 360-degree view of an EL's progress and needs, Amira ISIP provides explicit, clear guidance to teachers on how to use assessment data to help ELs accelerate their growth. Section 4.3 below describes the Amira ISIP approach in more detail.

English Assessment Norms for English Language Learners

We norm each year on the entire Spanish-English bilingual population who take Amira ISIP's English screener. These students are those who speak Spanish as their native language and constitute the vast majority of our EL population. The total sample size was 52,280 students across K-3 who were enrolled in both Spanish and English configurations of Amira ISIP, collected during the school year 2023–2024. Details on district/school count for each grade are shown in Table 3.1 below. Schools and districts came from a variety of states, with representation from every U.S. census region and school type (public, private, and charter).

Table 3.1: Unique students, districts, and schools contributing to the EL norms.

Grade	N	Number of Districts	Number of Schools
Kindergarten	4951	124	465
Grade 1	6714	160	569
Grade 2	7591	160	634
Grade 3	7029	165	615

The benchmarks for the Amira ISIP Screener composite score (the Amira ISIP Reading Mastery, or ARM Score), specifically for the EL population in Amira ISIP's 2023–2024 usage, are shown in Table 3.2 below.

Table 3.2 ARM Score benchmarks, for the EL population, on the Amira ISIP English Screener.

Grade	Term	≤ 30th	31st-74th	≥ 75th
Kindergarten	Fall	< -0.28	-0.28 - 0.27	> 0.27
Kindergarten	Winter	< -0.14	-0.14 - 0.29	> 0.29
Kindergarten	Spring	< 0.08	0.08 - 0.65	> 0.65
1st Grade	Fall	< 0.22	0.22 - 1.1	> 1.1
1st Grade	Winter	< 0.44	0.44 - 1.38	> 1.38
1st Grade	Spring	< 0.65	0.65 - 1.72	> 1.72
2nd Grade	Fall	< 1.16	1.16 - 2.06	> 2.06
2nd Grade	Winter	< 1.43	1.43 - 2.44	> 2.44
2nd Grade	Spring	< 1.5	1.5 - 2.83	> 2.83
3rd Grade	Fall	< 1.87	1.87 - 2.91	> 2.91
3rd Grade	Winter	< 2.18	2.18 - 3.17	> 3.17
3rd Grade	Spring	< 2.34	2.34 - 3.73	> 3.73

3.6 UX Studies

Amira Learning has undertaken a range of initiatives to ensure a positive UX for its assessment and reporting platform, guided by thoughtful design principles and continuous feedback loops to enhance usability, accessibility, and engagement for both students and educators.

For the assessment interface, a key component of this work is deliberate design that prioritizes engagement and efficiency. Amira ISIP's test design carefully balances reliability, validity, and test-taker fatigue, aiming to keep assessments concise—typically under 20 minutes—while still covering all necessary standards and maintaining user engagement. The student experience is designed to be comforting and engaging, featuring a friendly and patient Amira avatar that offers positive reinforcement, such as "keep going" or "good job,

now onto the next task." In addition, Amira harnesses artificial intelligence to provide real-time support grounded in the Science of Reading, contributing to an equitable and innovative user experience.

Amira assesses students primarily by having them read aloud. This approach leverages cutting-edge AI and speech recognition technology to provide an accurate, efficient, and comprehensive evaluation of reading skills. Instead of relying heavily on multiple-choice questions, Amira prioritizes production tasks using a student's natural voice that mirrors how students authentically learn to read. This method allows Amira to analyze how students decode, pronounce, and comprehend text using their own voices.

This is made possible by its highly accurate AI scoring technology. This approach ensures that the data teachers receive in reports reflects students' real, expressive reading abilities rather than just their test-taking skills.

3.7 Amira's Real-Time Calibration Design

Amira ISIP represents a new generation of educational assessment—combining the sophistication of artificial intelligence with the human-centered capability of listening to student reading. Amira ISIP is based on the direct observation of oral reading and makes in-test adjustments based on real-time performance. Like other modern assessments, Amira ISIP is scored consistently by software and utilizes modern psychometrics to weight item importance. In essence, Amira ISIP is a hybrid that brings together the strengths of earlier assessment models while transcending their limitations.

When innovative tools emerge, there is often a tendency to mistake them for familiar. While Amira is a digital assessment, it is not a traditional Computer Adaptive Test (CAT). And while it draws on fluency constructs and oral reading observations, it is not a traditional Curriculum-Based Measurement (CBM). Amira represents a third generation—intelligently combining and improving on both approaches.

Here is a high-level, holistic overview of how Amira's design compares with 1st-Generation CBM Assessments (e.g., DIBELS, Acadience, AIMSweb) and 2nd-Generation CAT Assessments (e.g., NWEA, iReady, Renaissance Star).

Table 3.3 Comparative Analysis of Amira's Assessment Design vs. First-Generation CBM and Second-Generation CAT Assessment

Design Attribute	Amira	Like	Unlike
Uses start and stop logic to collect appropriate information	Yes	DIBELS	
Uses Item Response Theory (IRT) to weight	Yes	iReady	CBMs

Design Attribute	Amira	Like	Unlike
items based on difficulty and discrimination			
Consistent scoring without human bias	Yes	iReady	CBMs
Listens to oral reading	Yes	DIBELS	CATs
Test Administration < 20 Minutes	Yes	DIBELS	CATs
Extracts non-binary information like response time	Yes		CATs CBMs
Enables state- and district-level configuration to achieve standards coverage	Yes		CATS CBMS
Enables teacher review and correction	Yes	DIBELS	CATs
Ensures all students receive the same foundational items	Yes	DIBELS	CATs
Minimizes lost instructional time	Yes	iReady	CBMs

This chart demonstrates how fundamentally different Amira's 3rd-generation design is from its predecessors. Amira is not a traditional CBM. Amira is not a CAT. Instead, Amira embraces and extends the important and key features of the old approaches while transcending their weaknesses.

Modern Psychometrics

Amira ISIP is grounded in IRT in order to accurately measure student mastery (see Section 4). IRT delivers on the principle of invariance and improves the accuracy of estimated ability. Per Baker's seminal "The Basics of Item Response Theory" (2001), IRT means "Results are not dependent upon the group of examinees, and the parameters that describe the examinee are not dependent on the test items. This property of *invariance of item and person parameters* is the single most important advantage of item response theory over classical test theory." This invariance enables improved precision. Per Embretson & Reise (2000), "The most important feature of item response theory is that the precision of measurement is not assumed to be the same for all examinees, as it is in classical test theory. Instead, the precision of measurement is a function of the examinee's trait level."

Two-Level Calibration Design: State/District Configurability & Start/Stop Rules

Amira ISIP's testing is designed around a powerful two-level decision-making process that ensures each student is tested appropriately.

Level 1: Configurable Construct Access

The first level of configurability is choosing the constructs most appropriate for each grade. Rather than administering every possible task type to every student, Amira enables educators to group constructs by grade into three categories:

- **Constructs considered essential**: administered to all students
- **Grade inappropriate tasks**: Constructs considered grade inappropriate and administered to no students
- **Constructs that are optional**: administered depending on the student's demonstrated readiness or psychometric situation

All students begin assessment with the benchmark tasks appropriate to their enrolled grade level, following standard screening protocols. This ensures:

- Consistent baseline measurement across all students
- Identification of students performing significantly above or below grade expectations
- Compliance with screening requirements for early identification

Level 2: Stop & Start Rules

In Amira, start and stop rules are predetermined within the assessment blueprint to ensure consistent administration across all students while maintaining efficiency and accuracy. These rules automatically control task initiation and discontinuation without administrator intervention, reducing student frustration and preventing error practice.

Stopping Rules

The assessment employs several stopping criteria to ensure efficient and appropriate testing.

Floor Rules:

- If a student scores at or below the floor threshold on the initial testlet, the assessment discontinues that task per standard protocol
- The student is referred to more developmentally appropriate prerequisite skill assessments

Ceiling Rules:

- Assessment continues until the student reaches the maximum performance ceiling or completes the full range of available difficulty levels
- Ensures accurate measurement of high-performing students

Standard Error Criterion:

- Testing stops when the standard error of measurement falls below a predetermined threshold, indicating sufficient precision has been achieved
- Balances accuracy with testing efficiency

Time Limits:

- Maximum administration time per task prevents extended testing sessions
- Maintains student engagement and reduces assessment fatigue

Consecutive Error Rules:

- Assessment may stop if the student makes a specified number of consecutive errors, indicating the material has exceeded their current instructional level

Starting Rules

Starting rules enable the collection of additional information when assessment tasks are insufficient for accurate psychometric differentiation. For example, if a student's performance places them on the cutline for dyslexia risk, an additional testlet may be triggered to provide more data.

These start and stop rules ensure that Amira employs a consistent baseline of items across all students while preventing frustration when tasks exceed a student's current developmental level.

Consistent Machine Scoring With Teacher Review

Amira ISIP employs consistent digital scoring. Research indicates that assessments that rely on human judgment are inevitably inequitable and subject to bias. Amira's models score student performance, oblivious to that child's demographics. However, unlike CAT tests, Amira's design allows for educator judgment to override the scoring, providing a safety net.

Bias Associated with Human Scoring

Thorndike (1920) first identified the halo effect phenomenon. In his and much subsequent research, the tendency of scorers to be influenced by previous experience with a testee is a substantial source of measurement error.

Even more concerning, a landmark meta-analysis by Brian A. Nosek (2002) and his colleagues on implicit bias revealed that most people, regardless of their own race, hold implicit biases that can affect their judgments and actions. More specific to education, research by David B. Quinn (2020) demonstrated that teachers' scoring was biased by their awareness of their students' race and ethnicity, with black students receiving lower scores from white teachers than they did from black teachers.

Minimizes Frustration and Lost Instructional Time

Like CAT tests and unlike CBMs, Amira can conduct and proctor many assessments simultaneously. Consequently, there is almost no instructional time lost. Equally importantly, like CBMs and unlike CATs, assessments take less than 20 minutes, minimizing student stress.

Accuracy: Rich Measurement

The most important and unique attribute of Amira's third-generation design is the ability to

collect an enormous number of measurement points in short order. Unlike either CAT or CBM, Amira transcends the primitive notion of binary item outcomes. Just as in assessing math mastery, it is vital to observe the student's reasoning and logic; in assessing reading mastery, a test should assess reading production. Amira measures speed, automaticity, self-correction, prosody, and a range of other "audio features" to more accurately gauge a student's ability.

Amira's real-time calibration design achieves measurement efficiency through systematic starting and stopping rules, reducing testing time by 40-60% compared to paper-based assessments while maintaining equivalent reliability. The application of these rules maintains appropriate levels of challenge for students across the ability spectrum. This approach provides granular estimates of ability within specific skill domains, enabling the identification of targeted intervention areas. The assessment accommodates the full range of student abilities through systematic application of these rules, ensuring both floor and ceiling effects are appropriately managed to maximize measurement precision across all performance levels.

3.8 Administration

Amira ISIP is designed for flexible administration in 1:1, small group, or whole-class settings, depending on district preference and student needs. The assessment is computer-based and requires only that students have access to a device with a microphone. Students read aloud while Amira listens, prompts, and scores automatically in real time. This design eliminates the need for teacher-led timing or scoring and minimizes administration burden. Amira's self-proctoring and self-scoring capabilities eliminate the need for specialized training or certification among educators administering the assessment. Teachers simply launch the assessment session and monitor student engagement while Amira manages delivery, timing, and scoring. Each administration typically takes approximately 20 minutes per student from start to finish, ensuring minimal instructional disruption.

Amira ISIP serves as both a Universal Screener and a Dyslexia Risk Screener, identifying students at risk for reading difficulties or dyslexia through adaptive tasks that measure phonological awareness, rapid automatized naming, decoding, and oral reading fluency. Results are immediately available upon completion, allowing teachers to review risk levels and plan interventions without delay.

Amira ISIP is generally administered at regular intervals throughout the school year. Amira ISIP supports 12 testing periods in a school year.

- Fall - August 1st through November 30th
    - Periods 1 through 4
- Winter - December 1st through March 31st
    - Periods 5 through 8
- Spring - April 1st through July 31st

- Periods 9 through 12

Amira ISIP Assessments are typically administered three times a year during designated testing windows. Periods 1, 5, and 9 align with the BOY, MOY, and EOY testing windows, respectively. These are known as Benchmark Assessments and are used to evaluate a student's overall reading ability, including fluency, vocabulary, and foundational skills.

Some districts may conduct additional administrations within each period to enable more frequent monitoring, especially for students receiving interventions. The specific timing and frequency of these administration periods can vary by school district, state requirements, and local policies.

4. MEASUREMENT MODEL

4.1 IRT Model

Amira ISIP employs a sophisticated item calibration process for its reading assessment items, which is essential for placing items on a difficulty continuum and providing meaningful data for educational decision-making. This process is foundational to Amira ISIP's CAT model, which dynamically adjusts the difficulty of test items based on a student's responses in real-time.

Following item design and field testing, items are calibrated using a 2-parameter logistic (2PL) Item Response Theory (IRT) model (Hambleton, Swaminathan, & Rogers, 1991). The 2PL is widely used to describe the relationship between a person's latent ability and their probability of correctly answering a test item, defined by the following equation.

$$P(X = 1|\theta, a, b) = \frac{1}{1+exp(-a^*(\theta-b))}$$

Where $P(X = 1|\theta, a, b)$ is the probability of a correct response, θ (theta) represents the student's latent ability, *a* is the discrimination parameter of the item, and *b* is the difficulty parameter of the item.

The 2PL model assumes that items vary in both their difficulty and their ability to discriminate between different ability levels. The discrimination parameter determines how well an item differentiates between people with different ability levels. Higher values indicate that the item is better at distinguishing between high- and low-ability individuals. The difficulty parameter represents the ability level at which a student has a 50% probability of answering the item correctly. Items with higher values are more difficult and require greater ability to succeed.

4.2 Item Calibration

The calibration process requires applying the 2PL measurement model to a dataset and estimating person and item parameters simultaneously. Student responses from field testing are calibrated using specialized software (R with the mirt package) before being incorporated into the item pool. Importantly, items being calibrated are not used in generating student scores during this phase.

4.2.1 Data Collection and Processing

The calibration process begins with comprehensive data collection from operational assessments. Data is systematically extracted from Amira ISIP's data storage systems, targeting specific assessment windows and student populations using structured SQL

queries.

The system processes various assessment types, each requiring specialized processing due to different response formats and scoring mechanisms. Raw assessment data undergoes sophisticated processing to extract item-level responses. Lastly, multiple validation steps are taken to ensure data integrity.

4.2.2 Calibration Process

The calibration process follows a systematic approach. Response data is prepared and organized into a wide-format matrix. Special handling is applied for different item types. Items with insufficient sample sizes are excluded from the calibration to ensure stable parameter estimates. All current items had a sample size of at least 5,000 responses. The 2PL model is fitted using maximum likelihood (ML) estimation. Items are rigorously analyzed to determine how well they fit the chosen IRT model, identifying misfitting items that don't perform as expected. Items are examined for potential bias or DIF to ensure they perform equivalently across different subgroups (e.g., gender, ethnicity). Items showing significant DIF are reviewed, revised, or discarded to ensure fairness and effectiveness for all students, including those with diverse accents, dialects, or developmental speech patterns.

As items are used in operation, their performance is continuously monitored, enabling recalibration or replacement as needed. Calibrated items are stored in an item pool, categorized by difficulty and discrimination, which forms the basis for Amira ISIP's selection algorithm.

4.3 EAP Scoring

Amira ISIP employs Expected A Posteriori (EAP) estimation for estimating student latent trait levels (θ). EAP is a Bayesian method for estimating θ in IRT that treats the student's latent ability (θ) as a random variable with a prior distribution and computes the posterior expected value given the student's response pattern. A standard normal distribution N(0,1) is assumed for the prior distribution.

$$\widehat{\theta_{EAP}} = \int_{x} \theta * P(\theta|X)\, d\theta$$

Where $P(\theta|X)$ is the posterior distribution of ability θ given the response vector X. The integral computes the weighted average of all possible values of ability θ, weighted by how probable each value is given the data. EAP considers the entire posterior distribution rather than just finding the mode, resulting in stable estimates. EAP estimation is widely used and well-suited to handling short test lengths and extreme response patterns.

The EAP scoring system is implemented using a sophisticated Python-based framework that incorporates several key features. Rather than using a standard normal prior for all students, Amira ISIP employs grade-specific prior means that reflect typical ability levels for

each grade.

All grades use a standard deviation of 1.0 for the prior distribution. These grade-specific priors are empirically derived from large-scale calibration studies and reflect the expected growth trajectory in reading ability across grade levels. EAP estimates are computed using a theta grid ranging from -5 to 5 with increments of 0.05, providing high precision while maintaining computational efficiency.

In addition to overall ability estimates, Amira ISIP computes subscore estimates for specific skill domains. Subscore estimates use the student's overall theta estimate as the prior mean, providing more precise domain-specific estimates while maintaining coherence with overall ability. Items may contribute to multiple subscores based on their content, allowing for comprehensive skill assessment across overlapping domains. The system tracks which specific items contribute to each subscore estimate, enabling transparent interpretation of domain-specific results.

The EAP method is implemented using custom algorithms optimized for Amira ISIP's specific requirements, providing both ability estimates and their associated measurement precision with enhanced computational efficiency compared to standard implementations.

4.4 Vertical Scaling

Vertical scaling is the process of mapping performance across various test (or grade) levels onto a single scale score (Kolen & Brennan, 2010). It allows measurement of student ability across grade levels using a common scale. This allows for comparability across grades and longitudinal growth tracking.

4.4.1 Scaling Methodology

Amira ISIP employs a common-item design approach for vertical scaling, in which items administered across multiple grade levels serve as linking items to establish relationships between grade levels. The vertical scaling process involves several key steps:

1. Base Scale Establishment: Data from the foundational grades (Kindergarten and Grade 1) are combined to establish a unified base scale through concurrent calibration.
2. Within-Grade Calibration: Higher grade levels (Grades 2-5) undergo separate within-grade calibrations to establish grade-specific item parameters.
3. Linking Transformation: The Stocking-Lord linking method is used to transform higher-grade parameters to the common base scale established for Grades K-1.

4.4.2 Stocking-Lord Linking Procedure

The Stocking-Lord method is implemented through an iterative process to ensure optimal

linking:

1. Initial Anchor Set: Common items between grade levels are identified as potential linking items.
2. Iterative Evaluation: An iterative process (maximum 5 rounds) evaluates the quality of linking items using two criteria:
    a. Beta Differences: The difference in marginal probabilities between transformed and base scale parameters must be negligible.
    b. d^2 Distance: The Euclidean distance (d^2) between transformed and base scale parameters must be negligible.
3. Anchor Set Refinement: Items that do not meet both criteria are removed from the anchor set, and the linking is re-estimated using the refined set.
4. Convergence: The process continues until all remaining anchor items meet both quality criteria.

The final Stocking-Lord transformation is defined by constants A (slope) and B (intercept), where:

- Transformed discrimination: discrimination_original/A
- Transformed difficulty: (B + A) * difficulty_original

These constants are preserved and applied consistently across all operational assessments to maintain scale consistency.

4.4.3 Quality Assurance

The vertical scaling process includes comprehensive quality checks. Item Characteristic Curves (ICCs) are plotted and compared between transformed and base scale parameters for all linking items. Multiple calibration studies validate the consistency of linking transformations. See the following section detailing Amira ISIP's large-scale calibration studies. Lastly, when multiple parameter estimates exist for the same item, selection is based on parameter precision (minimum standard error) and the sample size difference between estimates.

4.5 Calibration Studies

Amira Learning conducted a large-scale calibration in January/February 2025. This study was a crucial part of Amira ISIP's robust item calibration process, designed to ensure the accuracy of item parameters along the scale. The primary goal of this study was to calibrate new and existing items planned for operational use in 2025 as well as support vertical scaling across grade levels. It also aimed to establish a link between the new scale scores and legacy scores, ensuring consistency between old and new assessment results.

Study Design

The study involved a nationwide field test. A large and diverse sample of over 55,000 students across grades Kindergarten to 5 was utilized. The sample was designed to represent the demographic characteristics and ability range of the national student population. See Table 4.1 for sample size by grade level.

Table 4.1 Calibration Study Sample Size by Grade

Grade	N Student
K	9,632
1	10,982
2	11,927
3	9,421
4	6,951
5	6,807
Total	**55,720**

Additional data from an April pilot study (approximately 500 students per grade) were used to validate the initial linking functions established from the January/February calibration study. Statistical comparisons showed that the linking functions from the April data were nearly identical to the original linking functions, validating the use of the established links.

The calibrated items from this study are now part of Amira ISIP's item pool, categorized by their difficulty and discrimination. These calibrated items form the foundation of Amira ISIP's CAT algorithm, which dynamically adjusts test difficulty in real time based on a student's performance to provide an optimal challenge and accurately measure their abilities. Items may be revised, re-field tested, removed, or assigned to operational forms after review.

Amira Learning's assessment platform includes a calibrated item bank of over 4000 items spanning Grades K–8 across ten literacy domains. As such, we maintain IRT item parameter values in our bank. A summary of the IRT-based item difficulty (b-value) and item discrimination (a-parameters) in the tables below.

Average IRT Item Difficulty (b-values) by Grade and Subtest

Grade	ORF	Decoding	PA	RC	Spelling	Letter Knowledge	Vocab	Listening	Overall
K	0.40	0.63	-0.12	0.64	—	-0.47	-0.22	1.04	**0.27**
1	0.66	1.05	0.73	1.33	1.42	—	0.47	1.78	**1.06**
2	1.06	1.77	1.10	1.86	2.32	—	0.74	—	**1.47**
3	1.41	2.45	—	2.25	2.93	—	1.22	—	**2.05**
4	1.68	2.16	—	2.65	4.89	—	1.03	—	**2.48**
5	1.61	—	—	3.61	5.13	—	1.41	—	**2.94**
6	2.30	—	—	3.41	4.40	—	—	—	**3.37**
7	2.18	—	—	3.73	4.74	—	—	—	**3.55**
8	3.48	—	—	4.50	4.90	—	—	—	**4.29**
Overall	**1.59**	**1.45**	**0.09**	**2.83**	**4.26**	**-0.47**	**0.73**	**1.56**	**2.39**

Note: ORF=Oral Reading Fluency, PA = Phonemic Awareness, RC = Reading Comprehension, Vocab=Vocabulary, LC=Listening Comprehension

Average IRT Discrimination (a-parameter) by Grade and Subtest

Grade	ORF	Decoding	PA	RC	Spelling	Letter Knowledge	Vocab	Listening	Overall
K	1.92	1.44	1.50	1.27	—	2.37	0.53	1.00	**1.43**
1	2.07	1.51	0.93	1.67	1.99	—	0.77	1.00	**1.42**
2	2.20	1.55	0.74	1.17	2.01	—	0.68	—	**1.39**
3	2.16	1.47	—	1.23	1.82	—	0.69	—	**1.48**
4	2.11	1.63	—	0.91	1.30	—	0.81	—	**1.35**
5	1.84	—	—	0.73	1.19	—	0.70	—	**1.12**
6	2.28	—	—	1.09	1.64	—	—	—	**1.67**
7	2.19	—	—	1.03	2.33	—	—	—	**1.85**
8	2.87	—	—	1.90	3.45	—	—	—	**2.74**
Overall	**2.15**	**1.51**	**1.37**	**1.29**	**1.84**	**2.37**	**0.70**	**1.00**	**1.74**

Note: ORF=Oral Reading Fluency, PA = Phonemic Awareness, RC = Reading Comprehension, Vocab=Vocabulary, LC=Listening Comprehension

Equipercentile Linking
Since the theta scale from the calibration study is a newly established measurement scale, it is necessary to apply an equipercentile linking method to align it with the legacy ISIP scale. This linking process leverages a group of common students—those who have taken both the legacy ISIP assessments and the new calibration study assessments. By comparing score distributions among these shared participants, equipercentile linking enables mapping the new theta scale onto the existing ISIP scale, ensuring continuity and interpretability across assessment versions.

4.6 Differential Item Functioning

Amira ISIP employs comprehensive differential item functioning (DIF) analysis, using multiple methodologies and large-scale samples, to ensure that assessment items function equitably across demographic groups. Two complementary DIF approaches—IRT-based analysis during item calibration (2020) and Zumbo & Thomas effect-size analysis on operational data (2023–24)—provide robust evidence that items measure reading ability consistently across gender, race/ethnicity, socioeconomic status, and other demographic dimensions.

4.6.1 IRT-Based Differential Item Functioning Analysis (2020)

Methodology
Item response theory models were employed for detecting DIF. IRT permits comparisons of item functioning between groups by assessing the probability that the performance of each item differs between groups at the same level of ability.

To conduct these analyses, an IRT model was constructed to estimate item parameters for each group of interest (e.g., ethnicity) and to compare the parameters obtained with those from a model that ignores group membership. If the models are not different, this indicates that the differences between groups on an item are best explained solely by ability and that group membership does not contribute to differential performance on the item. This would indicate that the item is not biased.

Statistical Framework
It is important to recognize that some items will show evidence for DIF solely by chance. To account for this, the goal was to keep the total number of items indicating DIF at or below 5%, the expected level of false positive findings across a large item pool.

Study Window: End-of-Year (EOY) Winter 2020 administration
Grade Levels: Kindergarten, Grade 1, Grade 2
Sample Sizes by Grade:

- Kindergarten: n = 686–743 per subtest
- Grade 1: n = 694–735 per subtest
- Grade 2: n = 682–694 per subtest

Total Items Analyzed:

- Kindergarten: 124 items across 14 subtests
- Grade 1: 123 items across 15 subtests
- Grade 2: 100 items across 10 subtests
- **Total: 347 items**

Results: Kindergarten

Table 4.2 shows the DIF results for kindergarten. Out of the 124 items on the original calibration Kindergarten form of the Amira ISIP Assessment, two (1.61%) showed gender differences. Six items (4.84%) showed differences between Blacks and Whites. There was one item that showed DIF for the White/Hispanic analysis (< 1%). For the two items that showed gender DIF, both favored girls. The overall rate was at or below 5.0% for each analysis. As is typical with many DIF investigations, different subtests showed patterns of DIF relative to prior test administrations. The items exhibiting DIF were removed from the original calibration set.

