Hammill Institute on Disabilities

Learning Disability Quarterly
2021, Vol. 44(3) 145–157

DOI: 10.1177/0731948720931870
journals.sagepub.com/home/ldq
SAGE

Early Detection of Dyslexia Risk: Development of Brief, Teacher-Administered Screens

Jack M. Fletcher, PhD[1], David J. Francis, PhD[1], Barbara R. Foorman, PhD[2], and Christopher Schatschneider, PhD[2]

Abstract

Many states now mandate early screening for dyslexia, but vary in how they address these mandates. There is confusion about the nature of screening versus diagnostic assessments, risk versus diagnosis, concurrent versus predictive validity, and inattention to indices of classification accuracy as the basis for determining risk. To help define what constitutes a screening assessment, we summarize efforts to develop short (3–5 min), teacher-administered screens that used multivariate strategies for variable selection, item response theory to select items that are most discriminating at a threshold for predicting risk, and statistical decision theory. These methods optimize prediction and lower the burden on teachers by reducing the number of items needed to evaluate risk. A specific goal of these efforts was to minimize decision errors that would result in the failure to identify a child as at risk of dyslexia/reading problems (false negatives) despite the inevitable increase in identifications of children who eventually perform in the typical range (false positives). Five screens, developed for different periods during kindergarten, Grade 1, and Grade 2, predicted outcomes measured later in the same school year (Grade 2) or in the subsequent year (Grade 1). The results of this approach to development are applicable to other screening methods, especially those that attempt to predict those children at risk of dyslexia prior to the onset of reading instruction. Without reliable and valid early *predictive* screening measures that reduce the burden on teachers, early intervention and prevention of dyslexia and related reading problems will be difficult.

Keywords

dyslexia, early screening, TPRI, early intervention

The key to intervening with children at risk of dyslexia and early reading problems is early intervention (Fletcher et al., 2019). When risk is identified in kindergarten (KG), Grade 1 (G1), and Grade 2 (G2), many studies report reductions of risk from 20% to below 5% of children depending on the quality and intensity of the instruction (Mathes et al., 2005; Torgesen, 2000). Other studies report that when identification and intervention begin in Grade 3, more time in intervention is required to accelerate gains compared with both core instruction (Connor et al., 2013) and supplemental intervention (Lovett et al., 2017) in G1 and G2. Outcomes for students identified in adolescence are much poorer (Vaughn et al., 2010). Crucial to implementing early identification is accurate identification of risk. This article addresses the process of early screening and methods for the development and evaluation of screeners using The Primary Reading Inventory as an example.

Early Screening

Because of the importance of early intervention, more than 40 states mandate screening for dyslexia and/or reading problems in KG, G1, and sometimes G2 and G3 (Petscher et al., 2019). For screening, the focus should be on identification of risk, with a goal of an accurate binary decision of at-risk, not at risk. This risk is for reading problems in general as children who show early reading problems typically have word-level difficulties (Leach et al., 2003). It is difficult to separate children identified with dyslexia from other children with word-level problems, calling some to question the utility of the dyslexia label (Elliott & Grigorenko, 2014). There is little evidence that different interventions are needed to remediate reading difficulties for children identified with dyslexia or any word-level problem (Miciak & Fletcher, 2020).

[1]University of Houston, Houston, USA
[2]Florida State University, Tallahassee, USA

Corresponding Author:
Jack M. Fletcher, University of Houston, 3695 Cullen Boulevard, Room 126, Houston, TX 77204-5022, USA.
Email: JackFletcher@uh.edu

Problems with legislated early screening will emerge because the legislation too often confuses screening and diagnosis. Screening should be defined as rapid triage of entire classrooms to identify risk, which corresponds to universal screening in a multitiered system of support (MTSS). Diagnosis requires more extensive assessment that can be costly and time consuming, completed by assessment professionals, not teachers. An effective screening program can reduce the burden of diagnostic assessment by allocating more costly diagnostic assessment to those students identified as at-risk through screening, with diagnostic assessment delayed until after a period of core instruction or intervention with progress monitoring.

To mitigate risk and the development of severe reading difficulties that can result from delaying early intervention, teachers who have immediate access to the child should screen for risk with a short probe that minimizes demands on teachers' time. Consistent with MTSS practices, children who are at risk should immediately enter a progress-monitoring system to determine how well they respond to core instruction in the classroom. Alternatively, subsequent to screening, a reading inventory could be administered to assess the at-risk student's development of reading-related skills and determine those skills that need instruction. A diagnostic assessment should not be provided until the child has received adequate core instruction and intervention early in schooling to document progress (most reliably after G1, where the strongest data for intervention efficacy resides; Fletcher et al., 2019). Many children will respond to early intervention and static diagnostic assessments are primarily useful for inadequate responders. Thus, the concepts of risk and prevention are vital. Not all children who show evidence of risk will manifest reading problems if they are provided early intervention, explicit core instruction, and supplemental intervention when the response to core instruction and early intervention are insufficient to mitigate risk (Miciak & Fletcher, 2020).

Predictive Versus Concurrent Validity

The most common practices for early screening are to use universal screeners, or a set of tests that assess domains in which dyslexia may be manifested (e.g., phonological awareness [PA], rapid naming), and subdivide the distribution of scores to indicate risk. Such approaches do not explicitly measure how well the instrument *predicts* subsequent risk, which is especially true for dyslexia screening in KG and G1. Children change rapidly during the first few years of schooling. Validity information that is concurrent with screening, that is, addresses only the child's status at the time of the assessment, does not account for these changes and provides only indirect information about the validity of the screen for *predicting* the child's status in subsequent grades.

The predictive validity of decisions about risk based on concurrent assessment of the screening assessment is not well studied for decisions for individual students, nor for educational systems as a whole (e.g., as evidenced by the distribution of outcomes for all students in a school, district, or state). For example, where to set thresholds for decisions based on concurrent assessments to optimize prediction of future outcomes is not adequately explored in the screening literature. Universal screeners, which are brief probes of skills such as letter–sound naming, and word or passage reading administered at baseline in a progress monitoring system, have particular promise as early screeners. Such measures have demonstrable predictive validity (e.g., Gotffreda et al., 2009), but many studies are based on small samples and require more attention to the selection of cut points to optimize decision accuracy (VanDerHeyden, 2013).

Predictive Screens: Some Examples

The technology for early screening emerged more than 40 years ago with the development of assessments designed to *predict* which KG children will develop dyslexia (reviewed in Benton & Pearl, 1978). For example, the Florida Longitudinal Project (Satz et al., 1978) followed several hundred KG children up to Grade 6 to evaluate KG precursors that would predict reading disability. This project resulted in a 20-min KG assessment (Satz & Fletcher, 1982) that included four measures: perceptual-motor skills, perceptual-discrimination skills, vocabulary, and alphabet knowledge, with the latter emerging as the best single predictor.

