National Center on Improving Literacy

Screening for Dyslexia

This report was authored and prepared by the members of the National Center on Improving Literacy including: Yaacov Petscher, Ph.D.; Hank Fien, Ph.D.; Christopher Stanley, Ph.D.; and Brian Gearin. Authorship is also attributed to Nadine Gaab, Ph.D. at Boston Children's Hospital/Harvard Medical Center, Jack M. Fletcher, Ph.D. at the University of Houston, and Evelyn Johnson, Ph.D. at Boise State University. We also acknowledge Drs. Donald L. Compton and Christopher W. Schatschneider at Florida State University for initial feedback in the development of this work.

RECOMMENDED CITATION:

Petscher, Y., Fien, H., Stanley, C., Gearin, B., Gaab, N., Fletcher, J.M., & Johnson, E. (2019). Screening for Dyslexia. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. Retrieved from improvingliteracy.org.

The research reported here is funded by a grant to the National Center on Improving Literacy from the Office of Elementary and Secondary Education, in partnership with the Office of Special Education Programs (Award #: S283D160003). The opinions expressed are those of the authors and do not represent views or policies of OESE, OSEP, or the U.S. Department of Education. You should not assume endorsement by the Federal government. © National Center on Improving Literacy.

In recent years, there has been a sharp increase in the number of states seeking to reform education for students with dyslexia. Although states vary in their approaches to reform, state legislation has tended to promote (1) a common definition of dyslexia, (2) universal screening during elementary school, (3) academic, preventative intervention in the early grades, (4) the use of evidence-based interventions, (5) the use of explicit and/or structured sequences of instruction, (6) professional development to facilitate these objectives, and (7) appointing advisory boards to assist with the implementation challenges at the district level (Gearin, Turtura, Kame'enui, Nelson, & Fien, 2018; see also Youman & Mather, 2013, 2015, 2018). Each of these legislative considerations has potential to add to existing challenges and opportunities or to create new ones for school systems.

Universal screening for dyslexia risk is one of the most promising, but most challenging elements of the dyslexia education reform effort. Although the Individual with Disabilities Education Act (IDEA, 2004), includes dyslexia as a part of the specific learning disability definition and while there is some degree of definitional agreement, there is still variation in how states implement IDEA when providing screening specifically for dyslexia. At the time of this writing, twenty-one states have opted legislation inclusive of universal screening systems for risk for dyslexia (National Center on Improving Literacy, 2018). Many other states are considering such. Previous initiatives, particularly *Reading First*, emphasized universal screening and shared other objectives with the newer dyslexia initiatives.

The term *screening* refers to the brief evaluation of a defined population of individuals to identify the risk for performing below a specified threshold or benchmark on a specified outcome (Morabia, 2004). Screenings in the medical community are commonplace, such as the routine evaluation of blood pressure being measured as a screen for hypertension or blood being drawn to screen for diabetes or risk for heart disease via levels of cholesterol. In these examples, screening for health disorders requires a balance of benefits and risks. The main benefit of screening for health risk is that individuals who present with a positive

test result, such as high blood pressure, can be provided with early treatment such as recommended lifestyle changes that may reduce the risk for heart disease. A positive screening for a disorder or its risk may also lead to additional evaluation before interventions are provided, such as the use of medications to reduce risk for heart disease (e.g., anti-hypertensives or cholesterol reducing drugs). It is important to consider other contextual factors associated with screening such as the accuracy and utility of screening devices, financial costs (e.g., for buying screening materials or training personnel), increased anxiety for the tested individual, the provision of unnecessary treatments to those who are incorrectly screened as at risk for a disorder, or the failure to provide necessary treatment to individuals who are incorrectly screened as not at risk for a disorder.

The purpose of universal *screening* for dyslexia risk is very different than the purpose of *diagnosing* dyslexia. Screening determines the level of risk for reading problems in general and the potential risk of having or developing dyslexia. It is not appropriate to use screening results to formally diagnose whether an individual actually has dyslexia. Accordingly, universal screening for dyslexia risk follows a qualitatively different process than the process used for diagnosing dyslexia. Screening procedures for dyslexia risk should be efficient and inexpensive and should be used for all students in a classroom. In contrast, diagnosis or identification of dyslexia requires a more comprehensive, time consuming, and expensive evaluation procedures and should only be applied to individuals in the population that have demonstrated elevated levels of risk demonstrated by screening results, have not responded adequately for generally effective early reading intervention, or both.

Valid screening methods should go hand in hand with valid identification and diagnostic methods. The chief benefit of universal screening for dyslexia risk is that it could prevent the reading problems associated with early common, but often under-identified reading disability. Reading intervention in early elementary school clearly reduces the risk for a reading problem in general, and specifically word-level reading problems epitomized by dyslexia. *It is of paramount importance that states, districts, and schools take action to improve services for students with dyslexia, which begins with efforts at early screening and preventative intervention.*

Although universal screening for dyslexia risk could potentially help many students struggling with reading, it also poses risks and challenges for school systems and the students within them. Implementing an effective universal screening system to understand students' risk for reading disabilities, including dyslexia, is not a simple matter of selecting and administering a one-time test to select children. Rather, all students in all grades should be screened multiple times a year. The most important issue is the reliability and validity of decisions made by professionals and families based on the screening method. Developing an effective screening system for dyslexia requires an inherent trade-off between correct and incorrect classifications from the screener, and more importantly, the risks associated with two types of incorrect screening results; one where the screener says the student is at risk but doesn't perform below a specified threshold or benchmark on a specified outcome (false positive error) and one where the screener says the student is not at risk but ends up performing below a specified threshold or benchmark on a specified outcome (false negative error). The issue with false positives is resource driven. If the screener generates too many false positive errors, it can drain and potentially become a waste of resources to provide intervention and/or additional assessment to students who did not need the extra support. However, the consequences of a false negative error, which could mean lack of access to early reading intervention, is perhaps more serious because of the need for students with or at risk for dyslexia to receive explicit reading instruction as early as possible in the period where reading acquisition is prioritized (kindergarten – Grade 2). Accounting for the reliability and validity of scores from screener assessments, with special attention to the types of errors that are made in screening systems is critical. Scores from a given screener may be unreliable because of its poor psychometric properties or lack validity due to the methods not being sensitive to the risk characteristics associated with dyslexia. In these cases, the resultant decisions about students may be uninformative or misinformative. Alternatively, a screener may have good psychometric properties, but be used by educators or families in ways that are not supported by research or the purposes of the screener. For example, a valid screener for determining a student's risk associated with developing social skills may not be valid for determining if a child is at risk for a reading disability or dyslexia.

Description of Our Paper Sections

Given the importance of the early identification of dyslexia, the present paper aims to provide an overview and some insight into what is known about screening for dyslexia risk. Section I provides a brief overview of "what is dyslexia" and the importance of screening for dyslexia risk. In Section II of this paper, we discuss the neurological and behavioral aspects relevant to dyslexia as well as the emerging research in both areas. Section III provides a robust presentation of viewpoints and considerations for best practices in behavioral screening. Section IV provides a brief overview of key statistical considerations one should consider when evaluating a screener with a companion technical report provided in Appendix A. Section V concludes with a checklist to support teachers, school psychologists, and school-based assessment teams in evaluating and choosing universal screeners.

SECTION I

WHAT IS DYSLEXIA?

Dyslexia is the most common learning disability, historically reported as affecting 5-17% of children (Cortiella & Horowitz, 2014; Shaywitz, 1998). It is commonly understood as a brain-based specific learning disability that impairs a person's ability to spell words in isolation accurately or to read single words fluently (Peterson & Pennington, 2015). It cannot be explained by poor vision or hearing or lack of motivation or educational opportunities. Dyslexia can have a secondary impact on reading fluency and comprehension on the sentence or paragraph level. In addition to its proximal impact upon reading skills, dyslexia has been linked with decreases in self-esteem and amount of time reading outside of school contexts, which may contribute to widening of gaps in reading ability, vocabulary, and background knowledge (Cunningham & Stanovich, 1998; Undheim, 2003). The reading impairments associated with dyslexia are unexpected in that individuals with dyslexia demonstrate otherwise typical learning growth. Although there is some debate about the precise definition of dyslexia (Peterson & Pennington, 2015), states increasingly use the International Dyslexia Association definition of dyslexia:

> *Dyslexia is a specific learning disability that is neurobiological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge (Lyon, Shaywitz, & Shaywitz, 2003). p. 2.*

It is worth highlighting that this definition recognizes the role of the brain in acquiring reading skills (i.e., neurological) as well as dyslexia's primary symptoms being reflected by poor performance in spelling and fluent word reading (i.e., behavioral).

THE IMPORTANCE OF EARLY SCREENING FOR DYSLEXIA RISK

Students may be at risk for not attaining full literacy skills for a variety of reasons. For example, students may be at risk because they are English learners who are struggling to learn literacy skills in two languages simultaneously (Gersten & Brengelman, 1996). Their low performance on literacy tests may be a reflection of a language acquisition challenge rather than an indicator of an underlying disability. Of course, single-cause explanations rarely capture the complexity behind a student's struggle to develop strong literacy skills (Maughan & Carroll, 2006; Snowling & Hulme, 2012). Multiple risk factors may be interacting with each other to make literacy problems more pronounced than they might be if only one risk factor was present (Muter & Snowling, 2009).

