EarlyBird

EarlyBird Grade 1 & 2 Dyslexia and Early Literacy Assessment

Technical Manual

October 2024

Table of Contents

Please note that this technical manual provides information about the Grade 1 and Grade 2 Assessments. For data specific to the EarlyBird assessments designed for PreK and Kindergarten, please see Chapters 5 and 6 in those respective technical manuals.

Acknowledgements

This Technical Manual for the EarlyBird Grade 1 Assessment was written and based on the research of:

Nadine Gaab, PhD.*, Associate Professor of Education at the Harvard Graduate School of Education, formerly of Boston Children's Hospital/Harvard Medical School.*

Yaacov Petscher, PhD.*, Professor of Social Work at Florida State University, Associate Director of the Florida Center for Reading Research, Director the Quantitative Methodology and Innovation Division at FCRR.*

Boston Children's Hospital Early Literacy Screener

The research for the development of the Boston Children's Hospital Early Literacy Screener was funded through generous grants provided by the following family foundations:

- *The Heckscher Foundation for Children*
- *The Oak Foundation*
- *The Poses Family Foundation*
- *The Peter and Elizabeth C. Tower Foundation*
- *The Emily Hall Tremaine Foundation*
- *And extensive in-kind donations from Boston Children's Hospital.*

The authors would like to thank these funders for their support of this project across a multi-year research study, as well as the many teachers, school and district administrators, and children who participated in this research. Additionally, many experts and leaders in the fields of literacy, education, school administration, educational policy, technology, developmental medicine and neuroscience have served as advisors to this project, helping to ensure the development of a screener that is both scientifically sound, and tailored to the needs of today's educators.

Florida Center for Reading Research Reading Assessment

The items, dynamic flow, computer-adaptive algorithms, creation of the development application, and psychometric work for this component skills battery (called the Florida Center for Reading Research Reading Assessment; FRA) were funded by grants from the Institute of Education Sciences (IES) to Florida State University [Barbara Foorman, Ph.D. (PI), Yaacov Petscher, Ph.D., Chris Schatschneider, Ph.D.) :

Institute of Education Sciences, USDOE ($4,447,990), entitled "Assessing Reading for Understanding: A Theory-Based, Developmental Approach," subcontract to the Educational Testing Service for five years (R305F100005), 7/1/10-6/30/15 (Foorman, PI on subcontract).

Institute of Education Sciences, USDOE (R305A100301; $1,499,741), entitled "Measuring Reading Progress in Struggling Adolescents," awarded for four years, 3/1/10-2/28/14. (Foorman, PI; Petscher and Schatschneider, Co-Is).

We would like to acknowledge the following individuals for their leadership in executing the work funded by the above two IES grants: Dr. Adrea Truckenmiller, Karl Hook, and Nathan Day. We also would like to thank the numerous school districts, administrators, and teachers who participated in the research funded by these two grants.

Reach Every Reader's Interstellar Express Assessment System

The items, dynamic flow, computer-adaptive task flow and placement algorithms, and psychometric work for the Reach Every Reader tasks (called the Interstellar Express Assessment System, created by Dr. Yaacov Petscher, Dr. Hugh Catts, and Cody Diefenthaler, were funded by a grant from the Chan Zuckerberg Initiative. We would like to acknowledge the many project coordinators, schools, teachers, families, children, artists, software engineers, and staff who participated in the many facets of research that contributed to the product development.

Chapter 1: Introduction

The development of basic reading skills is one major goal during the first years of elementary school. However, in the United States, 65% of 4th 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 3rd grade remain poor readers throughout their educational career (Foorman, Francis, Shaywitz, Shaywitz, & Fletcher, 1997). 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, Stoddard-Dare, & Workman-Crenshaw, 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 & Gaab, 2016a,b). 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%–92% 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.

We aimed to develop an assessment for the identification of children at risk for atypical reading and language skills in PreK through Grade 2. This technical manual describes the research, validation studies, and development of our Grade 1 & 2 assessments (technical manuals for the PreK and K assessments are also available). We are fortunate to have several consensus documents that review decades of literature about what predicts reading success (National Research Council, 1998; National Institute of Child Health and Human Development, 2000; Rand, 2002; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001).

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, Stone, & Jacobs, 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, Nunes, Stahl, & Willows, 2001; Foorman & Connor, 2011).

In the EarlyBird Grade 1 & 2 Assessment, the alphabetic principle is assessed with individually-administered tasks that measure letter-sound knowledge, word and nonword reading, and spelling.

Comprehending Written Language

Knowledge of word meanings

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*, *sooth*, *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 EarlyBird Grade 1 & 2 Assessment, 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).

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 EarlyBird Grade 1 & 2 Assessment, 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).

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, 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 EarlyBird Grade 1 & 2 Assessment, there are Reading Comprehension passage available to administer as a supplemental subtest. 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, Francis, Davidson, Harm, & Griffin, 2004; Foorman, 2009; Fitzgerald et al., 2014).

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 multifactorial strategies for screening and interventions (Catts & Petscher, 2020; Ozernov-Palchik et al., 2016a,b). Numerous longitudinal research studies (for an overview see Ozernov-Palchik et al., 2016a) 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 EarlyBird tool incorporates all of these skills, as outlined below.

Description of EarlyBird App

The EarlyBird gamified mobile app is easy, quick, accessible, and child-centered. It is self-administered in small groups with teacher oversight and, depending on the subtests administered, the Grade 1 and Grade 2 assessments each take 20-30 minutes per child. The assessments address literacy milestones that have been found to be predictive of subsequent reading success in each grade. No trained adult administration is needed to administer the EarlyBird app. Scoring is largely automated. EarlyBird includes screening at BOY, MOY, and EOY for low and moderate expected word reading performance at the end of the school year. The assessment incorporates subtests that were validated across four different validation studies. Appendices A-D present information pertaining to each of these studies separately, though the assessment is streamlined in the EarlyBird administration process.

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 assessment 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.

Subtests can be administered at the beginning of the school year (in fall), middle of the year (in winter), and end of the year (in spring). With the exception of RAN (which is normed based on one time of year only), all subtests have time of year-specific norms. To enable the most appropriate use of the assessment, recommendations will provide guidance on which subtests should be administered given the time of year and/or which subtests provide the appropriate follow-on should a child demonstrate weakness in select subtests.

Chapter 2: Subtest Information

Description of Subtests

Phonemic Awareness

Deletion (Grade 1) – Elephant: 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.

Nonword Repetition (Grade 1 & 2) - Ostrich: 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.

Phonics (including Alphabet Knowledge)

Letter Sound (Grade 1) - Giraffe: 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 3 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 (Grade 1 & 2) - Walrus: 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 VC and CVC to VCe words.

