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 2025–2026 school year. It details how the tool functions as a universal screener, benchmark assessment, and progress monitoring system—leveraging AI-powered speech recognition and adaptive testing to assess early literacy skills, detect dyslexia risk, and support differentiated instruction. Grounded in leading research frameworks, the guide explains the constructs measured, scoring methodology, and technical standards ensuring validity, reliability, and equity across diverse student populations.

DATE

2025-2026 School Year

AMIRA
LEARNING

Table of Contents

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, incorporating them to identify signals and markers of reading struggle, particularly aligning with grade-specific screening recommendations for kindergarten, first grade, and higher grades. 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 solely focused 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 Multiple Deficit Model (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 and speech recognition technology, including

Reading Error Detection (RED) Models, Multimodal Learning Analytics, Natural Language Processing (NLP), and Machine Learning for Adaptive Assessment. 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 here for a fuller description: Amira ISIP Theoretical Framework.

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.

1. Universal Screener
Amira ISIP's universal screener is designed as an online computer-based solution primarily for Kindergarten to Grade 3 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 the year assessment and is completely configurable at the LEA or SEA level. This initial assessment differs from other instances of the assessment in that inclusion of the Rapid Automatized Naming (RAN) task is highly encouraged in order to produce the most valid and accurate classification outcomes. This screener integrates early literacy assessment with dyslexia screening, aligning with state requirements.

2. 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 Reading Error Detection (RED), multimodal learning analytics (audio analysis, eye-tracking, keystroke data), Natural Language Processing (NLP) for comprehension, and

Machine Learning for Adaptive Assessment (CAT). 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, aiming for a median administration time of 15-18 minutes, to minimize test-taker fatigue and maximize 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 for equitable access.

3. Progress Monitoring

Amira ISIP's progress monitoring tools are integrated within 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, where 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, where teachers simply assign students to the automated tutor system. In all cases, Amira generates comprehensive data following each administration, including overall scores and subscores commensurate with the assessment configuration.

2. Constructs Measured

The following sections provide descriptions of the Amira ISIP Assessment design as well as links and screenshots to illustrate Amira ISIP in action.

2.1 Phonological Awareness

Amira ISIP uses several different tasks to assess Phonological and Phonemic awareness. These specific tasks were selected based on research evidence of efficacy in predicting dyslexia, as well as success of task administration and scoring within the Amira ISIP screening context.

2.1.1 Phoneme blending

In this task, spoken words are presented as sequences of individual phonemes. The student must blend the provided phonemes together into the full word. The task begins with Amira explaining the student will hear the individual sounds that make up a word. The student is then prompted to blend these sounds seamlessly into a word typically mastered at the student's grade level.

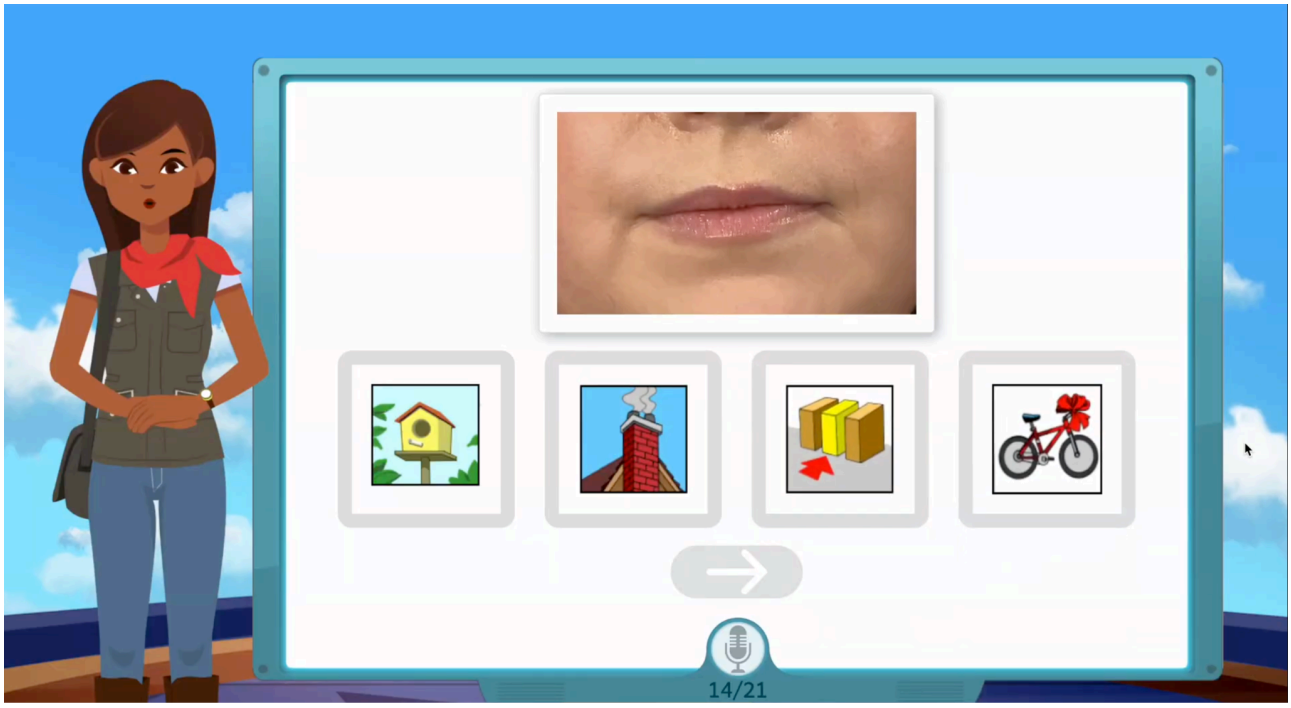


Figure 2.1: Screenshot of the Kindergarten Phoneme Blending Task

The phoneme blending is structured as follows:

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 phonemes: “/b/ /ī/ /s/ /ə/ /k/ /əl/.”
4. The student blends the phonemes to form the word bicycle and selects the corresponding image.
5. The responses are scored using Amira’s machine learning models.

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.

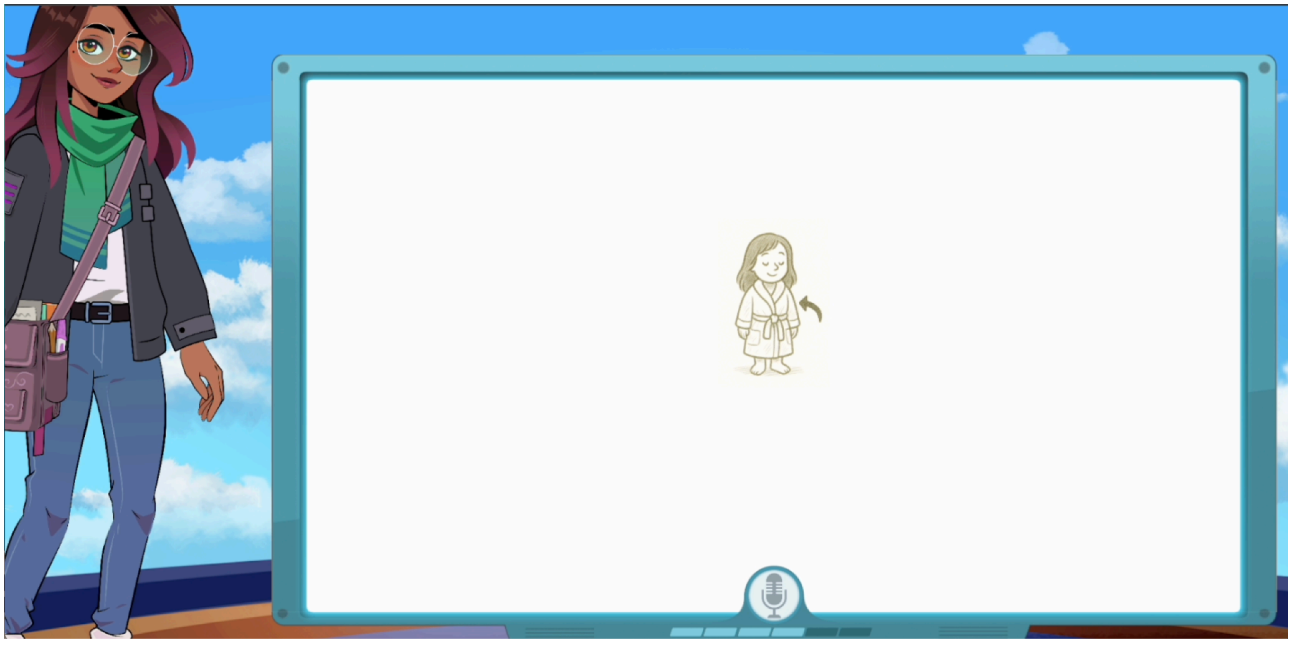


Figure 2.2: Screenshot of the Phoneme Segmentation Task

The single phoneme segmentation task is structured as follows:

1. Amira provides directions to the student 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.

See an example of the segmentation task here.

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 from a word. For half of the words, the deletion occurs at the beginning of the word, and for the other half of the words, the deletion occurs at the end of the word.

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 say the word.
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 like dyslexia often exhibit deficits in working memory, which can impact 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 are predictive of later reading difficulties.

Amira ISIP utilizes a Pseudo-word (Non-word) Repetition task to assess phonological working memory. In this task, a video is shown of an adult vocalizing a sequence of syllables that string together to produce 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).

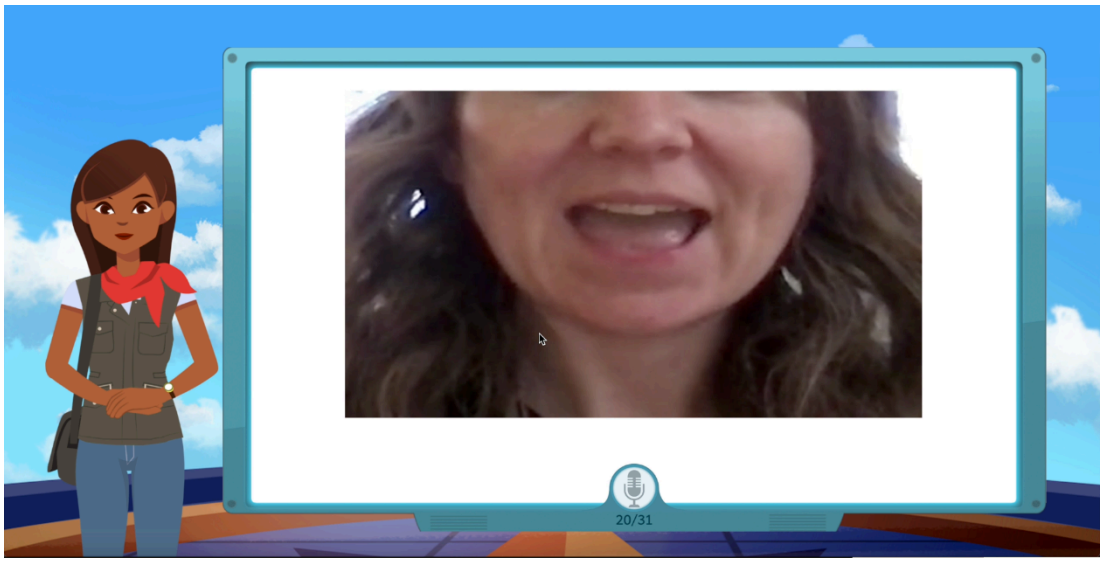


Figure 2.3: Screenshot of the Phonological Working Memory Task

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 the pseudo-word syllable sequence.
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)

This task requires students to be able to perform a phoneme substitution to 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 which phoneme it should be changed to. Then, the student is asked to say the resulting word after the phoneme manipulation.

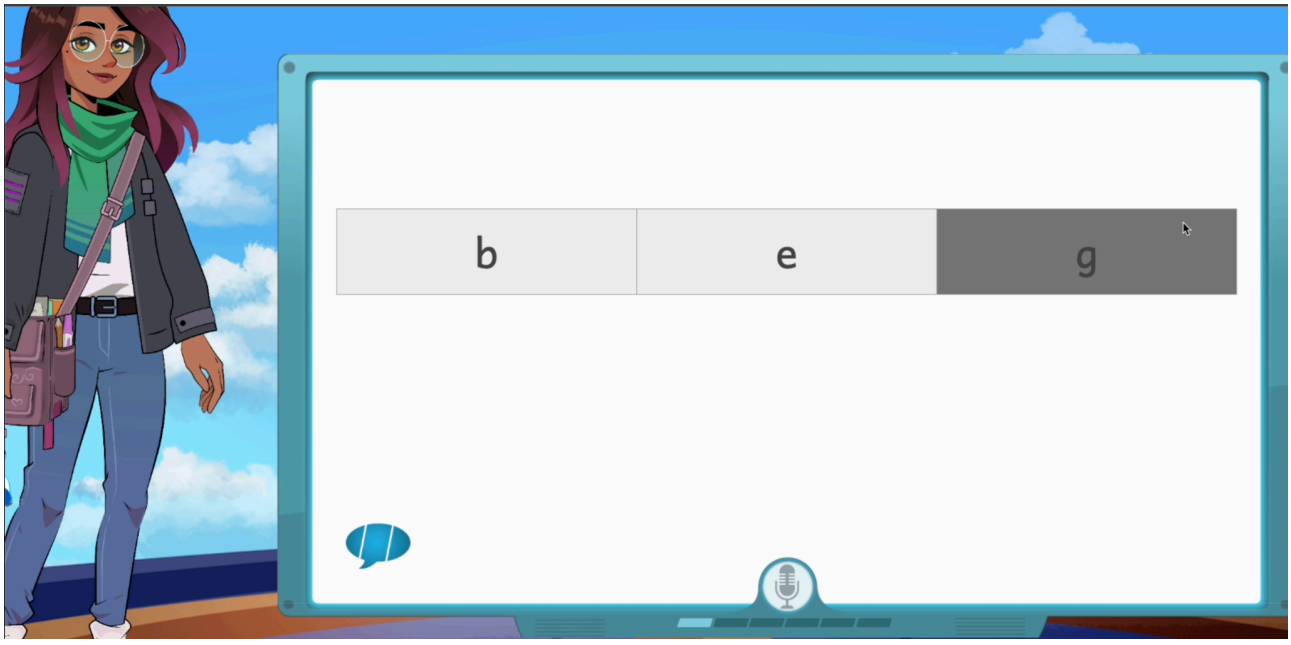


Figure 2.4: Screenshot of the Phoneme Manipulation

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 show up 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 shows the letters of the alphabet in text form on the screen, one at a time, and requires students to verbally name the letters within a certain time window per letter.

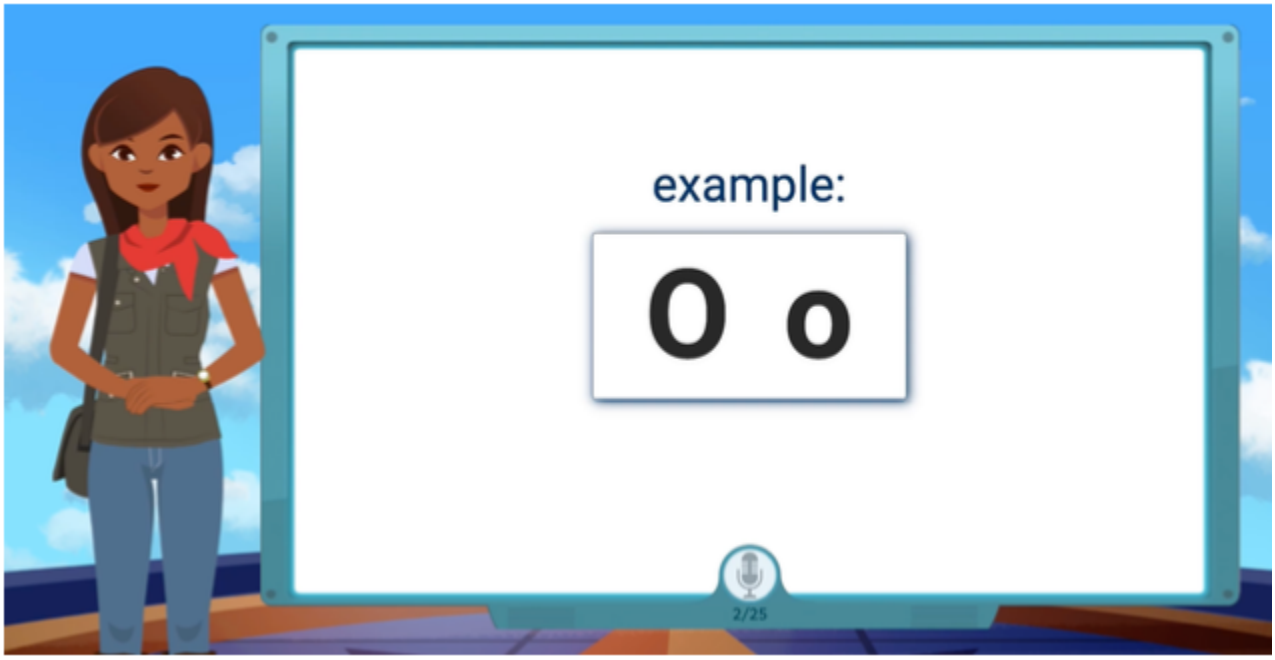


Figure 2.5: Screenshot of the Warm-up Example for the Letter Name Fluency

This task utilizes Amira's ability to listen to speech, enabling the software to emulate the typical approaches that teachers use to assess alphabetic knowledge mastery. 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 in text form.
4. The student is asked to say the name of the letter shown on the screen.

5. The student has a configured interval of time in which they are given 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 shows the letters of the alphabet in text form on the screen, one at a time, and requires students to produce the sound that the letter makes within a certain time window per letter.

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 upper and lower case 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.

