ASHA American Speech-Language-Hearing Association

AJSLP

Review Article

A Review of Screeners to Identify Risk of Developmental Language Disorder

Xue Bao,[a] Rouzana Komesidou,[a] and Tiffany P. Hogan[a]

[a] School of Health and Rehabilitation Sciences, MGH Institute of Health Professions, Boston, MA

ARTICLE INFO

Article History:
Received August 2, 2023
Revision received November 13, 2023
Accepted December 15, 2023

Editor-in-Chief: Erinn H. Finke
Editor: Anny Castilla-Earls

https://doi.org/10.1044/2023_AJSLP-23-00286

ABSTRACT

Purpose: This review aims to comprehensively summarize, compare, and evaluate screeners used to identify risk for developmental language disorder (DLD), a common learning disability that is underidentified. Screening for DLD is a cost-effective way to identify children in need of further assessment and, in turn, provides much needed supports.
Method: We identified 15 commercially available English language DLD screeners in North America. We then characterized each screener on 27 aspects in three domains, including (a) accessibility information (acronym, subtest, website, cost, materials included, publish year, examiner qualification, age range, administration time, and administration format), (b) usability features (dialect compatibility, progress monitoring function, actionable follow-up instruction, group assessment capability, and online administration availability), and (c) technical standards (the availability of a technical manual, conceptual definition, the sample size used in classification accuracy calculation, sample distribution, year of sample collection, outcome measure, sample base rate, cutoff score, sensitivity, specificity, positive predictive value, and negative predictive value).
Results: We obtained sufficient accessibility information from 14 out of 15 (93%) screeners. In contrast, none of the screeners (0%) included comprehensive usability features. Ten screeners (67%) included a range of classification accuracy (70%–100% sensitivity and 68%–90% specificity). We provided areas of strength and weakness for each screener as a quick reference for users and generated screener recommendations for five practical scenarios.
Conclusions: Our findings presented some DLD screeners that meet most standards and highlight numerous areas for improvement, including improving classification accuracy and clarifying follow-up instructions for children who are identified with DLD risk. Screening for DLD is critical to provide timely early identification, intervention, and classroom support, which in turn facilitates student outcomes.

Children with developmental language disorder (DLD) present with difficulty understanding and/or producing spoken language, not explained by a neurological condition or differential diagnosis (Bishop et al., 2017). DLD is commonly referred to as "a hidden disorder" (Nation et al., 2004) because its features are not consistently and readily apparent and are often only discovered following a low score on a language assessment (McGregor, 2020). However, currently, language assessment is not part of routine well-child visits conducted by pediatricians (Ebert et al., 2020), nor is it consistently part of universal, school-based screening (Adlof & Hogan, 2019). Signs and symptoms of DLD are less evident than other communication disorders, such as speech sound disorders and reading difficulties, which are often identified earlier by parents and teachers (Chan et al., 2022; Hendricks et al., 2019). Moreover, the presentation of DLD varies widely across different language domains (e.g., syntax, morphology, semantics, pragmatics, verbal memory; Lancaster & Camarata, 2019). This variation, with its associated strengths and weaknesses, further complicates the identification of DLD during routine interactions.

Correspondence to Tiffany P. Hogan: thogan@mghihp.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Difficulty identifying DLD has widespread impacts, in particular, given that it is a common learning disability occurring in approximately 7.58% of children (Norbury et al., 2016; Tomblin et al., 1997). Furthermore, it is associated with numerous negative impacts on academic outcomes (Cross et al., 2019; Dubois et al., 2020; see Ziegenfusz et al., 2022, for a review). In practical terms, this equates to approximately two children with DLD in each classroom of 30 students. Therefore, identifying children with DLD accurately and promptly must become a priority in school-based practice. Screening is the essential first step toward further testing to identify those who could benefit from support to mitigate long-term negative academic and functional outcomes.

Before children start formal education, one of the main approaches to identifying risk for DLD is by parent report. Pediatricians sometimes inquire about children's language development during wellness visits, but the inquiry is mainly based on parent interviews (Tanski et al., 2010). However, studies show that parents sometimes are not aware of their children's language difficulties (Calder et al., 2023; Chan et al., 2022; Hendricks et al., 2019; Porter et al., 2020). At the onset of formal education, schools play an essential role in identifying risk for conditions that affect academic achievement, including vision, hearing, language, reading, and communication-related impairments (Catts & Hogan, 2020). To do this, schools in the United States routinely administer screening measures in the early school grades. However, unlike vision, hearing, and reading, screening for DLD is not part of routine school practice (Adlof & Hogan, 2019). Instead, risk for DLD is generally driven by teacher or parent concerns and subsequent referral for speech-language pathology evaluation (Gallagher et al., 2019). Additionally, early screening often focuses on word reading abilities (Council of Chief State School Officers, n.d.), with reading skills receiving more attention and a higher likelihood of early screening as compared to oral language, under current education frameworks (Archibald, 2019). Thus, nearly half of children with DLD who develop adequate word reading skills (Adlof & Hogan, 2018) are likely to be missed during literacy screening.

Screening for language delays in young children has been a longstanding area of research, especially within monolingual, English-speaking children. For example, Klee et al. (1998, 2000) have developed an early language screening using parent questionnaires, achieving high specificity (96%) and sensitivity (91%) in identifying the risk of DLD as early as 2 years old. Recently, Ebert et al. (2020) have designed a DLD parent report screener specifically for English-speaking children aged 6–8 years old that can be completed within 2 min in primary care settings. Hendricks et al. (2019) have offered a brief, whole-classroom screener that demonstrated acceptable classification accuracy for identifying risk for DLD. Redmond et al. (2019) have further substantiated the feasibility of using English sentence repetition and past tense marking as effective indicators for identifying children at risk of language delay. Of note, the majority of studies of risk for DLD has included monolingual English-speaking children, though some work has focused on identification of risk for DLD in multilingual children (Hunt et al., 2022).

Despite these promising findings, implementing DLD screening measures in regular educational practice is not common. A significant barrier has been the limited access clinicians and educators have to these research-based measures. Moreover, even when screening tools developed specifically for research within this demographic are available, school-based guidance often prioritizes the use of commercially available measures (Palincsar et al., 2001). It is not enough to include commercially available DLD screeners in practice and assume they have met high standards for use. Commercially available screening measures should meet specific criteria to be deemed useful for identifying risk for DLD in the study of Johnson et al. (2009). For example, in the medical field, screeners must meet specific criteria to be classified as high quality. This includes the capacity to detect a high proportion of a disease, affordability, ease of administration, validity, reliability, reproducibility of results, provision of follow-up options following a positive result, and wide availability to the designated population (Institute for Quality and Efficiency in Health Care, 2006). In this review article, we evaluate commercially available DLD screeners using standards from the medical field to determine their quality. Specifically, we comprehensively summarize, compare, and evaluate accessibility information, usability features, and technical standards domains of the screeners. First, we collected "accessibility information." This included available subtest(s), website, cost, publication year, examiner qualification, age range, administration time, and administration format. This information provides general information that can inform initial decisions about a screener's relevance for specific school contexts. Accessibility information serves as the foundational layer of any screener evaluation. By assessing parameters such as administration time, age range, cost, and the year of publication, educational institutions and professionals can promptly determine the screener's feasibility and relevance for their specific setting or child demographic (Gaab & Petscher, 2022). This information acts as a preliminary filter for making informed decisions about the applicability of a screener. Additionally, the screener's publication year offers information about its relevance to contemporary settings.

Then, we examined "usability features": dialect compatibility, progress monitoring function, actionable follow-up instructions, group assessment capability, and online administration availability. Usability features help determine whether a screener fits a specific purpose and addresses the needs of users (Lu et al., 2022). Every educational setting, whether a district school or a specialized institution, grapples with unique challenges and demographics (Gaab & Petscher, 2022). Elements such as dialect compatibility help mitigate the chance that children will fail a screener for linguistic variation instead of risk for DLD (Hendricks & Adlof, 2017). In today's digital age, capabilities such as online administration and group assessments fit into the logistical and technological realities of contemporary education (Law & Heintz, 2021). However, usability goes beyond mere operation; it ensures the translation of screener results into meaningful outcomes. Screeners equipped with actionable follow-up instructions not only flag concerns but also provide guidance for follow-up assessments and interventions (Sansavini et al., 2021). Moreover, a screener with a built-in progress monitoring function provides further utility to track student progress over time (Hadley & Mendez, 2021; Petersen & Spencer, 2014).

Lastly, we analyzed "technical standards," including the presence of a technical manual, conceptual definition of the disorder, sample size used for the classification accuracy calculation, sample distribution, year of sample collection, outcome measure, sample base rate, cutoff score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Technical standards are critically important to ensure a screener's accuracy for the primary purpose of screening—to accurately determine risk (Petscher & Suhr, 2022). Elements such as a comprehensive technical manual and a clear conceptual definition of the disorder provide assurance to educators and clinicians about the rigorous standards underpinning the tool (Plante & Vance, 1994). However, credibility alone is not sufficient. Metrics such as sensitivity, specificity, PPV, and NPV are vital indicators of a screener's ability to indicate risk (Shreffler & Huecker, 2023). Furthermore, the sample distribution, size, and the year of collection help determine the screener's relevance to specific populations (Martínez-Mesa et al., 2016; Suresh et al., 2011). Additionally, clarity on the outcome measure and the base rate of DLD in the test sample show a screener's benchmarks for determining risk (Frost et al., 2007; Meisels, 1989). We use this information in four practical scenarios to illustrate how it can inform DLD screening. Our goals are to (a) provide clinicians and educators with a wide range of information to help them choose DLD screeners for specific contexts and (b) identify areas that need improvement for future DLD screener development.

Method

For this review, we first selected commercially available DLD screeners by determining search key words, choosing search databases, and setting inclusion and exclusion criteria. We then characterized each screener according to accessibility information, usability features, and technical standards.

Screener Selection Process

We utilized an adapted version of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to carry out the present review (Page et al., 2021). The process was depicted in Figure 1 and was carried out in four distinct phases: screener identification, first round, second round, and third round. All authors partook in the screener identification and selection procedure. A total of 15 DLD screeners were chosen for review post the application of inclusion and exclusion criteria (see Table 1).

Identification Phase

Given the extensive variation in terminology used by screener developers to denote DLD (Georgan et al., 2023), the key word "language screening test" was employed for screener identification. The identification search spanned major test publishers' websites including Brookes Publishing, ProEd, Pearson Assessments, Super Duper Publications, Western Psychological Services, and Ventris Learning. Additionally, expert advice was sought, and searches on Google.com were conducted. The initial identification was completed in September 2020, with follow-up identifications in November 2020, January 2021, April 2022, and July 2023 to encompass recently published screeners. A total of 260 potentially relevant measures were identified, though a significant number were either tangentially related or irrelevant to the focus of our review.

First Round

The first round entailed the removal of redundant screeners. In cases where multiple versions of a screener existed (e.g., Clinical Evaluation of Language Fundamentals–Fourth Edition [CELF-4] and CELF-5), the most recent version was selected. Out of the 260 initially identified tests, eight were excluded primarily due to redundancies and outdated versions, leaving 252 screeners for further examination.

Second Round

In the second round, we started with 252 screeners from the initial identification phase. Due to the broad range of key words and filters used in screener

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses chart showing the screener selection and screening process.