Nonword Spelling (Grade 1 & 2) - Penguin: 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 1 to 2 syllables, including CVC and VCe patterns, and include digraphs and blends.

Fluency

Rapid Automatized Naming (Grade 1 & 2) - Polar Bear: In Grade 1, the Letter RAN board has 5 uppercase letters *(R, M, K, E, S)* repeated in random order over 5 rows of 10 items each. The Grade 2 Letter RAN board has 6 lowercase letters (h, t, m, s, o, i) repeated in random order over 5 rows of 10 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. The student's response is recorded to the dashboard and available to the teacher for later confirmation of time and accuracy.

Word Reading (Grade 1 & 2) - Lion: 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 (Grade 1 & 2) - Supplemental Subtest: 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.

Vocabulary

Expressive Vocabulary (Grade 1 & 2) - Bison: 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 (Grade 1 & 2) - Gorilla: 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?").

Comprehension

Follow Directions (Grade 1 & 2) - Zebra: 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 (Grade 1 & 2) - Supplemental Subtest: 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 multiple choice questions, 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.

Chapter 3: Score Definitions

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.

Potential For Word Reading (PWR)

The Potential for Word Reading score is the probability, expressed as a percentage, that reflects the likelihood at each testing period that a student will reach grade-level expectations in word reading by the end of the year, presuming the student does not receive appropriate, evidence-based remediation. It is determined by an algorithm that was created through a multifactorial analysis of all available subtests for each grade and time period, and driven by those subtests statistically determined to be most predictive of end-of-year reading outcomes in each grade. Predicted PWR performance is displayed in the form of a percent and categorized into one of three groups: high performance, moderate performance, and low performance. High performance, for the purposes of this analysis, is defined as likely scoring above the 40th percentile on the KTEA-3 Letter and Word Recognition subtest at end-of-year. Low performance, for the purposes of this analysis, is defined as likely scoring at or below the 16th percentile on the KTEA-3 Letter and Word Recognition subtest at end-of-year. Moderate performance includes all students who do not meet either of the above criteria for the high or low performance categories.

Subtest Score Percentiles

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.

Ratios

In addition to the subtest score percentile, the Letter Sound subtest also yields 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 29, the ratio on the data dashboard would show 20/29.

Chapter 4: Psychometric Approaches

Item Response Theory (IRT)

Scores from the EarlyBird Assessments were analyzed through a combination of measurement frameworks and techniques. Traditional testing and analysis of items involves estimating the difficulty of the item (based on the percentage of respondents correctly answering the item) as well as discrimination (how well individual items relate to overall test performance). This falls into the realm of measurement known as classical test theory (CTT). While such practices are commonplace in assessment development, IRT holds several advantages over CTT. When using CTT, the difficulty of an item depends on the group of individuals on which the data were collected. This means that if a sample has more students that perform at an above-average level, the easier the items will appear; but if the sample has more below-average performers, the items will appear to be more difficult. Similarly, the more that students differ in their ability, the more likely the discrimination of the items will be high; the more that the students are similar in their ability, the lower the discrimination will be. One could correctly infer that scores from a CTT approach are entirely dependent on the makeup of the sample.

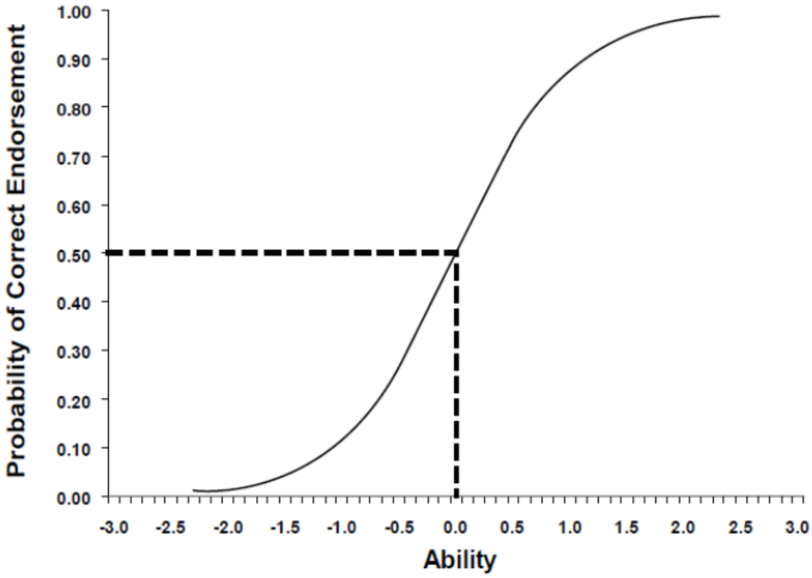
The benefits of IRT are such that 1) the difficulty and discrimination are not dependent on the group(s) from which they were initially estimated, 2) scores describing students' ability are not related to the difficulty of the test, 3) shorter tests can be created that are more reliable than a longer test, and 4) item statistics and the ability of students are reported on the same scale.

Item difficulty. The difficulty of an item (*b*) 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 -3.0 to +3.0.

Figure 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.

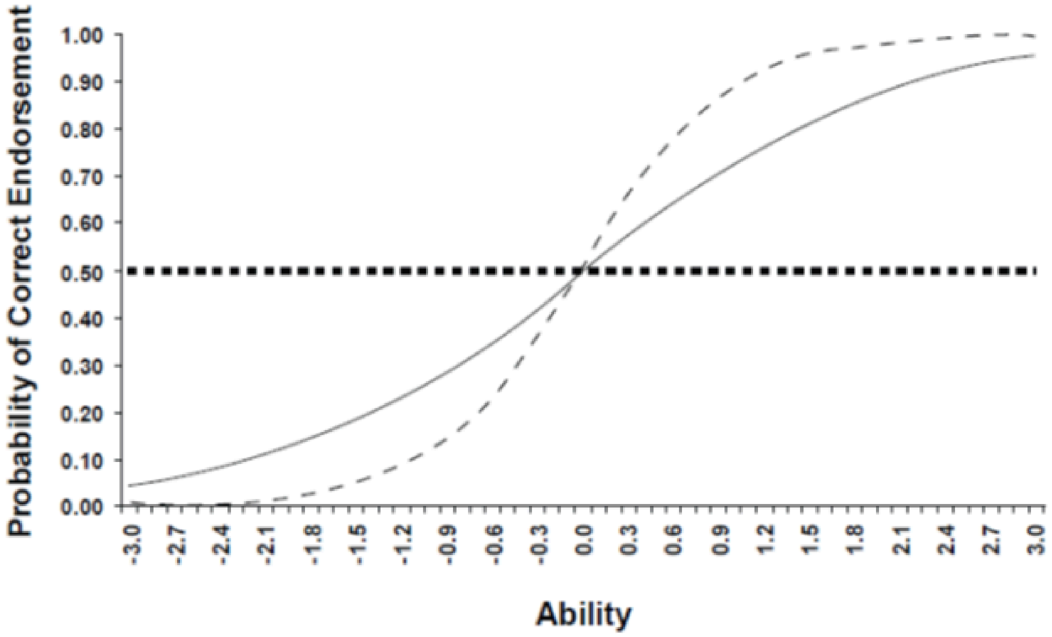
Along the y-axis is the probability of a correct response, which varies across the level of difficulty. Item difficulty is defined as the value on the x-axis at which the probability of correctly endorsing the item is 0.50. As demonstrated for the sample item in Figure 1, the difficulty of this item 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: Sample Item Characteristic Curve