Table 4.2 DIF Results for Grade K Students

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Letter Name Fluency	743	0	0	0	10
Letter Sound Fluency	743	0	2	0	10
Pseudo-word/Non-word Decoding	743	1	0	0	8
Vocabulary	743	0	0	0	8
Phonological Awareness					
Segmentation Initial Sound	686	0	1	0	5
Segmentation - Final Sound	687	0	0	0	5
Phoneme Blending	686	0	0	0	5

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Deleting Initial Sounds Task	689	0	1	1	5
Deleting Final Sounds Task	689	0	0	0	5
Rapid Automatized Naming	501	1	0	0	36
Spelling/Encoding	224	0	0	0	6
Listening Comprehension	230	0	2	0	6
Word Reading	689	0	0	0	10
Reading Comprehension	228	0	0	0	5
Total		**2**	**6**	**1**	**124**
Total Percent		**1.61**	**4.84**	**0.08**	

Interpretation: The minimal DIF rates indicate no systematic bias by gender or race/ethnicity. Items exhibiting DIF were removed from the operational assessment to ensure that the Kindergarten form used in schools is as equitable as possible.

Results: Grade 1

The details of the Grade 1 DIF analysis are shown in Table 4.3. Only two of 123 original calibration items demonstrated DIF between genders on the Grade 1 Amira ISIP Dyslexia Screener. One item favored males, and the other favored females. In the White/Black comparison, three items showed DIF, with two of them favoring White students. For the White/Hispanic comparison, five items showed DIF, three of which were advantageous to Hispanic students. The overall DIF rates for each analysis were below 5% across all comparisons. The items exhibiting DIF were removed from the original calibration set of items.

Table 4.3 DIF Results for Grade 1 Students

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Letter Name Fluency	731	0	0	0	10

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Letter Sound Fluency	731	0	0	0	10
Pseudo-word/Non-word Decoding	735	0	0	2	12
Phonological Awareness					
Segmentation Initial Sound	694	0	0	1	5
Segmentation - Final Sound	731	0	0	0	5
Phoneme Blending	694	0	0	0	5
Deleting Initial Sounds Task	694	0	0	0	5
Deleting Final Sounds Task	694	1	0	1	5
Rapid Automatized Naming	694	0	1	0	36
Word Reading					
Set 1	694	0	0	0	5
Set 2	694	0	1	0	5
Set 3	694	1	1	1	5
Set 4	694	0	0	0	5
Comprehension					
Reading Comprehension	339	0	0	0	5
Listening Comprehension	322	0	0	0	5
Total		**2**	**3**	**5**	**123**
Total Percent		**1.62**	**2.44**	**4.00**	

Interpretation: Gender DIF showed a balanced pattern (one item favoring males and one favoring females), indicating no systematic gender bias. Black/White DIF was minimal at 2.44% with no consistent directional pattern. White/Hispanic DIF was at 4.00%, with three of five items favoring Hispanic students, again indicating no systematic advantage to either group. Items exhibiting DIF were removed from the operational assessment.

Results: Grade 2

A similar pattern was apparent for second grade, shown in Table 4.4. Of the 100 calibration items investigated, one item displayed DIF by gender; five items showed DIF in the White/Black comparison, with two favoring Blacks; and three items showed DIF in the White/Hispanic comparison, with one item favoring Hispanics. The overall flagging rates were 5% or less for each analysis, with no systematic pattern of DIF or bias of significant concern. The items exhibiting DIF were removed from the original calibration set.

Table 4.4 DIF Results for Grade 2 Students

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Graphophonemic Knowledge					
Set 1	682	1	1	0	5
Set 2	682	0	2	0	5
Set 3	682	0	0	0	5
Set 4	682	0	0	0	5
Word Reading					
Set 1	694	0	0	0	5
Set 2	694	0	1	2	5
Set 3	694	0	1	1	5
Set 4	694	0	0	0	5
Rapid Automatized Naming	694	0	0	0	36
Comprehension					
Listening Comprehension	470	0	0	0	12

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Reading Comprehension	370	0	0	0	12
Total		**1**	**5**	**3**	**100**
Total Percent		**1.00**	**5.00**	**3.00**	

Interpretation: Gender DIF was minimal at 1.00%. Black/White DIF was at the 5% threshold, with two of five items favoring Black students, indicating balanced directionality. White/Hispanic DIF was at 3.00%, again showing no systematic bias against any group. Overall Conclusion (IRT-Based Analysis): There is no evidence of systematic item bias by ethnicity or gender for any of the forms used in the Amira ISIP Assessment. The overall DIF rates for any specific comparison were consistently 5% or below. The items affected tend to be on different tasks, supporting the absence of systematic bias by item or task. Finally, items exhibiting DIF in the analysis were removed from the screener as it stands today.

4.6.2 Contemporary Zumbo & Thomas DIF Analysis (2023–24)

Methodology

To provide comprehensive, ongoing bias monitoring beyond the initial 2020 development phase, Amira conducted a contemporary Differential Item Functioning (DIF) analysis using the Zumbo & Thomas (ZT) classification system with logistic regression on a large, nationally representative sample of operational data.

Analytical Approach

The Zumbo & Thomas method compares item performance across demographic groups using effect-size measures (Nagelkerke's R^2), which provides both statistical and practical significance information. Items are classified using R^2 thresholds:

- **Category A (Negligible DIF):** $R^2 \leq 0.13$ — Item functions equivalently across groups
- **Category B (Slight-to-Moderate DIF):** $0.13 < R^2 \leq 0.26$ — Item shows some differential functioning; expert review recommended
- **Category C (Moderate-to-Large DIF):** $R^2 > 0.26$ — Item shows meaningful differential functioning; flagged for expert review

Sample and Scope

Study Window: 2023–24 school year, ongoing operational assessments
Grade Span: Pre-K through Grade 8 (emphasis on K–3 for this submission)
Total Items Analyzed: 2,598 items across 10 subtests

Subtests Covered:

- Phonological Awareness
- Letter Sound Correspondence
- Decoding
- Spelling/Encoding
- ORF/Passage Reading
- Word Identification
- Reading Comprehension
- Listening Comprehension
- Vocabulary
- Rapid Automatized Naming (RAN)

Sample Size: 65,000 students from three large districts across three states

Quality Threshold: Items with fewer than 100 responses excluded from analysis

Matching Criteria: DIF was examined using two independent matching criteria:

1. Overall reading ability scores (ARM composite)
2. Subscale scores (domain-level ability)

Using multiple matching criteria reduces Type I error and validates findings across different ability measurement approaches. Pearson correlations between overall and subscale scores ranged from 0.72 to 0.90 (R^2 = 0.52–0.81), indicating strong construct coherence.

Demographic Groups Analyzed:

- Gender: Male and Female students
- Race/Ethnicity: White, Black/African American, and Hispanic/Latino students
- Socioeconomic Status: Assessed using available demographic markers

Results Summary:
The contemporary DIF analysis revealed exceptionally low levels of differential item functioning across all demographic subgroups:
Item Classification Results:

- 99% of items: Negligible DIF (Category A, $R^2 \leq 0.13$) — ~2,573 items retained
- <1% of items: Slight-to-moderate DIF (Category B, $0.13 < R^2 \leq 0.26$) — ~20 items flagged
- <1% of items: Moderate-to-large DIF (Category C, $R^2 > 0.26$) — ~5 items flagged
- Items Removed: Approximately 0.20% (~5 items) identified as exhibiting construct-irrelevant bias

Expert Review Process:
All flagged items (Categories B and C) underwent structured expert review by Amira curriculum specialists. Reviewers determined whether observed DIF reflected:

1. Construct-irrelevant variance (genuine bias requiring removal)
2. Legitimate construct-related performance differences (expected and acceptable)

Reviewers evaluated:

- Item content and cultural appropriateness
- Reading level and linguistic demands
- Potential barriers created by the item context
- Whether differential performance reflected ability differences vs. measurement bias

Expert Review Outcome:
Approximately 0.20% of items (~5 items) were identified as exhibiting genuine construct-irrelevant bias and were removed from the operational pool. All remaining flagged items were deemed to measure legitimate reading constructs and to show acceptable differential performance.

Interpretation:
The removal of biased items ensures that the final assessment maintains technical rigor and equity across diverse student populations. The extremely low percentage of items requiring removal (0.20%) demonstrates that the vast majority of items function equivalently across demographic subgroups, providing strong evidence for test fairness. Items showing differential performance almost always reflect legitimate differences in the underlying construct (reading ability) rather than measurement bias.

Construct Validity Support: Pearson correlations between overall and subscale scores ranged from 0.72 to 0.90 (R^2 = 0.52–0.81), indicating strong construct coherence and supporting the validity of the matching-criteria approach.

4.6.3 Consistency Across Time and Methodology

Both the IRT-based DIF analysis conducted during assessment calibration (2020) and the contemporary Zumbo & Thomas analysis (2023–24) yield consistent findings:

- **2020 IRT-Based Analysis:** ≤5% DIF flagging rate across all grades (K–2) and demographic groups; biased items removed during calibration
- **2023–24 Zumbo & Thomas Analysis:** 99% of 2,598 items showed negligible bias; <1% flagged; 0.20% removed

This consistency across five years, two distinct methodologies, and different sample sizes provides strong triangulated evidence that Amira ISIP assessments function equitably across demographic subgroups. The findings demonstrate that fairness validation is built into assessment development from the calibration phase forward and is maintained through ongoing operational monitoring.

4.6.4 Compliance with Fairness Standards

Amira's comprehensive DIF analysis aligns with the standards established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014). All demographic groups examined (gender,

race/ethnicity, socioeconomic status) show minimal differential item functioning, confirming that assessment results are comparable and fair for all students regardless of background.

References

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

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for analyzing test bias. *Educational and Psychological Measurement, 57*(6), 900–913.

5. SCORING & REPORTS

Amira ISIP provides a comprehensive suite of scores and reports designed to offer deep insights into a student's reading abilities and progress. Amira ISIP reports a primary norm-referenced and criterion-referenced score as well as various subscores.

5.1 Norm-Referenced vs. Criterion-Referenced Scores: ARM and MAST

Amira ISIP Assess produces two distinct categories of composite scores that serve complementary but methodologically different purposes:

The ARM (Amira Reading Metric) is a norm-referenced composite score. It situates each student's reading performance relative to a nationally representative sample of same-grade peers. ARM scores are expressed as percentile ranks (1–99) and as scale scores on a common vertical scale spanning Kindergarten through Grade 6. ARM is the primary score used for universal screening risk classification and benchmark comparisons. Because it is norm-referenced, a student's ARM score reflects relative standing — not whether specific skills or standards have been mastered.

The MAST Score (Mastery of Academic Standards & Targets) is a criterion-referenced score that summarizes where a student is on the path toward mastering the grade-level reading standards for their state and grade. Unlike ARM, which is norm-referenced and compares a student's reading performance with that of national peers, MAST is criterion-referenced: it reflects performance against grade-level standards rather than performance relative to other students.

Amira ISIP Assess is therefore not solely norm-referenced. The tool provides both norm-referenced (ARM) and criterion-referenced (MAST) scores. Districts and states may use ARM for population-level risk identification and MAST for standards-based progress monitoring and mastery determination.

The foundational reading skill domains assessed and reflected in MAST at each grade level (K–3) are:

- **Kindergarten:** Phonological Awareness (Blending, Segmentation), Alphabetic Knowledge (Letter Names, Letter Sounds), Phonics/Decoding (Pseudoword, Word Identification)

- **Grade 1:** Phonological Awareness, Alphabetic Knowledge, Phonics/Decoding, Oral Reading Fluency, Vocabulary, Reading Comprehension

- **Grade 2:** Phonics/Decoding, Oral Reading Fluency, Vocabulary, Reading Comprehension, Spelling

- **Grade 3:** Oral Reading Fluency, Reading Comprehension, Vocabulary, Spelling

Domain coverage at Grades 4–6 follows the same framework with increasing emphasis on fluency, comprehension, and vocabulary constructs. Standards alignment is state-configurable.

5.2 ARM

The Amira ISIP Assessments produce a single primary composite score, the ARM score. ARM scores use a universal scale that assigns a score to students at all levels of reading ability, ranging from those who cannot yet read connected text to those who can read it fluently and understand what they have read. It is reported on a Grade Level Equivalent (GLE) scale with clear benchmarks.

If a student's ARM score is 1.1 and they are a third grader in month 1 of the school year, then that student is two full grades behind. If you have one student who scores 6.34 and another who scores 6.14, the first is two months further along in mastery than the second.

The ARM score combines a student's performance across multiple tasks into a single measure. It is derived from two main components: (1) a theta, or ability estimate, generated through IRT models using all tasks except ORF, and (2) WCPM from the ORF task. This score is continuously updated with data collected by Amira ISIP across the school year—including screening, benchmark, and practice sessions—so it reflects both the student's current proficiency and their projected reading ability.

It is a norm-referenced score that compares a student's performance to a nationally representative reference group. The contribution of each assessment task to the composite ARM percentile rank (PR) is empirically determined through predictive analyses of end-of-year reading outcomes. In the early grades, weights are distributed more evenly across tasks to capture a broad range of foundational skills. As students advance and become more proficient at reading connected text, ORF is assigned progressively greater weight, reflecting its stronger predictive value for later reading achievement. Amira ISIP's aim with the ARM score is to provide a measurable continuum for all readers, including those not yet reading connected texts.

The ARM score enables you to compare scores for every single student (pre-readers included), providing a basis for:

- Comparing reading ability across students within a grade

- Measuring an individual student's growth
- Placing every child in a class into instructional groups

Unlike oral reading fluency (WCPM) scores, an ARM score will be produced for every student who completes the screening process, even if they cannot yet read connected text and are still building foundational skills.

5.3 MAST

The Amira Mastery of Academic Standards & Targets (MAST) Score is a criterion-referenced score that summarizes where a student stands on the path toward mastering the grade-level reading standards for their state and grade. Rather than comparing students to their peers — as norm-referenced measures like ARM do — MAST reflects a student's performance against fixed grade-level criteria.

MAST is calculated from a student's work in Assessment, Tutor, and Instruct. Amira combines results from these sources into a single overall score; however, only work connected to grade-level reading standards factors into the calculation, so simply completing an activity does not automatically change a student's score. Some standards carry more evidentiary weight than others, and MAST accounts for this when calculating the overall score — it does not simply average the standards-level subscores.

Score Format
MAST is reported on a within-grade, equal-interval scale, not a 0–100% scale, and it is not a percentage of standards mastered. For Grades 1–5, a MAST score consists of a grade digit followed by a three-digit score ranging from 000–999 (e.g., 3380 = Grade 3, score of 380). Kindergarten scores omit the grade prefix and are reported as a three-digit score only. MAST is currently available for Kindergarten through Grade 5; scores are not available for Pre-K or Grades 6–8.

Because the scale uses equal intervals, a given point difference (e.g., 100 points) represents the same amount of change anywhere on the scale for that grade. Scores should only be compared within the same grade — a Grade 3 score and a Grade 4 score reflect different sets of standards and should not be treated as points on a single vertical scale.

Status Labels and Cutlines
MAST scores are classified into one of three status labels:

- **Emerging (Red)** — the earliest mastery range
- **Approaching (Yellow)** — the middle range
- **Proficient (Green)** — the highest range

Cutlines defining each range are set per grade and are not evenly spaced across the 000–999 scale (for example, the Proficient cutline sits closer to 500 than to an evenly spaced midpoint). The same cutline pair applies at Beginning, Middle, and End of Year — the boundaries do not move during the year, though a student's score and status may change as new evidence accumulates. These cutlines are currently provisional and subject to change.

Relationship to Standards-Level Results

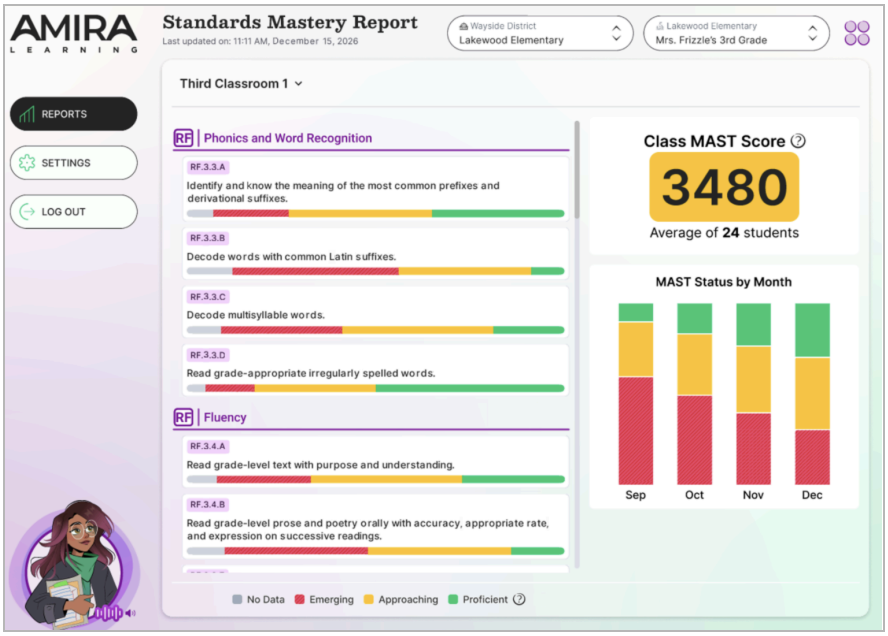
The overall MAST score and its status label provide the big-picture view of a student's progress within their grade. Each standards-level result, by contrast, describes the student's current status on one specific standard — the same label (e.g., Approaching) may appear at both the overall and standards level, but at the standards level it applies only to that individual result. The overall score is calculated from the student's underlying performance data; it is not derived by counting, adding, or averaging the standards-level labels.

When a Score Is Limited or Unavailable

A MAST score is only displayed once a student's available evidence is sufficient to support it. If a student does not yet have enough qualifying evidence, the report shows a limited-evidence indicator rather than a score. This is not a zero and does not indicate low mastery — it simply means Amira does not yet have enough eligible evidence for that grade's standards. A score should become available as the student completes more qualifying work in Assessment, Tutor, or Instruct.

Recommended Use

MAST is intended to be used in three steps: (1) start with the overall score and status label for a general sense of the student's progress; (2) review standards-level labels to identify class and student strengths and instructional needs; and (3) select an individual standard and review students grouped by status to plan targeted or small-group instruction. The overall score alone does not identify which standards need attention — that detail lives in the standards-level results.



5.4 Estimated Mastery

Estimated Mastery (also referred to as the Estimated Mastery ARM Score) is the current best estimate of a student's reading ability today, based on all data collected to date. It is a dynamic ARM score — not a skill-level probability metric — that updates automatically with every interaction a student has within the Amira ecosystem, including both assessments and practice activities.

Format
Estimated Mastery uses the same two-part format as the standard ARM score: a Peer Grade (the number before the decimal) and a Peer Month (the number after the decimal). For example, an Estimated Mastery score of 2.4 indicates the student is performing on par with the average second grader nationally in the fourth month of the school year.

How It Updates
Unlike a standard ARM score, which reflects a single point-in-time assessment, Estimated Mastery updates continuously as the student completes assessments and practices with Amira — making it possible to monitor a student's progress between formal assessment windows without administering an additional benchmark.

Where to View It
Estimated Mastery is displayed in the Class Progress Report, in the far-right column following assessment scores. It is also available in the Progress Report, where teachers can

view it by hovering over the Progress Line or checking the Reading Mastery summary below the Progress graph.

5.5 DRI

The Dyslexia Risk Indicator (DRI) is designed to identify students who are at risk of reading difficulties, including dyslexia. Amira ISIP's screening process aligns with the International Dyslexia Association (IDA) guidelines for identifying dyslexia risk, incorporating every recommended construct at each grade level. This index is supported by research, including a RAN task, which is a highly valid indicator of dyslexia risk and a predictor of word-reading development. The report provides two indicators: a risk score on a scale from 1 to 99 and a binary classification of "low risk" or "at risk" (with "at risk" further differentiated by "stronger signals" or "weaker signals"). Students classified as "at risk" are flagged for further assessment and monitoring. Amira ISIP's accuracy in identifying high-risk students has received the highest rating from the National Center on Intensive Intervention (NCII). Additional details on the classification accuracy and sensitivity are located in Section 8.

Other Key Subscores and Metrics
Amira ISIP also reports on a granular level across the various "threads of the Reading Rope":

- **ORF/WCPM**: Measures a student's reading speed and accuracy. The ORF passage is a cornerstone of the universal screener, and Amira ISIP can adapt passages up or down in difficulty to ensure accurate data. WCPM is a key component of the ARM score.
- **Reading Accuracy**: Indicates the percentage of words read correctly out of the total words read, crucial for overall reading comprehension.
- **Comprehension**: Assesses a student's ability to understand and recall information from text. This includes both Listening Comprehension (where no text reading is required) and Reading Comprehension (following an ORF passage). The relative performance between listening and reading comprehension can indicate a risk of dyslexia versus other factors, such as being an English learner.
- **Decoding Skills/Phonics**: Evaluates a student's ability to apply phonics rules to read new or unfamiliar words. This includes tasks such as Letter-Sound Knowledge, Word Decoding (WIF), and Pseudoword (Nonsense Word) Fluency (NWF), which require students to rely on decoding rather than memory.
- **Vocabulary**: Assesses students' ability to understand and recognize the meanings of words and phrases in context.
- **RAN**: A standalone measure assessing speed and automaticity, not just identification. Amira ISIP offers three forms (numbers, colors, objects) and records total processing time and item-level accuracy, computing a RAN speed score.
- **Spelling/Encoding**: Assesses a student's ability to apply phonics and spelling conventions through a dedicated task in which students type words.
- **Lexile Level**: Provides a standardized measure of a student's reading ability and the

text complexity they can comprehend.

- **Visual Attention**: Assessed through tasks where students identify target images in a grid-like display, measuring speed, accuracy, and completeness.
- **Phonological Working Memory (Nonword Repetition):** Assessed through a task where students repeat pseudo-words vocalized in a video.
- **Phonological Awareness**: Includes subtasks such as Phoneme Blending, Phoneme Segmentation, and Phoneme Manipulation (Substitution), designed to align with early literacy standards.
- **Percentile Rankings (PRs)**: Amira ISIP generates percentile ranks for almost all metrics, enabling comparative analysis of student achievement and growth relative to nationally representative norms.

6. LINKING & EQUATING

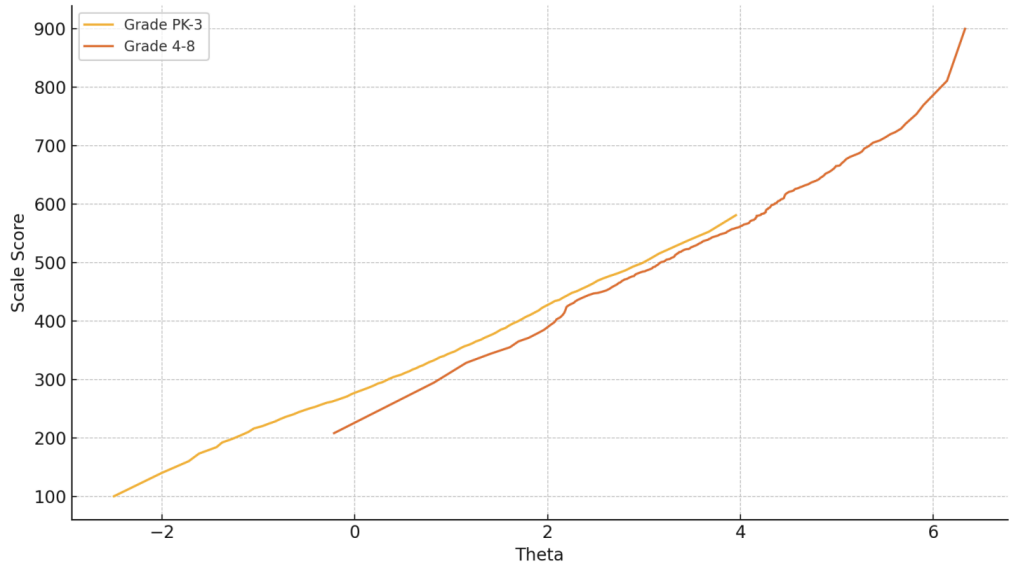
6.1 Linking to Legacy ISIP

To support longitudinal continuity and interpretability, the newly developed Amira–ISIP metric was linked to the established Legacy ISIP scale using an equipercentile linking procedure. For this initial calibration study (see section 4.5), we treated the Amira ISIP metric as a new latent scale and established a correspondence with Legacy ISIP scores through a concurrent-administration design.

Each student in the calibration sample had both an estimated Amira–ISIP ability (theta, θ) score and a corresponding Legacy ISIP scale score obtained at the same time. The Amira–ISIP theta estimates and Legacy ISIP scores were each converted to percentile ranks (0–99), enabling the construction of a raw quantile-to-quantile concordance. To ensure smoothness and mitigate the impact of sampling variability—especially in sparse regions of the score distribution—a LOESS smoothing procedure (Cleveland, 1979) was applied to both distributions. This yielded a continuous, monotonic concordance function, which serves as the operational linking transformation between Amira ISIP theta scores and Legacy ISIP scale scores.

This smoothed concordance allows any new theta value to be mapped onto the Legacy ISIP scale via interpolation, providing score continuity without altering the underlying structure of either assessment system. Note that the Legacy ISIP scores exist on two separate scales: PK through grade 3 and grades 4 through 8.

Figure 6.1 illustrates the resulting theta-to-ISIP scale score transformation curve; Amira ISIP Theta to Legacy ISIP Scale Score Linking Relationship



To evaluate the coherence of the linked scales, we examined both global and grade-specific agreement between the theta-based linked scores and their original Legacy ISIP counterparts. Key indicators included Pearson correlations, root-mean-square deviation (RMSD), mean bias, and absolute bias.

Across all grades, the overall correlation was strong (r = 0.887), and the average deviation remained modest (RMSD = 0.755 logits; |bias| = 0.560). However, as expected in vertical scale linking, performance varied by grade. In early grades (K–3), correlations were generally moderate to strong, and error metrics remained within acceptable bounds. In contrast, coherence diminished in Grades 6–8, likely reflecting reduced calibration sample sizes and greater variability in ability levels at the upper grade range.

Table 6.1 Strength of Linking Relationship

Grade	Correlation (r)	RMSD (logits)	Mean Bias	Absolute Bias
Overall	0.887	0.755	+0.041	0.560
Kindergarten	0.603	0.679	+0.103	0.512
Grade 1	0.730	0.589	−0.005	0.440
Grade 2	0.780	0.633	−0.030	0.464
Grade 3	0.764	0.673	−0.101	0.503
Grade 4	0.651	0.884	−0.252	0.695
Grade 5	0.608	0.967	−0.211	0.779
Grade 6	0.436	0.917	+0.376	0.678
Grade 7	0.349	1.052	+0.411	0.804
Grade 8	0.339	1.072	+0.527	0.863

6.2 WCPM Equating

Amira ISIP's ORF assessment is conducted at regular intervals throughout the school year. While we make every effort to maintain consistent text complexity among passages for each screening at a specific grade level, there will inevitably be some variability in passage difficulty.