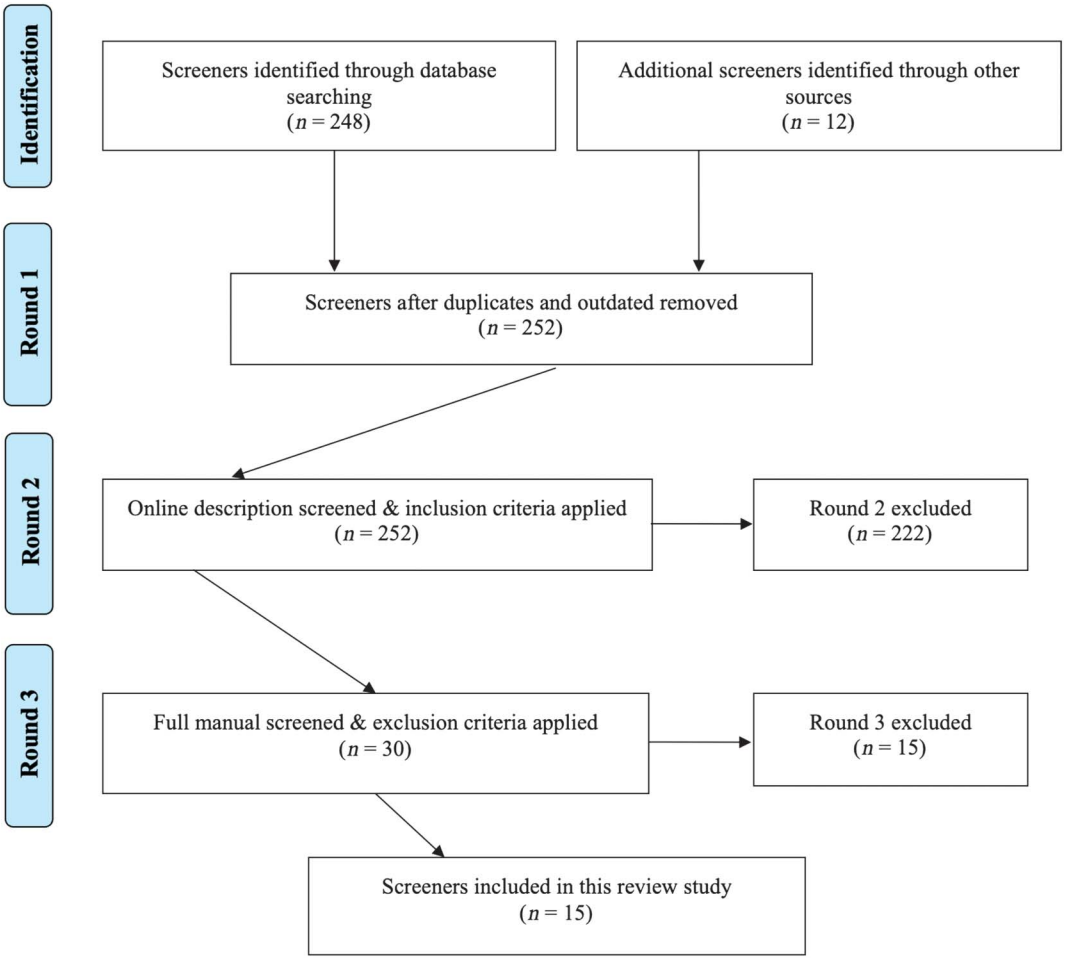


Table 1. The included developmental language disorder screeners.

Name of screener
Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test (Bayley & Aylward, 2019)
Children's Communication Checklist–2: United States Edition (Bishop, 2006)
Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test (Wiig et al., 2013, 2021)
Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition (Prizant & Wetherby, 2002, 2022)
CUBED Narrative Language Measures (Petersen & Spencer, 2016)
Language Use Inventory (O'Neill, 2009)
Preschool Language Scales–5 Screening Test (Zimmerman et al., 2012)
Quick Interactive Language Screener (Golinkoff et al., 2017, 2022)
Speech and Language Screener Quick Take Along (Snyders & Kjesbo, 2014)
Student Language Scale Screener for Language & Literacy Disorders (Nelson et al., 2018)
Diagnostic Evaluation of Language Variation–Screening Test (Seymour et al., 2003)
Early Screening Inventory, Third Edition (Meisels & Marsden, 2019)
Predictive Early Assessment of Reading and Language Screener (Petersen & Spencer, 2014)
Rice/Wexler Test of Early Grammatical Impairment (Rice & Wexler, 2001)
Well Screening, Research Edition (Ekelman et al., 2021)

identification, a majority of the results were either tangentially related or not directly pertinent to our review's focus, leading to a large pool of irrelevant results from various publishers and search engines. To address this, we examined screener descriptions on official publisher websites, applying specific criteria to refine our selection: (a) screeners published in English in North America, (b) target age range from birth to 21 years, and (c) mention of the terms "screening" or "screen" in the screener title or in its online description. Consequently, we excluded 222 tests/screeners, mostly for not meeting the stipulated criteria, leaving 30 tests for further review.

Third Round

In the third round, the screener manuals were meticulously reviewed with the following criteria: (a) exclusion of screeners published more than 20 years prior to the review initiation in 2021, (b) exclusion of screeners solely focusing on phonological abilities to maintain our emphasis on DLD (Catts et al., 2015), (c) priority to screeners from commercial publishers, and (d) consideration only for screeners clearly stating screening as their primary purpose within the manuals. The acknowledgment of the Flynn effect (Baker et al., 2015) led to the emphasis on updated norms, ensuring the screeners accurately mirror the current abilities of children. The comprehensive offline review of the complete manuals of the remaining 30 tests allowed for a more in-depth examination. Following this review, 15 tests were finalized for inclusion.

Data Extraction

We extracted data from three sources: screener manuals, peer-reviewed studies by the screener authors, and direct communication with the screener authors. To obtain accessibility information, usability features, and technical standards, we first examined the entire screener kit, primarily including stimulus booklets and items, record forms, administration manuals, and technical manuals. When we collected information on technical standards, we found that some screener developers and researchers investigated the classification accuracy of the screeners in a separate study. In those cases, we included relevant information published in peer-reviewed studies by the screener authors. In the case of missing information, after examining these two sources, the first author contacted the screener developers directly to obtain missing information. Supplemental Material S1 displays detailed data collected for each screener in the form of a "fact sheet." To maximize access to screener information by the general public, we have included sources (i.e., webpage addresses and page numbers) in each fact sheet. This fact sheet has been uploaded to OSF Preprints for public access (https://doi.org/10.31219/osf.io/cajp5).

Screener Evaluation

To summarize and evaluate the characteristics and quality of the DLD screeners, we categorized 27 aspects under three domains: accessibility information, usability features, and technical standards. Table 2 provided operational definitions of all 27 aspects.

Accessibility Information

Accessibility information included the screener acronym, subtest(s), website, cost, year of publication, examiner qualifications, age range, administration time, and administration format. The purpose of the accessibility information was to provide a general description of the screener. Furthermore, this information informed the determination of suitability for use in specific contexts (see Table 2).

If the screener targeted multiple domains (i.e., cognition, motor, social), only the subtest aimed at screening language ability was included. Examiner qualification was a relevant factor for screening selection. Many assessment publishers used a tiered system to determine the examiner qualification, representing the minimum qualification required to use a screener. In line with the recommendations from the standards for educational and psychological testing (Pitts & Naumenko, 2016) and the American Psychological Association's guidelines for test user qualifications (Turner et al., 2001), there were three commonly used levels of examiner qualification: Level A required no qualification in purchasing the test; Level B required a master's degree or equivalent clinical certifications; Level C required a doctoral degree or equivalent professional certifications. While Level A tests allowed for a more comprehensive user population and require fewer result interpretation skills, Level B tests gave test users more professional and comprehensive results in examinees' performance, which likely drove publishers to require higher test administration and interpretation skills.

Ideally, a screener has been quick, inexpensive, and easy to administer; therefore, we summarized the time needed to administer each screener (Herman, 2006). The administration time did not include the time needed to interpret screener results. Most screeners provided a range of administration times in their manual. To compare the administration time of all the included screeners, we took the longest time when the manual reported a range of times. The last aspect of accessibility information we evaluated was the administration format. Screeners could be administered in various formats, including but not limited to direct assessment, indirect assessment (information regarding the examinee's performance was collected via observation or caregiver/teacher report), and mixed (direct and indirect assessment in the DLD screener).

Table 2. The operational definition of the aspects used in the screener evaluation.

Name of the aspects	Operative definition and standard
Accessibility information	
Acronym	The official short name of the developmental language disorder (DLD) screener.
Subtest	For the DLD screeners that measure skills in multiple domains (i.e., motor, cognitive, communication, etc.), the names of the subtests used to screen language skills.
Website	The website address to access information for the screener.
Cost	The cost to purchase a complete kit of the screener, usually including stimulus books, manuals, and record forms.
Publication year	The year when the latest version of the DLD screener was published.
Examiner qualification	The training and expertise screener user must have to administer the DLD screener and interpret the results.
Age range	The age of the examinees who can be screened using the DLD screener.
Administration time	The estimation of time to complete the DLD screener.
Administration format	The format that the DLD screeners are administered.
Usability features	
Dialect compatibility	Instructions written in the screener manual explaining the aspects examiners should consider when they administer the DLD screener and interpreting the screener results of an examinee who communicates using English dialects.
Progress monitoring function	Whether the DLD screener can measure individual changes in the risk level of DLD over time.
Actionable follow-up instruction	Whether the DLD screener provides clear and actionable written instructions on recommendations for examinees screened as at risk of DLD in the screener manual.
Group assessment capability	Whether the examiner can administer the DLD screener to a group of examinees ($\geq$ 2).
Online administration availability	Whether the DLD screener can be administered remotely online.
Technical standards	
Technical manual	Whether the DLD screener has a technical manual that includes the psychometric properties of the screener.
Conceptual definition	The definition written in the screener manual of the disorder(s) that the DLD screener is designed to measure.
Sample size used in classification accuracy calculation	The total number of participants used to calculate sensitivity, specificity, positive predictive value, and negative predictive value of the DLD screener.
Sample distribution	Distribution of race or ethnicity groups of the sample in the standardization research
Year of sample collection	The year when the normative/standardization research data were collected
Outcome measure	Measurements used to determine the status of the sample in the standardization research.
Sample base rate	A predetermined percentage of the designated population that has the condition. If not provided by the screener, the base rate will be the percentage of children with DLD in the sample.
Cutoff score	The minimum scores the examinee must achieve to be considered having typical language development.
Sensitivity	The percentage of people at risk of DLD correctly identified by the DLD screener.
Specificity	The percentage of people not at risk of DLD correctly identified by the DLD screener.
Positive predictive value	The percentage of people at risk of DLD among the people identified as at risk of DLD by the screener.
Negative predictive value	The percentage of people not at risk of DLD among the people identified as not at risk of DLD by the screener.

Usability Features

Usability features included the availability of special features, such as dialect compatibility, progress monitoring, actionable follow-up instructions, group assessment capability, and online administration availability. It was essential to clarify the distinction between dialect compatibility and the inclusion of diverse racial and ethnic groups in standardization research samples. By "dialect compatibility," we meant that the screener has been designed with explicit provisions to accommodate children who may speak one or more dialects that differ from the normative research sample. This encompassed the screener having clear guidelines on how to administer the test to these children, how to accurately interpret their responses, and how to effectively communicate the results to the children and their caregivers. Importantly, screeners that were not dialect-compatible are at a heightened risk of over- or underidentification. Given the varied dialects in the United States and a growing number of multilingual children (Gatlin-Nash et al., 2023; Rose, 2023), the implications of such misidentification are significant and warrant caution. As for progress monitoring, although screeners

could be used repetitively, it was not recommended to use the same screener to monitor risk for DLD over time. Without clear guidance and psychometrics related to progress monitoring, it would be hard to distinguish whether children improve their language skills due to maturation or familiarity with screener items.

