Classification Accuracy of the Quick Interactive Language Screener for Preschool Children with and without Developmental Language Disorder

Amy Pace[1]

Maura Curran[2]

Amanda Owen Van Horne[3]

Jill de Villiers[4]

Aquiles Iglesias[3]

Roberta Michnick Golinkoff[3]

Mary S. Wilson[5]

Kathy Hirsh-Pasek[6]

Author Note

Departmental affiliation. 1. University of Washington; 2. MGH Institute of Health Professions; 3. University of Delaware; 4. Smith College; 5. Laureate Learning Systems; 6. Temple University

Acknowledgements. This research was supported by a Royalty Research Fund award (PI: Pace) and awards from the Institute Education Sciences (R305A110284, R324A160241; PI: Golinkoff). Study 1 data collection was supported by National Institute on Deafness and Other Communication Disorders (Grant Number: K23DC013291-01A1) and the National Science Foundation (Grant number 1748298) awarded to Karla McGregor.

Correspondence concerning this article should be addressed to Amy Pace, 1417 NE 42nd St., Seattle, WA 98105. Email: amypace@uw.edu

Classification Accuracy of the Quick Interactive Language Screener for Preschool Children with and without Developmental Language Disorder

1. Introduction

Developmental language disorder (DLD) is among the most common neurodevelopmental difficulties during early childhood, with long term impacts on learning and wellbeing (Conti-Ramsden et al., 2018). At kindergarten entry, approximately 7 to 10% of children will experience a language impairment severe enough to hinder academic progress (Norbury et al., 2016; Zablotsky et al., 2019). Children with language difficulties at or prior to school entry are more likely to have poor reading outcomes in grade school than children with typical language development, and this risk persists into adulthood with negative impacts on literacy, mental health, and even employment (Law et al., 2009; Tomblin et al., 2000). For these reasons, early and widespread assessment of language skills in preschool or earlier may be crucial to identify children who are struggling, to monitor children who are at risk, and to initiate timely intervention services (Lipkin & Macias, 2020).

Children experience optimal outcomes when language disorders are identified as early as possible (Guralnick, 2019). Despite its high prevalence and pervasive impact on functional outcomes, DLD often remains unnoticed until elementary school (Hendricks et al., 2019; McGregor, 2020). One potential solution for improving the detection and remediation of childhood disorders like DLD is widespread developmental screening in community settings or preschool programs for children ages 3 through 6 (Lipkin et al., 2020; McGregor et al., 2020; Rice, 2020). Although screening for language disorders is increasingly recommended as part of routine health services, several barriers persist. These may include high costs for testing materials, lack of time and resources for administration and scoring, limited tools for identifying

disorder within dialect (Oetting et al., 2016), in addition to limited access to trained personnel who can administer screeners (Lipkin et al., 2020b; Schonwald et al., 2009). Thus, there is a continuing need to develop and validate direct screening instruments appropriate for widespread screening of oral language skills in preschool children from diverse backgrounds (Adlof & Hogan, 2019; Redmond et al., 2019).

Screening for DLD has many benefits, but also poses challenges in selecting appropriate instruments. Most commercially available screeners for children in preschool or kindergarten are designed to be administered to individual students by speech-language pathologists and require at least 15–20 minutes per student in addition to time for scoring and interpretation (Weiler et al. 2018). Because this approach requires significant resources, universal screening of oral language skills using direct assessment (rather than indirect methods like parent or teacher report) is rare in practice but holds great clinical promise (Hendricks et al., 2019; Komesidou & Summy, 2020). Another critical factor is that many screening tools are not valid for use with children whose language systems differ from the established standard such as speakers of African American English (AAE) or other dialectal variations (Washington & Seidenberg, 2022). Thus, conventional screening approaches often result in disproportionately high failure rates because oral language screeners are not sensitive to dialect features (Oetting et al., 2016; Weiler et al., 2018). The current study examined a new tool called the Quick Interactive Language Screener (QUILS), which was developed to address several of the potential barriers to widespread language screening (Golinkoff et al., 2017).

