EarlyBird

EarlyBird Kindergarten Dyslexia and Early Literacy Assessment

Technical Manual

August 2024

Table of Contents

Please note that this technical manual provides information about the kindergarten assessment system. For data specific to the EarlyBird assessments designed for other grades, please see Chapters 5 (Reliability) and 6 (Validity) in those respective technical manuals.

Acknowledgements

This Technical Manual for the Early Bird Kindergarten 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 to 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.

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 received intensive reading instruction, 56%–92% (across six research studies) 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 kindergarten. 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 Assessment, the alphabetic principle is assessed with individually-administered tasks that measure letter-sound knowledge, phonological awareness, and word reading (for more advanced kindergarteners).

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 Assessment, knowledge of word meanings is measured in kindergarten by two vocabulary tasks: 1) a word matching task called Word Matching and 2) a receptive vocabulary task called Vocabulary. During the Word Matching task, the child's task is to tough the two out of three words (or pictures) which are also presented orally that go together (e.g., blue, triangle, yellow). During the Vocabulary task, students hear a spoken word and need to decide which one of the four presented pictures represents that word.

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 Assessment, there are two tasks in kindergarten that address oral listening comprehension/syntactic awareness . One is called Following Directions and 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). The other task is called Oral Listening Comprehension and requires that the student listen to a sentence and touch the one of four pictures which best represents the sentence (e.g., point to the picture of the bird flying away from the nest).

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, including vocabulary and oral listening comprehension (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, vocabulary and oral listening skills. The EarlyBird tool incorporates all of these skills as outlined below.

Description of EarlyBird Game System

The EarlyBird Kindergarten Assessment is a gamified mobile app, that is easy, quick, accessible, and child-centered, and can be completed prior to formal reading instruction. It is self-administered in small groups with teacher oversight and, depending on the subtests administered, takes 20-40 minutes per child. The assessments address literacy milestones that have been found to be predictive of subsequent reading success in kindergarten aged children. No trained adult administration is needed. Scoring is largely automated. EarlyBird includes screening for severe reading risk (hereafter referred to as Dyslexia Screener) and moderate reading risk (hereafter referred to as Potential for Word Reading, or PWR screener). The technical documentation is presented separately for each screener system even though the assessment system 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, and demonstrate each assessment before the child attempts them. At the end of each game, the child is rewarded with a virtual prizeand 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

Rhyming – Moose: 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: 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: 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: 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 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 Name - Crocodile: 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: 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) at the end of year. 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: 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.

Fluency

Rapid Automatized Naming - Polar bear: The Rapid Automatized Naming task uses a set of five objects *(house, door, cat, ear, bed)* that are repeated in random order in five rows, totaling 50 objects. The student is measured on how fast he or she is able to name each object out loud across each row. The number of seconds it takes for the student to name all 50 objects 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 - 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. These words represent a wide range of difficulty, from single-syllable to multi-syllable words, with a mix of decodable words and sight words.

Vocabulary

Vocabulary – Alpaca: 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.

Word Matching - Gorilla: Word Matching is a computer adaptive task that measures the ability to perceive relationships between words that are related by semantic class features where 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

Oral Sentence Comprehension - Rhino: 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: 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 names, characteristics, and order of mention of pictures; and identify from among several choices the targeted objects. Items consist of an array of objects on the screen and a set of audio instructions. Students respond to the directions by touching the specified objects on the screen, as instructed (e.g., "Click on the cat and then click on the heart").

Chapter 3: Score Definitions

Several different kinds of 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. A percentile rank is provided for each computer adaptive task at each time point.

Potential For Word Reading (PWR)

The PWR score is the probability, expressed as a percentage, that a student will reach grade-level expectations in word reading by the end of the year. How it works: An analysis was done to determine which subtest scores, for the particular time of year (fall, winter or spring), are most predictive of achieving targeted grade-level performance at the end of the year. Reaching expectations, for the purposes of this analysis, is defined as performing above 40th percentile on the SESAT – 10 (Stanford Early School Achievement Test): a reasonable standard for measuring grade level expectation word reading in Kindergarten. The PWR score is a multi-factoral calculation that involves a selection of the most predictive subtests and an aggregation and weight averaging of that data according to degree of predictability to generate a single output score. For Kindergarten, the screening tasks include phonological awareness blending, letter sound knowledge and word matching. The PWR score appears in the data dashboard at middle of year, as it is based on normative sample data from the winter testing period.