Technical Standards

We first evaluated if a technical manual or chapter was present in the screener. We then examined how each screener defined "language disorder." DLD has been referred to using many terms, such as specific language impairment (SLI), primary language impairment, language learning impairment, language disorder, developmental delay, and specific learning disability (Bishop, 2017). Even though these terms were often used interchangeably, the specific impairments represented by these labels varied. For example, one common label was SLI; to qualify as SLI, a person needed to have typical nonverbal IQ scores. However, the definition of "normal" varied by study (Gallinat & Spaulding, 2014). In this study, we use the definition described by the Criteria and Terminology Applied to Language Impairments Synthesizing the Evidence (CATALISE) Consortium as the reference for the definition of DLD (Bishop et al., 2017). The CATALISE Consortium defined DLD as "cases of language disorder with no known differentiating condition, including intellectual disability" (Bishop et al., 2017).

Next, we evaluated the sample used in the standardization or normative research, including the sample size used in classification accuracy calculation, sample distribution, and year of sample collection. Sample size and distribution allowed us to examine the ability to generalize the norms from the standardization research to the population. The year of sample collection was used to examine whether the psychometric standards and the norms concluded from the sample represent children's abilities now. This was an important consideration because, as mentioned earlier, the abilities of children measured by standardized tests have been shown to increase over time, that is, the Flynn effect, which refers to the increase of average human intelligence over time (Baker et al., 2015). Because of this, commonly administered psychoeducational tests, such as the Wechsler Intelligence Scale for Children (Wechsler, 2003), update their norms based on new samples periodically (Grégoire et al., 2016). Taking this into consideration, we selected screeners published within the last 20 years (Trahan et al., 2014).

We also evaluated the outcome measure used to determine the risk for DLD. Because each screener included participants with and without DLD, the procedures used to diagnose participants were essential in determining the composition of the sample. We then summarized the screener's sensitivity and specificity with the given cutoff score. Sensitivity is referred to the screener's ability to identify people with DLD and is also referred to as the true positive rate. For example, if a screener has a sensitivity of 80%, then 80% of children at risk of DLD are correctly identified by the screener as having risk. Specificity refers to the proportion of people not at risk of DLD who are determined to be not at risk by the screener. This is also referred to as the true negative rate. For example, if a screener has a specificity of 80%, the screener will identify 80% of children without DLD as not at risk. The calculation of sensitivity and specificity is provided in the following equations, and relevant terminology is explained by the two-by-two table depicted in Figure 2.

Figure 2. Definitions of statistical metrics used in this research. DLD = developmental language disorder.

		Condition of the Person		
		With DLD	*Without DLD*	
Screener Result	*At Risk*	(TP) True Positive	(FP) False Positive	Positive Predictive Value (PPV) $\frac{TP}{TP+FP}$
	Not at Risk	(FN) False Negative	(TN) True Negative	Negative Predictive Value (NPV) $\frac{TN}{FN+TN}$
		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	

$$\text{Sensitivity} = \frac{\text{(TP) True Positive}}{\text{(TP) True Positive} + \text{(FN) False Negative}} \quad (1)$$

$$\text{Specificity} = \frac{\text{(TN) True Negative}}{\text{(FP) False Positive} + \text{(TN) True Negative}} \quad (2)$$

Sensitivity and specificity are statistical measures of classification accuracy of screeners. When multiple measures of classification accuracy were presented in a screener manual, according to varied cut-points, we presented the highest values. The criteria adopted in the current study, as suggested by Plante and Vance (1994), primarily pertain to diagnostic testing. According to these standards, tests with sensitivity and specificity at or above .9 have been deemed to have good classification accuracy. Those within the range of .8–.89 have been seen as having fair accuracy, and tests with rates below .8 are considered to have an unacceptable misidentification rate (Plante & Vance, 1994). However, it was essential to highlight the distinction between diagnostic tests and screeners. While diagnostic tests aim to conclusively determine whether a child has DLD, screeners primarily identify risk for impairment and determine the need for further evaluation (Herman, 2006). For screeners, the crucial factor was ensuring that children marked as not at risk are genuinely without DLD because those children are not likely to receive further evaluation. As such, screeners can, in practice, tolerate slightly lower specificity. The primary aim was not definitive diagnosis but preliminary identification of risk. Risk identified by a screener would be followed up with progress monitoring or more comprehensive diagnostic testing to confirm or refute the presence of DLD. The base rate was another critical psychometric property used to evaluate the quality of a test. Base rate was a concept that has been often applied in the medical field. It referred to the predetermined percentage of the designated population with the condition (Fischhoff & Bar-Hillel, 1984). Base rates could vary depending on the amount of children with DLD in a specific sample. For example, in the general population, the base rate of DLD has been equal to the prevalence of the disorder, approximately 7.58% (Norbury et al., 2016; Tomblin et al., 1997). In normative research, the concentration of individuals at risk for DLD might be higher since many test developers recruit a matched sample, in which 50% of the sample has DLD and the other 50% does not. Base rate has been an essential influencer of how the classification accuracy fluctuates in the general population's base rate. Inflated base rates could lead to inflated classification accuracy, referred to as the base rate fallacy (Bar-Hillel, 1980). The base rate fallacy referred to a phenomenon when people prefer specific information (such as predictive power measured in a 50% matched sample in our cases) rather than considering the classification accuracy in the population, according to prevalence statistics. Base rates of DLD could vary substantially based on specific contexts, and this variation could influence the classification accuracy of screeners. Consider the following example: School A represented a typical school mirroring the general population. Given the prevalence of DLD, we anticipated that approximately 7.58% of its students would have the disorder. However, if a screener developed using a matched sample (with 50% controls and 50% DLD) was employed here, it might be predisposed to overidentify students. This meant that, rather than identifying close to the expected 7.58%, the screener might classify a significantly higher percentage, say 30% or even 40% of the students as having DLD. This would result in a large overestimation. Conversely, School B, specializing in language support, had a concentration of students at risk for DLD—let's say 40% of its students have the disorder. If the same screener, developed with the 50% matched sample, was applied here, it might produce more accurate screening results, classifying close to 40% of the students as having DLD. However, this was not because the screener is accurate; it was because the base rate in this specific context aligns with the screener's sample. This phenomenon exemplifies the base rate fallacy and emphasized the importance of developing screeners that reflect the base rate of DLD.

In the past, the base rate was seldom mentioned in research on language assessments, primarily because of the lack of comprehensive prevalence data on childhood language disorders (Kennedy et al., 1997). Here, we used the population prevalence of 7.5% provided by Norbury et al. (2016) as the reference for the population base rate. We used 20% as the reference for the screening base rate. We based this decision on best practices for tiered instruction and intervention in the response-to-intervention model (RTI). In an RTI model, approximately 20% of the students have been identified as at risk and as such receive additional support (Shapira, n.d.). RTI has been widely used in the education system (Adlof & Hogan, 2019) and involved monitoring progress on academic skills and providing tiers of supports to those who were not making progress over time (VanDerHeyden et al., 2007).

When assessing the accuracy and utility of screeners, especially those developed under varying base rates, it was crucial to go beyond mere sensitivity and specificity. One of the most direct ways to understand the real-world implications of these biases was by examining the PPV and NPV of each test. Unlike sensitivity and specificity, which did not fluctuate with the base rate, PPV and NPV changed when the base rate changed (Baldessarini et al.,

1983). PPV and NPV showed the likelihood of a DLD screener correctly identifying people with or without DLD (Baldessarini et al., 1983). PPV was the probability that, given a positive test result, a child truly had DLD. Conversely, NPV was the probability that a child did not have DLD given a negative test result. If a DLD screener had a PPV of 80%, 80% of children marked "at risk" by this screener had DLD. Similarly, NPV represented the DLD screener's ability to identify people who were not at risk of DLD (Baldessarini et al., 1983). On the other hand, if a screener had an NPV of 80%, 80% of the people marked as "not at risk" by the DLD screener did not have the disorder in real life. Because PPV and NPV were associated closely with the prevalence of a disorder, this study calculated and compared PPV and NPV using the sample base rate provided by the screeners (ratio of people with DLD and people without DLD in the sample) and a 20% screener base rate (DLD at risk rate). Using the sensitivity, specificity, and base rates of the included DLD screeners, we calculated the ratio of people in the true positive, true negative, false positive, and false negative groups. Then using the ratios, we calculated the PPV and NPV with the following formulas:

$$\text{PPV} = \frac{(\text{TP})\ \text{True Positive}}{(\text{TP})\ \text{True Positive} + (\text{FP})\ \text{False Positive}} \quad (3)$$

and

$$\text{NPV} = \frac{(\text{TN})\ \text{True Negative}}{(\text{TN})\ \text{True Negative} + (\text{FN})\ \text{False Negative}} \quad (4)$$

Results

After completing the screener selection and evaluation process, we extracted information from each of the 15 selected DLD screeners (see Table 1), including the Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition (CSBS DP; Prizant & Wetherby, 2002), the CUBED Narrative Language Measures (NLM; Petersen & Spencer, 2016), the Language Use Inventory (LUI; O'Neill, 2009), the Preschool Language Scales–5 Screening Test (PLS-5 Screening Test; Zimmerman et al., 2012), the Quick Interactive Language Screener (QUILS; Golinkoff et al., 2022), the Speech and Language Screener Quick Take Along (Snyders & Kjesbo, 2014), the Student Language Scale Screener for Language & Literacy Disorders (SLS Screener; Nelson et al., 2018), the Diagnostic Evaluation of Language Variation–Screening Test (DELV–Screening Test; Seymour et al., 2003), the Early Screening Inventory, Third Edition (ESI-3; Meisels & Marsden, 2019), the Predictive Early Assessment of Reading and Language Screener (PEARL; Petersen & Spencer, 2014), the Rice/Wexler Test of Early Grammatical Impairment (TEGI; Rice & Wexler, 2001), and the Well Screening, Research Edition (Ekelman et al., 2021). We summarized this information in Supplemental Material S1, which provides information as to where the information could be accessed (i.e., official website, technical manual, and peer-reviewed manuscripts).

Accessibility Information

Using the information provided on the official website and in the screener manual, we summarized the accessibility information in Table 3. Most of the DLD screeners were constructed with the sole purpose of screening for DLD. However, seven of the included (46%) screeners screen for other skills, such as reading, cognition, and motor disorders. For these seven screeners, we only reviewed the language-related subtests. This impacts cost-effectiveness (i.e., screeners that include more than a language focus cost more) and the usage of the screener (i.e., the cost may be acceptable because the screeners can be used more broadly). Out of the 15 screeners in this study, six (40%) were at zero or low cost (less than $100.00) to purchase the entire screener kit (CSBS DP, NLM, Speech and Language Screener Quick Take Along, SLS Screener, TEGI, Well Screening). In addition, the screener kit and manuals of the CSBS DP (checklist and scoring), the NLM, and the TEGI could be downloaded from their official website for free. On the other hand, the cost to purchase the entire kit of the LUI, the PLS-5 Screening Test, the QUILS, the DELV–Screening Test, and the PEARL was within the medium range (between $100.00 and $200.00), and the cost of the Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test (Bayley-4 Screening Test), the Children's Communication Checklist–2: United States Edition (CCC-2), the CELF-5 Screening Test, and the ESI-3 exceeded $200.00, with the CELF-5 Screening Test being the most expensive screener to purchase at $299.00. Furthermore, four screeners (27%) were published before 2010, and the rest were published after 2010. Among all the included screeners, the TEGI was published the earliest (in 2001), and the Well Screening was published the latest (in 2021). Among all the screeners we reviewed, eight of them (53%) require Level A qualifications, and the other half require Level B qualifications, providing access to screener users at different levels.

