Imagine+ EarlyBird Screener Technical Report

Imagine Learning Assessment Research

January 8, 2026



Table of contents

Introduction **3**

About the Imagine+ EarlyBird Screener 3

1 Test Design and Development **5**

1.1 Background 5

1.2 Mastering the Alphabetic Principle 6

1.3 Overview of Subtests 8

1.4 Design of the Assessments 11

2 Scaling and Initial Validation Studies **16**

2.1 Reach Every Reader Validation Study 16

2.2 Boston Children's Hospital Validation Study 45

2.3 Florida State University Validation Study 54

2.4 EarlyBird 2022 Validation Study 59

3 Score Reporting **67**

3.1 Reading Readiness (Grade Pre-K) 67

3.2 Potential For Word Reading (Grade K, 1, 2, and 3) 68

3.3 Subtest Score Percentiles (Grade Pre-K, K, 1, 2, and 3) 69

3.4 Standard Scores 69

3.5 Developmental Scale Scores 69

3.6 Ratios (Grade Pre-K, K, 1) 70

References **71**

Introduction

About the Imagine+ EarlyBird Screener

We aimed to develop an assessment for the identification of children at risk for atypical reading and language skills in Pre-K through Grade 3. Specifically, the Imagine+ EarlyBird Screener (formerly known as EarlyBird Literacy/Dyslexia Screener) identifies severe word reading issues including risk of dyslexia and deficits in foundational reading skills based on the science of reading. The screener addresses literacy milestones that have been found to be predictive of subsequent reading success in each grade, and multiple validation studies provide support for these predictive measures, all of which are described in the Scaling and Initial Validation Studies chapter of this report. The Imagine+ EarlyBird Screener includes screening at the beginning, middle, and end of the school year for low and moderate expected word reading performance at the end of the school year.

The Imagine+ EarlyBird Screener uses a gamified mobile app that is easy, quick, accessible, and child-centered. It is self-administered in small groups with teacher oversight, and no trained adult is needed to administer the Imagine+ EarlyBird Screener app. The subtests needed for the screening risk score take no more than ten minutes, with the full comprehensive assessment taking around 15 to 45 minutes depending on the grade level and season the test is administered. Scoring is largely automated with most results available immediately upon completion of the assessment.

In the game, the child views a map of a city and is told that they can go on a journey in order to reach the pond to sail their toy sailboat. The child is paired with a feathery friend, named Pip, who will travel with them and act as a guide as they meet new animal friends, who demonstrate each subtest before the child attempts it. At the end of each game, the child is rewarded with a virtual prize and travels farther along the path, getting closer to their final destination at the pond. When the child finishes the game, a score report is created on the teacher's web-based dashboard.

This technical manual describes the research, validation studies, and development of the Imagine+ EarlyBird Screener for Pre-K through Grade 3. Several consensus documents review decades of literature about what predicts reading success (National Institute of Child Health and Human Development, 2000; National Research Council, 1998; RAND Reading Study Group, 2002; Rayner et al., 2001). This technical report adheres to the guidance provided in the *Standards for Educational and Psychological Testing* (hereafter *Standards*), which call for test developers to document the procedures used to develop and maintain assessments as well as technical characteristics of their assessments based on empirical data (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Each section of this technical report introduces the relevant chapter(s) of the *Standards* that the section addresses.

Imagine Learning recommends using multiple measures to make informed decisions on student learning. Imagine+ EarlyBird Screener scores are one of these measures. Whenever possible, Imagine+ EarlyBird Screener results should be combined with other sources of information (e.g., teacher observation, end-of-grade test scores) before making decisions about students. In particular, when decisions are high stakes (e.g., deciding to promote a student to the next grade level), it is critical that multiple measures of student performance be considered. Imagine Learning does not support the use of Imagine+ EarlyBird Screener for teacher evaluation.

Chapter 1

Test Design and Development

The *Standards* devote considerable attention to test design and development because they are crucial to an assessment's validity. Chapter 4 of the *Standards* is exclusively dedicated to test design and development, and other chapters make frequent reference to test development practices. In essence, the *Standards* specify that all phases of the test development life cycle should be designed to support the validity of the test based on the test's specific purpose and population, and test developers should thoroughly document test development procedures. This section of the Imagine+ EarlyBird Screener technical report provides such documentation.

1.1 Background

The development of basic reading skills is one major goal during the first years of elementary school. However, in the United States, 65% of fourth graders are not reading on grade-level according to studies conducted by the National Center for Education Statistics (McFarland et al., 2019), and it has been shown that 70% of children who are poor readers in third grade remain poor readers throughout their educational career (Foorman et al., 2013). Furthermore, difficulties with learning to read have been associated with a cascade of socioemotional difficulties in children, including low self-esteem, depression, and feelings of shame, inadequacy, and helplessness (Valas, 1999). Children with learning disabilities are less likely to complete high school and are increasingly at risk of entering the juvenile justice system (Mallett et al., 2011). Despite the cascade of negative consequences, most children are currently identified only after they fail over a significant period of time and outside of the window for most effective interventions, which has been termed the "dyslexia paradox" (Ozernov-Palchik et al., 2016; Ozernov-Palchik & Gaab, 2016). Research has shown that the most effective window for early reading interventions is in kindergarten and first grade (Wanzek & Vaughn, 2007), most likely even earlier. When at-risk beginning readers (across six research studies) received intensive reading instruction, 56% to 92% of students achieved average reading ability (Torgesen, 2004). Early literacy milestone screening moves this from a reactive to a proactive model and – if evidence-based response to screening is implemented – enables a preventive educational approach.

1.2 Mastering the Alphabetic Principle

What matters the most to success in reading words in an alphabetic orthography such as English is mastering the alphabetic principle, the insight that speech can be segmented into discrete units (i.e., phonemes) that map onto orthographic (i.e., graphemic) units (Ehri et al., 2001; Rayner et al., 2001). Oral language is acquired largely in a natural manner within a hearing/speaking community; however, written language is not acquired naturally because the graphemes and their relation to phonological units in speech are invented and must be taught by literate members of the community. The various writing systems (i.e., orthographies) of the world vary in the transparency of the sound-symbol relation. Among alphabetic orthographies, the Finnish orthography, for example, is highly transparent: phonemes in speech relate to graphemes in print (i.e., spelling) in a highly consistent one-to-one manner. Similarly graphemes in print relate to phonemes in speech (i.e., decoding) in a highly consistent one-to-one manner. Thus, learning to spell and read Finnish is relatively easy. English, however, is a more opaque orthography. Phonemes often relate to graphemes in an inconsistent manner and graphemes relate to phonemes in yet a different inconsistent manner. For example, if we hear the "long sound of a" we can think of words with many different vowel spellings, such as crate, brain, hay, they, maybe, eight, great, vein. If we see the orthographic unit –ough, we may struggle with the various pronunciations of cough, tough, though, bough. The good news is that 69% of monosyllabic English words — those Anglo-Saxon words most used in beginning reading instruction — are consistent in their letter to pronunciation mapping (Ziegler et al., 1997). Most of the rest can be learned with grapheme-phoneme correspondence rules (i.e., phonics), with only a small percentage of words being so irregular in their letter-sound relations that they should be taught as sight words (Ehri et al., 2001; Foorman et al., 2013).

In the Imagine+ EarlyBird Screener, the alphabetic principle is assessed with individually-administered tasks that measure letter-sound knowledge, word and nonword reading, and spelling.

1.2.1 Comprehending Written Language

Mastering the alphabetic principle is a necessary, but not sufficient, condition for understanding written text. We may be able to pronounce printed words, but if we don't know their meaning, our comprehension of any text is likely to be impeded significantly. Hence, our knowledge of word meanings is crucial to comprehending what we read. Grasping the meaning of a word is more than knowing its definition in a particular passage. Knowing the meaning of a word means knowing its full lexical entry in a dictionary: pronunciation, spelling, multiple meanings in a variety of contexts, synonyms, antonyms, idiomatic use, related words, etymology, and morphological structure. For example, a dictionary entry for the word exacerbate says that it is a verb meaning: 1) to increase the severity, bitterness, or violence of (disease, ill feeling, etc.); aggravate or 2) to embitter the feelings of (a person); irritate; exasperate (e.g., foolish words that only exacerbated the quarrel). It comes from the Latin word exacerbātus (the past participle of exacerbāre: to exasperate, provoke), equivalent to ex + acerbatus (acerbate). Synonyms are: intensify, inflame, worsen, embitter. Antonyms are: relieve, soothe, alleviate, assuage. Idiomatic equivalents are: add fuel to the flame, fan the flames, feed the fire, or pour oil on the fire. The more a reader knows about the meaning of a word like exacerbate, the greater the lexical quality the reader has and the more likely the reader will be able to recognize the

word quickly in text, with full comprehension of its meaning (Perfetti & Stafura, 2014). In the Imagine+ EarlyBird Screener, knowledge of word meanings is measured by a vocabulary task called Expressive Vocabulary and a semantic relationships task called Word Matching. Expressive Vocabulary asks the student to label pictures using nouns (e.g., "What is this?" "An island") or verbs (e.g., "What is he doing?" "Measuring"). During the Word Matching task, the child's task is to touch the two out of three words (or pictures, which are also presented orally) that go together (e.g., blue, triangle, yellow).

1.2.2 Oral Listening Comprehension / Syntactic Awareness

In addition to understanding word meanings, another important aspect of successful reading acquisition is the ability to understand complex sentences which includes morphological and syntactic awareness. Syntax or grammar refers to the rules that govern how words are ordered to make meaningful sentences. Children typically acquire these rules in their native language prior to formal schooling. However, learning to apply these rules to reading and writing is a goal of formal schooling and takes years of instruction and practice. In the Imagine+ EarlyBird Screener, the Following Directions task addresses oral listening comprehension / syntactic awareness. The Following Directions task requires that the student touch the objects on the screen as prescribed by the directions (e.g., click on the cat and then click on the heart; click on the book after clicking on the airplane; before clicking on the book, click on the smallest cat).

1.2.3 Reading Comprehension

If a student can read and understand the meanings of printed words and sentences, then comprehending text should not be difficult, given the emphasis above on achieving the alphabetic principle, lexical quality, and syntactic awareness. Individual differences in readers' background knowledge, motivation, and memory and attention will create variability in word recognition skills, vocabulary knowledge, and syntactic awareness and this variability, in turn, will create variability in reading comprehension. Furthermore, genre differences—informational or literary text—may interact with reader skills to affect reading comprehension. For example, some students may have better inferential language skills so critical to comprehending informational text; other students may have better narrative language skills of discerning story structure and character motivation and, therefore, be good comprehenders of literary text. Because reading comprehension is affected by the interactions of variables related to reader and text characteristics (RAND Reading Study Group, 2002), tests of reading comprehension typically consist of informational and literary passages and provide as much relevant background information within the passage as possible.

For students taking the Imagine+ EarlyBird Screener, there are Reading Comprehension passages available to administer as a supplemental subtest for grades 1 and 2. Because texts are necessarily short in length for primary-grade children, the number of questions the teacher can ask is limited to five. Given the limited number of questions asked and the dominance of other component reading skills predicting success in early reading, the reading comprehension questions are administered for qualitative/descriptive and formative purposes only. Narrative and informational texts are written with attention to the features of text complexity in primary-grade text, such as word structure, word meanings, and syntactic features within and across sentences (Foorman et al., 2013).

1.2.4 Etiology of Reading Difficulties

It is important to note that atypical reading development has a multifactorial etiology. Causes can be observed on biological, psychological, and/or environmental levels and the identification of children who exhibit atypical reading development requires multi-factorial strategies for screening and interventions (Catts & Petscher, 2020). Numerous longitudinal research studies (for an overview see Ozernov-Palchik & Gaab (2016)) have identified behavioral precursors of typical/atypical reading development. In general, research has established that successful reading acquisition requires the integration of the "mechanics" of reading (e.g., decoding skills which require letter sound knowledge and phonological awareness) and oral language skills (Scarborough, 2001). Early pre-literacy skills related to these two components have been shown to predict reading skills and these include phonological awareness, phonological memory, letter sound/name skills, rapid automatized naming, and oral language skills. The Imagine+ EarlyBird Screener incorporates all of these skills, as outlined below.

1.3 Overview of Subtests

1.3.1 Phonemic Awareness

Rhyming – Moose (Grade Pre-K & K): Rhyming is a computer adaptive task that presents three pictures at a time, naming each one. After the student listens to the three words, he or she identifies the two rhyming words by tapping the rhyming pictures. For example, "Which two words end with the same sound?" The words with pictures 'duck', 'man', and 'fan' are presented. After the student listens to the three words, he or she identifies the two rhyming words as 'man' and 'fan'.

First Sound Matching – Tiger (Grade Pre-K & K): First Sound Matching is a computer adaptive task that measures a student's ability to isolate and match the initial phonemes in words. This task presents one picture as a stimulus, asking the student to listen to the first sound in that word. Three additional pictures are presented asking the student to touch the picture with the matching first sound. For example, "This is a dog. Hand, toy, doll. Which one starts with the same sound as 'dog'?" The student touches the picture of the dog to identify the correct matching first sound.

Blending – Kangaroo (Grade Pre-K & K): The Blending task is a computer adaptive task that requires students to listen to a word that has been broken into parts and then blend them together to reproduce the full word. The items in this task include compound words, words that require blending of the onset and rime, and words requiring the blending of three or more phonemes (e.g., "What would the word be if I say: /h/ /orn/").

Nonword Repetition – Ostrich (Grade K, 1 & 2): Nonword Repetition is a computer adaptive task that presents sounds in a spoken word form for the student to listen to and repeat. This can be in the form of a one- to five-syllable nonsense word. The student hears phonemes in a sequence that they have not heard before and asked to repeat the sequence. For example, a student hears the word 'tav' and is asked to repeat. The student must rely on their phonological short-term memory to repeat the sequence correctly.

Deletion – Elephant (Grade K & 1): Deletion is a computer adaptive task that measures the ability to remove phonological segments from spoken words to form other

words. The items in this task include compound words (e.g., "Say ladybug" "Now, say ladybug without saying /bug/"), words that can be segmented into the onset and rime (e.g., "Say sit" "Now, say sit without saying /s/"), and words including three or more phonemes.

1.3.2 Phonics (including Alphabet Knowledge)

Letter Name – Crocodile (Grade Pre-K & K): Letter Name is a fixed form task that assesses the student's knowledge of the name of each letter in the alphabet. The letters are presented one at a time and are ordered from easiest to hardest, based on research. The student is asked to verbally provide the name of each letter, as it is shown.

Letter Sound – Giraffe (Grade Pre-K, K & 1): Letter Sound is a fixed form task that assesses the student's knowledge of the sound made by each letter in the alphabet, as well as three digraphs (CH, SH, TH). The letters are presented one at a time and are ordered from easiest to hardest, based on research. The student is asked to verbally provide the sound that each letter makes, as it is shown.

Nonword Reading – Walrus (Grade K, 1, 2 & 3): Nonword Reading is a computer adaptive task in which a decodable nonsense word is presented in lowercase on the screen, and the student attempts to read it aloud. These nonsense words range from vowel-consonant (VC) and consonant-vowel-consonant (VCC) to vowel-consonant-silent-e (VCe) words.

Nonword Spelling – Penguin (Grade 1 & 2): Nonword Spelling is a computer adaptive task in which the student uses letter tiles to spell verbally-presented nonsense words. Foil tiles (letters that are not part of the word) are included in the available letter bank, and the student can drag any number of letter tiles to the answer field. The target nonsense words range from one to two syllables, including CVC and VCe patterns, and include digraphs and blends.

1.3.3 Fluency

Rapid Automatized Naming (RAN) – Polar Bear (Grade Pre-K, K, 1, 2 & 3): In Pre-K and Kindergarten, the RAN task uses a set of five objects (house, door, cat, ear, bed) that are repeated in random order in four rows (totaling 40 objects) for Pre-K or five rows (totally 50 objects) for Kindergarten. The student is measured on how fast they are able to name each object out loud across each row. The number of seconds it takes for the student to name all 40 or 50 objects provides the data for the final score.

In first grade, the Letter RAN board has five uppercase letters (R, M, K, E, S) repeated in random order over five rows of ten items each. The second and third grade Letter RAN board has six lowercase letters (h, t, m, s, o, i) repeated in random order over five rows of ten items each. The student is measured on how quickly and accurately they are able to name the letters out loud across each row. The number of seconds it takes for the student to name all 50 letters provides the data for the final score.

Regardless of grade, the student's response is recorded to the dashboard and available to the teacher for later confirmation of time and accuracy.

Word Reading – Lion (Grade K, 1, 2 & 3): Word Reading is a computer adaptive task in which a word is presented on the screen, and the student attempts to read it

aloud accurately. These words represent a wide range of difficulty, from single-syllable to multi-syllable words, with a mix of decodable words and sight words.

Oral Reading Fluency – Supplemental Subtest (Grade 1, 2 & 3): Oral Reading Fluency is a fixed form, paper-and-pencil task, administered and scored by a teacher, and designed to assess the student's ability to accurately and fluently read connected text. The student is instructed to read the passage out loud as quickly as they can in one minute. The teacher marks any errors or omissions the student makes before the one-minute timer runs out, and records the total number of errors, total words read correctly, and total words read.

1.3.4 Vocabulary

Receptive Vocabulary – Alpaca (Pre-K): Vocabulary is a computer adaptive task that measures a student's receptive vocabulary skills. Students listen to one word and select which picture from a field of four choices best represents the word.

Expressive Vocabulary – Bison (Grade K, 1, 2 & 3): Expressive Vocabulary is a computer adaptive task that asks the child to verbally label pictures depicting either a noun (e.g., "What is this a picture of?" "An island") or a verb (e.g., "What is he doing?" "Measuring"). The items range in terms of familiarity and specificity.

Word Matching – Gorilla (Grade K, 1 & 2): Word Matching is a computer adaptive task that measures the ability to perceive relationships between words that are related by semantic class features. Three written words (or pictures) appear on the screen and are pronounced by the app. The student then selects the two words that go together best (e.g., "Fish, Moon, Sun: Which two go together best?").

1.3.5 Comprehension

Oral Sentence Comprehension – Rhino (Grade Pre-K & K): The Oral Sentence Comprehension task is a computer adaptive receptive syntactic measure in which the student selects the one picture out of the four presented on the screen that depicts the sentence given by the computer (e.g., "Click on the picture of the bird flying towards the nest").

Follow Directions – Zebra (Grade K, 1 & 2): The Follow Directions task is a computer adaptive task that requires students to listen to and interpret spoken directions of increasing length and complexity, remember the characteristics and order of mention of pictures, and identify the targeted pictures from among several choices. Items consist of an array of pictures (including foils) on the screen and a set of audio instructions. Students respond to the directions by touching the specified pictures on the screen, as instructed (e.g., "Click on the cat and then click on the heart.").

Reading Comprehension – Supplemental Subtest (Grade 1 & 2): Reading Comprehension is a fixed form, paper-and-pencil task, administered and scored by a teacher, and designed to assess the student's ability to understand a passage of written text. The student is instructed to read the passage and then answer the corresponding open questions from the teacher, which are a mix of literal and inferential questions to check for understanding. The teacher should take note of qualitative information (e.g., number of words read correctly) when interpreting the student's score.

Table 1.1 summarizes the subtests available in each grade.

Table 1.1: Summary of Subtests by Grade

Subtest	Pre-K	K	Grade 1	Grade 2	Grade 3
Rhyming	X	X			
First Sound Matching	X	X			
Blending	X	X			
Deletion		X	X		
Nonword Repetition		X	X	X	
Letter Name	X	X			
Letter Sound	X	X	X		
Nonword Reading		X	X	X	X
Nonword Spelling			X	X	
RAN	X	X	X	X	X
Word Reading		X	X	X	X
Oral Reading Fluency			X	X	
Receptive Vocabulary	X				
Expressive Vocabulary		X	X	X	X
Word Matching		X	X	X	
Oral Sentence Comprehension	X	X			
Follow Directions		X	X	X	
Reading Comprehension			X	X	

1.4 Design of the Assessments

The majority of Imagine+ EarlyBird Screener tasks are based on computer adaptive test (CAT) algorithms that leverage an item response theory (IRT) framework to optimally match students to items. Because IRT item difficulties and person ability estimates are co-located on the same scale, algorithms are able to move students through individual assessments according to their response on individual items within a task. Correct responses to items typically result in students being administered relatively more difficult items based on the student's ability whereas incorrect responses to items typically result in students being administered relatively easier items based on the student's ability. The advantage of CAT is that the student generally receives items that are never too difficult or too easy based on ability and tasks can be administered quickly to obtain reliable information.

1.4.1 Psychometric Concepts Used in Algorithm

As is common for most educational assessments, the Imagine+ EarlyBird Screener functions with IRT as the underlying mechanism to estimate student ability. In IRT, item difficulty and student ability can be estimated using logistic models. There are three common IRT models used, named accordingly to the number of parameters used in the model, as the one parameter, two parameter, or three parameter model. These models are typically denoted 1 PL, 2 PL, or 3 PL. The 1 PL model is synonymously referred to as the Rasch model.

The 3 PL model is:

$$P(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}} \tag{1.1}$$

where

θ is the student ability
a is the item discrimination parameter
b is the item difficulty parameter
c is the item pseudo-guessing parameter

In the 2 PL model, c is set to 0. In the 1 PL model, in addition to setting c to 0, a is set to 1. Thus, the 1 PL and 2 PL models are simplifications of the 3 PL model.