Satz et al. (1978) also introduced statistical decision theory, including base rates, false-positive and false-negative errors, and conditional probabilities (Meehl & Rosen, 1955) that led to equations for predicting the risk of a reading problem in G2 based on performance on the 20 min assessment (see Table 1). Not surprisingly, *all screens yield decision errors*. Some children identified as "at risk" will not go on to develop a reading problem (false-positive error). Other children will be missed, that is, identified as not "at risk" when they in fact develop reading problems later in development (false-negative error). To illustrate, the Florida Kindergarten Screening Battery accurately identified risk status in 75% of the children screened, with a false-positive (Fp) rate of .27 and a false-negative (Fn) rate of .21.

Such errors are a necessary consequence of imperfect correlation between the criterion and the assessment on which decisions of risk are based. Although their existence is inevitable, the rate of both kinds of errors depends on the placement of the decision thresholds, which are related: as false positives go up, false negatives go down, and vice versa (see Figure 1). Therefore, it is important to determine what types of errors are less desirable and select a threshold that balances the two kinds of decision errors. When the

Table 1. Definition of Test Decisions.

	Identifications			
Outcomes	No risk	At risk	Case totals	Definitions
No problem	Correct d	False positive b	**d + b**	**False positive rate** **b/(d + b)**
Problem	False negative c	Correct a	**c + d**	**False negative rate** **c/(c + a)**
Identification totals	a + c	b + d	**a + b + c + d**	

Note. Sensitivity = a/(c + a) = 1—false-negative rate.
Specificity = d/(d + b) = 1—false-positive rate.

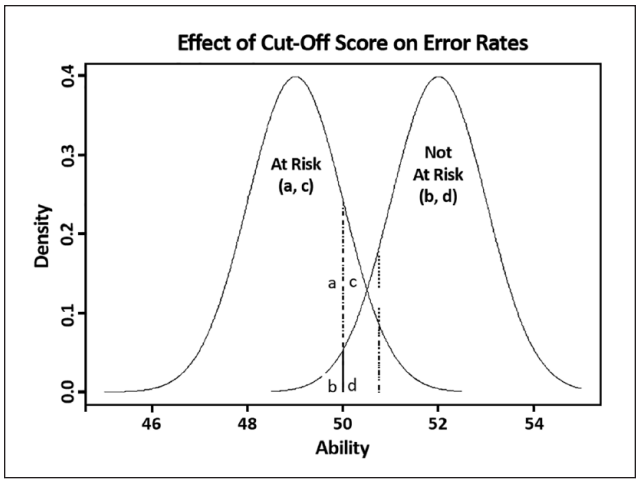

Figure 1. Trade-off of false-positive and false-negative areas around the risk threshold.
Note. By adjusting the threshold, the ratio of false positives and false negatives can be adjusted depending on which error type is more desirable.

costs associated with each type of error can be determined, and the base rate of problems in the population is known, it is possible to determine an optimal decision rule, where optimal is determined to be the rule that balances the consequences of the two kinds of errors. In screening for reading problems, the consequences of a false-negative error are more serious than a false-positive error. If a child who is at risk is not identified (false-negative error), the detection of risk is delayed during a period when reading instruction may be most effective, leading to a lifetime of academic difficulty for the student, with potential negative economic and social consequences. If a child not at risk is identified by a screen as at risk (false-positive error), the student may be subjected to unnecessary intervention, which could also constrain school resources and teachers' capacity. However, follow-up in the form of progress monitoring or a reading inventory could be used to identify false-positive errors and allow reallocation of resources when students are judged on-track. The Florida screen had Fn of .21, which means that many children who went on to have reading problems in G2 were missed by the screening.

There is now considerable evidence that assessing precursors of early reading can predict subsequent achievement (see reviews by Badian, 2000; Scarborough, 1989), but the predictors employed are different from those used by Satz et al. (1978): measures of phonemic awareness, rapid naming of letters and numbers, alphabetic knowledge (e.g., knowledge of letter–sounds), and vocabulary. One predictive effort that incorporated these advances led to the development of the predictive assessment of reading (Wood et al., 2005). This battery was developed from a longitudinal study of 485 first-grade children who received cognitive performance tests and reading skill evaluations in Grades 1, 3, and 8, with 220 children evaluated at all three time points. A second nationally representative sample of 500 children was tested in Grades KG–G3 using a more abbreviated version of the constructs predictive in the first study.

In the first study, averaging scores from tests in four domains created measures of four constructs: PA, rapid naming, vocabulary, and single-word reading. Woods et al. reported strong relations with criterion reading outcomes, with Fp rates around 25% and Fn rates of about 9–14% in the different grades. In the second study, Woods et al. (2005) created shorter measures of the predictive measures. This 15-min assessment also showed strong predictive validity for the reading composite from the beginning to the end of each grade. Fp and Fn rates were similar to those reported for the first predictive study.

Another predictive effort was the KG–G2 component of the Florida Center for Reading Research (FCRR) Reading Assessment (FRA; Foorman et al., 2015). The KG–G2 component consists of computer-adaptive alphabetic (PA, letter sounds, word reading, spelling) and oral language screening tasks (vocabulary, following directions, sentence comprehension) that provide a probability of literacy success linked to grade-level performance on the word reading in KG and reading comprehension subtests of the Stanford Achievement Tests (SESAT; SAT-10) in G1 and G2. Thus, the FRA provides universal screening and diagnostic tasks in a computer-adaptive format, with psychometric characteristics

and normative information derived from a large sample of students' representative of Florida demographics. The cut point selected for the FRA kept Fn ≤ .18, but Fp ranged from .53 in KG to around .30 in G1 and G2 (see Table 6 of Foorman et al., 2015).

The Primary Reading Inventory

In 1997, Texas passed a law mandating early screening for reading problems, including dyslexia, in KG and G1 and G2. The Texas Education Agency contracted with the authors of this article to develop a screening instrument and an inventory. The result was the The Primary Reading Inventory (TPRI), which provided 3- to 5-min screens to identify children at risk of reading problems, including dyslexia, in KG–G2 and a 30 min inventory to determine what reading concepts needed to be taught. The National Institutes of Child Health and Human Development (NICHD)-funded early assessment of reading skills (EARS)[1] project, a longitudinal assessment of reading precursor skills children, was used to develop the screens.