Early screening and intervention services are critical for students with undiagnosed literacy-related disabilities, including dyslexia. Screening and intervention can focus on early literacy skills as well as areas of self-regulation and executive control that can hinder the development of reading, writing, language processing, and comprehension. Schools should determine the areas of risk or literacy skills to measure and then choose technically adequate screening measures. Our best scientific evidence indicates that effective *prevention* and early reading intervention services should focus on the literacy-related problems that students who demonstrate risk experience. This includes providing intervention to students with (yet) undiagnosed literacy-related disabilities, including dyslexia, as well as those students who are experiencing literacy-related difficulties for other underlying reasons (Shaywitz, 2014). Whether the literacy-related difficulty is caused by dyslexia or a disability other than dyslexia, another factor (e.g. low oral language skills), or a combination of factors, early and intense intervention to address the difficulties is the best way to prevent early problems from becoming more severe over time (Connor et al., 2014).

The basis for long-term difficulties in learning to read is neurologically mediated. Reading is an acquired skill with at least two pathways that need to be programmed. One is generally referred to as a dorsal or sublexical route that deals with parts of words and involves the capacity to link what is known about the sound structure of language (phonological awareness) with print (the alphabetic principle). The second is a ventral or lexical route that deals with rapid orthographic processing of larger chunks of words and whole words based on the statistical or computational probabilities in which the visual symbols making up written language are ordered. The dorsal part of the reading network largely involves brain regions naturally associated with language processing; the ventral part of the reading network involves areas that naturally deal with visual expertise (e.g., letters and words). The brain is not naturally programmed to read, but makes use of systems designed for other purposes; however, exposure to print, usually through instruction, is necessary to program these areas to mediate reading.

In people with little print exposure, the brain must associate print and sounds through the sublexical, dorsal system. As phonological awareness develops, the person must deal with increasing large chunks of words to develop a repertoire of sight words instantly recognized with direct access to meaning, which leads to minimization of the need for sublexical decoding and is an indirect and inefficient pathway to meaning. Experience is essential for the ventral experience. Both the dorsal and ventral parts of the reading network operate in processing words but are differentially activated depending on the properties of the words and the experience of the reader. The key is early access to print so that the vental system can be programmed for rapid processing. In people with dyslexia, it is more difficult to program these systems, and earlier, more explicit instruction is required. If instruction is not early enough, the person at risk for dyslexia will not get the print exposure necessary to program these systems and develop the capacity for automatic reading that is essential for comprehension. Neuroscience helps explain the demonstrable efficacy of early reading intervention in preventing dyslexia and the need for long-term remedial intervention that has to be even more explicit and long-term in cases in which the reading problem was not detected, or the person did not respond adequately to intervention. *Early screening becomes the key*. Although neurobiological methods (i.e., neuroimaging and genetic) are not yet available or affordable for mass screening, behavioral methods that utilize the types of tasks used to activate the brain are available. However, the neurobiological research

field is essential for the discovery of the etiology of language-based learning disabilities. It can characterize when along a child's developmental trajectory brain development becomes atypical and can examine potential differences between subsequent good and struggling readers before behavior can be measured.

SECTION II

NEUROLOGICAL CONSIDERATIONS AND SCREENING

Considerable research has been conducted to clarify the neurobiological aspects of reading development, including the brain network involved in pre-reading skills, single and complex text reading, reading fluency and comprehension. Furthermore, many research studies have examined components of the reading networks that may be altered, and how typical and atypical reading networks develop over time using structural magnetic resonance imaging (MRI) of the brain (e.g., Martin, Kronbichler, & Richlan, 2016; Richlan, Kronbichler, & Wimmer, 2013; Lyytinen, Erksine, Hamalainen, Torppa, & Ronimus, 2015). Neuroimaging research has given us a first glimpse into the developing reading circuit which is primarily, but not exclusively, located in the left hemisphere of the brain. For instance, it has been shown that in the initial stages of reading development, specifically single word reading, superior temporal regions that are involved in oral language processing during early childhood start forming connections with temporo-parietal and subsequently occipito-temporal brain regions (e.g., Pugh et al., 2001). Although the temporo-parietal regions (also may be referred to as dorsal) have been shown to be supporting the integration of phonology and orthographical patterns, the occipito-temporal (also may be referred to as ventral) region, also often called the visual-word-form area, supports the (rapid) identifications of letters and words (e.g., Dehaene & Cohen, 2011; Price & Devlin, 2011). With increasing reading experience, inferior frontal regions are integrated into the reading circuit to support lexical access, semantics and executive functioning, important for reading fluency and comprehension (e.g., Price, 2012; Rimrodt et al., 2008). Structural and functional differences between individuals with and without dyslexia have been reported in all components of the reading network. Studies employing MRI have shown reduced gray matter volume indices and cortical thickness in children and adults with dyslexia compared to their peers (e.g., Lyytinen et al., 2015). Furthermore, functional hypoactivation (reduced activation) have been reported in the components of the reading network for a variety of tasks including phonological processing, letter recognition and word identification. These differences can even be observed in comparison to younger children with similar reading levels to suggest that individuals with dyslexia exhibit fundamental alterations within the reading

network and they do not have delayed brain maturation as frequently suggested (Hoeft et al., 2006; Hoeft et al., 2007). Additionally, alterations in white matter tracts have been reported for individuals with dyslexia compared to their peers, suggesting that atypical brain development can be observed on the cortical level as well as for structural and functional connections between cortical regions (e.g., Klingberg et al., 2000, Wang et al., 2016; Rimrodt, Peterson, Denckla, Kaufmann, & Cutting, 2010; Yeatman et al., 2011). These fundamental differences in atypical brain development between individuals with dyslexia and their peers have also been reported in studies using electroencephalography (EEG; Ozernov-Palchik & Gaab, 2016a) and have shown that the observed differences between children who subsequently develop dyslexia or do not are present prior to the onset of formal reading development. This finding suggests that dyslexia is not a result of the daily struggle to learn to read but predates the onset of reading instruction (Im, Raschle, Smith, Ellen Grant, & Gaab, 2015). Differences between children who subsequently develop reading impairments, including dyslexia or who have a familial risk, have been observed as early as infancy and preschool (e.g., Langer et al., 2017; Leppanen et al., 2012; Molfese, 2000; Molfese, Molfese, & Modgline, 2001; Raschle, Chang, & Gaab, 2011).

Behavioral Considerations in Screening

The simple view of reading (SVR; Gough & Tunmer, 1986) is the prevailing view that understanding individual differences in reading comprehension is a function of decoding and linguistic comprehension skills (Lonigan, Burgess, & Schatschneider, 2018). As such, universal screening processes in kindergarten appear to be most successful when inclusive of phonological awareness tasks (e.g., phoneme segmentation, blending, onset and rime), rapid automatic naming tasks (e.g., letter naming fluency), letter-sound association, and phonological memory (Catts et al., 2015). In first grade, reading and language screening may also include tasks associated with phoneme awareness and segmentation and be expanded to include letter manipulation, nonword repetition, oral vocabulary, and word recognition fluency (Compton et al., 2010). In second grade, assessment may move towards word identification (i.e., real *and* nonsense words), oral reading fluency, and reading comprehension (Universal Screening, 2017). These screening

assessments are brief and, with training, relatively easy to administer to all children, and they can help capture each child's reading and language strengths and weaknesses in key early stages of development.

STATE AND SCHOOL-LEVEL APPROACHES TO SCREENING.

There is evidence that diverse approaches to behavioral screening are being employed by schools based on current legislation. First, state laws and state-issued dyslexia handbooks vary in terms of their screening recommendations and requirements. For instance, Alabama requires screening in kindergarten on *letter naming skills, letter sound skills, phoneme segmentation skills, and nonsense word fluency skills* (Alabama State Board of Education, 2016). In grades 1-2, students should be screened on *accuracy of word reading, spelling skills, phonemic decoding efficiency skills, and sight word reading efficiency*. In contrast, Nevada requires that grades K-3 screening include: *phonological and phonemic awareness; sound-symbol recognition; alphabet knowledge; decoding skills; rapid naming skills;* and *encoding skills* (Nevada Department of Education, 2015). Certain states (e.g., Oregon and Washington) make similar recommendations, but also recommend that family history be considered as part of the screening process. Thus, while states are broadly similar in terms of the constructs they measure and when, there is not one common approach to screening.

Secondly, states differ in terms of the resources they provide for locating screeners. Many state education agencies have issued handbooks, resource guides, or other documents meant to help school districts locate potentially useful screeners. It is somewhat difficult to generalize about the contents of these documents because over 20 different screeners are recommended across states and it is not always clear what screener is being recommended due to inconsistent and incomplete references. States have also taken different approaches with the regard to specificity with which they recommend screeners. For instance, some states recommend assessment systems (e.g., DIBELS; University of Oregon, 2018) whereas others recommend subtests by target construct (e.g., DIBELS Letter Naming Fluency for assessing alphabet knowledge). Both "brand-name" (e.g., CORE Phonics Survey; Scholastic, 2002) and state-developed measures (e.g., Arkansas Rapid Automated Naming; Arkansas Department of Education, 2017) are represented across state lists. State lists also recognize screeners that vary in terms of their

approaches to administration and scoring. For instance, both behavioral measures to be completed by the student and observational measures to be completed by teachers have been recognized as potentially useful. Similarly, individually-administered and grouped-administered tests and computerized and paper-and-pencils screeners have been recognized as useful. AIMSWeb (Shinn & Shinn, 2002) and DIBELS are among the measures that appear on the majority of state lists.