The Imagine+ EarlyBird Screener adaptive algorithm leverages the 2PL model. In this model, the key parameters are item difficulty (b) and item discrimination (a). The difficulty of an item has traditionally been described for many tests as a p-value, which corresponds to the percent of respondents correctly answering an item. Values from this perspective range from 0% to 100% with high values indicating easier items and low values indicating hard items. Item difficulty in an IRT model does not represent proportion correct, but is rather represented as estimates along a continuum of approximately -3.0 to +3.0.

Figure 1.1 demonstrates a sample item characteristic curve which describes item properties from IRT. Along the x-axis is the ability of the individual. As previously mentioned, the ability of students and item statistics are reported on the same scale. Thus, the x-axis is a simultaneous representation of student ability and item difficulty. Negative values along the x-axis will indicate that items are easier, while positive values describe harder items. Pertaining to students, negative values describe individuals who perform below average, while positive values identify students who perform above average. A value of zero for both students and items reflects average level of either ability or difficulty. Along the y-axis is the probability of a correct response, which varies across the level of difficulty. Item difficulty is defined as the value on the x-axis at which the probability of correctly endorsing the item is 0.50. As demonstrated for the sample item in Figure 1.1, the difficulty of this item would be 0.0. Item characteristic curves are graphical representations generated for each item that allow the user to see how the probability of getting the item correct changes for different levels of the x-axis. Students with an ability (θ) of -3.0 would have an approximate 0.01 chance of getting the item correct, while students with an ability of 3.0 would have a nearly 99% chance of getting an item correct.

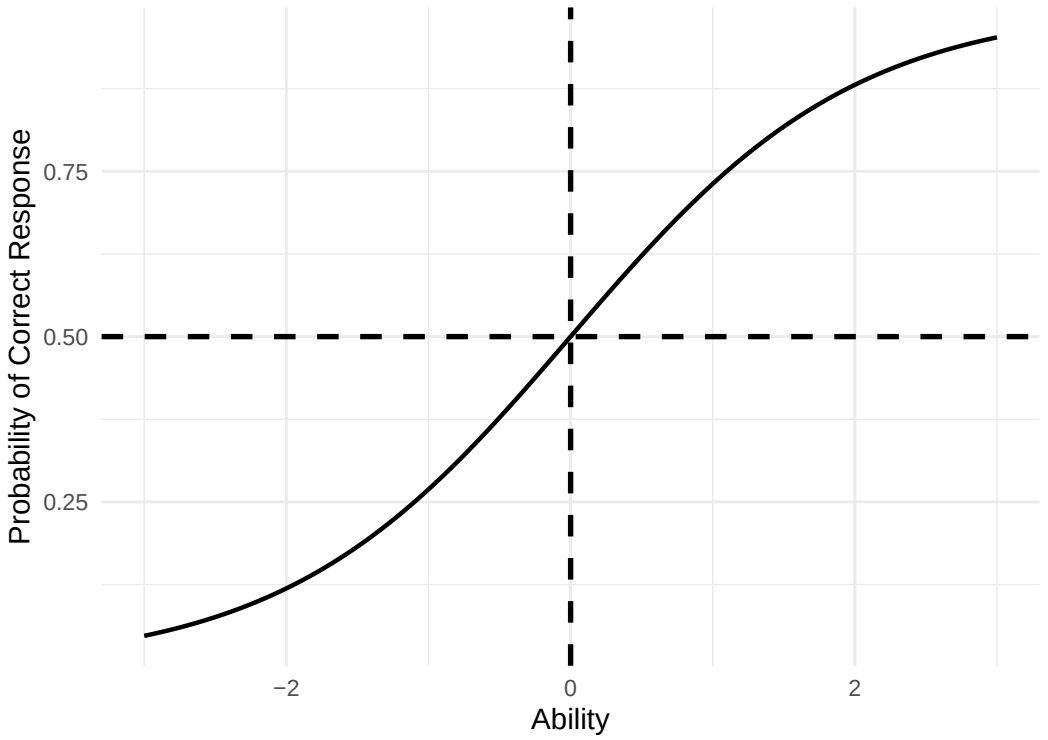


Figure 1.1: Sample Item Characteristic Curve

Item Discrimination (a) is related to the relationship between how a student responds to an item and their subsequent performance on the rest of a test. In IRT it describes the extent to which an item can differentiate the probability of correctly endorsing an item across the range of ability (i.e., -3.0 to +3.0). Figure 1.2 provides an example of how discrimination operates in the IRT framework. For all three items presented in Figure 1.2, the difficulty has been held constant at 0.0, while the discriminations are variable. The dashed line (Item 1) shows an item with strong discrimination, the solid line (Item 2) represents an item with acceptable discrimination, and the dotted line (Item 3) is indicative of an item that does not discriminate. It is observed that for Item 3, regardless of the level of ability for a student, the probability of getting the item right is the same. Both high ability students and low ability students have the same chance of doing well on this item. Item 1 demonstrates that as the x-axis increases, the probability of getting the item correct changes as well. Notice that small changes between -1.0 and +1.0 on the x-axis result in large changes on the y-axis. This indicates that the item discriminates well among students, and that individuals with higher ability have a greater probability of getting the item correct. Item 2 shows that while an increase in ability produces an increase in the probability of a correct response, the increase is not as large as is observed for Item 1, and is thus a poorer discriminating item.

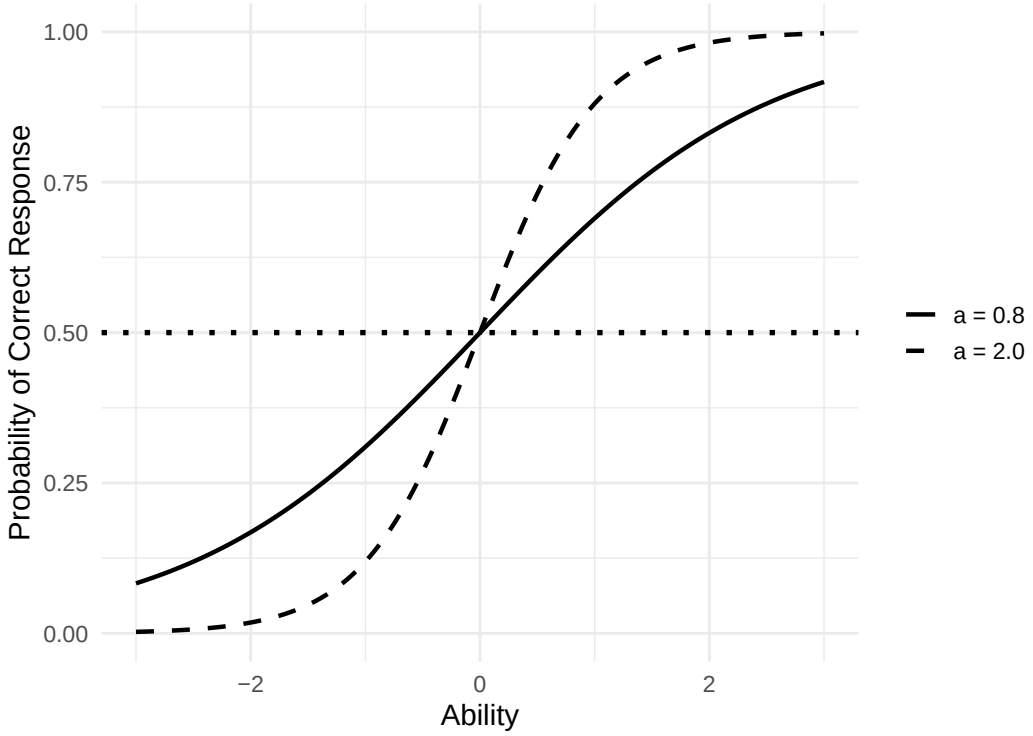


Figure 1.2: Sample Item Characteristic Curve

Within the IRT framework, the concept of *item information* refers, conceptually, to how much information an item provides about the ability level of a student. Generally, items will provide more information about a particular student's ability when the student has a 50% chance of answering the item correctly. In other words, items that are too easy or too hard for a student will not provide as much information as items that are "just right" in difficulty for the student. The Imagine+ EarlyBird Screener calculates item information according to the method known as maximum Fisher information. Mathematically, information from item i for a student of ability θ, denoted $I(\theta)$, is calculated as:

$$I(\theta) = \frac{P_i^*(\theta)^2}{P_i(\theta)(1 - P_i(\theta))} \tag{1.2}$$

where

$P_i^*(\theta)$ represents the first derivative of the item response curve evaluated at a particular ability level.

Once item information is calculated for each item, *test information* is calculated by summing the item information for every item the student received. Finally, the *standard error of measurement* (SEM) is calculated as the reciprocal of the square root of test information. Mathematically, this calculation is:

$$SEM(\theta) = \frac{1}{\sqrt{\sum_{n=1}^{I} I(\theta)}} \tag{1.3}$$

where I is the total number of items administered.

Conceptually, SEM indicates the level of precision in the student's score. Greater item information yields greater test information, which yields lower SEM and hence greater precision. Conversely, larger SEMs indicate less precision.

The Imagine+ EarlyBird Screener tasks are administered in the following ways: 1) the student is administered a set of five fixed items to calibrate their initial ability score; 2) the ability of the student after the first set of items is estimated along with the standard error of measurement (SEM) of ability; 3) the student SEM is compared to a target SEM threshold (associated with reliability = .80) where student SEM less than target SE results in the task terminating and moving to the next task; 4) when student SEM greater than target SEM, the student is administered another item according to | θ – b |. Steps 2-4 continue until the target SEM is reached or until a predetermined number of items have been administered.

For inclusion in the initial item pool, several criteria were used to evaluate item performance. Following the descriptive analysis of item performance, difficulty and discrimination values from the IRT analyses were used to further identify items which were poorly functioning. Items were flagged for item revision if the item discrimination was negative or the item difficulty was greater than +4.0 or less than -4.0. Secondary criteria were used in evaluating the retained items, which was comprised of a differential item function (DIF) analysis. DIF refers to instances where individuals from different groups with the same level of underlying ability significantly differ in their probability to correctly endorse an item. Unchecked, items included in a test which demonstrate DIF will produce biased test results. Items exceeding criteria for indications of DIF were flagged for further study and considered for removal.

Chapter 2

Scaling and Initial Validation Studies

The development and calibration of the initial item pool for the Imagine+ EarlyBird Screener were carried out through a rigorous, multi-phase research process spanning multiple institutions and geographic locations. The goal was to ensure that each item included in the final assessment was not only developmentally appropriate and psychometrically sound, but also predictive of later reading success. This process was anchored in best practices in psychometric theory, including classical test theory (CTT), IRT, and modern approaches to fairness and validity evaluation.

Three key research partners contributed to the item developing and scaling process. Boston Children's Hospital (BCH) contributed significantly to the development of Rhyming, First Sound Matching, Nonword Repetition, and Object RAN subtests, particularly for Kindergarten. The Word Matching, Oral Sentence Comprehension, and Follow Directions subtests were validated through efforts at Florida State University. Reach Every Reader (RER) further expanded the item pool by contributing items in areas such as Blending, Deletion, Nonword Repetition, Nonword Reading, Nonword Spelling, Word Reading, Letter RAN, and Expressive Vocabulary. Across many subtests and grade levels, additional internal research for validation and item calibration efforts were performed.

2.1 Reach Every Reader Validation Study

Note: Throughout this section, an earlier version of the Imagine+ EarlyBird Screener is referred to as the Interstellar Express Assessment System (IEAS). Several research studies, including the one described in this section, have informed the current version of the Imagine+ EarlyBird Screener.

Across the 2018-19 through 2022-23 academic school years, a common-item non-equivalent groups design was used for the purpose of vertical equating and scaling. In this design, the groups exposed to forms of items were not assumed to be equivalent in ability, however, the presence of common items across forms within and between grade levels provided data to serve as the statistical bridge needed to place scores from the different forms onto a common scale.

2.1.1 Procedures

2.1.1.1 Administration

Pre-COVID Assessment Administration

Before the pandemic, students completed all assessment tasks on iPads in one-on-one sessions with trained proctors. Proctors received in-person and written training, followed by at least two weeks of guided and independent practice. They were required to pass a fidelity test with 100% accuracy within three attempts to be approved for data collection. Assessments were conducted at school during teacher identified non-instructional time in Fall, Winter, and Spring. Schools provided quiet testing areas and proctors traveled with portable tables for when schools did not provide them. iPads were positioned upright to ensure a consistent viewing distance of approximately 18 inches from the student when seated. Proctors obtained verbal assent from each student before beginning the assessment. Students who withdrew assent were returned to class. After three refusals or withdrawals of assent, students were removed from the study. When a student was returned to class, they received a small incentive (e.g., sticker, eraser, pencil).

Post-COVID Assessment Administration

When schools closed due to COVID-19, the assessment platform was adapted for remote use. Developers enabled cross-platform functionality, allowing students to participate remotely on personal devices (e.g., iPad, tablet, PC, laptop) by connecting with the proctor's iPad. The addition of picture-in-picture video chat enabled real-time proctor/student interaction during the assessment. Remote testing required significant parental support, including scheduling appointments, downloading the app, and helping students connect at their scheduled time. Parents were asked to provide a quiet, distraction-free environment and not assist or coach students during the assessment. Unlike in-person testing, students did not receive incentives after each session. Instead, they received a set of age-appropriate books by mail upon completing all sessions.

2.1.1.2 Sample Plan

A total of 3,908 students across kindergarten through third grade served as participants across the five years of the study . The students were recruited from both public and private schools in the states of Florida (n = 2,206), Georgia (n = 256), South Carolina (n = 808), Massachusetts (n = 335), Oregon (n = 265), and online enrollment (unknown location, n = 38). Some students participated in the project across multiple years, resulting in data collected from more than one grade for 648 students. The overall sample by grade included 1,955 students assessed in kindergarten, 1,283 in first grade, 931 in second grade, and 451 in third grade. The grade-level distribution of participants by state was as follows:

- Florida – Kindergarten: 878, Grade 1: 826, Grade 2: 604, Grade 3: 295
- Georgia – Kindergarten: 240, Grade 1: 10, Grade 2: 6
- Massachusetts – Kindergarten: 124, Grade 1: 124, Grade 2: 74, Grade 3: 32
- Oregon – Kindergarten: 240, Grade 1: 16, Grade 2: 19, Grade 3: 1
- South Carolina – Kindergarten: 465, Grade 1: 292, Grade 2: 215, Grade 3: 117
- Online Enrollment – Kindergarten: 8, Grade 1: 15, Grade 2: 13, Grade 3: 6

Through a series of partnerships with state agencies (i.e., Florida and South Carolina) and the University of California – San Francisco, we implemented computer adaptive versions of the IEAS in kindergarten, grade 1 and grade 3. Approximately 1,830 students participated in the Florida implementation studies (Kindergarten N = 858, Grade 1 N = 972); 980 students participated in the South Carolina implementation study (Kindergarten N = 325, Grade 1 N = 397, Grade 3 N = 258); and 2835 students participated in the California implementation study (Kindergarten N = 1396, Grade 1 N = 1439).

2.1.2 Description of Subtests

The subtests used in the validation study vary slightly from the subtests in the current operational version of the Imagine+ EarlyBird Screener. The subtests used in the study include:

Letter Naming Fluency: This static measure assesses the number of letters students can name correctly in one minute. A sample of five letters is presented to the students before they begin the task. Students are then presented with one or more pages of upper- or lower-case letters in random order across five rows of ten items each and are asked to name each letter as quickly as possible. The number of letters named correctly in one minute is calculated.

Letter Sound Fluency: This static measure assesses the number of letters for which a student can provide the corresponding sound in one minute. A sample of five letters is presented to the students before they begin the task. Students are then presented with one or more pages of upper- or lower-case letters in random order across five rows with ten letters each and are asked to provide the corresponding sound as quickly as possible. The number of correct letter sounds provided in one minute is calculated. The parallel form reliability of this task is .90.

Rapid Automatized Naming - Objects (RAN-O): This task measures the time it takes for students to name a repeated set of six familiar objects. These six items include: book, cake, chair, fish, leaf, and sun. The items are presented randomly in an array of five items across five rows over two pages. The score is the time it takes (in seconds) to name the items on both pages.

Rapid Automatized Naming – Letter (RAN-L): This task measures the time it takes for students to name a repeated set of six letters. These six letters are: h m o i t s. The letters are presented randomly in an array of five items across five rows over two pages. In total, students label a total of 50 letters. The score is the time it takes (in seconds) to name the items on both pages.

Word Reading (WRE): Students are presented with printed words displayed individually on an iPad and prompted to read them. Prior to beginning this task, students are provided with a demonstration item and a practice item. There is a total of 311 items for this task. To select words for this task, several lists and databases were analyzed. First, words were selected from the Dulch and Fry high frequency word lists which are often utilized in classrooms. Additional frequency databases, such as the Corpus of Contemporary American English, the Children's Printed Word Database, and the Educator's Word Frequency Guide, were also examined. Following analysis of those lists, additional "regular" and "irregular" words were selected that were not high frequency words.

Nonword Reading (NRE): Students are presented with printed nonwords displayed individually on an iPad and prompted to read them. Prior to beginning this task, students are provided with a demonstration item and a practice item. There is a total of 248

items for this task. To develop nonwords, we utilized data from Ziegler et al. (1997) to determine the most frequently occurring rimes and constructed nonwords based on those rimes. VC, CVC, VCe, CVCC, CCVC, CCVCC, CCCVC, CVVC, CCVVCC, CCVVC, and multisyllabic words were created. We utilized both frequent and infrequent rimes for nonword construction. Along with frequency, regularity was also considered to ensure most words yielded one correct pronunciation.

Spelling (SPE): Students are presented with a prerecorded nonword and asked to spell it using a small set of five to seven letters displayed on an iPad. Prior to beginning the task, students are shown how to drag letters from the top to the bottom of the screen to spell the target word. Items are scored correct or incorrect. Items were single syllable and based on the same nonword reading data. There are 103 items in this task.

Deletion (DEL): Students are presented with a prerecorded word and are asked to delete a morpheme, syllable, or phoneme and produce the remaining word. Prior to beginning the task, students are provided with a demonstration item and two practice items. There is a total of 94 items for this task. Items were selected in which an initial or final morpheme (i.e., BASEball), syllable (i.e., penCIL), or phoneme (i.e., Seat) could be deleted and a word known to students would remain. For the most difficult items, students are required to delete a phoneme from an initial or final consonant cluster (i.e., sPoon).

Blending (BLE): Students are presented with a prerecorded target word that has been broken into a series of morphemes, syllables, onset/rimes or individual phonemes (approximately 800 milliseconds pause between segments) and asked to blend the segments together to produce the target word. Prior to beginning the task, students are provided with a demonstration item and two practice items. There is a total of 73 items for this task, which is for Kindergarten students only. Items were selected based on their familiarity to students and ability to be divided into pronounceable morphemes (i.e., base ball), syllables (i.e., pa per), onset/rimes (i.e., s oap), or phonemes (i.e., s n a ke).

Expressive Vocabulary (EVO): Students are presented with a single image displayed on an iPad and prompted to name 1) the image, 2) the action displayed in the image, or 3) a certain part of the image. Prior to beginning this task, students are provided with a demonstration item and a practice item. There is a total of 200 items for this task. Items were selected based on their varying familiarity to children in Kindergarten through Grade 3.

Receptive Vocabulary (RVO): Students are presented with a prerecorded audio prompt of a target word with four accompanying images displayed on an iPad and prompted to select the image (finger touch) corresponding to the target word. There is a total of 225 items for this task. Items were selected based on their varying familiarity to children in Kindergarten through Grade 3.

Synonyms (SYN): Students are presented with a prerecorded audio prompt of a target word and prompted to provide another word that means the same thing as the target word. There is a total of 140 items for this task. Items were selected based on their varying familiarity to children in Kindergarten through Grade 3.

Nonword Repetition (NWR): Students are presented with a prerecorded audio prompt of a nonword and were asked to repeat the nonword verbatim. The task was scored at the syllable level, but the final score is computed as item correct versus incorrect. There is a total of 246 items for this task. This task has been presented in both a computer adaptive and fixed item format. Nonwords were selected to range from one to five syllables in length and utilize phoneme combinations that are admissible in English but occur at a low frequency, such as $/f/$ and $/ak/$. This selection approach was used to minimize

the influence of prior language knowledge because children with larger vocabularies can benefit from syllable familiarity (Dollaghan & Campbell, 1998).

Sentence Repetition (SRT): Students are presented with a prerecorded audio prompt of a target sentence and asked to repeat the sentence verbatim. Prior to beginning the task, students are provided with two practice items. The task is scored at the word level; however, the final score is computed as item correct versus incorrect. The task has a total of 213 items and has been presented in both computer adaptive and fixed order manner. Sentences include complex syntactical structures (i.e., passives, subordinate clauses, expanded noun phrases) as previous research has shown that students with language difficulties demonstrate problems comprehending and producing these complex structures (Gillam et al., 2019; Riches et al., 2010).

2.1.3 Psychometric Approaches

Multiple Group Item Response Theory (MG-IRT) is an extension of IRT, a statistical framework used to model the relation between individuals' latent traits (e.g., ability, proficiency) and their response to test items, as previously described in Test Design and Development. MG-IRT extends IRT to account for multiple groups (e.g., students from different schools or grades) simultaneously, allowing for the comparison of item performance and latent trait distributions across these groups. This allows for testing whether item characteristics are the same across groups or differ.