1.1. The Quick Interactive Language Screener (QUILS)

The QUILS is a computerized touchscreen tool for preschool-aged children that can be completed in a single session (~15 min). The screener is unique in assessing three domains:

vocabulary across several word classes (Vocabulary Product), the syntax of several constructions (Syntax Product), and how children learn new words from context and extend linguistic forms (Language Process). It requires no formal training and can be automatically administered and scored using its software program. It was normed on a US sample of 898 geographically, socioeconomically, and demographically diverse, monolingual English-speaking, typically-developing children, ages 3;0 to 6;11. Within this normative sample, the screener has been shown to have good concurrent validity against standardized assessments of vocabulary (Peabody Picture Vocabulary Test 4th Edition; PPVT-4; Dunn & Dunn, 1997) and receptive language development (Preschool Language Scales, 4th Edition Auditory Comprehension; PLS-4:AC; Zimmerman et al., 2011), and excellent internal reliability (high Cronbach's alpha; Golinkoff et al., 2017).

According to the user's manual (Golinkoff et al., 2017), the QUILS was designed in part to address the need for screening tools that can be implemented with linguistically, socioeconomically, and culturally diverse populations. Thus, all items were selected to be culturally and dialectically neutral, meaning that they do not place children who speak dialects such as AAE at a disadvantage. For example, even though the past-tense *-ed* is known to be a clinical marker for DLD in speakers of Standard American English, it is not obligatory in AAE (Pruitt & Oetting, 2009). For this reason, the QUILS assesses an alternative linguistic structure, the past tense copula and auxiliary verb "was" since it is obligatory in both SAE and AAE ("Where *was* the hat?" rather than, "What happen*ed* to the hat?"). Additionally, the QUILS examines children's skill in learning new words and structures (Language Learning Process), which may be less dependent on prior knowledge or experience than tests of existing vocabulary

knowledge (Campbell et al., 1997; Pace et al., 2018). Therefore, the QUILS has the potential to reduce identification bias if used as a screener with diverse populations.

Previous research with the QUILS (Aravind et al., 2018; Levine et al., 2018) has examined data from the normative sample, which only included children with typically developing language in alignment with recommended practice for test development when identification (rather than the documentation of severity) is the goal (Peña et al., 2006). The clinical utility of a language screener, however, relies on its ability to accurately classify children into groups "at risk" or "not at risk" for language disorder (Dollaghan, 2007). To date, the classification accuracy of the QUILS has not been investigated with clinical populations. A primary aim of the current research was to describe the performance of diverse preschool-age children with and without language disorder on the QUILS and to examine the classification accuracy of the QUILS in two independent samples that differed in prevalence of disorder and in the diagnostic reference standard used to identify the presence of disorder. In the following section, we describe standard measures used to determine a tool's classification accuracy applied in the current research.

1.2. Standard Measures of Classification Accuracy

The present study utilizes discriminant analyses to examine the clinical utility of the QUILS as a language screener in two diverse samples of preschool-aged children. Two of the key indices for determining a screening tool's classification accuracy are sensitivity and specificity. *Sensitivity* refers to a measure's ability to correctly identify children with language disorders as having language disorders (i.e., the proportion of children with confirmed DLD who are identified as such by the measure). *Specificity* refers to the measure's ability to correctly reject the presence of DLD in those children who do not have it (i.e., the proportion of children

with confirmed typical language development who are accurately identified as such by the measure). In practice, sensitivity is often the more important criteria for screeners so as to catch the greatest number of children who need further evaluation, although ideally a balance between sensitivity and specificity is struck so that school districts and families are not overburdened by unnecessary referrals (Bishop, 2017). Population based studies support the practice of over-identification with screenings, aiming for ~30% failure rate (i.e., children who will be referred for a full diagnostic evaluation) given an expected 7-10% prevalence of DLD (Tomblin et al., 1997). Sensitivity and specificity are known to be susceptible to the participant characteristics within a sample and are therefore not recommended as the only measures of classification accuracy (Dollaghan, 2007).

One statistical approach that determines discriminability when used in combination with sensitivity and specificity is signal detection analysis (Treat & Viken, 2012). Within this approach, Receiver Operator Characteristic (ROC) curve plots can be used to determine accuracy in identifying children at risk for DLD against accuracy in identifying their typically developing peers using a specific instrument such as the QUILS (Swets, 2014). The Area Under the Curve (AUC) measures the entire two-dimensional area underneath the curve of the ROC plot (from the coordinates 0,0 to 1,1) and can be interpreted as the average sensitivity over all points on the curve. Thus, an AUC of .5 would indicate no discrimination (i.e., the tool is unable to detect children with and without DLD), .7 to .8 is considered acceptable, .8 to .9 is considered excellent, and an AUC of 1 would be an indication of perfect discrimination.

