EarlyBird

EarlyBird PreK
Early Literacy Assessment

Technical Manual

February 2025

Table of Contents

Please note that this technical manual provides information about the PreKindergarten assessment system. For data specific to the EarlyBird assessments designed for other grades (K, Grade 1 & Grade 2), please see Chapters 5 and 6 in those respective technical manuals.

Acknowledgements

This Technical Manual for the EarlyBird PreK Early Literacy 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.

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 PreK assessment. 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 PreK Early Literacy Assessment, the alphabetic principle is assessed with individually-administered tasks that measure letter-sound knowledge and phonological awareness.

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 PreK Assessment, knowledge of word meanings is measured by a receptive vocabulary task called Receptive Vocabulary. During the Receptive 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 PreK Assessment, this skill is assessed through the Oral Sentence Comprehension subtest. The Oral Sentence Comprehension task 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. (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 App

The EarlyBird gamified mobile app is easy, quick, accessible, and child-centered and designed to be completed prior to formal reading instruction. It is self-administered with teacher oversight. Depending on the subtests administered, the assessment takes 15-30 minutes per child. The assessment addresses literacy milestones that have been found to be predictive of subsequent reading success in kindergarten aged children. No trained adult administration is needed to administer the EarlyBird app. Scoring is largely automated. The EarlyBird PreK Screener provides end-of-year screening for Reading Readiness (i.e. the likelihood that an end-of-year PreK student will have the foundational skills to learn how to read). The full assessment incorporates subtests that were validated across three different validation studies. Appendices A-C 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); 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* - w*hich one starts with the same sound as dog?"* The student touches the picture of the *doll* to identify the correct matching first sound.

Blending – Kangaroo: Blending 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/*").

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. 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.

Fluency

Rapid Automatized Naming - Polar bear: The Rapid Automatized Naming task for PreK uses a set of five objects *(house, door, cat, ear, bed)* that are repeated in random order in four rows, totaling 40 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 40 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.

Vocabulary

Receptive 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 best represents the word.

Comprehension

Oral Sentence Comprehension (OSC) - 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 bird flies towards the nest").

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.

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

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

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

Subtest Score Percentiles
Students' performance is displayed in the form of normed percentiles. Normed percentiles are created based on 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. The distribution of scores were created from a representative 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 would have obtained a score better than about 60% of the students in the

representative 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

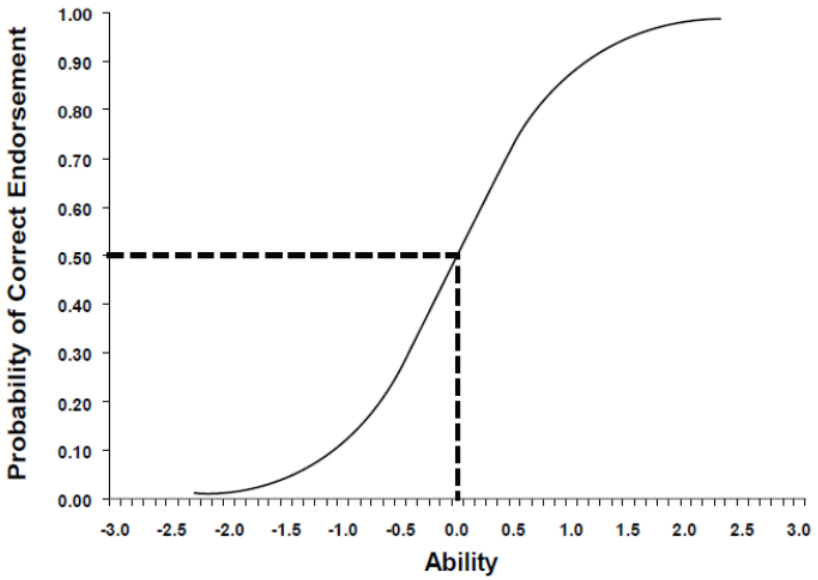
The difficulty of an item (*b*) has traditionally been described for many tests as a "p-value", which corresponds to the percent of respondents correctly answering an item. Values from this perspective range from 0% to 100% with high values indicating easier items and low values indicating hard items. Item difficulty in an IRT model does not represent proportion correct, but is rather represented as estimates along a continuum of approximately -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.