Dyslexia Risk Flag

The Dyslexia Risk Flag indicates the likelihood that a student will be at risk for reading struggles as determined by poor phonological processing skills at the end of the school year, presuming the student doesn't receive appropriate remediation. This risk calculation is based on a compilation of research conducted by the authors and other leaders in the field, revealing that those with severe word reading risk profiles are likely to have dyslexia. How it works: An analysis was done to determine which subtest scores are most predictive of the targeted performance at the end of the year. For the purposes of the analysis, dyslexia risk is defined as performing at or below the 16th percentile on the KTEA-3 Phonological Processing subtest (inclusive of blending, rhyming, sound matching, deletion, and segmenting items). The calculation involves logistic regression and receiver operating characteristic curve analyses with a selection of our most predictive subtests (rhyming, nonword repetition, and follow directions) and an aggregation and weight averaging of that data according to degree of predictability to generate a single output score which is conveyed as a "flag". That flag indicates the likelihood that a student would score poorly on the KTEA-3 task. Any child flagged for dyslexia risk is at high risk for low phonological processing skills and therefore subsequent low reading proficiency and needs intensive instruction targeted to the student's skill weaknesses. The Dyslexia Risk Flag can be administered/calculated any time of year, recognizing that it is based on a normative sample performance for the late fall/early winter period.

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 kindergarten student who scored at the 60th percentile would have obtained a score 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 Name and Letter Sound 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 name inventory of 26, the ratio on the data dashboard would show 20/26.

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.

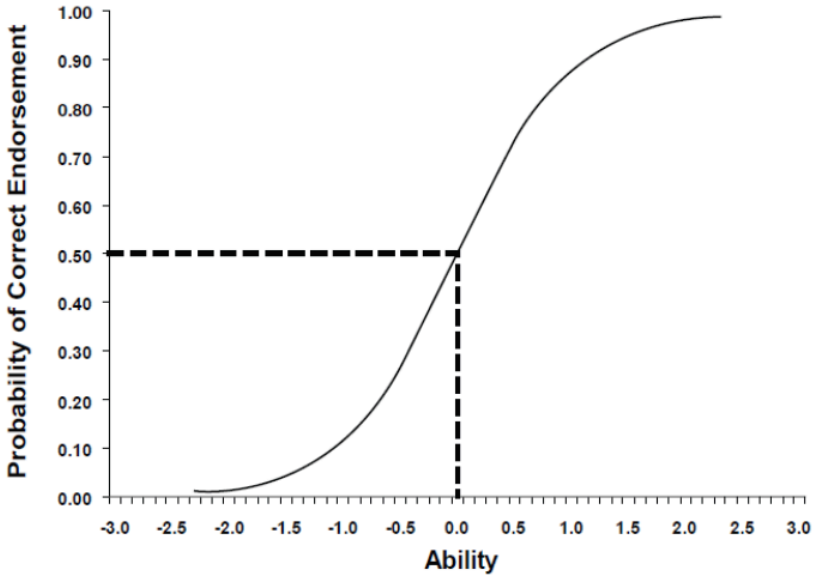
Item difficulty. The difficulty of an item (*b*) has traditionally been described for many tests as a "p-value", which corresponds to the percent of respondents correctly answering an item.

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.