SECTION III

CONSIDERATIONS FOR BEST PRACTICES IN UNIVERSAL SCREENING

There is broad agreement that schools should implement early screening (e.g. as early as preschool and kindergarten) and intervention programs. However, there are converging and diverging viewpoints on *how* this implementation should occur. For instance, state laws and expert opinion generally favor the use of universal screening within a multi-tiered system of support (MTSS; Gearin et al., 2018; Youman & Mather, 2015). Within such frameworks, students are screened early and at multiple timepoints to assess risk for dyslexia and reading disabilities in general (e.g., twice per year in grades K-3). Scores that assess risk for dyslexia are then used to make instructional decisions, such as grouping or the delivery of intensive intervention specially designed to address individual student needs. There is also widespread agreement that school systems should carefully attend to the foundational elements of MTSS when implementing universal screening. These elements include: (a) the technical adequacy and efficiency of screeners, (b) contextual factors affecting implementation (e.g., cost and knowledge requirements), and (c) the ways in which screening data will ultimately inform placement and instruction (Adlof, Scoggins, Brazendale, Babb, & Petscher, 2017; Pentimonti, Walker, & Zumeta, 2017; Poulsen, Nielsen, Juul, & Elbro, 2017). Finally, there is basic agreement about how to gauge the technical adequacy of screeners. The National Center on Intensive Intervention conducts independent, standardized reviews of screening measures to help schools select an appropriate screening tool.

On the other hand, viewpoints related to the screening issues tend to *diverge* when it comes to the many specific decisions that schools must make when implementing universal screening in an MTSS. Virtually every aspect of the implementation process has been subject to debate over the past few decades, from the selection of screeners, to the best use of screening scores (e.g., Breaux et al., 2017; Gillis, 2017; VanDerHeyden & Burns, 2017). Several factors likely contribute to the diverging viewpoints. The first factor is the sheer number of decisions school systems must make when implementing universal screening. Effectively implementing universal screening is not a simple matter of selecting and administering a test.

School administrators and educators must consider when the screener will be administered, to whom, and how the scores will be used. Because each one of these decisions must be evaluated against the local context, and because school systems differ in terms of their student populations, financial resources, technical infrastructure, and schedules, what is best-practice for one school and its students might not be deemed best practice for another.

A second factor complicating the identification of best-practices is that screening for any condition usually involves tradeoffs in risks, costs, and benefits. For instance, one screener might be highly sensitive and occasionally indicate the presence of a condition in individuals who do not actually have it (i.e., a false positive), while another screener may be less sensitive and occasionally fail to indicate the presence of a condition (i.e., a false negative). The first scenario may lead to unnecessary treatment for the condition, while the latter scenario may lead to the lack of necessary treatment. Because screening involves a balance between values, it can be difficult to gain universal support for a particular practice. Much of the on-going research on dyslexia screeners aims to elucidate how various aspects of screening (e.g., the number of measures, the level of cut scores, the administration protocol) can be adjusted to achieve a better balance between cost and benefits (e.g., Andrade, Andrade, & Capellini, 2015; Johnson, Jenkins, Petscher, & Catts, 2009; Piasta, Petscher, & Justice, 2012). Together, there is an extensive array of legislative, practical, and contextual considerations related to identifying and supporting students with dyslexia. While some accord exists between states, a complex picture of screening for dyslexia risk emerges.

The recommendation to administer dyslexia screenings in schools before 3rd grade is based on research on the prevention and early remediation of reading problems, including dyslexia intervention research. Three empirical findings support the use of screening for dyslexia risk in the early grades. First, reading problems can be prevented, and early problems remediated, through early identification. Early identification through screening assessments allows interventions to be implemented effectively as soon as possible. Second, patterns of reading development are established early once school begins and are stable over time unless interventions are implemented to increase student progress (Good, Kaminski, Simmons, & Kame'enui, 2001; Juel, 1988; Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992;

Torgesen, 2000; Torgesen et al., 2001). Third, without intense interventions, struggling readers do not eventually "catch up" to their average performing peers—in fact, the gap between strong and weak readers increases over time (Torgesen, 2000; Torgesen et al., 2001). Reading interventions that begin in third grade and beyond are likely to be less successful and less cost-effective than interventions that begin in the earlier grades. The later interventions begin, the longer they take to work, the longer they need to be implemented each day, and the less likely they are to produce desired effects (Adams, 1990; Good, Simmons, & Kame'enui, 2001; Snow, Burns, & Griffin, 1998; Stanovich, 1986; Torgesen, 2000; Torgesen et al., 2001). All of these findings are undergirded by the neurobiological research reviewed above.

The purpose of a dyslexia screening assessment is to identify students at risk for dyslexia and reading difficulties and students on track for successful reading outcomes. Screening data are used to make decisions about the level of instructional support students need. Students at high risk of dyslexia and poor reading proficiency—that is, students well below grade level reading expectations—should receive more instructional support than students who are on track for solid reading proficiency. Schools should provide at least three levels of instructional support for students, based on the risk that students face for poor overall reading proficiency:

1. Core classroom instruction for students reading at or above grade level (i.e., low risk for dyslexia and reading problems)—these students meet or exceed reading proficiency expectations;
2. Moderate additional support for students reading somewhat below grade level expectations (moderate risk for dyslexia and reading problems)—these students nearly meet reading expectations or are below reading proficiency expectations. These students would benefit from small group interventions (e.g., Tier 2) that include phonemic awareness and phonics instruction that is tailored to their needs;
3. Intense additional support for students reading well below grade level expectations (at high risk for dyslexia and reading problems)—these students are well below reading proficiency expectations. These students would benefit from small group intensive interventions (e.g., Tier 3) that include phonemic awareness and phonics instruction.

In terms of screening students for dyslexia and reading problems, the recommendation is that, in the absence of a progress monitoring system that assesses students' response to intervention instruction and growth multiple times per year, a screening assessment should be administered to all students in K-3 two times per year (e.g., beginning and end of the school year). The first screening assessment of the school year should be administered as early as possible (e.g., within two weeks to one month of the start of school) so that the information can be acted on right away. Although screening this early may be useful as a baseline to capture kids early, it is critical to note that as students are developing in their skills, many kindergarten screening assessments present with floor effects (e.g. Catts et al., 2008) resulting in very high false positive rates. The need to collect screening data early in the school year, and the need to collect it frequently in most grades and with all students, means that screening assessments should be efficient to administer. Fortunately, there are screening measures available that are efficient to use in K-3, and that provide strong information about level of student reading risk (NCII, 2018).

Dyslexia screening assessments should directly measure students' proficiency on essential reading content or essential pre-literacy measures (depending on the student's grade level/skill level). In the early elementary grades (K-3), screening assessments should focus on the development of a number of different foundational skills necessary for skillful reading. In kindergarten, knowledge of the alphabet, assessed through letter naming, is the most valuable screening tool (Adams, 1990). Also, early in kindergarten, students' developing awareness of the phonemic structure of spoken words is a good predictor of reading, and thus a strong screening measure (Adams, 1990; O'Connor & Jenkins, 1999; Spector, 1992). Assessing both letter knowledge and phonological awareness skills early in kindergarten should be part of a screening system in reading. By the middle and end of kindergarten, schools should screen students for problems with alphabetic understanding (phonics) as well as oral language. In first, second, and third grades regular assessments of word and/or reading fluency should be used to screen students for problems with fluent reading and for likely problems with reading comprehension. In 4th grade through the beginning of high school, it is recommended that reading fluency or computer adaptive assessments be administered two to three times per year. Particularly for students reading below grade level, fluency assessments may help determine if fluency problems are contributing to reading

comprehension problems. A recommended screening protocol will look as following; however, many school districts should carefully consider and identify their respective personnel and resources available to follow these protocols:

- Select screening assessment(s) through a careful process that take into account the population of interest, the scope of the assessment, the reliability and validity of scores, and the classification accuracy of the screener relative to specified outcome (see Appendix A for details)
- Administer screening assessment(s) at necessary intervals.
- Immediately following each screening assessment, a designated staff person should enter the data into a database and print the screening reports.
- A grade level team meeting in the primary grades, and cross-discipline team meetings in upper grades, should occur after each school-wide screening assessment to analyze the screening reports and determine instructional grouping and placement decisions for each student.
- Engage parents/families in decision making and keep them updated on child performance.

SECTION IV

CORE STATISTICAL CONSIDERATIONS IN BEHAVIORAL SCREENING

This section provides a brief presentation of tenets that underpin the reliability, validity, and classification accuracy of screening scores to support the process of choosing and using appropriate assessments. Appendix A presents an expanded view of this section along with statistical concepts of screeners more frequently seen in academic journals, book chapters, and technical reports. We provide this information as a way to expose the reader to processes that should be present in a screener assessments' technical report so that careful, robust evaluation of tools may take place to gauge the appropriateness for one's local screening context. Core concepts and statistical underpinnings of school-based screening have been covered extensively in educational and psychological literature (e.g., Glover & Albers, 2007; Schatschneider, Petscher, & Williams, 2008). In many ways, these exemplar sources take advantage of, or presume, a reader's knowledge of more basic conceptual and statistical concepts that themselves undergird the screening process. These concepts include reliability, validity, and classification accuracy.

RELIABILITY

The most basic definition of reliability is the consistency of a set of scores for a measure, yet this definition may be deceptively simplistic in the context of psychometrics due to the number of ways it can be estimated. Different forms of reliability of include internal consistency, alternate-form, test-retest, split-half, and inter-rater. Careful evaluation of a screener's reliability is necessary as not all forms of reliability are created equal (see Appendix A).

VALIDITY

Just as reliability is multifaceted in nature, so is the concept of validity to the point that we may be able to provide an alphabetized and non-exhaustive sample of forms of validity that include aetiological, conclusion, concurrent, construct, content, convergent, criterion, discriminant, ecological, external, face,

factor, hypothesis, in situ, internal, nomological, predictive, translational, treatment, and washback. At its core, validity is simply concerned with the extent to which something measures what it purports to measure. A word reading test should measure word reading and not receptive vocabulary. An historical perspective of validity was that three independent forms of validity existed (i.e., content, criterion, and construct validity) and could be readily interchanged (Messick, 1995). Content validity is primarily established by the consistency of expert judgments that test content is related to its described use. A classical definition of criterion validity is the simple correlation between a test score and an outcome score, and construct validity is concerned with the interpretation and use of scores (Messick, 1995). Messick (1989) sought to reconceptualize all forms of validity as forming a cohesive, unified framework of *construct* validity. This framework includes six areas that should be evaluated to measure a test, including a screener's, construct validity (see Appendix A).