Thus, to obtain precise measurements of student fluency growth, it is imperative to equate

these passages. Equating ensures that scores from all passages are placed on the same scale for accurate comparison.

Similar to ISIP linking, Amira ISIP employs an equipercentile linking method for WCPM equating. The Fall ORF passages were equated to the Winter ORF scale using students who took both forms. Likewise, using a matched sample, the Spring ORF passages were equated to the Winter ORF scale. Percentile ranks were calculated for each raw WCPM score on both forms, and the equating transformation maps scores with identical percentile ranks to the same scale score.

This method ensures that adjustments to scores are made across the entire percentile rank distribution, not just based on averages. This creates a smooth, monotonic transformation that preserves examinees' relative standing within their respective testing populations while accounting for differences in form difficulty and score distributions. Adjustments can vary across score ranges, reflecting differences in item difficulty for students at different proficiency levels. For instance, two passages may be comparable in difficulty for students reading at 200 WCPM, but one may pose more challenges for those reading at 50 WCPM. In such cases, greater adjustments are made for scores around 50 WCPM. The overarching objective of these adjustments is to provide the most precise estimate of fluency, accounting for variations across different passages.

Lastly, LOESS smoothing was applied to create a smooth, monotonic transformation function between forms. The resulting transformation function maintains the essential characteristics of the equipercentile relationship while eliminating artifacts that could lead to counterintuitive score conversions.

7. DEVELOPMENT OF NATIONAL NORMS

National norms facilitate the evaluation of student performance relative to other students nationwide. Establishing national norms for student performance is crucial for ensuring educational equity and consistency. Amira ISIP's norms serve as a benchmark against which individual student performance can be measured, enabling educators to identify areas where students are excelling or may need additional support.

National norms for the Amira ISIP Benchmark Assessment are presented as percentile ranks and were determined by grade and testing window. The three testing windows are defined as follows:

- Fall - August 1st through November 30th
- Winter - December 1st through March 31st
- Spring - April 1st through July 31st

The Amira-ISIP national norms are based on a nationally representative sample of over 800,000 student assessments from over 1,000 districts across all four regions of the United States. The sample was collected at three assessment time points (fall, winter, spring) during the 2024-2025 academic year. Demographic information for the final sample was aggregated at the district level from the National Center on Education Statistics, the US Census Bureau, ED Data Express, and the US Department of Education.

To ensure the continued relevance of the norms, a 2025 norming validation study was conducted using a fresh sample of student data. The results confirmed the stability and generalizability of the original 2024 norms, supporting their continued use with only minor adjustments and no significant changes to the national sample structure or demographic representation.

To achieve a nationally representative sample, data were stratified by district-level demographic composition and a socioeconomic index. Normalized weights associated with each strata were calculated using DeBell and Krosnick's (2009) approach to trim extreme weights. Percentile ranks were then calculated for the weighted sample at each grade and testing window. This post-stratification process helps reduce bias in sampling inherent to the Amira user population.

7.1 Stratification

Stratification involves dividing the population into distinct subgroups based on important characteristics. By using stratification, researchers ensure that each subgroup is represented in the sample in proportion to its share of the broader population. This approach reduces sampling error and allows for more precise and reliable estimates for each stratum (Rivers & Lieberman, 2021).

The stratification process for the Amira sample was designed to ensure a comprehensive and representative distribution at the district level. The districts were first stratified into four geographic regions: Northeast, Midwest, South, and West. The section level of stratification was based on community type: urban, suburban, and rural, to capture a broad range of educational contexts. Within these districts, the sample was further stratified to reflect varying socioeconomic backgrounds, ensuring that the distribution accurately reflects students from low-, middle-, and high-income families. Additionally, racial and ethnic diversity was a key consideration, with deliberate efforts to include students from multiple racial/ethnic groups (NCES, 2022). This stratification strategy enables the Amira sample to be nationally representative, providing robust and equitable insights into student performance across different district types, socioeconomic statuses, and demographic backgrounds (Hill & Roberts, 2017).

7.2 Weighting

Weighting is applied to adjust the influence of specific sample responses so that the sample more accurately reflects the population. If certain groups are overrepresented or underrepresented in the sample, weighting corrects these discrepancies.

The weighting process for the Amira sample was carefully designed to ensure that the final data accurately reflect the broader population, even if some groups were over- or underrepresented in the initial sampling. After stratifying the sample by district type (urban, suburban, rural), socioeconomic background, and racial/ethnic groups, weights were calculated for each participant based on their subgroup's representation in the national population using the following steps:

1. **Assign Initial Weights Based on Sampling Probability:** Each subgroup within the sample was assigned an initial weight corresponding to the inverse of its probability of selection. This ensures that participants from smaller or less-represented groups contribute proportionally more to the final dataset than those from larger, overrepresented groups.
2. **Adjust for Socioeconomic and Demographic Representation:** To address any imbalances across socioeconomic and racial/ethnic groups, additional weighting adjustments were made. For instance, if a specific racial group or socioeconomic tier was oversampled, the weight for participants within that group was reduced accordingly. Conversely, weights were increased for underrepresented groups, so their responses carried appropriate influence in the final results.
3. **Apply Post-Stratification Adjustments:** After initial weighting, post-stratification adjustments were conducted to ensure that the weighted sample aligns as closely as possible with national demographic benchmarks. This step ensures that the weighted data accurately mirrors the national distribution for key demographic and socioeconomic factors.

4. **Validation and Calibration:** The final weights were validated by comparing the weighted sample distributions to national data. If necessary, minor calibration adjustments were made to fine-tune the representation, ensuring that the Amira sample accurately reflects the composition of students across the United States (Thum & Hauser, 2015).

Through this detailed weighting process, Amira's final dataset provides a balanced representation of the diverse student population, enabling meaningful and generalizable insights for educators and stakeholders (AERA, APA, NCME, 2014).

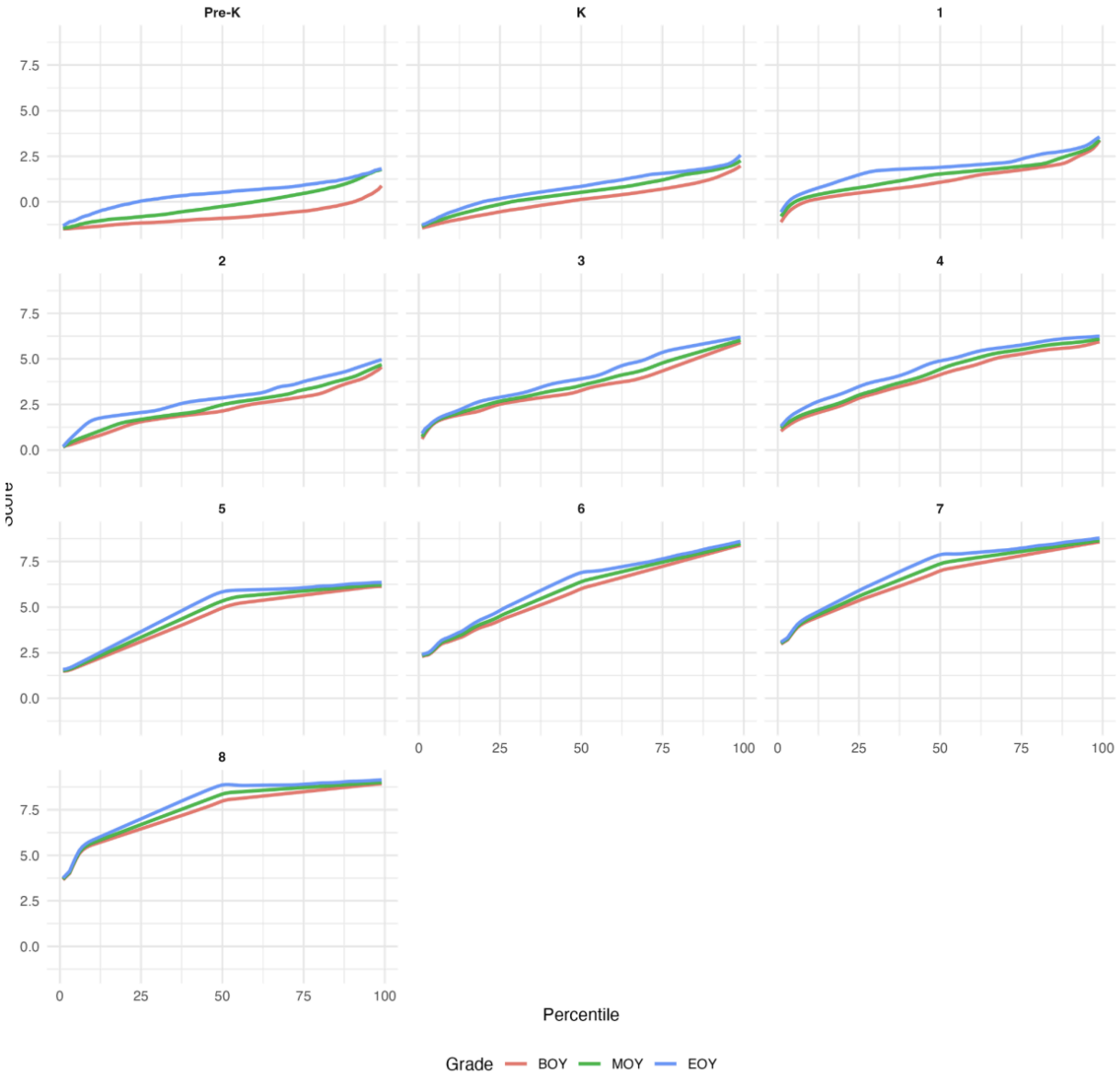
Table 7.1 Counts of Students, Districts, and States

Grade	Window	Number of Students	Number of Districts	Number of States
Pre-K	BOY	72658	2452	50
Pre-K	MOY	85004	2694	50
Pre-K	EOY	17556	801	40
K	BOY	557350	10440	50
K	MOY	579335	10195	50
K	EOY	210002	3494	50
1	BOY	556737	10315	50
1	MOY	565958	9442	50
1	EOY	230084	3705	50
2	BOY	595658	10595	50
2	MOY	599606	9547	50
2	EOY	240725	3712	50
3	BOY	593030	9976	50
3	MOY	584892	8846	50
3	EOY	221554	3487	50
4	BOY	481787	8802	50
4	MOY	458322	7690	50

Grade	Window	Number of Students	Number of Districts	Number of States
4	EOY	169138	2865	50
5	BOY	464683	8193	50
5	MOY	427299	7117	50
5	EOY	148431	2458	50
6	BOY	102254	2478	50
6	MOY	101148	2347	50
6	EOY	22327	700	41
7	BOY	71914	1452	50
7	MOY	56080	1150	50
7	EOY	17217	415	34
8	BOY	56166	1153	50
8	MOY	46833	952	50
8	EOY	10965	292	27

The normed percentiles associated with each reported score indicate that the student performed better than that percentage of students for their grade and testing window. A student who achieved a percentile rank of 70 on the Amira ISIP Benchmark in the spring testing window scored better than 70 percent of the nationally representative group of students who took the Amira ISIP Benchmark in the spring. These norms can be used to evaluate student progress throughout the school year, flag potential reading delays or the need for intervention, and identify areas of excellence. National norms are reported for the overall ARM score, domain-level subscores, and SODAs. See the progression of ARM score norms by grade and window in the figure below.

Figure 7.1 National ARM Score Norms by Grade and Window



8. CLASSIFICATION ACCURACY

Universal screening is paramount for identifying students at risk for academic difficulty in a Response to Intervention (RTI) model, the core of which is to provide students with multi-tiered support based on the level of academic risk they face. One primary component in RTI is *assessment*. A universal screening assessment in a particular content domain is typically administered multiple times a year. If a student scores below an established benchmark at a given time point, they are considered at risk for learning difficulties in that content domain and in need of intervention. For an assessment to be an effective universal screener, it is important to establish benchmarks through a scientifically designed and evidence-based process.

Amira Learning conducts studies to provide the most up-to-date evidence of the effectiveness of the Amira ISIP Benchmark. This research and supporting evidence follow the NCII's rating rubrics, which delineate technical standards (NCII, 2020a), and its call for submissions, which provides criteria for submitting evidence-based universal screening tools (NCII, 2020b). These NCII guidelines are not static from year to year, and the Amira ISIP Benchmark changes over time, requiring new research and supporting evidence. The research on universal screening, therefore, is regularly updated in response to these changes. The 2023-2024 Amira ISIP Benchmark norms were released in August 2023, which serve as the basis for this updated study.

This study documents the process the Amira Learning team followed to determine and validate the cut scores for Fall, Winter, and Spring to identify students in Grades K-3 with severe learning difficulties who need intensive reading intervention. The NCII ratings rubrics (NCII, 202a) guidelines were followed to establish the universal screening cut scores for the Amira ISIP Benchmark assessment. The 20th PR cut line on the NWEA MAP Growth assessment at the end-of-year summative was defined as the criterion measure across all grades in the classification accuracy analyses.

Amira ISIP has been rated with full bubbles for classification accuracy in grades K through 3, suggesting “convincing evidence” by NCIIa. Amira ISIP achieved half-bubbles, indicating “partially convincing evidence” in grades 4 through 6 and in the fall window in Kindergarten.

8.1 Student Sample

Table 8.1 presents the number of students in the study sample by grade, district, state, and the U.S. census division to which each state belongs. The sample included students in grades K–3 from Arizona, California, Maryland, Indiana, and Illinois. This distribution spans at least three of the nine U.S. Census Divisions at each grade level, aligning with NCII guidelines for a representative national sample.

Amira Reading Mastery scores from fall, winter, and spring from the academic year leading up to the criterion assessments (NWEA MAP) were extracted for both study samples and used in the classification accuracy analysis.

Table 8.1 Number of students in the study sample, by grade, state, census division, and school district.

Grade	State	U.S. Census Division	District	# Students
Kindergarten	AZ	Mountain	Leading Edge Academy	460
	IN	East North Central	MSD of Steuben County	341
	IL	East North Central	Evansville-Vanderburgh County SD	1340
	MD	Middle Atlantic	Baltimore County SD	6077
			Kindergarten Total Count	**8218**
1	AZ	Mountain	Leading Edge Academy	638
	CA	Pacific	Guadalupe Union SD	127
	IL	East North Central	Evansville	1483
	MD	Middle Atlantic	Baltimore County SD	6148
			First Grade Total Count	**8396**
2	AZ	Mountain	Amphitheater Public Schools	686
	AZ	Mountain	Leading Edge Academy	212
	CA	Pacific	Guadalupe Union SD	129

Grade	State	U.S. Census Division	District	# Students
	IL	East North Central	Evansville	1458
	MD	Middle Atlantic	Baltimore County SD	6405
			Second Grade Total Count	**8890**
3	AZ	Mountain	Amphitheater Public Schools	686
	AZ	Mountain	Leading Edge Academy	212
	CA	Pacific	Guadalupe Union SD	134
	MD	Middle Atlantic	Baltimore County SD	6654
			Third Grade Total Count	**7686**

8.2 Candidate Amira ISIP Screener Cut Scores

Establishing the ARM cut score that constitutes severe learning needs is a key step in an RTI process. While there is no clear consensus on how to identify students at risk of severe learning needs, a recommended approach is to use national norms for the screening assessment (Crawford, 2014). Because the development of national norms typically uses larger and more representative samples, they typically provide accurate and reliable information about an individual student's relative standing compared with their peers. If a student's score is below an established cut score based on a national norm, the student may require intensive intervention.

Based on research findings from the RTI literature, this study considered the Amira ISIP Benchmark scores corresponding to the 10th, 20th, and 30th PRs (for each grade and window) from the 2022 Amira ISIP norms as the candidate cut scores in our search, using the primary sample data. If a student's ARM score is below a given candidate cut score within the associated window, they are flagged as at-risk in the classification accuracy analysis. Table 8.2 presents the candidate ARM cut scores by grade and window.

Table 8.2 Candidate ARM Cut Scores by PR

Grade	Window	10th PR	20th PR	30th PR
Kindergarten	Fall	-0.3	-0.2	-0.1
Kindergarten	Winter	-0.07	0.07	0.21
Kindergarten	Spring	0.15	0.31	0.47
1st Grade	Fall	0.3	0.5	0.7
1st Grade	Winter	0.46	0.66	0.94
1st Grade	Spring	0.67	1.07	1.43
2nd Grade	Fall	0.62	1.2	1.71
2nd Grade	Winter	0.83	1.4	1.97
2nd Grade	Spring	1.15	1.76	2.19
3rd Grade	Fall	1.18	1.99	2.5
3rd Grade	Winter	1.5	2.48	2.92
3rd Grade	Spring	1.93	2.89	3.19
4th Grade	Fall	1.74	2.87	3.28
4th Grade	Winter	2.29	3.12	3.58
4th Grade	Spring	2.69	3.33	3.82
5th Grade	Fall	3.36	4.06	4.41
5th Grade	Winter	3.76	4.32	4.71
5th Grade	Spring	4.15	4.56	4.97
6th Grade	Fall	4.36	5.06	5.41
6th Grade	Winter	4.76	5.32	5.71
6th Grade	Spring	5.15	5.56	5.97

Table 8.3 Example of 2 x 2 Classification Table

	True At-Risk Status		
Predicted At-Risk Status		Students Actually At-Risk	Students Actually Not-At-Risk
	Students Classified as At-Risk	*True Positive (TP)*	*False Positive (FP)*
	Students Classified as Not-At-Risk	*False Negative (FN)*	*True Negative (TN)*

8.3 Methodology

The degree to which the Amira ISIP Screener can accurately identify students who need intensive intervention was evaluated using classification accuracy statistics based on the Amira ISIP cut scores that show the proportion of students correctly classified by their ARM scores as at-risk or not-at-risk, and the criterion measure cut scores that show whether students actually need intensive intervention. The classification accuracy analysis was conducted as follows:

1. Compare an individual student's (a) ARM score and the candidate ARM cut score, and (b) their score on the criterion measure and the criterion measure cut score. Assign 1 in one of the four designations demonstrated in the two-by-two classification table in Table 8.3.
2. Aggregate the designations to obtain the total counts in each cell for students in the sample.
3. Compute the statistics in Table 8.4.
4. These steps were repeated for the candidate ARM cut scores at the 20th, 25th, and 30th percentiles for each grade/window. The highest scoring cut scores as judged by the lower bound of the Area Under the Curve (AUC) was then selected for each grade level.

Table 8.4 Description of Classification Accuracy Summary Statistic

Statistic	Description	Interpretation
Overall Classification Accuracy Rate	(TP + TN) / (total sample size)	Proportion of the study sample whose classification by the ARM cut scores was consistent with classification by the

Statistic	Description	Interpretation
		criterion measure.
False Negative (FN) Rate	FN / (FN + TP)	Proportion of not-at-risk students identified by ARM scores in those observed as at-risk students on the criterion measure.
False Positive (FP) Rate	FP / (FP + TN)	Proportion of at-risk students identified by ARM scores in those observed as not at-risk students on the criterion measure.
Sensitivity	TP / (TP + FN)	Proportion of at-risk students identified by ARM scores in those observed as such on the criterion measure.
Specificity	TN / (TN + FP)	Proportion of not-at-risk students identified by ARM scores in those observed as such on the criterion measure
Area Under the Curve (AUC), including the lower and upper bounds of the 95% confidence interval	Area under the receiver operating characteristics (ROC) curve	How well ARM scores separate the study sample into at-risk and not-at-risk categories that match those from the criterion measure cut scores.

8.4 Results

After conducting the classification accuracy analyses for each grade and window, the results were evaluated against the NCII rating rubrics (NCII, 2020a). The conclusion was that, based on national norms, the candidate ARM cut scores corresponding to the 30th PR performed best at identifying students in need of intensive intervention. This conclusion is based on the assumption that students scoring below each criterion measure's recommended cut score are indeed students requiring intensive intervention. Thus, the

candidate cut scores corresponding to the 30th PR are recommended as the ARM universal screening cut scores to identify students at severe risk and in need of intensive intervention.

The recommended ARM universal screening cut scores yield sensitivity, specificity, and the lower bound of the area under the Receiver Operating Characteristic (ROC) curve (AUC-LB) of at least 0.7 in Kindergarten Fall, meeting the half-bubble criteria for NCII evidence (NCII, 2020a).

For Kindergarten Winter and Spring, and all windows of First, Second, and Third grades, the results showed sensitivity ≥ 0.8, specificity ≥ 0.8, and AUC-LB ≥ 0.8, satisfying the full bubble criteria for NCII evidence (NCII, 2020a).

The classification accuracy results for the recommended Amira ISIP ARM cut score against each criterion measure, for each grade and window, are provided in the following Tables 8.5 (Kindergarten), 8.6 (Grade 1), 8.7 (Grade 2), and 8.8 (Grade 3).

Table 8.5 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Kindergarten

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB	AUC-UB
Fall	-0.1	NWEA MAP	20th PR	0.8	0.77	0.8	0.77	0.8
Winter	0.21	NWEA MAP	20th PR	0.81	0.84	0.81	0.81	0.83
Spring	0.47	NWEA MAP	20th PR	0.82	0.84	0.82	0.82	0.84

Table 8.6 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Grade 1

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB	AUC-UB
Fall	0.7	NWEA MAP	20th PR	0.86	0.82	0.87	0.83	0.85
Winter	0.94	NWEA MAP	20th PR	0.88	0.84	0.88	0.85	0.88
Spring	1.43	NWEA MAP	20th PR	0.88	0.81	0.89	0.84	0.86

Table 8.7 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Grade 2

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB	AUC-UB
Fall	1.71	NWEA MAP	20th PR	0.87	0.804	0.88	0.84	0.86
Winter	1.97	NWEA MAP	20th PR	0.86	0.83	0.9	0.84	0.86
Spring	2.19	NWEA MAP	20th PR	0.87	0.84	0.91	0.84	0.87

Table 8.8 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Grade 3

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB	AUC-UB
Fall	2.5	NWEA MAP	20th PR	0.86	0.84	0.88	0.85	0.87
Winter	2.92	NWEA MAP	20th PR	0.88	0.83	0.9	0.85	0.88
Spring	3.19	NWEA MAP	20th PR	0.9	0.84	0.91	0.87	0.89

For Grades 4, 5, and 6, using Hasbrouck & Tindal 2017 WCPM norms as the criterion measure yields an AUC lower bound of at least 0.7 across all windows in Grades 4 and 6, satisfying the half-bubble criteria. For Grade 5, the benchmarks result in sensitivity ≥0.8, specificity ≥0.8, and an AUC lower bound of at least 0.8, satisfying the full bubble criteria for NCII evidence (NCII, 2020a). The cross-validation study results were consistent with those from the primary sample, providing evidence that the recommended universal screening cut scores are valid. The specific results are in Tables 8.9 (Grade 4), 8.10 (Grade 5), and 8.11 (Grade 6).

Analyses completed for grades 4 and 5 using the NWEA Map, based on 2023 data, produced results similar to those reported by Hasbrouck & Tindal (see Tables 3.10 and 3.11). In grade 4, the ROC curve AUC was 0.75 in the fall, 0.78 in the winter, and 0.77 in the spring. Sensitivity was highest in winter, and specificity was highest in spring. For Grade 5, the AUC was 0.74 in the fall, 0.77 in the winter, and 0.75 in the spring. As in grade 4, sensitivity was highest in winter, and specificity in spring.

Table 8.9 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Grade 4

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB
Fall	3.58	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.8	0.78	0.8	0.71
Winter	4.02	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.8	0.79	0.81	0.72
Spring	4.24	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.76	0.79	0.75	0.71
Fall	3.28	NWEA MAP	20th PR	0.74	0.71	0.8	0.75
Winter	3.58	NWEA MAP	20th PR	0.79	0.81	0.74	0.78
Spring	3.82	NWEA MAP	20th PR	0.78	0.71	0.84	0.77

Table 8.10 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Grade 5

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB
Fall	4.73	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.79	0.8	0.79	0.73
Winter	5	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.82	0.84	0.82	0.7

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB
Spring	5.5	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.85	0.86	0.83	0.71
Fall	4.41	NWEA MAP	20th PR	0.74	0.74	0.73	0.74
Winter	4.71	NWEA MAP	20th PR	0.78	0.81	0.72	0.77
Spring	4.97	NWEA MAP	20th PR	0.76	0.71	0.81	0.75

Table 8.11 Classification Accuracy Results Based on the Recommended ARM Universal Screening Cut Scores for Grade 6

Window	ARM Cut (30th PR)	Criterion Measure	Criterion Cut Score	Classification Accuracy	Sensitivity	Specificity	AUC-LB
Fall	5.73	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.76	0.78	0.75	0.7
Winter	6	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.75	0.85	0.72	0.71
Spring	6.5	Hasbrouck & Tindal 2017 WCPM norms	10th PR	0.75	0.81	0.73	0.71

8.5 Classification Accuracy Study by Subgroup

Amira's screener demonstrates clear and compelling evidence of classification accuracy, reliably identifying students at risk for reading difficulties across subgroups representative of the student population. This includes detailed assessments by language background, English language proficiency levels, exceptionality status, and prior education experience.

The efficacy of Amira's classification is supported by rigorous analysis following the methodology recommended by the National Center on Intensive Intervention (NCII). Sensitivity and specificity rates for Amira's assessments are consistently high, indicating strong alignment with intended outcomes. Sensitivity measures the proportion of at-risk students correctly identified by Amira, while specificity measures the proportion of not-at-risk students accurately classified. These metrics, along with the Area Under the ROC Curve (AUC), confirm that Amira effectively differentiates between students who are and are not at risk for reading difficulties, ensuring reliable and valid screening across all relevant subgroups and assessment time points.

The efficacy of the Amira screener in precisely identifying students in need of intensive intervention has been rigorously assessed using methodology recommended by the National Center on Intensive Intervention (NCII). This approach quantifies Amira's ability to accurately differentiate between students at risk for reading difficulties, including dyslexia, and those who are not. The proportion of students correctly classified based on their Amira scores was meticulously compared against outcomes derived from external criterion measures, namely NWEA MAP Reading. Four statistics were evaluated for this analysis:

1. Classification accuracy: measures the proportion of students whose classification according to their Amira scores was consistent with classification by the criterion measure

2. Sensitivity: evaluates the proportion of at-risk students identified by Amira who are also recognized as at-risk on the criterion measure, while specificity assesses the proportion of not-at-risk students identified by Amira who are likewise classified as not-at-risk on the criterion measure

3. Specificity: measures the proportion of true negatives—students who are not at risk—who are accurately identified by the Amira assessment. A high specificity rate indicates that Amira effectively avoids misclassifying students who are not at risk, minimizing false positives and ensuring that students who do not need intervention are not incorrectly flagged.