Item parameters were estimated using a multiple group two-parameter logistic (2PL) IRT model, which accounts for both item difficulty and discrimination. Response data were analyzed using a maximum-likelihood multiple group analysis implemented in the multipleGroup function in the mirt package (Chalmers, 2012) with group designated by grade and slopes and intercepts fixed across group but means freed. The standard Expectation Maximization algorithm with fixed quadrature was used for the estimation algorithm, and the cross-product estimation method was used for calculating the parameter information matrix for computing standard errors. The 2PL model was selected to allow variation in item discrimination, enabling more accurate assessment of how well each item distinguishes between individuals of differing ability levels. Difficulty parameters (b) reflect the location on the latent trait continuum where a respondent has a 50% chance of answering the item correctly, while discrimination parameters (a) indicate how sharply the item differentiates between examinees near that ability level. Items with poor fit or extreme parameter estimates (e.g., low a) were reviewed and excluded from the final item pool. An expectation of the multiple-group IRT analyses is that resulting reliability estimates would be lower than in a computer adaptive testing environment. Fixed forms present the same items to everyone, so difficult items that provide information for high-ability students must be answered by low-ability students and vice versa for easy items and high-ability students.

Following the multiple group IRT approach to item bank calibration for task, many of the measures are presented in a computer adaptive format. As such, these assessments are tailored to each student's ability in real-time. Each computer adaptive assessment utilizes an item bank which has test items across the range of proficiency levels, where appropriate across Kindergarten through Grade 3 levels (Georgiadou et al., 2006). Approximately 4,000 students in Kindergarten through Grade 3 participated in the item bank development. Each item was administered to approximately 200 children in subsets of 12 to 15 items (one to three items per child) including three to four linking items. Administration of a computer adapted measure begins with a fixed number of items, ranging from three

to five items that range in difficulty. These items determine the difficulty of the following items, and the process continues until a reliable estimate of the child's ability is determined. On average, students received seven to nine items per task to reach a reliable estimate.

Table 2.1 provides the number of items dropped for each IEAS measure for the following reasons: less than 100 children attempted that item, negative or low item discrimination values, or poor item fit based on the posterior predictive RMSEA for the Q1 statistic.

Table 2.1: Number of Items Dropped for Each IEAS Measure Due to Less Than 100 Item Attempts, Discrimination p>.05, or Poor Item Fit

Task	Number of Items
BLE	3
DEL	4
WRE	77
NRE	32
RVO	61
SPE	23
SYN	37
EVO	37
NWR	19
SRT	35

Note: BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

2.1.4 Missing Data Design and Imputation Process

Following the multiple-group IRT analyses in the calibration sample, and as part of the missing data design, multiple imputation procedures were used in order to estimate ability scores across tasks. This multiple imputation procedure was conducted in R using the mice package (Van Buuren & Groothuis-Oudshoorn, 2011) using 500 imputations. Imputation occurred for only the IEAS measures and without the presence of the end of year gold standard measures present in the dataset. Imputation was completed for each grade separately, and data were imputed using a semester split of first half of the year (fall) and second half of the year (winter and spring). In order to be imputed, variables must have had at least 30% complete data.

2.1.5 Results

2.1.5.1 Empirical Reliability

Empirical reliability refers to the consistency or stability of test scores, as measured using observed data from a specific sample. It allows for the quantification of how much variability in test scores reflects true differences in the measured construct versus random error due to factors like measurement imprecision, test conditions, or respondent variability. It is a critical indictor or a test's quality because it affects the precision of score interpretations and decisions based on those scores. Reliability values range from

0 to 1, with values greater than or equal to 0.80 considered acceptable for research and greater than or equal to 0.90 for clinical decision-making (Nunnally & Bernstein, 1994).

Table 2.2 presents the empirical reliability estimates for the IEAS measures from the fixed form administrations across Kindergarten through Grade 3, given at different time points (fall, winter, spring) during Phase 1 of the study. Sample sizes are included in parentheses. The measures generally demonstrate moderate to high reliability, with WRE, NRE, BLE, and DEL being the most consistent. RVO, NWR, and EVO show more variability, particularly with smaller samples or in later grades. Reliability estimates range from 0.45 to 0.93 across measures, grades, and time points. WRE (0.82–0.93), NRE (0.79–0.88), BLE (0.84–0.90), and DEL (0.81–0.86) consistently show strong reliability across grades and time points. SYN (0.68–0.84), SRT (0.67–0.81), SPE (0.76–0.84), and RVO (0.45–0.78) show moderate reliability, with RVO having the lowest reliability in some cases (e.g., 0.45 in Kindergarten for spring). NWR (0.53–0.74) and EVO (0.53–0.80) often show lower reliability, particularly in Grade 2 and Grade 3 or with smaller samples.

Table 2.2: Phase 1 Fixed Form, Empirical Reliability Estimates by Grade, Time, and Task (Sample Size)

Task x Time	BLE	DEL	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
K Fall	.89 (1053)	.83 (802)	—	—	.68 (982)	—	.79 (629)	.79 (711)	.70 (661)	.75 (769)
K Winter	.90 (109)	.86 (210)	.86 (258)	.84 (256)	.61 (54)	.83 (138)	.79 (158)	.80 (208)	.70 (210)	.81 (260)
K Spring	.84 (41)	.83 (192)	.87 (376)	.86 (362)	.45 (146)	.84 (150)	.81 (219)	.75 (193)	.74 (363)	.79 (335)
G1 Fall	—	—	.88 (734)	.88 (717)	.78 (551)	.80 (513)	.80 (563)	.80 (651)	.71 (452)	.77 (598)
G1 Winter	—	—	.84 (137)	.91 (273)	.88 (247)	—	.80 (259)	.83 (107)	.76 (105)	.80 (111)
G1 Spring	—	—	.83 (121)	.90 (197)	.87 (196)	—	.77 (185)	.71 (61)	.72 (71)	.71 (75)
G2 Fall	—	—	.86 (374)	.86 (382)	.68 (507)	.76 (167)	.77 (354)	.79 (375)	.57 (124)	.67 (319)
G2 Winter	—	—	.81 (44)	.82 (135)	.85 (59)	.71 (113)	.76 (22)	.68 (135)	.67 (113)	.53 (59)
G2 Spring	—	—	.86 (114)	.86 (213)	.84 (78)	.73 (169)	.84 (179)	.77 (209)	.70 (80)	.81 (213)
G3 Fall	—	—	.87 (394)	.88 (452)	.76 (272)	—	.70 (327)	.62 (256)	—	.67 (323)
G3 Winter	—	—	.88 (125)	.79 (28)	—	—	.68 (125)	.53 (97)	—	.70 (122)
G3 Spring	—	—	.93 (104)	.81 (29)	—	—	.75 (104)	.55 (75)	—	.70 (101)

Note: Reliability is not calculated for samples sizes less than 22. BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Across three strategic implementation studies using the computer adaptive version of the assessments, we obtained empirical reliability estimates (Table 2.3, Table 2.4, and Table 2.5) to understand the level of precision and the efficiency of the assessments in classroom settings. Across all grade levels and subtests, reliability was extremely high. Additionally, a high percentage of students were able to terminate the task based on a reliability threshold as opposed to a maximum item threshold for each subtest.

Table 2.3: California Implementation Study (2023) Empirical Reliability

Task	Grade	Term	Reliability	N	% SE
BLE	Kindergarten	Fall	0.81	812	94.46%
		Winter	0.82	1,195	86.95%
		Spring	0.82	903	83.39%
DEL	Kindergarten	Fall	0.78	729	98.63%
		Winter	0.80	1,189	98.40%
		Spring	0.80	1,105	98.64%
	Grade 1	Fall	0.82	873	94.85%
		Winter	0.82	1,264	91.53%
		Spring	0.79	962	90.23%
EVO	Kindergarten	Fall	0.95	901	73.14%
		Winter	0.95	1,390	67.27%
		Spring	0.95	1,228	59.28%
	Grade 1	Fall	0.94	980	70.82%
		Winter	0.94	1,435	64.88%
		Spring	0.94	1,257	60.46%
NRE	Grade 1	Fall	0.89	810	90.49%
		Winter	0.88	1,203	86.53%
		Spring	0.86	975	84.72%
NWR	Kindergarten	Fall	0.92	920	78.70%
		Winter	0.93	1,389	77.39%
		Spring	0.93	1,229	74.13%
	Grade 1	Fall	0.91	978	87.01%
		Winter	0.92	1,433	80.46%
		Spring	0.91	1,256	80.02%
RVO	Kindergarten	Fall	0.91	914	83.48%
		Winter	0.91	1,313	78.67%
		Spring	0.90	1,032	74.22%
	Grade 1	Fall	0.87	968	82.23%
		Winter	0.87	1,382	81.19%
		Spring	0.88	1,089	74.93%
SRT	Kindergarten	Fall	0.96	892	99.10%
		Winter	0.96	1,314	98.93%
		Spring	0.96	1,069	98.88%
	Grade 1	Fall	0.97	946	65.43%
		Winter	0.97	1,379	70.27%
		Spring	0.97	1,093	75.39%
WRE	Grade 1	Fall	0.96	803	80.70%
		Winter	0.96	1,212	75.08%
		Spring	0.96	998	75.65%

Note: BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition. %SE = percentage of students in the sample who terminated the CAT based on standard error threshold.

Table 2.4: Florida Implementation Studies (2023, 2024) Empirical Reliability

Time	Task	Grade	Reliability	N	% SE
Winter 2023	DEL	Kindergarten	0.82	517	86%
Winter 2023	DEL	Grade 1	0.78	591	86%
Winter 2023	BLE	Kindergarten	0.82	312	81%
Winter 2023	EVO	Kindergarten	0.95	590	82%
Winter 2023	EVO	Grade 1	0.92	671	99%
Winter 2023	RVO	Kindergarten	0.89	413	91%
Winter 2023	RVO	Grade 1	0.84	415	94%
Winter 2023	SRT	Grade 1	0.94	403	92%
Winter 2023	SYN	Kindergarten	0.88	335	81%
Winter 2023	SYN	Grade 1	0.90	376	92%
Winter 2023	WRE	Grade 1	0.97	639	93%
Winter 2024	DEL	Kindergarten	0.89	267	88%
Winter 2024	DEL	Grade 1	0.88	295	92%
Winter 2024	BLE	Kindergarten	0.87	275	87%
Winter 2024	EVO	Kindergarten	0.89	291	94%
Winter 2024	EVO	Grade 1	0.87	315	93%
Winter 2024	WRE	Kindergarten	0.94	268	88%
Winter 2024	WRE	Grade 1	0.97	301	85%
Winter 2024	NRE	Grade 1	0.92	299	92%

Note: BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SYN=Synonyms, EVO=Expressive Vocabulary. %SE = percentage of students in the sample who terminated the CAT based on standard error threshold.

Table 2.5: South Carolina Implementation Study (2024)

Task	Grade	Reliability	N	% SE
DEL	Kindergarten	0.91	323	79%
DEL	Grade 1	0.89	397	89%
BLE	Kindergarten	0.89	325	78%
EVO	Kindergarten	0.87	374	92%
EVO	Grade 1	0.87	388	92%
EVO	Grade 3	0.88	270	95%
WRE	Kindergarten	0.95	306	77%
WRE	Grade 1	0.97	389	81%
WRE	Grade 3	0.95	258	98%
NRE	Grade 1	0.93	372	83%
NRE	Grade 3	0.95	258	91%

Note: BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, EVO=Expressive Vocabulary. %SE = percentage of students in the sample who terminated the CAT based on standard error threshold.

2.1.5.2 Construct Validity

Construct validity refers to the degree to which an assessment accurately represents the theoretical construct it is intended to measure. Evidence for construct validity is typically accumulated through convergent validity and discriminant validity. The former is observed when a measure correlates moderate to strongly with other assessments of

the same or related constructs, and the latter is the demonstration that the measure shows weak or negligible correlations with assessments of unrelated constructs.

2.1.5.2.0.1 Convergent Validity Correlations Table 2.6 provides a summary of how convergent validity, as measured by average correlations, varies by construct and time point, aligning with the following theoretical relations: Phonological Awareness (BLE, DEL, NWR), Reading and Spelling (WRE, NRE, SPE), and Oral Language (EVO, RVO, SYN, SRT). The data show distinct convergent validity patterns across the theoretical constructs. Values ranging from 0.10 to 0.29 can be interpreted as weak, values from 0.30 to 0.49 as moderate, and greater than or equal to 0.50 as strong.

Phonological Awareness. Moderate validity is demonstrated with values ranging from 0.26 (small) to 0.41 (moderate), peaking in Grade 1 Spring (0.41) and lowest in Kindergarten Spring and Grade 2 Spring (0.26). Data is sparse, appearing only in Kindergarten and parts of Grade 1 and Grade 2, with moderate convergence overall. Within the time point specific tables, BLE-DEL correlations are consistently stronger and NWR weaker indicating some variability within the construct. Across all time points, there is a moderate correlation (0.34), reflecting reasonable but not robust alignment due to NWR's weaker links.

Reading and Spelling. Strong convergent validity is demonstrated with consistently large average correlations, ranging from 0.51 to 0.78, indicating that these measures effectively capture a unified construct at all measured time points. Across all time points, the average correlation of 0.68 demonstrates excellent and consistent convergence.

Oral Language. Correlations range from moderate (0.42 in Grade 3 Fall) to large (0.58 in Grade 2 Spring). The average correlation across all time points is strong (0.50), showing good and uniform alignment across grades and terms.

All Measures. The Reading and Spelling construct shows the strongest and most consistent validity, with large correlations across all time points, suggesting WRE, NRE, and SPE reliably measure a cohesive construct of core reading and spelling skills. Phonological Awareness exhibits moderate validity, with BLE and DEL aligning well but NWR's weaker correlations (e.g., 0.08 in Kindergarten Spring) indicating some divergence within the construct. Oral Language demonstrates large validity overall (0.50), with large correlations in most tables, particularly Grade 2 Winter Spring File (0.58) and Grade 1 Winter (0.56). These findings highlight the reliability of Reading and Spelling, moderate consistency in Phonological Awareness, and strong but variable convergence in Oral Language across grades and terms.

The average correlation across all grades and constructs is 0.53. This average of 0.53 reflects strong convergent validity overall, indicating strong cohesion among the measures within each construct across the time points. The value is influenced by the consistently large correlations in Reading and Spelling (mean: 0.68), the large correlations in Oral Language (mean: 0.50), and the moderate correlations in Phonological Awareness (mean: 0.34).

Table 2.7 through Table 2.18 provide time-point specific convergent validity correlations to support these findings across all grades and time points.

Table 2.6: Average Convergent Validity Correlations, By Construct Across Grades and Time Points

Grade	Time Point	Sample Size	All Measures	Phonological Awareness (BLE, DEL, NWR)	Reading & Spelling (WRE, NRE, SPE)	Oral Language (EVO, RVO, SYN, SRT)
Kindergarten	Fall	841	0.41	0.35	—	0.47
Kindergarten	Winter	260	0.50	0.38	0.66	0.47
Kindergarten	Spring	377	0.48	0.26	0.69	0.49
Grade 1	Fall	720	0.63	—	0.70	0.55
Grade 1	Winter	346	0.57	0.39	0.76	0.56
Grade 1	Spring	197	0.55	0.41	0.71	0.52
Grade 2	Fall	550	0.64	—	0.72	0.56
Grade 2	Winter	179	0.42	0.30	0.51	0.46
Grade 2	Spring	257	0.48	0.26	0.60	0.58
Grade 3	Fall	385	0.54	—	0.65	0.42
Grade 3	Winter	162	0.56	—	0.66	0.46
Grade 3	Spring	141	0.67	—	0.78	0.55
All	All	4415	0.53	0.34	0.68	0.50

Note: BLE=Blending, DEL=Deletion, NWR=Nonword Repetition, WRE=Word Reading, NRE=Nonword Reading, SPE=Spelling, EVO=Expressive Vocabulary, RVO=Receptive Vocabulary, SYN=Synonyms, SRT=Sentence Repetition. Not all measures are present at all time points, resulting in some values being partial correlations. See the convergent validity correlation tables by grade and time point for time-specific measures.

Table 2.7: Fall Kindergarten Convergent Validity Correlations (N=841)

Measure	BLE	DEL	RVO	SYN	EVO	NWR	SRT	LNC
BLE	1.00	—	—	—	—	—	—	—
DEL	0.65	1.00	—	—	—	—	—	—
RVO	0.33	0.20	1.00	—	—	—	—	—
SYN	0.49	0.59	0.38	1.00	—	—	—	—
EVO	0.41	0.45	0.48	0.59	1.00	—	—	—
NWR	0.13	0.27	0.15	0.26	0.31	1.00	—	—
SRT	0.46	0.58	0.32	0.52	0.52	0.31	1.00	—
LNC	0.47	0.39	0.24	0.39	0.36	0.20	0.33	1

Note: BLE=Blending, DEL=Deletion, RVO=Receptive Vocabulary, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition, LNC=Letter Names in Context.

Table 2.8: Winter Kindergarten Convergent Validity Correlations (N=260)

Measure	BLE	DEL	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
BLE	1.00	—	—	—	—	—	—	—	—	—
DEL	0.60	1.00	—	—	—	—	—	—	—	—
WRE	0.47	0.48	1.00	—	—	—	—	—	—	—
NRE	0.56	0.58	0.71	1.00	—	—	—	—	—	—
RVO	0.31	0.49	0.17	0.28	1.00	—	—	—	—	—
SPE	0.59	0.58	0.57	0.72	0.22	1.00	—	—	—	—
SYN	0.56	0.60	0.37	0.56	0.24	0.53	1.00	—	—	—
EVO	0.38	0.51	0.28	0.36	0.47	0.40	0.54	1.00	—	—
NWR	0.27	0.27	0.21	0.24	0.24	0.21	0.31	0.33	1.00	—
SRT	0.45	0.57	0.33	0.49	0.47	0.52	0.50	0.59	0.43	1

Note: BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.9: Spring Kindergarten Convergent Validity Correlations (N=377)

Measure	BLE	DEL	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
BLE	1.00	—	—	—	—	—	—	—	—	—
DEL	0.56	1.00	—	—	—	—	—	—	—	—
WRE	0.45	0.52	1.00	—	—	—	—	—	—	—
NRE	0.51	0.57	0.76	1.00	—	—	—	—	—	—
RVO	0.34	0.37	0.17	0.25	1.00	—	—	—	—	—
SPE	0.57	0.56	0.62	0.70	0.19	1.00	—	—	—	—
SYN	0.48	0.64	0.51	0.54	0.30	0.50	1.00	—	—	—
EVO	0.40	0.46	0.29	0.28	0.52	0.31	0.52	1.00	—	—
NWR	0.08	0.15	0.20	0.17	0.17	0.05	0.21	0.27	1.00	—
SRT	0.40	0.46	0.35	0.38	0.43	0.42	0.56	0.59	0.36	1

Note: BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, LCP=Listening Comprehension, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.10: Fall Grade 1 Convergent Validity Correlations (N=720)

Measure	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
WRE	1.00	—	—	—	—	—	—	—
NRE	0.74	1.00	—	—	—	—	—	—
RVO	0.31	0.41	1.00	—	—	—	—	—
SPE	0.62	0.73	0.37	1.00	—	—	—	—
SYN	0.53	0.61	0.61	0.52	1.00	—	—	—
EVO	0.40	0.51	0.72	0.37	0.66	1.00	—	—
NWR	0.09	0.01	-0.01	-0.01	0.03	-0.09	1.00	—
SRT	0.38	0.41	0.48	0.42	0.46	0.37	0.25	1

Note: WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.11: Winter Grade 1 Convergent Validity Correlations (N=346)

Measure	DEL	WRE	NRE	SPE	SYN	EVO	NWR	SRT
DEL	1.00	—	—	—	—	—	—	—
WRE	0.75	1.00	—	—	—	—	—	—
NRE	0.73	0.81	1.00	—	—	—	—	—
SPE	0.66	0.72	0.76	1.00	—	—	—	—
SYN	0.48	0.60	0.62	0.60	1.00	—	—	—
EVO	0.35	0.49	0.52	0.49	0.58	1.00	—	—
NWR	0.39	0.34	0.29	0.20	0.22	0.15	1.00	—
SRT	0.51	0.48	0.49	0.45	0.60	0.50	0.31	1

Note: DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.12: Spring Grade 1 Convergent Validity Correlations (N=197)

Measure	DEL	WRE	NRE	SPE	SYN	EVO	NWR	SRT
DEL	1.00	—	—	—	—	—	—	—
WRE	0.62	1.00	—	—	—	—	—	—
NRE	0.70	0.76	1.00	—	—	—	—	—
SPE	0.60	0.66	0.71	1.00	—	—	—	—
SYN	0.39	0.50	0.53	0.55	1.00	—	—	—
EVO	0.23	0.33	0.35	0.43	0.55	1.00	—	—
NWR	0.41	0.30	0.32	0.31	0.27	0.19	1.00	—
SRT	0.44	0.26	0.41	0.38	0.54	0.48	0.34	1

Note: DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.13: Fall Grade 2 Convergent Validity Correlations (N=550)

Measure	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
WRE	1.00	—	—	—	—	—	—	—
NRE	0.75	1.00	—	—	—	—	—	—
RVO	0.49	0.48	1.00	—	—	—	—	—
SPE	0.72	0.68	0.55	1.00	—	—	—	—
SYN	0.51	0.50	0.60	0.46	1.00	—	—	—
EVO	0.56	0.53	0.69	0.61	0.58	1.00	—	—
NWR	0.39	0.24	0.36	0.41	0.28	0.37	1.00	—
SRT	0.45	0.48	0.49	0.45	0.48	0.55	0.22	1

Note: WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, LCP=Listening Comprehension, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.14: Winter Grade 2 Convergent Validity Correlations (N=179)

Measure	DEL	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
DEL	1.00	—	—	—	—	—	—	—	—
WRE	0.61	1.00	—	—	—	—	—	—	—
NRE	0.42	0.67	1.00	—	—	—	—	—	—
RVO	0.47	0.36	0.32	1.00	—	—	—	—	—
SPE	0.35	0.50	0.37	0.11	1.00	—	—	—	—
SYN	0.47	0.42	0.41	0.59	0.34	1.00	—	—	—
EVO	0.35	0.35	0.20	0.57	-0.01	0.50	1.00	—	—
NWR	0.30	0.34	0.26	0.27	-0.23	0.24	0.17	1.00	—
SRT	0.42	0.33	0.30	0.32	0.03	0.39	0.40	0.28	1

Note: DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.15: Spring Grade 2 Convergent Validity Correlations (N=257)

Measure	DEL	WRE	NRE	RVO	SPE	SYN	EVO	NWR	SRT
DEL	1.00	—	—	—	—	—	—	—	—
WRE	0.66	1.00	—	—	—	—	—	—	—
NRE	0.56	0.79	1.00	—	—	—	—	—	—
RVO	0.49	0.49	0.47	1.00	—	—	—	—	—
SPE	0.43	0.56	0.44	0.17	1.00	—	—	—	—
SYN	0.54	0.61	0.61	0.68	0.47	1.00	—	—	—
EVO	0.49	0.46	0.42	0.69	0.09	0.66	1.00	—	—
NWR	0.26	0.38	0.34	0.33	0.17	0.26	0.22	1.00	—
SRT	0.50	0.48	0.46	0.44	0.14	0.50	0.50	0.42	1

Note: DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SPE=Spelling, SYN=Synonyms, EVO=Expressive Vocabulary, NWR=Nonword Repetition, SRT=Sentence Repetition.