EARS leveraged advances in the statistical modeling of individual growth to improve the identification of children at-risk for reading problems (see Boscardin et al., 2008) and to replicate and/or update current findings regarding the best KG/G1 predictors of word reading, fluency, and reading comprehension skills in G1/G2. Research identifying the cognitive correlates of dyslexia has not changed significantly as the EARS study was conducted (Fletcher et al., 2019). Measures of PA, letter and letter–sound naming, vocabulary, and rapid naming of alphanumeric stimuli remain the most robust predictors and correlates of word-level reading skills depending on the grade level of the child. The other measures of perceptual skills were included because of evidence from older studies indicating that they were also predictive (Satz et al., 1978). The measures were administered multiple times each year because of the focus on the measurement of individual growth (Boscardin et al., 2008). However, the use of a growth trajectory across the entire KG year has limits as a screening device because it precludes KG intervention and requires sufficient time to lapse permit reliable estimation of growth parameters, a function of the number, and spacing of time points.

A dominance analysis (Schatschneider et al., 2004), which determined the value of each given predictor within the context of all possible combinations of predictors, showed that three KG measures consistently predicted word reading, reading fluency, and reading comprehension at the end of G1 and G2: letter–sound knowledge, PA, and rapid naming of letters. Measures of perceptual skills, vocabulary, rapid naming of objects, and language measures in KG involving syntax did not *uniquely* predict reading outcomes in G1and G2. These null results do not mean that such tasks are not correlated or predictive of reading, just that they are less *uniquely* predictive. Letter–name knowledge at the beginning of kindergarten was predictive of G1 and G2 reading performance, but reached a ceiling by end of KG and was no longer predictive. Letter–sound knowledge, PA, and rapid letter naming were closely related to G1 and G2 reading outcomes. For word reading and reading comprehension outcomes in G1 and G2, rapid letter naming, and PA at KG end-of-year (EOY) were comparably predictive, but letter–sound knowledge was also uniquely predictive in some models. For fluency in G1 and G2, rapid letter naming was more predictive than PA measures, reflecting that rapid naming and reading fluency are timed tasks involving print. The relations of predictors and reading outcomes diminished as the time interval increased, but predictive validity was still strong at the longest intervals (beginning of KG to end of G2).

Schatschneider et al. (2004) addressed relations of tests to outcomes, but not the predictive validity of the tests for predicting individual risk for reading problems. The purpose of this article was to show how we used the EARS study to develop the screens for the TPRI. Five challenges guided the development of these screens:

1. *Creating a short, parsimonious screen*. What is the minimal amount of time required to establish an accurate and predictive screen? We proposed screens that took 3–5 min to administer and kept Fn below 10%.
2. *Predicting reading achievement prior to formal reading instruction*. Can KG measures from the assessment battery predict individual G1 and G2 reading outcomes with sufficient validity and classification accuracy to permit accurate early screening of reading problems?
3. *Identifying when reading tasks become maximally predictive*. At some point, the best predictors may not be tasks like PA and rapid naming. Simple reading tests may become most predictive and it would not be necessary to assess cognitive functions.
4. *Modifying tasks and decision rules to enhance classification accuracy*. Can we select items and decision rules to achieve the desired Fn and Fp rates with an even briefer assessment?
5. *Ensuring fairness*. Would there be gender, ethnic, and age bias in the predictions?

Method

Sample

The EARS sample, followed from 1992 to 1996, was obtained from three schools in an ethnically and culturally diverse suburban school district in Texas. The participants were 945 children in KG, G1, and G2 general education classrooms

Table 2. Time of Assessments, Endpoint Times, and Sample Sizes for Development of Screens.

Time point	Outcome	Sample size
1. Middle KG	End of G1	421
2. End of KG	End of G1	421
3. Beginning of G1	End of G1	599
4. End of G1	End of G2	376
5. Beginning of G2	End of G2	540

Note. KG = kindergarten; G1 = grade 1; G2 = grade 2.

randomly selected from the students whose parents returned signed consent forms (approximately 95% of students in the schools). Table 2 presents the sample sizes used to relate each of the five screening and outcome assessments, indicating when the assessments were conducted. Children were excluded because of severe emotional problems, uncorrected vision problems, hearing loss, acquired neurological disorders, or designation as limited in English proficiency. Eligibility for free and reduced-price lunch in the three schools was 13%, 15%, and 30%, respectively. Boys and girls were roughly equally represented in the sample. All major race/ethnicities were represented: White (54%), African American (18%), Latinx (15%), and Asian-American (12%). The sample was representative of the demographics of the three schools.

Measures

A large assessment battery including intellectual, achievement, cognitive, and sociodemographic indices was administered up to four times each year during each of the EARS study years (see Boscardin et al., 2008; Schatschneider et al., 2004). We focused on the measures that demonstrated the best predictive validity (Schatschneider et al., 2004). The following assessments were administered four times per year:

Phonological processing. A prepublication version of the comprehensive test of phonological processes (CTOPP; Wagner et al., 1999) provided seven measures of PA. Total scores were used, created by adding correct responses based on an item response theory (IRT) model for all the phonological processing measures in this version of the CTOPP. Despite the task structure, performance on PA tasks represents a unitary dimension with tasks ordered by the distribution of item difficulty (Schatschneider et al., 1999).

Rapid naming. Denckla and Rudel's (1976) rapid automatized naming test for letters consists of high-frequency lowercase letters (i.e., a, d, o, s, p) repeated 10 times in random sequences on a large paper. The child named each stimulus as quickly as possible. The correct number of responses named in 60 s was converted to number correct per second.

Alphabetic knowledge. Knowledge of letter names and sounds was assessed in KG and G1 BOY by presenting printed cards with both the upper and lowercase letters. The child named the letter and then gave the sound of the letter. Credit was given for no more than one correct sound per letter. After G1 BOY, this task was not administered because of ceiling effects.

Vocabulary. The Peabody Picture Vocabulary Test-Revised (Dunn & Dunn, 1981) is a norm-referenced receptive vocabulary measure. The child hears a stimulus word and is shown four pictures. The child chooses the one picture that depicts the word.

Word reading. To assess growth in word reading, children in G1 and G2 read aloud 50 words presented one at a time on 4 × 6 cards. The words, matched for frequency and consistency (Carroll et al., 1971), represented a diversity of linguistic features, and spanned G1–3 difficulty levels. This word list, based on a count of five million words from school texts for the American Heritage, is still in use. It recently was praised for incorporating range, frequency, and dispersion into the word list (Nation, 2016). Students in each grade read 50 words, 16 of which were common across the two grades, allowing placement of the 84 unique items across the two forms on a common scale using IRT for equating.