Classification Accuracy

The language around classification accuracy and the process by which students are correctly or incorrectly identified as at risk is diverse in the same ways as reliability and validity. Classification accuracy is a form of concurrent and predictive validity that looks at how a sample of individuals falls into one of two outcome groups (i.e., within an educational context, those who perform at or above a cut point on an outcome and those who perform below a cutpoint on an outcome. In the medical literature, the classic two outcome groups are those who are noted as failing the outcome or passing the outcome) based on two screener groups (i.e., at risk or not at risk on the screen). When a sample of individuals are given a screener assessment and a have a score on an outcome measure that would be considered the best available benchmark (typically called a gold standard outcome), a 2x2 contingency matrix (Table 1) can be created from which one is able to mathematically calculate important classification accuracy indexes.

Table 1. *Sample 2x2 contingency table*

Screen	Outcome	
	Fail	Pass
At Risk	**A:** True Positive	**B:** False Positive
Not At Risk	**C:** False Negative	**D:** True Negative

Four cells characterize student performances on the screen and outcome measure: Cell A individuals are called *True Positives* as these are individuals who were identified as at risk on a screener and performed below the set threshold on the outcome (e.g., below the 20th percentile of a standardized reading test); Cell B individuals are called *False Positives* as they were classified as at risk on the screener but ultimately performed at or above the set threshold on the outcome (e.g., above the 20th percentile of a standardized reading test); individuals in Cell C are *False Negatives* as they were identified as at risk on the screener and performed below the set threshold on the outcome; and Cell D are the *True Negative* individuals who were not at risk on the screener and performed at or above the set threshold on the outcome. Each cell by itself provides meaningful information about base classifications; however, there are ancillary computations that result in statistics that researchers and practitioners use to evaluate the screening efficiency at given screener-outcome cut-point selections and may be broadly classified as population-based indices, sample-based indices, and overall contextual indices (see Appendix A).

SECTION V

DECISION MAKING

The number of considerations when creating, evaluating, choosing, or using a screener for dyslexia can be overwhelming. In this final section, we provide guidance and questions to consider when selecting screener assessments. When reviewing a screener technical report, tool chart, or summary of the assessment, we recommend the following list of questions to guide your discussions:

POPULATION OF INTEREST

1) How is the population defined?
    a) What is the intended age range for the assessment?
    b) How is the outcome (e.g., dyslexia, learning disability) defined?
2) When the screener was normed, did the sample reflect your intended population?
    a) How similar is the norming sample to your local environment?
    b) Is the sample size for validating the screener sufficient for the analyses?
    c) Were multiple sites, states, or regions used to validate the screener?

SCOPE OF ASSESSMENT

3) How is the outcome from question 1b operationally defined?
    a) What is the outcome by which students are judged to have a skill deficiency (e.g., standardized word reading test)?
    b) What cut-point is used on the outcome from question 3a to define “failure”?
    c) Is the cut-point from 3b reasonable for your local environment?
    d) Is the content on the screener reflective of what should be measured?
    e) Is the screener a measure of accuracy (e.g., total score) or automaticity (e.g., fluency)?
        i) If the screener is computer adaptive, is the content developmentally appropriate for your local environment?
    f) Does the screener use more than one assessment?
        i) If yes, does the assessment provide guidance on how to use the scores in combination with each other?
        ii) If yes or no, does there appear to be good conceptual alignment between the screener and the outcome?

Statistical Considerations

Reliability

4) What type(s) of reliability are reported?
    a) If the screener is item-based, is internal consistency reported?
    b) If test-retest is reported, what is the spacing between testing occasions?
    c) If alternate-form or split-form reliability is reported, is another form of reliability reported?
    d) Are at least two forms of reliability reported?
    e) What level of reliability is reported?
        i) If the screener is not computer adaptive, is the internal consistency
            (1) At least .80 (important for research decisions)?
            (2) At least .90 (important for decision-making purposes)?
        ii) If the screener is computer adaptive
            (1) Is only marginal reliability reported (i.e., overall)?
            (2) Is reliability across a range of ability reported?
            (3) What is the level of reported reliability?

Validity

5) Content Validity
    a) Has the domain been well defined (see question 1)?
    b) Is the domain relevant as defined?
    c) Is the content appropriate for the local environment (see question 3.e.i)?
6) Substantive Validity
    a) Is there a reporting of how the test design matches the construct?
7) Structural Validity
    a) Are there tests of the factor structure/dimensionality reported (e.g., exploratory or confirmatory factor analysis)?
8) Generalizability
    a) For Bias, has one of the following types of analyses been used to test that the screener is not biased against subgroups (e.g., sex, race, poverty, students with disabilities, dual language learnings)
        i) Item-level bias analysis (e.g., differential item functioning)
        ii) Test-level bias analysis (e.g., differential classification accuracy)
9) External
    a) Convergent Validity
        i) Are correlations reported between the screener score and scores from an assessment on a related construct?
        ii) Are the correlations at least .60?
    b) Discriminant Validity
        i) Are correlations reported between the screener score and scores from an assessment on an unrelated construct?
        ii) Are the correlations no greater than .20?
    c) Predictive Validity

i) Are correlations reported between the screener score at one time point and scores on an assessment at a later time point?
ii) Are the correlations at least .20?

10) Consequential Validity
    a) Does the report document any intended or unintended side effects for those who are identified or misidentified based on the selected cut-points?

Classification Accuracy

11) Is Sensitivity reported?
    a) Is it at least .80?
    b) Is a confidence interval reported and is the lower bound of the confidence interval at least .80?
12) Is Specificity reported?
    a) Is it at least .80?
    b) Is a confidence interval reported and is the lower bound of the confidence interval at least .80?
13) What is the Area under the curve?
    a) Is it at least .80?
    b) Is a confidence interval reported and is the lower bound of the confidence interval at least .80?
14) What is the False Positive rate?
15) What is the False Negative rate?

The totality of these 15 organizing considerations are to provide a deeper dive into the world of behavioral screening and many of the guiding principles that researchers use when they evaluate, create, choose, or use screening assessments.

Decision-making framework. Screening for dyslexia risk should be enveloped with a decision-making framework that answers four fundamental questions.

1. *Is the student at risk for dyslexia or not meeting essential pre-reading and reading goals?* Student assessments can screen students for dyslexia or reading difficulties, and the data help determine the level of reading risk students face. Students with moderate to high risk for dyslexia should be provided validated interventions that focus on explicit phonemic awareness and phonics.
2. *Is the student making enough reading progress to read proficiently and reach important reading goals?* Frequent reading assessments can monitor the progress students are

making toward overall proficient reading and important reading goals. These assessments can help determine if students at risk for dyslexia are responding adequately to validated interventions and/or if interventions should be modified or intensified.

3. *Is the student reading with sufficient proficiency to meet grade level reading expectations and essential reading goals?* Summative or outcome assessments can determine if students are reading proficiently and are reaching important reading goals.
4. *For students not making adequate reading progress despite intense intervention, what additional intervention approaches have the best chance of improving the rate of reading progress?* Diagnostic assessments can provide detailed information about students' reading skills for the purpose of developing and implementing individualized interventions for students.

Assessments are needed to answer each of these four questions, and the information is used to make specific educational decisions (Consortium on Reading Excellence, 2008; Kamil et al., 2008). Often, an assessment measure a school uses for one purpose can also be used for additional purposes. In particular, the same assessment measure, administered at different points in time, can frequently be used to *screen* students for dyslexia or reading problems, monitor reading progress over time, and determine if students have met important reading outcomes.

With competing approaches and definitions to dyslexia, along with an array of screening tools and areas, it is important to be aware of and employ empirically-supported and evidenced-based practices. First, it is important to know the scientific traditions from which dyslexia research has emerged. Developmental neuroscientists along with behavioral and cognitive researchers, have advanced considerably our understanding of how structural and functional impairments can negatively impact information processing directly related to reading skills. For instance, it has been shown that the functional and structural brain characteristics of dyslexia are present prior to the onset of formal reading instruction provided in school. Through fruitful collaborations between cognitive and behavioral scientists and developmental cognitive neuroscientists, a greater understanding of the brain-based etiology and behavioral and cognitive symptomology of dyslexia has emerged. For instance, Gabrieli (2009) noted brain-

based assessments and markers enhance the accuracy with which behavioral measures alone are able to predict dyslexia. While the traditions complement one another in many ways, neuroscience researchers are careful to note the additional contributions neural measures offer are moderate and lack sensitivity and specificity, and therefore may not yet warrant the cost and logistical issues associated with their use with young children (Ozernov-Palchik & Gaab, 2016b).

CONCLUSION

There is little debate as to whether the early identification of students is a useful mechanism by which students who are at-risk for reading problems, including dyslexia, can be routed to appropriate next steps such as intensive early interventions (in pre-school or kindergarten) or more in-depth diagnostic testing for diagnoses of reading disabilities. The current paper serves to highlight core considerations and questions one may use in choosing a screener for use in elementary school settings. A final recommendation we provide to summarize the presented work is to underscore the importance of researcher-practitioner partnership. As practitioners look to use screeners in their local contexts that may present with varying student characteristics, there may be challenges in how one chooses a screener. The process of screener selection can seem daunting when considering the number of technical quality standards presented here. Through researcher-practitioner partnerships, the complexities may be mitigated such that researchers may learn about the local context to then support the practitioner in weighing the trade-offs such as: 1) characteristics of students in the school compared to those in the screener validation study; 2) the scope of assessment related to content and the screening needs in the school; and 3) which forms of reliability, validity, and classification accuracy should be preferred and at what level of reporting. The intended fruit of such partnerships would be a smaller set of screening tools identified that demonstrate close correspondences between intent of the screener and appropriateness for the context, and ideally a working collaborative between front-line educators in schools and researchers to continue to enhance screening practices.