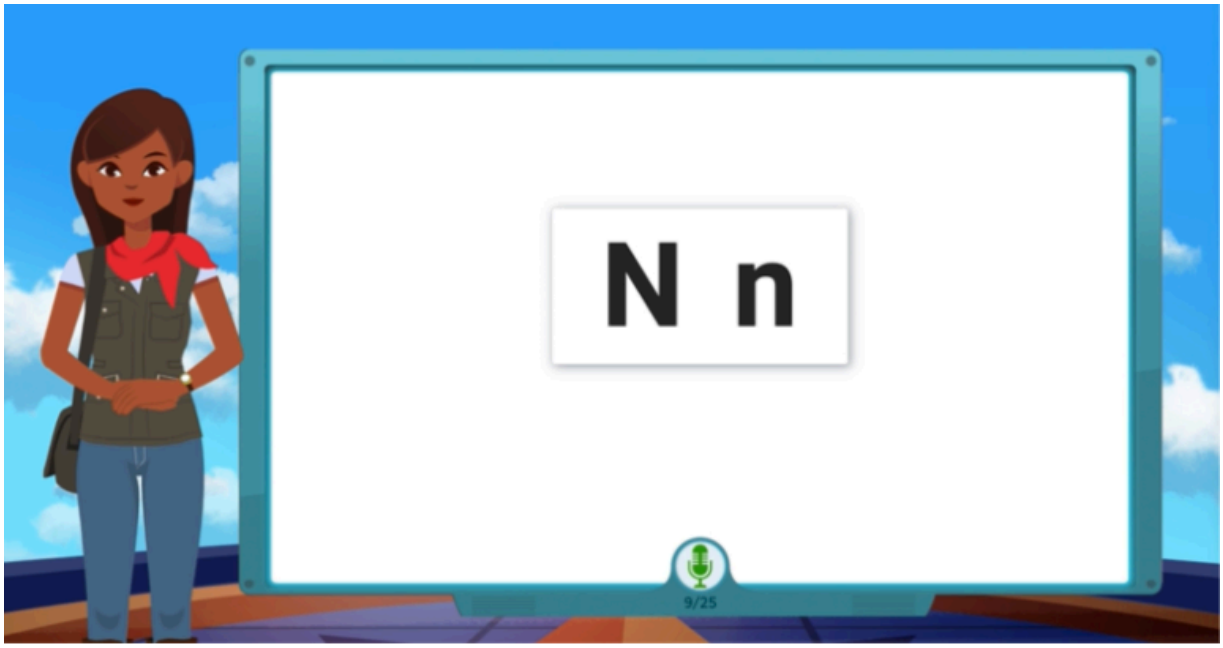


Figure 2.6: Screenshot of the Letter Sound Task

A student is typically presented with six to ten items in this 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/ELL 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 mono-syllabic. Words used in the task conform to standard and typical patterns within 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 of time to articulate 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

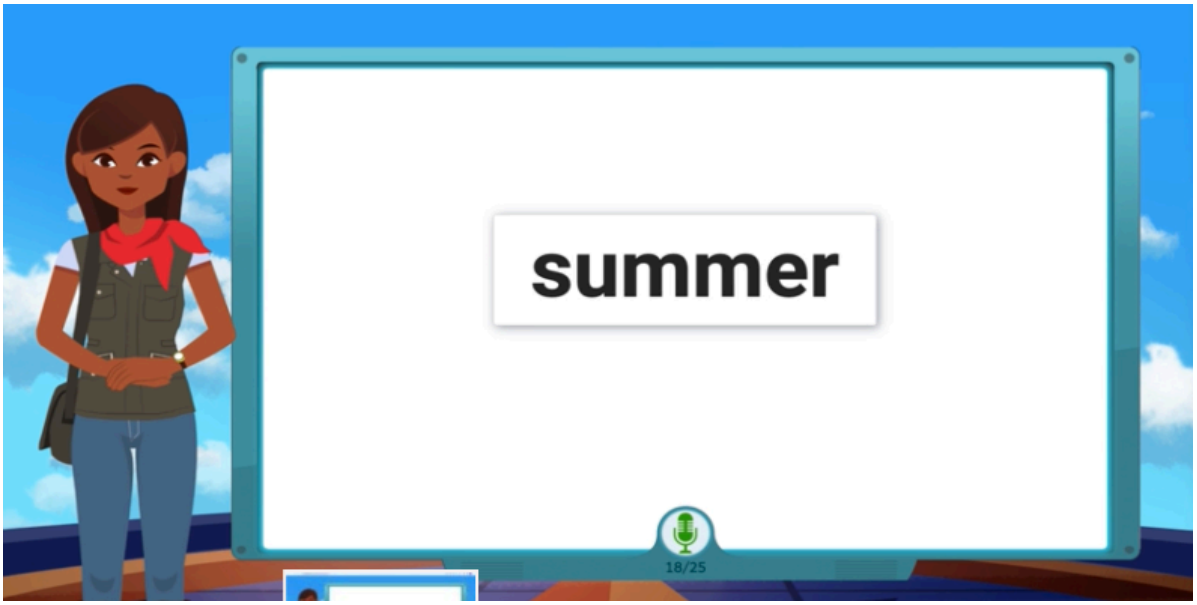


Figure 2.8: Screenshot of Word Identification Fluency task for Grade 2

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 of time to read the complete word.
6. Amira scores the item.

The Word Identification Fluency task is supported for Grades K to 6. The number of items presented varies from 4-20, 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. Typically, Amira provides enough text for the student to read for 90 seconds to 4 minutes. Once the student finishes a section, Amira allows them to move on to the next block of text.

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.

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 Oral Reading Fluency (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 is rooted in its detailed assessment of critical reading constructs , as outlined in the "reading rope" model, which includes word recognition, decoding, and comprehension.

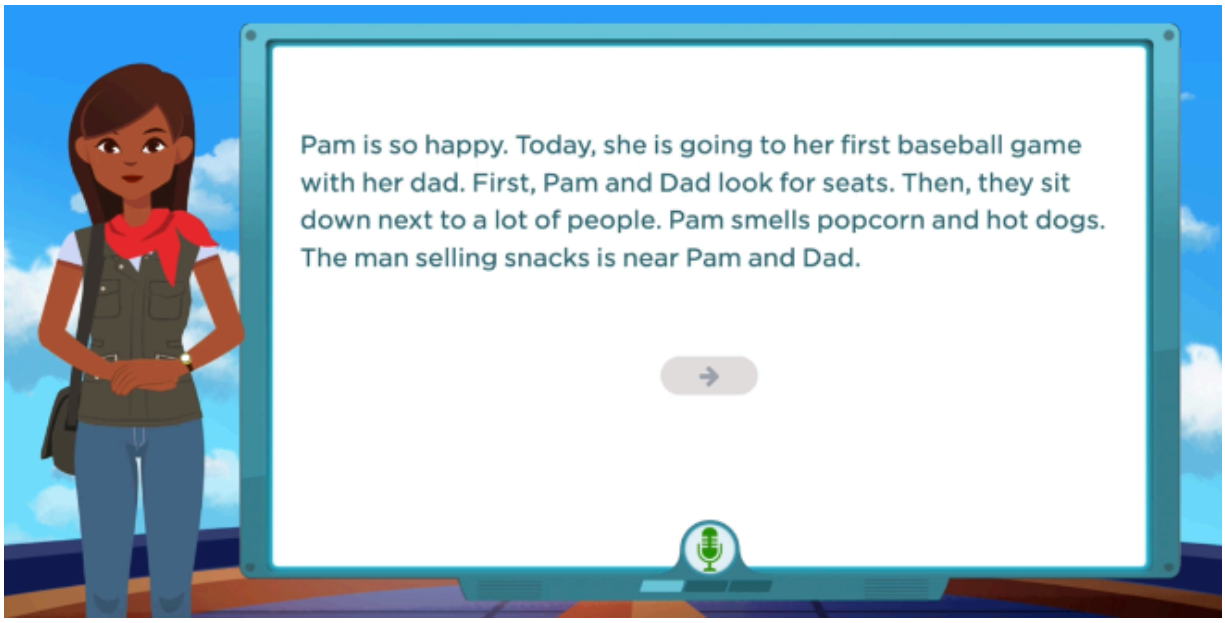


Figure 2.9: Screenshot of the Oral Reading Fluency Task

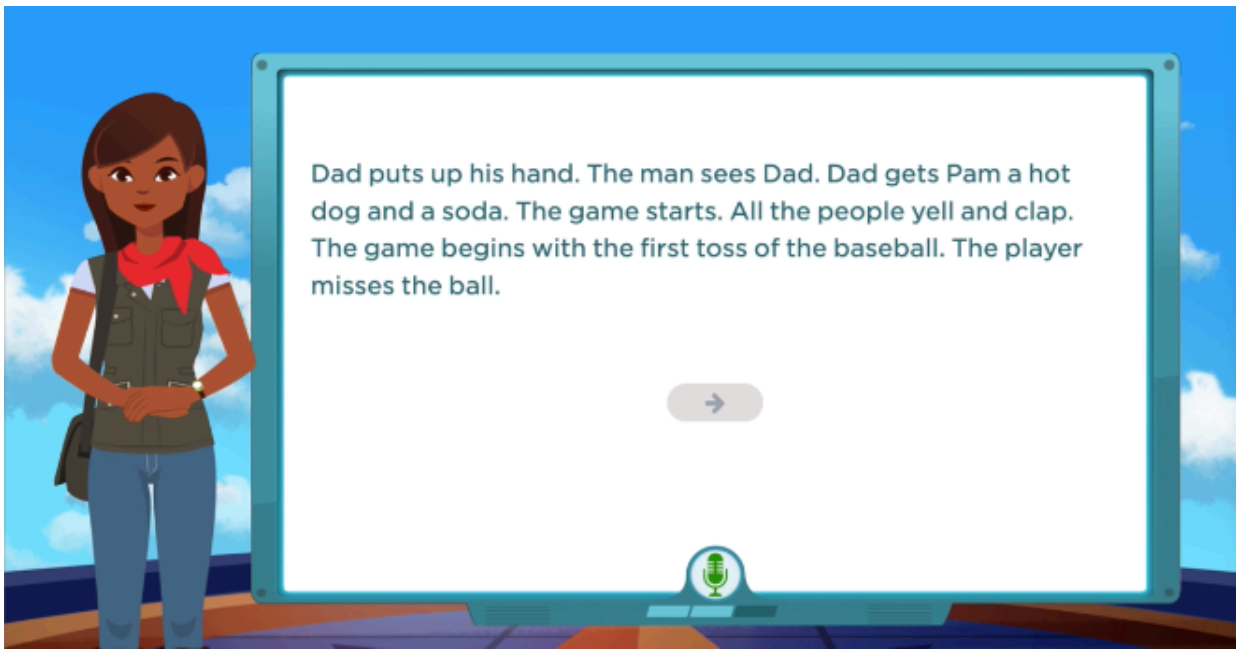


Figure 2.10: Screenshot of the Oral Reading Fluency Task

When a student completes a block within the passage, Amira enables the student to move on 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.

While rate (speed) and accuracy (correct) are calculated as expected, the calculation of prosody is as follows.

Prosody
Amira ISIP assesses prosody using a research-based algorithm that leverages Amira ISIP's AI and ORF capabilities. Amira ISIP generates a prosody score based on the student's oral reading fluency subtest. There is no need for an additional assessment task – the ORF generates all of the inputs required for a sophisticated, SoR-grounded measure of Prosody. The Prosody score supplements Amira ISIP's WCPM, Accuracy and Reading Speed metrics which are also derived from the oral reading assessment.

Amira ISIP's Prosody Score uses the four classic research-based measures in combination. Prosody is a function of smoothness, meaning-driven pitch variation, appropriate pausing and expression. Amira ISIP measures oral reading in all 4 of these dimensions.

Smoothness Rating (SR):
Derived from human ratings or software like Praat. Higher ratings reflect fewer hesitations and smoother reading.

1. Pitch Variation Index (PVI): Quantifies variation in pitch, critical for measuring expressiveness.
2. Pauses per Minute (PPM): Frequent pauses disrupt fluency. Normalize this by expressing pauses relative to time.
3. Expression Inconsistency (EI): Penalizes inconsistent prosody. For example, abrupt shifts in volume or stress lower this score.

The quantitative approach to measuring prosody, including components like pitch variation, pauses, and smoothness, is backed by research in linguistics, speech processing, and reading fluency.

Prosody Quantification Formula
Amira ISIP's calculated Prosody Score (PS) based on measurable components of reading fluency:

PS=(SR/5+PVI+PPM−EI)

Where:

1. SR = Smoothness Rating (scale: 1–10)
2. PVI = Pitch Variation Index (scale: 0–1)
    a. Reflects changes in pitch or intonation across utterances.
    b. Calculated as: PVI=Sum of pitch changesNumber of utterancesPVI = \frac{\text{Sum of pitch changes}}{\text{Number of utterances}}PVI=Number of utterancesSum of pitch changes.

Pitch variation is

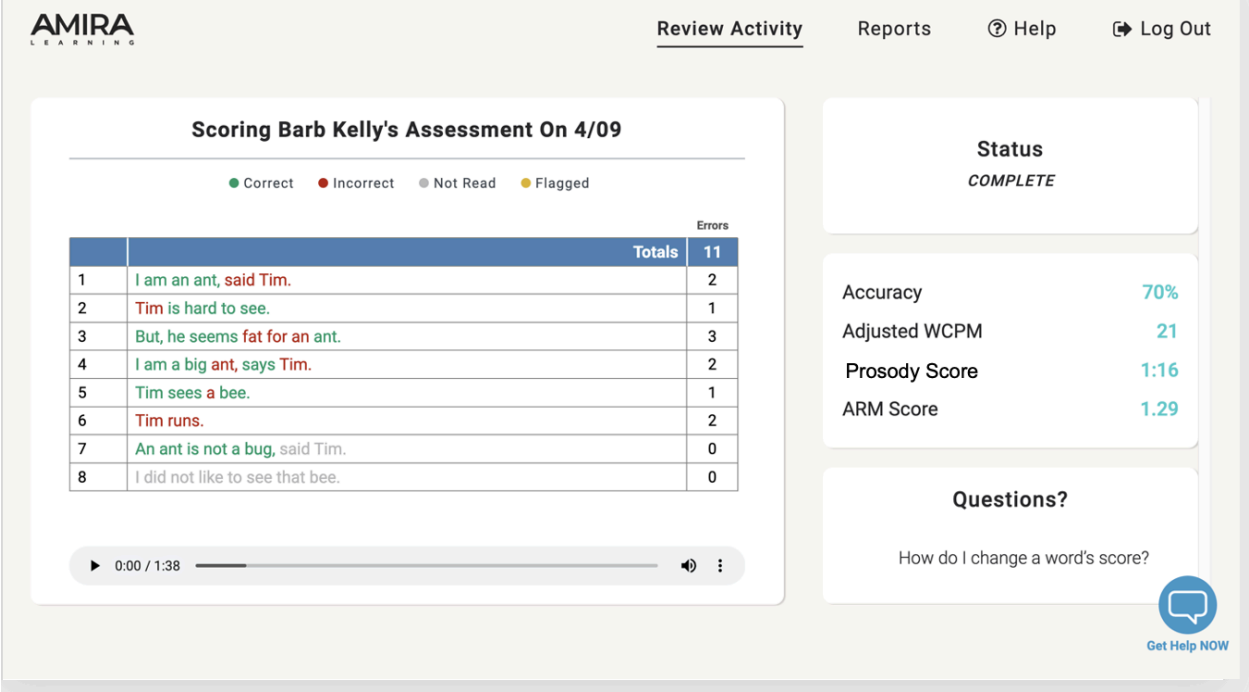
1. **PR** = (Sessions/Pauses) * 5
    a. Measures the number of pauses per minute. Fewer pauses indicate better fluency.
    b. Calculated as: PPM=60×Number of pausesTotal reading time (seconds)PPM = 60 \times \frac{\text{Number of pauses}}{\text{Total reading time (seconds)}}PPM=60×Total reading time (seconds)Number of pauses
2. **EI** = Expression Inconsistency (scale: 0–1)

This is a measure of latency derived from the student's reading. The metric is derived from the "lapse" time where there is no audible articulation occurring during the reading process. Extremes in either direction from the centroid of the distribution signal prosody issues. A student at the mean of the distribution is 1 and a student 1 standard deviation or more from the 50th PR is 0.

All four measures are automatically derived and scored by the AI as an interpretation of the digital recording of the child's reading out loud.

Display Of Prosody

Amira provides the teacher with prosody information on the Review Screen as shown below:



Scores range from 0.00 to 4.99, with students demonstrating the highest level of prosody at the top of the scale.

Below are key studies and their contributions to these metrics:

1. Pitch Variation and Expressiveness

- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000).
  "Feeltrace: An instrument for recording perceived emotion in real time."
  In this study, pitch variation was linked to emotional expressiveness and fluency in speech. While primarily used in emotion detection, their methods demonstrate that pitch variation is measurable and correlates with interpretive fluency in reading.
- Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010).
  "Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Comprehension."
  The authors emphasize that intonation and pitch modulation are critical

components of prosody and can be objectively measured for fluency assessments.

2. Pauses and Reading Fluency

- Benjamin, R. G., & Schwanenflugel, P. J. (2010).
  "Text complexity and oral reading prosody in young readers."
  This research highlights the relationship between pauses during reading and prosody, demonstrating that fluent readers have fewer and shorter pauses. Pauses per minute (PPM) serves as a reliable measure of prosodic fluency.
- Rasinski, T. V. (2004).
  "Assessing Reading Fluency."
  Rasinski discusses the role of pausing and phrasing in prosody and links these elements to comprehension and overall fluency. Though the paper is more qualitative, it provides foundational evidence for using pauses as a prosodic indicator.

3. Smoothness and Prosody

- Rasinski, T., Rikli, A., & Johnston, S. (2009).
  "Reading fluency: More than automaticity? More than a concern for the primary grades?"
  This study emphasizes smoothness (the absence of hesitations or disruptions) as an essential component of prosody, supporting its use as a measurable element in fluency assessment.
- Wood, C. (2006).
  "Metrical stress sensitivity in young children and its relationship to phonological awareness and reading."
  This research links stress patterns (a key component of smoothness and phrasing) to reading development, reinforcing the role of prosodic smoothness in reading fluency.

4. Quantitative Prosody Measurements in Speech Analysis

- Shriberg, E. (2001).
  "To 'Errrr' is Human: Ecology and Acoustics of Speech Disfluencies."
  This foundational study on disfluencies in speech links acoustic measures like pitch, pauses, and smoothness to expressive reading and oral fluency, providing a basis for quantitative metrics.
- Breen, M., Kaswer, L., Van Dyke, J. A., Krivokapic, J., & Landi, N. (2016).
  "I know what you're reading: Prosodic cues to syntactic processing."
  This study highlights how prosodic elements like pitch variation and pausing

aid in syntactic and semantic processing, validating their use in assessing reading prosody.

2.6 Vocabulary

In this task, Amira presents a word and asks the student to choose which word *goes best with* the target word from an array of three options. Amira reads the target word out loud and can read each of the multiple-choice options out loud on mouse-over, avoiding the need for students to be able to read the words in order to complete the task. The goal of this task is to measure on-grade vocabulary skills, with each item chosen to represent a class of words that should be in the vocabulary of learners progressing at the state's expected pace.This task is supported in Grades K-6.
See an example video of the vocabulary task here.

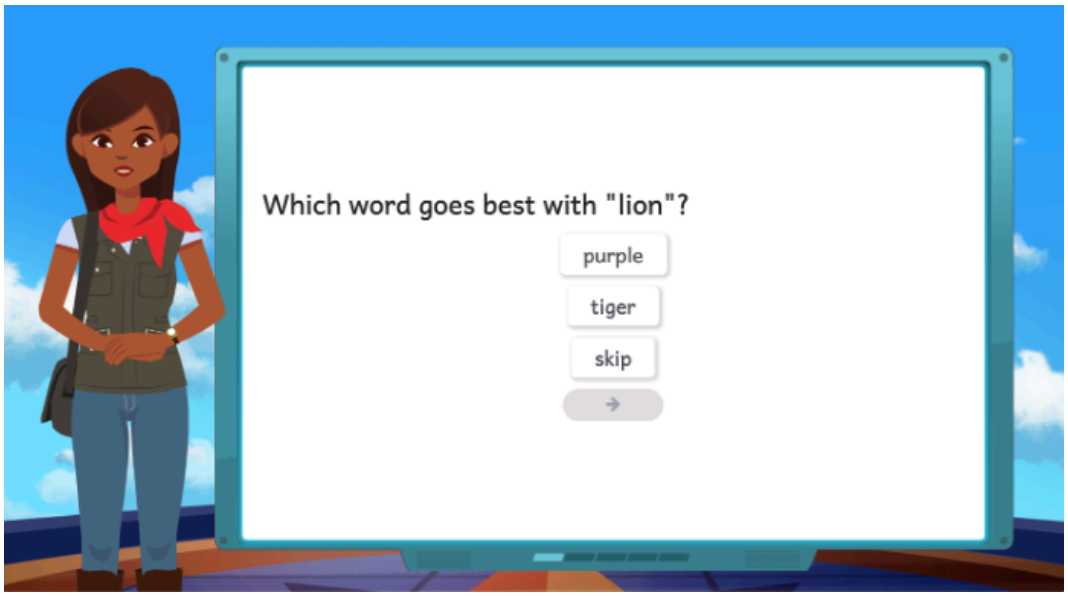


Figure 2.11: Screenshot of the Vocabulary Task - Words

Amira also supports a configurable version of the vocabulary task whereby the choices are presented in the form of 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.