Finally, likelihood ratios (LR) indicate the degree of confidence that a given test score would be expected for a child with the target disorder (in this case, DLD) compared to the likelihood that the same test score would be expected in a child without the target disorder and

are robust against the proportion of disorder in a given sample (Dollaghan, 2007). A positive LR [LR+; sensitivity/(1- specificity)] indicates how much the odds of a diagnosis increase given a positive diagnosis with the specific instrument (Mant, 1999). For example, a LR+ of 3 indicates that a child identified as language disordered on the predictor is three times more likely to have a language disorder than to have language in the typical range. Positive LR values ≥ 10 are highly informative, from 5 to 10 are moderately informative, and from 2 to 5 are modestly informative to "rule in" the disorder; an LR+ over 10 is very strong evidence to support the likelihood of the disorder (i.e., DLD). A negative LR [LR-; (1-sensitivity)/specificity)] gives the odds of having a diagnosis in children with a negative test. Negative LR values between 0 and 1 provide less evidence for the presence of DLD; values closer to zero are more likely to "rule out" the disorder. For instance, a LR- of .1 indicates that a child identified with typical language development on an instrument is 1/10 as likely to have a language disorder. LR– values ≤ .10 are highly informative, between .1 to .2 are moderately informative, and between .2 to .5 are modestly informative.

1.3. Factors that Influence Classification Accuracy

Various factors can influence the sensitivity and specificity of screening instruments. For example, classification accuracy depends on the selection of the diagnostic reference standard that is used to identify whether a child truly is or is not at risk for impairment (Dollaghan, 2007). Although there is no single gold-standard measure for diagnosing DLD, validation studies have reported a variety of evidence-based approaches including informed clinical judgement, parent and teacher report, enrollment in speech and language intervention, performance on standardized tests or a combination of these methods (Dockrell & Marshall, 2015). Because diagnostic criteria may vary widely across context, region, and sample (Nitido & Plante, 2020), the current study

compared the classification accuracy of the QUILS in two clinical samples using different gold-standard diagnostic instruments to determine the presence or absence of disorder (Umemneku et al., 2019).

In addition to the gold-standard measure that is selected, classification accuracy is also known to be influenced by the cutoff for eligibility decisions and the composition of the sample, including the prevalence and severity of disorder (Simel et al., 1991; Tomblin, 2010). The QUILS screener is intended to be implemented widely with diverse preschool children across educational, clinical, community, and research contexts, and it is therefore important to evaluate the potential for identifying language disorder within different clinical populations. To further evaluate the clinical utility of the QUILS, the present research investigated performance on the QUILS screener in children with and without language impairment and examined the instrument's classification accuracy in two independent samples.

2. Study 1

Prior research has described performance on the QUILS in typically developing children only. A primary aim of the present study was to describe performance on the QUILS in a high-prevalence clinical sample including children diagnosed with DLD ($N = 54$) and children with typically developing language (TD; $N = 13$). We hypothesized that children diagnosed with DLD would perform significantly below children with TD on all QUILS indices. A second key aim of this study was to examine classification accuracy of the QUILS using the established cutoff from the standardization sample (below 25th percentile) to identify children with and without impairment. We used discriminant analysis to determine sensitivity, specificity, and likelihood ratios at the specified cutoff. We asked:

1) Does performance on the QUILS vary significantly between children with typically developing language and those diagnosed with DLD?
2) Do QUILS index scores accurately predict group membership (TD vs. DLD) using the recommended cutoff from the standardization sample (<25th percentile)?