Item Discrimination. 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 2 provides an example of how discrimination operates in the IRT framework. For all three items presented in Figure 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 2: Sample Item Characteristic Curves with Varied Discriminations



Computer Adaptive Testing (CAT)

The majority of EarlyBird tasks are based on computer adaptive algorithms that leverage an 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. The CAT in EarlyBird tasks are administered in the following ways: 1) the student is administered a set of 5 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 (SE) of ability; 3) the student SE is compared to a target SE threshold (associated with reliability = .80) where student SE < target SE results in the task terminating and moving to the next task; 4) when student SE > target SE the student is administered another item according to | θ – *b* |. Steps 2-4 continue until the target SE is reached or until a predetermined number of items have been administered.

Guidelines for Retaining Items

Several criteria were used to evaluate item performance. The first process was to identify items which demonstrated strong floor or ceiling effects in response rates >= 95%. Such items are not useful in creating an item bank as there is little variability in whether students are successful on the item. In addition to evaluating the descriptive response rate, we estimated item-total correlations. Items with negative values are indicative of poor functioning such that it suggests individuals who correctly answer the question tend to have lower total scores. Similarly, items with low item-total correlations indicate the lack of a relation between item and total test performance. Items with correlations <.15 were flagged for removal.

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 consisted 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.

Marginal Reliability

Reliability describes how consistent test scores will be across multiple administrations over time, as well as how well one form of the test relates to another. Because the EarlyBird assessment uses Item Response Theory (IRT) as its method of validation, reliability takes on a different meaning than from a Classical Test Theory (CTT) perspective. The biggest difference between the two approaches is the assumption made about the measurement error related to the test scores. CTT treats the error variance as being the same for all scores, whereas the IRT view is that the level of error is dependent on the ability of the individual. As such, reliability in IRT becomes more about the level of precision of measurement across ability, and it may sometimes be difficult to summarize the precision of scores in IRT with a single number. Although it is often more useful to graphically represent the standard error across ability levels to gauge for what range of abilities the test is more or less informative, it is possible to estimate a generic estimate of reliability known as marginal reliability (Sireci, Thissen, & Wainer, 1991) with:

$$\bar{\rho} = \frac{\sigma_{\theta}^{2} - \overline{\sigma_{e*}^{2}}}{\sigma_{\theta}^{2}}$$

where σ_{θ}^{2} is the variance of ability score for the normative sample and $\overline{\sigma_{e*}^{2}}$ is the mean-squared error.

Construct Validity

Construct validity describes how well scores from an assessment measure the construct it is intended to measure. Components of construct validity include convergent validity, which can be evaluated by testing relations between a developed assessment and another related assessment, and discriminant validity, which can be evaluated by correlating scores from a developed assessment with an unrelated assessment. The goal of the former is to yield a high association which indicates that the developed measure converges, or is empirically linked to, the intended construct. The goal of the latter is to yield a lower association which indicates that the developed measure is unrelated to a particular construct of interest.

Predictive Validity

The predictive validity of scores to the selected criterions were 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 and among each of the EarlyBird Assessments and norm-referenced tests. Second, a multiple regression was run to estimate the total amount of variance that the linear combination of selected predictors explained in selected criterions.

Classification Accuracy

Logistic regressions were used, in part, to calibrate classification accuracy. Grade 1 and Grade 2 study participants' performance on the selected criterions were coded as '1' for performance above the 40th percentile for each grade on the KTEA-3 Letter and Word Recognition subtest (for "high performance" PWR) or at or below the 16th percentile for each grade on the KTEA-3 Letter and Word Recognition subtest (for "low performance" PWR), and '0' for scores that did not meet these criteria. In this way, the "high performance" PWR represents a prediction of success and the "low performance" PWR is a prediction of severe risk. Each dichotomous variable was then regressed on a combination of EarlyBird Assessments. 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, EarlyBird reports the likelihood of reading success both in terms of a percent and by classifying students into categories based on their performance.

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., 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. Optimal models were selected based on a combination of theory, parsimony, and classification accuracy values. Model parameters were set to simultaneously maximize sensitivity while minimizing the difference between sensitivity and specificity. This approach was utilized to maximize the ability to identify students who exceeded thresholds while attempting to strike a balance with attempting to identify students who failed to reach a given threshold.

Technical Documentation

The following sections provide technical documentation of the Reliability and Validity of the assessment as well as additional details about the samples, data collection efforts, and associated findings of multiple studies conducted by Boston Children's Hospital (BCH), Florida State University (FSU), EarlyBird, and Reach Every Reader (RER) in Appendices A, B, C, and D respectively.

The subtests that comprise the EarlyBird Grade 1 Assessment were developed and validated through a combination of these studies. The app-based Letter Sounds, Deletion, Word Matching, and Following Directions subtests and the supplemental Reading Comprehension subtest were all originally created and validated by FSU. Further validation of all of the first grade subtests originally created by BCH and FSU was conducted by EarlyBird in 2022. The uppercase Letter RAN subtest was created by EarlyBird and validated in the 2022 study (see Appendix C). The app-based Word Reading, Nonword Reading, Nonword Repetition, Nonword Spelling, and Expressive Vocabulary subtests and the supplemental Oral Reading Fluency subtest were all developed and validated by RER (see Appendix D).

The subtests that comprise the EarlyBird Grade 2 Assessment were developed and validated through a combination of studies as well. The app-based Word Matching and Follow Directions subtests and the supplemental Reading Comprehension subtest were all originally created and validated by FSU (see Appendix B). The app-based Word Reading, Nonword Reading, Nonword Repetition, lowercase Letter RAN, Expressive Vocabulary, and Nonword Spelling subtests and the supplemental Oral Reading Fluency subtest were all developed and validated by RER (see Appendix D).