Table 2.16: Fall Grade 3 Convergent Validity Correlations (N=385)

Measure	WRE	NRE	RVO	SYN	EVO	SRT
WRE	1.00	—	—	—	—	—
NRE	0.65	1.00	—	—	—	—
RVO	0.46	0.45	1.00	—	—	—
SYN	0.43	0.37	0.69	1.00	—	—
EVO	0.37	0.30	0.45	0.21	1.00	—
SRT	0.27	0.29	0.44	0.35	0.38	1

Note: WRE=Word Reading, NRE=Nonword Reading, RVO=Receptive Vocabulary, SYN=Synonyms, EVO=Expressive Vocabulary, SRT=Sentence Repetition.

Table 2.17: Winter Grade 3 Convergent Validity Correlations (N=162)

Measure	WRE	NRE	SYN	EVO	SRT
WRE	1.00	—	—	—	—
NRE	0.70	1.00	—	—	—
SYN	0.52	0.47	1.00	—	—
EVO	0.37	0.24	0.56	1.00	—
SRT	0.30	0.15	0.42	0.47	1

Note: WRE=Word Reading, NRE=Nonword Reading, SYN=Synonyms, EVO=Expressive Vocabulary, SRT=Sentence Repetition.

Table 2.18: Spring Grade 3 Convergent Validity Correlations (N=141)

Measure	WRE	NRE	SYN	EVO	SRT
WRE	1.00	—	—	—	—
NRE	0.78	1.00	—	—	—
SYN	0.60	0.55	1.00	—	—
EVO	0.43	0.34	0.63	1.00	—
SRT	0.43	0.37	0.50	0.52	1

Note: WRE=Word Reading, NRE=Nonword Reading, SYN=Synonyms, EVO=Expressive Vocabulary, LCP=Listening Comprehension, SRT=Sentence Repetition.

2.1.5.3 Criterion Validity

Concurrent Validity: IEAS measures and standardized assessments.

Correlations between the IEAS measures and recognized standardized assessments are found in Table 2.19 and demonstrate moderate to strong associations within and across grade levels. Most correlations (8 out of 10) are strong (greater than or equal to 0.50), indicating that IEAS measures generally align well with standardized assessments across grades, supporting their concurrent validity. Nonword Reading consistently shows strong correlations (0.66 to 0.75) across all grades, suggesting stable validity for decoding skills. Sentence Repetition correlations strengthen from Grade 1 (0.52) to Grade 2 (0.77), possibly reflecting improved language processing skills. Expressive Vocabulary correlations vary, with a dip in Grade 1 (0.31, moderate) but strong performance in Kindergarten

(0.51) and Grade 2 (0.77). Only two correlations are moderate (Kindergarten Receptive Vocabulary: 0.46; Grade 1 Expressive Vocabulary: 0.31), suggesting areas where IEAS measures may have less alignment with standardized tests, possibly due to task differences or developmental factors. Sample sizes range from 146 (Kindergarten Receptive Vocabulary) to 376 (Kindergarten Nonword Reading/Expressive Vocabulary), with larger samples generally associated with stronger correlations. However, small samples (e.g., 146 for Receptive Vocabulary) still yield meaningful results. These results show the IEAS measures are valid for assessing reading, language, and vocabulary skills in early elementary students.

Table 2.19: Concurrent Validity Between IEAS Measures and Selected Standardized Assessments (Spring Sample)

Grade	IEAS Measure	Outcome	Correlation	Sample size
Kindergarten	NRE	KTEA NWD	0.66	376
	SRT	CELF RS	0.69	363
	EVO	WPPSI PN	0.51	376
	RVO	WPPSI RV	0.46	146
Grade 1	NRE	KTEA NWD	0.75	192
	SRT	CELF RS	0.52	194
	EVO	WPPSI PN	0.31	196
Grade 2	NRE	KTEA NWD	0.70	209
	SRT	CELF RS	0.77	205
	EVO	EVT	0.77	206

Note: WRE=Word Reading, NRE=Nonword Reading, SRT=Sentence Repetition, EVO=Expressive Vocabulary, RVO=Receptive Vocabulary, KTEA LWR=Kauffman Test Educational Achievement Letter Word Recognition, KTEA NWD=Kauffman Test Educational Achievement Nonword Decoding, WPPSI PN=WPPSI Picture Naming, WPPSI RV=Receptive Vocabulary, CELF RS=Clinical Evaluation of Language Fundamentals Recalling Sentences, CELF SC=Clinical Evaluation of Language Fundamentals Sentence Comprehension, EVT=Expressive Vocabulary Test.

Predictive and Concurrent Validity: IEAS measures and KTEA Word Reading

Table 2.20 reveals that reading measures (Word Reading, Nonword Reading) exhibit the strongest predictive and concurrent validity with KTEA Word Reading (a measure of decoding and sight word recognition), improving across grades and peaking in concurrent spring assessments. Phonological awareness measures (Deletion, Blending) are moderately predictive in early grades, supporting their role in foundational reading development, while oral language measures (Expressive Vocabulary, Synonyms) show weaker but still relevant associations, indicating they tap related but distinct constructs from word reading. Trends suggest a developmental progression where predictive power strengthens as students approach higher grades and reading fluency solidifies.

General Trends Across Measures

Range of Correlations. Predictive validity (fall, winter) correlations range from weak (e.g., 0.10 for Nonword Reading in Grade 1 Fall) to very strong (e.g., 0.89 for Word Reading in Grade 1 Winter). Concurrent validity correlations range from weak (e.g., 0.14 for Nonword Reading in Kindergarten Spring) to very strong (e.g., 0.91 for Word Reading in Grade 3 Spring). Stronger correlations tend to appear in later grades and with certain measures (e.g., Word Reading, Nonword Reading). Winter correlations are often higher than fall for predictive validity (e.g., Word Reading increases from 0.82 in Grade 1 Fall to 0.89 in Winter), indicating closer proximity to spring testing enhances prediction. Spring

(concurrent) correlations are typically strong for direct reading measures (Word Reading, Nonwording Reading), reflecting their alignment with KTEA Word Reading at the same time point.

Specific Measure Trends

Word Reading: Word Reading serves as the strongest and most consistent predictor across all grades and time points. Its predictive validity increases from kindergarten (0.69 winter) to grade 1 (0.82 fall, 0.89 winter), then slightly dips in grade 2 (0.79 fall) and grade 3 (0.71 fall), but remains high. Concurrent validity peaks at 0.91 in Grade 3 Spring, showing near-perfect alignment with KTEA Word Reading.

Nonword Reading. Nonword Reading exhibits solid predictive validity, ranging from 0.59 (Grade 2 Winter) to 0.77 (Grade 1 Winter). Concurrent validity is strongest in Grade 3 (0.72 spring) and grade 1 (0.68 spring). It is slightly less predictive than Word Reading but still trends with consistent moderate-to-strong correlations.

Deletion. The predictive validity of Deletion scores is moderate in Kindergarten (0.47–0.49), jumps to 0.71 in Grade 1 Winter and 0.60–0.63 in Grade 2. Concurrent validity is moderate-to-strong at 0.47 in Kindergarten Spring, 0.64 in Grade 1 Spring, and 0.63 in Grade 2 Spring.

Blending and Letter Naming in Context. Blending demonstrates moderate predictive validity in Kindergarten (0.40 to 0.47), and concurrent at 0.36. Letter Naming in Context also demonstrates moderate predictive validity in Kindergarten at 0.46. Both of these early phonological awareness measures are not administered beyond Kindergarten.

Spelling. Spelling scores provide moderate predictive validity, ranging from 0.50 (Grade 2 Winter) to 0.68 (Grade 1 Winter). Concurrent validity is also moderate, with 0.51–0.61 in Kindergarten and Grade 1 Spring.

Oral Language Vocabulary Measures (Expressive Vocabulary, Receptive Vocabulary, Synonyms). Vocabulary measures are less predictive than reading skills, with some strengthening in Grade 2, but weaker overall. Expressive Vocabulary and Receptive Vocabulary predictive correlations are weak to moderate (0.18 to 0.47), and concurrent correlations peak in Grade 2 Spring (0.45 and 0.48, respectively). Synonyms demonstrates greater validity, with moderate-to-strong predictive validity correlations ranging from 0.37 to 0.57, and concurrent correlations ranging from 0.44 in Kindergarten Spring to 0.64 in Grade 2 Spring.

Oral Language General Measure (Sentence Repetition). This language-based measure shows modest predictive and concurrent validity, less tied to KTEA Word Reading. Correlations show weak-to-moderate results (0.26 to 0.48).

Nonword Repetition. This measure is the weakest predictor across grades and timepoints (0.10–0.31 across grades and time points) and shows minimal alignment with KTEA Word Reading with concurrent correlations ranging from 0.14 to 0.27.

Table 2.20: Predictive (Fall and Winter) and Concurrent Validity (Spring) Between IEAS Measures and KTEA Word Reading by Grade and Time Point

Grade	Time	N	LNC	BLE	DEL	WRE	NRE	SPE	EVO	RVO	SYN	SRT	NWR
Kindergarten	Fall	841	0.46	0.40	0.47	—	—	—	0.28	0.18	0.38	0.31	0.12
Kindergarten	Winter	260	—	0.47	0.49	0.69	0.65	0.59	0.24	0.23	0.41	0.41	0.20
Kindergarten	Spring	377	—	0.36	0.47	0.76	0.66	0.51	0.23	0.19	0.44	0.27	0.14
Grade 1	Fall	720	—	—	—	0.82	0.72	0.62	0.36	0.30	0.52	0.43	0.10
Grade 1	Winter	346	—	—	0.71	0.89	0.77	0.68	0.47	—	0.57	0.43	0.27
Grade 1	Spring	197	—	—	0.64	0.86	0.68	0.61	0.33	—	0.45	0.26	0.27
Grade 2	Fall	550	—	—	—	0.79	0.65	0.63	0.46	0.44	0.37	0.48	0.29
Grade 2	Winter	179	—	—	0.60	0.87	0.59	0.50	0.37	0.39	0.46	0.32	0.31
Grade 2	Spring	257	—	—	0.63	0.83	0.70	0.55	0.45	0.48	0.64	0.40	0.27
Grade 3	Fall	385	—	—	—	0.71	0.69	—	0.38	0.46	0.38	0.27	—
Grade 3	Winter	162	—	—	—	0.81	0.63	—	0.28	—	0.39	0.29	—
Grade 3	Spring	141	—	—	—	0.91	0.72	—	0.39	—	0.54	0.39	—

Note: LNC=Letter Names in Context, BLE=Blending, DEL=Deletion, WRE=Word Reading, NRE=Nonword Reading, SPE=Spelling, EVO=Expressive Vocabulary, RVO=Receptive Vocabulary, SYN=Synonyms, SRT=Sentence Repetition, NWR=Nonword Repetition.

2.1.5.4 Classification Accuracy

Classification accuracy is the percentage of correct predictions made by a model on a test dataset, where accuracy is determined by the number of correct predictions out of the total number of predictions. However, in a testing context, there are common indices beyond accuracy that are used for assessment screeners. One approach to generating these indices is the use of a confusion matrix. From the matrix, one can calculate sensitivity (the proportion of students who are identified as at risk on the screening assessment among all students who fail the outcome), specificity (the proportion of students who are identified as not at risk among all students who pass the outcome), positive predictive power (the proportion of students who fail the outcome assessment among all students who are identified as at risk on the screening assessment), and negative predictive power (the proportion of students who pass the outcome assessment among all students who are identified as not at risk on the screening assessment). Another approach is to use the area under the curve (AUC) calculated through a Receiver Operating Characteristic (ROC) methodology, which evaluates a test's ability to classify across all possible classification thresholds.

Jenkins (2003) suggested that screening assessments should demonstrate a negative predictive power of 0.90 to 0.95 and a sensitivity level of 0.90 to 0.95. As a rule of thumb, AUC values between 0.70 and 0.80 are usually considered acceptable, values between 0.80 and 0.90 are considered excellent, and values above 0.90 are outstanding.

Table 2.21 summarizes the classification accuracy of the IEAS screening algorithms for identifying students at risk for word reading difficulties (PWR) and dyslexia (DYS) across Kindergarten through Grade 3, at three time points (fall, winter, spring). The table reports metrics including Area Under the Curve (AUC), sensitivity, specificity, negative predictive power (NPP), and positive predictive power (PPP), along with sample sizes for each grade and time point.

AUC ranges from 0.76 (Kindergarten Fall, DYS) to 0.96 (Grade 2 Spring, PWR), indicating acceptable to outstanding classification accuracy, with higher values in later grades and time points. Sensitivity is generally high (0.71 to 0.92), improving from Kindergarten to Grade 2, meeting or approaching the recommended 0.90 to 0.95 for screening assessments by Jenkins (2003). Specificity ranges from 0.70 to 0.91, with stronger performance in Grade 1 and Grade 2. NPP is consistently high (0.78 to 0.99), often meeting or exceeding Jenkins' 0.90 to 0.95 threshold, especially for DYS. PPP is lower for DYS (0.29 to 0.61) compared to PWR (0.62 to 0.87), reflecting challenges in predicting dyslexia risk accurately.

Overall, accuracy improves from fall to spring and from Kindergarten to Grade 2, with Grade 3 showing stable but slightly lower metrics. PWR algorithms generally outperform DYS in PPP. These data demonstrate that IEAS demonstrates strong classification accuracy, particularly for PWR, with excellent sensitivity and NPP across grades, though DYS predictions have lower PPP, indicating some limitations in identifying true dyslexia cases.

Table 2.21: Classification Accuracy for Screening Algorithms by Grade, Time Point, and Algorithm Type

Grade	Time Point	N	Screening Algorithm	AUC	Sensitivity	Specificity	NPP	PPP
Kindergarten	Fall	798	PWR	0.77	0.71	0.70	0.78	0.62
	Fall	790	DYS	0.76	0.76	0.71	0.92	0.29
	Winter	798	PWR	0.83	0.75	0.73	0.95	0.83
	Winter	790	DYS	0.83	0.83	0.74	0.95	0.37
	Spring	325	PWR	0.87	0.81	0.77	0.86	0.71
	Spring	329	DYS	0.89	0.86	0.77	0.97	0.42
Grade 1	Fall	637	PWR	0.92	0.85	0.81	0.89	0.75
	Fall	604	DYS	0.93	0.86	0.82	0.97	0.46
	Winter	637	PWR	0.94	0.85	0.86	0.90	0.78
	Winter	604	DYS	0.95	0.86	0.85	0.97	0.51
	Spring	197	PWR	0.95	0.90	0.85	0.92	0.80
	Spring	343	DYS	0.92	0.87	0.83	0.97	0.48
Grade 2	Fall	455	PWR	0.91	0.85	0.86	0.90	0.81
	Fall	523	DYS	0.94	0.90	0.86	0.98	0.54
	Winter	455	PWR	0.95	0.91	0.89	0.93	0.84
	Winter	523	DYS	0.95	0.92	0.86	0.98	0.55
	Spring	457	PWR	0.96	0.90	0.91	0.93	0.87
	Spring	523	DYS	0.95	0.90	0.89	0.98	0.61
Grade 3	Fall	270	PWR	0.91	0.85	0.82	0.89	0.75
	Fall	380	DYS	0.93	0.93	0.82	0.99	0.49
	Winter	270	PWR	0.92	0.87	0.81	0.90	0.75
	Winter	380	DYS	0.91	0.85	0.80	0.97	0.45
	Spring	124	PWR	0.92	0.87	0.81	0.90	0.75
	Spring	137	DYS	0.91	0.85	0.80	0.97	0.45

Note: PWR=Potential for Word Reading, DYS=Dyslexia Risk Flag, AUC = Area Under the Curve, NPP = Negative Predictive Power, PPP = Positive Predictive Power.

2.1.5.5 Differential Test Functioning

Differential Test Functioning (DTF) refers to a situation in which a test item or a set of items functions differently across subgroups of test-takers (e.g., based on gender, ethnicity, or socioeconomic status) despite equivalent levels of the underlying trait or ability being measured. It is a way to detect potential bias or unfairness in testing by examining whether certain groups have different probabilities of answering items correctly, even when they have the same ability level.

To test whether the IEAS provides valid scores equally well across different demographic groups, a series of logistic regressions were used to predict success on the KTEA Word Reading in spring. The independent variables included a variable that represented whether students were identified as not at-risk (coded as '0') or at risk (coded as '1') based on the identified cut-point on a combination score of the screening tasks, a variable that represented a selected demographic group, as well as an interaction term between the two variables. A statistically significant interaction term indicates differential accuracy in predicting end-of-year risk status by PWR or DYS. Differential accuracy was separately tested for Black versus White students, Latino versus non-Latino students, female versus male students, and dual language learners versus non-dual language learners.

The differential classification accuracy test between Black and White students (Table 2.22 through Table 2.25), yielded no significant effect for the interaction term between the dichotomous PWR or DYS indicator and racial status was observed at any grade level. Likewise, there was no significant effect found in comparisons between Latino and non-Latino students, or dual language learners versus non-dual language learners. A significant interaction between female and male students and the PWR indicator was observed in Grade 3 (Table 2.25) but no significant interaction was observed in other grades or with the DYS indicator. These results suggest that the assessment works equally well for all the studied comparisons, with the noted exception.

It is noted that some estimates have high standard errors, indicating instability of the model coefficients.