2.1 Methods

2.1.1 Participants

This clinical sample was derived from two projects conducted in the Eastern US: a summer camp for children diagnosed with or at risk of Developmental Language Disorder (DLD); and a clinical intervention program that enrolled children with DLD as well as a comparison group of children with typical language development. Participants were identified through a combination of community recruitment and clinical referral. A total of 67 children aged 3;0 to 6;9 participated (*M* = 52.7 months; *SD* = 11.0; range = 36.0 to 81.5 months). Four participants were excluded from the final sample because they were younger than 36 months (*N* = 2) or they did not provide QUILS data (*N* = 2). Participant demographics are presented in Table 1 for the full sample as well as the TD and DLD groups. Participants were monolingual English-speakers according to parent report, although one participant received exposure to 20% Patois. We opted to retain this participant's data because group assignment criteria followed best practice for differentiating DLD from TD in linguistically and dialectically diverse samples (Li'el et al., 2019). SES was determined by primary caregiver education reported in years: 12 years or fewer was classified as low-SES (equivalent to a high school degree); 13 years or greater was classified as mid- to high-SES (equivalent to some college or above). On average, primary caregivers reported 15.9 years of education (*SD* = 2.52; range 12 to 22).

2.1.2 Materials

Quick Interactive Language Screener (QUILS). The QUILS is a receptive touch-screen instrument measuring comprehension and developed from the latest research in child language acquisition (Golinkoff et al., 2017). Twelve subtests across three language components sample preschool children's receptive knowledge of specific language constructs (Table 2). Norms are currently available for children ages 3;0 through 6;11. The QUILS can be administered by non-language specialists (e.g., classroom aides) in addition to educators and clinicians and takes approximately 15 minutes to complete. Test items present information visually and aurally, using static illustrations as well as dynamic animations. The QUILS begins with three training items to familiarize children with the touchscreen interface and the general format of the assessment. Children respond to all test items by selecting one of the options on the screen with a finger press. Between each subtest, children view short, animated scenes (e.g., a giraffe flying a plane) and are presented with verbal encouragement from the screener (e.g., "Great job! Let's do some more!").

The Vocabulary Product component measures children's comprehension of open-class words (Nouns, Verbs) and closed-class words (Prepositions, Conjunctions). Open-class words are lexically and semantically meaningful content words and are so-called because this category is constantly adding new members, whereas closed-class words modulate the meaning of sentences by performing grammatical operations and belong to a finite set of words that do not typically accept new members. For example, an item on the Prepositions subtest prompts children to "Find a dog *behind* a black table" while presenting pictures of three distinct scenarios: a dog behind a black table (target), a black dog behind a brown table (foil), and a dog in front of a table (foil; Figure 1).

The Syntax Product component measures children's comprehension of various syntactic structures that emerge during the preschool period such as sentences that refer to past actions and locations, prepositional phrases, embedded clauses, and wh-questions. For example, an item from the Past Tense subtest began by presenting the following animated scenario: "Look! The boy is wearing a hat! Oh no! The hat blew off!". The narrator then asked children, "Where *was* the hat?" and three potential locations, each surrounded with a yellow box, were presented on the computer screen: the boy's head (target), a hat on the girl's head (foil), and the hat in the air (foil; Figure 2).

The QUILS not only measures children's comprehension of vocabulary and syntax, but also includes a Process component which assesses how children learn *new* vocabulary words and syntactic structures. Each Process subtest was developed from the experimental methods used to investigate language acquisition in the empirical literature (Carey & Bartlett, 1978; Dollaghan, 1985; Fisher, 2002; Golinkoff et al., 1996; Naigles, 1990; Naigles et al., 1993). The Noun Learning subtest measures children's skill at fast mapping and extending novel nouns to unfamiliar objects; the Adjective Learning subtest measures children's skill at fast mapping and extending novel adjectives to novel properties of familiar objects; the Verb Learning subtest measures children's skill at inferring the meaning of a novel verb from the syntactic structure of the sentence and extending the new verb to an unfamiliar context (e.g., syntactic bootstrapping); and the Converting Actives to Passives subtest measures children's skill at comprehending the conversion of novel verbs from active voice to passive voice.

An item from the Noun Learning subtest, for example, presents four images on the screen and prompts children: "The *fep* is blue. Show me the blue *fep*". Children are expected to use the process of mutual exclusivity to fast map the novel noun to one of four potential objects: the blue

fep (target); a blue crayon (foil); a blue stroller (foil); a novel golden object which does not meet the color criterion (foil; Figure 3a). To succeed on this item, children must eliminate known nouns (crayon and stroller) and the unfamiliar item that does not meet the criterion (i.e., golden object) from the set of possible referents. After their mapping selection, on the following screen children are presented with four new extension options and asked, "Can you find another *fep*": a green *fep* (target); a blue cup (foil); a lightbulb (foil); and a robot (foil) (Figure 3b). Children must extend the novel noun "*fep*" to a novel exemplar that is perceptually distinct from the original (green rather than blue). Children are required to answer correctly on both the mapping and extension trials to receive credit for each item on the Noun Learning and Adjective Learning subtests. On all other subtests, items include a single trial.