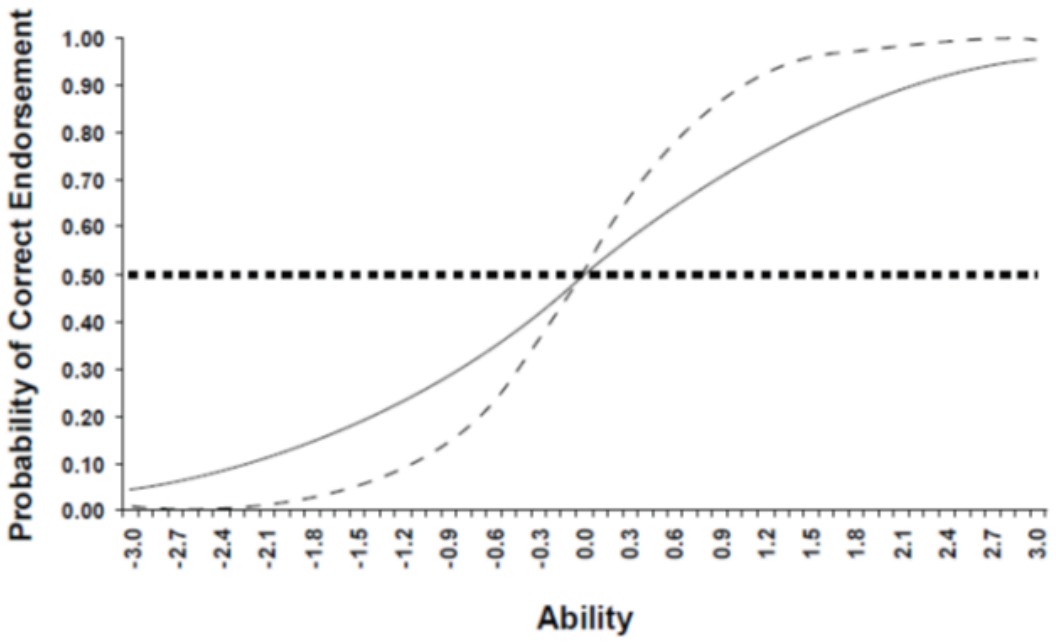
Figure 1: Sample Item Characteristic Curve



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

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 analyses. First, a correlation analysis was used to evaluate the strength of relations between and among each the EarlyBird Assessments and norm-referenced tests. Second, logistic regressions were composed to predict the binarized outcome variable of interest to provide information for selecting an optimal predictive algorithm. Statistically significant relations to the outcome were flagged for examination. Criteria for selecting the optimal algorithm included selecting the smallest (most parsimonious) combination of tests that best predicted the outcome, were statistically significantly related to the outcome, and were theoretically plausible. Finally, OLS regression analysis was composed to estimate the variance that the linear combination of selected predictors explained in the continuous variable criterion. The total amount of variance, squared variance (variance accounted for), and 95% confidence intervals were computed and reported. The results obtained via the predictive validity analyses were used to finalize a selection of the best algorithm for predicting Reading Readiness.

Classification Accuracy

Based on the combination of tests selected previously, logistic regressions were used, in part, to calibrate classification accuracy. Students' performance on the selected criterions were coded as '1' for performance at or above the 40th percentile on the CTOPP-2 Phonological Awareness Composite Score (Reading Readiness) and '0' for scores that did not meet these criteria. In this way, Reading Readiness represents a prediction of success. Each dichotomous variable was then regressed on a combination of EarlyBird Assessments that were identified via the predictive validity analyses. As such, students could be identified as not at-risk on the multifactorial combination of screening tasks via the joint probability and demonstrating adequate performance on the criterion (i.e., specificity or true-negatives), at-risk on the combination of screening task scores via the joint probability and not demonstrating adequate performance on the criterion (i.e., sensitivity or true-positives), not at-risk based on the combination of screening task scores but at-risk on a criterion (i.e., false negative error), or at-risk on the combination of screening task scores but not at-risk on the criterion (i.e., false positive error). Classification of students in these categories allows for the evaluation of cut-points on the combination of screening tasks to determine which were the cut-point maximizing selected indicators. The concept of risk or success can be viewed in many ways, including the concept as a "percent chance" which is a number between 1 and 99, with 1 meaning there is a low chance that a student may develop a problem, and 99 being there is a high chance that the student may develop a problem. When attempting to identify children who are "at-risk" for poor performance on some type of future measure of reading achievement, 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., accuracy, sensitivity, specificity, positive and negative predictive power; see Schatschneider, Petscher & Williams, 2008) and each was considered in light of the other in choosing appropriate cut-points. Area under the curve, sensitivity, and specificity estimates from the final logistic regression model were bootstrapped 1,000 times in order to obtain a 95% confidence interval of scores using the *cutpointr* package in R statistical software.