REFERENCES

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Adlof, S. M., Scoggins, J., Brazendale, A., Babb, S., & Petscher, Y. (2017). Identifying children at risk for language impairment or dyslexia with group-administered measures. *Journal of Speech, Language, and Hearing Research, 60*(12), 3507-3522.

Alabama State Board of Education. (2016). *Alabama Dyslexia Resource Guide.* Retrieved from Alabama Department of Education website: https://www.alsde.edu/sec/ari/dyslexia/dyslexia%20resource%20guide.pdf

Anastopoulos, A. D., Beal, K. K., Reid, R. J., Reid, R., Power, T. J., & DuPaul, G. J. (2018). Impact of child and informant gender on parent and teacher ratings of attention-deficit/hyperactivity disorder. *Psychological Assessment, 30*(10), 1390-1394. http://dx.doi.org/10.1037/pas0000627

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*(5), 732-740.

Andrade, O. V., Andrade, P. E., & Capellini, S. A. (2015). Collective screening tools for early identification of dyslexia. *Frontiers in Psychology, 5,* 1581.

Arkansas Department of Education. (2017). *Arkansas Dyslexia Resource Guide.* Retrieved from http://www.arkansased.gov/public/userfiles/Learning_Services/Dyslexia/DRG-Final-12-13-17-JS1.pdf

Breaux, K., Scheller, A., & Eichstadt, T. (2017). Dyslexia screening myths: A response to VanDerHeyden and Burns. *Communiqué*, *46*(1), 1–27.

Catts, H. W., Bridges M. S., Little T. D., & Tomblin J. B. (2008). Reading achievement growth in children with language impairments. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1569–1579. https://doi.org/10.1044/1092-4388(2008/07-0259)

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*, 281-297.

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education, 3*(1), 1-8.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*(1), 55-75.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., ... & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327.

Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., ... & Bouton, B. (2012). Accelerating chronically unresponsive children to tier 3 instruction: What level of data is necessary to ensure selection accuracy? *Journal of Learning Disabilities, 45*(3), 204-216.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*, 327-340.

Connor, C. M., Alberto, P. A., Compton, D. L., & O'Connor, R. E. (2014). *Improving reading outcomes for students with or at risk for reading disabilities: A synthesis of the contributions from the institute of education sciences research centers* (No. NCSER 2014-3000). Washington, D.C: National Center for Special Education Research.

Connor, C. M. (2019). Using technology and assessment to personalize instruction: Preventing reading problems. *Prevention Science, 20*(1), 89-99.

Consortium on Reading Excellence. (2008). Core phonics surveys. In B. Honing, L. Diamond, & R. Nathan (Eds.), *Assessing reading: Multiple measures for kindergarten through eighth grade* (2nd ed., pp. 63–80). Novato, CA: Arena.

Cortiella, C., & Horowitz, S. H. (2014). *The state of learning disabilities: Facts, trends and emerging issues*. New York: National Center for Learning Disabilities.

Cummings, K. D., Biancarosa, G., Schaper, A., & Reed, D. K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology, 52*(4), 361-375.

Cummings, K. D., Park, Y., & Bauer Schaper, H. A. (2013). Form effects on DIBELS Next oral reading fluency progress-monitoring passages. *Assessment for Effective Intervention, 38*(2), 91-104.

Cunningham, A.E., & Stanovich, K.E. (1998). What reading does for the mind. *American Educator, 22*, 8-15.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

Davis, G. N., Lindo, E. J., & Compton, D. L. (2007). Children at risk for reading failure; constructing an early screening measure. *Teaching Exceptional Children, 39*(5), 32-37.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in cognitive sciences*, *15*(6), 254-262.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Minneapolis, MN: Pearson.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3), 315-342.

Gabrieli, J. D. E. (2009). Dyslexia: A new synergy between education and cognitive neuroscience. *Science, 325*, 280-283.

Gearin, B., Turtura, J., Kame'enui, E. J., Nelson, N. J., & Fien, H. (2018). A multiple streams analysis of recent changes to state-level dyslexia education law. *Educational Policy*, https://doi.org/10.1177/0895904818807328

Gersten, R., & Brengelman, S. U. (1996). The quest to translate research into classroom practice: The emerging knowledge base. *Remedial and Special Education*, *17*(2), 67-74. http://dx.doi.org/10.1177/074193259601700202

Gilbert, J. K., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2012). Early screening for risk of reading disabilities: Recommendations for a four-step screening system. *Assessment for Effective Intervention, 38*(1), 6-14.

Gillis, M. (2017). How RTI supports early identification of students with different reading profiles. *Perspectives on Language and Literacy*, *43*(3), 41–45.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135.

Good, R. H., Kaminski, R. A., Simmons, D., & Kame'enui, E. J. (2001). Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an Outcomes-Driven Model: Steps to Reading Outcomes. *OSSC Bulletin*, *44*(1).

Good, R. H., Simmons, D. C. , & Kame'enui, E. J. (2001). The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*, 257—288.

Gough, P. and Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*, 6–10.

Hoeft, F., Hernandez, A., McMillon, G., Taylor-Hill, H., Martindale, J. L., Meyler, A., ... & Whitfield-Gabrieli, S. (2006). Neural basis of dyslexia: a comparison between dyslexic and nondyslexic children equated for reading ability. *Journal of Neuroscience*, *26*(42), 10700-10708.

Hoeft, F., Meyler, A., Hernandez, A., Juel, C., Taylor-Hill, H., Martindale, J. L., ... & Deutsch, G. K. (2007). Functional and morphometric brain dissociation between dyslexia and reading ability. *Proceedings of the National Academy of Sciences*, *104*(10), 4234-4239.

Im, K., Raschle, N. M., Smith, S. A., Ellen Grant, P., & Gaab, N. (2015). Atypical sulcal pattern in children with developmental dyslexia and at-risk kindergarteners. *Cerebral cortex*, *26*(3), 1138-1148.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 § (2004).

International Dyslexia Association (2002). *Definition of dyslexia*. Retrieved from https://dyslexiaida.org/definition-of-dyslexia/

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131-140.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*(4), 174-185.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*(4). 437-447. http://dx.doi.org/10.1037/0022-0663.80.4.437

Kamil, M. L., Borman, G. D., Dole, J., Kral, C. C., Salinger, T., and Torgesen, J. (2008). *Improving adolescent literacy: Effective classroom and intervention practices: A Practice Guide* (NCEE #2008-4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47*(2), 125-135.

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.

Klingberg, T., Hedehus, M., Temple, E., Salz, T., Gabrieli, J. D., Moseley, M. E., & Poldrack, R. A. (2000). Microstructure of temporo-parietal white matter as a basis for reading ability: evidence from diffusion tensor magnetic resonance imaging. *Neuron*, *25*(2), 493-500.

Langer, N., Peysakhovich, B., Zuk, J., Drottar, M., Sliva, D. D., Smith, S., ... & Gaab, N. (2017). White matter alterations in infants at risk for developmental dyslexia. *Cerebral Cortex*, *27*(2), 1027-1036.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575.

Leppänen, P. H. T., Hämäläinen, J. A., Guttorm, T. K., Eklund, K. M., Salminen, H., Tanskanen, A., ... & Lyytinen, H. (2012). Infant brain responses associated with reading-related skills before school and at school age. *Neurophysiologie Clinique/Clinical Neurophysiology*, *42*(1-2), 35-41.

Lonigan, C. J., Burgess, S. R., & Schatschneider, C. (2018). Examining the simple view of reading with elementary school children: Still simple after all these years. *Remedial and Special Education, 39*(5), 260–273. https://doi.org/10.1177/0741932518764833

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia*, *53*(1), 1–14. https://doi.org/10.1007/s11881-003-0001-9

Lyytinen, H., Erskine, J., Hämäläinen, J., Torppa, M., & Ronimus, M. (2015). Dyslexia—Early identification and prevention: Highlights from the Jyväskylä longitudinal study of dyslexia. *Current Developmental Disorders Reports*, *2*(4), 330-338.

Martin, A., Kronbichler, M., & Richlan, F. (2016). Dyslexic brain activation abnormalities in deep and shallow orthographies: A meta-analysis of 28 functional neuroimaging studies. *Human Brain Mapping*, *37*(7), 2676-2699.

Maughan, B. & Carroll, J. (2006). Literacy and mental disorders. *Current Opinions in Psychiatry*, *19*, 350-354. https://doi.org/10.1097/01.yco.0000228752.79990.41

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28-50.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3).

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., (pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

Molfese, D. L. (2000). Predicting dyslexia at 8 years of age using neonatal brain responses. *Brain and language*, *72*(3), 238-245.

Molfese, V. J., Molfese, D. L., & Modgline, A. A. (2001). Newborn and preschool predictors of second-grade reading scores: An evaluation of categorical and continuous scores. *Journal of Learning Disabilities*, *34*(6), 545-554.

Morabia, A. (2004). History of medical screening: from concepts to action. *Postgraduate Medical Journal*, *80*(946), 463–469. https://doi.org/10.1136/pgmj.2003.018226

Muter, V. and Snowling, M. J. (2009). Children at familial risk of dyslexia: Practical implications from an at-risk study. *Child and Adolescent Mental Health*, *14*, 37-41. https://doi.org/10.1111/j.1475-3588.2007.00480.x

National Center on Improving Literacy. (2018). State of Dyslexia. Retrieved from https://improvingliteracy.org/state-of-dyslexia

National Center on Intensive Intervention. (2018). Academic Screening Tools Chart. Retrieved from https://charts.intensiveintervention.org/chart/academic-screening

Nevada Department of Education. (2015). *Nevada's Dyslexia Resource Guide*. Retrieved from Nevada Department of Education website: http://www.doe.nv.gov/uploadedFiles/ndedoenvgov/content/Special_Education/508_DyslexiaResoGuideComplete.pdf

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (McGraw-Hill Series in Psychology) (Vol. 3). New York: McGraw-Hill.