In the commercially available version of the QUILS, the computer software automatically converts a child's raw score (i.e., the total number correct out of 48 items) to an age-based standard score and corresponding percentile rank based on the full normative sample. Norm-referenced standard scores and percentile ranks are reported for the child's overall performance (i.e., all 48 items) as well as for each of the three component areas assessed (16 items on each QUILS component: Vocabulary, Syntax, and Process). Since the overarching purpose of a screener is to identify risk and not to diagnose, the QUILS relies on a 25th percentile cut score to reflect the recommended population rates of identification (Tomblin et al., 1997). Failure is obtained by performance below the 25th percentile on: 1) both the Vocabulary and Syntax components; 2) the Process component; and/or 3) the Overall percentile score (average of the three component scores Vocabulary, Syntax, and Process). Notably, vocabulary alone is not considered to be a robust indicator of language impairment (Gray et al., 1999) and therefore, a percentile score below 25 on only the Vocabulary component of the QUILS does not justify

referral. Percentiles were selected for referral recommendations instead of standard scores because percentiles are well known metrics of children's development (e.g., used in well-child visits, etc.) and the QUILS is designed to be implemented by educators or paraprofessionals who may not have in-depth training in psychometrics.

For discriminant analyses in the present study, a fourth index representing children's minimum percentile score (i.e., MIN Score) on any one of the three QUILS indices was created to capture heterogeneity in children's language skills. For example, if a child scored at the 30th percentile on Vocab + Syntax, the 15th percentile on Process, and the 28th percentile overall, their MIN Score would be the 15th percentile, reflecting their lowest score across these three indices.

Structured Photographic Expressive Language Test (SPELT-P2 and SPELT 3). The Structured Photographic Expressive Language Test–Preschool 2nd Edition (SPELT-P2; Dawson et al., 2005), evaluates morphosyntax skills from 3;0 through 5;11 years old. The Structured Photographic Expressive Language Test–3rd Edition (SPELT-3; Dawson et al., 2003), evaluates morphosyntax skills from 4;0 through 9;11. In both tests, children view color photographs and are verbally prompted to provide spoken responses for target structures. For example, one test item on both tests shows a photograph of a girl with a glass of juice. The examiner says, "This girl has some juice. What do you think will happen next?" For this item, a response containing either "will" or "is going to" receives credit. Many of the items overlap across the two tests, but the SPELT-P2 provides additional scaffolding for responses compared to the SPELT-3. In a clinical study of 4- and 5-year-old children, a sensitivity of .90 and specificity of 1.00 was reported for the SPELT-P2 with a standard score cutoff point of 87 used to determine group membership (Greenslade et al., 2009). A study of diagnostic accuracy using the SPELT-3 showed sensitivity of .9 and specificity of 1.00 for 4- and 5-year-old children with a cutoff score

of 95 (Perona et al., 2005). Children who were reported to be speakers of Standard American English (SAE) and/or below age 4 received the SPELT-P2 ($N = 21$) and SAE speakers above age 4 received the SPELT-3 ($N = 15$).

Diagnostic Evaluation of Language Variation – Norm Referenced (DELV-NR). The DELV-NR (Seymour et al., 2003) is a diagnostic test for children aged 4 through 9 that provides a standardized approach to distinguishing between speech and language differences versus disorders and is designed to be suitable for children who speak a dialect of English other than Standard American English (SAE), such as African American English (AAE), as well as for those who speak SAE. The Syntax Subtest from the DELV assesses deep syntactic knowledge of sentence structure and was administered to all children above age 4;0 who were reported to be speakers of a dialect of American English other than SAE ($N = 31$).

Receptive and Expressive Vocabulary. Children's receptive vocabulary was measured with the Peabody Picture Vocabulary Test – 4th Edition (PPVT-4; Dunn & Dunn, 1997); expressive vocabulary was measured with the Expressive Vocabulary Test – 2nd Edition (EVT-2). These standardized instruments are used to describe children's vocabulary development but not used diagnostically for classification of participants (Gray et al., 1999).