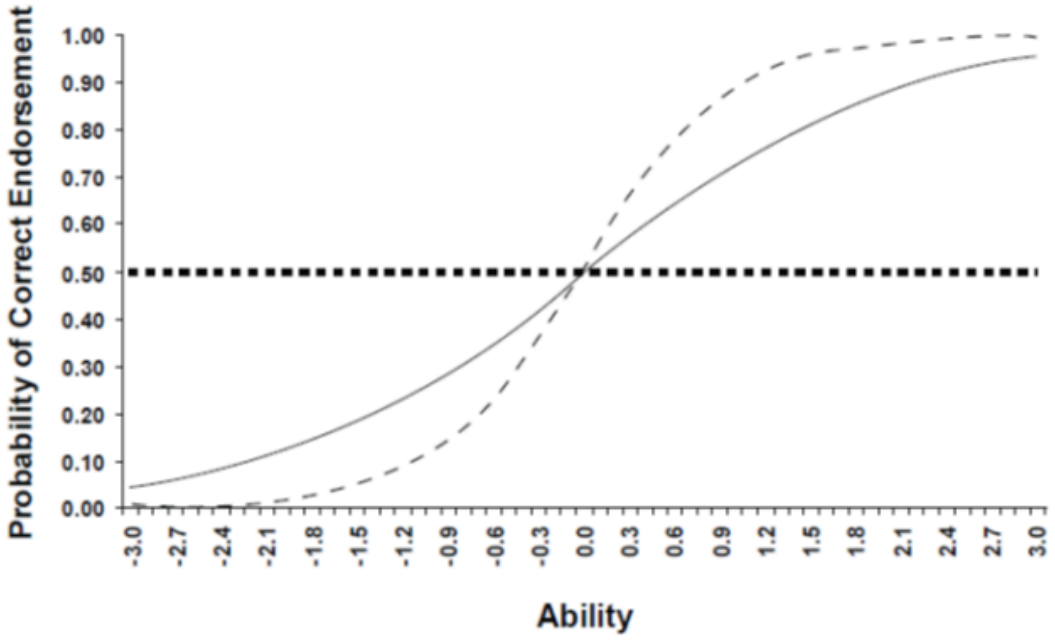
Figure 1: Sample Item Characteristic Curve



Item Discrimination. Item Discrimination 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



2PL models were fit using mirt package (Chalmers, 2012) and were evaluated using local fit (i.e., performance of the individual items) and goodness-of-fit based onβ the M_2 statistic (Maydeu-Olivares, 2013), the root mean square error of approximation based on M_2 ($RMSEA_2$), the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI). 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].

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 tasks. 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 $| \theta - b |$. Steps 2-4 continue until the target SE is reached or until a pre-determined 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 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. For the PWR assessments, 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.

DIF testing in the PWR study 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.

DIF testing in the Dyslexia Risk 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.

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 PWR 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 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. Students' performance on the selected criterions were coded as '1' for performance at or above the 40th percentile on the SESAT (for PWR) or below the 16th percentile on the KTEA-3 Phonological Processing (for Dyslexia Risk flag), and '0' for scores that did not meet these criteria. In this way, the PWR represents a prediction of success and the Dyslexia flag is a prediction of 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 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., 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.

Technical Documentation

The following sections provide technical detailing of the samples, data collection efforts, and associated results with the Dyslexia Risk and PWR Screener systems in the kindergarten assessment. Information about the EarlyBird assessments designed for other grades can be found in their respective technical manuals. The kindergarten Dyslexia Risk system was validated through efforts at Boston Children's Hospital and the kindergarten PWR system was validated through efforts at Florida State University; thus, validation processes are described within each system along with evidence from an integration study led by Boston Children's Hospital. To facilitate access to pertinent technical information, Part I provides a brief summary of results approximately commensurate with criteria from the National Center on Intensive Intervention's (NCII) academic screening tool chart criteria. Part II provides detailed documentation of procedures and results from the Dyslexia Risk Screener Validation Study. Part III provides detailed documentation of procedures and results from the PWR Screener Validation Study.

Summary of Marginal Reliability of Validation Data (BCH and FSU Studies)

Model-Based Marginal Reliability with Confidence Interval from 2PL Unidimensional, Item Response Theory

Task	Marginal Reliability	95% CI LB	95% CI UB
First Sound Matching	0.88	0.87	0.91
NonWord Repetition	0.91	0.89	0.92
Word Matching	0.91	0.90	0.92
Letter Names*	0.85	0.83	0.87
Letter Sounds*	0.97	0.96	0.97
Phonological Awareness Blending	0.99	0.98	0.99
Phonological Awareness Deletion	0.94	0.93	0.95
Word Matching	0.87	0.84	0.89
Following Directions	0.94	0.93	0.94
Word Reading	0.98	0.97	0.99
Sentence Comprehension	0.89	0.88	0.90
RAN	-	-	-

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

*Letter Name and Letter Sound subtests were receptive and computer-adaptive for the validation study. In the current product, each of these subtests is an expressive inventory - see below for reliability based on customer data using the expressive inventory version of these subtests.