Technical Documentation

The following sections provide technical documentation of the Reliability (Chapter 5) and Validity (Chapter 6) 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) and EarlyBird, in Appendices A, B, and C, respectively.

The Rhyming, First Sound Matching, Vocabulary, and RAN subtests that appear in the EarlyBird PreK Early Literacy Assessment were validated in part through efforts at Boston Children's Hospital, with additional validation of newly developed, PreK-appropriate items by EarlyBird. The Letter Name, Letter Sound, Blending, and Oral Sentence Comprehension subtests were validated through efforts at Florida State University, with additional analyses using Classical Test Theory (CTT) and Item Response Theory (IRT) to calibrate items for PreK as part of the EarlyBird study.

Chapter 5: Reliability

Summary of Marginal Reliability of Validation Data
(EarlyBird PreK Validation Study, 2022; n = 274 PreK students)

Subtest	Marginal Reliability
First Sound Matching	.78
Rhyming	.79
Vocabulary	.57
Letter Name	.65
Letter Sound	.82
Blending	.78
Oral Sentence Comprehension	.79
RAN*	-

*RAN is a time-limited task and does not have a marginal reliability estimate.
Note: See Appendix C for full study details

Summary of Empirical Reliability of PreK Computer Adaptive Subtests
(EarlyBird Customer Data, August 2022 - June 2023; n = 3,399 PreK students)

Subtest	BOY	MOY	EOY
Rhyming	0.78	0.80	0.81
First Sound Match	0.83	0.79	0.79
Blending	-	-	0.85
Vocab	0.87	0.88	0.86
Oral Sentence Comprehension	0.82	0.79	0.78
RAN*	-	-	-

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

Summary of Empirical Reliability of PreK Subtests at BOY and MOY
(EarlyBird Customer Data, August 2023 - March 2024; n = 1,849 PreK students)

Subtest	BOY	MOY
Letter Name*	0.99	0.99
Letter Sound*	0.98	0.96
Rhyming	0.91	0.94
First Sound Match**	-	0.95
Vocab	0.87	0.88
Oral Sentence Comprehension	0.86	0.86
RAN***	-	-

* Letter Name and Letter Sound were expressive inventory subtests; all other subtests were computer-adaptive tests

**First Sound Match subtest recommended for administration starting at MOY

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

Summary of Concurrent and Construct Validity *(EarlyBird PreK Validation Study, Winter 2022*)*

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

vals	M	SD	BL	OSC	FS	RHM	VCB	RAN	LN	LS	SC	RLI	Eli	BW	SM	RON	PA	RV
BL	-1.62	1.19																
OSC	-0.49	1.40	.24															
FS	0.14	1.51	.40	.25														
RHM	0.09	1.75	.37	.23	.49													
VCB	0.37	1.87	.33	.17	.38	.36												
RAN*	65.90	22.73	-.19	-.16	-.24	-.15	-.20											
LN	2.79	2.92	.25	.11	.31	.35	.29	-.02										
LS	-0.87	1.76	.26	.26	.33	.24	.13	-.14	.29									
SC	100.20	14.07	.39	.30	.40	.37	.46	-.20	.27	.24								
RLI	99.77	17.38	.41	.38	.51	.38	.33	-.35	.28	.35	.76							
Eli	97.09	12.84	.43	.29	.48	.58	.35	-.15	.35	.27	.59	.62						
BW	95.17	12.80	.50	.25	.45	.52	.27	-.17	.35	.30	.55	.58	.68					
SM	97.05	12.03	.39	.44	.59	.48	.39	-.15	.31	.29	.39	.41	.50	.52				
RON	99.56	15.96	.26	.37	.28	.25	.23	-.40	.25	.28	.29	.56	.30	.35	.27			
PA	96.66	13.67	.50	.36	.58	.63	.38	-.19	.39	.32	.60	.65	.84	.87	.78	.37		
RV	103.46	17.19	.37	.41	.33	.39	.57	-.21	.38	.36	.57	.65	.57	.52	.54	.46	.63	
LID	103.36	13.53	.33	.18	.42	.34	.20	-.18	.42	.35	.29	.44	.54	.41	.47	.49	.54	.45

Note. BL = Blending; OSC = Oral Sentence Comprehension; FS = First Sounds; RHM = Rhyming; VCB = Vocabulary; RAN = Rapid Automatized Naming of Objects; LN = Letter Name; LS = Letter Sound; SC = CELF Preschool-3 Sentence Comprehension; RLI = CELF Preschool-3 Receptive Language Index; Eli = CTOPP-2 Elison; BW = CTOPP-2 Blending Words; SM = CTOPP-2 Sound Matching; RON = CTOPP-2 Rapid Object Naming; PA = CTOPP-2 Phonological Awareness; RV = PPVT-5 Receptive Vocabulary; LID = WRMT-III Letter ID.