Cognitive Measure. Children were administered the Matrices subtest ($N = 36$) or the Picture Similarities ($N = 31$) subtest from the Differential Abilities Scales–II (DAS II; Elliott et al., 2018). To rule out cognitive impairment, participants had to receive *t*-scores greater than 35 on these measures in accordance with the clinical guidelines for impairment.

2.1.3 *A Priori* Classification

Children were assigned to the *a priori* DLD group based on standardized assessment scores in combination with clinical expertise and a history of language intervention or parental

concern. Children enrolled in the summer camp for children diagnosed with DLD (n = 37) were classified on the basis of standard scores on the SPELT-3 that fell below 95 (according to the empirically derived cutoff score recommended in the SPELT-3 manual) or scaled scores on the DELV-NR Syntax subtest that fell below 7, in combination with a history of language intervention and/or scores on a NonWord Repetition task that were low for their age band (Dollaghan & Campbell, 1998). Children who did not meet these classification criteria were excluded from the present study because descriptive testing, including the QUILS, was not completed. Children in the clinical intervention program were classified in the DLD group if their standard scores on the SPELT-P2 fell below 87 (the empirically derived cutoff score recommended in the SPELT-P2 manual; Greenslade et al., 2009) or their scaled scores on the DELV-NR Syntax subtest fell below 7 (n = 17). Children who did not meet these criteria were classified as typically developing (TD; n = 13). Children in the TD group also passed hearing and cognitive screens and had no parental or clinical concerns about language disorder. Thus, a combined total of 54 participants met the criteria for DLD (~80%) and 13 participants met the criteria for TD (~20%). Targeted recruitment of children with or at risk for DLD yielded a rate of impairment that was substantially higher than would be expected in a population sample of the disorder (Rice & Hoffman, 2015).

2.1.4 Procedure

Each participant was first tested individually at a location convenient to the family (e.g., Lab, school, public library, etc.) prior to the start of the camp. Test administrators held a minimum of a two years of college and included research assistants, laboratory coordinators, graduate student clinicians, postdoctoral fellows, and research SLPs. In general, a DLD diagnosis was confirmed first and then the QUILS was administered. Testing might stretch over several

days to include DLD and all descriptive testing at a pace comfortable for the child and family. Testing sessions usually lasted 45-60 min. For the QUILS, the administrator sat beside the child at a small table and asked the child to practice touching the computer screen before beginning the assessment to ensure that the child knew the appropriate way to press the screen so that each response would be registered by the program. Once the training items had begun, the administrator did not provide any additional feedback; if the child requested information about a test item or failed to select a response, the administrator simply encouraged the child to continue playing or make their best guess. The DELV or SPELT was administered within a four-week period of the QUILS; standard administration and scoring procedures were followed for all participants.

2.2 Planned Analyses

To answer Research Question 1, descriptive statistics were calculated to describe children's performance on the QUILS and inferential statistics examined group differences (TD vs. DLD). We also created language risk profiles to examine whether the QUILS index scores were useful in differentiating children at no risk from children at mild (i.e., below cutoff on one index), moderate (i.e., below cutoff on two indices), or high risk (i.e., below cutoff on all indices) for impairment in this high-prevalence clinical sample. To answer Research Question 2, we used discriminant analyses to examine the classification accuracy of the QUILS in a high-prevalence clinical sample using the norm-referenced cutoff established in the standardization sample (< 25^{th} percentile). Because we were interested in evaluating a single cutoff score, we used sensitivity and specificity analyses only (rather than ROC curves, which depict the trade-off between sensitivity and specificity for every possible cut-off score).

2.3 Results

Table 3 summarizes the descriptive results for DLD and TD participants in Study 1. Mean scores for children in the TD group closely approximated the sample mean for the SPELT-P2, SPELT-3, and the DELV-NR and were higher than the sample mean for the PPVT and EVT. In comparison, mean scores for the DLD group fell more than 1 SD below the mean on either the SPELT or the DELV, and the group average was slightly lower than the sample mean for the PPVT and EVT. Descriptive statistics for all QUILS index scores can also be found in Table 3. Mean percentile score on all QUILS indices in the DLD group were below those of the TD group (falling more than -.5 SD below the group mean), but there was also substantial within-group variability resulting overlap in the score distribution. One-way ANOVAs confirmed significant differences across all QUILs indices between diagnostic groups, all *F*s (1, 65) = 20.9-29.3, *p*s = .000. QUILS scores in this sample did not vary by SES, $F(1, 61) = .74$, $p = .394$.