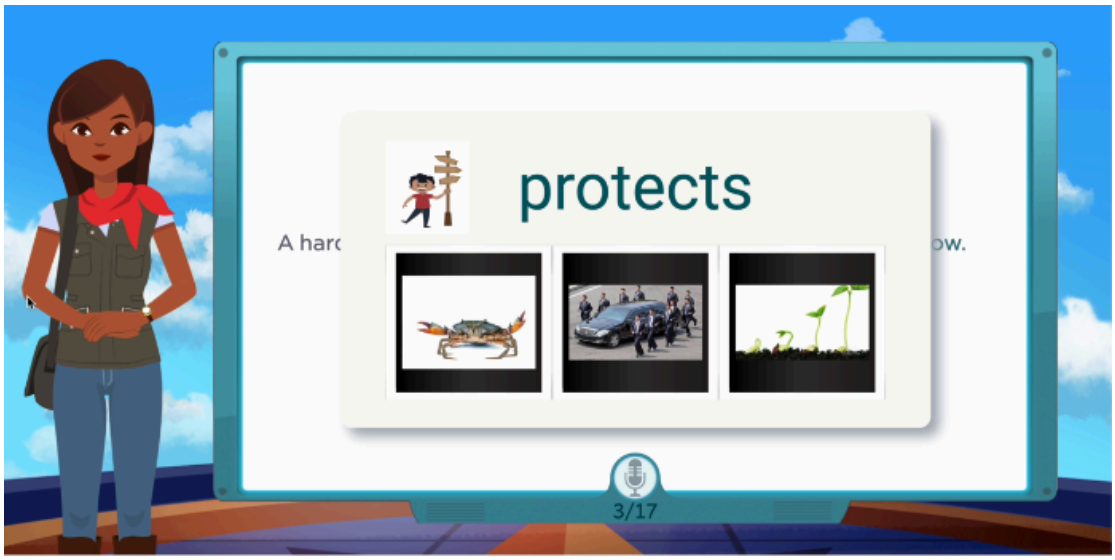


Figure 2.12: Screenshot of the Vocabulary Task - Pictures

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.

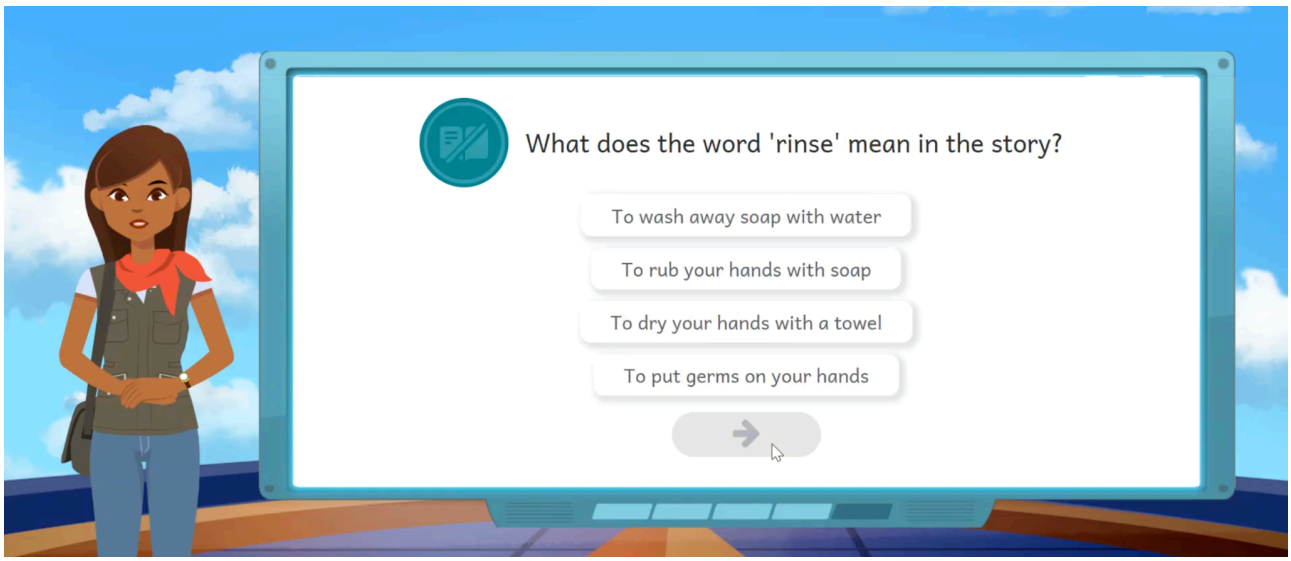


Figure 2.13: Screenshot of the Vocabulary Task - Comprehension

2.7 Spelling/Encoding

Amira ISIP utilizes a dedicated Spelling/Encoding task to assess this construct, and this task is supported in grades K through 3. 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, and the student can move forward at their own pace. If too much time elapses, Amira will automatically move to the next item.

Words vary in difficulty level, and the amount of time a student has to respond is adaptive within 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.

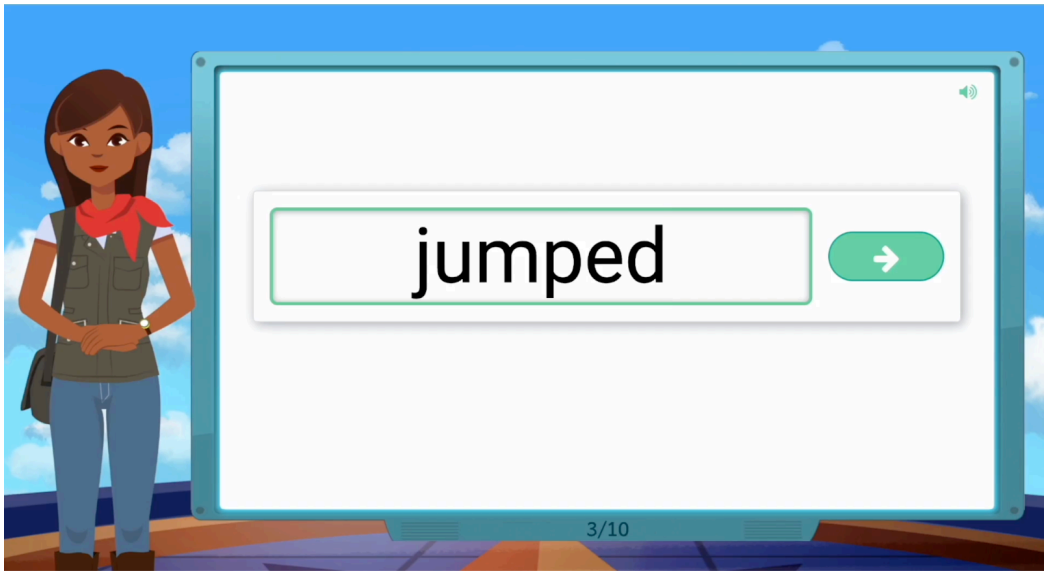


Figure 2.14: Screenshot of the Spelling/Encoding Task

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.

2.8 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 to their performance on reading-based subtests, helps indicate the risk of dyslexia (or the absence of it). 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 like 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 out loud for 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 correctly answer the question.
5. Amira scores the item automatically.

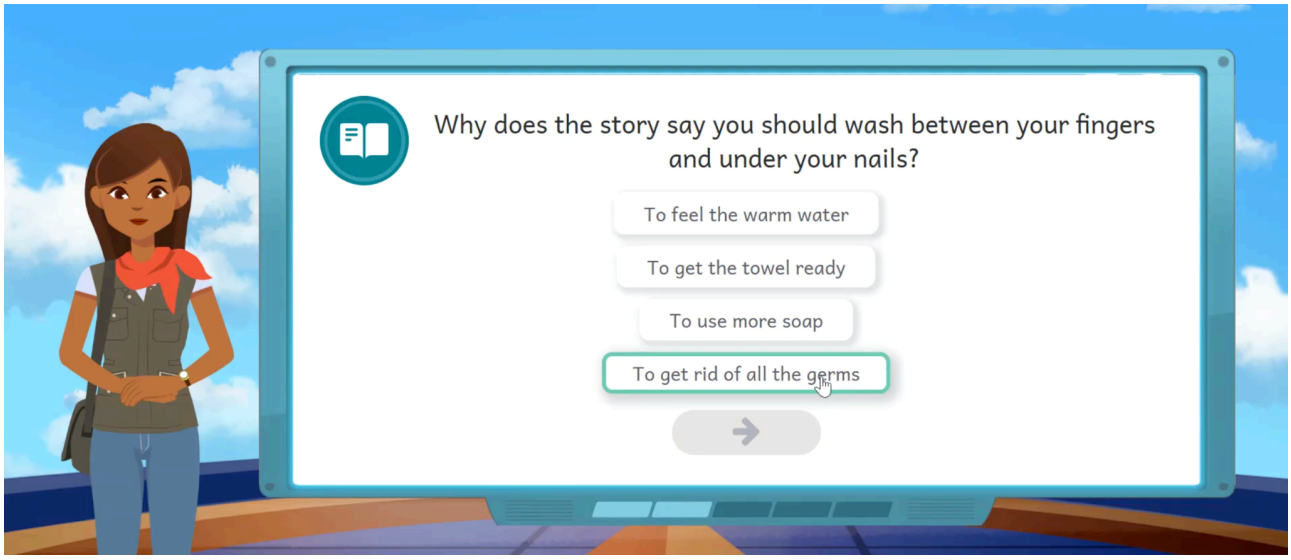


Figure 2.15 Screenshot of the Reading Comprehension task.

See an example video of the reading comprehension task here.

2.9 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, 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.

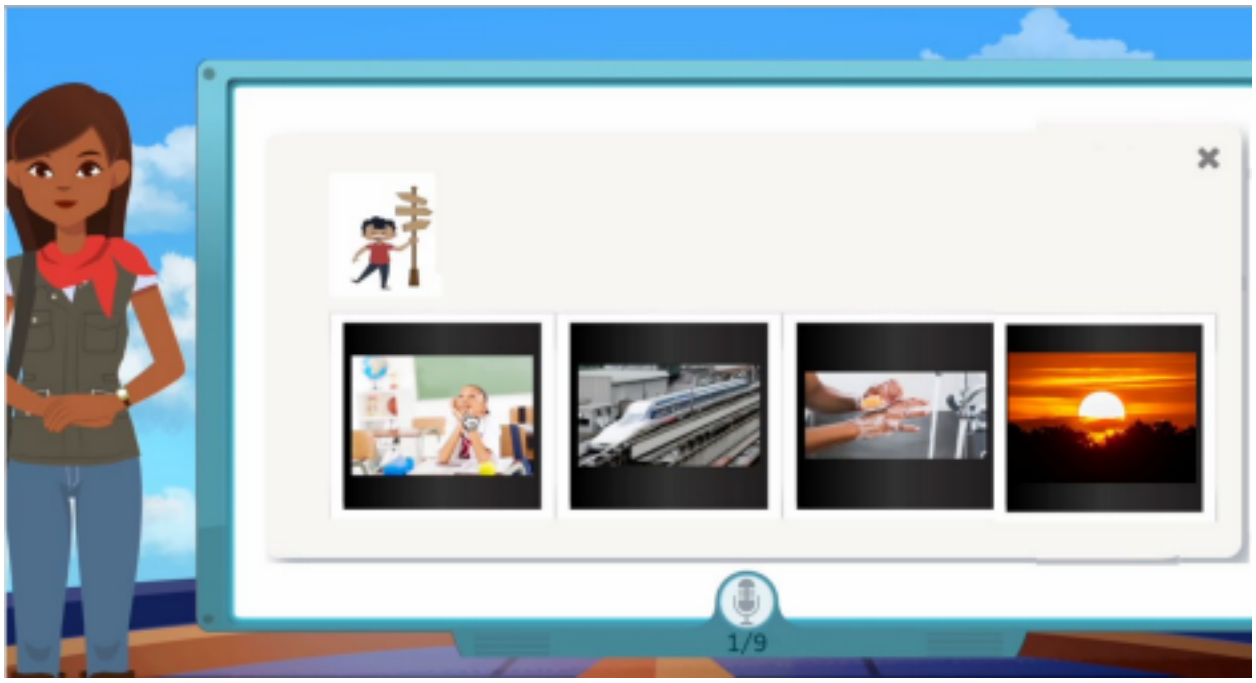


Figure 2.16. Screenshot of the Oral Language Vocabulary Task.

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 offers valuable insights into a child's language comprehension, which is 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 PVT 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 translated into wide acceptance as a measure of Oral Language. For example, Hoffman et al. (2011) found that PPVT scores were predictive of later reading comprehension abilities, 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 with no text that can range from 60 to 75 seconds long. 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.

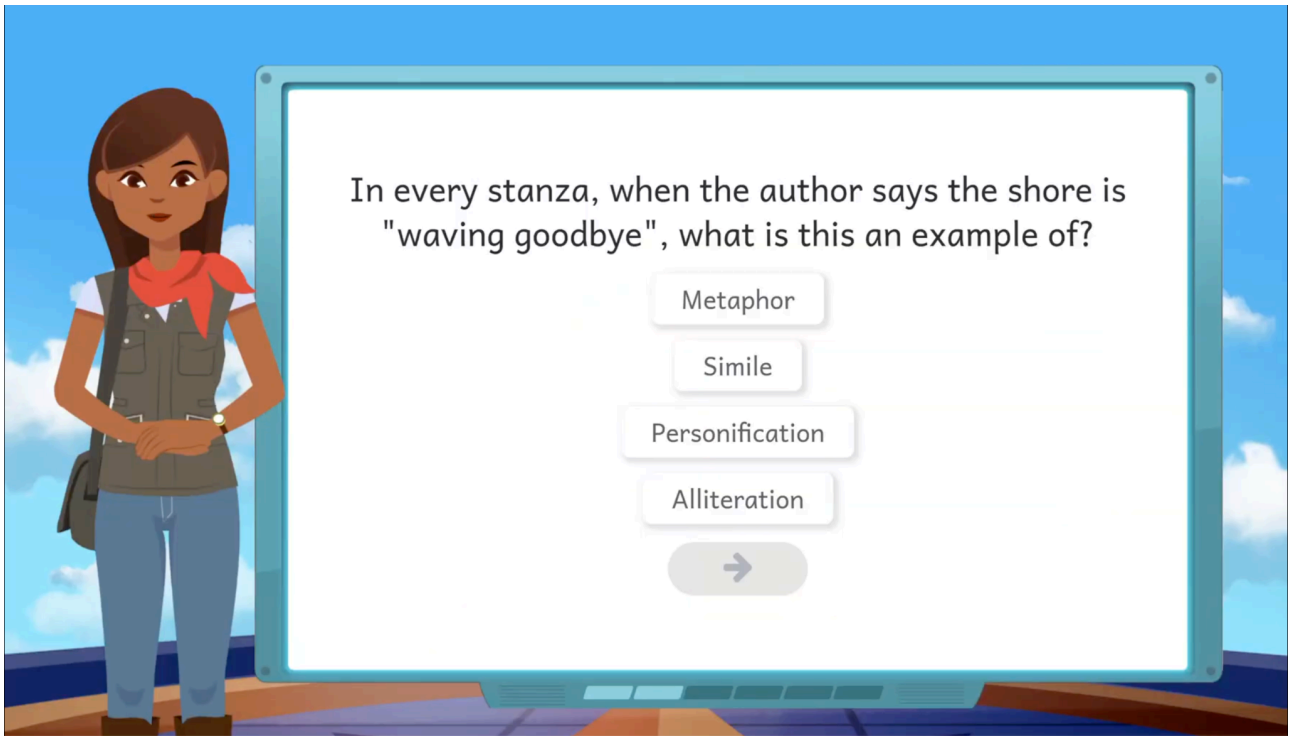


Figure 2.17 Screenshot of an Oral Multiple Choice Question

This task is shown after a spoken passage. All text is verbalized so no reading is required.

See an example video of the oral language comprehension task here.

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. GPT models analyze the language employed by the student to measure their comprehension and understanding of the passage.
5. GPT models analyze the level of the language employed by the student to measure their overall receptive vocabulary and oral language proficiency.

Figure 2.18. Screenshot of the Oral Language Retell Task.

In this task, the AI analyzes a student's oral response to questions relating to a short story told orally. 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 AI models to evaluate key language skills such as vocabulary usage, sentence structure, narrative coherence, and comprehension. 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 and syntax but also reflects their capacity to recall and manipulate language 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 signal 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 that are 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.

In all forms of the RAN task, the stimuli are those that are likely to be known by children at very early ages. For example, in the number RAN task, the stimuli used are numbers between one and nine, always avoiding the use of zero. In the object RAN task, the stimuli used are line drawings of common objects (e.g., dog, book, hand, star). The Benchmark affords two dimensions of customizability: the RAN task type can be configured by the school district or by the software as a function of 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 total processing time required to complete the task. Students are timed and total time to completion is recorded by the system, with item-level accuracy also recorded. Amira ISIP also computes a RAN speed by dividing the number of items accurately named sequentially by the total processing time.

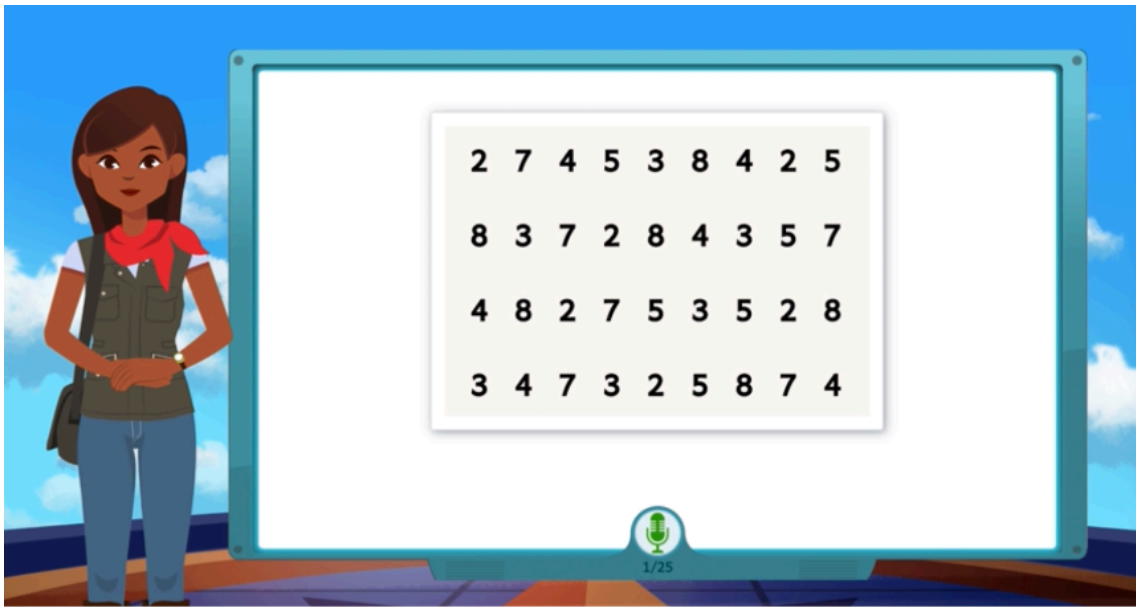


Figure 2.19: Screenshot of the Number RAN Task

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 RAN as a predictive measure of reading ability. Within each task type (numbers, colors, or objects), six different stimulus items per task type are repeated in random order for a total of 36 stimuli, arrayed in four horizontal rows of nine items per row. The RAN task is structured as follows:

1. Amira provides directions to the student to identify the stimuli from left to right, starting with the top line and moving line by line from the top line 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 different stimulus items per task type repeated in random order for a total of 18 stimuli, arrayed in two horizontal rows of nine items per row.
4. The student then proceeds with the full task of 36 stimuli, arrayed in four horizontal rows of nine items per row.