*EarlyBird PreK Object RAN data (RAN) is from spring of 2022; All other data, including the CTOPP-2 Rapid Object Naming (RON) results, are from winter of 2022.

Note. See Appendix C for full EarlyBird study details

Summary of Predictive Validity and Classification Accuracy
(EarlyBird PreK Validation Study, 2022)

Spring Classification Accuracy for Reading Readiness
(40th percentile – CTOPP-2 Phonological Awareness Composite)
Mean Bootstrapped Area Under the Curve = .834 (95% confidence interval = .76, .91)
Mean Bootstrapped Sensitivity = .77
Mean Bootstrapped Specificity = .78
True Positive N = 92, True Negative N = 62, False Positive N = 18, False Negative N = 27.
Base rate = .60

Fall Predictive Validity Coefficient for Reading Readiness Based on Best Spring Algorithm
(CTOPP-2 Phonological Awareness Composite)
Multiple r = .63, 95% CI = .54, .69, n = 199

Note. See Appendix C for full study details

Differential Item Functioning (DIF)
Summary of DIF analysis (BCH Study, 2019-2020)
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. For full study details, see Appendix A.

Summary of DIF analysis (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 study details, see Appendix B.

Summary of DIF analysis (EarlyBird Study, 2022)
Across FS, Vocab, and Rhyming, eight items demonstrated DIF with at least a moderate effect size (i.e., deltaMH >= 1.0) on either race, gender, or both. A total of seven items (four vocab, two FS and one Rhyming) differed on race while four items (two rhyming, one FS, and one Vocab) differed on gender. Three of the eight items differed on both race and gender. These items were removed from the item bank for further study and testing. For full study details, see Appendix C.

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 years.

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 kindergarten students (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.

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.

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.

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.

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.

Differential Item Functioning (DIF)

DIF testing for subtests that were developed and validated by BCH 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 four 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.

Score Validity

Correlations

Correlations between and among EarlyBird Kindergarten ability scores with standardized outcomes ranged from -.41 between RAN and FSM to .92 between TOWRE SWE and K-LWR (Table 4).

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

Description of Calibration Sample

Data collection for the Kindergarten version of FSU's PWR Risk Screener began by testing item pools for the Screen tasks (i.e., Letter Sound, Phonological Awareness, Word Reading, Vocabulary Pairs - later renamed 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 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 5-7 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 Kindergarten 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

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 7 provides a breakdown of the sample sizes used for each of the PWR adaptive tasks.

Differential Item Functioning

DIF testing for subtests that were developed and validated by FSU was conducted comparing: Black-White students, Latino-White students, Black-Latino students, students eligible for Free or Reduced Priced Lunch (FRL) with students not receiving FRL, and English Language Learner to non-English Language Learner students. DIF testing in the 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.

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 kindergarten 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).

Appendix C: Technical Documentation of EarlyBird Validation Study

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

Procedures

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

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

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

The third and final phase of the PreK Validation Study took place between August and October 2022. 181 children, then in Kindergarten, were administered a predictive validity battery of paper/pencil tests, including the CTOPP-2, KTEA-3, TOPEL, and CELF Preschool-3. The remainder of the participants in the original sample were unable to be tested in fall 2022 due to changes in their geographic location and/or school enrollment.

Upon analyzing the data from the study, some items and subtests were removed from the PreK version of the EarlyBird assessment due to floor effects, DIF analysis results, and observations made during testing that indicated some tasks were not developmentally appropriate for the majority of PreK students. The remaining items and subtests are therefore highly vetted. Having two data points from 181 participants, from the app taken in the spring and from the psychometric assessments administered the following fall, allowed for the evaluation of the PreK screener's predictive validity.

Psychometric Results

Item Response Analytic Framework

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. Additionally, vertical scaling was performed for the Nonword Repetition subtest by fixing item slope and intercept parameters to link with Kindergarten item performance[3] . 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.