O'Connor, R.E., & Jenkins, J. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3,* 159–197.

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. https://doi.org/10.1002/wcs.1383.

Ozernov-Palchik, O., & Gaab, N. (2016b). Tackling the early identification of dyslexia with the help of neuroimaging. *Perspectives on Language and Literacy*. Winter 2016. The International Dyslexia Association.

Pentimonti, J. M., Walker, M. A., & Edmonds, R. Z. (2017). The selection and use of screening and progress monitoring tools in data-based decision making within an MTSS framework. *Perspectives on Language and Literacy, 43*(3), 34-40.

Peterson, R.L., & Pennington, B.F. (2015). Developmental dyslexia. A*nnual Review of Clinical Psychology, 11*, 283-307. https://doi.org/10.1146/annurev-clinpsy-032814-112842

Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (measurement) applications of curriculum-based measurement in reading. *Assessment for Effective Intervention, 38*(2), 71-75.

Petscher, Y., Kershaw, S., Koon, S., & Foorman, B. R. (2014). *Testing the Importance of Individual Growth Curves in Predicting Performance on a High-Stakes Reading Comprehension Test in Florida*. REL 2014-006. Regional Educational Laboratory Southeast.

Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49*(1), 107-129.

Piasta, S. B., Petscher, Y., & Justice, L. M. (2012). How many letters should preschoolers in public programs know? The diagnostic efficiency of various preschool letter-naming benchmarks for predicting first-grade literacy achievement. *Journal of Educational Psychology, 104*(4), 945.

Poulsen, M., Nielsen, A.M. V., Juul, H., & Elbro, C. (2017). Early identification of reading difficulties: A screening strategy that adjusts the sensitivity to the level of prediction accuracy. *Dyslexia*, *23*(3), 251–267. https://doi.org/10.1002/dys.1560

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, *62*(2), 816-847.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, *15*(6), 246-253.

Pugh, K. R., Mencl, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., ... & Shaywitz, B. A. (2001). Neurobiological studies of reading and reading disability. *Journal of Communication Disorders*, *34*(6), 479-492.

Raschle, N. M., Chang, M., & Gaab, N. (2011). Structural brain alterations associated with dyslexia predate reading onset. *Neuroimage*, *57*(3), 742-749. https://doi.org/10.1016/j.neuroimage.2010.09.055

Raschle, N.M., Stering, P.L., Meissner, S.N., & Gaab, N. (2014). Altered neuronal response during rapid auditory processing and its relation to phonological processing in prereading children at familial risk for dyslexia. *Cerebral Cortex, 25*, 2489-2501. https://doi.org/10.1093/cercor/bht104

Richlan, F., Kronbichler, M., & Wimmer, H. (2013). Structural abnormalities in the dyslexic brain: a meta-analysis of voxel-based morphometry studies. *Human Brain Mapping*, *34*(11), 3055-3065.

Rimrodt, S. L., Clements-Stephens, A. M., Pugh, K. R., Courtney, S. M., Gaur, P., Pekar, J. J., & Cutting, L. E. (2008). Functional MRI of sentence comprehension in children with dyslexia: beyond word recognition. *Cerebral Cortex*, *19*(2), 402-413.

Rimrodt, S. L., Peterson, D. J., Denckla, M. B., Kaufmann, W. E., & Cutting, L. E. (2010). White matter microstructural differences linked to left perisylvian language network in children with dyslexia. *Cortex*, *46*(6), 739-749.

Rovinelli, R. J., & Hambleton, R. K. (1976). *On the use of content specialists in the assessment of criterion-referenced test item validity.* Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976).

Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). *How to evaluate a screening process: The vocabulary of screening and what educators need to know*. NY, NY: Routledge.

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18(3), 308-315*.

Schatschneider, C., Wagner, R. K., Hart, S. A., & Tighe, E. L. (2016). Using simulations to investigate the longitudinal stability of alternative schemes for classifying and identifying children with reading disabilities. *Scientific Studies of Reading, 20*(1), 34-48.

Scholastic. (2002). *CORE Phonics Survey*. New York: Consortium on Reading Excellence. Retrieved from http://www.scholastic.com/dodea/module_2/resources/dodea_m2_tr_core.pdf

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly, 30*(4), 470.

Shaywitz, S. E. (1998). Dyslexia. *New England Journal of Medicine, 338*(5), 307-312.

Shaywitz, S.E. (2014). Testimony Before the Committee on Science, Space, and Technology, United States House of Representatives. Retrieved from: https://science.house.gov/imo/media/doc/Shaywitz%20Testimony.pdf

Shaywitz, S. E., Escobar, M. D., Shaywitz, B. A., Fletcher, J. M., & Makuch, R. (1992). Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. *The New England Journal of Medicine*, *326*(3), 145-150. http://dx.doi.org/10.1056/NEJM199201163260301

Shaywitz, S.E., & Shaywitz, B.A. (2008). Paying attention to reading: The neurobiology of reading and dyslexia. *Development and Psychopathology, 20*, 1329-1349. https://doi.org/10.1017/S0954579408000631

Shinn, M., & Shinn, M. (2002). *AIMSweb training workbook: Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement.* Eden Prairie, MN: Edformation.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1-3), 83-117.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Snowling, M. J. & Hulme, C. (2012). Annual research review: The nature and classification of reading disorders–a commentary on proposals for DSM-5. *J Child Psychol Psychiatry, 53*, 593–607, https://doi.org/10.1111/j.1469-7610.2011.02495.x

Spector, P. E. (1992). Summated rating scale construction: An introduction. *Sage university papers series: Quantitative applications in the social sciences, No. 82*. Thousand Oaks, CA, US: Sage Publications, Inc.

Spencer, M., Wagner, R. K., Schatschneider, C., Quinn, J. M., Lopez, D., & Petscher, Y. (2014). Incorporating RTI in a hybrid model of reading disability. *Learning Disability Quarterly, 37*(3), 161-171.

Stanovich, Keith E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *22,* 360-407.

Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement properties of DIBELS oral reading fluency in grade 2: Implications for equating studies. *Assessment for Effective Intervention, 38*(2), 76-90.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*(1), 55-64. http://dx.doi.org/10.1207/SLDRP1501_6

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, *34*(1), 33–58. https://doi.org/10.1177/002221940103400104

Undheim, A.M. (2003). Dyslexia and psychosocial factors: A follow-up study of young Norwegian adults with as history of dyslexia in childhood. *Nordic Journal of Psychiatry, 57*, 221-226. https://doi.org/10.1080/08039480310001391

Universal Screening: K-2 reading (2017). Retrieved from: https://dyslexiaida.org/universal-screening-k-2-reading/

University of Oregon. (2018). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®)*. Eugene, OR: University of Oregon. Retrieved from https://dibels.uoregon.edu/docs/materials/dibels_8_admin_and_scoring_guide_2018.pdf

Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

VanDerHeyden, A. M., & Burns, M. K. (2017). Four dyslexia screening myths that cause more harm than good in preventing reading failure and what you can do instead. *Communique, 45*(7), 1.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wang, Y., Mauer, M. V., Raney, T., Peysakhovich, B., Becker, B. L., Sliva, D. D., & Gaab, N. (2016). Development of tract-specific white matter pathways during early reading development in at-risk children and typical controls. *Cerebral Cortex*, *27*(4), 2469-2485.

Yeatman, J. D., Dougherty, R. F., Rykhlevskaia, E., Sherbondy, A. J., Deutsch, G. K., Wandell, B. A., & Ben-Shachar, M. (2011). Anatomical properties of the arcuate fasciculus predict phonological and reading skills in children. *Journal of cognitive neuroscience*, *23*(11), 3304-3317.

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2012). Relation between CBM-R and CBM-mR slopes: An application of latent growth modeling. *Assessment for Effective Intervention, 37*(3), 147-158.

Youman, M., & Mather, N. (2013). Dyslexia laws in the USA. *Annals of Dyslexia*, *63*(2), 133–153. https://doi.org/10.1007/s11881-012-0076-2

Youman, M., & Mather, N. (2015). Dyslexia laws in the USA: an update. *Perspectives on Language and Literacy*, *41*(4), 10–18.

Youman, M., & Mather, N. (2018). Dyslexia Laws in the USA: A 2018 Update. *Perspectives on Language and Literacy*, 37–41.

Zumeta, R. O., Compton, D. L., & Fuchs, L. S. (2012). Using word identification fluency to monitor first-grade reading development. *Exceptional Children, 78*(2), 201-220.

APPENDIX A

TECHNICAL CONSIDERATIONS FOR EVALUATING DYSLEXIA SCREENERS

An organizing heuristic to ground our discussion is displayed in Figure 1 and portrays behavioral screening concepts as a hierarchy. The base consideration of a screener is the population of interest (e.g., students with dyslexia, students with language disorders), followed by the scope of assessment (e.g., what type of risk is being screened, what kind of screener is being developed or used, what is its compatibility with needs), and then statistical considerations of reliability, validity, classification accuracy. It is the cumulative aspect of these components that should then inform the decision-making process.