In the abbreviated example Amira provides, she names the stimuli in the top row from left to right, followed by the stimuli in the second row from left to right. Amira tells the students it's their turn and presents a screen with six different stimulus items per task type, repeated in random order for a total of 36 stimuli, arrayed in four

horizontal rows of nine items per row. The student reads the stimuli serially from left to right in each subsequent line going from top to bottom. Total time to completion is recorded as well as item-level accuracy.

See an example video of the RAN task here.

2.10 Visual Attention

Amira ISIP utilizes a Visual Search task to assess Visual Attention, which is 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 concentrate on pertinent visual information and filter 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, which is a key factor influencing reading skills and serves as a predictor of reading difficulties in young children.

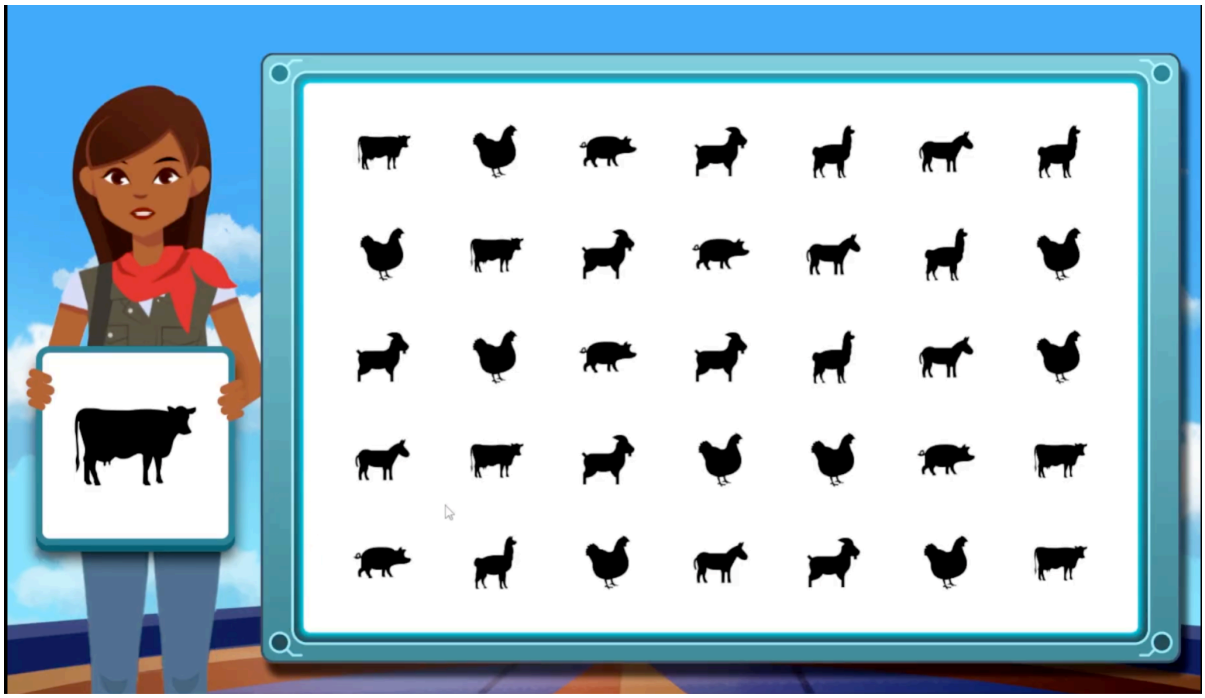


Figure 2.20: Screenshot of the Visual Attention Task

The visual attention task is structured as follows:

1. Students are presented with a visual display containing numerous items that resemble animals, arranged in a grid-like pattern across the screen.
2. Amira provides instructions and a brief example to the student, 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.

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 prekindergarten 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 was developed, grounded in national literacy standards, to outline 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.
- Accessibility and Sensitivity Checks: Items undergo thorough checks for compatibility with assistive technologies and for adherence to accessibility guidelines, ensuring readability and appropriateness.

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, distracter analysis is also conducted.
- Calibration with Item Response Theory (IRT): Items are calibrated using a 2-parameter model. IRT analysis identifies parameters for item difficulty and discrimination and item fit. Items that do not fit the model or that 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 item performance consistency across student subgroups, verifying that no item displays unintended bias toward any demographic group.
- Reliability and Consistency Checks: Items undergo tests for internal consistency and reliability, reviewed by psychometricians to confirm accurate functioning.
- 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 participate in reviewing 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 adaptive 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 sustained item pool growth 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 Item Review and Bias and Sensitivity Committees, providing expertise on 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 to IDOE: Summary reports offer IDOE 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 for additional time or breaks during assessments and can accommodate students taking assessments in both English and Spanish in separate administrations or all at once. Teachers can even configure whether English or Spanish is presented first. For oral reading fluency (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 universal design for learning (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 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 understanding of tasks and provide an equitable opportunity.
- Spanish Screener: Enables students to be screened for dyslexia in Spanish to prevent disproportionate flagging of English Language Learners (ELLs) 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 students who need extra preparation can achieve readiness 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 (through screen magnification), colored overlays, filters, lighting adjustments, tracking devices, and whisper phones.

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 assist teachers in translating 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 mastery levels for fluency, comprehension, word recognition, and overall reading ability.
2. **Norms Specifically for English Language Learners (ELLs)**: Amira ISIP provides information on the same set of measures normed specifically for English Language Learners nationally. 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 ELL Norms
- Proficiency in Spanish

By offering a 360-degree view of an ELL's progress and needs, Amira ISIP provides explicit and clear guidance to teachers on how to utilize assessment data to help ELLs 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 ELL 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 is 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 ELL 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) for specifically the ELL population of Amira ISIP's 2023–2024 usage are shown below in Table 3.2.

Table 3.2: ARM Score benchmarks, for the ELL 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 user experience (UX) for its assessment and reporting platform, guided by thoughtful design principles and continuous feedback loops aimed at enhancing usability, accessibility, and engagement for both students and educators.

For the assessment interface, a key component of this work involves deliberate design considerations that prioritize engagement and efficiency. Amira ISIP's test design carefully balances reliability, validity, and test-taker fatigue, with the goal of keeping 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 a user experience that is both equitable and innovative.

Amira assesses students primarily by having them read aloud. This approach leverages cutting-edge artificial intelligence (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 the student's real, expressive reading abilities and not just their test-taking skills.

3.8 Administration

Amira ISIP is intended to be administered in a 1:1 computer-based setting, where students read aloud into a microphone, and Amira listens and responds. 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—typically at the beginning (BOY), middle (MOY), and end (EOY) of the school year. Periods 1, 5, and 9 respectively align with the BOY, MOY, and EOY testing windows. 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 in each period for 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.

3.7 Computer Adaptive Design

The Amira ISIP Computer Adaptive Test (CAT) functionality is a core component of its assessment suite, designed to personalize the evaluation experience for each student while maximizing accuracy and efficiency.

Amira ISIP employs CAT, grounded in Item Response Theory (IRT) principles, to dynamically adjust the difficulty of assessment tasks in real-time based on a student's performance. This adaptive nature ensures that each student is assessed at their optimal challenge level, providing a personalized experience. The process of Amira ISIP's CAT/IRT approach involves several steps:

- Calibration of Item Bank: An extensive item bank is calibrated using IRT, where each test item is statistically analyzed to determine its unique parameters. Items being calibrated are not used in generating scores. See more detail on Amira ISIP's item calibration process in Section 4.2.
- Initial Estimation of Ability: At the start of an assessment, an initial estimate of the student's ability is made based on their grade level, guiding the selection of the first testlet, typically of medium difficulty.
- Adaptive Testlet Selection: As a student progresses, the CAT algorithm continuously refines its estimate of their ability based on previous responses, selecting subsequent groups of items that are optimally challenging.

- EAP Theta Refinement: EAP estimation is used to update the student's ability estimate after each response, enhancing the precision of the assessment. See Section 4.3 for details on EAP estimation.
- Termination Criteria: Each discrete task has specific termination criteria, including time limits, the ability to collect sufficient items, and measurement error.
- This adaptive design offers significant benefits, including precision and efficiency by quickly identifying a student's true ability with fewer items, which reduces test fatigue and overall testing time. The algorithm minimizes test length while maximizing accuracy. It also provides enhanced diagnostic capability through granular data, identifying specific areas of strength and weakness for targeted instructional interventions. Furthermore, Amira ISIP's CAT ensures fairness and accessibility for all students by presenting items that are appropriately challenging, thereby promoting a more accurate reflection of their true abilities and reducing frustration.

Two-Level Adaptive Design
Amira ISIP's adaptive testing system is designed around a thoughtful two-level decision-making process that ensures each student receives an assessment experience tailored to their individual learning profile and current ability level..

Level 1: Intelligent Task Access
The first level of adaptation focuses on determining which assessment tasks type (for example, whether a kindergarten student demonstrated readiness in oral reading fluency) are most appropriate for each student. Rather than administering every possible task type to every student, Amira ISIP uses gating criteria to make these decisions based on the student's demonstrated abilities across key literacy domains. Some tasks are considered essential and are administered to all students regardless of their current performance level. These core tasks provide fundamental data about each student's reading development. Other tasks are conditionally administered based on whether students have demonstrated readiness in prerequisite skills such as letter knowledge, basic decoding abilities, or phonological awareness.

This approach ensures that students are not overwhelmed with tasks that are far beyond their current developmental level, while also ensuring that advanced students are appropriately challenged with more complex assessment components.

Level 2: Precise Testlet Selection

Once the system determines which tasks a student should complete, the second level of adaptation focuses on selecting the most informative testlets within each task domain. In Amira ISIP's assessment design, approximately five items within a single domain (such as blending or vocabulary) are grouped together as testlets. Each testlet has a target theta value, which represents the ability level where that particular testlet provides the most reliable measurement information. This selection process matches students with testlets that are optimally suited to their current ability level, ensuring the most accurate assessment of their skills.

Throughout the assessment, the system continuously updates its understanding of each student's ability estimates after the completion of each testlet. This real-time EAP interim scoring allows the system to make increasingly precise decisions about which testlet to present next, creating a dynamic assessment experience that adapts moment by moment to the student's performance, ensuring that students are consistently working with items that provide maximum information about their true capabilities while maintaining an appropriate level of challenge.

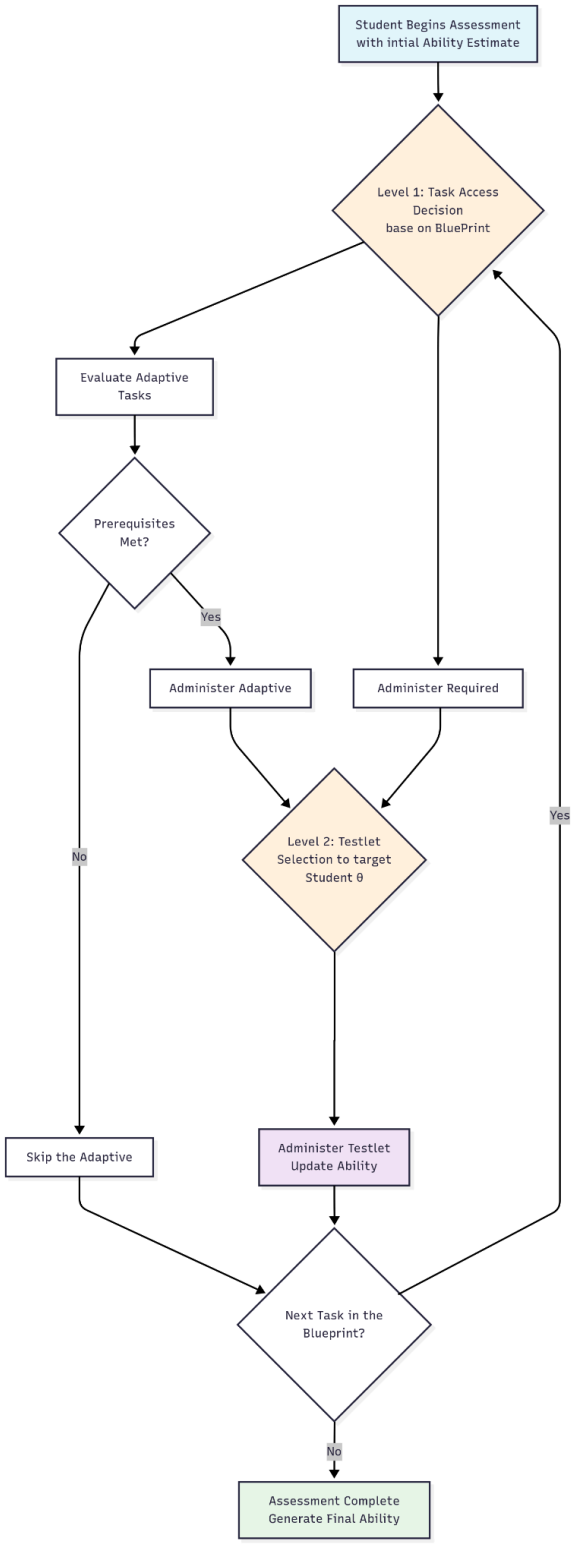


Figure 3.1 Adaptive Test Flow

4. Measurement Model

4.1 IRT Model

Amira ISIP employs a sophisticated item calibration process of 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 (see [Section xx]) and field testing (see [Section xx]) 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 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, requiring greater ability for successful completion.

4.2 Item Calibration

The process of calibration requires applying the 2PL measurement model to a set of data and concurrently estimating person and item parameters. 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 as each type requires 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 5,000 or more 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 differential item functioning (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 operationally, their performance is continuously monitored, allowing for 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 CAT 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 in adaptive testing and is equipped to handle 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 associating performance at various test (or grade) levels to a single scale score (Kolen & Brennan, 2010). It allows for the measurement of student ability across different grade levels using a common scale. This allows for comparability across grades as well as longitudinal growth tracking.

4.4.1 Scaling Methodology

Amira ISIP employs a common-item design approach for vertical scaling, where items that are administered across multiple grade levels serve as linking items to establish the relationship 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 from 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 not meeting both criteria are removed from the anchor set, and the linking is re-estimated with 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) as well as 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 were conducted and determined 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 for 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.

Equipercentile Linking

Since the theta scale produced by the calibration study represents 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 across these shared participants, equipercentile linking enables the mapping of the new theta scale onto the existing ISIP scale, ensuring continuity and interpretability across assessment versions.

4.6 Differential Item Functioning

Item response theory models were employed for detecting differential item functioning (DIF). IRT permits comparisons of item functioning between groups in terms of the probability that performance of that item for each group is different at the same level of ability. To conduct these analyses, an IRT model was constructed that estimated item parameters for each group of interest (e.g., ethnicity), and compares the parameters obtained for this model to a model in which group membership is ignored. 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 of an item. This would indicate that the item is not biased. It is important to recognize that some items will show evidence for DIF solely by chance. The goal is to keep the total number of items indicating DIF below 5%.

DIF was investigated for Grades K to 2 for the End-of-Year (EOY winter) window in 2020. Males were used as the reference group for the gender investigation while

White was used as the reference group for ethnicity. There is little evidence to support pervasive DIF across grades and time for any of the studied groups.

Table 4.2 shows the DIF results for kindergarten. Out of the 124 items on the original calibration Kindergarten form of 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 were in favor of 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 of items.

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
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	

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, there were five items showing DIF, but three of these were advantageous to Hispanic students. The overall rates of DIF for each analysis were below 5% for 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
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

Subtest	N	Gender	Black	Hispanic	Total Item Counts
Total		2	3	5	123
Percent		1.62	2.44	4.00	

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

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
Reading Comprehension		370	0	0	0	12
	Total		1	5	3	100

There is no evidence for systematic item bias by virtue of ethnicity or gender for any of the forms utilized in the Amira ISIP Assessment. The overall rates of DIF for any specific comparison were uniformly at 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.

5. Scoring and Reports

5.1 Reported Scores

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.

5.1.1 ARM

The Amira ISIP Assessments produces one primary composite score called the Amira Reading Mastery (ARM) score. ARM scores use a universal scale that assigns a score to students at all levels of reading ability, ranging from students who cannot yet read connected text to those who can read connected text 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 that scores a 6.34 and another that scores a 6.14, the first is two months further advanced in mastery than the second.

The ARM score synthesizes a student's performance across various measures, specifically from a theta or ability estimate derived from Item Response Theory (IRT) models (for tasks other than RAN and ORF), Rapid Automatized Naming (RAN) speed, and Words Correct Per Minute (WCPM) from Oral Reading Fluency. This score is continuously updated based on all data Amira ISIP collects throughout the school year, including screening, benchmark, and practice sessions, reflecting the student's current proficiency and predicted ability for the future.

It is a norm-referenced score, comparing a student's performance to a nationally representative reference group. The weighting of each screener and ORF task to produce the composite ARM percentile rank (PR) is empirically determined based on predictive analyses of end-of-year reading outcomes, with more weight placed on ORF passage reading as grade level increases. 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 comparison of 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 that completes the screening process, even if they cannot yet read connected text and are still building foundational skills.

5.1.2 MAST

The Amira Mastery of Academic Standards & Targets (MAST) Score is a criterion-referenced score that measures a student's likelihood of mastering state-specific, grade-level academic standards. Rather than comparing students to their peers like norm-referenced assessments, the MAST score focuses on whether individual students have acquired specific knowledge and skills within their current grade level. The score is expressed on a 0-100% scale, representing the percentage of all grade-level standards that a student has likely mastered.

The score calculation relies on sophisticated AI-driven skill mastery models that continuously analyze and synthesize data from all student interactions within the Amira ISIP platform. This comprehensive approach includes performance data from assessment activities, instructional sessions, and tutoring interactions. The AI model weighs multiple factors when estimating mastery likelihood, including task recency, difficulty level, and accuracy rates.

Each academic standard within the MAST framework is mapped to specific reading skills that Amira ISIP's system is trained to observe and evaluate. The mastery estimation process extends beyond simple percentage calculations to incorporate the confidence, consistency, and contextual appropriateness of student responses. This nuanced analysis allows the system to adapt its mastery estimates based on behavioral patterns observed over time.

Individual standard mastery status is communicated through a color-coded RYGG system, where Red, Yellow, Green, and Grey indicators represent different levels of mastery confidence. These status indicators are dynamically updated as new performance data becomes available, with transitions between mastery levels determined by statistical confidence thresholds that ensure reliable and meaningful progress tracking.

5.1.3 DRI

The Dyslexia Risk Index (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 Rapid Automatized Naming (RAN) task, which is a highly valid signal of dyslexia risk and predictive 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 is high, missing only 46 out of 4,506 screened in one study.

5.1.4 Other Key Subscores and Metrics

Amira ISIP also reports on a granular level across the various "threads of the reading rope":

- Oral Reading Fluency (ORF) / Words Correct Per Minute (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 dyslexia risk versus other factors like 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 like Letter-Sound Knowledge, Word Decoding (WIF), and Pseudoword (Nonsense Word) Fluency (NWF), which require students to rely on decoding skills rather than memory.
- Vocabulary: Assesses student’s ability to understand and recognize the meanings of words and phrases in context.
- Rapid Automatized Naming (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 using a dedicated task where students type words.
- Lexile Level: Provides a standardized measure of a student's reading ability and 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 various sub-tasks like 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 and 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 correspondence with Legacy ISIP scores through a concurrent administration design.