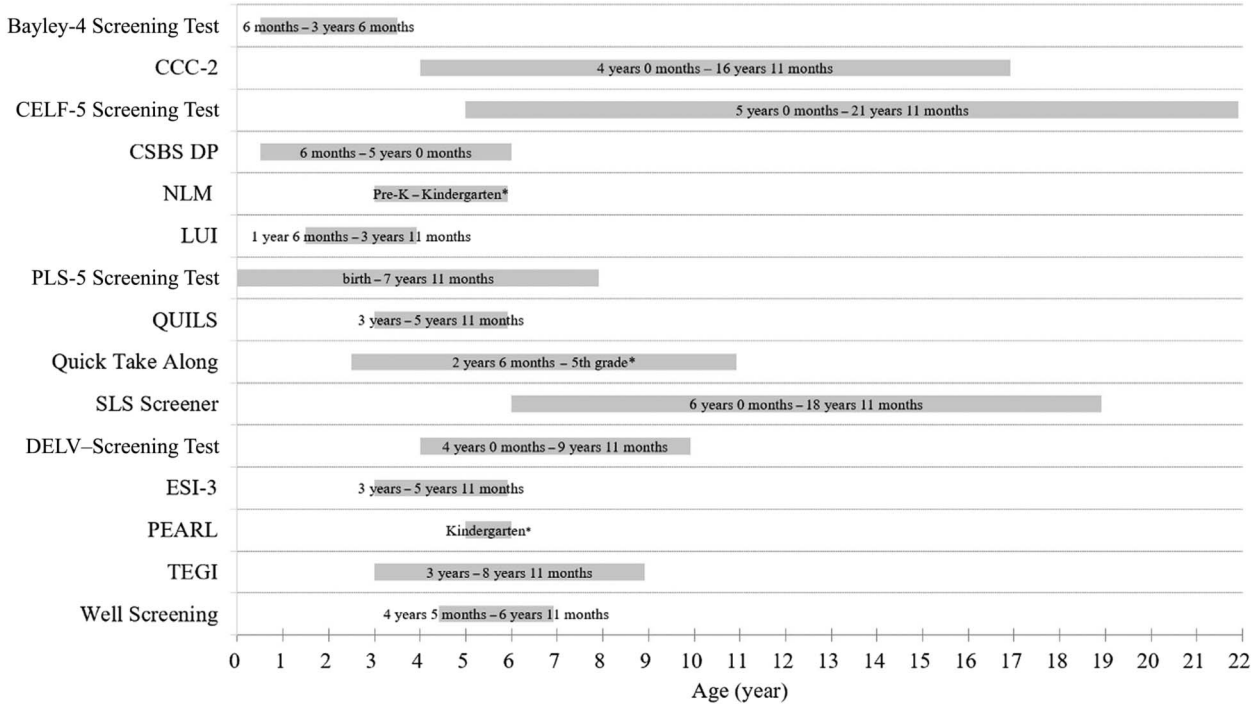
Figure 3 showed the age range distribution of the screeners, which illustrates that the DLD screeners included in this study cover a wide range of ages, from birth to 21 years of age. The majority (87%) had their administration time less than 20 min. The SLS Screener

Table 3. Summary of the accessibility information.

Name of screener	Acronym	Subtest	Cost of the screener kit (in USD)	Publication year	Examiner qualification	Age range (years;months)	Administration time (min)	Administration format
Bayley Scales of Infant and Toddler Development–Fourth Edition Screening Test	Bayley-4 Screening Test	Language subtest	$265.50	2019	Level B	0;6–3;6	15	Mixed
Children's Communication Checklist–2: United States Edition	CCC-2	N/A	$219.50	2006	Level B	4;0–16;11	5–10	Indirect
Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test	CELF-5 Screening Test	N/A	$299.00	2013	Level B	5;0–21;11	15	Mixed
Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition	CSBS DP	Infant–Toddler Checklist for Screening	Free (checklist and scoring) $399 (entire kit)	2002	Level A (checklist) Level B (questionnaire and behavior sample)	0;6–6;0 (estimate chronological age, 0;6–2;0 functional communication age)	5–10	Indirect
CUBED Narrative Language Measures	NLM	Listening subtest	Free	2016	Level A[a]	Pre-kindergarten to kindergarten	5	Direct
Language Use Inventory	LUI	N/A	$165	2009	Level A	1;6–3;11	20–30	Indirect
Preschool Language Scales–5 Screening Test	PLS-5 Screening Test	N/A	$180	2012	Level B	0;0–7;11	5–10	Mixed
Quick Interactive Language Screener	QUILS	Vocabulary, Syntax, Process	$139.95 (annual subscription)	2017	Level A	3;0–5;11	15–20	Mixed
Speech and Language Screener Quick Take Along	N/A	N/A	$13.95	2014	Level A	2;6–10;11 (5th grade)	N/A	Mixed
Student Language Scale Screener for Language & Literacy Disorders	SLS Screener	N/A	$79.95	2018	Level A	6;0–18;11	3	Indirect
Diagnostic Evaluation of Language Variation–Screening Test	DELV–Screening Test	Diagnostic Risk Status	$160	2018	Level B	4;0–9;11	15–20	Direct
Early Screening Inventory, Third Edition	ESI-3	Language and Cognition Domain	$203.5 (3;0–4;5 kit) $203.5 (4;6–5;11 kit)	2019	Level A	3;0–5;11	15–20	Mixed
Predictive Early Assessment of Reading and Language Screener	PEARL	N/A	$139.00	2014	Level A[a]	Kindergarten	5–15	Direct
Rice/Wexler Test of Early Grammatical Impairment	TEGI	N/A	Free	2001	Level B	3;0–8;11	10	Direct
Well Screening, Research Edition	N/A	Receptive Language; Expressive Language	$99.99	2021	Level B	4;5–6;11	20	Direct

[a]Training is needed to administer the screener.

Figure 3. The age range of screeners. School ages referred to the local standards of the United States. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; CCC-2 = Children's Communication Checklist–2: United States Edition; CELF-5 Screening Test = Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test; CSBS DP = Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test; QUILS = Quick Interactive Language Screener; Quick Take Along = Speech and Language Screener Quick Take Along; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; ESI-3 = Early Screening Inventory, Third Edition; PEARL = Predictive Early Assessment of Reading and Language Screener; TEGI = Rice/Wexler Test of Early Grammatical Impairment.



had the shortest administration time (3 min). The administration time for the LUI was relatively long compared to other screeners (i.e., up to 30 min). Additionally, screeners that assess children at a very young age (i.e., birth to 12 months old) typically involved more observation items than other screeners. These observation items were considered indirect items. Four screeners were exclusively parent/teacher interview questionnaires.

Usability Features

Table 4 presented the available features for each DLD screener. Only 20% of the screeners featured dialect compatibility. In addition, the CELF-5 Screening Test and the DELV–Screening Test had instructions and considerations for administration and interpretation when used with children who speak multiple dialects of English, including African American English, Appalachian English, Southern White English, Spanish-Influenced English, and Chinese-Influenced English.

The CSBS DP, the NLM, and the Well Screening used multiple different screener forms to monitor progress. As a DLD screening tool, the NLM had 16 progress monitoring forms designed to track students' performance from preschool to third grade. Test items of the NLM varied by grade level so that students' performance could be interpreted and compared across time. These screeners had progress monitoring built into their measure with multiple forms to neutralize the influence of maturation and screener familiarity (Dexter & Hughes, n.d.).

Nine of the included screeners (60%) have actionable follow-up instructions for use after the screener was administered, written in their screener manuals. All screeners with follow-up instructions recommended full diagnostic assessments if the screener showed that the child was at risk of DLD. Interestingly, in all screeners that had a diagnostic or norm-referenced assessment in the same series, their developers recommended using the diagnostic screeners in the same series, such as the Bayley-4 Screening Test, the CELF-5 Screening Test, the CSBS DP, the PLS-5 Screening Test, the SLS Screener, and the DELV–Screening Test. Additionally, none of the included screeners (0%) had the capacity to be group administered. Of the 15 measures, five (33%) could be administered

Table 4. Summary of the usability features of the developmental language disorder screeners.

Screener	Dialect compatibility	Progress monitoring function	Actionable follow-up instruction	Group assessment capability	Online administration availability
Bayley-4 Screening Test			✓		
CCC-2					
CELF-5 Screening Test	✓		✓		
CSBS DP		✓	✓		
NLM		✓			
LUI					✓
PLS-5 Screening Test			✓		✓
QUILS	✓		✓		✓
Quick Take Along					
SLS Screener			✓		
DELV–Screening Test	✓		✓		
ESI-3			✓		✓
PEARL					
TEGI					
Well Screening		✓	✓		✓

Note. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; CCC-2 = Children's Communication Checklist–2: United States Edition; CELF-5 Screening Test = Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test; CSBS DP = Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test; QUILS = Quick Interactive Language Screener; Quick Take Along = Speech and Language Screener Quick Take Along; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; ESI-3 = Early Screening Inventory, Third Edition; PEARL = Predictive Early Assessment of Reading and Language Screener; TEGI = Rice/Wexler Test of Early Grammatical Impairment.

online: LUI, PLS-5 Screening Test, QUILS, ESI-3, and Well Screening.

Technical Standards

The primary source of technical standards for each was the screeners' technical manuals. Fourteen of the screeners (93%) included a published, stand-alone technical manual or an individual section for technical information. Other resources of technical standards included peer-reviewed studies published by the authors and direct communication with the authors. Table 5 summarizes the technical standards of the screeners. In addition, sensitivity, specificity, sample base rate, PPV, and NPV were summarized separately in Figures 4 and 5.

Two screeners (13%) included children with an SLI diagnosis in their clinical sample. Six screeners (40%) used "language disorder" as a label. The rest of the screeners used language impairment (two screeners), communication and language delay, expressive and receptive language disorder, or language difficulties as the target disorder. The ESI-3 and the Speech and Language Screener Quick Take Along (13%) did not specify the terminology.

There were two primary outcome measures in determining the clinical group: clinical diagnosis and a diagnostic test or a battery of tests (see Table 5). Clinical diagnosis referred to the diagnosis or the label sample participants received from medical doctors, psychologists, clinicians, and school speech-language pathologists. Other screeners used multiple measures. For example, the LUI used the DELV–Norm Referenced, the CELF Preschool-2, the CCC-2, and parent interviews as the outcome measures.

Furthermore, inclusionary criteria to determine the clinical group varied by screener. For example, both CELF-5 Screening Test and PLS-5 Screening Test used "language disorder" as the label of their clinical sample. However, to meet the inclusion criteria of the clinical sample of the CELF-5 Screening Test, a participant needed to have a standard score below 76 on a comprehensive language measure. In contrast, participants needed to score below 78 on the PLS-5 to qualify for the clinical sample of the PLS-5 Screening Test. The sample used was critical in evaluating the classification accuracy and the generalizability of the screener. We first examined the number of participants used in the sensitivity and specificity calculation. For the screeners that covered areas outside language skills, only the sample size used to calculate the sensitivity and specificity of the DLD screening portion was used. Table 5 showed the sample size of each screener. Among the screeners, the PLS-5 Screening Test recruited the largest sample (1,440 participants), whereas the Bayley-4 Screening Test had the smallest sample (50 participants).

Table 5. Summary of the technical standards of the screeners.