Table 2.22: Differential Test Functioning – Kindergarten Tests for Black vs. White Students (Black), Latino vs. Non-Latino Students (Latino), Female vs. Male Students (Female) and Dual Language Learners vs. Non-Dual Language Learners (DLL)

Time-Status	Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fall - PWR	(Intercept)	-1.39	0.79	-1.75	0.080
Fall - PWR	fallpwrrisk	1.29	0.90	1.43	0.153
Fall - PWR	Black	-0.20	0.84	-0.23	0.815
Fall - PWR	fallpwrrisk*Black	0.15	0.98	0.15	0.880
Fall - DYS	(Intercept)	-18.57	1630.66	-0.01	0.991
Fall - DYS	falldysrisk	16.69	1630.66	0.01	0.992
Fall - DYS	Black	15.86	1630.66	0.01	0.992
Fall - DYS	falldysrisk*Black	-15.45	1630.66	-0.01	0.992
Winter - PWR	(Intercept)	-1.79	0.76	-2.35	0.019
Winter - PWR	winterpwrrisk	2.15	0.91	2.36	0.018
Winter - PWR	Black	0.50	0.81	0.62	0.534
Winter - PWR	winterpwrrisk*Black	-1.29	0.98	-1.32	0.188
Winter - DYS	(Intercept)	-18.57	1537.40	-0.01	0.990
Winter - DYS	winterdysrisk	16.86	1537.40	0.01	0.991
Winter - DYS	Black	15.78	1537.40	0.01	0.992
Winter - DYS	winterdysrisk*Black	-15.11	1537.40	-0.01	0.992
Fall - PWR	(Intercept)	-2.89	1.03	-2.81	0.005
Fall - PWR	fallpwrrisk	3.09	1.12	2.76	0.006
Fall - PWR	Latino	1.44	1.07	1.34	0.180
Fall - PWR	fallpwrrisk*Latino	-1.87	1.20	-1.56	0.119
Fall - DYS	(Intercept)	-2.53	0.73	-3.44	0.001
Fall - DYS	falldysrisk	0.13	1.28	0.10	0.920
Fall - DYS	Latino	-0.34	0.87	-0.39	0.693
Fall - DYS	falldysrisk*Latino	1.41	1.45	0.97	0.332
Winter - PWR	(Intercept)	-1.16	0.51	-2.27	0.023
Winter - PWR	winterpwrrisk	0.71	0.70	1.01	0.313
Winter - PWR	Latino	-0.09	0.59	-0.15	0.878
Winter - PWR	winterpwrrisk*Latino	0.04	0.82	0.05	0.959
Winter - DYS	(Intercept)	-2.27	0.61	-3.74	0.000
Winter - DYS	winterdysrisk	-15.30	1495.30	-0.01	0.992
Winter - DYS	Latino	-0.91	0.79	-1.15	0.251
Winter - DYS	winterdysrisk*Latino	17.87	1495.30	0.01	0.990
Fall - PWR	(Intercept)	-1.15	0.27	-4.26	0.000
Fall - PWR	fallpwrrisk	1.01	0.41	2.48	0.013
Fall - PWR	Female	-0.46	0.45	-1.01	0.315
Fall - PWR	fallpwrrisk*Female	0.51	0.62	0.82	0.413
Fall - DYS	(Intercept)	-3.07	0.51	-6.00	0.000
Fall - DYS	falldysrisk	0.95	0.80	1.19	0.234
Fall - DYS	Female	0.21	0.73	0.28	0.777
Fall - DYS	falldysrisk*Female	0.21	1.09	0.19	0.848
Winter - PWR	(Intercept)	-1.06	0.27	-3.99	0.000
Winter - PWR	winterpwrrisk	0.79	0.40	1.95	0.051
Winter - PWR	Female	-0.08	0.41	-0.20	0.840
Winter - PWR	winterpwrrisk*Female	-0.03	0.60	-0.05	0.961
Winter - DYS	(Intercept)	-2.77	0.42	-6.59	0.000
Winter - DYS	winterdysrisk	0.06	1.12	0.06	0.954
Winter - DYS	Female	-0.88	0.83	-1.06	0.291
Winter - DYS	winterdysrisk*Female	2.67	1.41	1.89	0.058
Fall - PWR	(Intercept)	-1.30	0.23	-5.64	0.000
Fall - PWR	fallpwrrisk	1.25	0.33	3.83	0.000
Fall - PWR	DLL	-0.15	0.60	-0.25	0.806

Fall - PWR	fallpwrrisk*DLL	-0.21	0.83	-0.25	0.805
Fall - DYS	(Intercept)	-2.94	0.39	-7.59	0.000
Fall - DYS	falldysrisk	1.18	0.56	2.09	0.037
Fall - DYS	DLL	-0.39	1.09	-0.36	0.722
Fall - DYS	falldysrisk*DLL	0.36	1.59	0.23	0.819
Winter - PWR	(Intercept)	-1.10	0.22	-5.03	0.000
Winter - PWR	winterpwrrisk	0.83	0.32	2.62	0.009
Winter - PWR	DLL	-0.24	0.55	-0.43	0.666
Winter - PWR	winterpwrrisk*DLL	0.16	0.83	0.20	0.844
Winter - DYS	(Intercept)	-3.19	0.42	-7.66	0.000
Winter - DYS	winterdysrisk	1.94	0.58	3.35	0.001
Winter - DYS	DLL	0.45	0.84	0.54	0.591
Winter - DYS	winterdysrisk*DLL	-15.76	1385.38	-0.01	0.991

Note: Sample size consists of 225 total students. Of the 225 students: 144 students=White, 31 students=Black, 39 students=Latino, 118 students=Female, and 36 students=Dual Language Learners.

Table 2.23: Differential Test Functioning – Grade 1 Tests for Black vs. White Students (Black), Latino vs. Non-Latino Students (Latino), Female vs. Male Students (Female) and Dual Language Learners vs. Non-Dual Language Learners (DLL)

Time-Status	Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fall - PWR	(Intercept)	-17.57	1021.48	-0.02	0.986
Fall - PWR	fallpwrrisk	18.48	1021.48	0.02	0.986
Fall - PWR	Black	15.75	1021.48	0.02	0.988
Fall - PWR	fallpwrrisk*Black	-15.82	1021.48	-0.02	0.988
Fall - DYS	(Intercept)	-18.57	1537.40	-0.01	0.990
Fall - DYS	falldysrisk	19.13	1537.40	0.01	0.990
Fall - DYS	Black	15.67	1537.40	0.01	0.992
Fall - DYS	falldysrisk*Black	-17.12	1537.40	-0.01	0.991
Winter - PWR	(Intercept)	-17.57	1057.33	-0.02	0.987
Winter - PWR	winterpwrrisk	18.26	1057.33	0.02	0.986
Winter - PWR	Black	15.38	1057.33	0.01	0.988
Winter - PWR	winterpwrrisk*Black	-15.14	1057.33	-0.01	0.989
Winter - DYS	(Intercept)	-18.57	1581.97	-0.01	0.991
Winter - DYS	winterdysrisk	18.90	1581.97	0.01	0.990
Winter - DYS	Black	14.96	1581.97	0.01	0.992
Winter - DYS	winterdysrisk*Black	-15.92	1581.97	-0.01	0.992
Fall - PWR	(Intercept)	-1.01	0.58	-1.73	0.083
Fall - PWR	fallpwrrisk	2.40	0.87	2.75	0.006
Fall - PWR	Latino	-0.81	0.66	-1.22	0.222
Fall - PWR	fallpwrrisk*Latino	0.02	0.97	0.02	0.982
Fall - DYS	(Intercept)	-2.77	1.03	-2.69	0.007
Fall - DYS	falldysrisk	2.62	1.17	2.24	0.025
Fall - DYS	Latino	-0.04	1.11	-0.04	0.971
Fall - DYS	falldysrisk*Latino	-0.90	1.32	-0.69	0.493
Winter - PWR	(Intercept)	-1.70	0.77	-2.22	0.027
Winter - PWR	winterpwrrisk	3.25	1.00	3.25	0.001
Winter - PWR	Latino	-0.65	0.86	-0.76	0.450
Winter - PWR	winterpwrrisk*Latino	-0.14	1.11	-0.12	0.902
Winter - DYS	(Intercept)	-18.57	1581.97	-0.01	0.991
Winter - DYS	winterdysrisk	18.72	1581.97	0.01	0.991
Winter - DYS	Latino	15.08	1581.97	0.01	0.992
Winter - DYS	winterdysrisk*Latino	-16.07	1581.97	-0.01	0.992
Fall - PWR	(Intercept)	-2.13	0.37	-5.68	0.000
Fall - PWR	fallpwrrisk	2.74	0.50	5.49	0.000
Fall - PWR	Female	0.24	0.52	0.47	0.636
Fall - PWR	fallpwrrisk*Female	-0.24	0.69	-0.35	0.729
Fall - DYS	(Intercept)	-3.76	0.72	-5.26	0.000
Fall - DYS	falldysrisk	3.39	0.82	4.15	0.000
Fall - DYS	Female	1.03	0.85	1.21	0.228
Fall - DYS	falldysrisk*Female	-1.29	1.02	-1.27	0.202
Winter - PWR	(Intercept)	-2.80	0.51	-5.44	0.000
Winter - PWR	winterpwrrisk	3.50	0.60	5.79	0.000
Winter - PWR	Female	0.30	0.69	0.44	0.663
Winter - PWR	winterpwrrisk*Female	-0.09	0.83	-0.11	0.910
Winter - DYS	(Intercept)	-19.57	1159.64	-0.02	0.987
Winter - DYS	winterdysrisk	19.36	1159.64	0.02	0.987
Winter - DYS	Female	16.60	1159.64	0.01	0.989
Winter - DYS	winterdysrisk*Female	-16.88	1159.64	-0.01	0.988
Fall - PWR	(Intercept)	-1.98	0.28	-7.20	0.000
Fall - PWR	fallpwrrisk	2.70	0.37	7.33	0.000
Fall - PWR	DLL	-0.16	0.80	-0.20	0.844

Fall - PWR	fallpwrrisk*DLL	-0.15	1.05	-0.14	0.886
Fall - DYS	(Intercept)	-3.00	0.39	-7.75	0.000
Fall - DYS	falldysrisk	2.61	0.48	5.45	0.000
Fall - DYS	DLL	-15.56	1360.06	-0.01	0.991
Fall - DYS	falldysrisk*DLL	15.26	1360.06	0.01	0.991
Winter - PWR	(Intercept)	-2.74	0.39	-7.02	0.000
Winter - PWR	winterpwrrisk	3.56	0.46	7.81	0.000
Winter - PWR	DLL	0.48	0.84	0.58	0.564
Winter - PWR	winterpwrrisk*DLL	-0.21	1.19	-0.17	0.863
Winter - DYS	(Intercept)	-3.57	0.51	-7.04	0.000
Winter - DYS	winterdysrisk	3.35	0.58	5.81	0.000
Winter - DYS	DLL	-15.00	1360.06	-0.01	0.991
Winter - DYS	winterdysrisk*DLL	14.53	1360.06	0.01	0.991

Sample size consists of 229 total students. Of the 229 students: 153 students=White, 29 students=Black, 30 students=Latino, 115 students=Female, and 29 students=Dual Language Learners.

Table 2.24: Differential Test Functioning – Grade 2 Tests for Black vs. White Students (Black), Latino vs. Non-Latino Students (Latino), Female vs. Male Students (Female) and Dual Language Learners vs. Non-Dual Language Learners (DLL)

Time-Status	Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fall - PWR	(Intercept)	-2.20	0.75	-2.95	0.003
Fall - PWR	fallpwrrisk	3.70	0.93	3.99	0.000
Fall - PWR	Black	-0.36	0.84	-0.42	0.672
Fall - PWR	fallpwrrisk*Black	-0.74	1.04	-0.71	0.477
Fall - DYS	(Intercept)	-18.57	1279.19	-0.01	0.988
Fall - DYS	falldysrisk	20.03	1279.19	0.02	0.988
Fall - DYS	Black	14.57	1279.19	0.01	0.991
Fall - DYS	falldysrisk*Black	-16.18	1279.20	-0.01	0.990
Winter - PWR	(Intercept)	-2.89	1.03	-2.81	0.005
Winter - PWR	winterpwrrisk	4.45	1.17	3.82	0.000
Winter - PWR	Black	-0.31	1.15	-0.27	0.788
Winter - PWR	winterpwrrisk*Black	-0.31	1.31	-0.23	0.816
Winter - DYS	(Intercept)	-20.57	3477.21	-0.01	0.995
Winter - DYS	winterdysrisk	22.03	3477.21	0.01	0.995
Winter - DYS	Black	0.00	3853.39	0.00	1.000
Winter - DYS	winterdysrisk*Black	-1.26	3853.39	0.00	1.000
Fall - PWR	(Intercept)	-2.74	0.73	-3.76	0.000
Fall - PWR	fallpwrrisk	3.19	0.88	3.65	0.000
Fall - PWR	Latino	0.27	0.84	0.32	0.749
Fall - PWR	fallpwrrisk*Latino	-0.20	1.02	-0.19	0.846
Fall - DYS	(Intercept)	-2.40	0.60	-3.98	0.000
Fall - DYS	falldysrisk	2.26	0.79	2.85	0.004
Fall - DYS	Latino	-2.08	1.17	-1.77	0.076
Fall - DYS	falldysrisk*Latino	2.15	1.33	1.61	0.107
Winter - PWR	(Intercept)	-3.56	1.01	-3.51	0.000
Winter - PWR	winterpwrrisk	4.94	1.20	4.11	0.000
Winter - PWR	Latino	0.34	1.17	0.29	0.774
Winter - PWR	winterpwrrisk*Latino	-0.81	1.38	-0.58	0.559
Winter - DYS	(Intercept)	-2.97	0.73	-4.10	0.000
Winter - DYS	winterdysrisk	4.36	1.07	4.06	0.000
Winter - DYS	Latino	-17.60	1879.42	-0.01	0.993
Winter - DYS	winterdysrisk*Latino	16.27	1879.42	0.01	0.993
Fall - PWR	(Intercept)	-3.12	0.59	-5.29	0.000
Fall - PWR	fallpwrrisk	4.09	0.67	6.13	0.000
Fall - PWR	Female	0.89	0.71	1.25	0.210
Fall - PWR	fallpwrrisk*Female	-1.21	0.84	-1.43	0.151
Fall - DYS	(Intercept)	-19.57	1194.89	-0.02	0.987
Fall - DYS	falldysrisk	20.01	1194.89	0.02	0.987
Fall - DYS	Female	16.57	1194.89	0.01	0.989
Fall - DYS	falldysrisk*Female	-16.89	1194.89	-0.01	0.989
Winter - PWR	(Intercept)	-3.57	0.72	-4.98	0.000
Winter - PWR	winterpwrrisk	4.81	0.79	6.05	0.000
Winter - PWR	Female	0.42	0.93	0.45	0.651
Winter - PWR	winterpwrrisk*Female	-0.47	1.05	-0.44	0.659
Winter - DYS	(Intercept)	-19.57	1166.44	-0.02	0.987
Winter - DYS	winterdysrisk	20.30	1166.44	0.02	0.986
Winter - DYS	Female	15.82	1166.44	0.01	0.989
Winter - DYS	winterdysrisk*Female	-15.91	1166.44	-0.01	0.989
Fall - PWR	(Intercept)	-2.64	0.37	-7.21	0.000
Fall - PWR	fallpwrrisk	3.45	0.43	7.98	0.000
Fall - PWR	DLL	0.29	0.83	0.35	0.727

Fall - PWR	fallpwrrisk*DLL	-0.41	1.22	-0.34	0.737
Fall - DYS	(Intercept)	-4.23	0.71	-5.95	0.000
Fall - DYS	falldysrisk	4.59	0.75	6.09	0.000
Fall - DYS	DLL	1.79	1.03	1.75	0.080
Fall - DYS	falldysrisk*DLL	-2.15	1.45	-1.48	0.139
Winter - PWR	(Intercept)	-3.19	0.46	-6.98	0.000
Winter - PWR	winterpwrrisk	4.45	0.53	8.43	0.000
Winter - PWR	DLL	-15.38	1458.51	-0.01	0.992
Winter - PWR	winterpwrrisk*DLL	14.80	1458.51	0.01	0.992
Winter - DYS	(Intercept)	-20.57	1457.43	-0.01	0.989
Winter - DYS	winterdysrisk	21.41	1457.43	0.01	0.988
Winter - DYS	DLL	18.17	1457.43	0.01	0.990
Winter - DYS	winterdysrisk*DLL	-19.42	1457.43	-0.01	0.989

Note: Sample size consists of 238 total students. Of the 238 students: 152 students=White, 42 students=Black, 51 students=Latino, 122 students=Female, and 29 students=Dual Language Learners.

Table 2.25: Differential Test Functioning – Grade 3 Tests for Black vs. White Students (Black), Latino vs. Non-Latino Students (Latino), Female vs. Male Students (Female) and Dual Language Learners vs. Non-Dual Language Learners (DLL)

Time-Status	Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fall - PWR	(Intercept)	-2.40	1.04	-2.30	0.022
Fall - PWR	fallpwrrisk	3.70	1.23	3.00	0.003
Fall - PWR	Black	0.10	1.12	0.09	0.932
Fall - PWR	fallpwrrisk*Black	-1.70	1.35	-1.26	0.208
Fall - DYS	(Intercept)	-2.48	1.04	-2.39	0.017
Fall - DYS	falldysrisk	2.64	1.18	2.24	0.025
Fall - DYS	Black	-0.76	1.20	-0.64	0.525
Fall - DYS	falldysrisk*Black	-0.02	1.39	-0.02	0.987
Winter - PWR	(Intercept)	-18.57	1966.65	-0.01	0.992
Winter - PWR	winterpwrrisk	19.95	1966.65	0.01	0.992
Winter - PWR	Black	15.68	1966.65	0.01	0.994
Winter - PWR	winterpwrrisk*Black	-16.99	1966.65	-0.01	0.993
Winter - DYS	(Intercept)	-18.57	1743.25	-0.01	0.992
Winter - DYS	winterdysrisk	19.26	1743.25	0.01	0.991
Winter - DYS	Black	14.90	1743.25	0.01	0.993
Winter - DYS	winterdysrisk*Black	-16.04	1743.25	-0.01	0.993
Fall - PWR	(Intercept)	-2.30	0.74	-3.10	0.002
Fall - PWR	fallpwrrisk	2.12	0.96	2.21	0.027
Fall - PWR	Latino	-0.13	0.88	-0.15	0.881
Fall - PWR	fallpwrrisk*Latino	-0.14	1.17	-0.12	0.906
Fall - DYS	(Intercept)	-3.04	1.02	-2.97	0.003
Fall - DYS	falldysrisk	2.48	1.20	2.07	0.038
Fall - DYS	Latino	-0.41	1.25	-0.32	0.746
Fall - DYS	falldysrisk*Latino	-0.05	1.52	-0.03	0.976
Winter - PWR	(Intercept)	-18.57	1360.06	-0.01	0.989
Winter - PWR	winterpwrrisk	19.41	1360.06	0.01	0.989
Winter - PWR	Latino	15.93	1360.06	0.01	0.991
Winter - PWR	winterpwrrisk*Latino	-17.18	1360.06	-0.01	0.990
Winter - DYS	(Intercept)	-19.57	2109.04	-0.01	0.993
Winter - DYS	winterdysrisk	20.48	2109.04	0.01	0.992
Winter - DYS	Latino	16.15	2109.04	0.01	0.994
Winter - DYS	winterdysrisk*Latino	-18.24	2109.04	-0.01	0.993
Fall - PWR	(Intercept)	-1.75	0.36	-4.86	0.000
Fall - PWR	fallpwrrisk	1.84	0.55	3.34	0.001
Fall - PWR	Female	-2.20	1.07	-2.05	0.040
Fall - PWR	fallpwrrisk*Female	2.48	1.22	2.04	0.041
Fall - DYS	(Intercept)	-2.71	0.52	-5.24	0.000
Fall - DYS	falldysrisk	2.51	0.68	3.66	0.000
Fall - DYS	Female	-16.86	1424.40	-0.01	0.991
Fall - DYS	falldysrisk*Female	16.62	1424.40	0.01	0.991
Winter - PWR	(Intercept)	-2.56	0.52	-4.94	0.000
Winter - PWR	winterpwrrisk	3.00	0.65	4.64	0.000
Winter - PWR	Female	-17.00	1450.07	-0.01	0.991
Winter - PWR	winterpwrrisk*Female	17.32	1450.07	0.01	0.990
Winter - DYS	(Intercept)	-3.37	0.72	-4.68	0.000
Winter - DYS	winterdysrisk	3.20	0.83	3.87	0.000
Winter - DYS	Female	-16.20	1388.34	-0.01	0.991
Winter - DYS	winterdysrisk*Female	16.17	1388.34	0.01	0.991
Fall - PWR	(Intercept)	-2.10	0.33	-6.28	0.000
Fall - PWR	fallpwrrisk	2.51	0.45	5.55	0.000
Fall - PWR	DLL	-16.46	1390.63	-0.01	0.991

Fall - PWR	fallpwrrisk*DLL	14.67	1390.63	0.01	0.992
Fall - DYS	(Intercept)	-3.17	0.51	-6.21	0.000
Fall - DYS	falldysrisk	3.06	0.61	5.06	0.000
Fall - DYS	DLL	-15.40	1390.63	-0.01	0.991
Fall - DYS	falldysrisk*DLL	-3.06	3231.53	0.00	0.999
Winter - PWR	(Intercept)	-3.02	0.51	-5.90	0.000
Winter - PWR	winterpwrrisk	3.63	0.59	6.15	0.000
Winter - PWR	DLL	-15.55	1304.53	-0.01	0.990
Winter - PWR	winterpwrrisk*DLL	14.94	1304.53	0.01	0.991
Winter - DYS	(Intercept)	-3.83	0.71	-5.36	0.000
Winter - DYS	winterdysrisk	3.69	0.78	4.75	0.000
Winter - DYS	DLL	-15.74	2109.04	-0.01	0.994
Winter - DYS	winterdysrisk*DLL	-3.69	10958.87	0.00	1.000

Note: Sample size consists of 164 total students. Of the 164 students: 103 students=White, 26 students=Black, 33 students=Latino, 84 students=Female, and 27 students=Dual Language Learners.

2.2 Boston Children's Hospital Validation Study

The Gaab Lab (then at BCH) designed and executed two validation studies for the Imagine+ EarlyBird Screener over the course of the 2018-19 (pilot study; results available upon request) and 2019-20 (validation study) academic school year. This section describes the validation study.

2.2.1 Methods

The BCH validation study was designed as a nationwide study to assess predictive validity of the screener. The goal of the study was to measure the extent to which the Imagine+ EarlyBird Screener can predict end-of-year language and literacy outcomes when administered at the beginning of the Kindergarten school year. The first phase of predictive validation was completed between August and November 2019. We assessed 419 children (215 female, 200 male, 4 unknown, average age of 5.08 years; 2.26, 2.27, and 2.28) in 19 schools and eight states in every region of the country including MT, MO, MA, NY, LA, PA, RI, and TX. Using the same exclusion/inclusion criteria as the 2018/2019 validation study, we tested 100 children with some degree of familial history of dyslexia or reading difficulty and 328 without a familial history. 22.83% of parents reported their combined income; approximately 39% of those parents reported a combined income of less than $100K. Of the 94% of parents who reported their child's race and ethnicity, 34.42% identified their children as non-white or multiracial. Children were tested within an eight-week window after their first day of Kindergarten using all twelve assessments in the app, developed at BCH as well as Florida State University's (FSU) Florida Center for Reading Research. We added items to multiple assessment components that were previously validated at FSU.