Each student in the calibration sample had both an estimated Amira–ISIP ability or 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 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 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.

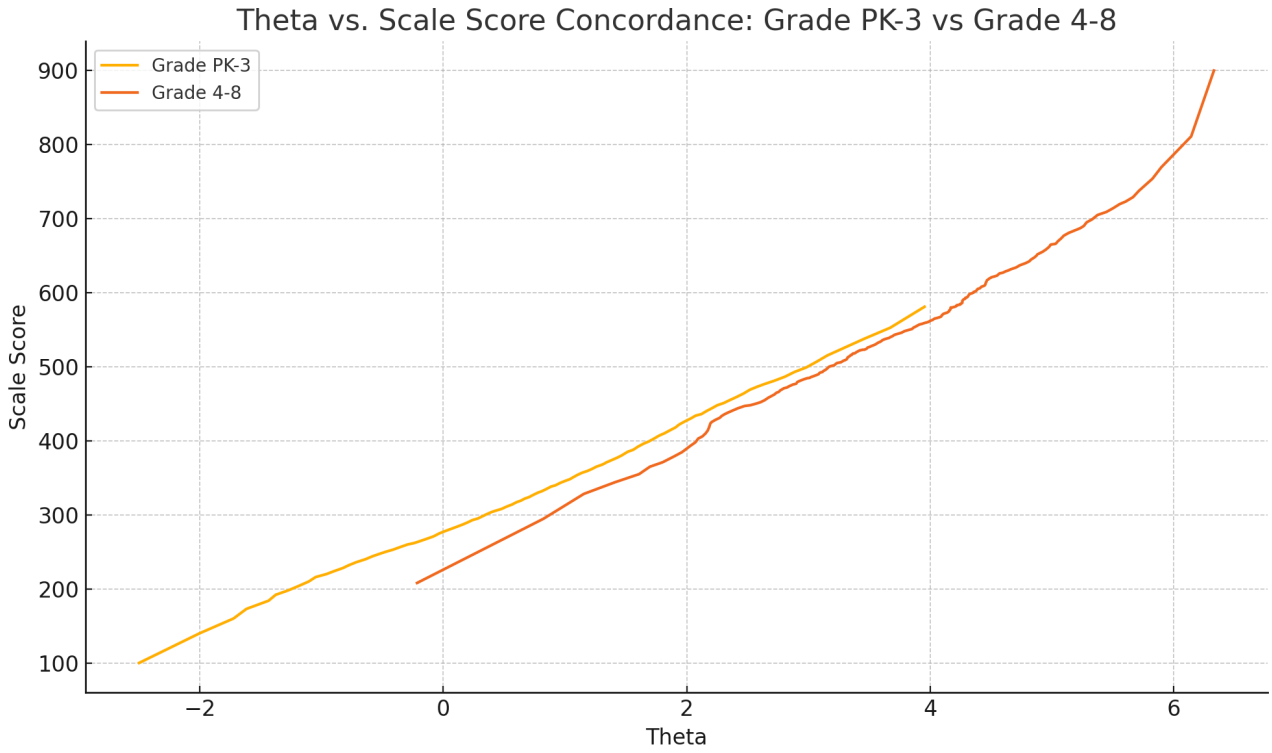


Figure 6.1 Amira ISIP Theta to Legacy ISIP Scale Score Linking Relationship

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

Grade	Correlation (r)	RMSD (logits)	Mean Bias	Absolute Bias
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 the relative standing of examinees within

their respective testing populations while accounting for differences in form difficulty and score distributions. Adjustments can differ at various score ranges, accounting for varying item difficulties for students at different proficiency levels. For instance, two passages may be comparable in difficulty for students reading at 200 words per minute (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 across the country. 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 on the Amira ISIP Benchmark Assessment are presented as percentile ranks and were determined according to grade and testing window. The three testing windows are defined as the following:

- 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 from three assessment time points (fall, winter, spring) in the 2024-2025 academic school year.

Table 7.1: Counts of Students, Districts, and States

Grade	Window	Number of Students	Number of Districts	Number of States
-1	BOY	72658	2452	50
-1	MOY	85004	2694	50
-1	EOY	17556	801	40
0	BOY	557350	10440	50
0	MOY	579335	10195	50
0	EOY	210002	3494	50
1	BOY	556737	10315	50
1	MOY	565958	9442	50

Grade	Window	Number of Students	Number of Districts	Number of States
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
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

Grade	Window	Number of Students	Number of Districts	Number of States
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 across the school year, flag for reading delays or intervention, and determine areas of excellence. National norms are available for the overall ARM score, as well as domain-level subscores.

8. Classification Accuracy

Universal screening is paramount in identifying students at risk for academic difficulty in a RTI model, the core of which is to provide students multi-tiered support based on the level of academic risk that students encounter. 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 for a given time point, they are considered to be 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 evidenced-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 guidelines from the National Center on Intensive Intervention (NCII) in their rating rubrics that delineate technical standards (NCII, 2020a) and their call for submission that provides criteria for submitting evidenced-based universal screening tools (NCII, 2020b). These NCII guidelines are not static across years, and the Amira ISIP Benchmark changes over time in ways that require new research and supporting evidence. The research on universal screening, therefore, gets regularly updated based on these changes. Most recently, the 2023-2024 Amira ISIP Benchmark norms were released in August 2023, which serves as the basis for this updated study.

This study documents the process the Amira Learning team followed in order to determine and validate the cut scores for Fall, Winter, and Spring that can be used to identify students in Grades K to 3 who have severe learning difficulties and need intensive intervention in reading. To establish the universal screening cut scores for the Amira ISIP Benchmark assessment, the NCII rating rubrics (NCII, 2020a) were followed using a sample consisting of students in Arizona, California, Maryland, Indiana, and Illinois – with coverage across at least three of nine geographical divisions defined by U.S. Census Bureau. According to the NCII rating rubrics, this constitutes a national sample. 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.

8.1 Student Sample

Table 8.1 presents the number of students in the study sample across grades by district, state, and the U.S. census division in which each state belongs. The sample

included students in grades K-3 spread across Arizona, California, Maryland, Indiana, and Illinois, covering at least 3 of the 9 U.S. census divisions for each grade level. 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
K	AZ	West	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
			Leading Edge Academy	212

Grade	State	U.S. Census Division	District	# Students
	CA	Pacific	Guadalupe Union SD	129
	IL	East North Central	Evansville	1458
	MD	Middle Atlantic	Baltimore County SD	6405
			Second Grade Total Count	**8890**
3	AZ	West	Amphitheater Public Schools	686
			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 what should be used to identify students at risk for severe learning needs, a recommended approach is to use national norms for the assessment used for the screening purpose (Crawford, 2014). Because the development of national norms tends to use larger and more representative samples, they typically provide accurate and reliable information about the relative standing of an individual student against their peers. If a student's score is lower than an established cut score based on a national norm, this 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 lower than a given candidate cut score within the associated window, they were 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	-0.30	-0.20	-0.10
Kindergarten	Winter	-0.07	0.07	0.21
Kindergarten	Spring	0.15	0.31	0.47
1st Grade	Fall	0.30	0.50	0.70
1st Grade	Winter	0.46	0.66	0.94
1st Grade	Spring	0.67	1.07	1.43
2nd Grade	Fall	0.62	1.20	1.71
2nd Grade	Winter	0.83	1.40	1.97
2nd Grade	Spring	1.15	1.76	2.19
3rd Grade	Fall	1.18	1.99	2.50
3rd Grade	Winter	1.50	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 percentile ranks of each grade/window. The highest scoring cut scores as judged by the lower bound of the 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 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.

Statistic	Description	Interpretation
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 in 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 the candidate ARM cut scores corresponding to the 30th PR based on the national norms performed the best for 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 result in the sensitivity, specificity, and the lower bound of the area under the ROC curve (AUC-LB) being at least 0.7 in Kindergarten Fall, satisfying 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
Fall	-0.1	NWEA MAP	20th PR	0.80	0.77	0.80	0.79
Winter	0.21	NWEA MAP	20th PR	0.81	0.84	0.81	0.82
Spring	0.47	NWEA MAP	20th PR	0.82	0.84	0.82	0.83

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
Fall	0.7	NWEA MAP	20th PR	0.86	0.82	0.87	0.84
Winter	0.94	NWEA MAP	20th PR	0.88	0.84	0.88	0.86
Spring	1.43	NWEA MAP	20th PR	0.88	0.81	0.89	0.85

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
Fall	1.71	NWEA MAP	20th PR	0.87	0.804	0.88	0.85
Winter	1.97	NWEA MAP	20th PR	0.86	0.83	0.90	0.85
Spring	2.19	NWEA MAP	20th PR	0.87	0.84	0.91	0.85

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
Fall	2.50	NWEA MAP	20th PR	0.86	0.84	0.88	0.86
Winter	2.92	NWEA MAP	20th PR	0.88	0.83	0.90	0.87
Spring	3.19	NWEA MAP	20th PR	0.90	0.84	0.91	0.88

For Grades 4, 5 and 6, using Hasbrouck & Tindal 2017 WCPM norms as the criterion measure, result in the lower bound of the area under the receiver operating characteristics (ROC) curve (AUC) being at least 0.7 in all windows of Grades 4 and 6, satisfying the half bubble criteria. For Grade 5, the benchmarks result sensitivity ≥0.8, specificity≥0.8, and the lower bound of AUC being 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 with data from 2023 produced similar results as the Hasbrouck & Tindal (see Tables 3.10 and 3.11). In grade 4 the receiver operating characteristics (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 highest for spring. For Grade 5, the AUC was 0.74 in the fall, 0.77 in the winter, and 0.75 in the spring. Similar to grade 4, the sensitivity was highest in winter and specificity highest for 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	0.80	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 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 multi-tiered support based on the level of academic risk that students encounter. 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 for a given time point, they are considered to be 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 evidenced-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 guidelines from the National Center on Intensive Intervention (NCII) in their rating rubrics that delineate technical standards (NCII, 2020a) and their call for submission that provides criteria for submitting evidenced-based universal screening tools (NCII, 2020b).

This study documents the process the Amira Learning team followed in order to validate the cut scores for Fall, Winter, and Spring that can be used to identify students in Kindergarten and First Grade who have severe learning difficulties and need intensive intervention in reading. To establish the universal screening cut scores for the Amira ISIP Benchmark assessment, the NCII rating rubrics (NCII, 2020a) were followed using a sample consisting of students in Texas, South Carolina, Kentucky, and Oklahoma. In 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, in order 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
		Vernon ISD	53
	South Carolina	Lancaster Co SD	57

Grade	State	District	N
		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

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	0.70	0.71	0.71
First Grade	Phonological Awareness	NWEA MAP: Phonological Awareness Domain	0.79	0.80	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	0.80	0.80	0.79
First Grade	Word or Pseudo Word Reading Fluency	NWEA MAP: Phonics/Word Recognition Domain	0.80	0.83	0.88
First Grade	Oral Reading Fluency	NWEA MAP: Oral Reading Fluency	0.87	0.90	0.95

9. Reliability and 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 two different reliability studies as applied to the ARM composite score.

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 the 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 ARM Composite Score

Grade	N	Number of Items	Cronbach's Alpha
K	14116	26	0.86
1	16609	29	0.85
2	14513	22	0.93
3	14546	29	0.93
4	14513	36	0.91
5	14544	40	0.91
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 is a measure of 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, as taken by the same student within two weeks in the Winter window in 2022-2023 school year. All correlation coefficients are at 0.7 or higher, which indicates that the Amira ISIP Benchmark and Progress Monitoring Assessments is 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
0	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 Parallel Forms Reliability

The third study examines Parallel Forms Reliability, measuring the correlation between scores of students who have taken two different forms within the same screening window and calculating the correlation between the scores. If the correlation is high, it indicates that the test is reliable.

Two forms of the benchmark oral reading fluency (ORF) assessment were administered to the same group of students to establish parallel forms reliability. The WCPM scores obtained on each ORF assessment version are then correlated to assess the degree of consistency between them.

Table 9.3 shows the Pearson correlation coefficients of the two WCPM scores computed based on Amira ISIP Benchmark ORF assessments taken in each grade's 2022-2023 SY window. The correlation coefficients were at 0.71 or higher, suggesting

that the Amira ISIP Benchmark ORF assessment consistently measures students' ORF ability.

Table 9.3: 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

9.1.4 Inter-rater Reliability

Amira ISIP's intelligent tutoring system employs a range of Artificial Intelligence (AI) and Machine Learning (ML) technologies. The common method to validate the reliability of an ML model is replicating the results via human classification. The AI system outcomes are compared to the judgment 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 Reading Error Detection (RED) model and those given by human raters using Pearson's correlation coefficient. The RED scoring system is an automated system that utilizes advanced algorithms to detect errors and assess the quality of students' speech. Experienced educators were employed to alternatively score the students' speech independently.

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 can be used as a reliable and 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 9.4: Inter Rater Reliability

Grade	Sample Size	Inter-Rater Reliability
K	6186	0.95
1	6771	0.98
2	8400	0.97

Amira ISIP's Screener demonstrates exceptional reliability across diverse student populations. Through comprehensive evaluations of internal consistency, test-retest reliability, alternate form reliability, and interrater agreement, Amira ISIP has consistently shown reliability estimates that meet or exceed a coefficient of 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, and interrater reliability as high as 0.98—underscore the consistency and accuracy of Amira ISIP's assessments. These findings provide compelling evidence that educators can trust Amira ISIP's data to inform critical instructional and intervention decisions with confidence.

9.1.5 Reliability of Subgroups

Amira ISIP's Screener demonstrates a high level of reliability across a diverse range of student subgroups, including those representative of California's student population. 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 a coefficient of 0.90 with many of them above .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.

Grade/Age

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

Gender

Grade	Gender	N (Number of Participants)	Internal Consistency	Interrater 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

English Learner Status

Grade	English Learner Status	N (Number of Participants)	Internal Consistency	Interrater Reliability
K	EL/MLL	1026	0.91	0.97

Grade	English Learner Status	N (Number of Participants)	Internal Consistency	Interrater Reliability
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

Exceptionality Status

Grade	Exceptionality Status	N (Number of Participants)	Internal Consistency	Interrater 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

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

Grade	Race	N (Number of Participants)	Internal Consistency	Interrater Reliability
K	Asian	315	0.93	0.96
K	Black	1653	0.90	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

Socio-Economic Status

Grade	Socio-Economic Status	N (Number of Participants)	Reliability Coefficient (Cronbach's Alpha)	Interrater Reliability
K	Low SES	500	0.90	0.98
K	High SES	964	0.90	0.97
1	Low SES	532	0.93	0.98
1	High SES	1012	0.93	0.98

Grade	Socio-Economic Status	N (Number of Participants)	Reliability Coefficient (Cronbach's Alpha)	Interrater Reliability
2	Low SES	575	0.92	0.98
2	High SES	949	0.90	0.97

Language Backgrounds

Grade	Home Language	N (Number of Participants)	Reliability Coefficient (Cronbach's Alpha)	Interrater Reliability
K	English	1018	0.90	0.97
K	Spanish	446	0.90	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 address the specific needs of each subgroup effectively.

9.1.6 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 Fluency (Words Correct Per Minute (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 9.5 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	1.3%

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 9.6 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 9.7 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 into high, typical and low just over 84% of the time.

9.2 Validity

To truly understand and enhance a student's reading skills, it is crucial to rely on psychometrically valid instruments that are grounded in rigorous scientific principles, ensuring accuracy and reliability of 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.

9.2.1 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 main premise of the model; that is, 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).

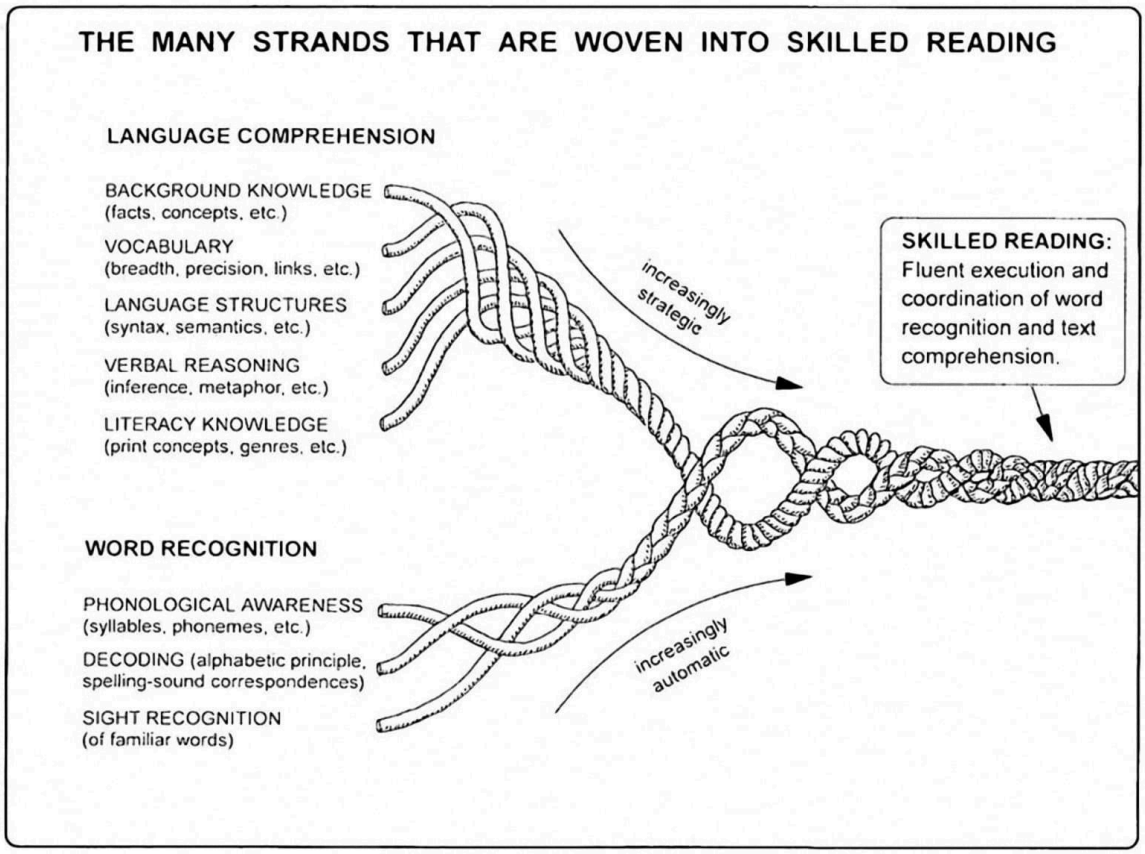


Figure 9.1: Simple View of Reading Framework from Scarborough (2001)

A key dimension of Amira ISIP's validity evidence is that the scores emanating from the assessment substantially explains 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:

- Words correct per minute (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 all phonemes have been pronounced; PA is an unweighted average over 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 a student sample (N) ranging from 23,023 – 291,492 per grade that had scaled scores for correlation, depending on the popularity of the Amira ISIP Assessment in terms of use for a particular grade in a particular language.