Screener	Technical manual	Conceptual definition	Outcome measure	Sample size used in classification accuracy calculation	Sample distribution (race/ethnicity groups included)	Year of sample collection	Cutoff score
Bayley-4 Screening Test	✓	SLI	Clinical diagnosis	50	African American, Asian, Hispanic, other, White	Not available	High risk: 0–2 percentile Borderline risk: 3–25 percentile
CCC-2	✓	SLI	Clinical diagnosis	116	African American, Hispanic, White, others	2005	1 *SD* below the mean
CELF-5 Screening Test	✓	Language disorder	Clinical diagnosis	332	White, Hispanic, African American, Asian, other	2012	Below criterion score of each age
CSBS DP	✓	Communication and language delays	Not available	180	White, Black, Asian, other; Hispanic	1997–2000	1.25 *SD* below the mean
NLM	✓	Language impairment	Clinical diagnosis	527	White, Hispanic, African American, Native American, other	2014–2015	1 *SD* below the mean
LUI	✓	Language difficulties	DELV–Norm Referenced, CELF Preschool-2, CCC-2, and parent interview	236	Not available	2012	1.64 *SD* below the mean
PLS-5 Screening Test	✓	Developmental language delay and language disorder	PLS-5 and clinical diagnosis	1440	African American, Asian, Hispanic, White, other	2010	Less than four items correct
QUILS	✓	Language delay or disorder (students with weaknesses in tense, vocabulary, syntax)	Not available	415	White, Black/African American, multiracial, Asian, other races	2014	Not available
Quick Take Along	Not available	Not available	Not available	Not available	Not available	Not available	Not available
SLS Screener	✓	Language impairment	Clinical diagnosis	269	Not available	2010–2015	Two or more ratings less than five on the first eight items
DELV–Screening Test	✓	Language disorder	Clinical diagnosis	74	African American, Caucasian, Hispanic, other	2001	Criterion varied by age
ESI-3	✓	Not available	Not available	Not available	African American, Asian, Hispanic, other, White	Not available	Not available
TEGI	✓	Language disorder	Clinical diagnosis, CELF-3, and CELF-Preschool	100	African American, Hispanic, other, White	2000	16th percentile

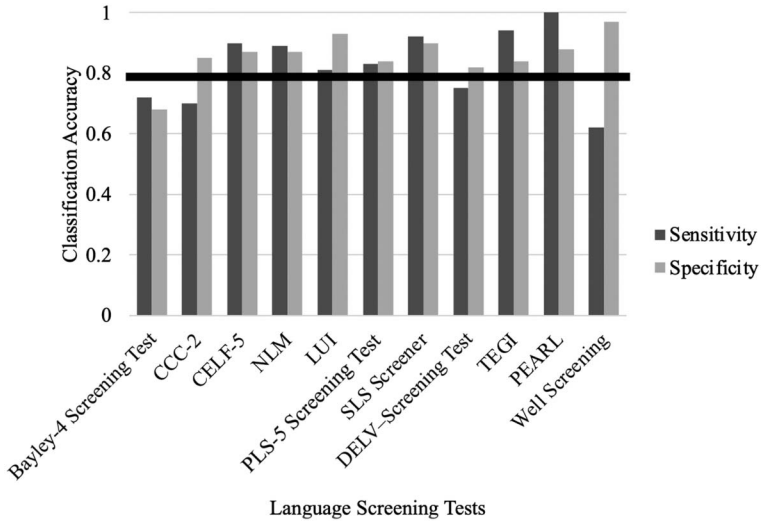
(*table continues*)

Table 5. *(Continued).*

Screener	Technical manual	Conceptual definition	Outcome measure	Sample size used in classification accuracy calculation	Sample distribution (race/ethnicity groups included)	Year of sample collection	Cutoff score
PEARL	✓	Language disorder	Receiving language service, story retell database, and parent/caregiver confirmation	42	Not available	Not available	One or lower on Mod-7
Well Screening	✓	Expressive and receptive language disorder	KDI-II, K-SEALS, and KBIT-2	188	White, Black, Hispanic, other	2014–2019	Greater than 1 *SD* below the mean

Note. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; SLI = specific language impairment; CCC-2 = Children's Communication Checklist–2: United States Edition; CELF-5 Screening Test = Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test; CSBS DP = Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; CELF Preschool-2 = Clinical Evaluation of Language Fundamentals Preschool–Second Edition; DELV–Norm Referenced = Diagnostic Evaluation of Language Variation–Norm Referenced; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test; QUILS = Quick Interactive Language Screener; Quick Take Along = Speech and Language Screener Quick Take Along; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; ESI-3 = Early Screening Inventory, Third Edition; TEGI = Rice/Wexler Test of Early Grammatical Impairment; PEARL = Predictive Early Assessment of Reading and Language Screener; KDI-II = Kindergarten Diagnostic Instrument, Second Edition; K-SEALS = Kaufman Survey of Early Academic and Language Skills; KBIT-2 = Kaufman Brief Intelligence Test, Second Edition.

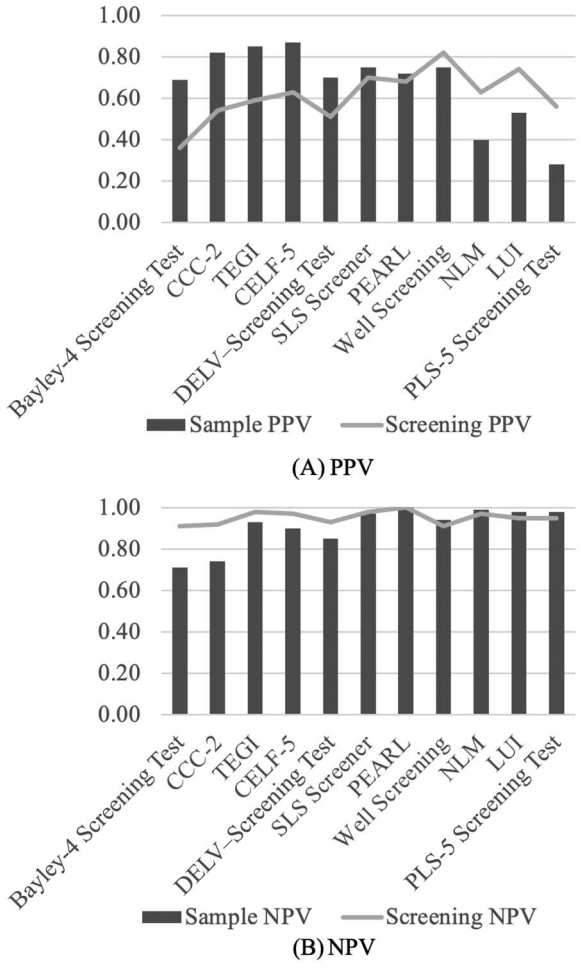
Figure 4. The sensitivity and specificity of the developmental language disorder screeners. The black line shows the .8 sensitivity and specificity. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; CCC-2 = Children's Communication Checklist–2: United States Edition; CELF-5 = Clinical Evaluation of Language Fundamentals–Fifth Edition; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; TEGI = Rice/Wexler Test of Early Grammatical Impairment; PEARL = Predictive Early Assessment of Reading and Language Screener.



Eleven out of the 15 screeners (73%) included information regarding sample distribution in the manuals or peer-reviewed articles. All but one (LUI, Canadian sample) of the screeners collected their sample in the United States. Table 5 presents race and ethnicity information for each screener's sample population. Because most of the screeners did not specify the race or ethnicity distribution for the sample used to calculate classification accuracy in identifying DLD, we presented the distribution data from the entire sample. Therefore, the screeners did not provide sufficient information for us to determine whether the entire sample and the sample used to calculate classification accuracy in identifying DLD had the same distribution. Thus, we did not include the percentage of each race or ethnicity group in the fact sheet. We also noted the year when the sample was collected. The samples used in the CSBS DP, the NLM, the SLS Screener, and the Well Screening were collected for more than 1 year. We took the time when the screeners started data collection for these four screeners. Among the 11 screeners with sampling dates, five screeners (45%) collected their normative data in the past 10 years, after 2012. Data collection years ranged from 2000 to 2014. Finally, we gathered information on the cutoff score. The cutoff score was closely associated with sensitivity and specificity. Therefore, depending on the different cutoff scores, the screener could have different classification accuracy. Cutoff scores for all screeners are presented in Table 5. The screeners used two methods to determine cutoff scores: criterion-referenced and norm-referenced. In the criterion-referenced screeners, cutoff scores were marked as the number of correct or incorrect items. For example, to meet the "not at risk" criteria in PLS-5 Screening Test, a child must correctly answer five or more items. For norm-referenced screeners, the cutoff score was based on a standard deviation, standard score, or percentile ranking. For example, to be identified as "at risk" in the CCC-2, a child must score 1 *SD* below the mean.

We then examined four aspects of classification accuracy (sensitivity, specificity, PPV, and NPV), with base rates as a reference. We summarized the sensitivity and specificity for the 11 screeners that included this information in percentages or a number between zero and one. This information is presented in Table 6. Figure 4 provides sensitivity and specificity for each screener for which this information is available. Among the 11 screeners that had sensitivity and specificity measured specifically for screening the risk of DLD, seven (64%) reached .8 in both sensitivity and specificity. These screeners were considered to have acceptable sensitivity and specificity. Only the SLS Screener reached excellent classification accuracy with a sensitivity of .92 and a specificity of .9, meaning that the SLS Screener identified 92% of children with DLD as "at risk" of DLD and 90% of children without DLD as "not at risk" of DLD. Thirty-six percent of the screeners had at least one unacceptable (below .8) sensitivity or specificity.

Figure 5. The positive predictive value (PPV) and negative predictive value (NPV) based on sample base rate and screening base rate (20%) of the developmental language disorder screeners, sorted by the sample base rate from highest to lowest. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; CCC-2 = Children's Communication Checklist–2: United States Edition; TEGI = Rice/Wexler Test of Early Grammatical Impairment; CELF-5 = Clinical Evaluation of Language Fundamentals–Fifth Edition; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; PEARL = Predictive Early Assessment of Reading and Language Screener; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test.



We also calculated the PPV and NPV using the base rate of the sample of the standardization research and the 20% screening base rate, in which we assumed 20% of the screened population was at risk of DLD. Before calculating the PPV and NPV, we summarized the sample base rate of all included screeners. The base rates ranged from 7% to 50%. Among the 11 screeners that recruited participants with DLD and controls, 36% (i.e., NLM, LUI, PLS-5 Screening Test, and Well Screening) recruited a sample close to the prevalence of DLD in the population (Norbury et al., 2016). The SLS Screener, DELV–Screening Test, and PEARL (27%) sample base rates were close to the 20% screening base rate. The rest of the screeners (36%) used a matched sample, in which the number of DLD participants was the same as controls. Figure 5 depicts the sample PPV and NPV and the screening PPV and NPV, sorted by the sample base rate of the included screeners from highest to lowest. Some of the included screeners presented PPV and NPV at different base rates in their technical manuals. We cross-checked our PPV and NPV calculations with the published values. In each case, our calculated value matched the values provided by these screeners.

When we examined the PPV separately, we concluded that it was inflated when the sample base rate was higher than the screening base rate. In other words, the PPV based on the sample was higher than the PPV based on the screening base rate. This is shown in Figure 5A. The higher the percentage of participants with DLD in the sample, the more inflated the PPV was. For example, the Bayley-4 Screening Test used a matched sample in its standardization research. The PPV calculated from their sample base rate was .69, meaning 69% of people identified as at risk of DLD by the screener were truly at risk of DLD. Therefore, 31% of children were misidentified by the screener as false positives. When we changed the base rate to 20% and recalculated the PPV, the PPV changed from .69 to .36, which indicated that in a given context where 20% of people have the risk of DLD, 36% of people identified with risk by the Bayley-4 Screening Test truly have DLD. Sixty-four percent of people would therefore be misidentified. As a result, even though the same screener was used, the percentage of misidentification in terms of false positives increased from 31% to 64% because of the change in the assumed DLD base rate.