Summary of Empirical Reliability of Kindergarten Letter Name and Letter Sound Subtests (Expressive Inventory version) at BOY

(EarlyBird Customer Data, August 2023 - November 2023; n = ~9,300)

Task	Marginal Reliability
Letter Name*	0.99
Letter Sound*	0.90

**Expressive inventory*

*Fall/Winter Classification Accuracy for Dyslexia Risk (**16th percentile** KTEA – Phonological Processing)*
Mean Bootstrapped Area Under the Curve = 0.85 (95% Confidence Interval = 0.80, 0.90)
Mean Bootstrapped Sensitivity = 0.81
Mean Bootstrapped Specificity = 0.80
True Positive N = 24, True Negative N = 124, False Positive N = 29, False Negative N = 7.
Base rate = 16.8%

*Fall/Winter Classification Accuracy for Potential for Word Reading Success (**40th percentile** –SESAT Word Reading)*
Mean Bootstrapped Area Under the Curve = 0.84 (95% Confidence Interval = 0.81, 0.88)
Mean Bootstrapped Sensitivity = 0.81
Mean Bootstrapped Specificity = 0.72
True Positive N = 60, True Negative N = 74, False Positive N = 29, False Negative N = 14.
Base rate = 41.8%

*Spring Classification Accuracy for Potential for Word Reading Success (**40th percentile** –SESAT Word Reading)*
Mean Bootstrapped Area Under the Curve = 0.88 (95% Confidence Interval = 0.86, 0.90)
Mean Bootstrapped Sensitivity = 0.83
Mean Bootstrapped Specificity = 0.87
True Positive N = 58, True Negative N = 79, False Positive N = 12, False Negative N = 13.
Base rate = 41.8%

Fall/Winter Predictive Validity Coefficient for Dyslexia Risk (KTEA – Phonological Processing)
Multiple r = .61, 95% CI = .50, .69, n = 184

Fall/Winter Predictive Validity Coefficient for Potential for Word Reading Success (SESAT Word Reading)
Multiple r = .67, 95% CI = .58, .74, n = 183

Spring Concurrent Validity Coefficient for Potential for Word Reading Success (SESAT Word Reading)
Multiple r = .73, 95% CI = .65, .79, n = 164

Correlations among Winter EarlyBird subtests, and Spring Standardized Assessment Battery