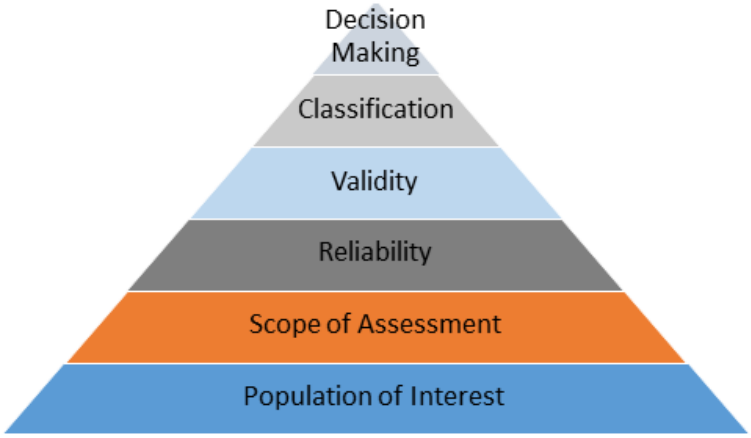

Figure 1. Foundational screening assessment considerations

POPULATION OF INTEREST. Although a seemingly intuitive part of behavioral screening, a well-defined population of interest informed by etiology and symptomatology is the critical foundation for evaluating, creating, or choosing a screener for direct services in a school setting. For example, if the intended population is, *children with dyslexia*, there is a necessary but insufficient aspect to this brief descriptor. The developmental part of the population has been identified (i.e., children) as well an identified outcome (i.e., dyslexia) that separates the individuals from other identified outcomes (e.g., students with language disorders, students with behavioral disabilities). However, it is lacking in specificity for the age range of the population and the expected symptomology associated with the identification outcome. A close correspondence between what is developmentally appropriate in content for measurement and the identified outcome safeguards against poor assessment decision making, such as screening for dyslexia risk in pre-school age children with a non-word fluency assessment or administering a letter name fluency task to fifth grade students.

SCOPE OF ASSESSMENT. This level of building or evaluating behavioral screeners is multidimensional with significant depth and is inclusive of issues such as: alignment between the operationalized outcome (e.g., what risk is being screened *for*; Keenan & Meenan, 2014) and the operationalized screener (e.g., what risk is being screened *by*; Glover & Albers, 2007), whether the screener is speed-based (e.g., time-limited) or power-based (e.g., fixed-item or computer adaptive accuracy measures), and whether single or multiple assessments should be combined (e.g., Compton et al., 2010). Assessment scope issues such as compatibility with service delivery needs, localizing screeners, and frequency of administration have been covered extensively elsewhere (e.g., Glover & Albers, 2007; Schatschneider, Petscher, & Williams, 2008).

SCREENER-OUTCOME ALIGNMENT. Alignment between how dyslexia is operationalized as an outcome (e.g., poor word reading or poor reading comprehension) and what a screener measures in content has important implications for the screening process, as not all assessments are created equally. To illustrate, even where reading comprehension assessments, such as the Gray Oral Reading Test

(GORT), Qualitative Reading Inventory – 3 (QRI), Woodcock-Johnson Passage Comprehension – 3 (WJPC), and Peabody Individual Achievement Test (PIAT) are viewed as standardized, norm-referenced tests of reading comprehension, individuals who are identified as having poor comprehension skills may vary according to which measure may be used. In a sample of 995 children, Keenan and Meenan (2014) found that for those students achieving at the lowest 10% of the distribution of each assessment, only 39%-56% of students were consistent across pairs of the assessment. That is, when categorizing students at or below the 10th percentile of the WJPC, only 39% of those students were found to be at or below the 10th percentile of the GORT. Not only can variability exist in who gets identified based on the operationalized outcome, but so too can the interplay between outcome and screener matter. Cutting and Scarborough (2006) found that the amount of variance explained in reading comprehension varied when using the same measures of decoding and oral language as independent variables but different standardized reading comprehension dependent variables. When using the Weschler Individual Achievement Test 72% of the variance was explained by the decoding and oral language predictors compared with 67% in the Gates-MacGinitie Reading Comprehension Test, and 49% in the GORT. The way a stimulus is designed to measure reading skills in the screener and the outcome assessment is critical to understand when evaluating screening assessments, as there are important implications for who is identified with dyslexia based on the alignment between the screener and the outcome.

SPEED-POWER ASSESSMENTS. The research on trade-offs between speed-based assessments, such as curriculum-based measurements (CBM) that measure speed and accuracy (i.e., fluency assessments), and power-based assessments, such as computer-adaptive assessments (CAA), is very much in its infancy. CBM in reading have long been used for universal screening due to the brevity in administration and psychometric properties of reliability and validity (Petscher, Cummings, Biancarosa, & Fien, 2013). Computer adaptive measures have more recently permeated the screening landscape as they are more reliable than fixed-item assessments at the individual level (Wainer, Dorans, Flaugher, Green & Mislevy, 2000), and leverage accuracy-based performance without regard to automaticity (Van der Linden & Glas, 2000). An important consideration for CAA is that they elevate the potential for estimating the true

correlation between assessments. That is, a known psychometric property of reliability is that a correlation between two assessments cannot exceed the square root of the product of the reliability for measures (Nunnally & Bernstein, 1994). Suppose that a researcher has a screener with a reliability of .75 and an outcome with a reliability of .85; in this case, the maximum correlation that can be estimated is $\sqrt{.75(.85)} = .798$. As the reliability of either measure changes, so does the maximum correlation. An implication for screening is that an advantage of CAA compared to CBM lies in the formers' ability to maximize student-level reliability that may then yield larger possible correlations and classification accuracy.

Direct comparisons in the predictive utility of CBM versus CAA are limited. Shapiro, Dennis, and Fu (2015) examined differential predictions of AIMSweb Mathematics CBM and the STAR-Math CAA to the Pennsylvania System of School Assessment (PSSA). Bivariate correlations between AIMSweb Concepts and Applications and the PSSA were r = .61, .24, and .49 in each of grades 3, 4, and 5 respectively; compared to the AIMSweb Computation-PSSA correlations of r = .61, .75, and .74, and the STAR-Math-PSSA correlations of r = .82, .88, and .70. Although no correlation contrast test was applied to these estimates, and one must take respective screener assessment content development, foci, and screener-outcome alignment into account, one may surmise that a potential explanatory mechanism for a stronger correlation of STAR-Math to PSSA is that the CAA produces student-level precision estimates allowing for the possibilities of estimating larger correlations. As more CAAs are commercially available for screening, greater attention is needed to their utility for screening for dyslexia risk.

Univariate or multivariate screening. The extent to which one or more screeners are necessary for dyslexia screening or broader reading risk is not a new discussion. As previously noted, a critical goal in the screening process is to ensure that false positives and false negatives are minimized so that more accurate screening results may be observed; however, one cannot simultaneously minimize both errors as it is a trade-off that test-makers make as to which of the types of errors is prioritized to minimize. Previous research has noted that univariate screening produces too many false positives (e.g., Johnson, Jenkins, & Petscher, 2010; Johnson, Jenkins, Petscher, & Catts, 2009). Conversely, research into multiple-screener methods reveals that multivariate screening models, such as

two-stage screening (Compton, Fuchs, Fuchs, & Bryant, 2006; Compton et al., 2010), four-step screening (Gilbert, Compton, Fuchs, & Fuchs, 2012), and hybrid model approaches (Schatschneider, Wagner, Hart, & Tighe, 2016; Spencer, Wagner, Schatschneider, Quinn, Lopez & Petscher, 2014) yields greater classification accuracy and longitudinal stability of classification by not only leveraging multiple screening assessments but also including measures of estimated growth on the screeners. In one of only a few existing screening studies for language impairment risk or dyslexia compared to typical word reading, Adlof, Scoggins, Brazendale, Babb, & Petscher (2017) found that a combined battery of word reading and listening comprehension was approximately equal in discriminatory power of identification (i.e., area under the curve; AUC = .79) over using only word reading (AUC = .78) in risk for language impairment for a group of second grade students. Moreover, risk of dyslexia was not improved by a combined battery (AUC = .85) compared to using only word reading (AUC = .86).

Another contextual consideration in multivariate screening is the extent to which multiple informants may improve screener decision. Despite calls from the field to include and evaluate how well teacher ratings may improve screening (e.g., Davis, Lindo, & Compton, 2007), few studies have done so. Compton et al. (2012) used a battery of universal screening measures, Tier 1 and Tier 2 progress monitoring data, teacher ratings of student attention and behavior, standardized tests of word reading and listening comprehension, and tutor ratings of attention and behavior of students in Tier 2 to evaluate how much data was necessary for screening for non-response to instruction or intervention across tiers. Results showed that adding both Tier 1 progress monitoring and teacher ratings improved AUC from .88 in a model with only screeners to .92 with added measurements. However, it is unclear the extent to which teacher ratings served as the active ingredient in the moving the AUC compared to the progress monitoring data, especially in light of research that shows the unique value of slopes in predicting outcomes above benchmark status measures (e.g., Kim et al., 2010; Petscher, Kershaw, Koon, & Foorman, 2014; Schatschneider et al., 2008; Yeo et al., 2012; Zumeta, Compton, & Fuchs, 2012).

It should be noted here that when viewing the scope of assessments, the tension of univariate or multivariate screening is related to process rather than measure. That is, screeners are developed

individually, are frequently administered as individual assessments, and have recommendations for screening at the individual assessment level. The decision as to whether univariate or multivariate scores should be used should be informed by how well those individual assessment scores may be combined in a meaningful way that collectively can improve screening beyond the utility of the individual measures. Further, given the previous sections that have been outlined related to considerations for scope of assessment, it is important that continued research evaluates alignment issues of screeners to outcomes and the impact of reliability on dyslexia risk screening.