4. Area Under the ROC Curve (AUC): quantifies how well Amira separates the at-risk and not-at-risk categories that match those from the criterion measure. ROC is a probability curve, and AUC represents the degree of separability and indicates how well Amira can distinguish between classes.

Refer to the table below for the classification accuracy results in the spring window, using NWEA MAP as the criterion measure.

Table 8.12 Classification Accuracy by Subgroup

Category	Subgroup	N	Classification Accuracy	Sensitivity	Specificity	AUC
Gender	Male	8,955	0.85	0.82	0.86	0.84
	Female	8,906	0.85	0.83	0.85	0.84
English Learner Status	EL	2,355	0.81	0.87	0.80	0.81
	Non-EL	15,507	0.85	0.80	0.86	0.83
Exceptionality Status	With Disabilities	2,016	0.81	0.87	0.75	0.81
	Without Disabilities	15,846	0.85	0.81	0.86	0.84
Major Racial/Ethnic Categories	White	5,718	0.87	0.87	0.87	0.87
	Hispanic	3,196	0.81	0.86	0.80	0.83
	Asian	1,204	0.88	0.79	0.87	0.79
	Black	6,513	0.84	0.80	0.85	0.82
	Other	5,631	0.83	0.83	0.84	0.83
Language Background	English	3,523	0.82	0.77	0.82	0.80
	Spanish	1,011	0.74	0.84	0.71	0.77
Prior Education (Pre-K)	Yes	1,009	0.85	0.72	0.86	0.79
	No	3,525	0.79	0.80	0.78	0.79

8.6 Classification Accuracy Study of Amira ISIP Subscores

Universal screening is paramount in identifying students at risk for academic difficulty in an RTI model, the core of which is to provide students with multi-tiered support based on the level of academic risk they face. One primary component in RTI is assessment. A universal screening assessment in a particular content domain is typically administered multiple times a year. If a student scores below an established benchmark at a given time point, they are considered at risk for learning difficulties in that content domain and in need of intervention. For an assessment to be an effective universal screener, it is important to establish benchmarks through a scientifically designed and evidence-based process.

Amira Learning conducts studies to provide the most up-to-date evidence of the effectiveness of the Amira ISIP Benchmark. This research and supporting evidence follow the NCII's rating rubrics, which delineate technical standards (NCII, 2020a), and its call for submissions, which provides criteria for submitting evidence-based universal screening tools (NCII, 2020b).

This study documents the process the Amira Learning team followed to validate the cut scores for Fall, Winter, and Spring to identify students in Kindergarten and First Grade with severe learning difficulties who need intensive reading intervention. To establish the universal screening cut scores for the Amira ISIP Benchmark assessment, the NCII rating rubrics (NCII, 2020a) were followed, using a sample of students from Texas, South Carolina, Kentucky, and Oklahoma. For each grade, students' corresponding subscores from the NWEA MAP Reading assessments were used as the criterion measures in the classification-accuracy analyses.

The classification accuracy analyses involved testing different candidate cut scores for the ARM score at each grade and window, and using a static cut score (the cut recommended by the criterion assessment) for each criterion measure, to identify the optimal benchmarks for identifying students in need of intensive intervention. Students who score below those benchmarks are likely at risk for severe learning difficulty and in need of intensive intervention.

Table 8.12 Classification Accuracy Subscore Sample

Grade	State	District	N
Kindergarten	Texas	Lancaster ISD	129
	Texas	Vernon ISD	53

Grade	State	District	N
	South Carolina	Lancaster Co SD	57
	South Carolina	York School District 1	16
	Oklahoma	Tulsa ISD	52
Kindergarten Total Count			**307**
1	Texas	Klein ISD	126
	Texas	Lancaster ISD	242
	Texas	Tuloso Midway ISD	97
	Texas	Vernon ISD	61
	South Carolina	Lancaster Co SD	153
	South Carolina	York School District 1	26
	Kentucky	Christian County PSD	24
	Oklahoma	Tulsa ISD	120
First Grade Total Count			**849**

Table 8.13 Classification Accuracy Subscore Results

Grade	Amira Subscore	Test or Criterion Measure	AUC for BOY Cut Point for Risk	AUC for MOY Cut Point for Risk	AUC for EOY Cut Point for Risk
Kindergarten	Phonological Awareness	NWEA MAP: Phonological Awareness Domain	0.72	0.76	0.76
Kindergarten	Letter-Sound Correspondence	NWEA MAP: Phonics/Word Recognition Domain	0.77	0.79	0.77
Kindergarten	Rapid Naming	NWEA MAP: Rapid Automatized Naming WCPM	0.7	0.71	0.71

First Grade	Phonological Awareness	NWEA MAP: Phonological Awareness Domain	0.79	0.8	0.84
First Grade	Letter-Sound Correspondence	NWEA MAP: Phonics/Word Recognition Domain	0.85	0.87	0.94
First Grade	Rapid Naming	NWEA MAP: Rapid Automatized Naming WCPM	0.8	0.8	0.79
First Grade	Word or Pseudo Word Reading Fluency	NWEA MAP: Phonics/Word Recognition Domain	0.8	0.83	0.88
First Grade	Oral Reading Fluency	NWEA MAP: Oral Reading Fluency	0.87	0.9	0.95

9. RELIABILITY & VALIDITY

9.1 Reliability of Forms: Universal Screener, Benchmark, and Progress Monitoring

Reliability describes the extent to which scores are internally consistent and relatively free from random error. For an assessment's scores to be considered valid for a particular interpretation and use, establishing that the scores are reliable is necessary. Here, we present data from reliability studies as applied to the ARM composite score and ORF WCPM score. IRT standard error of measurement (SEM) estimates are provided on theta scores. Classification accuracy analyses confirm that the assessment consistently classifies students into performance categories. Amira ISIP has been rated with full bubbles for reliability in grades K through 6, indicating "convincing evidence" from NCII and meeting high industry standards.

9.1.1 Internal Consistency Reliability

The first study examines Internal Consistency Reliability, measuring the consistency of scores across the items within a test. This is done using Cronbach's Alpha, which calculates the correlation between all pairs of items in a test. The practical significance of the reliability coefficients was evaluated as follows: poor (0-0.39), adequate (0.40-0.59), good (0.60-0.79), and excellent (0.80-1.0). These estimates of practical significance are arbitrary but conventionally used and provide a useful heuristic for interpreting reliability data.

Table 9.1 shows the reliability coefficients for the ARM composite score for the universal screener. All Cronbach's Alpha coefficients were in the excellent range.

Table 9.1 Cronbach's Alphas for the Amira ISIP Benchmark/Progress Monitoring ARM Composite Score

Grade	N	Number of Items	Cronbach's Alpha
K	14116	26	0.91
1	16609	29	0.93
2	14513	22	0.93
3	14546	29	0.93
4	14513	36	0.91
5	14544	40	0.91

Grade	N	Number of Items	Cronbach's Alpha
6	10588	40	0.91

9.1.2 Test-Retest Reliability

Test-Retest reliability was assessed by examining the correlations between scores from tests taken by the same students at different time points. We measure these correlations using Pearson's correlation coefficient, which quantifies the strength of the linear relationship between two variables.

Table 9.2 shows the Pearson correlations between the theta scores derived from IRT calibration for the Benchmark and Progress Monitoring assessments, obtained from the same student within two weeks during the Winter window of the 2022-2023 school year. All correlation coefficients are 0.7 or higher, indicating that the Amira ISIP Benchmark and Progress Monitoring Assessments are reliable across all supported grades.

Table 9.2 Test-Retest Reliability Results for the Amira ISIP Benchmark/Progress Monitoring Theta Score

Grade	N	Correlation Coefficient
K	955	0.84
1	6402	0.86
2	8460	0.87
3	7870	0.86
4	5813	0.87
5	4153	0.87
6	1505	0.87

9.1.3 Classification Consistency

Classification consistency was evaluated using a test-retest design across grades Pre-K through Grade 8 in the BOY window of the 2025-2026 school year. For each student, risk category assignments were compared across two test administrations that occurred within the Fall testing window with classifications based on percentile rank (PR) cut scores of the ARM composite score: students scoring below the 25th percentile were classified as Tier 3, those between the 25th and 74th percentiles as Tier 2, and those at or above the 75th percentile as Tier 1.

Results showed high consistency in classification, with percent agreement exceeding 91% at every grade level. Cohen's Kappa values ranged from 0.88 to 0.90, indicating almost perfect agreement based on standard interpretive thresholds. The highest agreement was observed in Grade 1 (Kappa = 0.901, 93.45% agreement), while Grade 3 showed the lowest, though still strong, consistency (Kappa = 0.88, 91.76% agreement). Across all grades, these estimates confirm the stability and reliability of the classification decisions, supporting the use of this instrument for consistently identifying students' risk status over time.

Table 9.3 Classification Consistency Results for the Amira ISIP Benchmark/Progress Monitoring ARM Composite PR

Grade	N	Percent Agreement	Cohen's Kappa
Pre-K	220	93.18	0.888
K	10650	92.85	0.889
1	18592	93.45	0.901
2	28392	93.28	0.899
3	24965	91.76	0.876
4	16615	92.89	0.893
5	15947	93.52	0.901
6	2716	92.42	0.885
7	1026	91.81	0.877
8	1055	95.92	0.935

9.1.4 Parallel Forms Reliability

Parallel forms reliability is a statistical measure used to assess the consistency of test scores between two different versions of the same test, known as parallel forms. This type of reliability is particularly useful when it is not possible or desirable to administer the same test to the same group of participants multiple times, as it allows estimation of the test's reliability without repeating the same questions.

Amira's parallel forms reliability was evaluated by administering two different versions of the ORF assessment to the same group of grade 1 and 2 students at three time periods:

1. BOY and MOY
2. MOY AND EOY
3. BOY and EOY

Table 9.3 shows the Pearson correlation coefficients of the two WCPM scores computed based on the Benchmark ORF task taken in each grade's 2022-2023 SY window. The high correlation coefficients, ranging from 0.71 to 0.86 in the table below, confirm that different forms of the ORF assessment consistently measure the same underlying construct.

Table 9.4 Parallel Forms Reliability Results for the Amira ISIP Benchmark Oral Reading Fluency (ORF) WCPM score

Grade	Window 1	Window 2	N	Correlation Coefficient
1	BOY	MOY	1596	0.73
1	MOY	EOY	1866	0.81
1	BOY	EOY	1642	0.71
2	BOY	MOY	300	0.75
2	MOY	EOY	342	0.86
2	BOY	EOY	295	0.8
3	BOY	MOY	359	0.79
3	MOY	EOY	398	0.78
3	BOY	EOY	347	0.74
4	BOY	MOY	533	0.79
4	MOY	EOY	680	0.8
4	BOY	EOY	609	0.79
5	BOY	MOY	576	0.76
5	MOY	EOY	713	0.79
5	BOY	EOY	634	0.74
6	BOY	EOY	469	0.81

9.1.5 Inter-rater Reliability

Amira ISIP's intelligent tutoring system employs a range of AI and Machine Learning (ML) technologies. The common method for validating the reliability of an ML model is to replicate the results through human classification. The AI system's outcomes are compared with the judgments of human experts. If the ML model demonstrates reliability comparable to that of a human, it can be considered a viable substitute for an expert (Kim, 2006; Chung, Jang, Yun, & Sa, 2008).

To evaluate the interrater reliability of Amira ISIP's Screener, a study was designed to determine the level of agreement between composite scores derived from the Amira ISIP RED model and those obtained from human raters, using Pearson's correlation coefficient. The RED scoring system is an automated system that uses advanced algorithms to detect errors and assess the quality of students' speech, generating scores based on the error detection model (EDM). Experienced educators were employed to independently score the students' speech.

The analyses revealed correlations of at least 0.95 or greater between the EDM scores and the average scores of the human raters across grade levels. This suggests a strong correlation, indicating that the EDM is largely consistent with human judgment. The study confirms that Amira ISIP's machine scoring is a viable tool for automated scoring and a reliable, efficient alternative to traditional scoring methods. The software has attained a level of accuracy that enables it to function as a virtually indistinguishable substitute for teacher scoring.

Table 9.5 Inter-Rater Reliability Grade N Inter-Rater Reliability

Grade	N	Inter-Rater Reliability
K	6186	0.95
1	6771	0.98
2-3	8400	0.97

The analyses revealed correlations of 0.95 or greater between the EDM scores and the average scores of human raters across grade levels. The Grade 1 correlation of 0.98 is particularly notable, representing near-perfect agreement between machine and human scoring. This consistency demonstrates that the EDM is functionally equivalent to expert human judgment.

Implications for Scoring Reliability and Equity
These inter-rater reliability findings validate several critical aspects of Amira's assessment system:

- **Machine Scoring as a Reliable Alternative:** The EDM is a viable, consistent, and efficient alternative to traditional teacher scoring. Educators can confidently rely on automated RED scores to identify students' reading proficiency without the variability inherent in subjective human judgment.
- **Elimination of Scoring Bias:** By using objective algorithms rather than subjective assessment, Amira's scoring eliminates common sources of scoring variability and bias, including:

  - **Halo effect:** Prior knowledge of students influencing scoring decisions
  - **Implicit bias:** Scoring variations based on student demographics, accents, or dialects
  - **Fatigue effects:** Inconsistency in human raters across extended scoring sessions
  - **Administrative variability:** Differences in scoring rigor across different educators

  Validation across diverse student populations (documented in Section 9.1.6) confirms that EDM scoring is equitable and consistent regardless of student background, race/ethnicity, English learner status, socioeconomic status, or special education status. This ensures that assessment scores can be used with confidence to inform educational decisions for all students.
- **Real-Time Scoring Capability:** Because the RED model achieves reliability comparable to that of human experts, Amira can provide immediate automated scoring and feedback without sacrificing accuracy. This enables educators to access results instantly upon assessment completion, facilitating rapid instructional decision-making and intervention placement without administrative burden. The study confirms that Amira ISIP's machine scoring is a reliable, equitable, and efficient alternative to traditional subjective scoring methods. The software has attained a level of accuracy that enables it to function as a virtually indistinguishable substitute for teacher scoring, with the added benefits of consistency, objectivity, and elimination of rating bias.

Generalizability of Inter-Rater Reliability Across Grades 3–8

While formal inter-rater reliability studies were conducted for Kindergarten through Grade 2, the findings from this cohort provide robust empirical support for the generalizability of RED model reliability to Grades 3 through 8. The Grade 2 results (r = 0.97) are particularly instructive, as Grade 2 represents a developmental transition point where students progress from early decoding skills to more complex reading tasks—a skill continuum that extends through Grade 8. All grade levels (K–8) employ identical technical architectures: the same Reading Error Detection algorithms, the same underlying machine learning models trained on the same large-scale, nationally representative speech datasets, and the same automated scoring infrastructure. The RED model's consistent demonstration of near-perfect inter-rater reliability across the K–2 span (r = 0.95–0.98) underscores the stability and robustness of the underlying machine learning system. This consistency across grades provides strong evidence for generalizability: because the RED model achieves exceptional agreement with expert human judgment across three consecutive grade levels with increasingly complex reading tasks, it is reasonable to infer that this reliability extends to Grades 3–8, which employ the identical technical system.

The inference that Grade 2 reliability generalizes to upper elementary and middle grades is supported by three lines of evidence: (1) technical equivalence—all grades use identical RED algorithms and ML models; (2) the pattern of K–2 consistency (0.95–0.98 across three

grades) suggests the RED model is inherently stable regardless of grade level; and (3) grade-level classification consistency data (Cohen's Kappas ranging from 0.876–0.901 across K–3; see Section 9.1.3, Table 9.3) provides independent validation that students are reliably classified into reading proficiency tiers across grades. For Grades 4–8, classification consistency metrics serve as the primary evidence of reliability, confirming that the RED model consistently and accurately places students into appropriate instructional categories. This multi-method approach to reliability—combining validated inter-rater studies for K–2 with strong empirical inference for Grades 3–8, supported by classification-consistency evidence across all grades—ensures robust and dependable measurement across the full K–8 assessment span.

Comprehensive Reliability Summary

Amira ISIP's Screener demonstrates exceptional reliability across diverse student populations and grade levels. Through comprehensive evaluations of internal consistency, test-retest reliability, alternate-form reliability, inter-rater agreement, and classification consistency, Amira ISIP has consistently demonstrated reliability estimates that meet or exceed 0.80, indicating robust and dependable performance. These high-reliability metrics—such as Cronbach's alpha values above 0.90, test-retest correlations ranging from 0.84 to 0.87, alternate form correlations between 0.71 and 0.86, inter-rater reliability as high as 0.98, and classification consistency (Kappas) ranging from 0.876 to 0.901—underscore the consistency and accuracy of Amira ISIP's assessments across all grade levels. These findings provide compelling evidence that educators can trust Amira ISIP's data to inform critical instructional and intervention decisions with confidence across the full K–8 assessment span.

9.1.6 Reliability of Subgroups

Amira ISIP's Screener demonstrates high reliability across a diverse range of student subgroups. The reliability estimates are disaggregated by key demographic factors such as grade/age, gender, English learner status, exceptionality status, major racial/ethnic categories, socio-economic status, and language backgrounds. These disaggregated reliability metrics are crucial for ensuring that the assessment tool produces consistent and accurate results across all student groups, thereby supporting equitable educational outcomes. The reliability estimates for most subgroups meet or exceed 0.90, with many exceeding 0.95, which is considered a very strong indicator of reliability, ensuring that the assessment scores can be confidently interpreted and used in high-stakes educational decisions.

The tables below detail the reliability measures for subgroups by grade/age, gender, English learner status, exceptionality status, and major racial/ethnic categories.

Table 9.6 Grade/Age

Grade	N (Number of Participants)	Internal Consistency	Inter-Rater Reliability
K	6186	0.91	0.95
1	6771	0.93	0.98
2	8400	0.92	0.97

Table 9.7 Gender

Grade	Gender	N (Number of Participants)	Internal Consistency	Inter-Rater Reliability
K	Male	3118	0.91	0.95
K	Female	3067	0.91	0.95
1	Male	4005	0.93	0.94
1	Female	3851	0.93	0.94
2	Male	4290	0.92	0.97
2	Female	4110	0.93	0.97

Table 9.8 English Learner Status English

Grade	English Learner Status	N (Number of Participants)	Internal Consistency	Inter-Rater Reliability
K	EL/MLL	1026	0.91	0.97
K	Not EL/MLL	6595	0.91	0.95
1	EL/MLL	1120	0.96	0.91
1	Not EL/MLL	7322	0.93	0.94
2	EL/MLL	1154	0.93	0.98
2	Not EL/MLL	7246	0.91	0.97

Table 9.9 Exceptionality Status Exceptionality

Grade	Exceptionality	N (Number of	Internal	Inter-Rater

	Status	Participants)	Consistency	Reliability
K	With	705	0.90	0.96
K	Without	5413	0.91	0.95
1	With	970	0.92	0.96
1	Without	7345	0.92	0.95
2	With	1080	0.92	0.98
2	Without	7306	0.91	0.97

Table 9.10 Major Racial/Ethnic Categories

Grade	Race	N (Number of Participants)	Internal Consistency	Interrater Reliability
K	White	3725	0.91	0.95
K	Hispanic	862	0.91	0.96
K	Asian	315	0.93	0.96
K	Black	1653	0.9	0.94
K	Other	506	0.91	0.96
1	White	2170	0.93	0.91
1	Hispanic	871	0.93	0.93
1	Asian	294	0.92	0.94
1	Black	1662	0.93	0.94
1	Other	850	0.93	0.96
2	White	1355	0.92	0.97
2	Hispanic	858	0.93	0.98
2	Asian	311	0.92	0.97
2	Black	1765	0.92	0.98

2	Other	255	0.91	0.98

Table 9.11 Socio-Economic Status

Grade	Socio-Economic Status	N (Number of Participants)	Reliability Coefficient (Cronbach's Alpha)	Interrater Reliability
K	Low SES	500	0.9	0.98
K	High SES	964	0.9	0.97
1	Low SES	532	0.93	0.98
1	High SES	1012	0.93	0.98
2	Low SES	575	0.92	0.98
2	High SES	949	0.9	0.97

Table 9.12 Language Backgrounds

Grade	Home Language	N (Number of Participants)	Reliability Coefficient (Cronbach's Alpha)	Interrater Reliability
K	English	1018	0.9	0.97
K	Spanish	446	0.9	0.97
1	English	1252	0.91	0.98
1	Spanish	419	0.92	0.98
2	English	1294	0.92	0.98
2	Spanish	246	0.91	0.98

Disabilities (e.g., Speech or Hearing)

Most students with accommodations are included in the calculations above, as we have insufficient sample sizes to produce separate reliability estimates for students with speech

or hearing impairments. Data collection is ongoing, and reliability metrics for these groups will be reported once a sufficient sample size is achieved.

The reliability data presented above for Amira ISIP's Screener provides compelling evidence of its robustness and consistency across various student subgroups. With most reliability estimates at 0.90 and many exceeding 0.95, well above the generally accepted threshold of 0.80, Amira ISIP ensures that its assessments are reliable tools for accurately measuring student performance across diverse populations. This commitment to reliable measurement across all student groups reinforces Amira ISIP's role as a trusted tool for educators, enabling them to make informed decisions that enhance student learning and effectively address the specific needs of each subgroup.

9.1.7 Growth Slope Accuracy

Amira ISIP also calculates a slope of estimated growth for all students as progress monitoring occurs. This slope estimates the weekly change in WCPM.

In estimating measurement error, 121,384 students had sufficient measurement points in the BOY window (at least 3) to qualify for inclusion in the study. The median projected growth over 36 weeks was 14.1 WCPM compared to the actual growth of 14.8 WCPM.

Table 9.13 Summary Statistics

Total N	167,047
Median Projected Growth	14.1 WCPM
Median Actual Growth	14.8 WCPM
Median Error	4.8 WCPM
Standard Deviation	4.6 WCPM
Measurement Error	1.30%

The resulting Reliability of Growth Slope Coefficient ranges from 0.57 in Kindergarten to 0.84 in grade 1, as shown in Table 9.6.

Table 9.14 Reliability of Growth Slope

Grade	N Size	Reliability of Slope Coefficient
Kindergarten	5,022	0.57
1st Grade	14,621	0.84

2nd Grade	40,231	0.76
3rd Grade	49,933	0.69
4th Grade	31,418	0.75
5th Grade	25,822	0.72
6th Grade	0	N/A

This classification analysis supports the accuracy of Amira ISIP's Progress Monitoring for all students and those with low, typical, and high growth, as described in Table 9.7.

Table 9.15 Accuracy of Prediction Overall and by Growth Pattern

Growth	Number Projected Correctly	Number Projected Incorrectly	Percent Accurately Predicted
All Students	86,625	16,294	84%
Typical	48,056	8,330	85%
Low	9,555	5,573	63%
High	32,996	3,982	89%

In summary, Amira ISIP's slope of growth has a Measurement Error of 1.3% and accurately classifies students' growth as high, typical, or low in just over 84% of cases.

9.2 Standard Error of Measurement

The Standard Error of Measurement (SEM) is a measure of the precision of students' ability estimates derived from the adaptive assessment component. The SEM quantifies the expected variability in a student's observed ability estimate due to measurement error. Lower SEM values indicate more precise ability estimates and are therefore preferable to higher values.
In the context of adaptive assessments, SEM values are influenced by multiple factors, including:

- Model-data fit: How well student response patterns conform to the underlying 2PL IRT model
- Student response consistency: The degree to which students respond in predictable patterns, given their ability level

- Student location on the ability continuum: Students at extreme ability levels (very high or very low) may have less precise estimates due to limited item availability in those ranges
- Item-ability match: The degree to which the adaptive algorithm successfully matches item difficulties to student ability
- Test length: The number of items administered, with more items generally providing more information and lower standard errors

High SEM values may occur when students perform erratically, have extreme response vectors (e.g., answering all items correctly or incorrectly), or when the adaptive algorithm serves items that are poorly matched to the student's ability level. Although there are no absolute benchmarks for acceptable standard errors, values are evaluated relative to theoretical expectations and are expected to remain relatively consistent across the ability distribution in a well-designed adaptive test.

9.2.1 Theoretical Minimum SEM

Given the adaptive nature of the assessment and the wide range of difficulty in the item bank, standard errors are expected to approach the theoretical minimum for a test of a given length. The theoretical minimum represents the lowest SEM achievable if the adaptive algorithm perfectly matched each item's difficulty to the student's ability estimate at every step. This minimum is constrained by test length—the more items administered, the lower the SEM can potentially be.

For a 2PL IRT model, the theoretical minimum SEM on the theta (θ) scale is given by:

$$\text{Min SEM}(\theta) = \sqrt{(1/I_max)},$$

where I_max represents maximum test information.

9.2.2 Adjustment for Prior Information

The ability estimates in this assessment are derived using EAP estimation (see Section 4.3) with grade-level prior distributions. EAP estimation incorporates two sources of information: (1) information from the administered items and (2) information from the prior distribution. The standard errors output from EAP estimation (EAP_SE) reflect the combined precision from both sources.

For technical reporting purposes, it is important to isolate the measurement precision attributable to the adaptive test itself, removing the contribution of the prior. This adjustment is accomplished by calculating test information as:

$$I_test = 1/EAP_SE^2 - 1/\sigma^2_h,$$

where σ_h is the standard deviation of the grade-level prior distribution. The test-only SEM is then:

$$SEM(\theta) = 1/\sqrt{I_test}$$

This adjusted SEM represents the precision of ability measurement derived exclusively from the adaptive item administration, providing a more accurate assessment of the test's measurement properties independent of external prior information. All SEM values reported in this section reflect this adjustment and represent test-only measurement precision.

In cases where I_test ≤ 0 (indicating the test added no information beyond the prior), the SEM is considered undefined. Such cases typically occur with extremely short tests or severe item-ability mismatches. Across all administrations analyzed, no cases with undefined SEM were observed, indicating that all adaptive assessments provided meaningful measurement information beyond the grade-level priors.