Grade 1 - Summary of Marginal and Empirical Reliability

Marginal Reliability of FSU Tasks
(FSU Study, 2014)

Subtest	Fall/BOY
Word Matching	0.86
Following Directions	0.93

Marginal Reliability of EarlyBird Validation Tasks
(EarlyBird Grade 1 Validation Study, Winter 2022; n = 385 Grade 1 students)

Task	Winter/MOY
Word Matching	.83
Following Directions	.85
RAN*	-

*RAN is a time-limited task and does not have a marginal reliability estimate.

Empirical Reliability of Computer Adaptive Subtests
(EarlyBird Grade 1 Customer Data, August 2022 - June 2023; n = 8,593 Grade 1 students)

Subtest	Fall/BOY	Winter/MOY	Spring/EOY
Word Matching	0.82	0.84	0.84
Following Directions	0.85	0.85	0.85
RAN*	-	-	-

*RAN is a time-limited task and does not have a marginal reliability estimate.

Empirical Reliability of Letter Sound (Expressive Inventory) Subtest
(EarlyBird Grade 1 Customer Data, August 2023 - November 2023; n = 5,802 Grade 1 students)

Subtest	Fall/BOY
Letter Sound	0.95

Marginal Reliability Estimates of RER Validation Tasks
(RER Grade 1 Validation Study, 2018-2024)

Task	Fall/BOY	Winter/MOY	Spring/EOY
Word Reading	0.88	0.94	0.90
Nonword Reading	0.88	0.88	0.88
(Nonword) Spelling	0.80	0.83	0.80
Expressive Vocabulary	0.94	0.94	0.94
Nonword Repetition	0.93	0.93	0.93
RAN*	–	–	–

*RAN is a time-limited task and does not have a marginal reliability estimate.

Grade 2 - Summary of Marginal and Empirical Reliability

Marginal Reliability of Computer Adaptive Subtests
(FSU Grade 2 Validation Study Data, Spring 2014)

Task	Spring/EOY
Word Matching	.85
Following Directions	.93
	-

Empirical Reliability of Computer Adaptive Subtests
(EarlyBird Grade 2 Customer Data, March-June 2024; n = 1,035 Grade 2 students)

Subtest	Spring/EOY
Word Matching	.82
Follow Directions	.88
RAN*	-

*RAN is a time-limited task and does not have a marginal reliability estimate.

Note: Empirical reliability from [2024] customer data from [n] Grade 2 students

Summary of Marginal Reliability Estimates of RER Validation Tasks
(RER Grade 2 Validation Study, 2018-2023)

Task	Fall/BOY	Winter/MOY	Spring
Word Reading	0.86	0.85	0.84
Nonword Reading	0.86	0.87	0.85
(Nonword) Spelling	0.76	0.85	0.82
Expressive Vocabulary	0.80	0.80	0.80
Nonword Repetition	0.70	0.70	–
RAN*	–	–	–

*RAN is a time-limited task and does not have a marginal reliability estimate.

Grade 1 - Summary of Concurrent and Construct Validity

Grade 1 Concurrent and Construct Validity
(EarlyBird Study, Spring 2022)

Descriptive Statistics and Correlation Table for EarlyBird Subtests and Standardized Assessments of Language and Literacy (N = 367)

Tests	M	SD	DEL	FD	WM	RAN	KTEA3	WID	WATT	WBSKSL	CRLN
DEL	1.02	1.55		.30	.24	-.31	.40	.40	.38	.41	.26
FD	-0.30	1.29			.34	-.22	.41	.35	.32	.35	.16
WM	0.18	1.25				-.27	.45	.40	.35	.39	.18
RAN	38.79	10.02					-.49	-.49	-.46	-.49	-.58
KTEA3	100.65	14.82						.89	.77	.88	.56
WID	98.25	16.47							.82	.96	.53
WATT	101.04	14.92								.95	.47
WBSKSL	99.52	16.02									.53
CRLN	96.82	8.06									

Note. *Note.* M = average theta score or value; DEL = Deletion, FD = Following Directions, WM = Word Matching; RAN = Letter RAN (time in seconds); KTEA3 = KTEA-3 Reading Comprehension; WID = WRMT III Word ID; WATT = WRMT III Word attack; WBSKSL = WRMT III Composite; CRLN = CTOPP2 Rapid letter naming.

Note. See Appendix C for full EarlyBird study details

Fall Grade 1 convergent validity correlations

(RER validation study, 2018-2023)

	WRE	NRE	SPE	EVO	NWR
WRE	1.00				
NRE	0.74	1.00			
SPE	0.62	0.73	1.00		
EVO	0.40	0.51	0.37	1.00	
NWR	0.09	0.01	-0.01	-0.09	1.00

Note. WRE= Word Reading, NRE = Nonword Reading, SPE = Spelling, EVO = Expressive Vocabulary, NWR = Nonword Repetition

Winter Grade 1 convergent validity correlations

(RER validation study, 2018-2023)

Measure	WRE	NRE	SPE	EVO	NWR
WRE	1.00				
NRE	0.81	1.00			
SPE	0.60	0.62	1.00		
EVO	0.34	0.29	0.22	1.00	
NWR	0.48	0.49	0.60	0.31	1.00

Note. WRE= Word Reading, NRE = Nonword Reading, SPE = Spelling, EVO = Expressive Vocabulary, NWR = Nonword Repetition, ORF = Oral Reading Fluency.

Spring Grade 1 convergent validity correlations

(RER validation study, 2018-2023)

Measure	WRE	NRE	SPE	EVO	NWR
WRE	1.00				
NRE	0.76	1.00			
SPE	0.50	0.53	1.00		
EVO	0.30	0.32	0.27	1.00	
NWR	0.26	0.41	0.54	0.34	1.00

Note. DEL = Deletion, WRE= Word Reading, NRE = Nonword Reading, SPE = Spelling, SYN = Synonyms, EVO = Expressive Vocabulary, NWR = Nonword Repetition, SRT = Sentence Repetition, RAO = RAN Objects, ORF = Oral Reading Fluency.

Concurrent validity between IEAS measures and selected standardized assessments
(RER validation study, 2018-2023)

Grade	RER	Outcome	*r*	*n*
G1	NRE	KTEA NWD	0.76	711
G1	EVO	WPPSI PN	0.47	717

Note. NRE = Nonword Reading, EVO = Expressive Vocabulary, KTEA NWD = Kauffman Test Educational Achievement Nonword Decoding, WPPSI PN = WPPSI Picture Naming, WPPSI, EVT = Expressive Vocabulary Test.