EOY G1 and G2 Outcome Measures

Outcome measures were the basic reading composite (G1) and broad reading composite in G2 from the Woodcock Johnson–Revised (WJR; Woodcock & Johnson, 1989) scale. To compute these composites, the following subtests were administered at the end of G1 and G2.

Word reading. Letter–word identification (real words) and word attack (pseudowords) subtests assessed students' sight word knowledge and decoding skills. These measures have internal consistency reliability estimates above .9 and extensive demonstrations of validity.

Reading comprehension. To assess a child's ability to derive meaning from text, we administered the passage comprehension subtest of the WJR ($\alpha = .95$). This test measures silent reading comprehension at the sentence level using a cloze procedure. Children fill in missing words, relying on what they read for context.

Procedures

Five screens were developed that varied depending on when they were administered (October, December, February, May). The sample for each screen represented all children available at a time point who had one of the criterion outcome assessments (see Table 2).

For screens administered in KG and at G1 BOY, we predicted outcomes at the end of G1 using the WJR basic reading composite. For screens administered at the end of G1 and at G2 BOY, we predicted performance on the G2 WJR broad reading composite. We used these composites because they are more reliable than individual measures and have more dispersion of ability. We chose the broad reading cluster in G2 because this outcome would have the potential to identify children at risk of reading comprehension problems as well as decoding problems. All three measures used in forming the composites were highly correlated in G1 and G2 (r = .70–.96).

To define outcomes, a criterion of 0.5 grade equivalents below grade level was used at the end of G1 and G2. This threshold is arbitrary, selected to indicate that children were at risk of reading one half-grade below grade level. For G1, this represented a grade equivalent of 1.4 or lower; for second graders, this represented a grade equivalent of 2.4 or lower. In G1, this score would be at the 22nd percentile for basic reading and 18th percentile for broad reading. In G2, the outcome score corresponded to the 35th percentile. The cut point identifies a higher percentage of students in G2 because of schools' failure to identify problems in G1, which results in increasing numbers of students falling behind. In addition, it reflects the widening of expectations as children get older. In a sense, it is more difficult for a child to be one half grade below expectations in G1 as compared with G2. We chose this type of criterion instead of a set threshold based on a percentage of the population because all such risk criteria are arbitrary. We viewed linking to grade-level expectations as less arbitrary than a percentile-based criterion because one half–grade level defined an amount that could potentially be caught up with intervention, even though one half grade is not a comparable magnitude at different grades.

Statistical Analysis

At each time point, the general approach was to run prediction models using discriminant function analysis of the binary outcomes (reading problem, no problem) that identified variables that contributed uniquely to the prediction of risk status. The first step in each analysis involved an examination of all possible combinations of the five predictors in predicting the criterion (i.e., reading problem, no problem). We examined both the squared canonical correlation, an index of the strength of the relation between the predictor and outcome variable, and the classification matrices (Table 1), resulting from the predicting outcome on a case-by-case basis. The classification matrices were jackknifed using a leaving-one-out method to obtain an unbiased estimate of classification accuracy through cross-validation. Variables selected exhibited both a high-squared canonical correlation and relatively low numbers of both false-positive and false-negative errors. Once predictors were selected, a cut point from the equation expressing the relation of the predictors and outcomes was established, plotting these relations to adjusting the thresholds to establish the lowest possible Fp error rate, whereas keeping Fn errors below 10%. As Figure 1 shows, these rates are tradeoffs and can be adjusted by manipulating the cut point that identifies risk. We then applied IRT to identify the fewest number of items needed to accurately estimate ability at the cut point. This step allowed us to distinguish those cases likely to be above the cut point in ability with the fewest possible items while maintaining Fn $<$ .10, again graphing the results using the abbreviated item sets.

Results

Kindergarten Screens

The KG screens were developed using performance in December and April (Table 2). December was selected because of the high probability that many children with fewer literacy experiences would do poorly on a KG screen administered at the beginning of school because they need time to acclimate to the school environment. We selected the end of KG to help the teacher identify children who would benefit from administration of the inventory to plan learning objectives for the summer and following year.

From this battery of tests described above, we selected the measures of letter names, letter–sounds, PA, rapid letter naming, and vocabulary. Other measures in the battery did not show independent contributions to reading outcomes in Schatschneider et al. (2004). We included vocabulary because it predicted reading comprehension in other studies (Wood et al., 2005).

Two KG measures consistently provided the best predictive discrimination: letter–sound naming and PA. Predictions to G1 BOY also were explored to ensure that excessive instability was not being introduced by the length of time from predictor to criterion. However, results were virtually identical for BOY and EOY G1 predictions, so we dropped the predictions from December of KG to October of G1, just using EOY G1.

Using the same 421 children, from middle-of-year (MOY) KG to EOY G1, Fp = .37 and Fn = .09, which means that 91% of the true positives were detected (also called sensitivity, and equal to 1-Fn) and 63% of the true negatives were detected (called specificity, and equal to 1-Fp). The total accuracy was .69, but this is misleading because we did not try to make the most accurate instrument; rather, we tried to minimize false-negative errors. From end KG to end G2, Fp = .33 and Fn = .11, so that sensitivity = .89 and specificity = .67. Overall accuracy was .72.

The next step involved determining the specific items for both the letter–sound identification and PA measures that would produce identification comparable to that of the

discriminant analyses using the full item set from measures. The goal was to obtain a shorter test that provided optimal discrimination around the cut point of the linear discriminant function, after which we could determine the cutoff scores for decisions based on the limited item sets.

Letter–sound identification. The 26 letters in the English alphabet represent a finite set of letter–sound items that cannot be expanded by creation of new items. The letter–sounds also display a clear ordering with respect to difficulty (Treiman et al., 1998). Not all sounds are equally predictive, so to score above the cut point determined by the discriminant analysis, children would have to be successful on items that are more difficult.

We evaluated a screen that consisted of the most difficult 13 letter–sounds based on error rates, but which yielded comparable discrimination to the 26-item set. We then determined the 10 most difficult items across both waves. The items were similar, but the order varied over time. From easiest to hardest, the 10 most difficult letter–sound items were:

December: N L O E I Q W X U Y April: L O N W E I Q U Y X

The letters Q and X present some difficulty from a linguistic perspective because neither letter has a clearly identifiable sound in isolation in English. The next most difficult items are R and H, which are not appreciably easier than L, O, and N. Thus, substituting R and H for X and Q, resulted in the following list of 10 items:

L O N I R E H W U Y

This list has a reliability (coefficient α) of .90 and a bivariate correlation with end of G1 basic reading of .51 (December) and .54 (April).