Item Level Results

For the First Sounds Matching test, item calibration was performed for 23 items with four of the 23 items set to be equivalent to linked K items for vertical scaling. The model fit indices were M_2 = 183.30, p = .99, RMSEA = .00, TLI = 1.05. The discrimination parameters (a) had an average of 1.79 and ranged from .61 to 3.8. The difficulty parameters (b) had an average of -.56 and ranged from -2.09 to .39.

For the Rhyming test, item calibration was performed for 24 items with four of the 24 items set to be equivalent to linked K items for vertical scaling. Model fit indices were M_2 = 275.22, p = .23, RMSEA = .02, TLI = .99. The discrimination parameters (a) had an average of 1.96 and ranged from .91 to 3.63. The difficulty parameters (b) had an average of -.62 and ranged from -1.33 to .56.

For the Vocabulary test, item calibration was performed for 24 items with four of the 24 items set to be equivalent to linked K items for vertical scaling. Model fit indices were M_2 = 161.93, p = 1.00, RMSEA = .00, TLI = 1.03. The discrimination parameters (a) had an average of 1.81 and ranged from .45 to 6.08. The difficulty parameters (b) had an average of -2.09 and ranged from -3.88 to .30.

Reliability

Marginal reliability is conceptualized as the level of precision for the measurement across the trait continuum. Values of .80 are typically viewed as acceptable for research purposes while estimates at .90 or greater are acceptable for clinical decision making (Nunnally & Berstein, 1994).

Marginal reliability for the fixed form reading tests, First Sound Matching (r_{xx} = .78, 95% CI = [.76, .79]), and Rhyming (r_{xx} = .79, 95% CI = [.77, .80]), approached an acceptable level of reliability while Vocabulary (r_{xx} = .57, 95% CI = [.52, .62]) did not. Marginal reliability for the adaptive subtests, Oral Sentence Comprehension, (r_{xx} = .79, 95% CI = [.76, .82]) and Letter Sound, (r_{xx} = .82, 95% CI = [.78, .85]), indicated acceptable levels of reliability while Blending, (r_{xx} = .78, 95% CI = [.74, .83]) approached an acceptable level, and Letter Name, (r_{xx} = .65, 95% CI = [.58, .71]) did not.

Score Validity

Correlations and Predictive Validity

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

Correlations between and among EarlyBird ability scores with standardized outcomes ranged from .08 between Deletion and Sound Matching to .87 between Phonological Awareness and Blending Words (Table 12). Among the EarlyBird ability scores, the correlations ranged from -.40 between RAN and RON (Rapid Object Naming) to .71 between First Direction subtest and CELF-3 Following Directions.

Logistic regression models were composed for the Reading Readiness Algorithm to identify the best model for predicting Phonological Awareness (Table 13). For the Reading Readiness model, First Sounds and Vocabulary were statistically significantly related to Phonological Awareness.

Multiple regression models were composed including the selected EarlyBird assessments that emerged from the logistic regression (Table 14). For Reading Readiness, the model for predicting Phonological Awareness included First Sounds and Vocabulary and resulted in an R^2 of .40 (multiple r = .63, 95% CI = .54, .69, n = 199).

Classification accuracy

Using the predictive validity model including the best model selected via the logistic regressions, AUC was estimated for the overall efficiency of discrimination of the log-odds from the multiple predictors for the Reading Readiness model (AUC = .83, 95% CI = .76, .91) with an optimal cut-point identified at .57 (Table 15). A total of 92 students (46.2%) were classified as true positives, 62 students (31.2%) were classified as true negatives, 18 students (9.0%) were classified as false positives, and 27 students (13.5%) were classified as false negatives. Mean bootstrapped sensitivity (.77, 95% CI = .63, .90) and specificity (.78, 95% CI = .60, .91) were estimated along with matrix-based negative predictive power (.70), positive predictive power (.84), and overall correct classification (.77) were computed. The base rate in the sample was 59.8%.

Differential Item Functioning

For the new PreK First Sound Matching, Vocab, and Rhyming subtest items, DIF testing was conducted comparing: Black-White students and comparing: Male-Female students. Values for DIF were computed via the difR package using the Mantel-Haenszel method (1959) for detecting uniform DIF. For each of the FS, Vocab, and Rhyming tasks, DIF was tested for two primary contrasts: 1) Male vs. female, and 2) White vs. Non-White. The Mantel-Haenszel chi-square statistic was reported for test by item and the chi-square was used to derive an effect size estimate (i.e., ETS delta scale; Holland & Thayer, 1988). Effect size values <= 1.0 are considered small, 1.0 – 1.5 is moderate, and >= 1.5 is considered large.