Table 9.8 Correlation Matrix among Amira ISIP English Scaled Scores for Kindergarten

N = 163,828	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.72	1.00			
PA	0.73	0.97	1.00		
HFW	0.75	0.93	0.92	1.00	
Vocab	0.83	0.88	0.88	0.95	1.00

Table 9.9 Correlation Matrix among Amira ISIP English Scale Scores for Grade 1

N = 267,891	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.74	1.00			
PA	0.76	0.98	1.00		
HFW	0.81	0.95	0.95	1.00	
Vocab	0.91	0.84	0.85	0.91	1.00

Table 9.10 Correlation Matrix among Amira ISIP English Scale Scores for Grade 2

N = 291,492	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.76	1.00			
PA	0.76	0.99	1.00		
HFW	0.75	0.97	0.97	1.00	
Vocab	0.88	0.86	0.86	0.86	1.00

Table 9.11 Correlation Matrix among Amira ISIP English Scale Scores for Grade 3

N = 265,800	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.72	1.00			
PA	0.72	0.99	1.00		
HFW	0.70	0.97	0.97	1.00	
Vocab	0.85	0.87	0.87	0.83	1.00

Table 9.12 Correlation Matrix among Amira ISIP English Scale Scores for Grade 4

N = 163,685	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.73	1.00			
PA	0.73	0.99	1.00		
HFW	0.70	0.91	0.90	1.00	
Vocab	0.85	0.91	0.91	0.78	1.00

Table 9.13 Correlation Matrix among Amira ISIP English Scale Scores for Grade 5

N = 137,542	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.70	1.00			
PA	0.70	0.99	1.00		
HFW	0.70	0.90	0.90	1.00	
Vocab	0.82	0.87	0.87	0.76	1.00

Table 9.14 Correlation Matrix among Amira ISIP English Scale Scores for Grade 6

N = 23,023	WCPM	Decode	PA	HFW	Vocab
WCPM	1.00				
Decode	0.71	1.00			
PA	0.71	0.99	1.00		

N = 23,023	WCPM	Decode	PA	HFW	Vocab
HFW	0.71	0.88	0.87	1.00	
Vocab	0.81	0.94	0.93	0.77	1.00

9.2.2 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 in a difficulty learning to read and decoding words (Ziegler & Goswami, 2005), even when presented with instruction that typically works to help students succeed. Much of the work and research that has gone into screener development has focused on isolating the components of being able to acquire word reading skills at a normal developmental progression. These components include the core skills of phonological awareness (PA), alphabetic knowledge (letter name fluency and letter sound fluency), decoding, and automaticity. In higher grades, more emphasis is placed on tasks that more directly measure how accurately and fluently kids can actually read words or pseudo-words that require actual decoding (to be distinguished from reading words by sight). There have been several evaluations of the full screener that demonstrate its validity in identifying children who are 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 to be included in the minimum configuration, we require the research to support 1) a link between the task to 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 dyslexic subjects (Boada & Pennington, 2016; Kovelman et al., 2012; Ramus et al., 2013; Scarborough, 1990; Swan & Goswami, 2007; Scarborough). 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 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 ability to decode pseudo-word/non-word words is essential screening information both for 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 lead to problems in reading acquisition (e.g., Goswami & Bryant, 1990).

Rapid Automatized Naming
The rapid automatized naming (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 automatization of verbal responses to visual stimuli, not restricted to symbols, correlates specifically 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 specific executive functions that support verbalizing sequences.

Furthermore, Boscardin et al. (2008) found that measurements 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, compared to 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 instruction in reading or are behind in formal instruction.

Word Identification Fluency and Oral Reading Fluency
Especially by Grade 1 and beyond, after students have had a chance to receive formal instruction in 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 will commonly persist in individuals with dyslexia even after they have caught up to on-grade level in reading through intensive instruction (Treiman, 1997), making the task particularly diagnostic for the range of 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 being severely taxed when a child with a reading disorder is parsing an unfamiliar text. Shankweiler et al. (1999) have similarly demonstrated the strong relationship between comprehension and decoding for children with reading difficulties.

9.2.3 Content Validity

Amira ISIP's content validation process, detailed in Table 9.15 below, includes comprehensive activities from content review to IRT modeling.

Table 9.15 Content Validation Process 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 exhibiting differential functioning for subpopulations. Items showing DIF are eliminated.

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

Firstly, 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.

9.2.4 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 to 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 of Amira ISIP's ARM scores in relation 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, for each associated external measure used. 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 9.16 Correlation Coefficient Between Amira ISIP Scale Scores and External Screener Scores

Grade	External Screener	N	Concurrent Validity
K	NWEA MAP	5861	0.75
1	NWEA MAP	6415	0.80
2	NWEA MAP	6696	0.80
3	iReady Diagnostic	1065	0.72

Grade	External Screener	N	Concurrent Validity
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.

9.2.5 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 students in Grade 3 who participated in Amira ISIP Assessments in the Fall and the iReady Diagnostic assessment in the Spring. We assessed the validity of Amira ISIP's ARM scores in comparison to 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, for each external measure utilized. 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 9.17 Predictive Validity Correlation Coefficients

Grade	External Screener	N	Predictive Validity
K	NWEA MAP	5309	0.71
1	NWEA MAP	6148	0.81
2	NWEA MAP	6405	0.79

Grade	External Screener	N	Predictive Validity
3	iReady Diagnostic	1162	0.73
3	NWEA MAP	1848	0.74

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 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 in comparison to these external criterion measures of 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 9.18 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

Grade	State	District	n
		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 9.19 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

Grade	Amira ISIP Subscore	Test or Criterion Measure	n	Coefficient
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

9.2.6 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 in a difficulty learning to read and decode words (Ziegler & Goswami, 2005), even when presented with instruction that typically works to help 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 to 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 who are 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.

9.2.6.1 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 9.20 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.

		Grammagio Not Flagged For SLI	Grammagio Flagged for SLI	Totals	
Amira Dyslexia Risk Index	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 9.21: 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

9.2.6.2 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. Spanish Screener

10.1 Subtests

All subtests on Amira ISIP's English screener are also available in Spanish. When developing the Spanish version of the assessment, special attention was given to cultural relevance and appropriateness of the skills in Spanish versus English. Amira ISIP's Spanish assessment is grounded in the *Tejas Lee*, built from the ground up as a test for Spanish reading. All Spanish items are authentic – none have been translated from the English item bank.

Amira Learning collaborated closely with Spanish reading scientists, bilingual education experts, and policymakers to develop an assessment that matches Amira ISIP English in quality and is fully rooted in the most effective evaluation of Spanish reading proficiency. Figure A1 in Appendix A showcases some of the experts who contributed to the design and quality assurance of Amira ISIP Spanish.

To ensure alignment with Spanish and bilingual curricula, the Spanish Assessment underwent a rigorous Standard Setting Process. Here are some key points to note:

1. All passages were originally crafted in Spanish.
2. Each task (sub-test) features items that are specifically tailored to Spanish language skills.
3. Every item undergoes multiple rounds of review by distinguished panels to ensure cultural sensitivity and appropriateness.

To view Amira ISIP's Spanish screener including the subsets of tasks, refer to the resources provided here. The *Tejas Lee* site can be accessed here.

Screenshots of each Spanish Screener task are provided below.

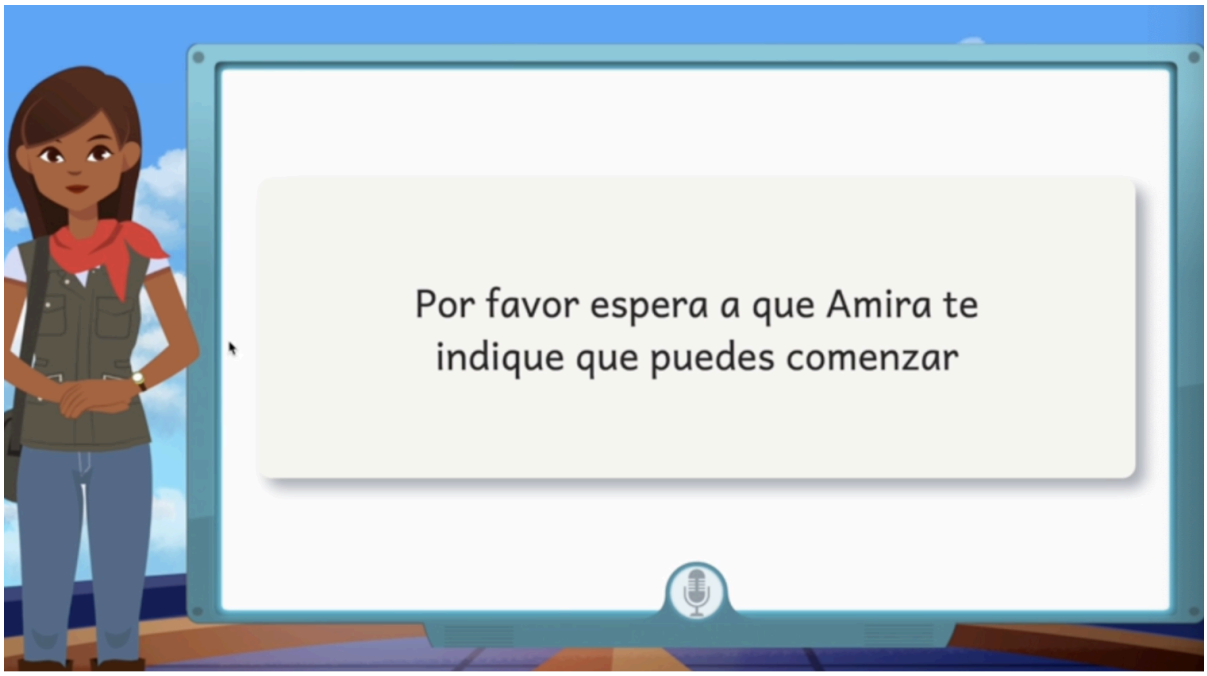


Figure 10.1 Screenshot of the Amira ISIP Spanish Screener

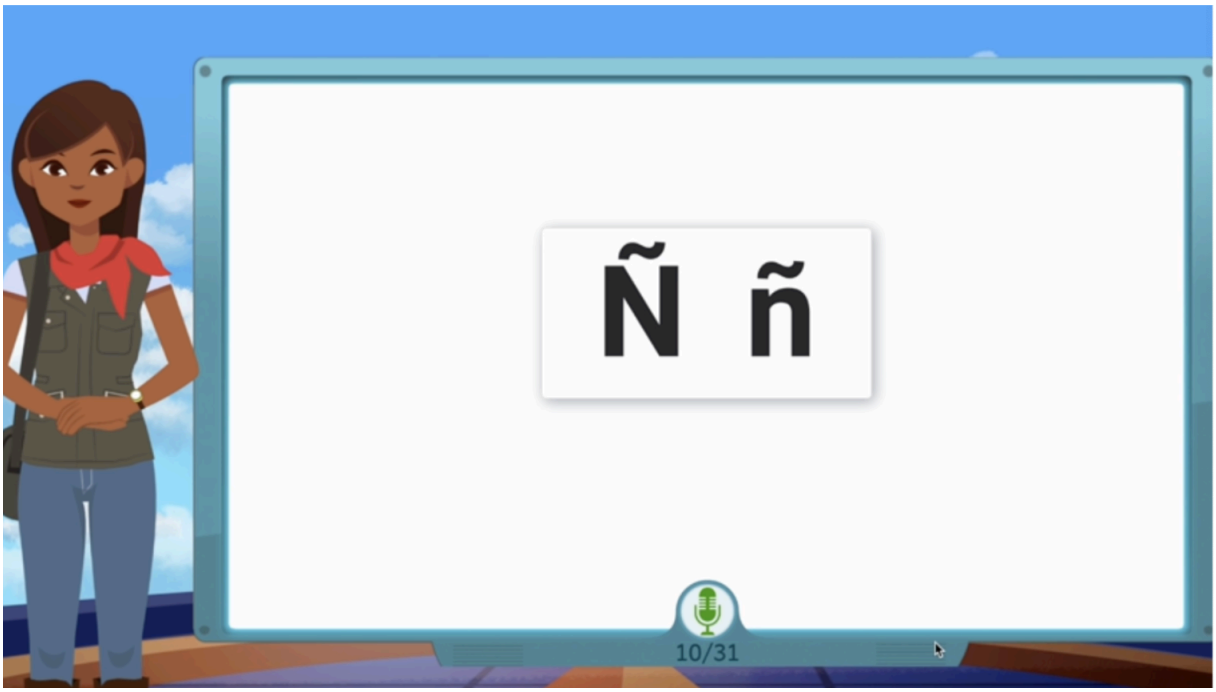


Figure 10.2 Screenshot of the Spanish Screener Letter Sound Fluency/Letter Name Fluency Task