STATISTICAL CONSIDERATIONS FOR SCREENERS. Having reviewed the population of interest and the multidimensional components related to the scope of the assessment, three of the final four components in the Figure 1 hierarchy of evaluating a screener are statistical in nature (i.e., reliability, validity, and classification accuracy). Each of these features of screener psychometrics themselves are necessary but insufficient ingredients in creating, choosing, and using a behavioral screener. In the following subsections we touch on each technical standard and raise key aspects one might be mindful of in evaluating tools.

RELIABILITY. The most basic definition of reliability is the consistency of a set of scores for a measure, yet this definition may be deceptively simplistic in the context of psychometrics due to the number of ways it can be estimated. Different forms of reliability of include internal consistency, alternate-form, test-retest, split-half, and inter-rater. Careful evaluation of the reliability of screener's scores is necessary as not all forms of reliability are created equal.

INTERNAL CONSISTENCY. Internal consistency is how well a set of item-level scores from an assessment correlate with each other. The importance of reporting this form of reliability is that one is able to quickly gauge the coherence of items for a screener and then view its potential impact on correlation and classification accuracy (see previous section on speed-power assessments). A known limitation of internal consistency is that researchers frequently report it via Cronbach's alpha, a statistic that has received criticism due to its easy-to-meet methodological assumptions and that it may be artificially inflated simply

by adding items (McNeish, 2017). When reporting internal consistency via Cronbach's alpha or alternative statistics such as omega total, Coefficient H, or the greatest lower bound, the ideal is for internal consistency to minimally exceed .80 for research purposes and .90 for clinical decision making (Nunnally & Bernstein, 1994).

ALTERNATE FORM RELIABILITY. Also referred to as parallel form reliability, screener technical reports frequently include alternate form reliability and is defined as the consistency of scores (i.e., the correlation) between two different versions of the same test. This form of reliability can be useful for characterizing the feasibility of using different forms across groups of individuals, or within a group across multiple waves of data collection. A strength in reporting alternate form reliability is that when its evidence is strong, the use of alternate forms allows practitioners to guard against practice effects or exposure effects (i.e., the likelihood for an individual to get an item right because of previous exposure to the same stimulus). A potential weakness of alternate form reliability is that the threshold for acceptable levels should be high to ensure that individual difference performances across forms are due to actual ability changes and not form effects. For example, an alternate form reliability of .70 might suggest a strong correlation but it also suggests significant non-overlap in measurement as a .70 estimate translates to only 49% shared variance in scores between two forms. Similarly, alternate form reliability of .90 points to very high overlap in the scores, but nearly 20% of the variance between the forms is unexplained. This does not by itself point to a fatality in form equivalence but speaks to a broader contextual issue that has emerged in last decade of screening research related to *if*, *when*, and *how* to adjust for lack of equivalence across forms of assessments (Christ & Ardoin, 2009; Cummings, Park, & Bauer Schaper, 2013; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Petscher & Kim, 2011; Stoolmiller, Biancarosa, & Fien, 2013).

TEST-RETEST RELIABILITY. The longitudinal consistency of scores is frequently reported where the screener is given at two, short-interval time points. Where retest reliability can be useful is in the very short time-frame of administration (e.g., 1 week) to demonstrate that the relative rank ordering of scores does not change over time. Two limitations of this form of reliability are temporal and growth-expectation factors. The former refers to the amount of space that occurs for test-retest reliability; a review of many screeners

that have been evaluated by the National Center on Intensive Intervention's academic screening tool chart show example retest spacing of 1-week, 2-week, and 4-week. The greater the amount of spacing between testing occasions, the more that maturation effects influence the strength of the correlation. In a related manner, the theoretical expectation for growth is also critical for evaluating test-retest. That is, beyond considering the retest spacing, does a researcher expect individuals in the sample to differentially change over the 1-week, 2-week, or 4-week period? To the extent that individual differences change over time, a low retest reliability may reflect such an expectation.

SPLIT-HALF RELIABILITY. Split-half reliability tests for how well one portion of the screener (e.g., odd items) correlates with another portion of the screener (e.g., even items). Although this form of reliability can provide a proxy for alternate form reliability, it is also limited due to the possibility of the manipulating how the halves are constructed in order to achieve optimal estimates (Chakrabartty, 2013).

INTER-RATER RELIABILITY. A final form of reliability worth evaluating is inter-rater reliability, and is the consistency of scores on a particular behavior between two or more raters. Inter-rater reliability is key when validating scores from observation tools such as teacher ratings of student behaviors (Anastopoulos, Beal, Reid, Reid, Power, & DuPaul, 2018) or observer ratings of oral language skills (Connor, 2019). Even within the context of direct student assessment, inter-rater reliability can be useful in understanding the extent to which differences among students in screener scores are due to administration or scoring errors. Cummings, Biancarosa, Schaper, and Reed (2014) evaluated the relation between examiner errors in scoring oral reading fluency probes and found that 16% of the variance in scores was due to examiner differences. Such findings underscore the potential importance of calibrating administrations of screeners to reduce scoring errors and misidentification.

The interplay among reliability types in creating, choosing, and using screeners for dyslexia risk is balance and purpose for evaluating reported statistics based on the need for each type. If one were to take a set of indexes, such as social security number, date of birth, and height collected data would demonstrate excellent test-retest reliability but poor internal consistency (McCrae, Kurtz, Yamagata, & Terracciano, 2011). Conversely, an assessment of stress might have good internal consistency and poor test-retest

reliability. The choice of which forms of reliability are most important for a dyslexia screener is inextricably tied to the scope of the assessment. CBMs screeners operate as speeded assessments and thus do not report information at the item level, thus, estimates of internal consistency are not provided. Instead, these types of assessments rely on alternate-form and test-retest reliability. CAA screeners operate with inherent equivalence across forms (i.e., all items are calibrated to the same scale). Accordingly, CAA reliability tends to be reported via marginal reliability (akin to internal consistency) and test-retest reliability.

VALIDITY. Just as reliability is multifaceted in nature, so is the concept of validity to the point that we may be able to provide an alphabetized and non-exhaustive sample of forms of validity that include aetiological, conclusion, concurrent, construct, content, convergent, criterion, discriminant, ecological, external, face, factor, hypothesis, in situ, internal, nomological, predictive, translational, treatment, and washback. At its core, validity is simply concerned with the extent to which something measures what it purports to measure. A word reading test should measure word reading and not receptive vocabulary. An historical perspective of validity was that three independent forms of validity existed (i.e., content, criterion, and construct validity) and could be readily interchanged (Messick, 1995). Content validity is primarily established by the consistency of expert judgments that test content is related to its described use. A classical definition of criterion validity is the simple correlation between a test score and an outcome score, and construct validity is concerned with the interpretation and use of scores (Messick, 1995). Messick (1989) sought to reconceptualize all forms of validity as forming a cohesive, unified framework of *construct* validity. This framework includes the six areas that should be evaluated to measure a test, including a screener's, construct validity.

CONTENT. Evidence for content validity includes characterizations of the content's relevance, the overall representativeness of the content (e.g., test items or stimuli), and the quality of the test items or stimuli. This form of validity is especially important when one is building an assessment, such as a screener, and is relevant to the *scope of the assessment* previously described because it provides a foundation by which score interpretations can be defended. That is, a domain that has been evaluated for content validity

via the domain's definitions, item representation, and domain relevance allow for interpretations and score use to be parsimoniously developed and defended (Sireci, 1998).

SUBSTANTIVE. A general perspective of substantive validity is that this form is established by describing the theoretical rationales that explain consistency in one's response to test items. Tasks such as rater judgment of items relative to an established taxonomy (Rovinelli & Hambleton, 1976), rater judgment of the extent to which a particular knowledge-base or skill is essential to successful item completion (Lawshe, 1975), or calculating the proportion of raters who assign an item to its theorized content (Anderson & Gerbing, 1991) have all been used to provide evidence of substantive validity.

STRUCTURAL. Structural aspects of validity are concerned with how well the structure of the assessment aligns with the construct domain and can be test via quantitative methods such as exploratory or confirmatory factor analysis.

GENERALIZABILITY. The interpretation of scores and how well they generalize across tasks, samples, and time points reflect the generalizability aspect of validity. It may be ascertained by a description of what the defined population and boundaries for that population are; the sample representativeness in the conducted study to validate the assessment; the employed design, data collection measures, procedures, and analyses within the validation study; a review of potential biases (e.g., sample selection bias or information bias) and confounds; as well as studies of replication.

EXTERNAL. External validity is concerned with quantitative evidences including convergent, discriminant, and predictive forms of validity. Convergent validity measures the degree to which scores that should be related are in fact related to each other. For example, a measure of uppercase alphabet letter knowledge should be strongly correlated with a measure of lowercase alphabet letter knowledge, and a researcher-developed measure of receptive vocabulary should be moderately to strongly correlated with a standardized measure of receptive vocabulary like the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007). Discriminant validity is characterized by how unrelated scores from two domains should be when they are expected to be unrelated. For example, a measure of alphabet letter knowledge should not be

correlated with one's intake of sugar sweetened beverages. Predictive validity is the longitudinal association between a test score at one time point and another test score at a later time point.

CONSEQUENTIAL. One of the more hotly debated forms of validity is consequential validity (Cizek, Rosenberg, & Koons, 2008), and in the area of screening for dyslexia risk it is unsurprising that this should be a hallmark of evaluating screeners. Due to the confluence of accountability testing, screening legislation, IEP provision, and instructional and intervention supports for at risk readers, there is a burden on screener developers and users to carefully take stock of implications of at risk and not at risk classifications on screeners specifically pertaining to what happens when correct decisions and decision errors occur. It is key that that score labels (e.g., high risk, moderate risk, low risk) are accurate and precise descriptors of what is being assessed (Messick, 1989), and that assessment developers and test users clearly describe, as well as possible, the potential and actual consequences of using a selected screener.