On the contrary, NPV would be higher with a lower DLD base rate. For example, the CCC-2 had a base rate of 50% in their sample, meaning half of the participants had DLD. The NPV based on the 50% base rate was .74, which meant 74% of people who received a "not at risk" label were not at risk of DLD. However, if a clinician used the CCC-2 in an environment where 20% of people were at risk of DLD, the screener performance in identifying people without DLD improved: 92% of people identified as "not at risk of DLD" are truly not at risk of DLD. Even though some screeners had less satisfactory NPV at their sample base rate used in corresponding screeners, all screeners have reached an NPV of .9 based on a uniform screening base rate of 20%.

Examining PPV and NPV simultaneously, the NPV of the included screeners was, on average, much higher

Table 6. Summary of sensitivity, specificity, sample base rate, positive predictive value (PPV), and negative predictive value (NPV).

Screener	Sensitivity	Specificity	Sample base rate	PPV based on sample base rate	PPV based on 20% screening base rate	NPV based on sample base rate	NPV based on 20% screening base rate
Bayley-4 Screening Test	.72[a]	.68[a]	50%	.69	.36	.71	.91
CCC-2	.70	.85	50%	.82	.54	.74	.92
CELF-5 Screening Test	.90[b]	.87[b]	50%	.87	.63	.90	.97
CSBS DP	Not available	Not available	Not available	Not available	Not available	Not available	Not available
NLM	.89[c]	.87[c]	9%	.40	.63	.99	.97
LUI	.81[d]	.93[d]	9%	.53	.74	.98	.95
PLS-5 Screening Test	.83[e]	.84[e]	7%	.28	.56	.98	.95
QUILS	Not available	Not available	Not available	Not available	Not available	Not available	Not available
Quick Take Along	Not available	Not available	Not available	Not available	Not available	Not available	Not available
SLS Screener	.92[f]	.90[f]	25%	.75	.70	.97	.98
DELV–Screening Test	.75[g]	.82[g]	36%	.70	.51	.85	.93
ESI-3	Not available	Not available	Not available	Not available	Not available	Not available	Not available
PEARL	1.00	.88	24%	.72	.68	1.00	1.00
TEGI	.94[h]	.84[h]	50%	.85	.59	.93	.98
Well Screening	.62	.97	14%	.75	.82	.94	.91

Note. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; CCC-2 = Children's Communication Checklist–2: United States Edition; CELF-5 Screening Test = Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test; CSBS DP = Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test; QUILS = Quick Interactive Language Screener; Quick Take Along = Speech and Language Screener Quick Take Along; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; ESI-3 = Early Screening Inventory, Third Edition; PEARL = Predictive Early Assessment of Reading and Language Screener; TEGI = Rice/Wexler Test of Early Grammatical Impairment.

[a]Taken from the expressive communication classification with high risk and borderline risk. [b]Taken from the 5:0–8:11 (years;months) age group. [c]Taken from the fall Pre-K group. [d]Taken from the 24–47 months age group. [e]Taken from the language disorder group. [f]Taken from the 9:0–9:11 age group. [g]Taken from the teacher-reported results. [h]Taken from the 5.00–5.11 age group at probe score of .80.

than PPV. Considering the broader purpose of screeners, which is to identify those at risk of DLD, administrators generally lean toward screeners with higher sensitivity and NPV, even if it means tolerating a lower PPV.

In essence, successful screeners need to effectively flag those at potential risk (i.e., high sensitivity). Achieving this often necessitates tolerating a lower specificity (i.e., more false positives) to ensure that all individuals with potential DLD are identified as at risk. The high specificity, on the other hand, assures that those who pass a screener are genuinely not "at risk."

This approach allows for tailored strategies such as progress monitoring, intervention, and, when necessary, further assessment. It is important to consider the varying models where screening is employed, such as in an RTI model, where screening is performed, some Tier Two support is provided, and then the effectiveness of the support is evaluated. This can lead to further evaluation if required, based on how individuals respond to the initial support. In this way, screening acts as a flexible tool, helping to triage and guide the most appropriate next steps in a variety of contexts to best serve the needs of individuals at risk for DLD.

Discussion

Overview

In this review, we summarized, compared, and analyzed commercially available North American–based DLD screeners for use with English-speaking children. Our goals were to (a) provide clinicians and educators with a wide range of information to help them choose DLD screeners for specific contexts and (b) identify areas that need improvement for future DLD screener development.

Using Screener Information to Make Clinical Decisions: Top Rated Screeners and Five Scenarios

Based on the criteria and information evaluated in this review, we highlight three screeners that received the best ratings in accessibility, usability, and psychometrics (see operational definitions of these aspects in Table 2).

- CELF-5 Screening Test
  - Published in 2013, which is relatively recent (see Table 3)

- Examiner qualification: Level B (see Table 3)
- Age range: 5;0–21;11 (years;months), covering a wide age range (see Figure 3)
- Sensitivity: .902, excellent (see Table 6 and Figure 4)
- Specificity: .872, acceptable (see Table 6 and Figure 4)
- Dialect compatibility is present (see Table 4)
- Clear outcome measure with a large and diverse sample size (see Table 5)

- SLS Screener
  - Published in 2018, which is recent (see Table 3)
  - Examiner qualification: Level A, accessible to more people (see Table 3)
  - Age range: 6;0–18;11, covering a wide age range (see Figure 3)
  - Sensitivity: .926, excellent (see Table 6 and Figure 4)
  - Specificity: .906, excellent (see Table 6 and Figure 4)
  - Actionable follow-up instruction is present (see Table 4)
  - Clear outcome measure with sample size information (see Table 5)
- PLS-5 Screening Test
  - Published in 2012, which is relatively recent (see Table 3)
  - Examiner qualification: Level B (see Table 3)
  - Age range: 0;0–7;11, covering a wide age range (see Figure 3)
  - Sensitivity: .835, acceptable (see Table 6 and Figure 4)
  - Specificity: .845, acceptable (see Table 6 and Figure 4)
  - Online administration availability is present (see Table 4)
  - Clear outcome measure with a large and diverse sample size (see Table 5)

Table 7 summarizes the usability features and classification accuracy as a quick access to clinicians and administrators. We also present four user scenarios and provide screener recommendations. In our selection of screeners for each scenario, we adopted a comprehensive approach that extended beyond mere consideration of sensitivity and specificity. While classification accuracy is undeniably critical, it is equally important to recognize the multifaceted nature of screeners, especially when applied in real-world contexts. Many recently published screeners are yet to have complete psychometric. Our recommendations consider screeners with clinically valuable features, providing utility in specific settings. This decision was informed by studies that emphasize the risks of an overreliance on strict sensitivity and specificity criteria (White & Schatschneider, 2023). Our guiding principle was to offer recommendations that would take into account both practicality and accuracy. While some recommended screeners might not exhibit ideal classification accuracy presently, their intrinsic features, such as their suitability for certain age ranges or adaptability to diverse linguistic backgrounds, make them the preferred screener in the contexts described in our scenarios. Ultimately, our goal is to provide a well-rounded perspective, balancing both quantitative metrics and the nuanced, practical utility of each screener.

Scenario 1—Preschool teachers screening children aged 3–5 years for DLD.

Recommendation: PLS, QUILS, CSBS DP, with ESI and TEGI as alternative options.

Rationale: Our primary recommendations of PLS, QUILS, and CSBS DP were based on their alignment with the 3–5 years age range, combined with their distinct, clinically valuable features. The PLS is commendable for its dual features: actionable follow-up instruction and online administration availability, along with its robust classification accuracy, as both its sensitivity and specificity surpass the 80% threshold. The QUILS, even though its sensitivity and specificity were not published specifically for the DLD population, stands out with three significant features: actionable follow-up instruction, online administration availability, and its compatibility across various dialects. The CSBS DP offers the invaluable benefit of progress monitoring, alongside its actionable follow-up instruction. This allows preschool teachers to diligently track language development progress over time and suggest timely intervention strategies. Additionally, while the ESI and the TEGI lack the previously mentioned features, their commendable classification accuracy, surpassing acceptable standards, makes them potential alternatives for this age range. Though the TEGI requires thoughtful interpretation because the items are susceptible to dialect variation.

Scenario 2—Elementary school teachers screening students ages 5–11 years for DLD.

Recommendation: Primarily CELF-5 Screening Test, with CCC-2 and Speech and Language Screener Quick Take Along as considerations. For specific age subgroups within this range, options include PLS-5 Screening Test,

Table 7. Summary of the usability features and classification accuracy.

Screener	Dialect compatibility	Progress monitoring function	Actionable follow-up instruction	Group assessment capability	Online administration availability	Acceptable sensitivity (above 80%)	Acceptable specificity (above 80%)
Bayley-4 Screening Test			✓				
CCC-2							✓
CELF-5 Screening Test	✓		✓			✓	✓
CSBS DP		✓	✓				
NLM		✓				✓	✓
LUI					✓	✓	✓
PLS-5 Screening Test			✓		✓	✓	✓
QUILS	✓		✓		✓		
Quick Take Along							
SLS Screener			✓			✓	✓
DELV–Screening Test	✓		✓				✓
ESI-3			✓		✓		
PEARL						✓	✓
TEGI						✓	✓
Well Screening		✓			✓		

Note. Bayley-4 Screening Test = Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test; CCC-2 = Children's Communication Checklist–2: United States Edition; CELF-5 Screening Test = Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test; CSBS DP = Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition; NLM = CUBED Narrative Language Measures; LUI = Language Use Inventory; PLS-5 Screening Test = Preschool Language Scales–5 Screening Test; QUILS = Quick Interactive Language Screener; Quick Take Along = Speech and Language Screener Quick Take Along; SLS Screener = Student Language Scale Screener for Language & Literacy Disorders; DELV–Screening Test = Diagnostic Evaluation of Language Variation–Screening Test; ESI-3 = Early Screening Inventory, Third Edition; PEARL = Predictive Early Assessment of Reading and Language Screener; TEGI = Rice/Wexler Test of Early Grammatical Impairment.

DELV–Screening Test, Well Screening, SLS Screener, NLM, TEGI, and PEARL.

Rationale: For the broader elementary (age range: 5–11 years), three screeners, namely, CCC-2, CELF-5 Screening Test, and Speech and Language Screener Quick Take Along, are appropriate options. Among these, the CELF-5 Screening Test emerges superior, showcasing commendable sensitivity and specificity metrics. It further excels with its usability features, including dialect compatibility and actionable follow-up instructions. In contrast, the CCC-2 solely achieves specificity above the 80% mark and lacks additional features. The Speech and Language Screener Quick Take Along, though covering the age range, is limited by the absence of normative data and defined quality features. For more restricted age ranges, an array of screeners is available. Notably, the PLS-5 Screening Test stands out, balancing excellent classification accuracy with its two notable features: actionable follow-up instructions and online administration availability. Both the DELV–Screening Test and the Well Screening possess two usability features, with the DELV–Screening Test focusing on dialect compatibility and actionable follow-up and the Well Screening emphasizing progress monitoring and online administration. However, their classification accuracy could be improved. The SLS Screener and the NLM shine with above 80% sensitivity and specificity; the former provides actionable follow-up, and the latter emphasizes progress monitoring. Lastly, the TEGI and the PEARL, although not featuring distinct usability characteristics, still excel with over 80% classification accuracy.