Grade 2 - Summary of Concurrent and Construct Validity

Fall grade 2 convergent validity correlations
(RER validation study, 2018-2023)

	WRE	NRE	SPE	EVO	NWR
WRE	1.00				
NRE	0.75	1.00			
SPE	0.72	0.68	1.00		
EVO	0.56	0.53	0.61	1.00	
NWR	0.39	0.24	0.41	0.37	1.00

Note. WRE= Word Reading, NRE = Nonword Reading, SPE = Spelling, EVO = Expressive Vocabulary, NWR = Nonword Repetition

Winter grade 2 convergent validity correlations
(RER validation study, 2018-2023)

Measure	*WRE*	*NRE*	*EVO*	*NWR*
WRE	*1.00*			
NRE	*0.67*	*1.00*		
EVO	*0.35*	*0.20*	*1.00*	
NWR	*0.34*	*0.26*	*0.17*	*1.00*

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

Spring grade 2 convergent validity correlations
(RER validation study, 2018-2023)

Measure	*WRE*	*NRE*	*SPE*	*EVO*	*NWR*
WRE	*1.00*				
NRE	*0.79*	*1.00*			
SPE	*0.56*	*0.44*	*1.00*		
EVO	*0.46*	*0.42*	*0.09*	*1.00*	
NWR	*0.38*	*0.34*	*0.17*	*0.22*	*1.00*

Note. 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.

Concurrent validity between IEAS measures and selected standardized assessments
(RER validation study, 2018-2023)

Grade	RER	Outcome	*r*	*n*
G2	NRE	KTEA NWD	0.72	542
G2	EVO	EVT	0.72	537

Note. NRE = Nonword Reading, EVO = Expressive Vocabulary, KTEA NWD = Kauffman Test Educational Achievement Nonword Decoding, WPPSI PN = WPPSI Picture Naming, WPPSI, EVT = Expressive Vocabulary Test.

Grade 1 & Grade 2 Summary of Classification Accuracy

Classification accuracy for screening algorithms by grade, time point, and algorithm type (RER validation study, 2018-2023)

Grade	*Time Point*	*Cut Point*	*AUC*	*SE*	*SP*	*NPV*	*PPV*
Grade 1	*Fall*	*high PWR*	*0.92*	*0.85*	*0.81*	*0.89*	*0.75*
		low PWR	*0.93*	*0.86*	*0.82*	*0.97*	*0.46*
	Winter	*high PWR*	*0.94*	*0.85*	*0.86*	*0.90*	*0.78*
		low PWR	*0.95*	*0.86*	*0.85*	*0.97*	*0.51*
	Spring	*high PWR*	*0.95*	*0.90*	*0.85*	*0.92*	*0.80*
		low PWR	*0.92*	*0.87*	*0.83*	*0.97*	*0.48*
Grade 2	*Fall*	*high PWR*	*0.91*	*0.85*	*0.86*	*0.90*	*0.81*
		low PWR	*0.94*	*0.90*	*0.86*	*0.98*	*0.54*
	Winter	*high PWR*	*0.95*	*0.91*	*0.89*	*0.93*	*0.84*
		low PWR	*0.95*	*0.92*	*0.86*	*0.98*	*0.55*
	Spring	*high PWR*	*0.96*	*0.90*	*0.91*	*0.93*	*0.87*
		low PWR	*0.95*	*0.90*	*0.89*	*0.98*	*0.61*

Note. PWR = potential for work reading, DYS = dyslexia risk, AUC = area under the curve, SE = sensitivity, SP = specificity, NPV = negative predictive value, PPV = positive predictive value.

Note. See Appendix D for study details

Grade 1 Predictive and Concurrent Validity Coefficients

Fall Predictive Validity Coefficient for low PWR/Dyslexia risk
(KTEA-3 Letter and Word Recognition)
Multiple r = .82, 95% CI - .79, .84, n = 447

Winter Predictive Validity Coefficient for low PWR/Dyslexia risk
(KTEA-3 Letter and Word Recognition)
Multiple r = .85, 95% CI - .82, .87, n = 627

Spring Concurrent Validity Coefficient for low PWR/Dyslexia risk
(KTEA-3 Letter and Word Recognition)
Multiple r = .86, 95% CI - .83, .89, n = 270

Grade 2 Predictive and Concurrent Validity Coefficients

Fall Predictive Validity Coefficient for low PWR/Dyslexia risk
(KTEA-3 Letter and Word Recognition)
Multiple r = .79, 95% CI - .73, .83, n = 288

Winter Predictive Validity Coefficient for low PWR/Dyslexia risk
(KTEA-3 Letter and Word Recognition)
Multiple r = .84, 95% CI - .78, .88, n = 509

Spring Concurrent Validity Coefficient for low PWR/Dyslexia risk
(KTEA-3 Letter and Word Recognition)
Multiple r = .84, 95% CI - .79, .88, n = 257

Differential Item Functioning (DIF)

Summary of DIF analysis results (BCH Study, 2019-2020)

Across all BCH-based tasks and comparisons, only 10 items (Word Matching subtest) demonstrated at DIF with at least a moderate effect size (i.e., ETS >= 1.0). 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. For full BCH study details, see Appendix A.

Summary of DIF analysis results (FSU Study, 2014)

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. For full FSU study details, see Appendix B.

Summary of DIF analysis results (RER Study, 2018-2023)

Differential test functioning was calculated for Black students, Hispanic students, Male (vs. Female) students, and students identified as English Language Learners (ELL). No significant differences were found for any demographic sub-group. See Tables 8-9. For full RER study details, see Appendix D.

Appendix A: Technical Documentation of Boston Children's Hospital (BCH) study

The Gaab Lab (then at Boston Children's Hospital) designed and executed two validation studies for BELS (now EarlyBird) over the course of the 2018/2019 (Pilot Study; results available upon request) and 2019/2020 (Validation Study) academic school year.

Procedures

BCH validation study was designed as a nationwide study. The first phase of validation was completed between August and November 2019. We assessed 419 children (215 female, 200 male, 4 unknown, average age of 5.08 years; Table 1 and 2) 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 exclusionary/inclusionary 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 Boston Children's Hospital (BCH) as well as Florida State University's (FSU) Florida Center for Reading Research. We added items to multiple screener components that were previously validated at FSU.

Differential Item Functioning (DIF)

DIF testing in the BCH study was estimated using the difR package (Magis, Beland, & Raiche, 2020) using the Mantel-Haenszel method (1959) for detecting uniform DIF. For each of the six MATRS tasks, DIF was tested for three primary contrasts: 1) Male vs. female, 2) White vs. Sample, and 3) Black vs. Sample. 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, 1988). 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.