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. BL (N=210)	-2.23	2.38																				
2. DEL (N=101)	-1.14	1.41	**.44**																			
3. FD (N=97)	-1.09	1.25	.17	.17																		
4. LN (N=94)	3.07	2.87	.14	**.26**	**.27**																	
5. LS (N=213)	-0.08	1.60	**.39**	**.31**	.18	**.23**																
6. SS (N=97)	-0.32	0.94	.17	.17	**.26**	.18	**.25**															
7. VP (N=213)	0.33	1.64	**.20**	**.23**	.09	**.22**	**.22**	**.34**														
8. FSM (N=191)	0.18	0.99	**.56**	**.37**	.15	**.26**	**.51**	.16	**.25**													
9. NWR (N=200)	0.12	0.85	**.48**	**.44**	.16	**.27**	**.37**	.16	**.24**	**.51**												
10. RHYM (N=195)	0.14	0.90	**.26**	**.23**	**.34**	**.47**	**.25**	.20	**.38**	**.39**	**.43**											
11. WM (N=204)	0.09	0.89	**.30**	**.32**	**.34**	**.40**	**.30**	**.48**	**.34**	**.46**	**.51**	**.39**										
12. RAN (N=175)	84.27	27.86	**-.30**	-.09	-.09	**-.24**	**-.28**	-.16	**-.22**	**-.41**	**-.23**	**-.23**	**-.25**									
13. K-LWR (N=139)	96.78	13.50	**.27**	.14	**.33**	.19	**.43**	**.28**	.13	**.45**	**.25**	**.34**	**.22**	**-.26**								
14. K-NWD (N=111)	92.45	13.01	**.23**	.12	.21	**.28**	**.31**	.26	.15	**.46**	**.27**	**.42**	**.24**	**-.33**	**.86**							
15. K-P (N=215)	100.02	14.85	**.39**	**.46**	**.33**	**.30**	**.35**	.17	**.27**	**.51**	**.46**	**.53**	**.36**	**-.20**	**.63**	**.72**						
16. K-D (N=103)	93.17	14.74	**.31**	.15	**.39**	.23	**.31**	.24	.12	**.50**	**.24**	**.43**	**.23**	**-.40**	**.84**	**.81**	**.84**					
17. WID (N=192)	97.16	16.29	**.22**	.05	**.34**	**.24**	**.34**	.17	**.18**	**.48**	**.23**	**.41**	**.24**	**-.26**	**.91**	**.77**	**.62**	**.84**				
18. WA (N=192)	99.31	14.45	**.30**	.03	**.29**	**.28**	**.27**	.15	**.23**	**.49**	**.34**	**.47**	**.32**	**-.24**	**.79**	**.75**	**.71**	**.80**	**.81**			
19. SWE (N=108)	94.03	16.44	**.31**	.23	**.44**	.19	**.40**	.20	**.22**	**.52**	**.30**	**.44**	**.26**	**-.27**	**.92**	**.79**	**.71**	**.88**	**.91**	**.77**		
20. PDE (N=108)	94.95	13.62	**.27**	.18	**.39**	.14	**.34**	.05	.18	**.54**	**.30**	**.44**	**.30**	**-.34**	**.82**	**.82**	**.71**	**.84**	**.77**	**.73**	**.85**	
21. CELF (N=219)	103.26	12.17	.12	.08	**.43**	.37	.25	.38	.28	.27	**.44**	**.43**	**.45**	-.02	.21	**.34**	**.46**	.31	.21	.31	.29	**.36**

Note: M = mean, *SD*= standard deviation, BL = Blending, DEL = Deletion, FD = Following Directions, LN = Letter Name Knowledge, LS = Letter Sound Knowledge, SS = Sentence Structure/Oral Sentence Comprehension), VP = Vocabulary/Vocabulary Pairs,, FSM = First Sound Matching, NWR = Nonword Repetition, RHYM = Rhyming, WM = Word Matching, K-LWR = KTEA Letter Word Recognition, K-NWD = KTEA Nonsense Word Decoding, K-P = KTEA Phonological Processing, K-D = KTEA Dyslexia, WID = WRMT Word Identification, WA = WRMT Word Attack, SWE = TOWRE Sight Word Efficiency, PDE = TOWRE Phonemic Decoding Efficiency, CELF = CELF Sentence Structure. Bold values indicates $p < .05$.

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 to assess predictive validity of the screener. The goal of the study was to measure the extent to which BELS 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; 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 assessment components that were previously validated at FSU.

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 be 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 8 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 3 unknown, average age 6 years 7 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 this spring, 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 the app last fall and from the psychometric assessments this year, allowed for the evaluation of the screener's predictive validity.

Psychometric Results

Classical Test Theory Results

First Sound Matching (FSM)

The mean p-value (i.e., percent correct) for FSM items was 0.59 (SD = 0.15) with a minimum of 0.39 and a maximum of 0.91.

NonWord Repetition (NWR)

The mean p-value (i.e., percent correct) for NWR items was 0.49 (SD = 0.22) with a minimum of 0.08 and a maximum of 0.91.

Rhyming (RHYM)

The mean p-value (i.e., percent correct) for RHYM items was 0.67 (SD = 0.14) with a minimum of 0.36 and a maximum of 0.89.

Word Matching (WM)

The mean p-value (i.e., percent correct) for WM items was 0.68 (SD = 0.23) with a minimum of 0.24 and a maximum of 0.98. Eleven items presented with ceiling effects (i.e., p-value >= .95) and were removed from the item bank.