Phonological awareness. To evaluate the contribution of the individual PA items to predictions, the total number correct out of the 10 letter–sound items was plotted along with the PA total score from the IRT analysis against reading classifications to determine the most appropriate cut point on the PA screen. The IRT-based score is an estimate of a student's PA ability, or "theta" score (p_theta). These scores are distributed with a mean of 0 and standard deviation of 1. By manipulating the cut points on the PA screen, the best discrimination resulted from adding the following decision rules to the ones established above:

MOY KG: December: letter–sounds = less than 4 of 10 and p_theta of −.80;

EOY KG: letter–sounds = less than 8 of 10 and p_ theta of −.37.

The level of PA and letter–sound knowledge required to pass the screen at the beginning of KG (i.e., to be determined *no*t at-risk) are lower for both skills than at the end of KG. The PA threshold increased by almost 0.5 standard deviations. This change reflects the rapid development of PA over the KG year and is consistent with growth mixture models on the EARS data by Boscardin et al. (2008), where children with low PA performance and flat growth in KG were most at risk of poor development of word reading skills in G1.

The next step involved the assembly of brief lists of items that maximally discriminated around the skill level cut point for PA for each screen (−.80 for MOY and −.37 for EOY) and to determine the number correct score on the reduced item set that optimally identified students as above or below the skill-level cut point. Because the PA item difficulties are on the same scale as the p_theta scores, the selection of items is straightforward. The eight items with difficulty parameters closest to the cutoff point were selected for each time point. The critical factor was the location of the item on the difficulty scale (i.e., item difficulty), not the nature of the task. Items came from three different PA tasks: blending word parts (BWP), blending phonemes into words (BPW), and blending phonemes in nonwords (BPN). The following items were selected in December: BWP: *P-ICK, M-ARK, F-IGHT, CH-IN, TH-ANK, S-AW;* BPW: *R-A-SH, W-I-SH.* In April, items were BWP: *W-ILL*, BPW: *S-OO-N, L-A-S-T, W-I-SH; BPN: V-AW, F-OO, W-OY, H-A-SS.* The items were different at each time point, but all involved blending onsets and rimes, or individual phonemes, rather than segmenting or manipulating phonemes. As can be seen from the list, several items at EOY included nonsense words. Number correct scales formed from these items have reliabilities (coefficient α) of .91 for both December and April. The bivariate correlations with end of G1 basic reading are .50 (December) and .48 (April).

To create decision rules for the screens, the total number of correct of the 10 letter–sound items was plotted against the total number of phonological items correct out of the 8 selected for each wave. Cutoff points on the sum score for the phonological items were manipulated and it was determined that the best discrimination resulted from the following decision rules adjusted to account for the developmental progression in the two time points:

Letter–sounds: decision rule middle KG: Less than 4 out of 10 correct

Decision rule end KG: Less than 8 out of 10 correct

If the child fails the letter–sounds screen, the child is at-risk; if the child passes the letter–sounds screen, PA items are administered and the child may be at risk.

Decision rule middle and end KG: Less than 6 out of 8 correct, child is at-risk.

Table 3. TPRI Screening Results (1998 Original/2010 Updated).

	Criterion time		False-positive rate		False-negative rate	
Screening time	Original	Update	Original	Update	Original	Update
Middle of KG	End G1	End KG	.44	.44	.05	.06
End of KG	End G1	End KG	.38	.39	.10	.09
Beginning of G1	End G1	End G1	.37	.29	.07	.07
End of G1	End G2	End G2	.27	.23	.04	.08
Beginning of G2	End G2	End G2	.15	.23	.09	.11

Note. KG = kindergarten; G1 = grade 1; G2 = grade 2.

Table 3 shows that Fn = .05 (MOY KG to EOY G1; sensitivity = .95) and .10 (EOY KG to EOY G1; sensitivity = .90), whereas the Fp rates declined from .44 to .38 (specificity = .56, and .62, respectively). To examine whether there was differential accuracy of the prediction equations across ethnic groups, a variable representing the identifications produced from the cut points for letter–sounds and PA was constructed. This variable was used with ethnicity to determine whether outcomes varied according to these two factors or their interaction. In April, there was a significant interaction for ethnicity by predicted classification, $F(3, 413) = 2.49$, $p < .06$. Post hoc examination indicated that classification of outcomes was more accurate for Latinx and African American than White students, so that bias due to ethnic differences was not in the direction of concern. The results for the MOY KG screen were similar.

BOY G1 Screen

The G1 and G2 screens followed the same development process as the KG screens and the primary results for Fn and Fp rates are in Table 3. Using the same procedures as in KG to select the best combination of predictors, the BOY G1 screen consisted of three measures used to predict EOY G1 broad reading, adding the word reading task to the set of possible predictors. Specific items can be found at www.tpri.org/index.html:

1. *Letter—sound task—identify the sounds of 10 letters (with a cutoff of 8 out of 10)*
2. *Word reading task -read 10 words (with a cutoff of 8 out of 10)*
3. *PA task—6 phoneme blending tasks (with a cutoff of 5 out of 6)*

The decision to include the letter–sound task was because KG was not mandatory in all states and to maintain continuity with the EOY KG screen. The letter–sound task in the BOY G1 screen was identical to the letter–sound task in the EOY KG screen. Children who do not meet the criterion on the letter–sound screen at BOY G1 should be considered at risk.

The items selected for the word-reading task were based on a discriminant function analysis of 599 children who had both October and May data for G1. The best predictor was the 50-item measure of word reading described above. The score used in the analysis was derived from an IRT model for the 50 items (w_theta). From this analysis, the identification rule was manipulated so that the Fn would fall below 10%. This criterion was met when a cutoff of −1.07 on w_theta was used. In this analysis, Fp =.42; Fn = .05. Overall accuracy was .65.

While this cutoff achieved the desired accuracy rate for Fn, the Fp error rate was high. To reduce Fp, we plotted the data and determined that PA scores reduced the Fp rate. Using a decision rule that identified children as being at risk if they scored below −1.07 on w_theta and below 0.0 on p_theta, we selected eight items from the word reading list and six items from the PA measure with difficulty parameters that were nearest to the cut points for the two measures. Using the selected items, cutoffs were established for each measure and the identification accuracy of the resulting screening instrument and decision rule was reexamined: Fp = .37; Fn = .07; overall accuracy = .70 (Table 3). Reliability (coefficient α) was .90 for the word-reading list and .91 for the PA items. Bivariate correlations with EOY G1 WJR broad reading cluster score were .81 and .67, respectively. The ethnicity by identification interaction was not significant, $F(3, 591) < 1$, $p = .58$.