Appendix B: Technical Documentation of Florida State University (FSU) study

Description of Calibration Sample

Data collection began by testing item pools for the subtests (i.e., Letter Sounds, Blending, Deletion, Word Reading, Vocabulary Pairs (later integrated with Word Matching) and Following Directions). A statewide representative sample of students that roughly reflected Florida's demographic diversity and academic ability (N ~ 2,400) was collected as part of a larger K-2 validation and linking study. 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. Tables 3-4 include the population values and derived weights applied to all analyses.

Linking Design & Item Response Analytic Framework

A common-item, non-equivalent groups design was used for collecting data in the 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.

Norming Studies

A statewide representative sample of students across multiple districts that roughly reflected the state's demographic diversity and academic ability (N ~ 28,000) was collected on students in Kindergarten through Grade 2. Table 5 provides a breakdown of the sample sizes used by grade level for each of the PWR adaptive tasks.

Differential Item Functioning

DIF testing 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. It was conducted with a multiple indicator multiple cause (MIMIC) 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 (ESSD) was generated, and was computed similar to a Cohen's (1988) *d* statistic. 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 PWR. 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 was that moderate correlations would be observed. Correlation results are reported in Table 7. 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 recent meta-analysis in the early grades (Weiser & Mathes, 2011).

Appendix C: Technical Documentation of EarlyBird Validation Study

The EarlyBird Grade 1 Validation Study, conducted during the 2021-2022 academic year, 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.

Participants and Procedure

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 EarlyBird-created items and content appropriate for Grade 1 students. Two subtests (Nonword Reading and Letter RAN) were developed by EarlyBird in conjunction with content area experts. (2024 update: note that the EarlyBird version of Nonword Reading and Nonword Repetition were later replaced by the RER version of those subtests for the 2024-2025 school year.)

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.

Psychometric Results

Item Response Analytic Framework

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 composed for the fixed form tests. All IRT models were estimated with flexMirt (Houts & Cai, 2020). Model quality was evaluated using local fit (i.e., performance of the individual items) and goodness-of-fit indices based on the M_2 statistic (Maydeu-Olivares, 2013), and the root mean square error of approximation based on M_2 ($RMSEA_2$). M_2 is often sensitive to sample size in terms of rejecting the fitted model, thus, the $RMSEA_2$ 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.

Appendix D: Technical Documentation of Reach Every Reader Study

The Interstellar Express Assessment System (IEAS; Petscher & Catts, 2021) is the set of screening, broad diagnostic, and targeted diagnostic of computer adaptive assessments (CAA) and curriculum-based measures (CBM) created and validated in the Reach Every Reader Project for the early identification of reading and language disabilities in kindergarten through grade 3 (KG-G3). It was validated through a longitudinal study conducted from 2018 to 2023.

Participants and Procedure

A total of 3,470 students from 35 schools across 7 districts participated. These students were distributed among 589 classrooms and about 29% of the assessments were conducted online.

Summary of Results

IEAS validation work on approximately 5,000 Kindergarten through Grade 3 students shows marginal reliability (>.80 for over 80% of students) and validity (r >.40 construct and concurrent validity) that is acceptable for children in this age range.

Differential test functioning was calculated for Black students, Hispanic students, Male (vs. Female) students, and students identified as English Language Learners (ELL). No significant differences were found for any demographic sub-group (Tables 8-11).

Additional details are available upon request.

Tables

Table 1

BCH sample characteristics Part I (BCH study, 2019-2020)

	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

BCH sample demographic characteristics, Part II (BCH study, 2019-2020)

		Sample	
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
	NA	271	62.16
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
	NA	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
	NA	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
	NA	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
	NA	244	55.96

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 3

U.S. population-based weight values (FSU study, 2014)

Race	FRL	ELL	Weigh t
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	18.3
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

Note. 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 4

U.S. population-based weight values (FSU study, 2014)
Sample weight 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.

Race	FRL	ELL	Letter Names & Letter Sounds	Blending_& Deletion	Following Directions	Word Match	Sentence Comprehension
White	Yes	Yes	1.063	1.098	1.098	1.098	1.117
White	Yes	No	0.824	0.800	0.802	0.802	0.796
White	No	Yes	0.891	0.854	0.854	0.854	0.854
White	No	No	0.681	0.672	0.672	0.672	0.675
Black	Yes	Yes	3.370	3.605	3.605	3.605	3.523
Black	Yes	No	1.442	1.395	1.386	1.386	1.375
Black	No	Yes	0.769	0.769	0.769	0.769	0.769
Black	No	No	0.935	0.921	0.932	0.932	0.927
Hispanic	Yes	Yes	1.507	1.972	1.972	1.912	1.903
Hispanic	Yes	No	1.565	1.528	1.520	1.520	1.535
Hispanic	No	Yes	2.836	2.754	2.754	2.754	2.714
Hispanic	No	No	1.298	1.352	1.369	1.369	1.342
Other	Yes	Yes	0.927	0.911	0.911	0.911	0.895
Other	Yes	No	0.617	0.609	0.610	0.622	0.640
Other	No	Yes	0.782	0.768	0.768	0.768	0.754
Other	No	No	0.604	0.570	0.553	0.571	0.582

Note. FRL = Free/reduced price lunch; ELL = English language learner. Note that Tables A.1 and A.2 should be used together. 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.

Table 5

Sample sizes (FSU study, 2014)

Grade	PA	LN/LS	SC	WM	FD	WR	Spelling
K	2,100	2,377	2,275	2,015	2,304	1,969	
1				2,115	2,247	2,372	
2				1,980	2,027	2,089	1,992
Total	2,100	2,377	2,275	6,110	6,578	6,430	1,992

Note. PA = phonological awareness blending and deletion, LN/LS = letter names and sounds, SC = sentence comprehension, WM = word matching, FD = following directions, WR = word reading.

Table 6

Marginal reliability coefficients (FSU study, 2014)

Grade	Task	Reliability
K	Blending	.99
	Deletion	.94
	Letter Sounds	.97
	Letter Names	.85
	Word Matching	.87
	Following Directions	.94
	Sentence Comprehension	.89
1	Vocabulary Pairs (Word Matching)	.86
	Following Directions	.93
2	Vocabulary Pairs (Word Matching)	.85
	Following Directions	.93

Note. CI = confidence interval

Table 7

Descriptive Statistics and Correlation Table for Grade 1 EarlyBird Subtests and Standardized Assessments of Language and Literacy *(EarlyBird study, 2022)*

n = 367

Tests	M	SD	DEL	FD	WM	RAN	KTEA3	WID	WATT	WBSKSL	CNWREP	CRLN
DEL	1.02	1.55		.30	.24	-.31	.40	.40	.38	.41	.23	.26
FD	-0.30	1.29			.34	-.22	.41	.35	.32	.35	.07	.16
WM	0.18	1.25				-.27	.45	.40	.35	.39	.04	.18
RAN	38.79	10.02					-.49	-.49	-.46	-.49	-.07	-.58
KTEA3	100.65	14.82						.89	.77	.88	.17	.56
WID	98.25	16.47							.82	.96	.21	.53
WATT	101.04	14.92								.95	.30	.47
WBSKSL	99.52	16.02									.28	.53
CNWREP	97.16	14.46										.15
CRLN	96.82	8.06										

Note. *Note.* M = average theta score or value; DEL = Deletion, FD = Following Directions, OSC = Oral Sentence Comprehension; WM = Word Matching; WR = Word reading; RAN = Letter RAN (time in seconds); KTEA3 = KTEA-3 Reading Comprehension; WID = WRMT III Word ID; WATT = WRMT III Word attack; WBSKSL = WRMT III Composite; CNWREP = CTOPP-2 nonword repetition; CRLN = CTOPP-2 Rapid letter naming.