9.2.3 Results

Table 9.16 presents the mean and standard deviation of SEM(θ) across all assessment administrations during the BOY 2025-2026 school year window. Statistics are reported separately by grade level to account for differences in typical test length and item bank characteristics

Table 9.16 Standard Error of Measurement by Grade

Grade	N	Mean Items	Mean SEM(θ)	SD SEM(θ)	Theoretical Min
Pre-K	5,768	30	0.24	0.04	0.09
K	282,689	35	0.26	0.05	0.09
1	391,799	46	0.23	0.05	0.07
2	422,061	34	0.25	0.07	0.09
3	391,635	24	0.31	0.1	0.1
4	296,115	22	0.36	0.11	0.11
5	270,993	22	0.32	0.1	0.11
6	53,338	21	0.31	0.09	0.11
7	34,740	19	0.23	0.15	0.12
8	31,295	18	0.32	0.935	0.12

Note: SEM(θ) values are reported on the theta scale with prior influence removed.

Across all grades and administrations, the mean test-only SEM(θ) was 0.28 (SD = 0.10). Standard errors ranged from 0.23 (Grade 1) to 0.36 (Grade 4), reflecting variations in test length across grades. The mean test length was 27 items, with considerable variability reflecting the adaptive nature of the assessment and potential differences in routing algorithms across grades.

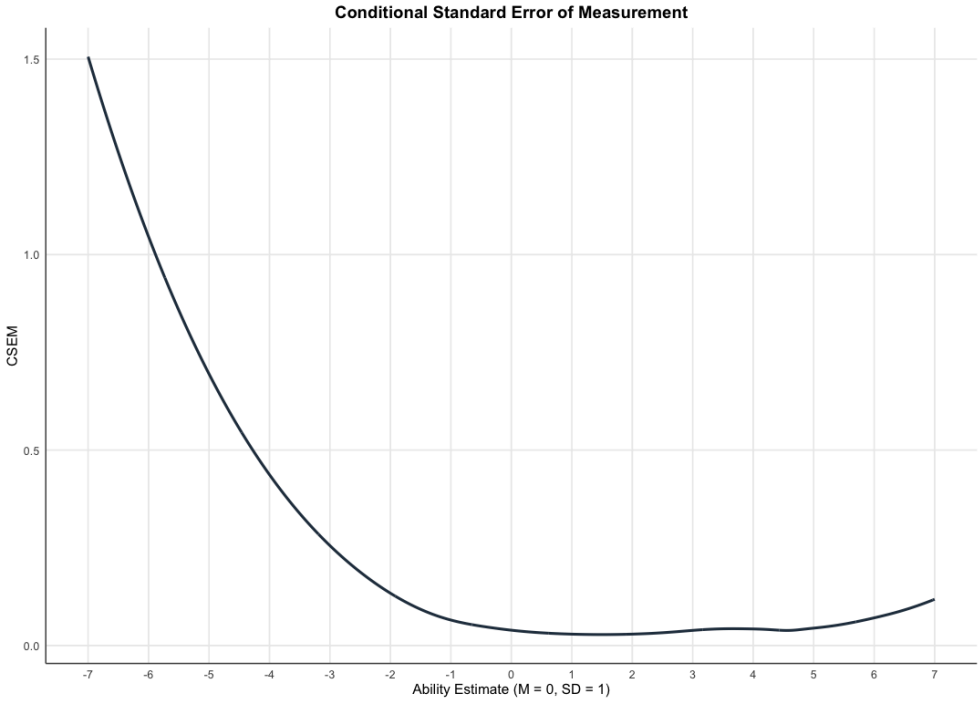
Relatively small differences were found in the test-only SEMs adjusted for EAP priors. The test-only SEMs are consistently larger than the EAP standard errors, with differences ranging from 0.007 to 0.024 logits. This pattern confirms that the grade-level priors contribute meaningful information to ability estimation. The relatively small differences suggest that the adaptive tests provide sufficient information that the prior contributes minimally to overall precision.

The SEM values reported here reflect measurement precision for the adaptive assessment component specifically and should be interpreted within the context of the broader composite scoring system. The composite scale score integrates IRT-based ability estimates with additional non-IRT measures, and overall composite score reliability is reported separately using internal consistency methods (see Section 9.1).

Conditional Standard Error of Measurement

The conditional standard error of measurement (CSEM) represents the standard error of the ability estimate conditional on the true ability level (represented by θ). Unlike classical test theory, where the standard error of measurement is assumed constant across all ability levels, the CSEM in CAT varies as a function of the test taker's true ability, reflecting the fundamental principle that measurement precision differs across the ability continuum based on how well the administered items target that particular ability level. The CSEM is mathematically defined as the reciprocal of the square root of the test information function at any given ability level, expressed as $CSEM(\theta) = 1/\sqrt{I(\theta)}$, where $I(\theta)$ represents the total information provided by all administered items at ability level, θ.

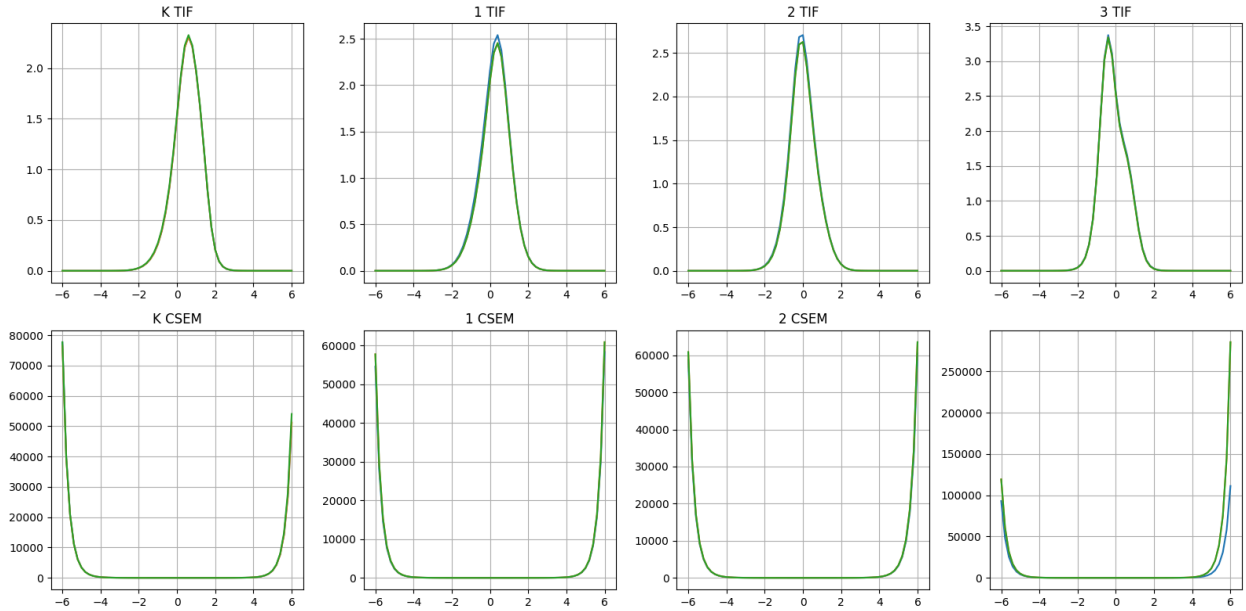
The figure below shows the observed mean CSEMs by overall θ score across the full vertical Amira scale. The results show that the measurement error is consistent and low across a wide range of scores, indicating high reliability.



Test Information Function (TIF)

The TIF is a statistical representation of measurement precision across the ability scale. It shows how effectively the assessment discriminates between students at different literacy levels. Higher information values indicate greater precision in distinguishing between students' true abilities; peaks in the TIF curve reveal where the test functions optimally. The TIF is complemented by the Conditional Standard Error of Measurement (CSEM), which quantifies the magnitude of measurement error at each ability level—lower CSEM values indicate more reliable score estimates.

In the diagram below, the Total TIF reflects a high-information environment across all three administration windows. Each testing occasion—Kindergarten (K), 1st Grade (1), 2nd Grade (2), and 3rd Grade (3)—demonstrates robust measurement precision concentrated around the modal ability range for that grade level.



Information Distribution Across the Ability Scale

The TIF curves display a consistent pattern: information peaks sharply near the center of the theta scale (approximately θ = 0 to 1), corresponding to typical developing readers within each grade, then tapers symmetrically at the extremes. This design reflects the assessment's focus on discriminating among students with typical to above-typical literacy skills. While minor differences appear at the lowest end of the scale—where early learners are just beginning their literacy journey and differentiation between students is naturally limited—the assessment maintains its diagnostic utility for the majority of test-takers.

Conditional Standard Error of Measurement (CSEM)

The CSEM panels below each TIF graph confirm this precision. Across all four testing windows (K CSEM, 1 CSEM, 2 CSEM, and 3 CSEM), the conditional error remains uniformly low across the theta scale, typically lowest in the central ability range. The CSEM curves show a characteristic U-shaped pattern, with slightly elevated error at the extreme ends of the scale (far left and far right), but this elevation is minimal relative to the low baseline error throughout the core measurement range.

This psychometric profile indicates that the assessment provides consistent, reliable differentiation of students' literacy skills across all measured grade levels, with particularly strong precision for students performing at or above grade-level expectations.

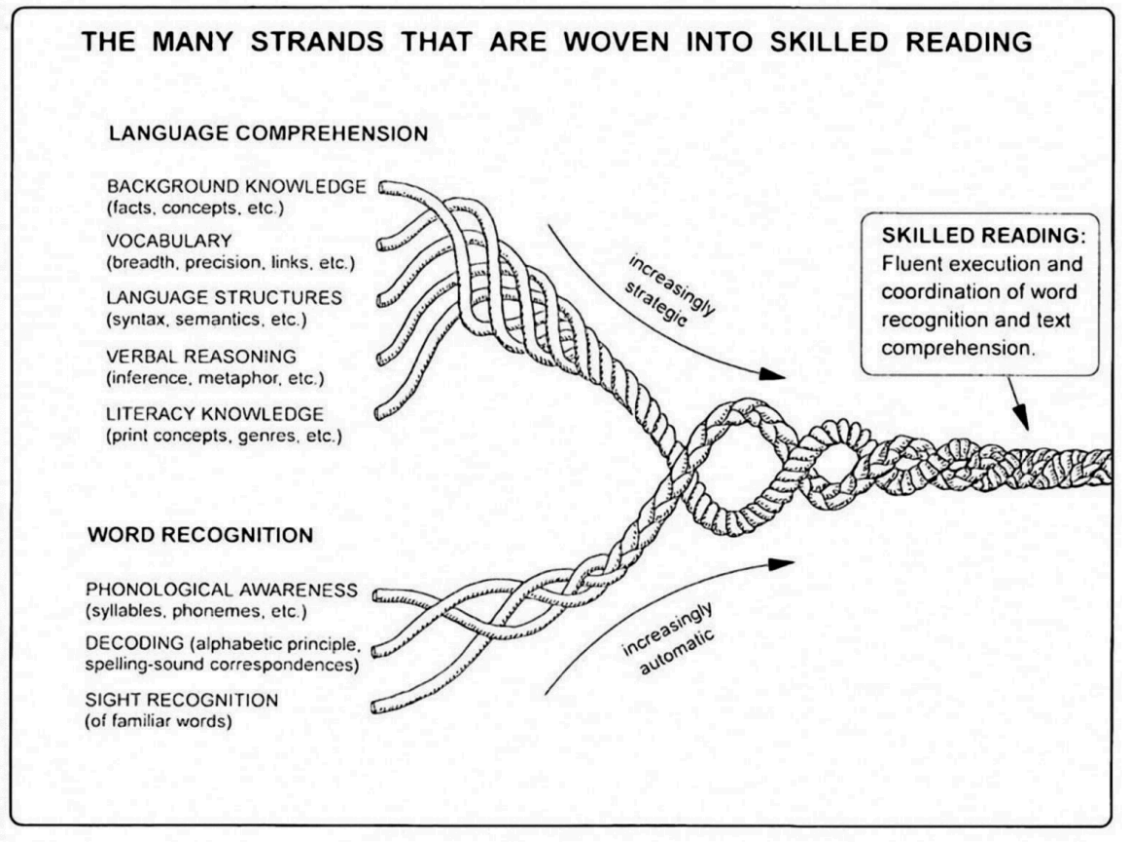
9.3 Validity

To truly understand and enhance a student's reading skills, it is crucial to rely on psychometrically valid instruments grounded in rigorous scientific principles, ensuring accurate and reliable results. Psychometrically valid assessments are designed with meticulous attention to detail, yielding results that can be confidently relied upon to make informed educational decisions. Validity refers to the degree to which evidence and theory support the interpretations of test scores for the proposed usage (AERA, APA, & NCME, 20141). The following sections detail the ongoing collection of validity evidence to support the usage of Amira ISIP's scores. Amira ISIP has been rated with full bubbles for reliability in grades K through 6, suggesting "convincing evidence" by NCII and meeting high industry standards.

9.3.1 Structural Validity

The internal structure of Amira ISIP is founded on the Simple View of Reading (SVR) Framework (SVR-F). Since the introduction of the SVR, hundreds of studies have used this model to guide their investigation and/or interpret their results. Many investigations have directly examined the model's main premise: that reading comprehension is the product of decoding and language comprehension. This work has confirmed that much of the variance in reading comprehension can be accounted for by individual differences in decoding and language comprehension (Catts et al., 2005; de Jong & van der Leij, 2002; Hoover & Gough, 1990). This has been shown to be the case in English readers as well as in readers of other alphabetic orthographies, including Greek (Protopapas et al., 2012), Hebrew (Joshi et al., 2015), and Italian (Tobia & Bonifacci, 2015), as well as non-alphabetic writing systems like Chinese (Ho et al., 2012; see Florit & Cain, 2011, for review). The SVR has also proved successful in explaining and accounting for differences in reading comprehension for second-language learners (Hoover & Gough, 1990; Verhoeven & van Leeuwe, 2012) and dual-language users (Bonifacci & Tobia, 2017).

Figure 9.1 Simple View of Reading Framework from Scarborough (2001)



A key dimension of Amira ISIP's validity evidence is that the scores emanating from the assessment substantially explain the observed variance in outcomes. Amira ISIP uses multiple measures to portray a student's reading ability. These measures show a high degree of correlation, consistent with the mass of reading science research and the SVR-F. But, also consistent with the SVR-F, the multiple measures of reading mastery utilized by Amira ISIP show meaningful divergence. The fact that the structure of Amira ISIP's measures is founded on the most researched and accepted model of reading provides evidence of validity.

Here is the comprehensive explanation of Amira ISIP's measures:

- **WCPM**: ORF measures a student's ability to read aloud with natural ease. WCPM incorporates accuracy (words correct) and speed (minutes spent reading aloud).
- **Decode**: Decode measures a student's ability to combine letter sounds for unfamiliar words. Person names and high-level vocabulary are considered uncommon, whereas sight words are considered familiar. Multi-syllabic words have greater weight in the calculation.
- **Phonological Awareness (PA)**: PA measures a student's ability to pronounce phonemes within words accurately. Students are scored on how well they pronounce all phonemes; PA is an unweighted average of those scores.

- **High Frequency Words (HFW)**: HFW measures the estimated percentage of high frequency words a student has mastered. It is an especially useful measure of reading fluency for younger readers.
- **Vocabulary**: Vocabulary is a measure of a student's ability to understand the meanings of words and phrases in context.

The following tables show the observed correlations among Amira ISIP measures, derived from student samples (N) ranging from 23,023 to 291,492 per grade, with scaled scores used for correlation, depending on the popularity of the Amira ISIP Assessment for a particular grade and language.

Table 9.17 Correlation Matrix among Amira ISIP English Scaled Scores for Kindergarten

N = 163,828	WCPM	Decode	PA	HFW	Vocab
WCPM	1				
Decode	0.72	1			
PA	0.73	0.97	1		
HFW	0.75	0.93	0.92	1	
Vocab	0.83	0.88	0.88	0.95	1

Table 9.18 Correlation Matrix among Amira ISIP English Scale Scores for Grade 1

N = 267,891	WCPM	Decode	PA	HFW	Vocab
WCPM	1				
Decode	0.74	1			
PA	0.76	0.98	1		
HFW	0.81	0.95	0.95	1	
Vocab	0.91	0.84	0.85	0.91	1

Table 9.19 Correlation Matrix among Amira ISIP English Scale Scores for Grade 2

N = 291,492	WCPM	Decode	PA	HFW	Vocab
WCPM	1				

N = 291,492	WCPM	Decode	PA	HFW	Vocab
Decode	0.76	1			
PA	0.76	0.99	1		
HFW	0.75	0.97	0.97	1	
Vocab	0.88	0.86	0.86	0.86	1

Table 9.20 Correlation Matrix among Amira ISIP English Scale Scores for Grade 3

N = 265,800	WCPM	Decode	PA	HFW	Vocab
WCPM	1				
Decode	0.72	1			
PA	0.72	0.99	1		
HFW	0.70	0.97	0.97	1	
Vocab	0.85	0.87	0.87	0.83	1

Table 9.21 Correlation Matrix among Amira ISIP English Scale Scores for Grade 4

N = 163,685	WCPM	Decode	PA	HFW	Vocab
WCPM	1				
Decode	0.73	1			
PA	0.73	0.99	1		
HFW	0.70	0.91	0.9	1	
Vocab	0.85	0.91	0.91	0.78	1

Table 9.22 Correlation Matrix among Amira ISIP English Scale Scores for Grade 5

N = 137,542	WCPM	Decode	PA	HFW	Vocab
WCPM	1				
Decode	0.70	1			

N = 137,542	WCPM	Decode	PA	HFW	Vocab
PA	0.70	0.99	1		
HFW	0.70	0.9	0.9	1	
Vocab	0.82	0.87	0.87	0.76	1

Table 9.23 Correlation Matrix among Amira ISIP English Scale Scores for Grade 6

N = 23,023	WCPM	Decode	PA	HFW	Vocab
WCPM	1				
Decode	0.71	1			
PA	0.71	0.99	1		
HFW	0.71	0.88	0.87	1	
Vocab	0.81	0.94	0.93	0.77	1

9.3.2 Construct Validity

The Amira ISIP suite of assessments and practice tasks is based on a range of activities supported by decades of research that validates their effectiveness in identifying the risk of reading difficulties. The constructs measured by Amira ISIP are rooted in the understanding that developmental reading deficiency primarily manifests as difficulty learning to read and decoding words (Ziegler & Goswami, 2005), even when students are presented with instruction that typically helps students succeed. Much of the work and research that has gone into screener development has focused on isolating the components involved in acquiring word-reading skills within a normal developmental progression. These components include the core skills of PA, alphabetic knowledge (letter-name and letter-sound fluency), decoding, and automaticity. In higher grades, more emphasis is placed on tasks that directly measure how accurately and fluently kids can decode words or pseudo-words (as opposed to reading words by sight). There have been several evaluations of the full screener that demonstrate its validity in identifying children at risk for reading difficulties (Boscardin et al., 2008; Fletcher et al., 2021; Schatschneider et al., 2004).

Each task that is in Amira ISIP's recommended configuration for universal screening has construct validity demonstrated by an extensive body of research. For any task that Amira ISIP recommends be included in the minimum configuration, we require the research to

support 1) a link between the task and behavioral and/or neural correlates of dyslexia or other developmental reading difficulties, and 2) statistically significant differences between the performance of individuals with dyslexia on the task as compared to age-matched controls. The following sections summarize the construct validity evidence for each of these task types.

Phonological Awareness
As established by extensive research, phonological processing is impaired in individuals with dyslexia (Boada & Pennington, 2016; Kovelman et al., 2012; Ramus et al., 2013; Scarborough, 1990; Swan & Goswami, 2007). It is widely accepted in the research community that a deficit of phonological awareness is a core correlate of dyslexia and that phonological/phonemic awareness is a strong predictor of learning to read.

Letter Name Fluency and Letter Sound Fluency
The ability to name and map letters to the sound(s) they make, often known as alphabetic knowledge, is a key precursor to learning to decode words and is highly predictive of later reading achievement. Students' knowledge of individual letter-sound correspondences and their ability to decode pseudo- and non-word words are essential screening information for both predicting risk and informing instruction (Brown, J. E., & Sanford, A. K., 2011). RTI for English language learners: Appropriately using screening and progress monitoring tools to improve instructional outcomes. Brown & Sanford, 2011; Stanovich, 1986).

Pseudo-word/Non-word Decoding
To separate decoding skills from fluency driven by High Frequency Word Recognition, Amira ISIP includes Pseudo-word/Non-word Decoding tasks for students of all grades. Significant research suggests that dyslexic children have specific impairments in decoding (i.e., the phonological deficit hypothesis, see Ramus et al., 2003). Because reading acquisition requires the child to learn the mapping between orthography and phonology (Share, 1995), problems in the representation and use of phonological information inevitably impede reading acquisition (e.g., Goswami & Bryant, 1990).

Rapid Automatized Naming
The RAN task, which requires rapid repetitive naming of stimuli such as numbers, letters, and colors, has been found to be a highly valid signal of dyslexia risk (Denckla & Rudel, 1976; Wolf & Bowers, 1999). Performance on the RAN task has been shown to significantly differentiate children with dyslexia not only from normal controls but also from other learning-disabled children with conditions distinct from dyslexia (Denckla & Rudel, 1976). A deficit in the automatization of verbal responses to visual stimuli, not restricted to symbols, specifically correlates with dyslexia. This study also demonstrated that the deficit is not explained by a generalized slowing of reaction time or lower intelligence quotient (IQ) but, rather, is specific to the executive functions that support verbalizing sequences.

Furthermore, Boscardin et al. (2008) found that measures of precursor reading skills, such as rapid naming, are highly predictive of word-reading learning trajectories in later grades.

In particular, students identified as having rapid naming difficulties in kindergarten exhibited slower development of word recognition skills in subsequent years of the study than age-matched controls. This makes RAN a particularly useful task for dyslexia risk identification in student populations (e.g., Kindergarten, emerging bilingual students) who have had little formal reading instruction or are behind in it.

Word Identification Fluency and Oral Reading Fluency
Especially by Grade 1 and beyond, after students have had a chance to receive formal instruction, word reading, word identification fluency, and oral passage reading are some of the most direct measures of whether a student is experiencing difficulty acquiring word-level reading skills. These skills are highly predictive of reading fluency and comprehension in later grades, including performance on standardized assessments.

Spelling/Encoding
There is a large body of research establishing spelling difficulty as a correlate of dyslexia (Coleman et al., 2009; Ise, 2010; Treiman, 1997; Van Bergen et al., 2012). In particular, spelling problems often persist in individuals with dyslexia even after they have caught up to grade-level in reading through intensive instruction (Treiman, 1997), making the task particularly diagnostic in the older grades.

Reading Comprehension
Deficits in reading comprehension have been directly linked to dyslexia and other specific language impairments (Crain et al., 1990; Hagtvet, 2003). Crain et al. (1990) and others (e.g., Yuill & Oakhill, 1991) have shown that comprehension of spoken sentences is partly a function of working memory, a skill that is severely taxed when a child with a reading disorder parses an unfamiliar text. Shankweiler et al. (1999) similarly demonstrated a strong relationship between comprehension and decoding in children with reading difficulties.

9.3.3 Content Validity

Amira ISIP's content validation process, detailed in Table 9.24 below, includes comprehensive activities from content review to IRT modeling.

Table 9.24 Content Validation Process Step Approach Step Approach

Step	Approach
Creation	Items are written in accordance with psychometrics, style, and cultural sensitivity guidelines.
Review	Each item is meticulously reviewed by experts representing diverse perspectives, and adjustments are made as necessary.
Predictiveness	Items are evaluated for their ability to predict performance outcomes.

IRT Reliability	When applicable, items are evaluated using standard IRT metrics.
Equating	When applicable, items are paired with others to facilitate score equating, ensuring validity across various administrations of the screener.
Bias Review	Differential item functioning (DIF) analysis is conducted to detect items that exhibit differential functioning across subpopulations. Items showing DIF are eliminated.

Key components of the Amira ISIP suite of assessments and practice include passages and words designed for students' oral reading. These passages are meticulously crafted to mirror typical reading development in students and adhere to specific guidelines.

First, passage items are constructed from words aligned with the core curriculum at each grade level. Each word is treated as an item aligned to standards and chosen to cover the standards taught at different points in each grade level's primary language of instruction (English or Spanish). The words in each passage align with the standards relevant to the student's grade.

Second, each story is crafted to conform to a set of standard narrative elements. These elements include:

- A main character(s): who or what the story is mainly about.
- A setting: where and when the story happens.
- A problem: what the main character wants or the problem the character must solve.
- A set of major events: the most important things that happen to solve the problem.
- An outcome: whether or not the problem is solved.

Additionally, expository texts conform to informational text structure and are included in each grade. To determine typical teaching, the content teams consulted reading series, district curriculum guides, and reading standards across various program types (single language, bilingual, immersion, etc.). Thus, word-level features for each passage do not reflect any one publisher's or district's scope and sequence but reflect general reading standards.

9.3.4 Concurrent Validity

Concurrent validity measures the extent of agreement between two distinct assessments measured at the same time. Amira ISIP's concurrent validity evidence was established by comparing ARM scores with those obtained from commonly used external assessments of reading ability: the NWEA MAP Reading assessment and the iReady Diagnostic assessment.

This comparison was conducted using data collected from students in Grades K-3 who took both the Amira ISIP Benchmark Assessment and the external assessment within the same screening window (Fall, Winter, or Spring).

We assessed the validity evidence for Amira ISIP's ARM scores relative to external measures of reading fluency using Pearson's correlation coefficient, which quantifies the strength of the linear relationship between two variables. The table below shows the sample sizes and correlation coefficients for each grade and associated external measure. Across all grades and external assessments included in the analysis, the correlation coefficients indicated a strong positive linear relationship between Amira ISIP's ARM score and the external reading fluency score.

Table 9.25 Correlation Coefficient Between Amira ISIP Scale Scores and External Screener Scores

Grade	External Screener	N	Concurrent Validity
K	NWEA MAP	5,861	0.75
1	NWEA MAP	6,415	0.8
2	NWEA MAP	6,696	0.8
3	iReady Diagnostic	1,065	0.72
3	NWEA MAP	965	0.78

Across all grades and external assessments included in the analysis, the correlation coefficients indicated a strong positive linear relationship between Amira ISIP's ARM score and the external reading fluency score.

9.3.5 Predictive Validity

This section explores the predictive validity of Amira ISIP's assessments by correlating ARM scores obtained in the Fall screening window with those from commonly used external assessments of reading ability administered in the Spring screening window: the iReady Diagnostic assessment and the NWEA MAP Reading assessment.

Data were gathered from students in Grades 1 to 3 who underwent assessments with Amira ISIP in the Fall and the NWEA MAP Reading assessment in the Spring. Additionally, data were obtained from Grade 3 students who participated in the Amira ISIP Assessments in the Fall and the iReady Diagnostic assessment in the Spring. We assessed the validity of Amira ISIP's ARM scores against these external criterion measures of reading fluency using Pearson's correlation coefficient, a measure of the strength of the linear relationship between two variables.