End of G1 Screen

The EOY G1 screen consisted of two measures predicting EOY G2 broad reading:

1. *Word reading task—read 10 words (with a cutoff of 8/10)*
2. *PA task—6 phoneme blending tasks (with a cutoff of 5 out of 6)*

The items on the word reading list were selected based on a discriminant function analysis of 376 children who had data from both G1 spring and end of G2 (WJR broad reading score). After analyzing a number of different models,

the best model included just the w_theta score. The PA items did not improve predictions. Because the Fn rate was .15, we adjusted the classification rule so that the percentage of Fn errors would fall below 10%. When the cutoff was set at w_theta equal to −0.2, the percentage of Fn errors dropped below 10%. With the rate of Fn errors reduced, we then tried to reduce the Fp rate using the PA score. Using a decision rule that identified children as being at risk if they scored below −0.2 on w_theta and below +0.80 on p_theta, Table 3 shows that Fp = .27; Fn = .04. Eight word-reading items and six PA items with difficulty parameters nearest to these respective cut points were selected.

The word list has reliability of .92 and a bivariate correlation with end of G2 WJR broad reading of .82. The PA scale had an estimated reliability of .92 and a bivariate correlation with end of G2 broad reading of .60. There were no significant interactions involving ethnicity, $F(3, 368) < 1, p = .83$.

BOY G2 Screen

The BOY G2 screen consisted of a word-reading task in which the child read 10 words (with a cutoff of 8 out of 10). No other variables uniquely predicted outcomes at BOY G2. As before, the items selected for the word-reading task were based on a discriminant function analysis of 537 children who had both beginning and end of G2 outcome data. After analyzing a number of different models, the best model included just the word reading w_theta score. Table 3 shows Fp = .15; Fn= .09. Overall accuracy was .77. This set of items has a reliability of .86 and a bivariate correlation with end of G2 WJ broad reading of .80. The ethnicity by identification interaction was not significant, $F(3, 529) < 2.08, p = .10$.

Additional Evaluations

The screens (not the original longitudinal study) underwent two additional evaluations summarized in technical manuals using samples studied in 1998–1999 and 2010. (https://www.tpri.org/index.html). The participants in the 1998–1999 samples came from a field study involving 32 classrooms of KG and G1 students, 128 KG students, and 144 G1 students randomly selected from each class in four elementary schools in a local urban school district. The purposes were to estimate the internal consistency of the tasks using classical test theory and generalizability theory, to collect concurrent validity data, and to evaluate teacher responses to the TPRI. The technical manual shows strong reliability for the screens, similar Fp and Fn as in the original development work, and differential item functioning analyses that identified a low rate of items on the screens that had ethnic or gender bias (<5%). Teachers were highly reliable in administration, regardless of whether the child being tested was from their own classroom or another teacher's classroom.

The 2010 study involved 3,821 children from 203 classrooms in 16 schools. The sample was comparable in gender representation and ethnically diverse, including 1,136 students who were African American, 1,178 who were Latinx, and 1244 who were White. We revalidated the existing TPRI screens. Measures were collected in both the fall and spring at every grade level. Outcome measures were administered at the end of the year and students were evaluated against whether they fell above or below the threshold previously established for the end of the year.

Across the four grades and five screen forms (KG-BOY, KG-EOY, G1-BOY, G1-EOY, and G2-BOY), we evaluated the screens with data from 4,581 outcomes. Table 3 contains the results for each of these forms of the screen. While the TPRI screens would have correctly identified risk status in more than 70%, it is more instructive to consider that the screens maintained Fn rates <10%. There was no need to adjust the cutoffs. Some items were replaced to deal with the age of the screens, but this was done with the IRT work outlined above by selecting items similar in difficulty levels. These findings represent strong cross-validation of the original decision rules.

Discussion

The purpose of this article was to show how we used the EARS study to develop the screens for the TPRI. We identified five challenges to the development of predictive screens that would identify risk for dyslexia/reading difficulties.

First Challenge

The first challenge was to establish an accurate and predictive screen that could be administered in less than 5 min. The five versions of the screens take 3–5 min to administer, with Fn rates below 10%. Our approach used IRT to facilitate and cluster items similar in difficulty around cut points to reduce the number of items required to determine that a student's ability reliably fell above or below the cut point. In this application of IRT, we are not trying to estimate student ability across a continuum of ability, but to determine whether the student's ability lies above or below the cut point.

We used firm thresholds to make reliable binary determinations of risk in children who are developing rapidly, especially in relation to instruction. The determination of who is *not at risk* is deliberately very reliable. In the 2010 study, the TPRI screens would have accurately identified more than 70% of the students screened. In the updated version of the TPRI, of 4,506 children screened across all grades, the TPRI failed to identify 46 children below the criterion measure at end-G1, about 1% of all children screened, or 10% of children identified on the criterion. Children are often at the floor of more traditional psychometric tests and it is easier to identify a child who is not at risk than a child who

is at risk. The net is cast broadly with the goal of not missing children who go on to develop problems.

Second Challenge

The second challenge involved the accuracy of KG precursor skills as predictors of G1 and G2 outcomes. The two KG screens had relatively high Fp (.44, .38) rates, which is also characteristic of other KG predictive screens (Foorman et al., 2015; Wood et al., 2005). Because few children who are at risk were missed, higher Fp rates may be inevitable. This means that almost half of the students are erroneously identified as at risk. However, a positive on the screen indicates a need for an inventory, progress monitoring, and possibly additional or more intensive instruction based on a very short assessment. It is not diagnostic of a reading problem and screening should be part of a sequential, recursive system where errors are corrected over time based on instructional response. Ideally, the types of screens we report can be used as universal screeners as part of a service delivery system based on a MTSS. In KG, diagnostic assessments cannot reliably identify who has or will have dyslexia. A period of instruction, possibly including intensive instruction and/or intervention, and careful monitoring of learning to read over time is needed (Miciak & Fletcher, 2020). For children at risk of dyslexia, MTSS seems ideal for preventing reading problems.

Third Challenge

The third challenge involved when different tasks would be maximally predictive. We found variation in which measures were the best predictors depending on the timing of the screen, sample, and outcomes. In KG, letter–sounds and PA were the best predictors. At BOY G1, letter–sound knowledge had reached a performance ceiling but was retained for children who had not gone to KG and any difficulty was considered a risk indicator. PA and reading simple words were also predictive at BOY G1, but by end G1, only word reading was needed as a screen. At this point, concurrently valid reading tests can be used to screen.