Table 8

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
	fallpwrrisk	18.48	1021.48	0.02	0.986
	Black	15.75	1021.48	0.02	0.988
	fallpwrrisk*Black	-15.82	1021.48	-0.02	0.988
Fall - DYS	(Intercept)	-18.57	1537.40	-0.01	0.990
	falldysrisk	19.13	1537.40	0.01	0.990
	Black	15.67	1537.40	0.01	0.992
	falldysrisk*Black	-17.12	1537.40	-0.01	0.991
Winter - PWR	(Intercept)	-17.57	1057.33	-0.02	0.987
	winterpwrrisk	18.26	1057.33	0.02	0.986
	Black	15.38	1057.33	0.01	0.988
	winterpwrrisk*Black	-15.14	1057.33	-0.01	0.989
Winter - DYS	(Intercept)	-18.57	1581.97	-0.01	0.991
	winterdysrisk	18.90	1581.97	0.01	0.990
	Black	14.96	1581.97	0.01	0.992
	winterdysrisk*Black	-15.92	1581.97	-0.01	0.992

Fall - PWR	(Intercept)	-1.01	0.58	-1.73	0.083
	fallpwrrisk	2.40	0.87	2.75	0.006
	Latino	-0.81	0.66	-1.22	0.222
	fallpwrrisk*Latino	0.02	0.97	0.02	0.982
Fall - DYS	(Intercept)	-2.77	1.03	-2.69	0.007
	falldysrisk	2.62	1.17	2.24	0.025
	Latino	-0.04	1.11	-0.04	0.971
	falldysrisk*Latino	-0.90	1.32	-0.69	0.493
Winter - PWR	(Intercept)	-1.70	0.77	-2.22	0.027
	winterpwrrisk	3.25	1.00	3.25	0.001
	Latino	-0.65	0.86	-0.76	0.450
	winterpwrrisk*Latino	-0.14	1.11	-0.12	0.902
Winter - DYS	(Intercept)	-18.57	1581.97	-0.01	0.991
	winterdysrisk	18.72	1581.97	0.01	0.991
	Latino	15.08	1581.97	0.01	0.992
	winterdysrisk*Latino	-16.07	1581.97	-0.01	0.992
Fall - PWR	(Intercept)	-2.13	0.37	-5.68	0.000
	fallpwrrisk	2.74	0.50	5.49	0.000
	Female	0.24	0.52	0.47	0.636
	fallpwrrisk*Female	-0.24	0.69	-0.35	0.729

Fall - DYS	(Intercept)	-3.76	0.72	-5.26	0.000
	falldysrisk	3.39	0.82	4.15	0.000
	Female	1.03	0.85	1.21	0.228
	falldysrisk*Female	-1.29	1.02	-1.27	0.202
Winter - PWR	(Intercept)	-2.80	0.51	-5.44	0.000
	winterpwrrisk	3.50	0.60	5.79	0.000
	Female	0.30	0.69	0.44	0.663
	winterpwrrisk*Female	-0.09	0.83	-0.11	0.910
Winter - DYS	(Intercept)	-19.57	1159.64	-0.02	0.987
	winterdysrisk	19.36	1159.64	0.02	0.987
	Female	16.60	1159.64	0.01	0.989
	winterdysrisk*Female	-16.88	1159.64	-0.01	0.988
Fall - PWR	(Intercept)	-1.98	0.28	-7.20	0.000
	fallpwrrisk	2.70	0.37	7.33	0.000
	DLL	-0.16	0.80	-0.20	0.844
	fallpwrrisk*DLL	-0.15	1.05	-0.14	0.886
Fall - DYS	(Intercept)	-3.00	0.39	-7.75	0.000
	falldysrisk	2.61	0.48	5.45	0.000
	DLL	-15.56	1360.06	-0.01	0.991
	falldysrisk*DLL	15.26	1360.06	0.01	0.991

Winter - PWR	(Intercept)	-2.74	0.39	-7.02	0.000
	winterpwrrisk	3.56	0.46	7.81	0.000
	DLL	0.48	0.84	0.58	0.564
	winterpwrrisk*DLL	-0.21	1.19	-0.17	0.863
Winter - DYS	(Intercept)	-3.57	0.51	-7.04	0.000
	winterdysrisk	3.35	0.58	5.81	0.000
	DLL	-15.00	1360.06	-0.01	0.991
	winterdysrisk*DLL	14.53	1360.06	0.01	0.991

Table 9

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
	fallpwrrisk	3.70	0.93	3.99	0.000
	Black	-0.36	0.84	-0.42	0.672
	fallpwrrisk*Black	-0.74	1.04	-0.71	0.477
Fall - DYS	(Intercept)	-18.57	1279.19	-0.01	0.988
	falldysrisk	20.03	1279.19	0.02	0.988
	Black	14.57	1279.19	0.01	0.991
	falldysrisk*Black	-16.18	1279.20	-0.01	0.990
Winter - PWR	(Intercept)	-2.89	1.03	-2.81	0.005
	winterpwrrisk	4.45	1.17	3.82	0.000
	Black	-0.31	1.15	-0.27	0.788
	winterpwrrisk*Black	-0.31	1.31	-0.23	0.816
Winter - DYS	(Intercept)	-20.57	3477.21	-0.01	0.995
	winterdysrisk	22.03	3477.21	0.01	0.995
	Black	0.00	3853.39	0.00	1.000
	winterdysrisk*Black	-1.26	3853.39	0.00	1.000
Fall - PWR	(Intercept)	-2.74	0.73	-3.76	0.000