Multiple-Group Item Response Modeling (MG-IRM)

First Sound Matching (FSM)

Model fit for the FSM unidimensional 2PL-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 =911.91, p <.001; Table 3); however, global fit suggested good model fit to the data, CFI = .98, TLI = .98, RMSEA = .031 (95% CI = .026, .036). The mean *b* value was -0.49 (SD = 0.79) with a minimum of -3.02 and a maximum of 0.39. The mean *a* value was 1.35 (SD = 0.51) with a minimum of 0.40 and a maximum of 2.46. Marginal reliability was .87.

NonWord Repetition (NWR)

Model fit for the FSM unidimensional 2PL-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 =996.47, p <.001; Table 3); however, global fit suggested good model fit to the data, CFI = .98, TLI = .98, RMSEA = .030 (95% CI = .025, .035). The mean *b* value was 0.07 (SD = 1.06) with a minimum of -2.16 and a maximum of 2.64. The mean *a* value was 1.35 (SD = 0.35) with a minimum of 0.57 and a maximum of 2.19. Marginal reliability was .87.

Rhyming (RHYM)

Model fit for the RHYM unidimensional 2PL-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 =925.36, p <.001; Table 3); however, global fit suggested good model fit to the data, CFI = .98, TLI = .98, RMSEA = .032 (95% CI = .027, .067). The mean *b* value was -0.73 (SD = 0.64) with a minimum of -1.94 and a maximum of 0.64. The mean *a* value was 1.55 (SD = 0.65) with a minimum of 0.63 and a maximum of 3.16. Marginal reliability was .89.

Word Matching (WM)

Model fit for the FSM unidimensional 2PL-IRM resulted in a rejection of the null hypothesis of a correctly specified model (M_2 =1958.60, p <.001; Table 3); however, global fit suggested good model fit to the data, CFI = .97, TLI = .97, RMSEA = .019 (95% CI = .016, .024). The mean *b* value was -0.77 (SD = 2.12) with a minimum of -3.66 and a maximum of 8.24. The mean *a* value was 1.24 (SD = 1.86) with a minimum of 0.12 and a maximum of 14.88. One item was removed (WM_drum) due to *a* = 14.88 exceeding conventional thresholds for local fit. Marginal reliability was .87.

Differential Item Functioning (DIF)

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.

Score Validity

Correlations and Predictive Validity

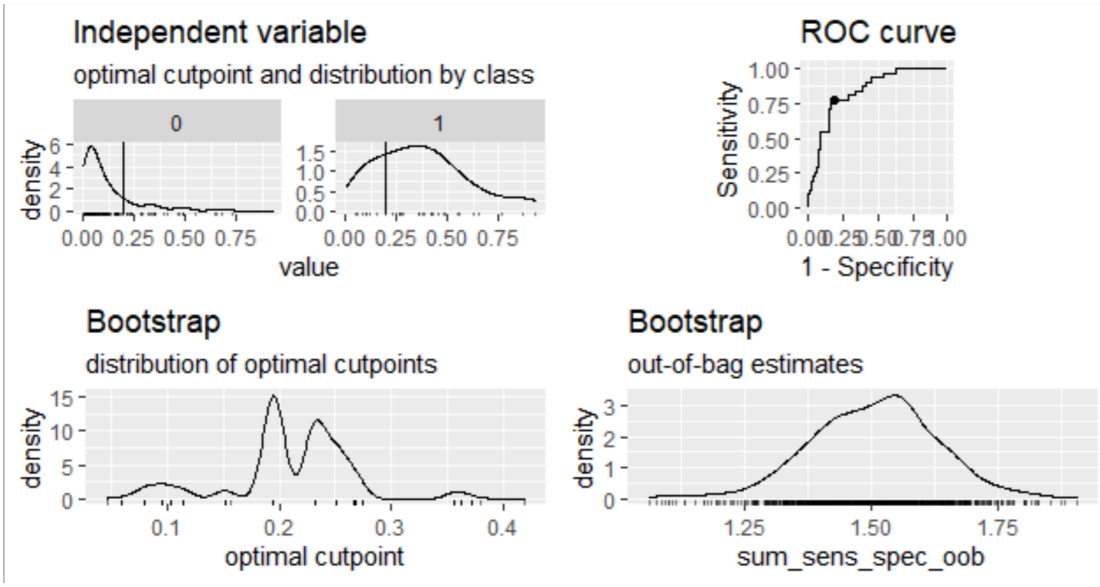
Correlations between and among EarlyBird ability scores with standardized outcomes ranged from -.41 between RAN and FSM to .92 between TOWRE SWE and K-LWR (Table 4). A series of multiple regression analyses tested for the additive and interactive relations between EarlyBird assessments and the K-PA outcome 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 R^2 of .37 (multiple r = .61, 95% CI = .50, .69, n = 184).