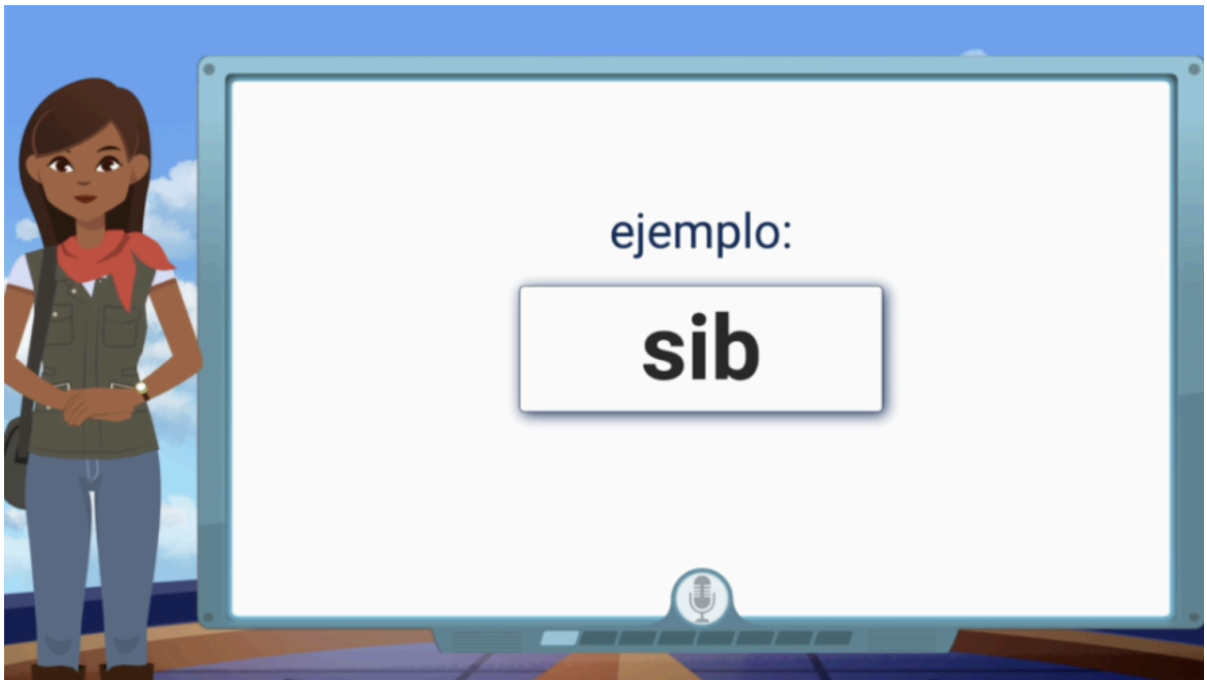


Figure 10.3 Screenshot of the Spanish Pseudo-word/Non-word Decoding

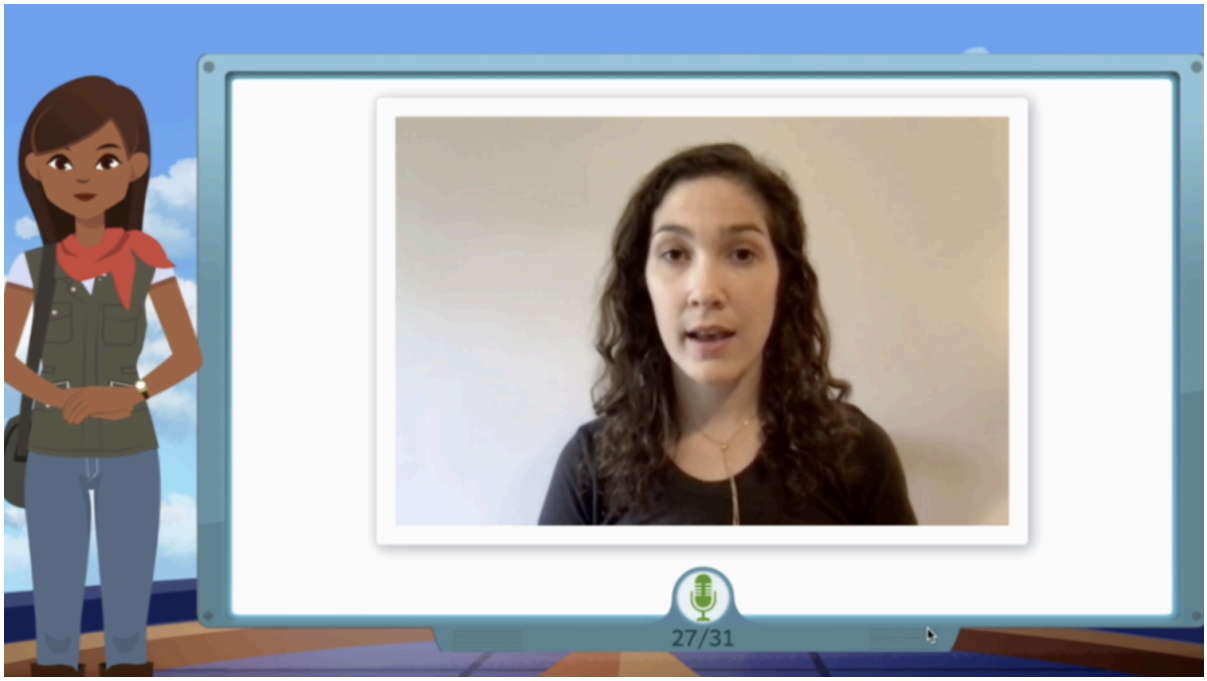


Figure 10.4 Screenshot of the Spanish Phonological Awareness Task

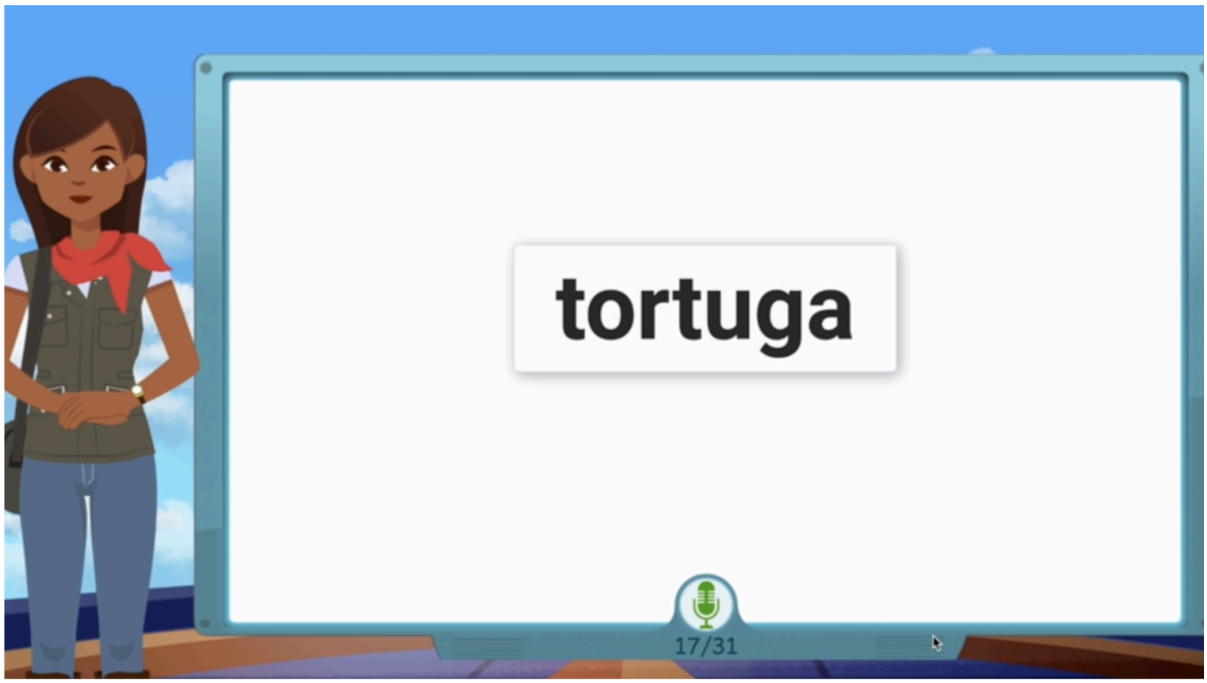


Figure 10.5 Screenshot of the Spanish Word Reading Task

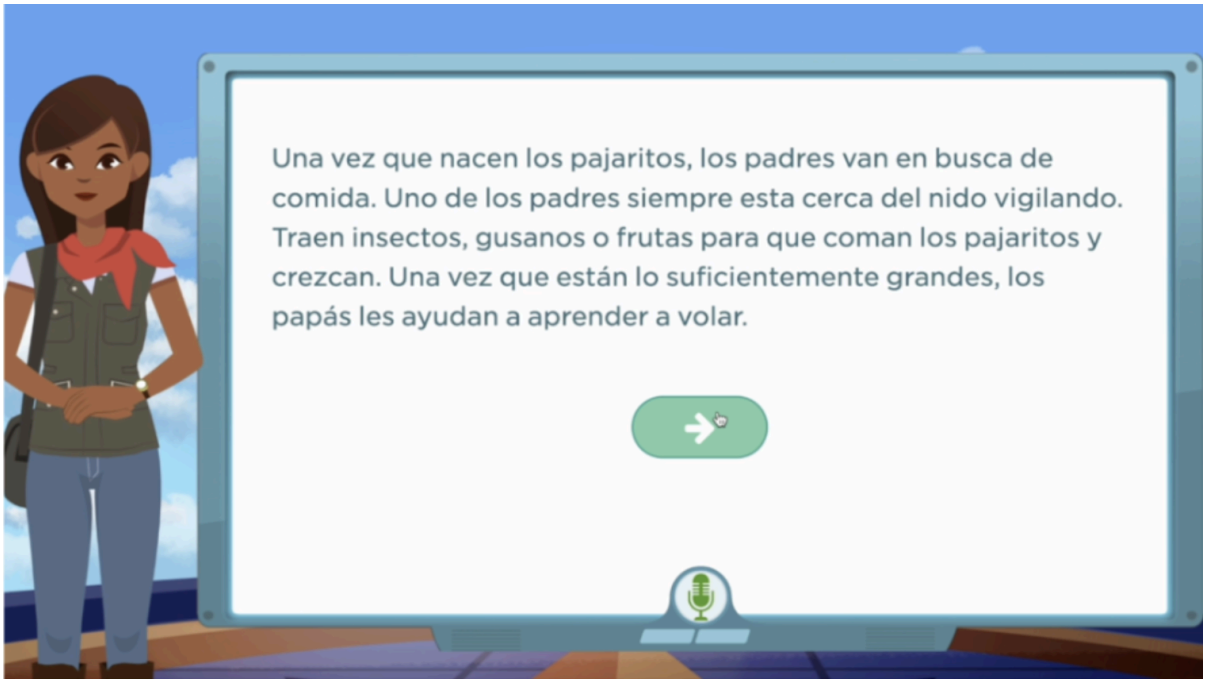


Figure 10.6: Screenshot of the Spanish Story Reading/ORF Task

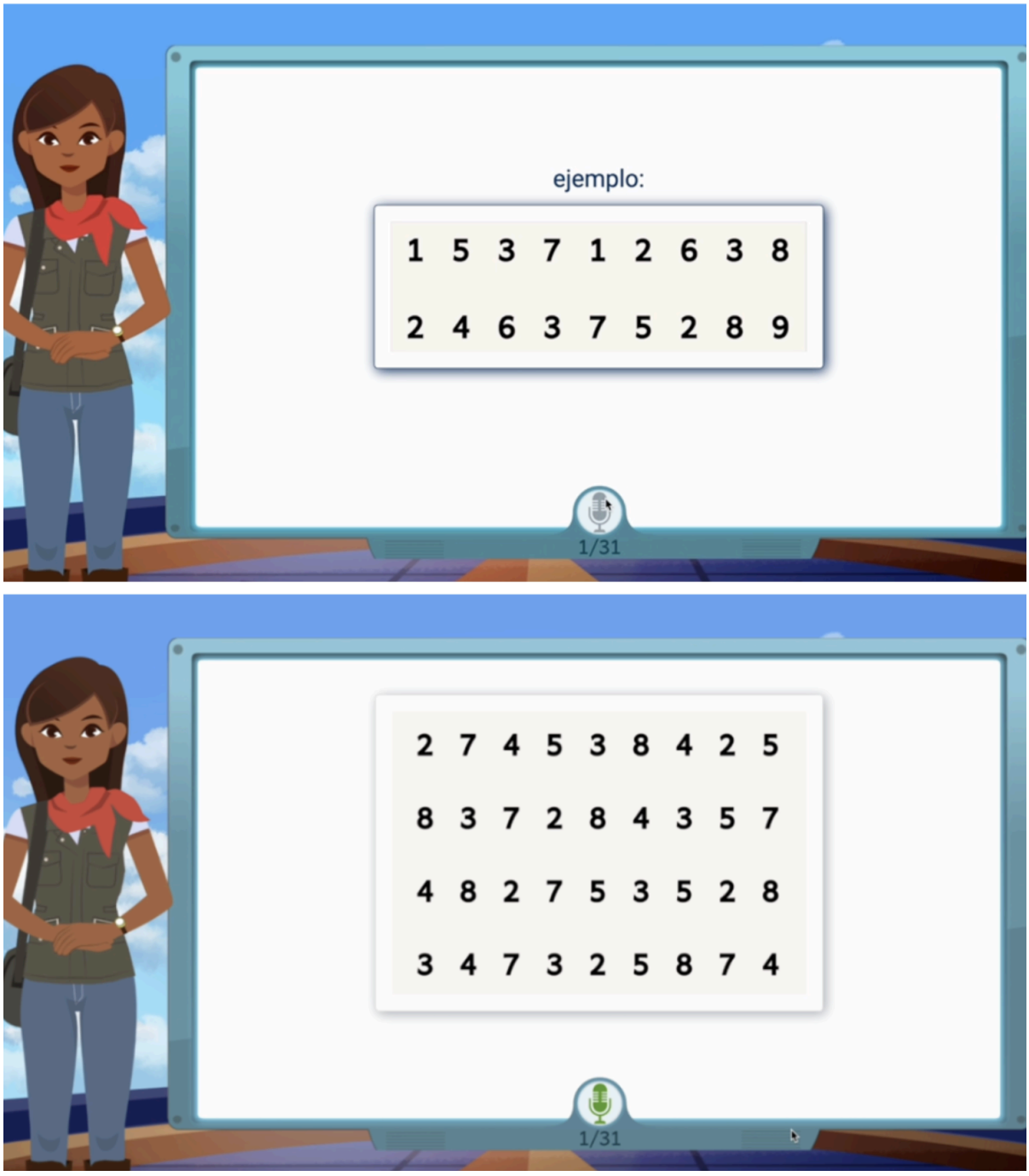


Figure 10.7 Screenshots of the Spanish RAN Task

10.2 Development of National Norms

Amira ISIP has sub-measure score norms for the Spanish screener tasks delineated in the previous section. The total sample size was 69,48669,486 students across Grades K to 5 from 2022 – 2023 school year. Table 10.1 below describes the features of the norming samples.

Table 10.1 Counts of Districts and Schools Used Amira ISIP Spanish Screener

Grade	Window	Number of Districts	Number of Schools	Number of Students
Kindergarten	BOY	114	384	5853
Kindergarten	MOY	175	619	8815
Kindergarten	EOY	160	576	8480
1st Grade	BOY	190	733	10687
1st Grade	MOY	212	865	12393
1st Grade	EOY	193	738	10714
2nd Grade	BOY	215	742	10571
2nd Grade	MOY	248	842	11667
2nd Grade	EOY	222	749	11027
3rd Grade	BOY	201	757	9586

Grade	Window	Number of Districts	Number of Schools	Number of Students
3rd Grade	MOY	216	815	9976
3rd Grade	EOY	174	656	7975
4th Grade	BOY	170	552	6819
4th Grade	MOY	183	616	7770
4th Grade	EOY	148	464	6062
5th Grade	BOY	145	496	6009
5th Grade	MOY	164	526	6354
5th Grade	EOY	101	383	4592
6th Grade	BOY	58	149	1158
6th Grade	MOY	66	150	1097
6th Grade	EOY	42	89	555

Amira ISIP produces Spanish Screener scores, norms, and assigns students PRs for each of the following sub-measures: WCPM, Decoding (Alphabetic Knowledge), HFW, Phonological Awareness, Vocabulary, Lexile, and Developmental Reading Assessment (DRA). The definitions of all Amira ISIP Spanish Screener constructs align with those of the Amira ISIP English Screener. See Tables B1 to B7 in Appendix B for score cuts associated with each measured construct.

10.3 Teacher Guidance for Interpreting Scores

This section provides guidance for teachers to interpret scores for bi- and multilingual students and/or English language learners (ELLs). Amira ISIP offers a worksheet designed to assist teachers in making informed decisions about supporting their ELL students. The guidance is structured and definitive, incorporating Amira ISIP's National and ELL benchmarks along with its English and Spanish dyslexia screeners.

The information provided on the worksheet is summarized as follows:

- Step 1: Determine if the student's Amira ISIP National PRs are above the intervention cut line. If so, no further analysis is necessary.
- Step 2: If the student's PRs fall within the intervention zone, check their ELL PRs. If these are above the intervention cut line for ELL students, no additional analysis is required.
- Step 3: If the student's ELL PR is below the cutline, assess the student's English DRI. If it is within the normal range, standard MTSS strategies should be implemented.
- Step 4: If the student's English DRI is high and they are a native Spanish speaker, evaluate their Spanish Screener score.
- Step 5: If the student's DRI for Spanish is low, standard MTSS support should be provided.
- Step 6: If the student is flagged for DRI in English and is not a native Spanish speaker, or if they are a native Spanish speaker flagged for DRI by the Spanish screener, support for students with phonological deficits should be implemented.

References

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

Baker, D. L., Good, R. H., Mross, A. P., McQuilkin, E., Watson, J., Chaparro, E., & Sanford, A. K. (2006). Fluidez en la lectura oral IDEL. In D. L. Baker, R. H. Good, N. Knutson, & J. M. Watson (Eds.), Indicadores dinámicos del éxito en la lectura (7th ed., pp. 36–45). Eugene, OR: Dynamic Measurement Group. Retrieved from http://dibels.uoregon.edu/

Biemiller, A. (2006). Vocabulary development and instruction: A prerequisite for school learning. Handbook of Early Literacy Research, 2, 41–51.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Part 5 of Lord & Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Boada, R. & Pennington, B. F. 2006. Deficient implicit phonological representations in children with dyslexia. Journal of Experimental Child Psychology, 95: 153–193.

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. Journal of Educational Psychology, 105(4), 1142–1151.

Bonifacci, P., & Tobia, V. (2017). The simple view of reading in bilingual language-minority children acquiring a highly transparent second language. Scientific Studies of Reading, 21, 109–119.

Boscardin, C. K., Muthén, B., Francis, D. J., & Baker, E. L. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. Journal of Educational Psychology, 100(1), 192.

Bourassa, D., & Treiman, R. (2003). Spelling in children with dyslexia: Analyses from the Treiman-Bourassa Early Spelling Test. Scientific Studies of Reading, 7(4), 309–333.
Breznitz, Z. (1987). Increasing first graders' reading accuracy and comprehension by accelerating their reading rates. Journal of Educational Psychology, 79(3), 236–242.

Brown, J. E., & Sanford, A. K. (2011). RTI for English language learners: Appropriately using screening and progress monitoring tools to improve instructional outcomes.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. Behavior Research Methods, 49, 1,520–1,523.

Buckingham, B. R., & Dolch, E. W. (1936). A combined word list. New York, NY: Ginn and Company.

Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A. G. Kamhi (Eds.), The connections between language and reading disabilities (pp. 25–40). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Crain, S., Shankweiler, D., Macaruso, P., & Bar-Shalom, E. (1990). Working memory and comprehension of spoken sentences: Investigations of children with reading disorder. In G. Vallar & T. Shallice (Eds.), Neuropsychological impairments of short-term memory (pp. 477–508). Cambridge University Press.

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

Crawford, L. (2014). The role of assessment in a response to intervention model. Preventing School Failure: Alternative Education for Children and Youth, 58(4), 230–236. https://doi.org/10.1080/1045988X.2013.805711
Cutting, L. E., & Denckla, M. B. (2001). The relationship of rapid serial naming and word reading in normally developing readers: An exploratory model. Reading and Writing, 14(7), 673-705.

DeBell, M., & Krosnick, J. A. (2009). Computing weights for American National Election Study survey data. ANES Technical Report series, no. nes012427. https://electionstudies.org/wp-content/uploads/2018/04/nes012427.pdf
de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. Scientific Studies of Reading, 6, 51–77.

Denckla, M. B., & Rudel, R. G. (1976). Rapid 'automatized' naming (RAN): Dyslexia differentiated from other learning disabilities. Neuropsychologia, 14(4), 471-479.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. Journal of Speech, Language, and Hearing Research, 41(5), 1136-1146.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach - assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23(4), 355–368.

Dorans, N. & Holland, P.W. (1993): DIF detection and description: Mantel–Haenszel and standardization. In Holland, P.W. and Wainer, H., editors, Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum, 35–66.

Duke, N. K., & Cartwright, K. B. (2021). The science of reading progresses: Communicating advances beyond the simple view of reading. Reading Research Quarterly, 56, S25-S44.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? Educational Psychology Review, 23(4), 553–576.

Frith, U. (1995). Dyslexia: Can we have a shared theoretical framework? Educational and Child Psychology, 12, 6–17.

Fry, E. (2004). The vocabulary teacher's book of lists. San Francisco, CA: Jossey-Bass.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. Scientific Studies of Reading, 5(3), 239–256.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). Bayesian data analysis. London: Chapman & Hall.

Georgiou, G. K., Parrila, R., & Papadopoulos, T. C. (2008). Predictors of word decoding and reading fluency across languages varying in orthographic consistency. Journal of Educational Psychology, 100(3), 566–580.

Goswami, U., & Bryant, P. (1990). Phonological skills and learning to read. Hove, UK: Erlbaum.

Graf Estes, K., Evans, J. L. & Else-Quest, N. M. 2007. Differences in nonword repetition performance of children with and without Specific Language Impairment: A meta-analysis. Journal of Speech, Language, and Hearing Research, 50: 177–195.

Hagtvet, B. (2003). Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills as well as phonological skills. Reading and Writing, 16, 505–539.

Hasbrouck, J., & Tindal, G. (2005). Oral reading fluency: 90 years of measurement. Technical report #33. Behavioral Research and Teaching, University of Oregon.

Hasbrouck, J., & Tindal, G. (2017). An update to compiled ORF norms. Behavioral Research and Teaching.

Ho, C. S.-H. (2014). Preschool predictors of dyslexia status in Chinese first graders with high or low familial risk. Reading and Writing: An Interdisciplinary Journal, 27(9), 1,673–1,701.

Ho, C. S.-H., Chow, B. W-Y., Wong, S. W-L., Waye, M. M. Y., & Bishop, D. V. M. (2012). The genetic and environmental foundation of the simple view of reading in Chinese. PLoS ONE, 7(10): e47872, https://doi.org/10.1371/journal.pone.0047872

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and Writing: An Interdisciplinary Journal, 2(2), 127–160.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how. Reading Teacher, 58(8), 702–714.

Ise, E. S. G. (2010). Spelling deficits in dyslexia: Evaluation of an orthographic spelling training. Annals of Dyslexia, 60(1), 18–39.

Jones, M. W., Branigan, H. P., & Kelly, M. L. (2008). Visual deficits in developmental dyslexia: relationships between non-linguistic visual tasks and their contribution to components of reading. Dyslexia, 14(2), 95-115.

Joshi, R. M., Ji, X. R., Breznitz, Z., Amiel, M., & Yulia, A. (2015). Validation of the simple view of reading in Hebrew—a Semitic language. Scientific Studies of Reading, 19, 243–252.

Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? Literacy, 42, 75–82.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking. New York: Springer.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44, 978–990.

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

Mantel N., & Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719–748.

McBride-Chang, C. (1999). The ABCs of the ABCs: The development of letter-name and letter- sound knowledge. Merrill-Palmer Quarterly, 45(2), 285–308.

McLaughlin, S. (1998). Introduction to language development. San Diego: Singular Publishing Group, Inc.

Melby-Lervåg, M., & Lervåg, A. (2012). Oral language skills moderate nonword repetition skills in children with dyslexia: A meta-analysis of the role of nonword repetition skills in dyslexia. Scientific Studies of Reading, 16(1), 1-34.

Melloni, C., Vender, M., Babatsouli, E., & Ball, M. J. (2020). Phonological processing and nonword repetition: a critical tool for the identification of dyslexia in bilingualism. An anthology of bilingual child phonology. Bristol, UK: Multilingual Matters.

Meng, X., Sai, X., Wang, C., Wang, J., Sha, S., & Zhou, X. (2005). Auditory and speech processing and reading development in Chinese school children: Behavioural and ERP evidence. Dyslexia, 11(4), 292–310.

Mostow, J. (2016). Project LISTEN's reading tutor. In S. A. Crossley & D. S. McNamara (Eds.), Adaptive educational technologies for literacy instruction (pp. 263–267). NY: Taylor & Francis, Routledge.

Mostow, J., Aist, G., Huang, C., Junker, B., Kennedy, R., Lan, H., Latimer, D., O'Connor, R., Tassone, R., Tobin, B., & Wierman, A. (2008). 4-Month evaluation of a learner-controlled reading tutor that listens. In V. M. Holland & F. P. Fisher (Eds.), The path of speech technologies in computer assisted language learning: From research toward practice (pp. 201–219). New York: Routledge.

Mostow, J., Nelson, J., & Beck, J. E. (2013). Computer-guided oral reading versus independent practice: Comparison of sustained silent reading to an automated reading tutor that listens. Journal of Educational Computing Research, 49(2), 249–276.

National Center on Intensive Intervention (NCII). (n.d.). Intensive intervention & multi-tiered system of supports (MTSS). Retrieved October 1, 2020, from https://intensiveintervention.org/special--pics/mtss.

National Center on Intensive Intervention (NCII) (2020a). Academic screening tools chart rating Rubrics. https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_ 2020-06-30.pdf

National Center on Intensive Intervention (NCII) (2020b). Call for submissions of academic screening tools. American Institutes for Research (AIR). https://intensiveintervention.org/sites/default/files/NCII_AcadScreen_CallForSubmissions_2 020-06-30.pdf

Neuhaus, G., Foorman, B. R., Francis, D. J., & Carlson, C. D. (2001). Measures of information processing in rapid automatized naming (RAN) and their relation to reading. Journal of Experimental Child Psychology, 78(4), 359–373.