Many states require multiple measures as "screens" for dyslexia (Petscher et al., 2019). This approach is redundant and results in potentially burdensome, unnecessary testing akin to giving a battery of universal screeners when one or two may be sufficient (Gotffreda et al., 2009) and multiple concurrent assessments may actually reduce accuracy (VanDerHeyden et al., 2018). Other states indicate that screening in KG–3 should include multiple measures with no specification of when or how these skills should be screened. Some states focus on PA and rapid naming without specifying what should be named: Alphanumeric symbols are more sensitive than objects/colors, likely because the former measure of early reading (Schatschneider et al., 2002).

Universal screens in commercially available progress monitoring systems commonly use timed letter naming tests. However, these are not the same as rapid naming tests because many letters are named, not the same ones repeatedly. Performance on these types of tests may be determined by the item level difficulty of the letter–sound and whether the child knows the alphabet. Rapid naming and knowledge of letter names and sounds are highly correlated in KG (Schatschneider et al., 2004). In the current study, examining individual classifications of risk showed no value-added component from adding letter naming speed at any age, so inclusion of rapid naming of letters would have only increased the amount of time needed for screening with negligible impact on the accuracy of decisions.

It is not that rapid letter or object naming, or vocabulary are not predictive. In this study, these measures provided no additional information beyond the other tasks in deciding if students scored above or below the cut point for decision-making. In the Boscardin et al. (2008) study, rapid naming of letters was a strong predictor of which students fell into the kindergarten PA growth mixture class that showed both low performance and low growth. Kim et al. (2010) used dominance analysis and found that oral reading fluency growth, vocabulary, and prior reading comprehension were the best predictors of reading comprehension in G1; oral reading fluency was the best predictor of reading comprehension in G2 and G3. Other studies showed that KG language measures, such as vocabulary (Wood et al., 2005), sentence imitation (Catts et al., 2001), and nonsense word repetition were uniquely predictive even with letter knowledge and PA in a model predicting from KG to end G2 (Catts et al., 2015). Oral language difficulties are often comorbid with dyslexia (Snowling & Melby-Lervag, 2016). To incorporate vocabulary, the goal would be not to incorporate an entire test, but to identify the items that maximized prediction, which was done with the development of the FRA (Foorman et al., 2015). We are providing the work around the TPRI as a guide for how to approach screen development.

Fourth Challenge

The fourth challenge was to determine when word-reading tasks would become uniquely predictive of risk. By the end of G1, screening could be accomplished just with a word-reading list of properly calibrated words. At this point, most children have experienced at least a year of reading instruction. Difficulties with word reading after instruction are clearly highly predictive of risk for dyslexia. By the end of G1, screening should focus on either timed or untimed reading of short word lists and passages. Many reliable norm-referenced measures of reading, beginning at EOY G1, may be reliable screens for dyslexia. Even at BOY G1, the ability to read words may be a strong indicator of the *absence* of dyslexia.

Fifth Challenge

The fifth challenge was to ensure that the screens were fair. There were no age or gender effects. Only in the KG assessments did an ethnicity interaction emerge, but it was not in the direction indicating bias. Subsequent assessment of the screening items in 2000 and 2010 showed little evidence of significant differential item functioning by gender or ethnicity.

Limitations and Future Research

We do not propose that the TPRI screens are the best or only way to screen for dyslexia and other early reading problems. Longitudinal datasets should be examined that contain letter naming and word/passage reading tests commonly used for progress monitoring. Including screening with progress monitoring would simplify screening because the same system could be used for screening and progress monitoring. Such measures have demonstrable predictive validity (e.g., Gotffreda et al., 2009), but the measures have not been evaluated using methods like those in the present article. Other alternatives once children reach the end of G1 are norm-referenced tests of timed and untimed word and pseudoword reading, but the thresholds for risk would need to be determined. Even failure on a state test may be indicative of a word-level reading problem such as dyslexia (Vaughn et al., 2010). States should be careful to minimize testing to marshal resources for the important job of instruction and intervention.

The TPRI does not have a national standardization, partly to keep the results from being used for accountability purposes. The IRT methods used to scale items and ability yield parameters that are invariant across differences in populations in the ability distributions, or in the difficulties of the items included. If the items and scales had been calibrated in a sample with a different ability distribution, the item difficulty values in that sample would be a linear transformation of the values obtained in the current samples. It is not clear what a national standardization would add given that the criterion was a national norm-referenced assessment of reading. A change in the ability distribution in the sample would shift the cut point commensurate with the shift in item difficulties, but the same items would discriminate at the cut point. Only if the relations of the predictors and the criterion changed would there be a shift in the performance of the screening, but our fairness analyses suggest there is little reason to suspect that these tests relate differently to reading as a function of demographic characteristics.

The data used in this study were collected in 1992–1996. It is possible that because of changes in instruction, improved measures, and other factors that these results would not be replicated. However, the TPRI screens were robust in two cross-validation samples. Traditional tests, including those used as universal screens, yield a range of scores in a normal distribution in which a cut point is introduced toward the lower end of the distribution. Decisions based on normally distributed, fallible test scores will always yield decision errors just because of the number of items used to make a decision of risk and the unreliability of observable test scores. The version of the WJR was developed in 1989 and a more contemporary version with updated norms might improve the reliability of the screens. The constructs that are measured would not change and all the constructs used in this study remain in use. To the extent that expectations for grade-level performance in G1 and G2 have changed, then it is possible that cut points would need to be adjusted to accommodate such changes in grade-level expectations.

More pertinent is the question of how well the decision rules perform in a situation where the base rate of risk status is higher or lower. We geared the TPRI screens to base rates of 20–25% (Mathes et al., 2005). In a low base rate situation, confidence in positive-screening decisions is weaker because more of the positive decisions will come from children who are not at risk than in a situation where the base rate is higher. Imagine a setting where the incidence of reading problems in 10% rather than 20%. If we tested 1,000 children in this situation and the false-positive rate for the test is 40%, and the false-negative rate is 10%, then the test would identify $.40 \times 900 + .90 \times 100 = 450$ individuals, of which only 100 (22%) are actually at risk. In contrast, if the base rate were 30%, then the test would identify $.40 \times 700 + .90 \times 300 = 550$ individuals, of which 270 (49%) are truly at-risk.

We would recommend careful attention to base rates and post test probabilities (VanDerHeyden, 2013) in implementing any screening, but most importantly in making decisions based on screening results. As the base rate of the problem goes down in the screened population, the probability of being at risk given that a student has failed the screen is less than in a high base rate situation. Although it is possible to monitor predictions to ensure that the cut points work as intended, to the extent students who fail the screens are provided with additional instruction and/or intervention, one should expect lower predictive validity of screening decisions. If students are afforded effective services, students who screen positive may not develop reading problems, which is precisely what one would hope for students at-risk for dyslexia and related reading problems.