Classification Accuracy

Using the predictive validity model, receiver operating characteristic curve analysis (Figure 3) 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 5) 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 3: ROC curve for prediction of severe phonological awareness risk



Chapter 7: Technical Documentation Part III - PWR Risk Screener

Description of Calibration Sample

Data collection for the PWR Risk Screener began by testing item pools for the Screen tasks (i.e., Letter Sounds, Phonological Awareness, Word Reading, Vocabulary Pairs, and Following Directions). A statewide representative sample of students that roughly reflected Florida's demographic diversity and academic ability (N ~ 2,400) was collected on students in Kindergarten 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 6-7 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 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 PWR Risk 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 (1983) to place the items on a common scale. All item analyses were conducted using Mplus software (Muthen & Muthen, 2008) with a 2pl independent items model.

Norming Studies

Data was collected on approximate 2,000 kindergarten students across multiple districts in Florida who participated in the calibration and validation studies. Table 8 provides a breakdown of the sample sizes used by grade level for each of the PWR adaptive tasks.

Reliability

Marginal reliability for the computer-adaptive tasks (Table 9) 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 & Berstein, 1994).

Validity

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 PWR tasks and the SESAT Word Reading task ranged from .38 to .59 (Table 10). Second, a multiple regression was run to estimate the total amount of variance that the linear combination of the predictors explained in SESAT (46%).

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 PWR 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 PWR 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 (i.e., PWR probability) to determine which cut-point maximizes predictive power.

Classification accuracy for the fall/winter screener (Table 11) 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 11) 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.

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 of the PWR Screening Tasks scores is that moderate correlations would be observed. Correlation results are reported in Tables 13. 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).

Tables

Table 1

BCH sample characteristics Part I

	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

Demographic	Category	Sample 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

Item response theory model fit

Task	M2	df	*p*	RMSEA	LB	UB	TLI	CFI
FSM	911.91	665	<.001	0.031	0.026	0.036	0.98	0.98
NWR	996.47	740	<.001	0.030	0.025	0.035	0.98	0.98
RHY	925.36	665	<.001	0.032	0.027	0.067	0.98	0.98
WM	1958.6	1710	<.001	0.019	0.016	0.024	0.97	0.97

Note. FSM = first sound matching, NWR = nonword repetition, RHY = rhyming, WM = Word Matching, 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.