The table below presents the sample sizes and correlation coefficients for each grade and external measure used. For all grades and external assessments included in the analysis, the correlation coefficient representing the relationship between Amira ISIP's ARM score from the Fall screening window and the external reading fluency score from the Spring screening window fell within the range indicating a strong positive linear relationship.

Table 9.26 Predictive Validity Correlation Coefficients

Grade	External Screener	N	Predictive Validity
K	NWEA MAP	5,309	0.71
1	NWEA MAP	6,148	0.81
2	NWEA MAP	6,405	0.79
3	iReady Diagnostic	1,162	0.73
3	NWEA MAP	1848	0.74

9.3.5.1 Predictive Validity Study of Amira ISIP Subscores

This study explores the predictive validity of Amira ISIP's assessments by correlating each of the Amira ISIP screener scores obtained in the Fall screening window with the corresponding subscore from the NWEA MAP Reading assessment that best matches the literary construct associated with each subscore. Data were gathered from students in Grades K and 1 who underwent assessments with Amira ISIP in the Fall and the NWEA MAP Reading assessment in the Winter. We assessed the validity of Amira ISIP's subscores against external criterion measures for each construct using Pearson's correlation coefficient, a measure of the strength of the linear relationship between two variables. Typically, correlation coefficient values fall between 0 and 0.3 for a weak linear relationship, between 0.3 and 0.7 for a moderate linear relationship, and between 0.7 and 1.0 for a strong linear relationship.

Table 9.27 Sample

Grade	State	District	n
Kindergarten	Texas	Lancaster ISD	129
		Vernon ISD	53
	South Carolina	Lancaster Co SD	57
		York School District 1	16
	Oklahoma	Tulsa ISD	52
Kindergarten Total			307
First Grade	Texas	Klein ISD	126

		Lancaster ISD	242
		Tuloso Midway ISD	97
		Vernon ISD	61
	South Carolina	Lancaster Co SD	153
		York School District 1	26
	Kentucky	Christian County PSD	24
	Oklahoma	Tulsa ISD	120
First Grade Total			849

For all grades and subscores included in the analysis, the correlation coefficient representing the relationship between Amira ISIP's subscore from the Fall screening window and the corresponding NWEA MAP subscore from the Winter screening window fell within the range of 0.7-1.0, indicating a strong positive linear relationship.

Table 9.28 Sample sizes and correlation coefficients for each grade, for each external subscore measure utilized.

Grade	Amira ISIP Subscore	Test or Criterion Measure	n	Coefficient
Kindergarten	Phonological Awareness	NWEA MAP: Phonological Awareness Domain Subscore	307	0.74
Kindergarten	Letter-Sound Correspondence	NWEA MAP: Phonics/Word Recognition Domain Subscore	307	0.73
Kindergarten	Rapid Naming	NWEA MAP: Rapid Automatized Naming WCPM	78	0.71
1st Grade	Phonological Awareness	NWEA MAP: Phonological Awareness Domain Subscore	699	0.78

1st Grade	Letter-Sound Correspondence	NWEA MAP: Phonics/Word Recognition Domain Subscore	701	0.70
1st Grade	Rapid Naming	NWEA MAP: Rapid Automatized Naming WCPM	308	0.73
1st Grade	Word or Pseudo Word Reading Fluency	NWEA MAP: Phonics/Word Recognition Domain Subscore	701	0.80
1st Grade	Oral Reading Fluency	NWEA MAP: Oral Reading Fluency Subscore	92	0.72

9.3.6 Externally Conducted Validation Studies

The Amira ISIP Benchmark Assessment is grounded in decades of research supporting its construct validity for flagging risk for reading difficulty. The constructs measured by Amira ISIP are rooted in the understanding that developmental reading deficiency primarily manifests as difficulty learning to read and decode words (Ziegler & Goswami, 2005), even when presented with instruction that typically helps students succeed. Consequently, Amira ISIP's screener is focused on directly observing reading and decoding. Amira ISIP's screening includes phonological awareness, advanced phonemic awareness, sound-symbol recognition, alphabet knowledge, decoding skills, encoding skills, rapid naming, and developmental language.

Each task that is in Amira ISIP's recommended configuration for universal and dyslexia screening has construct validity demonstrated by an extensive body of research. Each task in Amira ISIP's recommended minimum configuration has research to support:

- A link between the task and behavioral and/or neural correlates of dyslexia or other developmental reading difficulties; and
- Statistically significant differences between the performance of individuals with dyslexia on the task as compared to age-matched controls.

Multiple evaluations of the Amira ISIP screener demonstrate its validity in identifying children at risk for reading difficulties (Fletcher et al., 2021; Boscardin et al., 2008; Schatschneider et al., 2004). The following sections summarize the construct validity evidence for each of these task types.

9.3.6.1 Classification Accuracy: Evaluation of Ability to Identify Speech and Language Disorders

Supported by a National Institutes of Health Grant, Dr. Mabel Rice, a renowned reading scientist, and her team at Kansas University evaluated the Amira ISIP Screener's capacity to duplicate the predictive ability of specialized assessments such as Grammagio and the diagnosis of trained specialists in classifying students with reading and/or language disorders such as Dyslexia and Specific Language Impairment (SLI). Results showed that the Amira ISIP Screener scores predicted dyslexia and SLI outcomes with statistical significance ($p < 0.05$).

Classification accuracy analyses showed that Amira ISIP's universal screener scores have a sensitivity of 0.85, a specificity of 0.81, and an AUC of 0.91 in predicting Grammagio's scores. These results suggest that the Amira ISIP flag has high utility for predicting these types of disorders.

A summary of the research findings is presented in Tables 9.20 and 9.21.

Table 9.29 Amira ISIP versus Grammaggio Measures

		Grammaggio Not Flagged for SLI	**Grammaggio Flagged for SLI**	**Totals**	
Amira At-Risk Flag	0	17	3	20	FP=0.19
	1	3	13	16	FN=0.15
	Sum	**20**	**16**	**36**	

Note: Accuracy = 0.83; Sensitivity = 0.85; Specificity = 0.81; Area Under the Curve = 0.91; Phi correlation = 0.66; Tetrachoric correlation = 0.86; FP = false positive; FN = false negative.

		Grammaggio Not Flagged for SLI	**Grammaggio Flagged for SLI**	**Totals**	
Amira At-Risk Flag	0	19	8	27	FP=0.50
	1	1	8	9	FN=0.05
	Sum	**20**	**16**	**36**	

Note: Accuracy = 0.75; Sensitivity = 0.70; Specificity = 0.89; Area Under the Curve = 0.89; Phi correlation = 0.52; Tetrachoric correlation = 0.79; FP = false positive; FN = false negative.

Table 9.30: Parameter estimates from prediction of the Amira ISIP Flag on the Language and Reading Groups

		Estimate	Std. Error	Z Value	PR (>|z|)
Language					
	(Intercept)	-1.73	0.63	-2.77	0.01
	Amira Flag	3.20	0.90	3.57	0.00
Reading					
	(Intercept)	-2.94	1.03	-2.87	0.00
	Amira Flag	2.69	1.14	2.36	0.02

9.3.6.2 External Research on Predictive Accuracy

External research was conducted by Dr. David Francis at the University of Houston (for Dr. David Francis' professional information, please refer to Appendix A) to determine the screener's capacity to predict future reading difficulties.

Longitudinal research involving nearly 5,000 students revealed that the screener successfully identified all but 46 students (1% of the study population) who exhibited significant reading challenges by Grade 2. For Kindergarten students, accuracy ranged between 90% and 95%, demonstrating the screener's efficacy in identifying at-risk students at an early developmental stage. This comprehensive research endeavor generated a diverse array of psychometric evidence affirming the validity and reliability of the screener.

10: PRIVACY & SECURITY

10.1 Amira's Commitment to Privacy and Security

Amira exists to help every child learn to read. That mission depends on the trust of the schools, teachers, and families we serve, and Amira treats the protection of student information as a first-order product requirement — engineered into the platform, verified by independent auditors, and governed by binding contracts with every education agency we serve.

Amira's approach rests on five commitments. First, data minimization: Amira collects only the information necessary to deliver assessment, instruction, and tutoring — nothing more. Second, purpose limitation: student data is used solely for authorized educational purposes; Amira never sells student data, never serves advertising to students, and never builds profiles of children for any non-educational purpose. Third, district control: the local education agency (LEA) owns its student data, directs its use, and can require correction, export, or destruction at any time. Fourth, security by design: all data is encrypted in transit and at rest, access is controlled on the principle of least privilege, and the entire program is examined annually by independent SOC 2 Type II auditors. Fifth, closed-loop AI: Amira's speech and reading models run on Amira's own secured infrastructure; Amira never performs public LLM inference during student sessions and never shares student data with public AI platforms such as ChatGPT or Gemini.

Because Amira listens to children read aloud, we hold ourselves to a standard beyond baseline legal compliance for voice data specifically: recordings are stored only in de-identified form, are never used to create voiceprints or biometric identifiers, and are handled by a single infrastructure sub processor (Amazon Web Services) under strict contractual controls. Districts may disable audio retention entirely.

The remainder of this chapter documents each of these commitments in detail.

10.2 Privacy and Security Compliance at a Glance

The table below summarizes Amira's conformance with each element that education agencies, auditors, and privacy frameworks expect of a student data processor. Every element is independently verifiable through Amira's SOC 2 Type II report, executed Data Sharing Agreements, published policies, and district audit rights.

Amira privacy and security compliance elements:

Expected Privacy & Security Element	Amira
Written, board-approved information security program	YES
Independent SOC 2 Type II examination (AICPA Trust Services Criteria)	YES
Encryption of all data in transit (TLS/HTTPS)	YES
Encryption of all data at rest (AES-256)	YES
Role-based access control on the principle of least privilege	YES
Multi-factor authentication for internal system access	YES
Continuous security monitoring and intrusion detection/prevention	YES
Independent third-party penetration testing	YES
Documented, tested incident response plan	YES
Breach notification to the LEA within 72 hours of confirmation	YES
Data minimization: collection limited to educational necessity	YES
Audio stored exclusively in de-identified form	YES
No biometric templates or voiceprints created or used	YES
No sale of student data; no targeted advertising; no marketing profiles of children	YES
No public LLM inference during student sessions; no student data to public AI platforms	YES
Signed Data Sharing Agreement (DSA) offered to every LEA	YES
Subprocessor flow-down contracts imposing equivalent protections	YES
District-controlled data disposition, deletion, and audio-retention settings	YES
Parent/guardian access, review, and correction pathway (via the LEA)	YES
Employee confidentiality agreements and recurring privacy/security training	YES
Alignment with nationally recognized cybersecurity frameworks (NIST CSF, SP 800-53, SP 800-171; AICPA SOC 2)	YES
U.S. data residency for U.S. users; Canadian data residency for Canadian users	YES

EU-U.S. Data Privacy Framework and UK Extension certification for transatlantic data	**YES**
Recognition of Global Privacy Control (GPC) opt-out signals	**YES**

10.3 Compliance with Governing Standards: FERPA, COPPA, and GDPR

FERPA (Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g)
Amira operates under FERPA's "school official" exception, 34 C.F.R. § 99.31(a)(1), which permits an LEA to disclose education records to a contractor without individual parental consent when three conditions are met. Amira satisfies all three:

- Legitimate educational interest. Amira performs reading assessment, screening, and tutoring functions the LEA would otherwise perform with its own staff.
- Direct control. Under Amira's Data Sharing Agreement, the LEA retains ownership of all student data and Amira acts only at the LEA's direction.
- Redisclosure limitations. Amira adheres to 34 C.F.R. § 99.33: student data is never redisclosed except as the LEA directs, as the DSA permits, or as law requires.

Amira does not create biometric templates and does not use recordings to identify or authenticate any person; recordings are stored de-identified and used solely for educational purposes. Amira also supports LEA obligations under the Protection of Pupil Rights Amendment (PPRA), 20 U.S.C. § 1232h.

COPPA (Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501–6506)
Amira complies with COPPA and the FTC's school-authorization framework, under which schools may consent to the collection of personal information from children under 13 where the information is used solely for educational purposes and never for commercial exploitation. In Amira's implementation:

- Only authorized school officials — never students — create and manage student accounts; there is no open registration and no anonymous access.
- All collection is governed by the district agreement and used exclusively for school-authorized educational purposes.
- Amira does not sell children's data, does not use children's data for behavioral advertising, and does not build profiles of children beyond authorized educational purposes.
- If Amira learns it has collected a child's personal information without school consent, that information is deleted.

GDPR, UK GDPR, and the EU-U.S. Data Privacy Framework
For personal data originating in the European Union and the United Kingdom, Amira complies with the GDPR and UK GDPR and is certified to the EU-U.S. Data Privacy

Framework (DPF) and the UK Extension, as administered by the U.S. Department of Commerce and enforceable by the Federal Trade Commission. Amira's certification is publicly viewable at dataprivacyframework.gov. Key GDPR-aligned commitments include: a documented lawful basis for each processing activity; processor status under school authority with processing limited to school instructions; data subject rights to access, rectification, erasure, restriction, and objection with defined response timelines; breach notification to supervisory authorities within 72 hours; encryption and technical-organizational measures documented in Section "Security Provisions for Audio, Including Encryption"; a designated Privacy Officer; independent recourse mechanisms and binding arbitration for unresolved complaints; and controls that allow a specific education authority to exclude its data from model training upon written request.

10.4 Amira's Privacy Policy

Amira maintains a unified public privacy policy covering the United States, Canada, the European Union, and the United Kingdom, available at all times at amiralearning.com/amira-privacy. The U.S. School Experience policy is reproduced in substance below (last modified June 29, 2026); the Canadian policy (PIPEDA and provincial law) and the EU/UK policy (GDPR, UK GDPR, and DPF) follow the same architecture with jurisdiction-specific rights and are incorporated into this manual by reference.

Scope and Users
The policy explains how Amira Learning, Inc. collects, uses, and discloses information about users of the Amira App, website, and related services. Users include Children (students under 13 enrolled in districts that make Amira available, who may use only the learning portion of the Services) and Educators (teachers and administrators, who alone can create accounts and subscribe).

Children's Information
Amira does not knowingly collect personal information from Children under 13 without school consent, collects and uses Children's information only as needed for authorized educational purposes, deletes any information found to be collected without consent, never sells Children's personal information, never uses Children's information for behavioral ad targeting, and builds no profiles of Children other than to support authorized educational purposes.

Information Collected

- From Educators: registration information (name, email, school affiliation, grade/classes) and roster information for their students (name, school email/credentials, school, class, grade, language, and optionally a parent email address).
- From Children — audio recordings: when a Child interacts with the Amira App, Amira collects recordings of the Child's voice. Recordings enable Amira's algorithms to understand reading strengths and needs, evaluate progress, customize the

experience, and improve Amira's models and the App, subject to the de-identification and training controls described in subsequent sections.
- Automatically: device and log information and analytics from Educator use; and, for Children, in-app activity (stories read, taps, progress) used only to support the internal operations of the Services.

Use and Sharing
Information is used to operate, personalize, and improve the Services; to report reading progress to Educators; to provide support, notices, and (for Educators only) relevant communications; to protect security; and to meet legal obligations. Amira shares information only: with vendors and service providers bound to the commitments of the policy; in response to lawful process; to protect rights and safety; in a corporate transaction where the successor is bound to the policy's commitments; and as aggregated or de-identified data that cannot reasonably identify any individual. Amira never discloses a Child's personal information for marketing or promotional purposes.

Rights, Choices, and Retention
Eligible users may request access, correction, or deletion via Amira's Privacy Rights Request Form or trust@amiralearning.com; students and parents exercise these rights through their school, which verifies identity and authority. Amira honors Global Privacy Control opt-out signals. Unless law requires longer retention, Amira retains personally identifiable information only while the LEA maintains its subscription; upon cancellation or termination, PII is retained sixty (60) days to accommodate subscription lapses and is then destroyed. PII is also destroyed at any time upon LEA request. Amira's Privacy Officer may be reached at trust@amiralearning.com or Amira Learning, Inc., 490 Post St, Suite 500 PMB 2538, San Francisco, CA 94102.

10.5 Student Audio: De-Identified Storage on AWS, and Nowhere Else

Student voice recordings receive the most conservative treatment of any data class in the Amira platform. Two structural rules govern them:

- **Rule 1 — De-identified storage only.** Audio is stored exclusively in de-identified form. Recordings are separated from directly identifying roster information and maintained such that the audio itself does not reasonably identify a specific child. Amira does not create voiceprints or biometric templates from recordings, and recordings are never used to identify or authenticate any person.
- **Rule 2 — A single infrastructure subprocessor.** De-identified audio resides on Amazon Web Services (AWS) infrastructure in the United States (or, for Canadian users, on AWS infrastructure located in Canada). AWS is the only subprocessor with any custody of student audio. Audio files are not provided to, shared with, or made accessible to any other subprocessor, vendor, analytics provider, or third party of any kind. Amira's speech recognition and scoring models are Amira's own and run

within Amira's AWS environment — no external speech-to-text service ever receives student audio.

Districts retain full control: audio retention can be disabled entirely at the district's election, recordings are erased on district request, and audio is destroyed pursuant to the retention schedule in Section “Amira's Privacy Policy“. Teachers may replay a student's recordings for instructional purposes (for example, demonstrating growth at a parent-teacher conference) through authenticated, role-based access only.

10.6 Security Provisions for Audio, Including Encryption

Audio inherits every control in Amira's SOC 2 Type II security program, plus voice-specific safeguards:

- **Encryption in transit.** Every recording is encrypted using TLS from the moment it leaves the student's device until it reaches Amira's environment; all internal service-to-service transfers are likewise encrypted.
- **Encryption at rest.** Audio at rest is encrypted with AES-256 on AWS storage, with managed key rotation; backups are encrypted to the same standard.
- **Access control.** Access to audio is restricted by role-based access control on the principle of least privilege; Amira personnel access to student records requires senior approval and a specific need; all internal access requires multi-factor authentication and is logged.
- **De-identification at rest.** Stored audio is maintained in de-identified form, so that even in the event of unauthorized storage access, recordings are not linked to an identified child.
- **Monitoring and testing.** Intrusion detection and prevention systems monitor the environment continuously; independent third parties conduct penetration tests; vulnerabilities are scanned continuously and patched promptly.
- **Incident response.** A written, tested incident response plan governs any suspected incident, with notification to affected LEAs within 72 hours of confirmation and cooperation with all required regulatory notifications.
- **Secure destruction.** When audio reaches the end of retention or an erasure request is honored, deletion uses secure protocols (overwriting or cryptographic wiping), including purges of backup media on routine cycles.

10.7 Policy: No Audio Sharing with Any Subprocessor Except AWS

It is Amira's firm policy that student audio files are not shared with any subprocessor except Amazon Web Services, which provides the secured infrastructure on which the de-identified recordings reside. This is a categorical rule, not a default: no analytics vendor,

no support platform, no communications provider, no AI service, and no other third party receives student audio, in identified or de-identified form. Where other subprocessors support the platform, they operate on non-audio data classes only. Any future change to this policy would require: (1) advance written notice to every affected LEA; (2) execution of flow-down contractual protections at least as protective as Amira's DSA; and (3) the LEA's opportunity to object and direct otherwise before any transfer occurs. This policy is enforceable by every LEA through the DSA's audit rights and is within the scope of Amira's annual SOC 2 Type II examination.

10.8 Data Sharing Agreement (DSA) Policy and Model DSA

Amira offers — and encourages — a signed Data Sharing Agreement with every school, district, and state education agency it serves. The DSA is the controlling instrument for student data: where its terms conflict with any other writing, including the service agreement, terms of service, or privacy policy, the DSA controls. Amira's model DSA parallels the Student Data Privacy Consortium's National Data Privacy Agreement (NDPA) in structure and substance, and Amira executes NDPA-form agreements with state-specific exhibits wherever a district or state alliance requests. Core DSA commitments include: LEA ownership and control of all student data; Amira's status as a FERPA school official under LEA direction; enumerated authorized uses and a prohibition on all others; no sale of student data and no targeted advertising; parent access and correction through the LEA; law-enforcement request notification; breach notification within 72 hours; LEA audit rights; a cybersecurity framework requirement (NIST and SOC 2); data disposition within 30 days of request; and survival of obligations through any corporate succession, with the LEA holding termination rights over any successor. The complete model DSA follows.

Model Student Data Sharing Agreement (Amira Learning Standard DSA Template, v1.0)

STUDENT DATA SHARING AGREEMENT — Amira Learning Standard DSA Template, Version 1.0
This Student Data Sharing Agreement ("DSA") is entered into on the date of full execution (the "Effective Date") by and between _____ (the "LEA"), located at _____, and Amira Learning, Inc. (the "Provider"), located at 490 Post St., Suite 500, PMB 2538, San Francisco, CA 94102.

WHEREAS, the Provider is providing educational or digital services to the LEA; WHEREAS, the Provider and the LEA recognize the need to protect personally identifiable student information and other regulated data exchanged between them as required by applicable laws and regulations, such as the Family Educational Rights and Privacy Act ("FERPA") at 20 U.S.C. § 1232g (34 C.F.R. Part 99); the Children's Online Privacy Protection Act ("COPPA") at 15 U.S.C. §§ 6501-6506 (16 C.F.R. Part 312); the Protection of Pupil Rights Amendment ("PPRA") at 20 U.S.C. § 1232h; and the applicable student data privacy laws and regulations of the State of _____; and WHEREAS, the Provider and the LEA desire to enter into this

DSA for the purpose of establishing their respective obligations and duties in order to comply with applicable laws and regulations. NOW THEREFORE, for good and valuable consideration, the LEA and Provider agree as follows:

A description of the Services to be provided, the categories of Student Data that may be provided by the LEA to Provider, and other information specific to this DSA are contained in the Standard Clauses hereto. In the event there is conflict between the terms of this DSA and any other writing, including, but not limited to, the Service Agreement and Provider Terms of Service or Privacy Policy, the terms of this DSA shall control. This DSA shall stay in effect for the term of the LEA's subscription to Amira unless earlier terminated as provided herein. The services to be provided by Provider to the LEA pursuant to this DSA are detailed in Exhibit "A" (the "Services"). Notices may be given via e-mail or first-class mail to the parties' designated representatives. [Signature blocks and representative designations appear in the executable template.]

STANDARD CLAUSES — ARTICLE I: PURPOSE AND SCOPE
Purpose of DSA. The purpose of this DSA is to describe the duties and responsibilities to protect Student Data, including compliance with all applicable federal, state, and local privacy laws, rules, and regulations, all as may be amended from time to time. In performing the Services, the Provider shall be considered a School Official with a legitimate educational interest, performing services otherwise provided by the LEA. Provider shall be under the direct control and supervision of the LEA with respect to its use of Student Data. The LEA agrees that the Provider is a School Official acting on the LEA's behalf.

Student Data to Be Provided. In order to perform the Services described above, the LEA shall provide Student Data as identified in the Schedule of Student Data, attached hereto as Exhibit "B". DSA Definitions. The definition of terms used in this DSA is found in Exhibit "C". In the event of a conflict, definitions used in this DSA shall prevail over terms used in any other writing.

ARTICLE II: DATA OWNERSHIP AND AUTHORIZED ACCESS
Student Data Property of LEA. All Student Data transmitted to the Provider pursuant to the Service Agreement is and will continue to be the property of and under the control of the LEA. All copies of such Student Data, including any modifications or additions or any portion thereof from any source, are subject to the provisions of this DSA in the same manner as the original Student Data. As between the Parties, all rights, including all intellectual property rights in and to Student Data, shall remain the exclusive property of the LEA. For the purposes of FERPA, the Provider shall be considered a School Official, under the control and direction of the LEA as it pertains to the use of Student Data.

Parent Access. To the extent required by law, the LEA shall establish reasonable procedures by which a parent, legal guardian, or eligible student may review Education Records and/or Student Data, correct erroneous information, and procedures for the transfer of student-generated content to a personal account. Provider shall respond in a reasonably timely manner (and no later than forty-five (45) days from the date of the request, or sooner where state law requires) to the LEA's request for Student Data in a student's

records held by the Provider. If a parent or other individual contacts the Provider directly to review Student Data, the Provider shall refer them to the LEA. Separate Account. If Student Generated Content is stored by the Provider, Provider shall, at the LEA's request, transfer or provide a mechanism to transfer it to a student's personal account. Law Enforcement Requests. Should law enforcement or other government entities contact Provider with a request for Student Data, the Provider shall notify the LEA in advance of a compelled disclosure, unless lawfully directed otherwise.

ARTICLE III: DUTIES OF LEA

The LEA shall: provide Student Data in compliance with all applicable laws; include its "School Official" criteria in its FERPA annual notification of rights where applicable; take reasonable precautions to secure credentials and obtain consents as required by governing law; and promptly notify Provider of any known unauthorized access and assist in investigation and response.

ARTICLE IV: DUTIES OF PROVIDER

Privacy Compliance. The Provider shall comply with all applicable federal and state laws and regulations pertaining to data privacy and security, including FERPA, COPPA, PPRA, the applicable student data privacy laws of the State, and all other applicable laws, as amended. Authorized Use. Student Data, including persistent unique identifiers, shall be used for no purpose other than the Services in Exhibit "A", the purposes authorized in Section 6 of this Article IV, or as stated in the Service Agreement and otherwise authorized under the statutes referred to in this DSA. Provider Employee Obligation. Provider shall require all employees and agents with access to Student Data to comply with this DSA and shall maintain appropriate confidentiality agreements with each.
No Disclosure. Provider shall not re-disclose any Student Data or any portion thereof, including user content or other non-public or personally identifiable information, other than as directed or permitted by the LEA or this DSA. This prohibition does not apply to aggregate summaries of De-Identified information, disclosures pursuant to lawfully issued subpoena or other legal process, or to Subprocessors performing services on behalf of the Provider pursuant to this DSA. Provider will not Sell Student Data to any third party.

De-Identified Data. Provider agrees not to attempt to re-identify De-Identified Student Data. De-Identified Data may be used for purposes allowed under FERPA and for: (1) assisting the LEA or other governmental agencies in conducting research and other studies; (2) research and development of the Provider's educational sites, services, or applications, and to demonstrate the effectiveness of the Services; and (3) adaptive learning purposes and customized student learning. Provider's use of De-Identified Data shall survive termination of this DSA or any request to return or destroy Student Data. Except for Subprocessors, Provider agrees not to transfer De-Identified Student Data to any party unless (a) that party agrees in writing not to attempt re-identification, and (b) prior written notice has been given to the LEA, who has provided prior written consent. Prior to publishing any document naming the LEA, the Provider shall obtain the LEA's written approval of the manner in which De-Identified Data is presented.