Implications for Practice

The primary purpose of early screening is to identify at-risk children so that they can be given proper instruction and intervention to allow them to access print as early in development as possible. After screening, progress should be monitored so that instructional adjustments can be made in a sequential, recursive process that also helps reduce screening errors. Identification should be reserved for children who

show inadequate response to quality instruction. Dyslexia cannot be reliably *diagnosed* in the absence of explicit instruction for students who show risk characteristics (Miciak & Fletcher, 2020). The value of this approach is that screening can be done quickly and reduce the number of children who need monitoring. The screens are highly accurate for identification of children who are not at risk of dyslexia. These children do not require further assessment, but the screen should be periodically administered to ensure that children who are at risk are not missed. Fp rates are higher in KG, but decline through G2.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Jack M. Fletcher and Barbara R. Foorman receive royalties from the University of Texas, and David J. Francis receives royalties from the University of Houston, for sales of the TPRI outside of Texas.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported in part by grants HD28172, HD30995, and HD052117 from the National Institutes of Child Health and Human Development (NICHD), and H326M190008 from the U.S. Department of Education. The opinions expressed are the authors' and not the opinion of the NICHD or Department of Education. The TPRI is owned and copyrighted by the Texas Education Agency, The University of Texas Health Science Center at Houston, and The University of Houston. The FRA is owned and copyrighted by Florida State University (FSU) and has been licensed to Lexia Learning as the RAPID. The development of the FRA/RAPID was supported by the Institute of Education Sciences, U.S. Department of Education, through a subaward to FSU from grant R305F100005 to the Educational Testing Service as part of the Reading for Understanding Initiative. The opinions expressed are the authors' and not the opinion of IES.

Note

1. HD28172 was funded by NICHD in 1992–1996 and entitled "Detecting Reading Problems by Modeling Individual Growth," but was dubbed the Early Assessment of Reading Skills by the study team for ease of communication with schools and parents. Francis was the Principal Investigator; Foorman and Fletcher were Co-PIs.

References

Badian, N.A. (Ed.). (2000). *Prediction and prevention of reading failure*. York Press.

Benton, A. L., & Pearl, D. (Eds.). (1978). *Dyslexia: An appraisal of current knowledge*. Oxford University Press.

Boscardin, C. K., Muthén, B., Francis, D. J., & Baker, E. L. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology*, *100*, 192–208.

Carroll, J., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book*. Houghton Mifflin.

Catts, H. W., Fey, M. E., Zhang, S., & Tomblin, J. B. (2001). Estimating risk for future reading difficulties in kindergarten children: A research-based model and it's clinical Implications. *Language, Speech, and Hearing Services in Schools*, *31*, 38–50.

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities*, *48*(3), 281–297.

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, *24*(8), 1408–1019.

Denckla, M. B., & Rudel, R. E. (1976). Naming of objects by dyslexic and other learning- disabled children. *Brain and Language*, *3*, 1–15.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture Vocabulary Test-Revised*. American Guidance Service.

Elliott, J. G., & Grigorenko, E. L. (2014). *The dyslexia debate*. Cambridge University Press.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. (2019). *Learning disabilities: From identification to intervention*. Guilford Press.

Foorman, B., Petscher, Y., & Schatschneider, C. (2015). *Florida Center for Reading Research (FCRR) Reading Assessments (FRA) Kindergarten to Grade 2* [Technical manual]. http://www.fcrr.org/for-researchers/fra.asp

Gotffreda, C. T., DiPerna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools*, *46*, 539–551.

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652–667.

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, *95*, 211–224.

Lovett, M. W., Frijters, J. C., Wolf, M. A., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. *Journal of Educational Psychology*, *109*, 889–903.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). An evaluation of two reading interventions derived from diverse models. *Reading Research Quarterly*, *40*, 148–183.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *3*, 195–216.

Miciak, J., & Fletcher, J. M. (2020). The critical role of instructional response for identifying dyslexia and other learning disabilities. *Journal of Learning Disabilities*. Advance online publication. https://doi.org/10.1177/0022219420906801

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.

Petscher, Y., Fien, H., Stanley, C., Gearin, B., Gaab, N., Fletcher, J. M., & Johnson, E. (2019). *Screening for dyslexia*. Office of Special Education Programs, National Center on Improving Literacy. https://improvingliteracy.org/

Satz, P., & Fletcher, J. M. (1982). *The Florida Kindergarten Screening Battery*. Psychological Assessment Resources.

Satz, P., Taylor, H. G., Friel, J., & Fletcher, J. M. (1978). Some developmental and predictive precursors of reading disability. In A. L. Benton & D. Pearl (Eds.), *Dyslexia: An appraisal of current knowledge* (pp. 457–501). Oxford University Press.

Scarborough, H. S. (1989). Prediction of reading disability from familial and individual differences. *Journal of Educational Psychology*, *81*, 101–108.

Schatschneider, C., Carlson, C. D., Francis, D. J., Foorman, B. R., & Fletcher, J. M. (2002). Relationships of rapid automatized naming and phonological awareness in early reading development: Implications for the double-deficit hypothesis. *Journal of Learning Disabilities*, *35*, 245–256.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, *96*, 265–282.

Schatschneider, C., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Mehta, P. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology*, *91*(3), 439–449.

Snowling, M., & Melby- Lervag, M. (2016). Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychological Bulletin*, *142*, 498–545.

Torgesen, J. K. (2000). Individual responses in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice*, *15*, 55–64.

Treiman, R., Tincoff, R., Rodriguez, K., Mouzaki, A., & Francis, D. J. (1998). The foundations of literacy: Learning the sounds of letters. *Child Development*, *69*(6), 1524–1540.

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review*, *42*(4), 402–414.

VanDerHeyden, A. M., Burns, M. K., & Bonifay, W. (2018). Is more screening better? The relationship between frequent screening, accurate decisions, and reading proficiency. *School Psychology Review*, *47*(1), 62–82.

Vaughn, S., Cirino, P. T., Wanzek, J., Wexler, J., Fletcher, J. M., Denton, C. D., Barth, A., Romain, M., & Francis, D. J. (2010). Response to intervention for middle school students with reading difficulties: Effects of a primary and secondary intervention. *School Psychology Review*, *39*, 3–21.

Wagner, R., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Assessment of Phonological Processes*. Pro-Ed.

Wood, F. B., Hill, D. F., Meyer, M. S., & Flowers, D. L. (2005). Predictive assessment of reading. *Annals of Dyslexia*, *55*(2), 193–216.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery–Revised*. DLM Teaching Resources.