Table 4

Means, standard deviations, and correlations for validation subsample

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. BL (N=210)	-2.23	2.38																				
2. DEL (N=101)	-1.14	1.41	**.44**																			
3. FD (N=97)	-1.09	1.25	.17	.17																		
4. LN (N=94)	3.07	2.87	.14	**.26**	**.27**																	
5. LS (N=213)	-0.08	1.60	**.39**	**.31**	.18	**.23**																
6. SS (N=97)	-0.32	0.94	.17	.17	**.26**	.18	**.25**															
7. VP (N=213)	0.33	1.64	**.20**	**.23**	.09	**.22**	**.22**	**.34**														
8. FSM (N=191)	0.18	0.99	**.56**	**.37**	.15	**.26**	**.51**	.16	**.25**													
9. NWR (N=200)	0.12	0.85	**.48**	**.44**	.16	**.27**	**.37**	.16	**.24**	**.51**												
10. RHYM (N=195)	0.14	0.90	**.26**	**.23**	**.34**	**.47**	**.25**	.20	**.38**	**.39**	**.43**											
11. WM (N=204)	0.09	0.89	**.30**	**.32**	**.34**	**.40**	**.30**	**.48**	**.34**	**.46**	**.51**	**.39**										
12. RAN (N=175)	84.27	27.86	**-.30**	-.09	-.09	**-.24**	**-.28**	-.16	**-.22**	**-.41**	**-.23**	**-.23**	**-.25**									
13. K-LWR (N=139)	96.78	13.50	**.27**	.14	**.33**	.19	**.43**	**.28**	.13	**.45**	**.25**	**.34**	**.22**	**-.26**								
14. K-NWD (N=111)	92.45	13.01	**.23**	.12	.21	**.28**	**.31**	.26	.15	**.46**	**.27**	**.42**	**.24**	**-.33**	**.86**							
15. K-P (N=215)	100.02	14.85	**.39**	**.46**	**.33**	**.30**	**.35**	.17	**.27**	**.51**	**.46**	**.53**	**.36**	**-.20**	**.63**	**.72**						
16. K-D (N=103)	93.17	14.74	**.31**	.15	**.39**	.23	**.31**	.24	.12	**.50**	**.24**	**.43**	**.23**	**-.40**	**.84**	**.81**	**.84**					
17. WID (N=192)	97.16	16.29	**.22**	.05	**.34**	**.24**	**.34**	.17	**.18**	**.48**	**.23**	**.41**	**.24**	**-.26**	**.91**	**.77**	**.62**	**.84**				
18. WA (N=192)	99.31	14.45	**.30**	.03	**.29**	**.28**	**.27**	.15	**.23**	**.49**	**.34**	**.47**	**.32**	**-.24**	**.79**	**.75**	**.71**	**.80**	**.81**			
19. SWE (N=108)	94.03	16.44	**.31**	.23	**.44**	.19	**.40**	.20	**.22**	**.52**	**.30**	**.44**	**.26**	**-.27**	**.92**	**.79**	**.71**	**.88**	**.91**	**.77**		
20. PDE (N=108)	94.95	13.62	**.27**	.18	**.39**	.14	**.34**	.05	.18	**.54**	**.30**	**.44**	**.30**	**-.3**	**.8**	**.82**	**.7**	**.8**	**.7**	**.7**	**.8**	

21. CELF (N=219)	103.26	12.17	.12	.08	**.43**	.37	.25	.38	.28	.27	**.44**	**.43**	**.45**	**-.02**	**.21**	**.34**	**.46**	**.31**	**.21**	**.31**	**.29**	**.36**

Note. *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. Bold values indicates $p < .05$.

Table 5

2x2 confusion matrix of Dyslexia Risk classification

		<16th %ile	>=16th %ile	
		1	0	
At Risk	1	24	29	53
Not At Risk	0	7	124	131
		31	153	184

Table 6

PWR 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	18.3
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 7

PWR U.S. population-based weight values

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	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

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 8

Sample sizes for PWR tasks

Grade	PA	LN/LS	SC	WM	FD	WR
K	2,100	2,377	2,275	2,015	2,304	1,969

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 9

Marginal reliability coefficients for PWR tasks

Grade	Task	Reliability (95% CI)
K	Phonological Awareness Blending	.99 (.98, .99)
	Phonological Awareness Deletion	.94 (.93, .95)
	Letter Sounds	.97 (.96, .97)
	Letter Names	.85 (.83, .87)
	Word Match	.87 (.84, .89)
	Following Directions	.94 (.93, .94)
	Word Reading	.98 (.97, .99)
	Sentence Comprehension	.89 (.88, .90)

Note. CI = confidence interval

Table 10

Bivariate correlations between PWR Tasks and SESAT

Grade	PA-D	LS	WM	FD	WR	Total R^2
K	.59	.51	.38	.46	.48*	.46

Note. Correlations and multiple regression are a function of PWR Screening 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.

Table 11

Classification Accuracy of the Potential for Word Reading Success (PWR) in K

Time Point	AUC (CI)	SE	SP	PPP	NPP	OCC	Base Rate
Fall/Winter	.84 (.81, .88)	.81	.72	.67	.84	.76	41.8
Spring	.88 (.86, .90)	.81	.87	.83	.86	.85	41.8

Note. 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.

Table 12

Bivariate Associations among PWR 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

Note. 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.

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.

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.

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.

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.

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 in the window when intervention is most effective — before they formally learn to read.

For information, visit www.earlybirdeducation.com