Consent to Use of Data for Educational Purposes and Product Improvement. The LEA hereby consents to, and authorizes, Provider's use of Student Data for the following purposes, each of which the Parties agree constitutes a legitimate educational interest and use on behalf of the LEA: (1) to provide, deliver, maintain, support, and secure the Services; (2) for adaptive learning and customized or personalized student learning, including generating individualized instruction, practice, feedback, and personalized learning recommendations for students; (3) to generate assessments, screenings, reports, and analytics for the LEA, its teachers, and administrators, and, where applicable, for parents and legal guardians; (4) to maintain, develop, support, evaluate, diagnose, and improve Provider's educational products and services, including improving and training the accuracy and efficacy of Provider's reading assessment, speech recognition, and instructional models; and (5) to demonstrate the effectiveness of the Services. Where Student Data is used pursuant to subsections (4) or (5), Provider shall (a) apply the security and confidentiality protections of this DSA to all such data; (b) De-Identify such data prior to any use or disclosure outside of Provider and its Subprocessors; and (c) not use such data for Targeted Advertising or to Sell Student Data.

Disposition of Data. Upon written request from the LEA, Provider shall dispose of or provide a mechanism to transfer Student Data within thirty (30) days of the request. Upon termination of this DSA, if no written request is received, Provider shall dispose of all Student Data after reasonable prior notice to the LEA. The duty to dispose does not extend to De-Identified Data or data placed in a separate student account. The LEA may employ the "Directive for Disposition of Data" (Exhibit "D"). Advertising Limitations. Provider is prohibited from using, disclosing, or selling Student Data to (a) inform, influence, or enable Targeted Advertising; or (b) develop a profile of a student, family member/guardian, or group, for any purpose other than providing the Services. This does not prohibit adaptive learning, product recommendations to teachers or LEA employees, or notices to account holders about education product updates.

ARTICLE V: DATA PROVISIONS

Data Storage. Where required by applicable law, Student Data shall be stored within the United States. Upon request, Provider will provide a list of locations where Student Data is stored. Audits. No more than once a year, or following unauthorized access, upon written request with at least ten (10) business days' notice and execution of an appropriate confidentiality agreement, the Provider will allow the LEA to audit the security and privacy measures in place. The Provider will cooperate reasonably with the LEA and any local, state, or federal agency with oversight authority, providing reasonable access to facilities, staff, agents, Student Data, and records. Failure to reasonably cooperate is a material breach.

Data Security. The Provider agrees to utilize administrative, physical, and technical safeguards designed to protect Student Data from unauthorized access, disclosure, acquisition, destruction, use, or modification, shall adhere to applicable data security law, and shall implement an adequate Cybersecurity Framework based on nationally recognized standards set forth in Exhibit "E" (NIST SP 800-53; NIST CSF; NIST SP 800-171; AICPA SOC 2). Data Breach. In the event of an unauthorized release, disclosure, or acquisition of Student Data, the Provider shall notify the LEA's designated representative immediately, but not

longer than seventy-two (72) hours after confirmation of the incident, including a forensic report, an account of data that may have been compromised, and corrective actions. Notice shall be clear and conspicuous and include the statutory elements (nature of the incident, data types involved, protective steps, contact information, vigilance guidance, dates or date ranges, and any law-enforcement delay). Provider shall maintain a written incident response plan consistent with industry standards and provide a summary on request, and shall cooperate with the LEA to secure Student Data in the event of a breach originating from the LEA's use of the Service.

ARTICLE VI: MISCELLANEOUS

Termination and Survival. The Parties may terminate by mutual written consent; upon termination of the Service Agreement, Provider shall dispose of Student Data pursuant to Article IV, Section 7. Priority of Agreements. This DSA governs the treatment of Student Data and takes precedence over the Service Agreement, Terms of Service, Privacy Policies, or any other writing in the event of conflict. Entire Agreement; Counterparts; Electronic Signatures; Severability; Governing Law and Venue [State/County to be designated]. Successors Bound. This DSA binds successors in interest; the Provider shall give written notice within sixty (60) days after closing of any sale, merger, or disposition, with signed assurance that the successor assumes the DSA's obligations; the LEA may terminate if it disapproves of the successor. Authority; Waiver.

EXHIBIT A — DESCRIPTION OF SERVICES

Amira Learning licenses include access to Amira ISIP Assess and Amira Tutor. Amira Tutor is an AI-based literacy platform that listens to students read aloud, provides real-time feedback and coaching, and generates detailed reports on reading accuracy, fluency, and comprehension. Amira ISIP Assess includes screening, benchmarking, and progress-monitoring tools aligned with current research in the science of reading, including frameworks such as the Active View of Reading and the Multiple Deficit Model, supporting early identification of reading difficulties, including indicators of dyslexia. Participating schools receive access for all students in designated grades, along with teacher and staff training, implementation support, and access to student-level data and reporting tools.

EXHIBIT B — SCHEDULE OF STUDENT DATA

Categories and elements collected, used, or processed: Application Technology Metadata (IP addresses, cookies, browser user agent); Application Use Statistics (metadata on user interaction); Assessment (reading assessment, screening, benchmarking, and progress-monitoring observation data generated by the Services); Demographics (gender, ethnicity/race, English proficiency status, special education flag/IEP/504, and native or primary language — accepted via district roster if provided; not required); Enrollment (school enrollment, grade level, class enrollment); Schedule (teacher names); Student Identifiers (local district ID, state-assigned ID, provider/app-assigned ID, app username and password); Student Name (first and/or last); Student In-App Performance (reading level, accuracy, fluency, and comprehension measures generated by the Services). The Provider does not collect or process categories or elements other than those identified, and will notify the LEA and obtain any required consent or amendment before collecting any additional categories.

EXHIBIT C — DEFINITIONS (SUMMARY)
The template defines, among other terms: De-Identified Data (records from which all personally identifiable information has been removed or obscured such that remaining information does not reasonably identify a specific individual); Educational Records; LEA; Metadata; Operator; Provider; Student Generated Content; School Official (per 34 C.F.R. § 99.31(a)(1)(i)(B)); Service Agreement; Student Data (broadly defined to include, among other elements, voice recordings, biometric information, and any information that would identify a specific student; excluding anonymized or de-identified data); Subprocessor (a party other than the LEA or Provider whom Provider uses for data collection, analytics, storage, or other service to operate and/or improve its service and who has access to Student Data); Subscribing School; Targeted Advertising; and Third Party.

EXHIBIT D — DIRECTIVE FOR DISPOSITION OF DATA
A standing form by which the LEA directs disposition (partial or complete; by destruction/deletion or by transfer; on a designated schedule or as soon as commercially practicable), executed by the LEA's authorized representative and verified by the Provider's authorized representative.

EXHIBIT E — DATA SECURITY REQUIREMENTS
The Provider has implemented and maintains an adequate cybersecurity program based on the following nationally recognized Cybersecurity Frameworks: National Institute of Standards and Technology — NIST SP 800-53, the Cybersecurity Framework for Improving Critical Infrastructure Cybersecurity (CSF), and Special Publication 800-171; and the American Institute of Certified Public Accountants — Systems and Organization Controls (SOC 2). Any exclusions, variations, or exemptions, and any additional security programs and measures, are detailed in an attachment to Exhibit "E".

10.9 Consents Provided by Districts to Amira

Amira's legal authority to serve students flows from the district, and the consent architecture has three layers:

- **FERPA school-official designation.** By executing the DSA and service agreement, the LEA designates Amira a school official with a legitimate educational interest under 34 C.F.R. § 99.31(a)(1), authorizing the rostering of students and the sharing of education records — including reading performance data and voice recordings — without individual parental consent, subject to the LEA's direct control and FERPA's redisclosure limits.
- **COPPA school authorization.** The school provides consent on behalf of parents for collection of personal information from children under 13, valid because the information is used strictly for educational purposes and never for commercial exploitation, consistent with the FTC's school-consent framework.
- **Enumerated DSA consents.** Article IV, Section 6 of the DSA records the LEA's express consent to five defined uses: service delivery and security; adaptive and personalized learning; generation of assessments, reports, and analytics for

educators and families; maintenance, evaluation, and improvement of Amira's reading assessment, speech recognition, and instructional models; and demonstration of service effectiveness — with improvement uses conditioned on de-identification before any use or disclosure outside Amira and its subprocessors, and with targeted advertising and sale categorically prohibited.

Districts additionally control operational consents: enabling or disabling audio retention, directing erasure, authorizing (or declining) any use of their data in model training where state law or district policy so requires, and approving any publication that references the district. Amira supports district-led parent notification and opt-out processes and provides plain-language documentation for families on request.

The Family Educational Rights and Privacy Act (FERPA) generally prohibits SEAs & LEAs from disclosing personally identifiable information (PII) from education records without written parental consent. Congress and the U.S. Department of Education, however, have long recognized that schools do not perform every institutional function with W-2 employees: they engage contractors — for student information systems, learning platforms, assessments, transportation routing, food service, and more — and those contractors necessarily handle education records. The "school official" exception, 34 C.F.R. § 99.31(a)(1), is the legal mechanism that makes this possible while preserving FERPA's protections. The SEA/LEA exception permits disclosure without individual consent to a contractor that (1) performs an institutional service or function for which the school would otherwise use employees, (2) operates under the school's direct control with respect to education records, and (3) is bound by FERPA's redisclosure limitations.

The Exception is grounded in these three well-established characteristics. First, the exception is not a loophole but a deliberately designed, twice-codified regulatory framework — formalized in the Department's 2008 rulemaking, reaffirmed in 2011, and elaborated in a decade of Department guidance — that conditions the privilege of access on district control and educational purpose. Second, the Exception is the operating basis for the entire American education technology sector: thousands of companies, and by extension the digital infrastructure of nearly every U.S. school district, function as school officials. Amira relies on the industry-standard, Department-sanctioned architecture for student data use. Third, the framework is court-tested and court-sanctioned — in the Supreme Court, in federal district courts, in state attorney general actions, and at the Federal Trade Commission — and it has held. Courts and regulators have policed its boundaries, consistently punishing commercial exploitation of student data while leaving the core educational-outsourcing framework intact. No court has invalidated the school official regulation. The judicial record therefore defines a safe harbor with remarkable clarity: a provider under SEA/LEA control, using data solely for authorized educational purposes, with no advertising or sale, is allowed to receive, process and retain student data.

The school official exception is not a niche arrangement; it is the legal foundation of the American K-12 digital ecosystem. The scale is documented:

- **Tool counts.** Instructure/LearnPlatform's widely cited EdTech Top 40 research, based on billions of usage data points, has found that U.S. school districts access on

the order of 2,500 or more distinct edtech tools in a single school year, with thousands of unique products observed across districts nationally. Each tool that touches an education record — a roster, a score, a student work product — operates, if it operates lawfully without individual parental consent, as a school official under § 99.31(a)(1) or under a state-law contract regime built on the same structure.

- **Contract registries.** The Student Data Privacy Consortium's national registry and its state alliances catalog tens of thousands of executed data privacy agreements between districts and providers, the overwhelming majority built on the National Data Privacy Agreement — whose Article I, like Amira's DSA, expressly designates the provider a school official under the LEA's direct control. State registries required by statute (New York's § 2-d supplemental postings, Illinois SOPPA postings, Connecticut's registry) document the same population of vendors on the same basis.
- **The infrastructure layer.** Every category of core district software operates under this framework: student information systems (PowerSchool, Infinite Campus, Skyward), learning management systems (Canvas, Schoology, Google Classroom), productivity suites (Google Workspace for Education, Microsoft 365 Education), rostering and identity providers (Clever, ClassLink), assessment platforms (NWEA MAP, Curriculum Associates i-Ready, Renaissance Star, Amplify mCLASS, Istation/Amira), communications systems, transportation, nutrition, and special-education case management. The U.S. Department of Education's guidance assumes this reality: its 2014 online-services guidance and successors are written to help districts use the school official exception properly, not to discourage its use.
- **Voice and audio specifically.** Audio is not novel within the framework: districts have for decades administered assessments that record student voices through contractors — most prominently the speaking domains of federally required English learner assessments (e.g., WIDA ACCESS, administered and centrally scored by Data Recognition Corporation), along with world-language proficiency tests and reading-fluency tools. The school official/assessment-contractor structure has carried recorded student speech since long before AI entered the conversation.

The significance for any evaluation of Amira is proportionality: Amira's legal basis is the same one on which a typical district relies more than two thousand times over. The relevant question for a district or policymaker is therefore never whether the school official exception is legitimate — the entire sector, with the Department's blessing, answers that — but whether a given provider satisfies its conditions in fact. Sections 3 and 6 document that Amira does, with a contract that implements each condition expressly.

10.10 Challenges and Judicial Precedent

The FERPA framework, and the outsourcing model built on it, have been tested repeatedly across five decades. The precedent sorts into three groups: cases defining FERPA's enforcement architecture, challenges to the regulatory framework itself, and enforcement actions policing the boundaries of school-based consent. Together they yield a consistent doctrine: the framework stands; its abuse is punished.

Gonzaga University v. Doe, 536 U.S. 273 (2002). The Supreme Court held that FERPA's nondisclosure provisions create no personal rights enforceable under 42 U.S.C. § 1983 — there is no private right of action under FERPA. Enforcement runs exclusively through the Secretary of Education's funding-condition authority, administered today by the Student Privacy Policy Office, which investigates complaints and secures voluntary compliance, with fund termination as the ultimate sanction. For the school official framework, Gonzaga means the compliance relationship is institutional: districts answer to the Department, and vendors answer to districts through contract — which is exactly why a rigorous DSA matters so much, since the district's contractual remedies (and § 99.33(e) debarment) are the operative enforcement mechanism against a provider.

United States v. Miami University, 294 F.3d 797 (6th Cir. 2002). The Sixth Circuit confirmed the other half of the architecture: the United States may sue to enjoin violations of FERPA's conditions, and FERPA's protections are real, binding obligations — not precatory. The pair of 2002 decisions thus establishes that FERPA is enforced vertically (Department to institution) rather than through private tort-style litigation.

Owasso Independent School District v. Falvo, 534 U.S. 426 (2002). The Court held that peer-graded papers are not "education records" before a teacher records them, reading "maintained" to suggest records kept in a central custodial sense. Owasso matters here for its interpretive posture: the Court declined to read FERPA in a manner that would federalize every classroom practice, signaling judicial reluctance to convert FERPA into an instrument that disrupts ordinary educational operations — the same practical spirit that animates the school official exception.

Challenges to the Regulatory Framework

Electronic Privacy Information Center (EPIC) v. U.S. Department of Education, 48 F. Supp. 3d 1 (D.D.C. 2014). The most direct legal attack on the Department's modernized FERPA regulations. EPIC challenged the 2011 amendments under the Administrative Procedure Act, arguing the Department had exceeded its statutory authority in broadening information-sharing definitions. The district court dismissed for lack of Article III standing without reaching the merits. The practical upshot: the Department's regulatory framework — including the 2008 outsourcing codification, which has now governed for nearly two decades — has never been vacated, enjoined, or held unlawful by any court. A regulation of this consequence, in continuous nationwide operation since 2008, unsuccessfully challenged and repeatedly reaffirmed in Department guidance, occupies about as settled a position as administrative law affords.

Policing the Boundaries: Commercial Use Cases and Enforcement

In re Google Inc. Gmail Litigation, No. 13-MD-02430 (N.D. Cal. 2014), and the Apps for Education advertising controversy. Plaintiffs, including student users of Google Apps for Education, alleged that Google's scanning of email content for advertising purposes violated wiretap laws; the litigation revealed that student accounts were subject to content scanning. The court denied class certification, but the reputational and legal pressure

produced the structural result: in April 2014 Google announced it had permanently removed ads scanning from Apps for Education. The episode is the sector's foundational cautionary tale, and its lesson is a precise statement of the school official exception's boundary — a provider's authority over student data extends exactly as far as the educational purpose, and commercial exploitation forfeits the framework's protection along with the market's trust.

Balderas v. Google LLC (D.N.M., filed Feb. 2020; 489 F. Supp. 3d 1254 (D.N.M. 2020)). Of particular relevance to New Mexico stakeholders: the state's Attorney General sued Google over children's data collected through Chromebooks and G Suite for Education, alleging COPPA violations and related state-law claims, and contesting the sufficiency of school-based consent. The federal district court dismissed the COPPA claim at the pleading stage in September 2020, and the matter thereafter resolved with Google undertaking enhanced student-privacy commitments in New Mexico without any judicial ruling invalidating the school-consent framework. The takeaway is two-sided and honest: state attorneys general are active and willing to test edtech data practices — and the school-consent architecture emerged from the test intact, with the enforcement energy directed at transparency and use limits rather than at the framework itself.

United States v. Edmodo, LLC (N.D. Cal. 2023) and the FTC's May 2023 Policy Statement. The Federal Trade Commission's action against Edmodo is the controlling modern statement of where school-based authorization ends. The FTC alleged Edmodo collected students' personal information and used it for advertising, and unlawfully pushed its COPPA compliance duties onto schools without providing the required notices — resulting in a stipulated order.

State-law overlays. Finally, the states have legislated the same boundary rather than displacing the framework: the forty-plus SOPIPA-family statutes (California Bus. & Prof. Code § 22584 and its progeny), New York Education Law § 2-d, Illinois SOPPA, and their counterparts all presuppose that operators will hold student data under school engagement — and then prohibit precisely the conduct the Google and Edmodo matters involved: sale, targeted advertising, and non-educational profiling. The legislative record thus confirms the judicial one: fifty state legislatures examined vendor access to student data and, rather than abolishing it, codified its conditions.

Synthesis

Across the Supreme Court's FERPA cases, the failed APA challenge, the state AG litigation, and the FTC's enforcement program, one doctrine emerges with unusual clarity. The school official / school consent framework is lawful, settled, and universally relied upon; what courts and regulators punish is its abuse — the use of children's data for the provider's own commercial benefit. Every adverse outcome in the record involves advertising, monetization, or opacity. No adverse outcome involves a provider that confined itself to the district-authorized educational purpose under genuine district control. The precedent therefore does more than permit Amira's model; it maps the exact conduct that keeps a provider safe, and that conduct — no sale, no advertising, no non-educational profiling, district ownership and control, enumerated purposes, transparency, and enforceable

contracts — is a description of Amira's Data Sharing Agreement and published commitments.

Application to Amira
Amira operates within the school official exception the way the Department of Education's guidance instructs: the institutional function (reading assessment and instruction) is one districts otherwise perform with employees; direct control is established and operationalized by state or district contract, and § 99.33(a)'s use and redisclosure limits are implemented. Amira sits on the compliant side of every line the precedent draws: it does not advertise, does not sell, does not profile for non-educational purposes, does not shift its own duties onto schools, and documents its practices in instruments districts can enforce and auditors can examine.

When a parent, board member, or legislator asks "who allowed a company to receive my child's reading data without my signature?", the accurate answer has three parts. Congress did, in 1974, when it vested the school-official determination in the educational agency; the Department of Education did, in 2008 and since, when it codified and repeatedly reaffirmed the conditions under which contractors may serve as school officials; and the parent's SEA or LEA did, when it executed an agreement or contract with Amira. Amira is authorized by the same legal act the district performs for the thousands of other tools its students touch each year. The framework has been challenged and has held; its boundaries have been enforced, and Amira honors them. Amira's consent is in its established institutional form — the one federal law has provided for half a century, refined for the digital era, and entrusted to the educational authorities accountable to the communities they serve.

10.11 Subprocessor Policy

Amira engages a deliberately small set of subprocessors and holds each to the same commitments Amira makes to districts. The policy has six elements: (1) Necessity — a subprocessor is engaged only where required to operate, secure, or support the Services (e.g., cloud infrastructure, communications delivery, support tooling); (2) Due diligence — every subprocessor is vetted for security posture, certifications, and legal compliance before engagement and periodically thereafter; (3) Flow-down contracts — every subprocessor executes a data processing agreement imposing confidentiality, security, purpose-limitation, and breach-notification obligations at least as protective as Amira's DSA, with immediate notice required if the subprocessor can no longer meet them; (4) Purpose limitation — subprocessors may process data solely to perform their function for Amira, never for their own purposes, marketing, or advertising; (5) Data-class restriction — student audio is restricted to AWS alone, and no subprocessor receives more data than its function requires; and (6) Accountability — Amira maintains records of all subprocessor transfers, remains fully responsible and liable for subprocessor handling of personal data (including under the EU-U.S. DPF), and makes its subprocessor list available to LEAs, with notice of changes and the opportunity to object.

10.12 State Legislation: Landscape and Amira's Conformance

Student data privacy in the United States is governed by a lattice of state statutes layered atop FERPA, COPPA, and PPRA. More than forty states have enacted student privacy laws since California's SOPIPA (2014) established the operator model — no targeted advertising, no profiling, no sale, reasonable security, deletion on request — which most states have since adapted. A second wave, exemplified by New York's Education Law § 2-d and Connecticut's Public Act 16-189, added mandatory contract terms, parents' bills of rights, framework-based security requirements, and breach-notification duties. A third wave is now emerging that regulates AI in education specifically: in 2026 alone, 77 bills across 27 states address AI in classroom instruction, including procurement standards, approved-tool lists, and restrictions on training AI systems with student data. Amira's compliance strategy is to conform to the strictest common denominator — the DSA's protections apply nationally, not only where mandated — and to add state-specific riders wherever statute requires. The table below summarizes representative statutes and Amira's conformance.

The table below outlines representative state statutes and Amira conformance. This table is illustrative of Amira's approach across all states; Amira executes state-specific agreements and exhibits wherever required and monitors pending legislation (including New Mexico's LESC recommendations and similar 2026-27 proposals) through its policy team.

Jurisdiction / Law	Core Requirements	Amira Conformance
California — SOPIPA (Bus. & Prof. Code § 22584)	Prohibits targeted advertising, profiling, and sale of K-12 student data; requires reasonable security; deletion on district request.	Amira does not sell student data, serves no advertising, and builds no non-educational profiles; security program and deletion workflows satisfy SOPIPA's operator duties.
California — AB 1159 (2025)	Prohibits operators from using student data to train generative AI systems; expands student privacy protections; creates a private right of action.	Amira maintains per-LEA controls that exclude an LEA's data from model training absent specific written authorization; California deployments operate with training exclusions applied consistent with the statute.

New York — Education Law § 2-d / Part 121	Parents' Bill of Rights, supplemental information, NIST CSF alignment, encryption, breach notice, and data-privacy contract terms.	Amira executes § 2-d riders with New York LEAs, publishes supplemental information, encrypts data in transit and at rest, and aligns its security program to the NIST Cybersecurity Framework.
Connecticut — Public Act 16-189	Written contracts with prescribed student-privacy terms; breach notification to districts and the state.	Amira's DSA incorporates the required contract terms, and its 72-hour LEA breach-notification commitment meets the Act's timelines.
Texas — Education Code Ch. 32 (student data) and HB 18 (SCOPE Act)	Student data privacy duties for operators; duties regarding minors' online experience, harmful content, and data collection.	Amira collects only school-authorized educational data, provides no open registration, no advertising, and no unmoderated content, and supports district agreements under Texas law.
Illinois — SOPPA (105 ILCS 85) and BIPA (740 ILCS 14)	SOPPA: written agreements, posted subcontractor lists, breach notice, no sale. BIPA: consent and policy duties for biometric identifiers, including voiceprints.	Amira executes SOPPA-compliant agreements and posts required disclosures. Amira does not create or use voiceprints or any biometric template; recordings are never used to identify a person, placing Amira's processing outside BIPA's biometric-identifier scope while district consent frameworks provide additional cover.
Colorado — Student Data Transparency and Security Act	Contract provisions, transparency lists, collection limits, and destruction duties for school service providers.	Amira's DSA, published data schedule (Exhibit B), and disposition process (Exhibit D) implement each duty.
Idaho — SB 1227 (2026)	Requires AI tool procurement to comply with state and federal privacy law; authorizes an approved-tools list; prohibits AI from replacing teachers.	Amira, a statewide Idaho partner, is designed as a teacher-support tool and satisfies the procurement privacy requirements; Amira supports and participates in approved-tool vetting.
Nevada — AB 406 (2025)	Prohibits public schools from using AI to perform counselor, psychologist, or social-worker functions.	Not applicable by design: Amira performs literacy assessment and tutoring only and performs no

		counseling or mental-health functions.
New Mexico — PED AI Guidance; LESC 2026 policy recommendations (pending)	Non-binding guidance today; recommendations for AI procurement standards, privacy requirements, and a vetted-tools list are before the Legislature.	Amira operates statewide under PED's assessment program, supports districts' ability to disable audio retention, and has committed to participate in any vetting or evaluation framework New Mexico adopts.
NDPA / SDPC member states (MA, CA, TX, OH, IL, and others)	Standardized National Data Privacy Agreement with state-specific exhibits, adopted by state alliances of the Student Data Privacy Consortium.	Amira's model DSA parallels the NDPA's structure and obligations, and Amira executes NDPA-form agreements, including state exhibits, wherever an alliance district requests them.
Utah, Virginia, Nevada, Colorado, Connecticut (consumer privacy statutes)	State consumer-privacy rights (access, correction, deletion, opt-out) with education carve-outs and FERPA interplay.	Amira's Additional Privacy Notices for U.S. Residents document state-specific rights, and Amira honors verified requests, including Global Privacy Control signals.

10.13 SOC 2 as a Compliance Process, and Amira's Adherence

What SOC 2 Is

Systems and Organization Controls (SOC) 2 is the American Institute of Certified Public Accountants' (AICPA) attestation framework for service organizations that hold customer data. A SOC 2 examination is performed by an independent, licensed CPA firm against the AICPA's Trust Services Criteria: Security (required), plus Availability, Processing Integrity, Confidentiality, and Privacy as applicable. Amira's examination covers the Security, Availability, and Confidentiality trust services criteria. The framework evaluates the design and operation of controls across risk management, access control, change management, system operations, monitoring, vendor management, and incident response. Two report types exist: a Type I report assesses whether controls are suitably designed at a point in time; a Type II report — the more rigorous standard — tests whether those controls actually operated effectively over an extended observation period (typically 6–12 months), with the auditor sampling evidence throughout. A Type II report is therefore not a questionnaire or self-certification: it is a sustained, evidence-based, independent examination, renewed on a recurring cycle, resulting in a formal attestation report that customers may review under NDA.