Table 2.26: BCH Sample Characteristics Part 1

Category	MA	PA	RI	LA	MT	NY	MO	TX	Total
Phase 1	**117**	**84**	**40**	**23**	**43**	**40**	**47**	**25**	**419**
Female	54	46	23	14	23	18	27	10	215
Male	62	38	17	9	19	22	20	13	200
Sex N/A	1	0	0	0	1	0	0	2	4
FHD+	20	13	6	5	8	10	12	4	78
FHD−	97	71	34	18	35	30	35	21	341
Phase 2	**30**	**56**	**25**	**20**	**37**	**9**	**22**	**20**	**219**
Female	11	28	15	13	19	4	10	8	108
Male	19	28	10	7	17	5	12	10	108
Sex N/A	0	0	0	0	1	0	0	2	3
FHD+	6	10	4	4	7	2	6	3	42
FHD−	24	46	21	16	30	7	16	17	177

Note: MA = Massachusetts, PA = Pennsylvania, RI = Rhode Island, LA = Louisiana, MT = Montana, NY = New York, MO = Missouri, TX = Texas. FHD = Family History of Dyslexia. For the purpose of this paper, FHD+ is classified as participants with first degree relative with either a dyslexia diagnosis or reading difficulty. FHD- is classified as participants without first degree relative with either a dyslexia diagnosis or reading difficulty.

Table 2.27: BCH Sample Characteristics Part 2

Demographic	Category	N	%
Sex	Male	200	47.73
	Female	215	51.31
	N/A	4	0.95
Race/Ethnicity	White	339	75.50
	Black	58	12.92
	Asian	22	4.90
	Native American	11	2.45
	Native Hawaiian/Pacific Islander	4	0.89
	No Response	15	3.34
Hispanic/Latino/Spanish Origin	Yes	50	12.22
	No	329	80.44
	N/A	30	7.33
Family History	First degree relative - dyslexia	128	31.30
	Non first degree relative - dyslexia	0	0.00
	First degree relative - struggling reader	0	0.00
	Non first degree relative - struggling reader	0	0.00
	No diagnosis	29	7.09
	N/A	252	61.61
Language other than English	Yes	64	15.65
	No	344	84.11
	N/A	1	0.24
US Ladder*	1	5	1.15
	2	6	1.38
	3	5	1.15
	4	17	3.90
	5	48	11.01
	6	36	8.26
	7	33	7.57
	8	14	3.21
	9	1	0.23
	N/A	271	62.16

Note: *US Ladder: This question asked to place themselves on a scale from 1-9, relative to the other people in the United States, regarding money, education and job status. Higher the number, the closer they see themselves to people who have the most money, most education and most respected jobs. Likewise, lower the number, the closer they see themselves to people who have the least money, least education and least respected jobs or no job.

Table 2.28: BCH Sample Characteristics Part 2 (Cont.)

Demographic	Category	N	%
Household Occupation	Working full time	5	1.15
	Working part-time	29	6.65
	Unemployed or laid off	91	20.87
	Looking for work	52	11.93
	Keeping house or raising children full-time	14	3.21
	Retired	5	1.15
	N/A	240	55.05
Highest Degree Attained - Mother	8th grade or less	0	0.00
	Some high school	4	0.92
	High school diploma or GED	30	6.88
	Associate degree	38	8.72
	Bachelor's degree	70	16.06
	Master's degree	47	10.78
	Doctorate	2	0.46
	Professional	5	1.15
	N/A	240	55.05
Highest Degree Attained - Father	8th grade or less	1	0.23
	Some high school	6	1.38
	High school diploma or GED	67	15.37
	Associate degree	23	5.28
	Bachelor's degree	59	13.53
	Master's degree	31	7.11
	Doctorate	2	0.46
	Professional	2	0.46
	N/A	245	56.19
Family Combined Income	Less than $10,000	2	0.46
	$10,000 to $19,999	3	0.69
	$20,000 to $29,999	2	0.46
	$30,000 to $39,999	8	1.83
	$40,000 to $49,999	8	1.83
	$50,000 to $59,999	7	1.61
	$60,000 to $69,999	8	1.83
	$70,000 to $79,999	8	1.83
	$80,000 to $89,999	17	3.90
	$90,000 to $99,999	9	2.06
	$100,000 to $109,999	14	3.21
	$110,000 to $119,999	13	2.98
	$120,000 to $129,999	8	1.83
	$130,000 to $139,999	9	2.06
	$140,000 to $149,999	10	2.29
	$150,000 to $159,999	8	1.83
	$160,000 to $169,999	6	1.38
	$170,000 to $179,999	6	1.38
	$180,000 to $189,999	3	0.69
	$190,000 to $199,999	3	0.69
	$200,000 to $209,999	2	0.46
	$210,000 to $219,999	1	0.23
	$220,000 to $229,999	2	0.46
	$230,000 to $239,999	2	0.46
	$240,000 to $249,999	10	2.29
	$250,000 or greater	6	1.38
	Don't Know	17	3.90
	N/A	244	55.96

The initial plan (before COVID-19 restrictions) was to retest these participants in the spring of 2020 with the following comprehensive standardized early literacy outcomes assessment battery in order to assess predictive validity of the screener. With the onset of the COVID-19 pandemic, travel became restricted and schools were closed for in-person data collection. We were therefore unable to visit all of the locations in-person to collect data. Therefore, the team quickly adapted the entire psychometric assessments to be administered virtually over Zoom. After multiple pilot sessions, the virtual protocol was finalized in early May. The team used a combination of the screen share feature on Zoom and Q-global, which is Pearson's web-based administration system, to conduct these assessments. A detailed overview about the virtual assessment protocol can be found here: https://osf.io/wg4ef/. Children were tested either at home or in school and a variety of technology challenges were solved, especially in children from low-income family backgrounds.

From early May until the end of October, we were able to successfully collect follow-up data from 219 participants. Participants were tested in eight States (MA, NY, RI, MT, LA, MO, PA, TX) and 19 schools. Specifically, by spring we had 419 students still attending the schools where they were tested in the fall. Of the 419, we were able to test 219 either in-person or remotely, giving us a testing rate of 54%. 199 participants were not tested in this second phase of validation due to multiple reasons:

(1) Parents of the participants were not interested in participating in a virtual testing session or were unable to be reached by the research team. The original intent of the study was for the parents to have minimal participation in the study, as the bulk of the communication was to be done with schools. COVID-19 complications and school closures unexpectedly altered the study structure.

(2) Some schools were not interested in in-school virtual testing due to their increased work-load during remote schooling periods.

For the second phase of predictive validation, 219 participants (108 female, 108 male and three unknown, average age six years seven months) were tested with the follow-up paper-pencil psychometric battery in spring/summer 2020. There were 48 children with some degree of familial history of dyslexia or reading difficulty and 171 children without a familial history in the sample. 48% of parents reported their combined income and approximately 39% of those parents reported a combined income of less than $100k. 98% of parents reported their child's race and ethnicity. Of those parents, 12.3% identified their children as non-white or multiracial.

Participants were either tested in their homes virtually or in school virtually. Out of the 219 children tested in spring 2020, 105 were tested in person. These participants were tested by a psychometric tester in-person because they were either tested before schools were closed or tested by a local tester who didn't have travel/visiting restrictions at the school. All of the participants received a detailed score report that outlined the assessments that were administered with their respective scores. These score reports were also shared with the schools (if parents gave us the permission). Having two data points from 219 participants from both fall and spring within a single school year allowed for the evaluation of the screener's predictive validity.

Data were analyzed with classical test theory and multiple-group item response modeling (MG-IRM).

2.2.2 Results

Table 2.29 shows the CTT results for item p-values, which indicate items had acceptable percentages of correct responses. For this analysis, subtests were analyzed across grade levels.

For the Word Matching subtest, 11 items presented with ceiling effects (i.e., p-value greater than or equal to .95) and were removed from the item bank.

Table 2.29: Summary Statistics of CTT Results (p-values) for BCH Validation Study

Subtest	Mean	SD	Min	Max
First Sound Matching	0.59	0.15	0.39	0.91
Nonword Repetition	0.49	0.22	0.08	0.91
Rhyming	0.67	0.14	0.36	0.89
Word Matching	0.68	0.23	0.24	0.98

Table 2.30 summarizes results for the MG-IRM analyses. Again, analyses were performed across grade levels. For all subtests, model fit for the unidimensional 2PL-IRM resulted in a rejection of the null hypothesis of a correctly specified model; however, global fit suggested good model fit to the data. For Word Matching, one item was removed due to a = 14.88 exceeding conventional thresholds for local fit. Marginal reliability was .87 for the subtests of First Sound Matching, Nonword Repetition, and Word Matching; marginal reliability for Rhyming was .89.

Table 2.30: Summary of MG-IRM Results for BCH Validation Study

Subtest	M2	df	p	RMSEA	LB	UB	TLI	CFI
First Sound Matching	911.91	665	<.001	0.031	0.026	0.036	0.98	0.98
Nonword Repetition	996.47	740	<.001	0.030	0.025	0.035	0.98	0.98
Rhyming	925.36	665	<.001	0.032	0.027	0.067	0.98	0.98
Word Matching	1958.60	1710	<.001	0.019	0.016	0.024	0.97	0.97

M2 = M2 statistic, df = degrees of freedom, RMSEA = root mean square error of approximation, LB = RMSEA 95% confidence interval lower-bound, UB = RMSEA 95% confidence interval upper-bound, TLI = Tucker-Lewis index, CFI = comparative fit index.

Items were also analyzed for DIF. DIF testing for subtests that were developed and validated by BCH was estimated using the difR package (Magis et al., 2010) using the Mantel-Haenszel method (Mantel & Haenszel, 1959) for detecting uniform DIF. For each of the six MATRS tasks, DIF was tested for four primary contrasts: 1) male vs. female, 2) white vs. non-white, and 3) black vs. non-black. The Mantel-Haenszel chi-square statistic was reported for test by item and the chi-square was used to derive an effect size estimate (i.e., ETS delta scale; Holland & Thayer (1986)). Effect size values <= 1.0 are considered small, 1.0 – 1.5 is moderate, and >= 1.5 is considered large.

Across all tasks and comparisons, only 12 items demonstrated at DIF with at least a moderate effect size (i.e., ETS >= 1.0): 2 Nonword Repetition items, and 10 Word Matching items. These items were removed from the item bank for further study and testing. All remaining items presented with ETS delta values <1.00 indicating small DIF.

Following item analysis and decisions on retaining and removing items, the validation study also included an analysis of correlations among subtests scores, predictive validity, and classification accuracy.

2.2.3 Predictive Validity

The predictive validity of scores to the selected criterions were addressed through a series of analyses. First, a correlation analysis was used to evaluate the strength of relations between and among each the Imagine+ EarlyBird Screener test events and norm-referenced tests. Second, logistic regressions were composed to predict the binarized outcome variable of interest to provide information for selecting an optimal predictive algorithm. Statistically significant relations to the outcome were flagged for examination. Criteria for selecting the optimal algorithm included selecting the smallest (most parsimonious) combination of tests that best predicted the outcome, were statistically significantly related to the outcome, and were theoretically plausible. Finally, ordinary least squares regression analysis was composed to estimate the variance that the linear combination of selected predictors explained in the continuous variable criterion. The total amount of variance, squared variance (variance accounted for), and 95% confidence intervals were computed and reported.

Correlations between and among Imagine+ EarlyBird Screener ability scores with standardized outcomes ranged from -.41 between RAN and FSM to .92 between TOWRE SWE and K-LWR (Table 2.31. A series of multiple regression analyses tested for the additive and interactive relations between Imagine+ EarlyBird Screener test events and the K-PA outcome to find the fewest number of tasks that maximized the percentage of explained variance in K-PA. The final model included FD, NWR, and RHYM with R2 of .37 (multiple r = .61, 95% CI = .50, .69, n = 184).

Table 2.31: Means, Standard Deviations, and Correlations for Validation Sample

Subtest	M	SD	BL	DEL	FD	LN	LS	SS	VP	FSM	NWR	RHYM	WM	RAN	K.LWR	K.NWD	K.P	K.D	WID	WA	SWE	PDE
BL (N=210)	-2.23	2.38																				
DEL (N=101)	-1.14	1.41	0.44																			
FD (N=97)	-1.09	1.25	0.17	0.17																		
LN (N=94)	3.07	2.87	0.14	0.26*	0.27*																	
LS (N=213)	-0.08	1.60	0.39*	0.31*	0.18	0.23*																
SS (N=97)	-0.32	0.94	0.17	0.17	0.26*	0.18	0.25*															
VP (N=213)	0.33	1.64	0.20*	0.23*	0.09	0.22*	0.22*	0.34*														
FSM (N=191)	0.18	0.99	0.56*	0.37*	0.15	0.26*	0.51*	0.16	0.25*													
NWR (N=200)	0.12	0.85	0.48*	0.44*	0.16	0.27*	0.37*	0.16	0.24*	0.51*												
RHYM (N=195)	0.14	0.90	0.26*	0.23*	0.34*	0.47*	0.25*	0.2	0.38*	0.39*	0.43*											
WM (N=204)	0.09	0.89	0.30*	0.32*	0.34*	0.40*	0.30*	0.48*	0.34*	0.46*	0.51*	0.39*										
RAN (N=175)	84.27	27.86	-0.30*	-0.09	-0.09	-0.24*	-0.28*	-0.16	-0.22*	-0.41*	-0.23*	-0.23*	-0.25*									
K-LWR (N=139)	96.78	13.50	0.27*	0.14	0.33*	0.19	0.43*	0.28*	0.13	0.45*	0.25*	0.34*	0.22*	-0.26*								
K-NWD (N=111)	92.45	13.01	0.23*	0.12	0.21	0.28*	0.31*	0.26	0.15	0.46*	0.27*	0.42*	0.24*	-0.33*	0.86*							
K-P (N=215)	100.02	14.85	0.39*	0.46*	0.33*	0.30*	0.35*	0.17	0.27	0.51*	0.46*	0.53*	0.36*	-0.20*	0.63*	0.72*						
K-D (N=103)	93.17	14.74	0.31*	0.15	0.39*	0.23	0.31*	0.24	0.12	0.50*	0.24*	0.43*	0.23*	-0.40*	0.84*	0.81*	0.84*					
WID (N=192)	97.16	16.29	0.22*	0.05	0.34*	0.24*	0.34*	0.17	0.18*	0.48*	0.23*	0.41*	0.24*	-0.26*	0.91*	0.77*	0.62*	0.84*				
WA (N=192)	99.31	14.45	0.30*	0.03	0.29*	0.28*	0.27*	0.15	0.23*	0.49*	0.34*	0.47*	0.32*	-0.24*	0.79*	0.75*	0.71*	0.80*	0.81*			
SWE (N=108)	94.03	16.44	0.31*	0.23	0.44*	0.19	0.40*	0.2	0.22*	0.52*	0.30*	0.44*	0.26*	-0.27*	0.92*	0.79*	0.71*	0.88*	0.91*	0.77*		
PDE (N=108)	94.95	13.62	0.27*	0.18	0.39*	0.14	0.34*	0.05	0.18	0.54*	0.30*	0.44*	0.30*	-0.34*	0.82*	0.82*	0.71*	0.84*	0.77*	0.73*	0.85*	
CELF (N=219)	103.26	12.17	0.12	0.08	0.43*	0.37	0.25	0.38	0.28	0.27	0.44*	0.43*	0.45*	-0.02	0.21	0.34*	0.46*	0.31	0.21	0.31	0.29	0.36

M = mean, SD= standard deviation, BL = Blending, DEL = deletion, FD = following directions, LN = letter name knowledge, LS = letter sound knowledge, SS = sentence structure, VP = vocabulary pairs, FSM = first sound matching, NWR = nonword repetition, RHYM = rhyming, WM = Word Matching, K-LWR = KTEA letter word recognition, K-NWD = KTEA ?, K-P = KTEA phonological ?, K-D = KTEA Dyslexia, WID = WRMT Word Identification, WA = WRMT Word Attack, SWE = TOWRE Sight Word Efficiency, PDE = TOWRE Phoneme Deletion Efficiency, CELF = CELF Sentence Structure. *p < .05.

2.2.4 Classification Accuracy

Based on the combination of tests selected previously, logistic regressions were used, in part, to calibrate classification accuracy. Students' performance on the selected criterions were coded as '1' for performance at or above the 40th percentile on the CTOPP-2 Phonological Awareness Composite Score and '0' for scores that did not meet these criteria. In this way, the classification represents a prediction of success. Each dichotomous variable was then regressed on a combination of Imagine+ EarlyBird Screener assess-

ments that were identified via the predictive validity analyses. As such, students could be identified as not at-risk on the multifactorial combination of screening tasks via the joint probability and demonstrating adequate performance on the criterion (i.e., specificity or true-negatives), at-risk on the combination of screening task scores via the joint probability and not demonstrating adequate performance on the criterion (i.e., sensitivity or true-positives), not at-risk based on the combination of screening task scores but at-risk on a criterion (i.e., false negative error), or at-risk on the combination of screening task scores but not at-risk on the criterion (i.e., false positive error). Classification of students in these categories allows for the evaluation of cut-points on the combination of screening tasks to determine which were the cut-point maximizing selected indicators. The concept of risk or success can be viewed in many ways, including the concept as a "percent chance" which is a number between 1 and 99, with 1 meaning there is a low chance that a student may develop a problem, and 99 being there is a high chance that the student may develop a problem. When attempting to identify children who are "at-risk" for poor performance on some type of future measure of reading achievement, Imagine+ EarlyBird Screener uses a yes/no decision based upon a "cut-point" along a continuum of risk.

Decisions concerning appropriate cut-points are made based on the level of correct classification that is desired from the screening assessments. A variety of statistics may be used to guide such choices (e.g., accuracy, sensitivity, specificity, positive and negative predictive power; see Schatschneider, Petscher & Williams, 2008) and each was considered in light of the other in choosing appropriate cut-points. Area under the curve, sensitivity, and specificity estimates from the final logistic regression model were bootstrapped 1,000 times in order to obtain a 95% confidence interval of scores using the cutpointr package in R statistical software.

Using the predictive validity model, receiver operating characteristic curve analysis (Figure 2.1) estimated the AUC for the overall efficiency of discrimination of the log-odds from the multiple predictors (AUC = 0.85, 95% CI = 0.80, 0.90) with an optimal cut-point identified at .21. The resulting confusion matrix (Table 2.32) showed 24 students were classified as true positives, 124 students were classified as true negatives, 29 students were classified as false positives, and 7 students were classified as false negatives. Mean bootstrapped sensitivity (.81, 95% CI = .67, .94) and specificity (.80, 95% CI = .60, .90) were estimated along with matrix-based negative predictive power (.95), positive predictive power (.45), and overall correct classification (.80) were computed. The base rate in the sample was 16%.

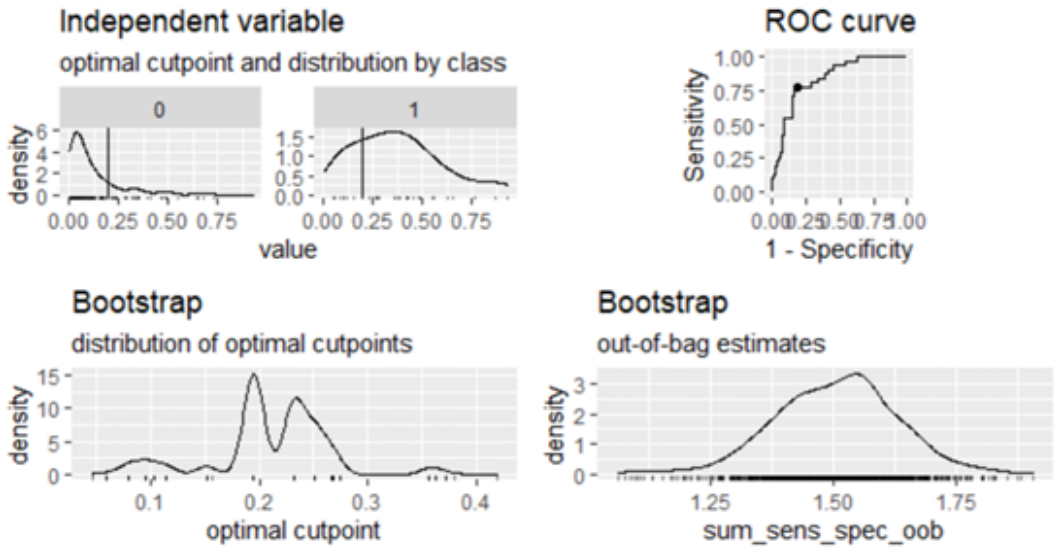


Figure 2.1: ROC Curve for Prediction of Severe Phonological Awareness Risk

Table 2.32: 2x2 Confusion Matrix of Dyslexia Risk Classification

Risk Status	<16th %ile	>=16th %ile	Row Total
At Risk (1)	24	29	53
Not At Risk (0)	7	124	131
Column Total	31	153	184

2.3 Florida State University Validation Study

An additional validation study was performed by Florida State University.