To examine whether the QUILS index scores were useful in differentiating children at no risk from children at mild, moderate, or high risk for impairment, we created language severity profiles based on the number of indices that fell below the 25th percentile (ranging from 0 to 3). Children who scored at or above the 25th percentile cutoff on all QUILS indices (Vocab + Syntax, Process, or Overall) were not considered to be at risk for DLD. Children who received a score below the 25th percentile on only one of the three QUILS index scores were considered to be at mild risk; children with scores below the 25th percentile on two index scores were considered to be at moderate risk; and children with scores below the 25th percentile on all three index scores were deemed high risk.

A majority of children in the TD group fell into the no-risk category ($n = 12$) and one child fell into the mild risk category because of a percentile score below the cutoff on the Process index. For children in the current sample accurately identified on the QUILS in the DLD group

using the 25th percentile cutoff across any index (n = 35), 9 (16.6%) fell into the mild risk category, 9 fell into the moderate risk category (16.6%), and 17 fell into the high risk category (31.5%). Notably, a one-way ANOVA revealed a main effect of QUILS risk profile (0, 1, 2, 3) on children's DELV-NR scores, $F(3, 27) = 3.7$, $p = .023$, with children at higher risk scoring significantly below children at lower risk. The same main effect emerged for children's scores on the SPELT-P2, $F(3, 17) = 15.2$, $p < .000$ and the SPELT-3, $F(2, 12) = 12.5$, $p = .001$; degrees of freedom differ because children who completed the SPELT-3 were classified in only three of the four risk categories.

To examine classification accuracy, we used the QUILS recommended cutoff score (below the 25th percentile) to differentiate between the *a priori* DLD and TD groups. We calculated sensitivity, specificity, and likelihood ratios for all four indices (V+S; P; O; MIN; Table 4). Out of the four index scores, MIN Score yielded the best classification accuracy at the 25th percentile cutoff with fair sensitivity at 65.0% and high specificity at 92.0%. These results yield a moderately informative positive likelihood ratio of 8.13 and a modestly informative negative likelihood ratio of .38. Thirty-five of the children who were classified as DLD based on the study criteria were correctly classified by the QUILS MIN Score, with 19 misclassified as TD at the 25th percentile cutoff.

To investigate potential reasons for under-identification, demographic variables and children's performance on measures of receptive and expressive vocabulary were examined. A scatterplot (Figure 4) depicts children's standard scores on the PPVT-4 and QUILS Minimum percentile scores (MIN Score) by QUILS classification group (True Positive or DLD, True Negative or TD, False Positive, and False Negative). Children who were misclassified as TD on the basis of the QUILS at the MIN Score 25th percentile cutoff had relatively high scores that fell

within the broad average range on receptive (n = 19, M = 101.1, SD = 9.4, range = 82 to 118) and expressive (n = 17, M = 98.9, SD = 10.3, range = 82 to 120) vocabulary measures compared with the averages for children with accurate classification as DLD (n = 35, PPVT M = 87.42, SD = 12.0, range = 61 to 117; EVT M = 91.3, SD = 10.5, range = 64 to 116), although there was significant variability in both groups. Demographic characteristics including age, race, and sex varied. In this group of 19 misclassified participants, the average age was 52.4 months (SD = 12.8 months); 14 were male; 14 were white, 3 were Black or African American, 1 was Native American or American Indian, and 1 was multiracial.

Twelve of the children who were classified as TD by the study criteria were accurately identified by the QUILS MIN Score at the 25th percentile cutoff, with one child incorrectly identified as having or at risk for DLD, yielding a false positive error rate of 8.3%. The misclassified child had standard scores in the high average range on the PPVT, EVT, and SPELT (SS = 105, SS = 117 and SS = 111, respectively). This child's QUILS percentile score on the Vocabulary + Syntax index was also in the high average range (percentile score = 74.2). However, this child's score on the QUILS Process index was more than -2 SD below the group mean (percentile score = 5.4), which classified them as at-risk for language impairment.