Scenario 3—Speech-language pathologists working with diverse populations of children to screen for DLD.

Recommendation: DELV–Screening Test, NLM, and QUILS.

Rationale: These screeners were developed for use with children who speak varied dialects and include culturally and linguistically diverse samples. The NLM has been validated on bilingual English-Spanish–speaking children (Spencer et al., 2023).

Scenario 4—Researchers evaluating language development in a longitudinal study and planning to screen for DLD over time.

Recommendation: CSBS DP, NLM, and LUI.

Rationale: The CSBS DP, the NLM, and the LUI offer progress monitoring, allowing researchers to track individual differences progress over time. These screeners provide detailed information in their technical manuals, which is valuable for understanding the screeners' psychometric properties and interpreting results in a research context.

Development of DLD Screeners Should Be Held to the Highest Technical Standards

This review highlights critical areas for improvement when developing DLD screeners. First, clear and consistent definitions and corresponding criteria to measure DLD are essential. Currently, varied definitions of DLD across screeners make it difficult to ensure these screeners are identifying children with or without DLD. Many screeners only provided a vague description of the disorders they measured, and labels included language impairment, language difficulty, and language disorder. Many screeners simply reported "clinical diagnosis" as their outcome measure, but the tools and standards used to validate the diagnosis were not clearly described in technical manuals. These inconsistencies and opaque descriptions are unsurprising given the wide variability in the terminology used to refer to DLD at the time of the publication of the screeners. Newly developed DLD screeners should adhere to current best practices for identifying DLD and should be clear and consistent in their use of terminology for referring to language disorders.

In terms of classification accuracy, DLD screeners should adhere to the highest standards of psychometrics (Ebert et al., 2020). Organizations have begun to evaluate screeners using these standards. An example is the "Tools Chart" (https://charts.intensiveintervention.org/ascreening). The chart is continually updated by soliciting data from publishers, which is reviewed by content experts and psychometricians. Each screener is rated on a variety of criteria as acceptable, half acceptable, and not acceptable. Over the decade that the tools chart has displayed screener ratings, publishers have improved screeners according to the feedback they receive from reviewers. Currently, only math and reading screeners are reviewed. This review article serves as a starting point for the creation of a similar resource for speech and language assessments, including DLD screeners, to provide clinicians, educators, administrator, and policymakers valuable information for decision making and publishers information to improve assessments.

Advancing DLD Screening: Actionable Steps and Cultural Sensitivity

For DLD screeners to be useful, users need actionable next steps, such as ways to use the results within an RTI model and explicit instructions on interpretation of results for varied dialects. A clear need identified by this review is the need for commercially available culturally and linguistically appropriate screeners sensitive to language variation and dialect use. Because DLD screeners are normed without consideration of the range of dialects spoken within the United States, there is an inherent bias that could lead to misidentification of DLD risk for some children (Oetting, 2018, 2020; Seymour & Pearson, 2004). Moreover, there is a dearth of DLD screeners available for languages other than English, underscoring an essential area for future research and development.

Limitations

Despite the intention of the study to provide a comprehensive and detailed evaluation of DLD screeners, the current study only included commercially available screeners for DLD, excluding screeners used in peer-reviewed studies (Hendricks et al., 2019; Redmond et al., 2019). Because of restricted access to these measures, we included screeners that were available for purchase. However, because more manuscripts and their measures are now open access, future reviews should consider including screeners published in research studies.

While this study provides an in-depth evaluation of DLD screeners, our focus was predominantly on classification accuracy, along with other aspects such as accessibility information and usability features. This was intentionally done to provide a balanced review, given that classification accuracy is a pivotal indicator of accurate screening. As such, traditional test psychometrics such as reliability and validity were not the central focus of our review. Looking toward future research, a deeper exploration into the validity and reliability of DLD screeners would be beneficial and could provide additional dimensions to this evaluation.

Lastly, we acknowledge that this review only captures screeners published through July 2023. Our presentation of a clearly replicable process allows for extension to newly published tools as they are available. Again, a tools chart would be a valuable resource that could be maintained and updated each year.

Conclusions

In this study, we comprehensively summarized, compared, and evaluated screeners to identify risk for DLD. We used the PRISMA approach for the search process and summarized data in open access fact sheets (see Supplemental Material S1). We examined 15 DLD screeners on three domains: accessibility information, usability features, and technical standards. Our findings indicate that a small number of DLD screeners meet most standards and highlight numerous areas for improvement. These include increasing classification accuracy; consideration of linguistic diversity; and more explicit, actionable follow-up guidance. Screening for DLD is critical for timely intervention and classroom support (Adlof & Hogan, 2019). It

is our aim that this review will help facilitate the decision-making process for DLD screener selection across numerous contexts with the downstream effect of increasing early identification and support for children with DLD.

Data Availability Statement

The data supporting the findings of this study are derived from publicly available screeners and direct correspondence with the creators of these screeners. The compiled data sets that inform this review are not publicly available due to copyright and confidentiality agreements. However, summaries of the data are presented within the article and the supplemental material. Supplemental Material S1 displays detailed data collected for each screener in the form of a fact sheet, which has been uploaded to OSF Preprints for public access (https://doi.org/10.31219/osf.io/cajp5). Researchers interested in accessing the detailed data should contact the corresponding screener authors.

Acknowledgments

This research was funded by Massachusetts General Hospital Institute of Health Professions, doctoral fellowship grant awarded to the first author. We thank Kate Radville for her comments and suggestions.

References

Adlof, S. M., & Hogan, T. P. (2018). Understanding dyslexia in the context of developmental language disorders. *Language, Speech, and Hearing Services in Schools, 49*(4), 762–773. https://doi.org/10.1044/2018_LSHSS-DYSLC-18-0049

Adlof, S. M., & Hogan, T. P. (2019). If we don't look, we won't see: Measuring language development to inform literacy instruction. *Policy Insights From the Behavioral and Brain Sciences, 6*(2), 210–217. https://doi.org/10.1177/2372732219839075

Archibald, L. (2019, September 23). Why is DLD hard to recognize? *DLD and Me*. https://dldandme.org/recognizing-dld/

Baker, D. P., Eslinger, P. J., Benavides, M., Peters, E., Dieckmann, N. F., & Leon, J. (2015). The cognitive impact of the education revolution: A possible cause of the Flynn effect on population IQ. *Intelligence, 49,* 144–158. https://doi.org/10.1016/j.intell.2015.01.003

Baldessarini, R. J., Finklestein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry, 40*(5), 569–573. https://doi.org/10.1001/archpsyc.1983.01790050095011

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bayley, N., & Aylward, G. P. (2019). *Bayley Scales of Infant and Toddler Development–Fourth Edition, Screening Test*. Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Cognition-%26-Neuro/Brief/Bayley-Scales-of-Infant-and-Toddler-Development-Screening-Test/p/100001998.html

Bishop, D. V. M. (2006). *Children's Communication Checklist–2: United States Edition*. Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-%26-Language/Children%27s-Communication-Checklist-2-%7C-U-S-Edition/p/100000193.html

Bishop, D. V. M. (2017). Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). *International Journal of Language & Communication Disorders, 52*(6), 671–680. https://doi.org/10.1111/1460-6984.12335

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE-2 Consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry, 58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Calder, S. D., Boyes, M., Brennan-Jones, C. G., Whitehouse, A. J. O., Robinson, M., & Hill, E. (2023). Do parent-reported early indicators predict later developmental language disorder? A Raine Study investigation. *International Journal of Language & Communication Disorders*. Advance online publication. https://doi.org/10.1111/1460-6984.12950

Catts, H. W., & Hogan, T. P. (2020). *Dyslexia: An ounce of prevention is better than a pound of diagnosis and treatment*. PsyArXiv. https://doi.org/10.31234/osf.io/nvgje

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities, 48*(3), 281–297. https://doi.org/10.1177/0022219413498115

Chan, J., Adlof, S. M., Duff, D., Mitchell, A., Ragunathan, M., & Ehrhorn, A. M. (2022). Examining the associations between parent concerns and school-age children's language and reading abilities: A comparison of samples recruited for in-school versus online participation. *Language, Speech, and Hearing Services in Schools, 53*(2), 431–444. https://doi.org/10.1044/2021_LSHSS-21-00080

Council of Chief State School Officers. (n.d.). *English language arts standards » Introduction » Key design consideration. Common core state standards initiative*. Retrieved August 3, 2021, from https://www.thecorestandards.org/ELA-Literacy/introduction/key-design-consideration/

Cross, A. M., Joanisse, M. F., & Archibald, L. M. D. (2019). Mathematical abilities in children with developmental language disorder. *Language, Speech, and Hearing Services in Schools, 50*(1), 150–163. https://doi.org/10.1044/2018_LSHSS-18-0041

Dexter, D. D., & Hughes, C. (n.d.). *Progress monitoring within a response-to-intervention model.* RTI Action Network. Retrieved August 5, 2021, from http://www.rtinetwork.org/learn/research/progress-monitoring-within-a-rti-model

Dubois, P., St-Pierre, M-C., Desmarais, C., & Guay, F. (2020). Young adults with developmental language disorder: A systematic review of education, employment, and independent living outcomes. *Journal of Speech, Language, and Hearing Research, 63*(11), 3786–3800. https://doi.org/10.1044/2020_JSLHR-20-00127

Ebert, K. D., Ochoa-Lubinoff, C., & Holmes, M. P. (2020). Screening school-age children for developmental language disorder in primary care. *International Journal of Speech-Language Pathology, 22*(2), 152–162. https://doi.org/10.1080/17549507.2019.1632931

Ekelman, B. L., Dutka, D. A., & St. Amour, K. (2021). *The Well Screening, Research Edition*. Brookes.

Fischhoff, B., & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Memory & Cognition, 12*(4), 402–410. https://doi.org/10.3758/BF03198301

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health, 10*(S2), S94–S105. https://doi.org/10.1111/j.1524-4733.2007.00272.x

Gaab, N., & Petscher, Y. (2022). Screening for early literacy milestones and reading disabilities: The why, when, whom, how, and where. *Perspectives on Language and Literacy, Winter,* 11–18.

Gallagher, A. L., Murphy, C., Conway, P., & Perry, A. (2019). Consequential differences in perspectives and practices concerning children with developmental language disorders: An integrative review. *International Journal of Language & Communication Disorders, 54*(4), 529–552. https://doi.org/10.1111/1460-6984.12469

Gallinat, E., & Spaulding, T. J. (2014). Differences in the performance of children with specific language impairment and their typically developing peers on nonverbal cognitive tests: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 57*(4), 1363–1382. https://doi.org/10.1044/2014_JSLHR-L-12-0363

Gatlin-Nash, B., Chow, J. C., & Evans, I. (2023). Addressing the needs of nonmainstream dialect speakers with learning disabilities. *Intervention in School and Clinic, 59*(1), 20–28. https://doi.org/10.1177/10534512221130066

Georgan, W. C., Archibald, L. M. D., & Hogan, T. P. (2023). Speech/language impairment or specific learning disability? Examining the usage of educational categories. *Journal of Speech, Language, and Hearing Research, 66*(2), 656–667. https://doi.org/10.1044/2022_JSLHR-21-00636

Golinkoff, R. M., de Villiaers, J., Hirsh-Pasek, K., Iglesias, A., & Wilson, M. S. (2017). *Quick Interactive Language Screeners*. Brookes.