2.3.1 Methods

Data collection for the screener began by testing item pools for the Screen tasks (i.e., Word Matching and Follow Directions). A statewide representative sample of students that roughly reflected Florida's demographic diversity and academic ability (N ~ 2,400) was collected on students K-2. Because the samples used for data collection did not strictly adhere to the state distribution of demographics (i.e., percent limited English proficiency, Black, White, Latino, and eligible for free/reduced lunch), sample weights according to student demographics were used to inform the item and student parameter scores. Table 2.33 and Table 2.34 include the population values and derived weights applied to all analyses.

Table 2.33: U.S. Population-Based Weight Values

Race	FRL	ELL	Weight
White	Yes	Yes	0.67
White	Yes	No	17.87
White	No	Yes	0.41
White	No	No	20.85
Black	Yes	Yes	1.55
Black	Yes	No	18.30
Black	No	Yes	0.10
Black	No	No	3.03
Hispanic	Yes	Yes	12.54
Hispanic	Yes	No	11.05
Hispanic	No	Yes	1.90
Hispanic	No	No	5.45
Other	Yes	Yes	0.51
Other	Yes	No	2.85
Other	No	Yes	0.43
Other	No	No	2.49

Population values for each grade for each of the sixteen demographic groups pertaining to race/ethnicity (i.e., White, Black, Hispanic, Other), free/reduced lunch status (eligible or ineligible), and English language learner (identified or not identified). Note that not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among race/ethnicity, free/reduced lunch status, and English language learner status.FRL = Free/reduced price lunch; ELL = English language learner.

Table 2.34: U.S. Population-Based Weight Values, by Subtest

Race	FRL	ELL	Letter Names Sounds	Blending Deletion	Following Directions	Word Match	Word Reading	Sentence Comprehension
White	Yes	Yes	1.063	1.098	1.098	1.098	1.634	1.117
White	Yes	No	0.824	0.800	0.802	0.802	0.871	0.796
White	No	Yes	0.891	0.854	0.854	0.854	1.640	0.854
White	No	No	0.681	0.672	0.672	0.672	0.698	0.675
Black	Yes	Yes	3.370	3.605	3.605	3.605	3.780	3.523
Black	Yes	No	1.442	1.395	1.386	1.386	1.340	1.375
Black	No	Yes	0.769	0.769	0.769	0.769	0.667	0.769
Black	No	No	0.935	0.921	0.932	0.932	0.977	0.927
Hispanic	Yes	Yes	1.507	1.972	1.972	1.912	1.365	1.903
Hispanic	Yes	No	1.565	1.528	1.520	1.520	1.469	1.535
Hispanic	No	Yes	2.836	2.754	2.754	2.754	6.333	2.714
Hispanic	No	No	1.298	1.352	1.369	1.369	1.219	1.342
Other	Yes	Yes	0.927	0.911	0.911	0.911	0.836	0.895
Other	Yes	No	0.617	0.609	0.610	0.622	0.604	0.640
Other	No	Yes	0.782	0.768	0.768	0.768	1.049	0.754
Other	No	No	0.604	0.570	0.553	0.571	0.563	0.582

FRL = Free/reduced price lunch; ELL = English language learner. Large sample weights reflect subgroups which needed to be weighted more in the analyses; however, a large value does not necessarily indicate gross under-sampling. Note that not all race/ethnicity subgroups are represented due to limited information provided when evaluating interactions among race/ethnicity, free/reduced lunch status, and English language learner status.

A common-item, non-equivalent groups design was used for collecting data in our pilot,

calibration, and validation studies. A strength of this approach is that it allows for linking multiple test forms via common items. For each task, a minimum of twenty-percent of the total items within a form were identified as vertical linking items to create a vertical scale. These items served a dual purpose of not only linking forms across grades to each other, but also linking forms within grades to each other. Because the tasks in the screener were each designed for vertical equating and scaling, we considered two primary frameworks for estimating the item parameters: 1) a multiple-group IRT of all test forms or 2) test characteristic curve equating. We chose the latter approach using Stocking and Lord (Stocking & Lord, 1983) to place the items on a common scale. All item analyses were conducted using Mplus software (Muthén & Muthén, 2008) with a 2PL independent items model.

For norming study analyses, data was collected on approximate 2,000 students in Kindergarten through Grade 2 across multiple districts in Florida who participated in the calibration and validation studies. Table 2.35 provides a breakdown of the sample sizes used by grade level for each of the adaptive tasks.

Table 2.35: Sample Size for Norms

Grade	PA	LN/LS	SC	WM	FD	WR
K	2100	2377	2275	2015	2304	1969

PA = Phonological Awareness Blending and Deletion, LN/LS = Letter Names and Sounds, SC = Sentence Comprehension, WM = Word Matching, FD = Following Directions, WR = Word Reading.

2.3.2 Results

Reliability

Marginal reliability coefficients for student performance in the spring Imagine+ EarlyBird Screener are reported in 2.36. Marginal reliability for the computer-adaptive tasks was quite high, ranging from .85 for Letter Names to .99 for Phonological Awareness - Blending. Values of .80 are typically viewed as acceptable for research purposes while estimates at .90 or greater are acceptable for clinical decision making (Nunnally & Bernstein, 1994).

Table 2.36: Marginal reliability coefficients for tasks

Grade	Task	Reliability
K	Phonological Awareness	0.94
K	Letter Sounds	0.78*
K	Vocabulary Pairs	0.86
K	Following Directions	0.94
K	Word Reading	0.96
K	Sentence Comprehension	0.90
1	Word Reading	0.94
1	Vocabulary Pairs	0.86
1	Following Directions	0.93
1	Sentence Comprehension	0.89
2	Word Reading	0.88
2	Spelling	0.91
2	Vocabulary Pairs	0.85
2	Following Directions	0.93

*Letter Sounds is not administered at the Spring assessment. Marginal reliability is provided as a function of performance on fixed items administered at the Fall and Winter assessments.

Predictive Validity

The predictive validity of the Screening tasks to the SAT-10 Word Reading (SESAT in K) was addressed through a series of linear and logistic regressions. The linear regressions were run two ways. First, a correlation analysis was used to evaluate the strength of relations between each of the Screening task ability scores with SESAT. Pearson correlations between screener tasks and the SESAT Word Reading task ranged from .38 to .59 (Table 2.37). Second, a multiple regression was run to estimate the total amount of variance that the linear combination of the predictors explained in SESAT (46%).

Table 2.37: Bivariate Correlations Between Screener Tasks and SESAT

Grade	PA-D	LS	WM	FD	WR	
K	0.59	0.51	0.38	0.46	0.48	0.46

Correlations and multiple regression are a function of sScreening tasks at the winter assessment and SESAT and SAT-10 in the spring. Kindergarten predictors for the multiple regression analysis include all predictors except word reading. Correlation is a function of Word Reading performance in the spring. PA-D = phonological awareness deletion, LS = letter sounds, WM = word match, FD = following directions, WR = word reading.

Classification Accuracy

For the logistic regressions, students' performance on the SESAT Word Reading test was coded as '1' for performance at or above the 40th percentile, and '0' for scores below this target. This dichotomous variable was then regressed on a combination of screener tasks. By dichotomizing scores on the screener as '1' for not at-risk for reading difficulties and '0' for at-risk for reading difficulties, students could be classified based on their dichotomized performances on both the screening tasks and the SESAT. As such, students could be identified as not at-risk on the combination of screening tasks and demonstrating grade level performance on the SESAT (i.e., specificity or true-negatives), at-risk on the combination of screening task scores and below grade level performance on the SESAT

(i.e., sensitivity or true-positives), not at-risk based on the combination of screening task scores but not at grade level on the SESAT (i.e., false negative error), or at-risk on the combination of screening task scores but at grade level on the SESAT (i.e., false positive error). Classification of students in these categories allows for the evaluation of cut-points on the combination of screening tasks to determine which cut-point maximizes predictive power.

Classification accuracy for the fall/winter screener (Table 2.38) included area under the curve (AUC) = .84 (95% CI = .81, .88), sensitivity = .81, specificity = .72, positive predictive power = .67, negative predictive power = .84, and overall correct classification = .76 with a sample base-rate of 41.8% approximating our 40th percentile normative cut-point on the SESAT. Classification accuracy for the spring screener (Table 2.38) included area under the curve (AUC) = .88 (95% CI = .86, .90), sensitivity = .81, specificity = .87, positive predictive power = .83, negative predictive power = .86, and overall correct classification = .85 with a sample base-rate of 41.8% approximating our 40th percentile normative cut-point on the SESAT.

Table 2.38: Classification Accuracy of the Screener Tasks in K

Time Point	AUC (CI)	SE	SP	PPP	NPP	OCC	Base Rate
Fall/Winter	.84 (.81, .88)	0.81	0.72	0.67	0.84	0.76	41.8
Spring	.88 (.86, .90)	0.81	0.87	0.83	0.86	0.85	41.8

AUC = area under the curve, CI = 95% confidence interval, SE= Sensitivity, SP = Specificity, PPP = Positive Predictive Power, NPP = Negative Predictive Power, OCC = Overall Correct Classification.

DIF

DIF testing for subtests that were developed and validated by FSU was conducted comparing: Black-White students, Latino-White students, Black-Latino students, students eligible for Free or Reduced Priced Lunch (FRL) with students not receiving FRL, and English Language Learner to non-English Language Learner students. DIF testing in the screener study was conducted with a multiple indicator multiple cause analysis in Mplus (Muthén & Muthén, 2008); moreover, a series of four standardized and expected score effect size measures were generated using VisualDF software (Meade, 2010) to quantify various technical aspects of score differentiation between the gender groups. First, the signed item difference in the sample (SIDS) index was created, which describes the average unstandardized difference in expected scores between the groups. The second effect size calculated was the unsigned item difference in the sample (UIDS). This index can be utilized as supplementary to the SIDS. When the absolute value of the SIDS and UIDS values are equivalent, the differential functioning between groups is equivalent; however, when the absolute value of the UIDS is larger than SIDS, it provides evidence that the item characteristic curves for expected score differences cross, indicating that differences in the expected scores between groups change across the level of the latent ability score. The D-max index is reported as the maximum SIDS value in the sample, and may be interpreted as the greatest difference for any individual in the sample in the expected response. Lastly, an expected score standardized difference was generated, and was computed similar to a Cohen's d statistic (Cohen, 2013). As such, it is interpreted as a measure of standard deviation difference between the groups for the expected score response with values of .2 regarded as small, .5 as medium, and .8 as large. Items demonstrating DIF were flagged for further study in order to ascertain why groups with the same latent ability performed differently on the items.

Differential Test Functioning

An additional component of checking the validity of cut-points and scores on the assessments involved testing differential accuracy of the regression equations across different demographic groups. This procedure involved a series of logistic regressions predicting success on the SESAT (i.e., at or above the 40th percentile). The independent variables included a variable that represented whether students were identified as not at-risk based on the identified cut-point on a combination score of the screening tasks, a variable that represented a selected demographic group, as well as an interaction term between the two variables. A statistically significant interaction term would suggest that differential accuracy in predicting end-of-year risk status existed for different groups of individuals based on the risk status identified by the screener. Differential accuracy was separately tested for Black and Latino students as well as for students identified as English Language Learners (ELL) and students who were eligible for Free/Reduced Price Lunch (FRL). No significant differential accuracy was found for any demographic sub-group (individual tables available upon request).

Concurrent Correlations

Reading and language skills tend to have moderate associations between them; thus, the expectation of the screening tasks scores is that moderate correlations would be observed. Correlation results are reported in Table 2.39. Word Matching, Following Directions, and Sentence Comprehension are receptive tasks and are therefore more highly related oral language measures. Additionally, the higher correlation was observed in a meta-analysis in the early grades (Weiser & Mathes, 2011).

Table 2.39: Bivariate Associations Among Screener Computer-Adaptive Tasks in Kindergarten

Assessment	PA-D	FD	WM	WR
PA-D	1.00	—	—	—
FD	0.44	1.00	—	—
WM	0.31	0.49	1.00	—
WR	0.45	0.35	0.29	1.00
SC	0.34	0.61	0.44	0.27

Correlations are estimated as a function of spring testing. PA-D = phonological awareness deletion, FD = following directions, WM = word match, WR = word reading, SC = sentence comprehension.

2.4 EarlyBird 2022 Validation Study

Additional validation was completed internally in 2022 with data from the 2021-22 school year.

For Pre-K, the EarlyBird Validation Study was designed to assess construct and predictive validity of the new Pre-K version of the EarlyBird assessment. Subtests originally developed at FSU were calibrated for Pre-K, and additional items were created. Additionally, a shorter version of Object RAN and new items were created for the Rhyming, First Sound Matching, and Vocabulary subtests (originally developed by BCH). These were designed and reviewed by content area experts to be developmentally appropriate for Pre-K students and were validated through the EarlyBird study.

For Grade 1, the EarlyBird Validation Study was designed to assess construct and predictive validity of the new first grade version of the assessment, as well as to validate two new subtests (Nonword Reading and Letter RAN) and new items for the Nonword Repetition subtest, developed for first grade students.

2.4.1 Participants and Procedures

For Pre-K, a total of 274 Pre-K students (142 female, 132 male) in 14 schools across 9 states spanning the major geographic regions of the United States (CA, GA, LA, MA, ME, MT, WI, MS, PA) participated in the 2022 EarlyBird study Table 2.40 summarizes the distribution of participants across districts and forms. A total of 78 participants (28% of the total sample) had a family history of reading difficulties. Of those, 50 participants (18% of the total sample) reported that a family member had been diagnosed with dyslexia. Socioeconomic status data was applied based on the school sites; approximately 56% of the students in the sample attended schools with Title I designation. The sample was made up of approximately 4% American Indian or Alaska Native students, 1% Asian students, 12% Black or African American students, 3% Hispanic or Latinx students, 65% White students, and 8% multiracial students. (7% of participants chose not to disclose this information.)

Table 2.40: Pre-K Item Calibration, Students from Each School Administered Forms A-D

District ID	School ID	Form A	Form B	Form C	Form D
17	170381	2	1	4	2
	170551	11	12	11	9
	170771	2	10	3	8
	171261	7	14	9	9
29	290070	6	12	9	6
	290282	15	10	7	12
	290841	8	6	10	9
	292291	10	6	9	7
	292441	8	10	8	6
	293521	11	16	8	11
	293641	9	7	6	10
	294201	9	12	12	5
	294921	7	15	13	12
99	990001	151	140	0	0
Total		256	271	109	106

The first phase of the validation study was completed between late January and early March 2022. Participants were administered 11 subtests. Of these, seven (Letter Name, Letter Sound, Blending, Deletion, Follow Directions, Oral Sentence Comprehension, and Word Matching) were computer adaptive and the other four (First Sound Matching, Rhyming, Vocab, and Object RAN) were fixed form tests. Each participant took either the Form A or Form B version of those subtests in the winter testing period and the other in the spring testing period. In addition to the EarlyBird app, 150 children were also administered a battery of paper/pencil tests to measure construct validity. The battery included the CTOPP-2, Preschool CELF-3, PPVT-5, and WRMT-III.

In the second phase of the study, between April and June 2022, participants were administered the app only. Based on feedback from teachers and testers during the first phase, the RAN subtest was modified to be shorter (40 items to name instead of 50).

The third and final phase of the Pre-K validation study took place between August and October 2022. 181 children, then in Kindergarten, were administered a predictive validity battery of paper/pencil tests, including the CTOPP-2, KTEA-3, TOPEL, and CELF Preschool-3. The remainder of the participants in the original sample were unable to be tested in fall 2022 due to changes in their geographic location and/or school enrollment. Upon analyzing the data from the study, some items and subtests were removed from the Pre-K version of the screener due to floor effects, DIF analysis results, and observations made during testing that indicated some tasks were not developmentally appropriate for the majority of Pre-K students. The remaining items and subtests are therefore highly vetted. Having two data points from 181 participants, from the app taken in the spring and from the psychometric assessments administered the following fall, allowed for the evaluation of the PreK screener's predictive validity. Table 2.41 summarizes the number of items and participants by form and subscale.

Table 2.41: Numbers of Items and Students in the Pre-K IRT Calibrations (Forms A-D)

Subscale	Form A	Form B	Form C	Form D	Anchor Items	Final Items
BL	11	10	8	10	3	30
(BL-Sample)	175	207	68	68	—	—
DEL	10	10	5	11	3	27
(DEL-Sample)	150	155	42	48	—	—
FD	22	21	20	22	8	61
(FD-Sample)	248	264	101	103	—	—
SS	21	23	20	19	8	59
(SS-Sample)	251	257	102	103	—	—
WM	23	23	18	14	8	54
(WM-Sample)	220	222	80	86	—	—
Total Items	87	87	71	76	30	231*

The original forms A-D contain 91 items each including anchor items. Only a single form was used for LN and LS and all items were anchor items from the K-2 item pool. A total of 231 items in Forms A-D were submitted to the PreK item pool for the BL, DEL, FD, SS, and WM tasks. All items in LN (26) and LS (23) were submitted to the PreK item pool.

A total of 385 first grade students (189 female, 194 male, 2 unknown) in 13 schools across 6 states spanning the major geographic regions of the United States (CA, FL, GA, LA, MA, NH) participated in the 2022 EarlyBird study. A total of 106 participants (28% of total sample) had a family history of reading difficulties. Of those, 47 participants (12% of the total sample) reported that a family member had been diagnosed with dyslexia. Socioeconomic status data was applied based on the school sites. Approximately 58% of the children in the sample attended schools with Title I designation. The sample was made up of approximately 7% Asian students, 13% Black or African American students, 3% Hispanic or Latinx students, 63% White students, and 8% multiracial students. (6% of participants chose not to disclose this information.)

The study was conducted in two phases. The first began in early February 2022 and ended in late March 2022. Each Grade 1 participant was administered nine subtests. Of these, six (Blending, Deletion, Word Reading, Following Directions, Word Matching, and Oral Sentence Comprehension) were developed at Florida State University's (FSU) Florida Center for Reading Research. One (Nonword Repetition) was based on a subtest developed at Boston Children's Hospital (BCH) for Kindergarten students but was expanded with new items and content appropriate for Grade 1 students. Two subtests (Nonword Reading and Letter RAN) were developed in conjunction with content area experts.

The second (and final) phase of the study began in early April and ended in early June 2022. 367 students participated in this phase of the study. In addition to taking the app again, each participant was administered a battery of standardized, paper-and-pencil assessments to aid in the development of algorithms. The paper-and-pencil battery included the following: Reading Comprehension subtest of the KTEA-3, Word Identification and Word Attack subtests of the WRMT-III, and Nonword Repetition and Rapid Letter Naming subtests of the CTOPP-2. With the data collected from the two study phases, a series of analyses were performed to examine the item-level properties and utilize the best items to create and evaluate a series of algorithms for predicting risk of dyslexia and word reading success at the end of the year. The best algorithms were then used to optimize the prediction.

2.4.2 Analyses

With the data collected from the two study phases, a series of analyses were performed to examine the item-level properties and utilize the best items to create and evaluate a series of algorithms for predicting risk of dyslexia and word reading success at the end of the year. The best algorithms were then used to optimize the prediction.

A common-item, non-equivalent groups design was used for collecting data in our study. A strength of this approach is that it allows for linking multiple test forms via common items.

Item difficulty and discrimination were estimated for fixed form items by outputting individual item responses (0 = incorrect; 1 = correct) for each item for each test and subjecting the responses to IRT modeling. 2PL models were fit for the fixed form tests. Additionally, vertical scaling was performed for the Nonword Repetition subtest by fixing item slope and intercept parameters to link with Kindergarten item performance.

All IRT models were estimated with flexMirt (Houts & Cai, 2015). Model quality was evaluated using local fit (i.e., performance of the individual items) and goodness-of-fit indices based on the M2 statistic (Maydeu-Olivares, 2013), and the root mean square error of approximation based on M2 (RMSEA2). M2 is often sensitive to sample size in terms of rejecting the fitted model, thus, the RMSEA2 is useful for determining adequate fit (<.089), close fit (<.05), or excellent fit [.05/(k -1), where k = number of categories]. Theta and standard error values were output for use in correlational analyses and validation.

2.4.3 Results

Item Level Results

For the First Sounds Matching test, item calibration was performed for 23 items with four of the 23 items set to be equivalent to linked K items for vertical scaling. The model fit indices were M2 = 183.30, p = .99, RMSEA = .00, TLI = 1.05. The discrimination parameters (a) had an average of 1.79 and ranged from .61 to 3.8. The difficulty parameters (b) had an average of -.56 and ranged from -2.09 to .39.

For the Rhyming test, item calibration was performed for 24 items with four of the 24 items set to be equivalent to linked K items for vertical scaling. Model fit indices were M2 = 275.22, p = .23, RMSEA = .02, TLI = .99. The discrimination parameters (a) had an average of 1.96 and ranged from .91 to 3.63. The difficulty parameters (b) had an average of -.62 and ranged from -1.33 to .56.

For the Vocabulary test, item calibration was performed for 24 items with four of the 24 items set to be equivalent to linked K items for vertical scaling. Model fit indices were M2 = 161.93, p = 1.00, RMSEA = .00, TLI = 1.03. The discrimination parameters (a) had an average of 1.81 and ranged from .45 to 6.08. The difficulty parameters (b) had an average of -2.09 and ranged from -3.88 to .30.