	fallpwrrisk	3.19	0.88	3.65	0.000
	Latino	0.27	0.84	0.32	0.749
	fallpwrrisk*Latino	-0.20	1.02	-0.19	0.846
Fall - DYS	(Intercept)	-2.40	0.60	-3.98	0.000
	falldysrisk	2.26	0.79	2.85	0.004
	Latino	-2.08	1.17	-1.77	0.076
	falldysrisk*Latino	2.15	1.33	1.61	0.107
Winter - PWR	(Intercept)	-3.56	1.01	-3.51	0.000
	winterpwrrisk	4.94	1.20	4.11	0.000
	Latino	0.34	1.17	0.29	0.774
	winterpwrrisk*Latino	-0.81	1.38	-0.58	0.559
Winter - DYS	(Intercept)	-2.97	0.73	-4.10	0.000
	winterdysrisk	4.36	1.07	4.06	0.000
	Latino	-17.60	1879.42	-0.01	0.993
	winterdysrisk*Latino	16.27	1879.42	0.01	0.993
Fall - PWR	(Intercept)	-3.12	0.59	-5.29	0.000
	fallpwrrisk	4.09	0.67	6.13	0.000
	Female	0.89	0.71	1.25	0.210
	fallpwrrisk*Female	-1.21	0.84	-1.43	0.151
Fall - DYS	(Intercept)	-19.57	1194.89	-0.02	0.987
	falldysrisk	20.01	1194.89	0.02	0.987

	Female	16.57	1194.89	0.01	0.989
	falldysrisk*Female	-16.89	1194.89	-0.01	0.989
Winter - PWR	(Intercept)	-3.57	0.72	-4.98	0.000
	winterpwrrisk	4.81	0.79	6.05	0.000
	Female	0.42	0.93	0.45	0.651
	winterpwrrisk*Female	-0.47	1.05	-0.44	0.659
Winter - DYS	(Intercept)	-19.57	1166.44	-0.02	0.987
	winterdysrisk	20.30	1166.44	0.02	0.986
	Female	15.82	1166.44	0.01	0.989
	winterdysrisk*Female	-15.91	1166.44	-0.01	0.989
Fall - PWR	(Intercept)	-2.64	0.37	-7.21	0.000
	fallpwrrisk	3.45	0.43	7.98	0.000
	DLL	0.29	0.83	0.35	0.727
	fallpwrrisk*DLL	-0.41	1.22	-0.34	0.737
Fall - DYS	(Intercept)	-4.23	0.71	-5.95	0.000
	falldysrisk	4.59	0.75	6.09	0.000
	DLL	1.79	1.03	1.75	0.080
	falldysrisk*DLL	-2.15	1.45	-1.48	0.139
Winter - PWR	(Intercept)	-3.19	0.46	-6.98	0.000
	winterpwrrisk	4.45	0.53	8.43	0.000
	DLL	-15.38	1458.51	-0.01	0.992

	winterpwrrisk*DLL	14.80	1458.51	0.01	0.992
Winter - DYS	(Intercept)	-20.57	1457.43	-0.01	0.989
	winterdysrisk	21.41	1457.43	0.01	0.988
	DLL	18.17	1457.43	0.01	0.990
	winterdysrisk*DLL	-19.42	1457.43	-0.01	0.989

References

Catts, H. W., & Petscher, Y. (2020). A Cumulative Risk and Protection Model of Dyslexia. https://doi.org/10.35542/osf.io/g57ph

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second ed.). Hillsdale: Lawrence Erlbaum Associates.

Ehri, L.C., Nunes, S., Stahl, S., & Willows, D. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*, 393-447.

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). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Foorman, B. R., Francis, D. J., Shaywitz, S. E., Shaywitz, B. A., & Fletcher, J. M.(1997). The case for early reading intervention. In B. A. Blachman (Ed.), Foundations of reading acquisition and dyslexia: Implications for early intervention (p. 243–264). Lawrence Erlbaum Associates Publishers.

Foorman, B. R., & Connor, C. (2011). Primary reading. In M. Kamil, P. D. Pearson, & E. Moje (Eds.), *Handbook on reading research* (Vol. 4, pp. 136–156). New York, NY: Taylor and Francis.

Foorman, B.R., Francis, D.J., Davidson, K., Harm, M., & Griffin, J. (2004). Variability in text features in six grade 1 basal reading programs. *Scientific Studies in Reading, 8*(2), 167 -197.

Houts, C. R., & Cai, L. (2020). flexMIRT user's manual version 3.52: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Magis, D., Beland, 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, 847-862.

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

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*. Los Angeles, CA: Muthén and 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). Washington, DC: U.S. Department of Health and Human Services.

National Research Council (1998). *Preventing reading difficulties in young children*. Committee on the Prevention of Reading Difficulties in Young Children, Committee on Behavioral and Social Science and Education, C.E. Snow, M.S. Burns, & P. Griffin, eds. Washington, D.C.: National Academy Press.

Norton, E. S., & Wolf, M. (2012). Rapid Automatized Naming (RAN) and Reading Fluency: Implications for Understanding and Treatment of Reading Disabilities. *Annual Review of Psychology*, *63*(1), 427–452. https://doi.org/10.1146/annurev-psych-120710-100431

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd Ed.). New York: McGraw-Hill.

Ozernov-Palchik, O., & Gaab, N. (2016a). "Tackling the 'dyslexia paradox': reading brain and behavior for early markers of developmental dyslexia." Wiley Interdisciplinary Reviews: Cognitive Science, 7(2), 156-176. doi: 10.1002/wcs.1383

Ozernov-Palchik, O., Yu, X., Wang, Y., & Gaab, N. (2016b) 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. DOI: 10.1080/10888438.2013.827687

RAND Reading Study Group (2002). *Reading for understanding*. Santa Monica, CA: 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.

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

Schatschneider, C., Petscher, Y., & Williams, K.M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know (pg. 304-317). In L. Justice & C. Vukelic (Eds.). *Every moment counts: Achieving excellence in preschool language and literacy instruction.* New York: Guilford Press.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

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.

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, & Computers, 36(3), 481–487. https://doi.org/10.3758/BF03195594

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's the pronunciation for _OUGH and the spelling for /u/? A database for computing feedforward and feedback inconsistency in English. *Behavior Research Methods, Instruments, & Computers, 29*, 600-618.

EarlyBird

The early bird gets to learn

About EarlyBird

EarlyBird transforms students’ lives through the early detection of reading difficulties, including dyslexia. Developed and scientifically validated at Boston Children’s Hospital in partnership with faculty at the Florida Center for Reading Research, EarlyBird brings together all the relevant predictors of reading in one easy-to-administer assessment. The cloud-based technology platform includes a game-based app for students and a dashboard that points teachers to customized action plans and evidence-based resources. With EarlyBird, educators can identify children at risk for reading difficulties starting in the window when intervention is most effective — before they formally learn to read.

For information, visit www.earlybirdeducation.com