Owens, R. E. (2004). Language development: An introduction (4th ed.). Boston, MA: Allyn and Bacon. Pennington, B. F., & Lefly, D. L. (2001). Early reading development in children at family risk for dyslexia. Child Development, 72(3), 816–833.

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. Cognition 101, 385–413.

Powell, D., Stainthorp, R., & Stuart, M. (2014). Deficits in orthographic knowledge in children poor at rapid automatized naming (RAN) tasks? Scientific Studies of Reading, 18(3), 192–207.

Protopapas, A., Simos, P. G., Sideridis, G. D., & Mouzaki, A. (2012). The components of the simple view of reading: A confirmatory factor analysis. Reading Psychology, 33(3), 217– 240.

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. Brain, 126(4), 841–865.

Rathvon, N. (2004). Early reading assessment: A practitioner's handbook. New York: Guilford Press.

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2;6 to 21 years of age. Journal of Speech, Language, and Hearing Research, 58(2), 345–359.

Roongpraiwan, R., Ruangdaraganon, N., Visudhiphan, P., & Santikul, K. (2002). Prevalence and clinical characteristics of dyslexia in primary school. Journal of the Medical Association of Thailand, 85(4), 1,097–1,103.

Samuels, S. J. (2002). Reading fluency: Its development and assessment. In A. E. Farstrup & S. J. Samuels (Eds.), What research has to say about reading instruction, 3rd ed., pp. 166–183. Newark, DE: International Reading Association.

Sander, E. K. (1972). When are speech sounds learned? Journal of Speech and Hearing Disorders, 37(1), 55–63.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), Handbook for research in early literacy (pp. 97–110). New York, NY: Guilford Press.

Secretaría de Educación Pública. (2014). Estándares nacionales de habilidad lectora. Estándares de lectura. Gobierno de México. https://www.gob.mx/sep/acciones-y-programas/estandares-nacionales-de-habilidad-lectora-estandares-de-lectura

Shankweiler, D., Eric Lundquist, E., Katz, L., Stuebing, K. K., Fletcher, J. M., Brady, S., Fowler, A., Dreyer, L. G., Marchione, K. E., Shaywitz, S. E., & Shaywitz, B. A. (1999). Comprehension and decoding: Patterns of association in children with reading difficulties. Scientific Studies of Reading, 3(1), 69–94.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. Cognition, 55, 151–218.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9–30). Baltimore, MD: Johns Hopkins University Press.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. Educational and Psychological Measurement, 70(3), 357–375.

Snowling, M. J. (2006). Nonword repetition and language learning disorders: A developmental contingency framework. Applied Psycholinguistics, 27: 588–591.

Stahl, S. A. (1999). Vocabulary development. Brookline, MA: Brookline Books.

Stanovich, K. E. (2000). Progress in understanding reading: Scientific foundations and new frontiers. New York: Guilford Press.

Tejas LEE. (2010). Welcome. http://www.tejaslee.org/

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (Eds.). (2001). Test scoring. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Tobia, V., & Bonifacci, P. (2015). The simple view of reading in a transparent orthography: The stronger role of oral comprehension. Reading and Writing, 28, 939–957.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1994). Longitudinal studies of phonological processing and reading. Journal of Learning Disabilities, 27(5), 276–286.

van Bergen, E., de Jong, P. F., Plakas, A., Maassen, B., & van der Leij, A. (2012). Child and parental literacy levels within families with a history of dyslexia. Journal of Child Psychology and Psychiatry, 53, 28–36.

van der Leij, A., van Bergen, E., van Zuijen, T., de Jong, P. F., Maurits, N., & Maassen, B. (2013). Precursors of developmental dyslexia: An overview of the longitudinal Dutch dyslexia programme study. Dyslexia, 19, 191–213.

van der Leij, A., & van Daal, V. H. P. (1999). Automatization aspects of dyslexia: Speed limitations in word identification, sensitivity to increasing task demands, and orthographic compensation. Journal of Learning Disabilities, 32(5), 417–428.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D.M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? Journal of Child Psychology and Psychiatry, 45, 2–40.

Verhoeven, L., & van Leeuwe, J. (2012). The simple view of second language reading throughout the primary grades. Reading and Writing, 25(8), 1,805–1,818.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal relations with reading. Psychological Bulletin, 101, 192-212.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). Comprehensive Test of Phonological Processing– 2nd ed (CTOPP-2). Austin: Pro-Ed.

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. Journal of educational psychology, 91(3), 415.

Yuill, N., & Oakhill, J. (1991). Children's problems in text comprehension: An experimental investigation. Cambridge University Press.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The educator's word guide. New York, NY: Touchstone Applied Science Associates, Inc.

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 Psychology Association and 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 Spanish Screener

Figure A1: Amira ISIP Spanish Partners

Appendix B

Spanish Sub-measure

Table B1: Benchmarks for the Spanish WCPM Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	0 – 0.24	0.25 – 0.76	0.77 – 33.53
Kindergarten	Winter	0 – 0.26	0.27 – 0.81	0.82 – 46.24
Kindergarten	Spring	0 – 0.32	0.33 – 1	8.1 – 61.95
1st Grade	Fall	0 – 0.37	0.39 – 21.18	22.4 – 73.68
1st Grade	Winter	0 – 0.48	0.5 – 30.92	31.85 – 86.5
1st Grade	Spring	0 – 0.77	0.8 – 45.06	46.01 – 104.16
2nd Grade	Fall	0 – 0.74	0.77 – 42.46	42.96 – 94.32
2nd Grade	Winter	0 – 2.22	3.45 – 54.54	55.29 – 106.11
2nd Grade	Spring	0 – 15.73	17.36 – 69.36	70.74 – 121.4
3rd Grade	Fall	0 – 9.85	11.27 – 57.33	58.19 – 108.73
3rd Grade	Winter	0 – 18.56	20.27 – 68.07	69.21 – 119.11
3rd Grade	Spring	0 – 27.26	29.27 – 78.81	80.23 – 129.49
4th Grade	Fall	0 – 28.2	29.27 – 69.89	70.53 – 117.89
4th Grade	Winter	0 – 31.57	32.78 – 82.81	83.6 – 129.53
4th Grade	Spring	0 – 34.93	36.29 – 95.72	96.66 – 141.18
5th Grade	Fall	0 – 31.57	32.78 – 82.81	84.04 – 139.68
5th Grade	Winter	0 – 34.93	36.29 – 95.72	97.11 – 151.33
5th Grade	Spring	0 – 38.3	39.79 – 108.64	110.17 – 162.97
6th Grade	Fall	0 – 31.57	32.78 – 82.81	84.04 – 139.68
6th Grade	Winter	0 – 34.93	36.29 – 95.72	97.11 – 151.33
6th Grade	Spring	0 – 38.3	39.79 – 108.64	110.17 – 162.97

Table B2: Benchmarks for Spanish Decoding (Alphabetic Knowledge) Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	0 – 0.24	0.25 – 0.76	0.77 – 88.62
Kindergarten	Winter	0 – 0.76	0.79 – 2.41	2.44 – 100
Kindergarten	Spring	0 – 1.51	1.58 – 4.8	4.87 – 100
1st Grade	Fall	0 – 0.26	0.27 – 0.81	0.82 – 100
1st Grade	Winter	0 – 0.78	0.81 – 2.46	2.49 – 100
1st Grade	Spring	0 – 3.03	3.16 – 87.09	88.05 – 100
2nd Grade	Fall	0 – 0.58	0.6 – 82.47	83.46 – 100
2nd Grade	Winter	0 – 2.76	2.88 – 91.76	92.3 – 100
2nd Grade	Spring	0 – 28.06	33.21 – 94.69	94.87 – 100
3rd Grade	Fall	0 – 17.87	24.76 – 92.11	92.59 – 100
3rd Grade	Winter	0 – 33.06	37.25 – 94.02	94.35 – 100
3rd Grade	Spring	0 – 51.9	53.81 – 95.93	96.12 – 100
4th Grade	Fall	0 – 48.05	50.23 – 93.3	93.57 – 100
4th Grade	Winter	0 – 61.2	62.86 – 94.47	94.72 – 100
4th Grade	Spring	0 – 74.35	75.49 – 95.64	95.87 – 100
5th Grade	Fall	0 – 72.13	73.3 – 95.99	96.14 – 100
5th Grade	Winter	0 – 75.36	76.4 – 96	96.15 – 99.93
5th Grade	Spring	0 – 78.59	79.51 – 96.01	96.15 – 99.86
6th Grade	Fall	0 – 72.13	73.3 – 95.99	96.14 – 100
6th Grade	Winter	0 – 75.36	76.4 – 96	96.15 – 99.93
6th Grade	Spring	0 – 78.59	79.51 – 96.01	96.15 – 99.86

Table B3: Benchmarks for the Spanish HFW Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	0 – 0.24	0.25 – 0.76	0.77 – 91.36
Kindergarten	Winter	0 – 0.76	0.79 – 2.41	2.44 – 91.61
Kindergarten	Spring	0 – 1.53	1.6 – 4.87	4.93 – 97.21
1st Grade	Fall	0 – 1.13	1.18 – 3.6	3.65 – 105.89
1st Grade	Winter	0 – 1.1	1.14 – 47.89	47.98 – 108.83
1st Grade	Spring	0 – 2.19	2.29 – 95.77	95.97 – 111.77
2nd Grade	Fall	0 – 1.85	1.93 – 99.14	99.24 – 99.9
2nd Grade	Winter	0 – 19.93	22.4 – 99.35	99.41 – 99.9
2nd Grade	Spring	0 – 39.86	44.79 – 99.56	99.58 – 99.9
3rd Grade	Fall	0 – 40.23	45.2 – 99.58	99.6 – 99.99
3rd Grade	Winter	0 – 61.43	65.21 – 99.62	99.64 – 99.99
3rd Grade	Spring	0 – 82.63	85.22 – 99.66	99.67 – 99.99
4th Grade	Fall	0 – 85.54	87.86 – 99.52	99.54 – 99.87
4th Grade	Winter	0 – 85.54	87.86 – 99.52	99.54 – 99.91
4th Grade	Spring	0 – 85.54	87.86 – 99.52	99.54 – 99.94
5th Grade	Fall	0 – 76.06	76.35 – 96.88	97.17 – 99.99
5th Grade	Winter	0 – 86.31	86.63 – 98.43	98.57 – 100.06
5th Grade	Spring	0 – 96.55	96.91 – 99.97	99.97 – 100.13
6th Grade	Fall	0 – 76.06	76.35 – 96.88	97.17 – 99.99
6th Grade	Winter	0 – 86.31	86.63 – 98.43	98.57 – 100.06
6th Grade	Spring	0 – 96.55	96.91 – 99.97	99.97 – 100.13

Table B4: Benchmarks for the Spanish Phonological Awareness Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	0 – 0.24	0.25 – 0.76	0.77 – 88.22
Kindergarten	Winter	0 – 0.76	0.79 – 2.41	2.44 – 100
Kindergarten	Spring	0 – 1.51	1.58 – 4.8	4.87 – 100
1st Grade	Fall	0 – 0.22	0.23 – 0.71	0.72 – 100.69
1st Grade	Winter	0 – 3.85	4.02 – 43.48	44 – 100.35
1st Grade	Spring	0 – 7.7	8.03 – 86.95	87.99 – 100
2nd Grade	Fall	0 – 0.58	0.6 – 82.76	83.64 – 100
2nd Grade	Winter	0 – 2.76	2.88 – 90.62	91.07 – 100
2nd Grade	Spring	0 – 24.67	29.14 – 94.12	94.36 – 100
3rd Grade	Fall	0 – 18.03	24.57 – 91.64	92.09 – 100
3rd Grade	Winter	0 – 34.01	38.73 – 93	93.36 – 100
3rd Grade	Spring	0 – 49.99	52.89 – 94.37	94.63 – 100
4th Grade	Fall	0 – 48.14	49.9 – 92.77	93.15 – 100
4th Grade	Winter	0 – 57.96	60.91 – 95.07	95.29 – 100
4th Grade	Spring	0 – 73.48	75.07 – 95.13	95.29 – 100
5th Grade	Fall	0 – 71.51	72.89 – 94.93	95.07 – 99.92
5th Grade	Winter	0 – 73.75	74.93 – 95.64	95.79 – 99.96
5th Grade	Spring	0 – 75.99	76.98 – 96.35	96.51 – 100
6th Grade	Fall	0 – 71.51	72.89 – 94.93	95.07 – 99.92
6th Grade	Winter	0 – 73.75	74.93 – 95.64	95.79 – 99.96
6th Grade	Spring	0 – 75.99	76.98 – 96.35	96.51 – 100

Table B5: Benchmarks for the Spanish Vocabulary Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	0 – 0.24	0.25 – 0.76	0.77 – 4744.92
Kindergarten	Winter	0 – 0.51	0.53 – 1.6	1.63 – 4828.81
Kindergarten	Spring	0 – 0.92	0.96 – 2.92	2.96 – 6624.01
1st Grade	Fall	0 – 0.26	0.27 – 0.83	0.84 – 6328.8
1st Grade	Winter	0 – 1.64	1.72 – 2983.01	3015.85 – 6501.43
1st Grade	Spring	0 – 3.03	3.16 – 5965.2	6030.87 – 7084.87
2nd Grade	Fall	0 – 0.92	0.96 – 8017.6	8059.53 – 9031.07
2nd Grade	Winter	0 – 934.72	1083.5 – 8040.1	8092.27 – 9478.44
2nd Grade	Spring	0 – 1515.73	1650.42 – 8062.6	8125 – 9925.8
3rd Grade	Fall	0 – 2399.41	2597.87 – 8551.48	8579.75 – 9440.73
3rd Grade	Winter	0 – 3295.2	3531.97 – 8625.21	8662.23 – 9691.53
3rd Grade	Spring	0 – 4191	4466.07 – 8698.93	8744.71 – 9942.33
4th Grade	Fall	0 – 3203.73	3288.39 – 8222.63	8287.81 – 9254.63
4th Grade	Winter	0 – 3835.7	3975.2 – 8924.28	8975.68 – 9676.02
4th Grade	Spring	0 – 4467.68	4662.01 – 9625.94	9663.56 – 10097.41
5th Grade	Fall	0 – 2720.6	2732.9 – 6502.31	6615.81 – 9528.19
5th Grade	Winter	0 – 4485.19	4568.93 – 8102.71	8173.73 – 9930.57
5th Grade	Spring	0 – 6264.78	6412.46 – 9793.15	9813.04 – 10082.52
6th Grade	Fall	0 – 2720.6	2732.9 – 6502.31	6615.81 – 9528.19
6th Grade	Winter	0 – 4485.19	4568.93 – 8102.71	8173.73 – 9930.57
6th Grade	Spring	0 – 6264.78	6412.46 – 9793.15	9813.04 – 10082.52

Table B6: Benchmarks for the Spanish Lexile Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	-400 – -395.44	-395.24 – -385.73	-385.54 – -245.73
Kindergarten	Winter	-200 – -197.72	-197.62 – -189.58	-185.93 – 137.62
Kindergarten	Spring	-200 – -128.12	-125 – 28.12	31.25 – 520.96
1st Grade	Fall	-400 – -397.27	-397.16 – -391.47	-391.35 – 110.79
1st Grade	Winter	-200 – -198.64	-198.58 – -12.4	-7.53 – 343.94
1st Grade	Spring	-200 – -87.3	-82.4 – 366.67	376.28 – 577.09
2nd Grade	Fall	-400 – -395.52	-395.33 – 591.43	605.87 – 847.44
2nd Grade	Winter	-200 – -197.76	-197.24 – 656.41	666.76 – 853.36
2nd Grade	Spring	-200 – -16	-8 – 721.38	727.65 – 859.28
3rd Grade	Fall	-402.42 – -225.06	-198.49 – 748.95	756.3 – 874.71
3rd Grade	Winter	-201.69 – -2.21	23.94 – 780.67	786.38 – 905.82
3rd Grade	Spring	-0.95 – 220.64	246.37 – 812.38	816.47 – 936.92
4th Grade	Fall	-482.86 – 129.13	160.62 – 853.81	858.58 – 948.11
4th Grade	Winter	-247.88 – 191.32	221.29 – 865.04	870.02 – 971.02
4th Grade	Spring	-12.9 – 253.5	281.97 – 876.27	881.47 – 993.93
5th Grade	Fall	-335.14 – 167.37	183.25 – 820.02	832.38 – 1080.7
5th Grade	Winter	-223.05 – 348.71	366.48 – 899.33	907.49 – 1077.07
5th Grade	Spring	-110.96 – 530.05	549.7 – 978.63	982.6 – 1073.45
6th Grade	Fall	-335.14 – 167.37	183.25 – 820.02	832.38 – 1080.7
6th Grade	Winter	-223.05 – 348.71	366.48 – 899.33	907.49 – 1077.07

Grade	Window	<=24th	25th-74th	>=75th
6th Grade	Spring	-110.96 – 530.05	549.7 – 978.63	982.6 – 1073.45

Note: The scale for this sub-measure ranges from -400 to 100.

Table B7: Benchmarks for the Spanish DRA Score

Grade	Window	<=24th	25th-74th	>=75th
Kindergarten	Fall	1 – 1.02	1.02 – 1.73	2.23 – 5.43
Kindergarten	Winter	1 – 1.02	1.02 – 1.73	2.47 – 16.29
Kindergarten	Spring	1 – 1.02	1.02 – 1.73	2.47 – 27.15
1st Grade	Fall	1 – 1.27	1.29 – 1.87	1.88 – 16.68
1st Grade	Winter	1 – 1.42	1.44 – 12	12.28 – 23.61
1st Grade	Spring	1 – 1.56	1.59 – 22.13	22.67 – 30.55
2nd Grade	Fall	1 – 1.04	1.05 – 32.67	33.33 – 40
2nd Grade	Winter	1 – 1.33	1.43 – 35.47	36 – 40
2nd Grade	Spring	1 – 1.67	1.87 – 38.27	38.67 – 40
3rd Grade	Fall	1 – 1.04	1.05 – 32.67	33.33 – 40
3rd Grade	Winter	1 – 1.33	1.43 – 35.47	36 – 40
3rd Grade	Spring	1 – 1.67	1.87 – 38.27	38.67 – 40
4th Grade	Fall	1 – 1.04	1.05 – 32.67	33.33 – 40
4th Grade	Winter	1 – 1.33	1.43 – 35.47	36 – 40
4th Grade	Spring	1 – 1.67	1.87 – 38.27	38.67 – 40
5th Grade	Fall	1 – 15.73	16.4 – 41.33	42 – 60
5th Grade	Winter	1 – 22.91	23.71 – 47	47.67 – 61.42
5th Grade	Spring	1 – 30.09	31.02 – 52.67	53.33 – 62.83
6th Grade	Fall	1 – 15.73	16.4 – 41.33	42 – 60
6th Grade	Winter	1 – 22.91	23.71 – 47	47.67 – 61.42
6th Grade	Spring	1 – 30.09	31.02 – 52.67	53.33 – 62.83

Appendix C

Criteria for Evaluating Item Quality

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 distracters Plausible distractors: each incorrect option chosen by ≥ 5-10%

Criteria Category	Evaluation Criteria
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
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 D

Amira ISIP Task and Time

Screen in Around Twenty Minutes

✔ = task is fully supported

o = task can be added, subject to adaptivity

(Time estimate, not same for every student)

AMIRA

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	