Reliability

Marginal reliability is conceptualized as the level of precision for the measurement across the trait continuum. Values of .80 are typically viewed as acceptable for research purposes while estimates at .90 or greater are acceptable for clinical decision making (Nunnally & Bernstein, 1994). First Sound Matching (rxx = .78, 95% CI = [.76, .79]) and Rhyming (rxx = .79, 95% CI = [.77, .80]) approached an acceptable level of reliability while Vocabulary (rxx = .57, 95% CI = [.52, .62]) did not. Marginal reliability for the adaptive subtests, Oral Sentence Comprehension, (rxx = .79, 95% CI = [.76, .82]) and Letter Sound, (rxx = .82, 95% CI = [.78, .85]), indicated acceptable levels of reliability while Blending, (rxx = .78, 95% CI = [.74, .83]) approached an acceptable level, and Letter Name, (rxx = .65, 95% CI = [.58, .71]) did not.

Correlations and Predictive Validity

The validity of scores of the selected criterions were addressed through examining the relations between and among EarlyBird ability scores and standardized outcomes. First, bivariate relations were computed to evaluate the strength of relations between and among each of the EarlyBird assessments and norm-referenced tests. Second, logistic regressions were composed to identify the fewest number of tasks that maximized the ability to explain changes in the CTOPP-2 Phonological Awareness Composite Score by the EarlyBird assessments. EarlyBird assessments were retained if they were statistically significant ($p < .05$) in a GLM with CTOPP-2 Phonological Awareness Composite Score as the outcome of interest. Finally, a multiple regression was run to estimate the total amount of variance that the linear combination of selected predictors explained in selected criterions. The selected predictors were based on the EarlyBird assessments that emerged as the best for predicting Phonological Awareness from the logistic regressions. Model adequacy and EarlyBird assessment contributions were determined by R-squared (i.e. variance accounted by the model), as well as multiple R (square root of R-squared) and the 95 percent confidence interval for multiple R.

Correlations between and among EarlyBird ability scores with standardized outcomes ranged from .08 between Deletion and Sound Matching to .87 between Phonological Awareness and Blending Words (Table 2.42). Among the EarlyBird ability scores, the correlations ranged from -.40 between RAN and RON (Rapid Object Naming) to .71 between Follow Direction subtest and CELF-3 Following Directions.

Table 2.42: Descriptive Statistics and Correlation Table for EarlyBird Subtests and Standardized Assessments of Language and Literacy (N = 150)

Subtest	M	SD	BL	DEL	FD	LN	LS	SS	VP	FSM	NWR	RHYM	WM	RAN	K.LWR	K.NWD	K.P	K.D	WID	WA	SWE	PDE
BL (N=210)	-2.23	2.38																				
DEL (N=101)	-1.14	1.41	0.44																			
FD (N=97)	-1.09	1.25	0.17	0.17																		
LN (N=94)	3.07	2.87	0.14	0.26*	0.27*																	
LS (N=213)	-0.08	1.60	0.39*	0.31*	0.18	0.23*																
SS (N=97)	-0.32	0.94	0.17	0.17	0.26*	0.18	0.25*															
VP (N=213)	0.33	1.64	0.20*	0.23*	0.09	0.22*	0.22*	0.34*														
FSM (N=191)	0.18	0.99	0.56*	0.37*	0.15	0.26*	0.51*	0.16	0.25*													
NWR (N=200)	0.12	0.85	0.48*	0.44*	0.16	0.27*	0.37*	0.16	0.24*	0.51*												
RHYM (N=195)	0.14	0.90	0.26*	0.23*	0.34*	0.47*	0.25*	0.2	0.38*	0.39*	0.43*											
WM (N=204)	0.09	0.89	0.30*	0.32*	0.34*	0.40*	0.30*	0.48*	0.34*	0.46*	0.51*	0.39*										
RAN (N=175)	84.27	27.86	-0.30*	-0.09	-0.09	-0.24*	-0.28*	-0.16	-0.22*	-0.41*	-0.23*	-0.23*	-0.25*									
K-LWR (N=139)	96.78	13.50	0.27*	0.14	0.33*	0.19	0.43*	0.28*	0.13	0.45*	0.25*	0.34*	0.22*	-0.26*								
K-NWD (N=111)	92.45	13.01	0.23*	0.12	0.21	0.28*	0.31*	0.26	0.15	0.46*	0.27*	0.42*	0.24*	-0.33*	0.86*							
K-P (N=215)	100.02	14.85	0.39*	0.46*	0.33*	0.30*	0.35*	0.17	0.27	0.51*	0.46*	0.53*	0.36*	-0.20*	0.63*	0.72*						
K-D (N=103)	93.17	14.74	0.31*	0.15	0.39*	0.23	0.31*	0.24	0.12	0.50*	0.24*	0.43*	0.23*	-0.40*	0.84*	0.81*	0.84*					
WID (N=192)	97.16	16.29	0.22*	0.05	0.34*	0.24*	0.34*	0.17	0.18*	0.48*	0.23*	0.41*	0.24*	-0.26*	0.91*	0.77*	0.62*	0.84*				
WA (N=192)	99.31	14.45	0.30*	0.03	0.29*	0.28*	0.27*	0.15	0.23*	0.49*	0.34*	0.47*	0.32*	-0.24*	0.79*	0.75*	0.71*	0.80*	0.81*			
SWE (N=108)	94.03	16.44	0.31*	0.23	0.44*	0.19	0.40*	0.2	0.22*	0.52*	0.30*	0.44*	0.26*	-0.27*	0.92*	0.79*	0.71*	0.88*	0.91*	0.77*		
PDE (N=108)	94.95	13.62	0.27*	0.18	0.39*	0.14	0.34*	0.05	0.18	0.54*	0.30*	0.44*	0.30*	-0.34*	0.82*	0.82*	0.71*	0.84*	0.77*	0.73*	0.85*	
CELF (N=219)	103.26	12.17	0.12	0.08	0.43*	0.37	0.25	0.38	0.28	0.27	0.44*	0.43*	0.45*	-0.02	0.21	0.34*	0.46*	0.31	0.21	0.31	0.29	0.36

M = mean, SD= standard deviation, BL = Blending, DEL = deletion, FD = following directions, LN = letter name knowledge, LS = letter sound knowledge, SS = sentence structure, VP = vocabulary pairs, FSM = first sound matching, NWR = nonword repetition, RHYM = rhyming, WM = Word Matching, K-LWR = KTEA letter word recognition, K-NWD = KTEA ?, K-P = KTEA phonological ?, K-D = KTEA Dyslexia, WID = WRMT Word Identification, WA = WRMT Word Attack, SWE = TOWRE Sight Word Efficiency, PDE = TOWRE Phoneme Deletion Efficiency, CELF = CELF Sentence Structure. *p < .05.

Logistic regression models were composed for the Reading Readiness Algorithm to identify the best model for predicting Phonological Awareness (Table 2.43). For the Reading Readiness model, First Sounds and Vocabulary were statistically significantly related to Phonological Awareness.

Table 2.43: Logistic Regression Results for Reading Readiness Model

Term	Estimate	SE	z	p
Intercept	0.28	0.19	1.46	.144
First Sounds	1.08	0.19	5.57	<.001
Vocabulary	0.23	0.10	2.42	.02

Multiple regression models were composed including the selected EarlyBird assessments that emerged from the logistic regression (Table 2.44). For Reading Readiness, the model for predicting Phonological Awareness included First Sounds and Vocabulary and resulted in an R2 of .40 (multiple r = .63, 95% CI = .54, .69, n = 199).

Table 2.44: OLS Regression Results for Reading Readiness Model

Term	Estimate	SE	t	P
Intercept	0.53	0.03	16.65	<.001
First Sounds	0.18	0.02	7.16	<.001
Vocabulary	0.04	0.02	2.70	.01

Classification accuracy

Using the predictive validity model including the best model selected via the logistic regressions, AUC was estimated for the overall efficiency of discrimination of the log-odds from the multiple predictors for the Reading Readiness model (AUC = .83, 95% CI = .76, .91) with an optimal cut-point identified at .57 (Table 2.45). A total of 92 students (46.2%) were classified as true positives, 62 students (31.2%) were classified as true negatives, 18 students (9.0%) were classified as false positives, and 27 students (13.5%) were classified as false negatives. Mean bootstrapped sensitivity (.77, 95% CI = .63, .90) and specificity (.78, 95% CI = .60, .91) were estimated along with matrix-based negative predictive power (.70), positive predictive power (.84), and overall correct classification (.77) were computed. The base rate in the sample was 59.8%.

Table 2.45: Classification Accuracy for Reading Readiness Model Predicting Phonological Awareness

tp	fp	tn	fn	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Cutpoint
92	18	62	27	0.77	0.83	0.77	0.78	0.84	0.7	0.56

tp = True positive; fp = False positive; tn = True negative; fn = False negative; AUC = Area under the curve; PPV = positive predictive value; NPV = Negative predictive value.

DIF

For the new Pre-K First Sound Matching, Vocab, and Rhyming subtest items, DIF testing was conducted comparing: Black-White students and comparing: Male-Female students. Values for DIF were computed via the difR package (Magis et al., 2010) using the Mantel-Haenszel method (Mantel & Haenszel, 1959) for detecting uniform DIF. For each of the FS, Vocab, and Rhyming tasks, DIF was tested for two primary contrasts: 1) Male vs. female, and 2) White vs. Non-White. The Mantel-Haenszel chi-square statistic was reported for test by item and the chi-square was used to derive an effect size estimate (i.e., ETS delta

scale; Holland & Thayer (1986)). Effect size values <= 1.0 are considered small, 1.0 – 1.5 is moderate, and >= 1.5 is considered large.

Across First Sound Matching, Vocabulary, and Rhyming, eight items demonstrated DIF with at least a moderate effect size (i.e., deltaMH >= 1.0) on either race, gender, or both. A total of seven items (four vocabulary, two First Sound Matching and one Rhyming) differed on race while four items (two Rhyming, one First Sound Matching, and one Vocabulary) differed on gender. Three of the eight items differed on both race and gender.

Chapter 3

Score Reporting

Chapter 8 of the *Standards* addresses test takers rights to timely and comprehensible score reports, and many other sections of the *Standards* address the information stakeholders should be provided about metrics on score reports. Standard 6.10 provides a succinct summary: "When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 119).

Multiple scores are provided in order to facilitate a diverse set of educational decisions. In this section, we describe the types of scores provided for each measure, define each score, and indicate its primary utility within the decision making framework. For most subtests, the Imagine+ EarlyBird Screener provides scores instantaneously to educators after a student completes the test; there is no wait time between when the student finishes the assessment and when the report is available. Score reports are provided in different views targeted to each stakeholder: views of a single student, class, school, or district. In all views, similar information is provided, and the view varies the level of aggregation of the data. Educators may prefer certain views over others depending on their role and the purpose in using the reports. The key metrics reported are: Reading Readiness, Potential for Word Reading (PWR), Subtest Score Percentiles, and Ratios, all of which are described further in the following sections.

3.1 Reading Readiness (Grade Pre-K)

The Reading Readiness indicator shows the likelihood that an end-of-year Pre-K student will have the foundational skills to learn how to read at the beginning of Kindergarten. Students are classified into one of two categories:

- **Reading Readiness** indicates the child demonstrates phonological awareness and oral language skills that are foundational in learning to read.
- **Emerging Readiness** indicates the child is still developing early skills of phonological awareness and oral language.

The Reading Readiness indicator was developed through an analysis to determine which subtests are most predictive of achieving targeted skills at the end of the year that are most predictive of being ready to learn to read in at the beginning of Kindergarten. For the purposes of this analysis, Reading Readiness is defined as performing above the 40th percentile on the CTOPP-2 Phonological Awareness Composite Score. The Reading Readiness score is a multi-factorial calculation that involves a selection of the most predictive subtests and an aggregation and weight averaging of that data according to the degree of predictability to generate a single output score. The screening tasks include First Sound Matching and Vocabulary.

3.2 Potential For Word Reading (Grade K, 1, 2, and 3)

The Potential For Word Reading (PWR) is a score stemming from the screener component of the Imagine+ EarlyBird Screener, and is the probability, expressed as a percentage, that a student will reach a pre-determined threshold of word reading achievement, by the end of the current academic year. At each time point of assessment (i.e., Fall, Winter, Spring) analyses at each grade level were conducted to identify a core set most predictive of achieving targeted grade-level performance at the end of the year. For the purposes of this analysis, grade-level performance was defined as scoring greater than or equal to the 40th percentile on the KTEA-3 Letter and Word Recognition subtest. At certain time-points by grade level, The PWR score is a multi-factorial calculation that involves more than one screening task. The PWR can be obtained three times per year: Fall, Winter and Spring.

The PWR also including a classification into Low, Moderate, and High PWR. Low PWR is determined by a separate Dyslexia Risk Indicator (DRI) algorithm. The DRI is a score stemming from the screener component of the Imagine+ EarlyBird Screener, and is the probability, expressed as a percentage, that a student will not perform above a pre-determined threshold of word reading achievement by the end of the current academic year. At each time point of assessment (i.e., Fall, Winter, Spring) analyses at each grade level were conducted to identify a core set most predictive of achieving targeted grade-level performance at the end of the year. For the purposes of this analysis, grade-level performance was defined as less than or equal to the 16th percentile on the KTEA-3 Letter and Word Recognition subtest. At certain time-points by grade level, the DRI score is a multi-factorial calculation that involves more than one screening task. Although the DRI numerical score is not reported to educators or students, it is used to inform classification of students into Low PWR when DRI is flagged. The PWR algorithm also classifies students into High PWR when scoring above a predetermined probability. All other students who are not classified as Low PWR by the DRI algorithm or as High PWR by the PWR algorithm are classified as Moderate PWR.

For beginning of the year Kindergarten (Fall), Deletion and Blending subtests are the inputs for both the PWR score and corresponding DRI classification into Low or High PWR. For middle of year Kindergarten (Spring), the Word Reading and Blending subtests are the inputs, also for both the PWR score and corresponding DRI classification into Low or High PWR. For all other grades and end of year Kindergarten (Spring), only the Word Reading subtest is used as input, again for both the PWR score and related Low or High PWR classification from the DRI.

3.3 Subtest Score Percentiles (Grade Pre-K, K, 1, 2, and 3)

Students' performance on each subtest is displayed in the form of normed percentiles. Normed percentiles are created based on distributions of raw scores of students from a nationally representative sample. The samples include students from all major geographic regions of the United States, attending a mix of public, private, and charter schools, with and without a familial history of diagnosed or suspected dyslexia, and from a range of socioeconomic backgrounds (as determined by the percentage of students receiving free or reduced price lunch at the participating schools). In terms of race and ethnicity, the samples closely match U.S. census data. They are periodically updated to reflect the most recent representative samples available.

Percentile ranks can vary from 1 to 99, and the distribution of scores were created from a large standardization sample and divided into 100 groups that contain approximately the same number of observations in each group. For example, a first grade student who scored at the 60th percentile performed as well as or better than about 60% of the students in the standardization sample. The percentile rank is an ordinal variable, meaning that it cannot be added, subtracted, used to create a mean score, or in any other way mathematically manipulated. The median is always used to describe the midpoint of a distribution of percentile ranks. Because this score compares a student's performance to other students within a grade level, it is meaningful in determining the skill strengths and skill weaknesses for a student as compared to other students' performance.

3.4 Standard Scores

Standard scores for Imagine+ EarlyBird Screener subtests are grade-level norms and can vary roughly between 55 and 145, although scores may sometimes fall below or above those limits. About 99% of all scores will fall between 55 and 145, and the average standard score is 100. A standard score provides information about how far above or below the mean of the standardization sample a given student's score falls. It is expressed in standard deviation units, so for measures like those in the Imagine+ EarlyBird Screener system that have been normed to have a mean of 100 and a standard deviation of 15, a score of 115 is one standard deviation above the mean, and a score of 85 is one standard deviation below the mean. Standard scores on different tests have essentially the same meaning in terms of how strong or weak (compared to the norming sample) a given score is, so they may be used to compare a student's performance across different tests. Standard scores can also be averaged and used in a variety of mathematical calculations.

3.5 Developmental Scale Scores

A developmental ability score is reported for the computer adaptive measures in the Imagine+ EarlyBird Screener to facilitate tracking growth in subtests over time within grade level and across grade levels as each task allows. Ability scores on item response theory based tests typically vary from -3 to +3, but this scale is typically transformed to make it easier to interpret as a measure of development. In the case of the tests in the Imagine+ EarlyBird Screener, the scores were transformed to have a mean of

500 and standard deviation of 100, with 100 representing the extreme low end of the scale, and 1000 representing the high end of the scale. Because the relative difficulty of items on many of the subtests have been estimated through linking studies for students from Kindergarten through Grade 3, the developmental ability score covers a range of ability from approximately Kindergarten at the low end to approximately Grade 3. It is important to note that the developmental span of subtests dictates the breadth of growth interpretation over time. That is, certain tasks are only administered in Kindergarten while others are eligible to be administered through Grade 3. A student who is at the high end of a Kindergarten-only task may look to have the same ability based on a numerical value (e.g., 1000) compared to a student who is at the high end of a task that spans Kindergarten through Grade 3; thus, it is important to use developmental scores to monitor growth within each task.

3.6 Ratios (Grade Pre-K, K, 1)

The Letter Sound (Grade Pre-K, K, 1) and Letter Name (Grade Pre-K & K) subtests also yield a ratio reflecting the total number of items the student answered correctly out of the full inventory of items given at that time period. For example, if a student could name 20 letters out of the total letter sound inventory of 26, the ratio on the data dashboard would show 20/26.

References

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

Catts, H. W., & Petscher, Y. (2020). *A cumulative risk and resilience model of dyslexia*. EdArXiv. https://doi.org/10.35542/osf.io/g57ph

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*, 1–29.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

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

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the national reading panel's meta-analysis. *Review of Educational Research*, *71*(3), 393–447.

Foorman, B. R., Francis, D. J., Shaywitz, S. E., Shaywitz, B. A., & Fletcher, J. M. (2013). The case for early reading intervention. In *Foundations of reading acquisition and dyslexia* (pp. 243–264). Routledge.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology*, *37*(2), 261–278.

Gillam, R. B., Montgomery, J. W., Evans, J. L., & Gillam, S. L. (2019). Cognitive predictors of sentence comprehension in children with and without developmental language disorder: Implications for assessment and treatment. *International Journal of Speech-Language Pathology*, *21*(3), 240–251.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the mantel-haenszel procedure. *ETS Research Report Series*, *1986*(2), i–24.

Houts, C. R., & Cai, L. (2015). *flexMIRT: Flexible multilevel multidimensional item analysis and test scoring user's manual version 3.6 RC*. https://vpgcentral.com/software/flexmirt.

Jenkins, J. R. (2003). Candidate measures for screening at-risk students. *Conference on Response to Intervention as Learning Disabilities Identification, Sponsored by the National Research Center on Learning Disabilities, Kansas City, MO*.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862.

Mallett, C., Stoddard-Dare, P., & Workman-Crenshaw, L. (2011). Special education disabilities and juvenile delinquency: A unique challenge for school social work. *School Social Work Journal*, *36*(1), 26–40.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Cataldi, E. F., Mann, F. B., & Barmer, A. (2019). The condition of education 2019. NCES 2019-144. *National Center for Education Statistics*.
Meade, A. W. (2010). A taxonomy of effect sizes for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*, 728–743.
Muthén, B., & Muthén, L. (2008). *Mplus User's Guide*. Muthén & Muthén.
National Institute of Child Health and Human Development. (2000). *Report of the national reading panel: Teaching children to read: Reports of the subgroups* (NIH Publication No. 00-4754). U.S. Department of Health; Human Services.
National Research Council. (1998). *Preventing reading difficulties in young children* (C. E. Snow, M. S. Burns, & P. Griffin, Eds.). National Academy Press.
Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
Ozernov-Palchik, O., & Gaab, N. (2016). Tackling the "dyslexia paradox": Reading brain and behavior for early markers of developmental dyslexia. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(2), 156–176. https://doi.org/10.1002/wcs.1383
Ozernov-Palchik, O., Yu, X., Wang, Y., & Gaab, N. (2016). Lessons to be learned: How a comprehensive framework of atypical reading development can inform educational practice. *Current Opinion in Behavioral Sciences*, *10*, 45–58.
Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*, 22–37. https://doi.org/10.1080/10888438.2013.827687
RAND Reading Study Group. (2002). *Reading for understanding*. RAND Corporation.
Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*(2), 31–74.
Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. *International Journal of Language & Communication Disorders*, *45*(1), 47–60.
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). Guilford Press.
Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.
Torgesen, J. K. (2004). Avoiding the devastating downward spiral: The evidence that early intervention prevents reading failure. *American Educator*, *28*(3), 6–19.
Valas, H. (1999). Students with learning disabilities and low-achieving students: Peer acceptance, loneliness, self-esteem, and depression. *Social Psychology of Education*, *3*(3), 173–192.
Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*, 1–67.
Wanzek, J., & Vaughn, S. (2007). Based implications from extensive early reading interventions. *School Psychology Review*, *36*(4), 541–561.
Weiser, B. L., & Mathes, P. G. (2011). Using encoding instruction to improve the reading and spelling performances of elementary students at-risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research*, *81*, 170–200.
Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What is the pronunciation for-ough and the spelling for/u/? A database for computing feedforward and feedback consistency in english. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 600–618.