Amira's Adherence
Amira maintains SOC 2 Type II compliance, verified by independent auditors. The examination scope covers the systems that process student data — including the storage and handling of de-identified audio — and encompasses: continuous security monitoring across infrastructure and systems; strict access controls to prevent unauthorized data exposure; ongoing risk assessments and vulnerability management; and formalized policies and procedures ensuring operational integrity and transparency. Amira's DSA (Exhibit E) contractually commits Amira to maintain a cybersecurity program based on SOC 2 and the NIST family of standards, making the audit not merely a marketing credential but an enforceable obligation to every LEA. Districts may request Amira's current SOC 2 Type II report through their account representative, and the DSA's audit clause provides an independent verification path.

10.14 Conclusion: All Reasonable Steps to Protect Student Privacy

The question any district, parent, or policymaker is entitled to ask of an education technology provider is simple: are you taking all reasonable steps to protect students? For Amira, the answer is yes, and it is documented rather than asserted. Amira collects the minimum data needed to teach a child to read. It binds itself to every district through a controlling Data Sharing Agreement that makes the district the owner and Amira the school official under district direction. It complies with FERPA, COPPA, and PPRA; with the GDPR, UK GDPR, and the EU-U.S. Data Privacy Framework; and with the full lattice of state student privacy and emerging AI statutes, conforming nationally to the strictest common standard rather than the local minimum.

For the most sensitive data class — a child's voice — Amira goes further than any law yet requires: recordings are stored only in de-identified form, encrypted in transit and at rest, held on a single infrastructure subprocessor (AWS) and shared with no other, never used to build voiceprints or biometric identifiers, never exposed to public AI platforms, erasable on district request, and subject to a district's right to disable audio retention altogether. Model improvement occurs only under express district consent, only on de-identified data, and with per-district training exclusions available and applied wherever law or district policy requires.

These commitments are verified by independent SOC 2 Type II auditors on a recurring cycle, enforced by district audit rights, backstopped by 72-hour breach notification, and administered by a dedicated privacy and security team under a designated Privacy Officer. No responsible steward of children's data can promise that risk is zero; what Amira promises — and demonstrates — is that every reasonable step available to a modern education technology provider has been taken, is being taken, and will continue to be taken. Protecting the children we serve is not a compliance obligation Amira discharges; it is a condition of the trust on which our mission depends.

APPENDIX A: Advisor Information

Dr. David J. Francis

Dr. David Francis is a renowned statistical psychologist and psychometrician. He is not and has not been an employee of Amira Learning, but does serve as an advisor. At the time of this research, he had no affiliation with the company. Dr. Francis is a Hugh Roy and Lillie Cranz Cullen Distinguished University Chair and a recipient of the University of Houston Teaching Excellence Award and a former member of the National Institute of Health's Behavioral Medicine Study Section. Dr. Francis is the Director of Texas Institute for Measurement, Evaluation and Statistics. He is a Fellow of Division 5 (Measurement, Evaluation, and Statistics) of the American Psychological Association and a current member of the Independent Review Panel for the National Assessment of Title I and the Technical Advisory Group of the What Works Clearinghouse. His areas of quantitative interest include modeling of individual growth, multi-level and mixture modeling, structural equation modeling, item response theory, and exploratory data analysis. Dr. Francis currently collaborates on multiple contracts and grants funded by NICHD, the Institute of Education Sciences of the U.S. Department of Education, the National Institute of Deafness and Communication Disorders, the Texas Education Agency, and the Houston Livestock Show and Rodeo.

Science Partners for Amira ISIP



Figure A1: Amira ISIP Partners

APPENDIX B: Criteria for Item Quality Evaluation

Criteria Category	Evaluation Criteria
1. Content Validity & Alignment	
Curriculum Alignment	Item aligns with state/national standards (Common Core, IAS, etc.)
Content Accuracy	Reviewed by at least two subject-matter experts
Instructional Relevance	Aligns with real-world applications and instructional use
2. Psychometric Properties	
Difficulty Level (P-Value)	Falls within target range between .25 and .90
Item Discrimination (Point-Biserial)	Item has a discrimination index ≥ 0.30
IRT Theta Values	Items fall between -3.50 to 3.50
Reliability Contribution	Supports overall test reliability (Cronbach's Alpha ≥ 0.80)
Bias & Fairness Check	No cultural, linguistic, or socioeconomic bias detected
Differential Item Functioning (DIF)	Effect size ≤ 0.30 (items above this are flagged for review)
3. Item Format & Technical Quality	
Multiple-Choice Item Quality	Single best answer, no positive point-biserial on distractors. Plausible distractors: each incorrect option chosen by ≥ 5-10%
Universal Design for Learning (UDL)	Compatible with screen readers & alternative response formats
4. Statistical Performance in Field Testing	
Field Testing Conducted	Tested with at least 1,000 students
5. Practical Considerations	
Automated Scoring Accuracy	Constructed-response items achieve ≥ 90% agreement with human raters

Criteria Category	Evaluation Criteria
Test Security & Exposure	Item is not overexposed (used in ≤ 25% of test forms)
Review & Approval	Reviewed by at least three experts (content specialists, psychometricians, educators)

APPENDIX C: Amira ISIP Tasks & Expected Times

Amira ISIP can be administered completely in approximately 20 minutes, depending on which specific tasks are included at which grade levels. The following table summarizes expected time on task by type of task by grade level.

Tasks	Pre-K	K	1	2	3	4	5	6	7	8	Time
Phonological Awareness (Segmentation, Blending, Deletion, Substitution) * we recommend configuring at least 2 of the PA subtasks for a valid subscore	✔	✔	✔	✔	✔	o	o	o	o	o	4 min
Phonological Working Memory	✔	✔	✔	o	o	o	o	o	o	o	1-2 min
Letter Name Identification (LNF)	✔	✔	✔	✔	o	o	o	o	o	o	1 min
Letter Sound Identification (LSF)	✔	✔	✔	✔	o	o	o	o	o	o	1 min
Listening Comprehension / Retell	✔	✔	✔	o	o	o	o	o	o	o	3 min
Expressive Vocabulary	✔	✔	✔	o	o	o	o	o	o	o	
Pseudoword Identification (NWF)	o	✔	✔	✔	✔	✔	✔	o	o	o	1 min
Word Identification (WIF)	o	✔	✔	✔	✔	✔	✔	✔	o	o	1 min
Spelling/Encoding	o	✔	✔	✔	✔	✔	✔	✔	✔	✔	2-4 min
Oral Reading Fluency (ORF)	o	✔	✔	✔	✔	✔	✔	✔	✔	✔	2-4 min
Reading Comprehension	o	✔	✔	✔	✔	✔	✔	✔	✔	✔	2 min
Receptive Vocabulary	o	✔	✔	✔	✔	✔	✔	✔	✔	✔	1-2 min
Structures & Reasoning	o	o	o	o	✔	o	o	o	o	o	2 min
Rapid Automatized Naming (RAN)	✔	✔	✔	✔	✔	✔	✔				1 min
Visual Attention	o	✔	✔	✔							1 min
Approx Times (in minutes)	**10 - 12**	**20-25**	**20-25**	**17 -23**	**15-20**	**11-16**	**11-16**	**7-12**	**6-11**	**6-11**	

✔ = task is fully supported

o = task can be added, subject to adaptivity

APPENDIX D: SODA Progress Monitoring

Overview
This section documents the psychometric analyses conducted for each **Subtest on Demand Assessment (SODA)** progress monitoring (PM) constructs. SODAs are specifically constructed to support PM activities that are targeted to specific reading skills and require much less administration time than administering the entire assessment battery. SODAs are available for 7 constructs (described below) and can be assigned across grade levels for students who need practice on those skills. The subscores produced by SODA mirror the reporting metrics available for assessment scores (see Technical Guide, Section 5).

Four types of reliability evidence are reported, each targeting a different source of score consistency: **internal consistency** (Cronbach's α), **marginal IRT reliability**, **test-retest reliability**, and **OLS growth slope reliability**. Validity evidence is provided in the form of the area under the receiver operating characteristic curve (AUC), using the Amira Reading Metric (ARM) Percentile Rank as an external criterion.

All analyses use SY2025 (academic year 2025–2026) SODA administration records. Only fully completed assessments are included. Item response theory (IRT) calibration uses a two-parameter logistic (2-PL) model fit to pooled item response data across all forms within each SODA construct, using a custom Bock-Aitkin EM algorithm with Bayesian MAP regularization.

Reliability and validity statistics are reported separately for each SODA construct and grade level (Kindergarten through Grade 6). Cells meeting fewer than the minimum sample size or precision requirements for a given analysis type are suppressed and indicated with a dash (—).

Construct Consolidation: 10 Task Types To 7 Sodas
SODA items are classified by task type, which are (in some cases) further consolidated into reportable scores. The consolidation reflects both psychometric and construct-validity considerations. Task types sharing a common underlying skill were concurrently calibrated on the same IRT vertical scale and jointly normed, ensuring that reported SODA subscores represent coherent reading competencies. Concurrent calibration also pools measurement evidence across constituent task types, yielding a more precise and stable construct estimate per assessment occasion than either task type alone would provide.

The Amira ISIP SODAs include 10 task types. Following psychometric review, several task types are co-administered and jointly calibrated as combined subscores. The final 7 SODA constructs and their constituent task types are: (1) Oral Reading Fluency (ORF / Text Fluency) — assessment task type; (2) Letter Knowledge — letterName + letterSound task types; (3) Phonological Awareness — blending + Up and Down (phoneme segmentation) task types; (4) Decoding — NWF (Nonsense Word Fluency, sometimes called "Pseudo Word

Fluency") + WIF (Word Identification Fluency) task types; (5) Reading Comprehension; (6) Spelling; and (7) Vocabulary.

For constructs that tend to appear only at certain grade levels (e.g., Letter Knowledge is a foundational skill that tends only to appear at grades K-2; see Technical Guide Section 2 and Appendix C), SODA content for off-grade administration is available. Students rostered above-grade level (e.g., a 3rd grade student who needs practice with letter sounds) will initially receive the highest level content available. Conversely, even though many Kindergarten students are pre-readers (especially early in the year), Amira has created and validated appropriate content at that reading level, with adaptively more difficult passages available for those students who demonstrate high performance at lower levels.

Equivalent/Adaptive Forms
All Amira ISIP SODA items are calibrated via 2-PL Bayesian MAP IRT. Item parameters were placed on a common vertical scale spanning grades K–6 by including overlapping anchor items across adjacent grade bands. Because all items share a common metric, scores remain statistically comparable regardless of which items a student receives. The item pool enables 20+ equivalent forms per grade per construct, enabling frequent progress monitoring without practice effects. Item exposure controls prevent repeated presentation of the same items to the same student across PM occasions. Grade coverage varies by construct; while all SODAs were administered across grades K-6, some reliability estimation procedures do not have adequate usage data to reflect meaningful coefficients. These instances are noted in the following tables of evidence.

Form Equivalency Methodology
Statistical equivalence is guaranteed by concurrent calibration on a common vertical theta scale: a change of one logit represents the same magnitude of growth regardless of which items a student receives. The SODA forms are administered as a stage-adaptive assessment; chosen based on performance to optimize each student's experience, aligning student ability to form difficulty. Every form is individually optimized for precision. Observed score differences across PM occasions therefore reflect true growth rather than form difficulty, supporting valid inference about individual learning trajectories.

Form equivalency is achieved through IRT calibration. All items within each SODA construct are calibrated on a common vertical scale. For each PM occasion, forms are selected based on the student's estimated ability (EAP theta), maximizing precision at that ability level. Because all items share a common metric, forms assembled from different item subsets are statistically equivalent with respect to measurement precision. See Technical Guide Sections 4.1–4.4.

Administration Format And Time
All SODA constructs are administered individually or in groups via computer. Average administration time is less than 5 minutes per construct per assessment occasion. See Technical Guide Appendix C (Amira ISIP Task and Time) for task-by-task time estimates.

Progress Monitoring Frequency
All SODA constructs have been developed and validated for biweekly administration (every two weeks). Monthly administration is also supported. The biweekly cadence provides sufficient data points to estimate reliable growth slopes and detect response to instruction in a timely manner. See Technical Guide Section 3.8 (Administration).
Reliability (Cronbach's Alpha By Soda × Grade)

Cronbach's coefficient α quantifies the degree to which items within a single administration measure the same latent construct. It is defined as:

$$\alpha = \kappa/(\kappa - 1) \times (1 - \Sigma\sigma^2{}_j / \sigma^2 total)$$

where κ is the number of items, σ^2 is the observed variance of item j across students, and $\sigma^2_o{}_a$ is the variance of students' total (sum) scores on the form.

Alpha is computed separately for each form using only students who provided a response to every item in the form. A minimum of 10 students is required for a form-level estimate to be retained. Form-level estimates are then aggregated to the **construct × grade** level using a student-count-weighted mean across all qualifying forms administered to students at that grade.

The consistently high alpha values confirm that items within each SODA construct measure a coherent underlying trait. For Oral Reading Fluency (ORF), each word in the reading passage is treated as a binary item (1 = read correctly, 0 = error), yielding 43–344 items per passage depending on passage length. Because alpha is a direct function of the number of items, ORF alpha values are consistently higher than those of shorter subtests and increase with grade level as longer, more demanding passages are administered.

All alpha values are computed using **grade-specific item responses only**: for each grade's estimate, only responses from students enrolled at that grade contribute to the item variance-covariance matrix. This approach yields estimates that reflect the measurement properties of the instrument as it is actually used with students at each developmental stage.

Cronbach's alpha (weighted-average) across SODA forms by grade.

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.77	0.74	0.77	0.69	0.72	0.70	0.73
N students	12,955	27,248	25,253	18,460	10,031	5,442	750
Letter Knowledge	0.78	0.81	0.82	0.82	0.82	0.81	0.85
N students	19,619	16,983	9,077	5,640	2,428	960	185
ORF	0.91	0.95	0.94	0.93	0.93	0.94	0.98
N students	307	3,690	9,644	7,031	6,860	4,743	489
Phonological Awareness	0.77	0.74	0.74	0.73	0.76	0.68	0.77
N students	19,419	25,275	18,700	12,883	5,984	2,692	403
Reading Comprehension	0.63	0.67	0.62	0.60	0.49	0.69	0.60
N students	928	5,184	11,254	12,005	9,415	5,857	1,202
Spelling	0.84	0.78	0.83	0.80	0.81	0.74	0.81
N students	1,481	4,782	6,520	4,908	3,291	2,125	533
Vocabulary	0.60	0.68	0.65	0.64	0.68	0.54	0.67
N students	2,530	5,144	7,422	6,933	5,617	3,670	813

Marginal IRT Reliability

Marginal reliability expresses the proportion of observed score variance attributable to true individual differences in ability, averaged over the empirical distribution of students at each grade. It is computed from theta estimates and their posterior standard errors using:

$$\rho marginal = Var(\theta) / [Var(\theta) + Mean(SEM^2)]$$

where *Var(θ)* is the empirical variance of EAP theta estimates across students within a construct × grade cell, and *Mean(SEM²)* is the mean squared posterior standard deviation (measurement error variance) across the same students.

Theta estimates (θ) and their standard errors (SEM) are obtained via **Expected A Posteriori (EAP) scoring**. For each student, the posterior distribution of θ given the observed item responses *u* and the calibrated 2-PL item parameters (a_i, b_i) is computed over a 61-point Gauss–Hermite quadrature grid [*-4, 4*], using a standard normal prior N(0, 1):

$$P(\theta \mid u, a, b) \propto P(u \mid \theta, a, b) \cdot \varphi(\theta)$$

$$\theta = \Sigma_k \theta_k \cdot P(\theta_k \mid u) \quad | \quad SEM = \sqrt{[\Sigma_k (\theta_k - \theta)^2 \cdot P(\theta_k \mid u)]}$$

Marginal IRT Reliability across SODA Forms by Grade

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.86	0.87	0.86	0.85	0.89	0.87	0.92
N sessions	3,168	6,480	5,117	3,312	2,202	1,287	205
Letter Knowledge	0.89	0.91	0.93	0.94	0.95	0.96	0.96
N sessions	3,620	2,860	1,656	936	700	317	136
ORF	—	—	—	—	—	—	—
N sessions	—	—	—	—	—	—	—
Phonological Awareness	0.82	0.83	0.84	0.85	0.90	0.89	0.90
N sessions	3,668	5,024	3,666	2,198	1,390	668	194
Reading Comp	0.85	0.81	0.79	0.80	0.78	0.81	0.78
N sessions	626	1,954	2,899	2,345	2,269	1,202	337
Spelling	0.90	0.88	0.82	0.90	0.90	0.88	0.90
N sessions	415	1,621	1,552	909	849	525	196
Vocabulary	0.82	0.86	0.80	0.84	0.86	0.80	0.90
N sessions	584	1,456	1,650	1,307	1,470	741	199

Note: ORF results are not available because the scaling metric is equated WCPM, not IRT

Test-Retest Reliability

Test-retest reliability estimates the stability of SODA scores over short time intervals during which the underlying construct is not expected to change appreciably. It is estimated as the Pearson product-moment correlation between two scores obtained by the same student on the same construct within a narrow time window:

$$r_{tt} = Cov(\hat{\theta_1}, \hat{\theta_2}) / [SD(\hat{\theta_1}) \cdot SD(\hat{\theta_2})]$$

where $\hat{\theta_1}$ and $\hat{\theta_2}$ are EAP theta estimates (or Equated WCPM for ORF) from the two administrations. Because all forms within a construct share a common calibration, students need not have taken the same specific form at both time points—theta estimates are directly comparable across forms.

The test-retest window is set at **4-14 calendar days**, thus pairs of observations were included if they were within 4-14 days of each other. Observation pairs occurring 3 or fewer days apart were rare and indicated that a student didn't complete the session and needed

to retest to get valid scores. The maximum time between sessions was capped at 14 days so that construct stability, rather than growth, would be captured. Calculations for test-retest reliability were only conducted if 10 or more students had valid pairs of observations meeting the inclusion criteria.

This interval is short enough that true growth in reading ability is unlikely to produce a meaningful change in the score. When a student has multiple qualifying pairs within this window, the pair with the smallest calendar gap is retained, and each student contributes at most one pair to the cell-level correlation.

Test-Retest Reliability (weighted-average) across SODA forms by Grade.

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.61	0.69	0.61	0.77	0.67	0.78	—
N pairs	79	323	261	196	103	51	—
Letter Knowledge	0.61	0.64	0.58	0.79	0.38	—	—
N pairs	95	84	40	18	15	—	—
ORF	0.66	0.79	0.78	0.78	0.82	0.82	0.80
N pairs	1,177	1,839	1,936	1,753	1,023	709	153
Phonological Awareness	0.52	0.63	0.57	0.69	0.57	0.72	—
N pairs	117	160	169	101	77	16	—
Reading Comprehension	0.80	0.64	0.73	0.70	0.49	0.62	0.77
N pairs	19	149	228	187	175	85	43
Spelling	0.75	0.81	0.75	0.87	0.92	0.70	0.77
N pairs	18	66	109	40	21	23	16
Vocabulary	0.60	0.78	0.66	0.76	0.68	0.58	0.35
N pairs	35	62	101	81	128	35	24

Slope Reliability

Slope reliability quantifies the degree to which differences among students in their observed growth trajectories reflect true inter-individual differences in growth rate, as opposed to measurement error. For each qualifying student, an ordinary least-squares (OLS) regression of theta on time is estimated:

$$\theta_{it} = \alpha_i + \beta_i \cdot t + \varepsilon_{it}$$

where t is the calendar day of each assessment relative to the student's first observation, β_i is the estimated growth rate (theta units per day) for student i, and ε_i is the residual. The OLS standard error of the slope estimate is:

$$SE(\beta_i) = \sqrt{[MSE_i / SS_{ti}]}$$

here $MSE_i = \Sigma(\hat{\varepsilon}_{it})^2 / (n_i - 2)$ is the mean squared residual and $SS_{ti} = \Sigma(t - \bar{t})^2$ is the sum of squared deviations of the assessment times from their mean. Slope reliability at the construct × grade level is then estimated using the variance-components formula:

$$\rho slope = Var(\beta) / [Var(\beta) + Mean(SE(\beta)^2)]$$

where $Var(\beta)$ is the empirical variance of individual slope estimates across students and $Mean(SE(\beta)^2)$ is the mean squared standard error of those estimates.

Slope reliability estimation used only observations that were at least 15 days apart (after culling any tests that occurred 3 days or fewer before the final (complete) test). That is, repeated observations were thinned so that consecutive retained observations are separated by at least a 15-day minimum gap. This prevents artificially high within-student autocorrelation from same-week administrations from inflating slope precision. The minimum required span of observations for progress monitoring slope reliability was 60 days, with some sessions spaced across the entire school year. Every student included in slope reliability calculations had at least four observations over the span of administrations. Lastly, calculations for slope reliability required at least 30 qualifying students to be calculated. Dashes indicate grades below the minimum observation threshold or with insufficient sample size for stable estimation.

Slope Reliability (weighted-average) across SODA forms by Grade.

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.75	0.93	0.78	0.69	0.73	0.71	0.71
N students	52	247	85	31	85	42	366
Avg N obs	9.2	6.7	12.2	8.7	10.9	12.1	8.8
Avg N weeks	18.6	17.7	26.0	26.0	21.3	24.9	32.9
Letter Knowledge	0.67	0.84	0.72	0.78	0.67	0.63	—
N students	74	97	35	36	152	51	—
Avg N obs	12.7	6.5	8.2	8.6	4.8	4.9	—
Avg N weeks	23.1	17.1	23.7	17.0	22.1	22.3	—
ORF	0.79	0.70	0.65	0.72	0.82	0.76	0.66
N students	32	30	61	121	55	148	38
Avg N obs	6.7	9.7	6.4	12.0	7.6	6.1	5.4
Avg N weeks	28.7	21.0	24.7	25.7	22.1	16.3	18.1
Phonological Awareness	0.66	0.69	0.59	0.62	0.70	0.68	0.71
N students	111	47	32	142	45	35	353
Avg N obs	11.1	6.3	9.5	5.1	8.0	7.5	8.8
Avg N weeks	22.6	25.3	25.4	21.9	20.1	17.6	32.9
Reading Comp	0.64	0.67	0.59	0.69	0.66	0.61	0.67
N students	124	37	61	113	96	31	34
Avg N obs	6.3	5.8	11.0	10.7	12.3	8.1	5.3
Avg N weeks	29.4	17.3	24.7	25.6	23.9	25.3	16.3
Spelling	0.80	0.77	0.67	0.70	0.63	0.64	0.75
N students	95	43	63	35	32	39	374
Avg N obs	6.2	6.8	6.2	8.6	7.5	4.8	8.9
Avg N weeks	29.3	16.6	17.4	19.6	23.3	17.3	32.9
Vocabulary	0.69	0.78	0.81	0.72	0.64	0.71	0.74
N students	37	48	54	59	44	34	32
Avg N obs	8.8	5.6	6.4	12.4	8.8	8.1	5.9
Avg N weeks	31.6	16.1	18.3	23.6	24.6	24.4	25.7

Validity Evidence: Classification Accuracy

The predictive validity of SODA was examined by comparing beginning-of-the-year (BOY) SODA scores to Amira Reading Mastery (ARM) scores. Classification accuracy of SODA was defined by the efficacy of BOYscores for identifying students who scored at or below the 20th percentile on the ARM benchmark assessment in SY2025. Predictive validity was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), along with sensitivity and specificity at the optimal decision threshold.

The criterion was the ARM score percentile rank (PR) from the last benchmark assessment administered in SY2025. Students with ARM PR ≤ 20 were classified as at-risk. This criterion captures a meaningful region of reading difficulty that aligns with benchmark-based identification practices. The SODA predictor was the EAP (Expected A Posteriori) theta estimate from a pooled Bayesian MAP 2-parameter logistic (2-PL) IRT calibration (or equated WCPM for ORF).

For each grade by subtest, a ROC curve was constructed from the SODA scores and the binary ARM at-risk criterion. AUC summarizes the probability that a randomly selected at-risk student has a lower SODA score than a randomly selected not-at-risk student (AUC = 0.50 indicates chance-level discrimination; AUC = 1.00 indicates perfect separation). The cut score for each cell was determined by Youden's J index, which maximizes sensitivity + specificity − 1. Cells with fewer than 20 students were excluded and are reported as "—". All values are rounded to two decimal places. The following tables AUC, sensitivity, and specificity for all subtests and grades.

Classification Accuracy: AUC from SODA at BOY predicting ARM risk status (PR ≤ 20).

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.65	0.76	0.82	0.83	0.81	0.84	0.68
N students	3,633	6,908	5,697	3,844	2,161	1,047	204
Letter Knowledge	0.69	0.66	0.61	0.60	0.61	0.61	—
N students	5,368	5,334	2,938	1,781	825	317	—
ORF	—	0.61	0.80	0.82	0.83	0.83	0.69
N students	—	2,870	6,131	4,144	4,189	2,826	343
Phonological Awareness	0.62	0.67	0.60	0.59	0.60	0.61	0.52
N students	4,970	6,965	4,873	3,067	1,508	682	133
Reading Comprehension	0.67	0.64	0.74	0.74	0.70	0.69	0.67
N students	547	3,224	5,951	6,554	4,846	2,750	644
Spelling	0.68	0.75	0.81	0.82	0.88	0.85	0.82
N students	1,023	3,243	4,147	3,227	2,167	1,147	308
Vocabulary	0.61	0.68	0.72	0.74	0.72	0.71	0.62
N students	1,728	3,343	4,532	4,122	3,153	2,138	476

Classification Accuracy: Sensitivity = P(SODA flags risk | student at-risk).

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.70	0.79	0.79	0.75	0.80	0.76	0.69
Letter Knowledge	0.67	0.66	0.60	0.50	0.62	0.49	—
ORF	—	0.62	0.73	0.74	0.81	0.78	0.68
Phonological Awareness	0.60	0.58	0.51	0.34	0.38	0.56	0.20
Reading Comprehension	0.68	0.56	0.74	0.61	0.69	0.61	0.58
Spelling	0.75	0.63	0.79	0.83	0.77	0.88	0.81
Vocabulary	0.55	0.63	0.67	0.62	0.64	0.58	0.36

Classification Accuracy: Specificity = P(SODA does not flag | student not at-risk).

SODA	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Decoding	0.55	0.62	0.73	0.79	0.73	0.78	0.64
Letter Knowledge	0.64	0.59	0.60	0.68	0.57	0.74	—
ORF	—	0.57	0.75	0.77	0.71	0.75	0.65
Phonological Awareness	0.60	0.70	0.65	0.79	0.78	0.61	0.90
Reading Comprehension	0.59	0.68	0.66	0.79	0.61	0.69	0.69
Spelling	0.59	0.78	0.70	0.71	0.86	0.71	0.78
Vocabulary	0.64	0.66	0.71	0.78	0.75	0.76	0.83