Golinkoff, R. M., de Villiaers, J., Hirsh-Pasek, K., Iglesias, A., & Wilson, M. S. (2022). *Quick Interactive Language Screeners*. https://quilscreener.com/screening-tools/

Grégoire, J., Daniel, M., Llorente, A. M., & Weiss, L. C. (2016). The Flynn effect and its clinical implications. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist–practitioner perspectives* (pp. 187–212). Elsevier. https://doi.org/10.1016/B978-0-12-404697-9.00006-6

Hadley, E. B., & Mendez, K. Z. (2021). Learning words that matter: Selecting vocabulary words for young children. *The Reading Teacher, 74*(5), 595–605. https://doi.org/10.1002/trtr.1978

Hendricks, A. E., & Adlof, S. M. (2017). Language assessment with children who speak nonmainstream dialects: Examining the effects of scoring modifications in norm-referenced assessment. *Language, Speech, and Hearing Services in Schools, 48*(3), 168–182. https://doi.org/10.1044/2017_LSHSS-16-0060

Hendricks, A. E., Adlof, S. M., Alonzo, C. N., Fox, A. B., & Hogan, T. P. (2019). Identifying children at risk for developmental language disorder using a brief, whole-classroom screen. *Journal of Speech, Language, and Hearing Research, 62*(4), 896–908. https://doi.org/10.1044/2018_JSLHR-L-18-0093

Herman, C. (2006). What makes a screening exam "good"? *The Virtual Mentor, 8*(1), 34–37. https://doi.org/10.1001/virtualmentor.2006.8.1.cprl1-0601

Hunt, E., Nang, C., Meldrum, S., & Armstrong, E. (2022). Can dynamic assessment identify language disorder in multilingual children? Clinical applications from a systematic review. *Language, Speech, and Hearing Services in Schools, 53*(2), 598–625. https://doi.org/10.1044/2021_LSHSS-21-00094

Institute for Quality and Efficiency in Health Care. (2006). Benefits and risks of screening tests. In *InformedHealth.org* [Internet]. Institute for Quality and Efficiency in Health Care. https://www.ncbi.nlm.nih.gov/books/NBK279418/

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. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base-rate fallacy in school psychology. *Journal of Psychoeducational Assessment, 15*(4), 292–307. https://doi.org/10.1177/073428299701500401

Klee, T., Carson, D. K., Gavin, W. J., Hall, L., Kent, A., & Reece, S. (1998). Concurrent and predictive validity of an early language screening program. *Journal of Speech, Language, and Hearing Research, 41*(3), 627–641. https://doi.org/10.1044/jslhr.4103.627

Klee, T., Pearce, K., & Carson, D. K. (2000). Improving the positive predictive value of screening for developmental language disorder. *Journal of Speech, Language, and Hearing Research, 43*(4), 821–833. https://doi.org/10.1044/jslhr.4304.821

Lancaster, H. S., & Camarata, S. (2019). Reconceptualizing developmental language disorder as a spectrum disorder: Issues and evidence. *International Journal of Language & Communication Disorders, 54*(1), 79–94. https://doi.org/10.1111/1460-6984.12433

Law, E. L.-C., & Heintz, M. (2021). Augmented reality applications for K–12 education: A systematic review from the usability and user experience perspective. *International Journal of Child–Computer Interaction, 30,* 100321. https://doi.org/10.1016/j.ijcci.2021.100321

Lu, J., Schmidt, M., Lee, M., & Huang, R. (2022). Usability research in educational technology: A state-of-the-art systematic review. *Educational Technology Research and Development, 70*(6), 1951–1992. https://doi.org/10.1007/s11423-022-10152-6

Martínez-Mesa, J., González-Chica, D. A., Duquia, R. P., Bonamigo, R. R., & Bastos, J. L. (2016). Sampling: How to select participants in my research study? *Anais Brasileiros de Dermatologia, 91*(3), 326–330. https://doi.org/10.1590/abd1806-4841.20165254

McGregor, K. K. (2020). How we fail children with developmental language disorder. *Language, Speech, and Hearing Services in Schools, 51*(4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003

Meisels, S. J. (1989). Can developmental screening tests identify children who are developmentally at risk? *Pediatrics, 83*(4), 578–585. https://doi.org/10.1542/peds.83.4.578

Meisels, S. J., & Marsden, D. B. (2019). *Early Screening Inventory, Third Edition*. Pearson.

Nation, K., Clarke, P., Marshall, C. M., & Durand, M. (2004). Hidden language impairments in children: Parallels between poor reading comprehension and specific language impairment? *Journal of Speech, Language, and Hearing Research, 47*(1), 199–211. https://doi.org/10.1044/1092-4388(2004/017)

Nelson, N. W., Howes, B. M., & Anderson, M. A. (2018). *Student Language Scale Screener for Language & Literacy Disorders*. Brookes.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *The Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

O'Neill, D. K. (2009). *Language Use Inventory*. Knowledge In Development Inc.

Oetting, J. B. (2018). Prologue: Toward accurate identification of developmental language disorder within linguistically diverse schools. *Language, Speech, and Hearing Services in Schools, 49*(2), 213–217. https://doi.org/10.1044/2018_LSHSS-CLSLD-17-0156

Oetting, J. B. (2020). *From my perspective/opinion: General American English as a dialect: A call for change. The ASHA LeaderLive*. https://leader.pubs.asha.org/do/10.1044/leader.FMP.25112020.12/full/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372,* Article n71. https://doi.org/10.1136/bmj.n71

Palincsar, A. S., Magnusson, S. J., Collins, K. M., & Cutter, J. (2001). Making science accessible to all: Results of a design experiment in inclusive classrooms. *Learning Disability Quarterly, 24*(1), 15–32. https://doi.org/10.2307/1511293

Petersen, D. B., & Spencer, T. D. (2014). Narrative assessment and intervention: A clinical tutorial on extending explicit language instruction and progress monitoring to all students. *Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse (CLD) Populations, 21*(1), 5–21. https://doi.org/10.1044/cds21.1.5

Petersen, D. B., & Spencer, T. D. (2016). *CUBED Narrative Language Measures*. Language Dynamics Group.

Petscher, Y., & Suhr, M. (2022). Considerations for choosing and using screeners for students with disabilities. In C. J. Lemons, S. R. Powell, K. L. Lane, & T. C. Aceves (Eds.), *Handbook of special education research* (Vol. II). Routledge. https://doi.org/10.4324/9781003156888-8

Pitts, R. T., & Naumenko, O. (2016). The 2014 standards for educational and psychological testing: What teachers initially need to know. *Working Papers in Education, 2*(1).

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15–24. https://doi.org/10.1044/0161-1461.2501.15

Porter, K. L., Oetting, J. B., & Pecchioni, L. (2020). Caregivers' perceptions of speech-language pathologist talk about child language and literacy disorders. *American Journal of Speech-Language Pathology, 29*(4), 2049–2067. https://doi.org/10.1044/2020_AJSLP-20-00049

Prizant, B. M., & Wetherby, A. M. (2002). *Communication and Symbolic Behavior Scales Developmental Profile: First Normed Edition*. Brookes.

Prizant, B. M., & Wetherby, A. M. (2022). *Communication and Symbolic Behavior Scales Developmental Profile, First Normed Edition*. Brookes. https://brookespublishing.com/product/csbs-dp/

Redmond, S. M., Ash, A. C., Christopulos, T. T., & Pfaff, T. (2019). Diagnostic accuracy of sentence recall and past tense measures for identifying children's language impairments. *Journal of Speech, Language, and Hearing Research, 62*(7), 2438–2454. https://doi.org/10.1044/2019_JSLHR-L-18-0388

Rice, M. L., & Wexler, K. (2001). *The Rice/Wexler Test of Early Grammatical Impairment*. Psychological Corporation.

Rose, S. T. (2023). *Reaching multilingual children in their home language.* The ASHA LeaderLive. https://leader.pubs.asha.org/do/10.1044/leader.FTR2.28012023.ai-treatment-slp.50/full/

Sansavini, A., Favilla, M. E., Guasti, M. T., Marini, A., Millepiedi, S., Di Martino, M. V., Vecchi, S., Battajon, N., Bertolo, L., Capirci, O., Carretti, B., Colatei, M. P., Frioni, C., Marotta, L., Massa, S., Michelazzo, L., Pecini, C., Piazzalunga, S., Pieretti, M., . . . Lorusso, M. L. (2021). Developmental language disorder: Early predictors, age for the diagnosis, and diagnostic tools. A scoping review. *Brain Sciences, 11*(5), Article 654. https://doi.org/10.3390/brainsci11050654

Seymour, H. N., & Pearson, B. Z. (2004). Evaluating language variation: Distinguishing development and dialect from disorder. Special issue. *Seminars in Speech and Language, 25*(1), Article 117.

Seymour, H. N., Roeper, T. W., & de Villiaers, J. (2003). *The Diagnostic Evaluation of Language Variation–Screening Test*. Ventris Learning.

Shapira, E. S. (n.d.). *Tiered instruction in a response-to-intervention model.* RTI Action Network. Retrieved August 5, 2021, from http://www.rtinetwork.org/essential/tieredinstruction/tiered-instruction-and-intervention-rti-model

Shreffler, J., & Huecker, M. R. (2023). Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. In *StatPearls [Internet]*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK557491/

Snyders, N. E., & Kjesbo, R. R. (2014). *Speech and Language Screener Quick Take Along*. Super Duper Publications.

Spencer, T. D., Thompson, M. S., Petersen, D. B., Liu, Y., & Restrepo, M. A. (2023). Reliability and validity evidence for the English and Spanish preschool narrative language measures-listening. *Early Childhood Research Quarterly, 64,* 148–161. https://doi.org/10.1016/j.ecresq.2023.02.005

Suresh, K., Thomas, S. V., & Suresh, G. (2011). Design, data analysis and sampling techniques for clinical research. *Annals of Indian Academy of Neurology, 14*(4), 287–290. https://doi.org/10.4103/0972-2327.91951

Tanski, S., Garfunkel, L. C., Duncan, P. M., & Weitzman, M. (2010). *Preventive services: A bright futures handbook*. American Academy of Pediatrics. https://brightfutures.aap.org/Bright%20Futures%20Documents/Forward%20and%20Introduction.pdf

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Trahan, L., Stuebing, K. K., Hiscock, M. K., & Fletcher, J. M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin, 140*(5), 1332–1360. https://doi.org/10.1037/a0037173

Turner, S. M., DeMers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist, 56*(12), 1099–1113. https://doi.org/10.1037/0003-066X.56.12.1099

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*(2), 225–256. https://doi.org/10.1016/j.jsp.2006.11.004

Wechsler, S. (2003). *Wechsler Intelligence Scale for Children–Fourth Edition*. APA PsycTests. https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft15174-000

White, C. M., & Schatschneider, C. (2023). Are two screeners better than one? A simulation study of correlation and classification in universal screening. *Contemporary School Psychology*. Advance online publication. https://doi.org/10.1007/s40688-023-00478-0

Wiig, E. H., Secord, W. A., & Semel, E. (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test*. Pearson. https://www.pearsonassessments.com/store/usassessments/en/

Store/Professional-Assessments/Speech-%26-Language/CELF-5-Screening-Test/p/100000078.html

Wiig, E. H., Secord, W. A., & Semel, E. (2021). *Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test*. Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-%26-Language/CELF-5-Screening-Test/p/100000078.html

Ziegenfusz, S., Paynter, J., Flückiger, B., & Westerveld, M. F. (2022). A systematic review of the academic achievement of primary and secondary school-aged students with developmental language disorder. *Autism & Developmental Language Impairments, 7*. https://doi.org/10.1177/23969415221099397

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2012). *Preschool Language Scales–5 Screening Test*. Pearson.