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

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)

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 (BCH Study, 2019-2020)

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 (BCH Study, 2019-2020)

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, 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

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

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 6

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 Name & Letter Sound	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 7

Sample sizes by task (FSU Study, 2014)

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 name and sound, SC = sentence comprehension, WM = word matching, FD = following directions, WR = word reading.

Table 8

Marginal reliability coefficients (FSU Study, 2014)

Grade	Task	Reliability (95% CI)
K	Phonological Awareness Blending	.99 (.98, .99)
	Phonological Awareness Deletion	.94 (.93, .95)
	Letter Sound	.97 (.96, .97)
	Letter Name	.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 9

Bivariate Associations among Computer-Adaptive Tasks in Kindergarten *(FSU Study, 2014)*

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.

Table 10

PreK Item Calibration - Students from Each School Administered Forms A-D (EarlyBird PreK Validation Study, 2022)

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

Table 11

Numbers of Items and Students in the PreK IRT Calibrations (Forms A-D) (EarlyBird PreK Validation Study, 2022)

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

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

Table 12

Descriptive Statistics and Correlation Table for EarlyBird Subtests and Standardized Assessments of Language and Literacy (N = 150)
(EarlyBird PreK Validation Study, 2022)

vals	M	SD	BL	OSC	FS	RHM	VCB	RAN	LN	LS	SC	RLI	Eli	BW	SM	RON	PA	RV
BL	-1.62	1.19																
OSC	-0.49	1.40	.24															
FS	0.14	1.51	.40	.25														
RHM	0.09	1.75	.37	.23	.49													
VCB	0.37	1.87	.33	.17	.38	.36												
RAN	65.90	22.73	-.19	-.16	-.24	-.15	-.20											
LN	2.79	2.92	.25	.11	.31	.35	.29	-.02										
LS	-0.87	1.76	.26	.26	.33	.24	.13	-.14	.29									
SC	100.20	14.07	.39	.30	.40	.37	.46	-.20	.27	.24								
RLI	99.77	17.38	.41	.38	.51	.38	.33	-.35	.28	.35	.76							
Eli	97.09	12.84	.43	.29	.48	.58	.35	-.15	.35	.27	.59	.62						
BW	95.17	12.80	.50	.25	.45	.52	.27	-.17	.35	.30	.55	.58	.68					
SM	97.05	12.03	.39	.44	.59	.48	.39	-.15	.31	.29	.39	.41	.50	.52				
RON	99.56	15.96	.26	.37	.28	.25	.23	-.40	.25	.28	.29	.56	.30	.35	.27			
PA	96.66	13.67	.50	.36	.58	.63	.38	-.19	.39	.32	.60	.65	.84	.87	.78	.37		
RV	103.46	17.19	.37	.41	.33	.39	.57	-.21	.38	.36	.57	.65	.57	.52	.54	.46	.63	
LID	103.36	13.53	.33	.18	.42	.34	.20	-.18	.42	.35	.29	.44	.54	.41	.47	.49	.54	.45

Note. BL = Blending; FS = First Sounds; RHM = Rhyming; VCB = Vocabulary; RAN = Rapid Automatized Naming of Objects; LN = Letter Names; LS = Letter Sounds; SC = CELF-3 Sentence Comprehension; RLI = Receptive Language Index; Eli = CTOPP Elision; BW = CTOPP Blending Words; SM = CTOPP Sound Matching; RON = CTOPP Rapid Object Naming; PA = CTOPP Phonological Awareness; RV = PPVT Receptive Vocabulary; LID = WRMT Letter ID.

Table 13
Logistic Regression Results for Reading Readiness Model (EarlyBird PreK Validation Study, 2022)

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

Table 14
OLS Regression Results for Reading Readiness Model (EarlyBird PreK Validation Study, 2022)

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

Table 15

Classification Quality for Reading Readiness Model Predicting Phonological Awareness (EarlyBird PreK Validation Study, 2022)

tp	fp	tn	fn	Accuracy	AUC	Sensitivity	Specificity	PPV	NPV	Cutpoint
92	18	62	27	.77	.83	.77	.78	.84	.70	.56

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

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.

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.

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