Using the QUILS cutoff from the normative sample (< 25th percentile) yielded high specificity but only fair sensitivity within this high-prevalence clinical sample. We consider possible explanations and clinical implications for these findings in the general discussion section. In the following study, the classification accuracy of the QUILS is examined in a community sample that approximated the type of preschool classroom context where screening would be likely to occur.

3. Study 2

The QUILS screener is intended for use in a classroom or community context, which would allow for widespread screening of language skills in children aged 3 through 6. In Study 2, we investigated the classification accuracy of the QUILS as a language-screening tool in a sample of 126 participants who completed the QUILS and a widely used standardized reference, the Preschool Language Scales, 5th Edition (PLS-5 Auditory Comprehension). We asked:

1) Does performance on the QUILS vary significantly between children with typically developing language and those diagnosed with DLD?
2) Do QUILS index scores accurately predict of group membership (TD vs. DLD) using the recommended cutoff from the standardization sample (<25th percentile) compared with (a) a balanced score which maximized both sensitivity and specificity and (b) a weighted cutoff which prioritized sensitivity over specificity and met the criteria of setting sensitivity at a minimum of .80 and specificity at a minimum of .70?

3.1 Method

3.1.1 Participants

The sample included a total of 126 children aged 3;1 to 5;11 (*M* = 55.9 months; *SD* = 8.4 months; range = 37.7 to 69.9 months) who completed the QUILS and the Preschool Language Scales Auditory Comprehension subtest (PLS-5 AC). Note that participants in this study were all younger than 6;0 because the original normative sample only included children between 3;0 and 5;11 (though the QUILS was later normed on children between 6;1 and 6;11; Jones & Lesaux, 2021). Participants were recruited from a university speech and hearing clinic, a public school with inclusive preschool and kindergarten classrooms, and preschool programs and Head Start centers in the areas surrounding four university sites in Delaware, Pennsylvania, Massachusetts,

and Washington State. Of these children, 14 were currently receiving clinical services for speech-language pathology; the remainder (n = 112) formed part of the norming sample for the QUILS, had no known history of speech or language disorder, and were included because they completed both the QUILS and the standardized assessment administered for validation (PLS-5 AC). Participant demographic data were collected with a questionnaire that was completed by a parent; these are reported in Table 5. According to parent report, all participants were monolingual English-speaking. Socioeconomic Status (SES) was determined by education level of the primary caregiver. Mid to high SES classification included caregivers who (1) earned a bachelor's degree or (2) earned a graduate degree. Low-SES classification included caregivers who (1) did not complete high school; (2) earned a high school degree or GED; or (3) attended trade school or earned an associate's degree.

3.1.2 Materials

QUILS. All participants completed the QUILS screener; administration procedures were identical to Study 1.

Preschool Language Scales 5th Edition (PLS-5). The Auditory Comprehension subtest of the Preschool Language Scale 5th Edition (PLS-5:AC; Zimmerman et al., 2011) was used as a reference standard for *a priori* group membership. Specific skills assessed on the PLS-5 AC include comprehension of basic vocabulary, morphology, and syntax, making this instrument a good match for the skills measured with the QUILS. This instrument was selected as a formal, standardized measure because evidence from previous studies has shown the PLS-5 to have fair to adequate levels of discriminant accuracy for preschool children and it was used in the normative sample to document convergent validity with the QUILS. The PLS-5 manual reports the sensitivity for Auditory Comprehension scores in a matched sample of children aged 3;0

through 7;11 with and without DLD to be .83 and specificity to be .77 with a cut score 1 SD below the mean (Zimmerman et al., 2011; p. 92). In the present study, children in the TD group had an average PLS AC standard score of 103.2 (*SD* = 11.9; range 86 to 136) whereas those in the DLD group had an average standard score of 78.4 (*SD* = 5.9; range 66 to 85).

3.1.3 *A Priori* Classification

Children with Standard Scores over 85 were assigned to the typically developing (TD) group (*n* = 101; *M* = 55.8 months; *SD* = 8.4 months) whereas children whose Standard Scores fell at or below 85 were assigned to the group with DLD (*n* = 25, *M* = 56.4 months; *SD* = 8.4 months), based on recommendations in the PLS-5 manual (Zimmerman et al., 2011). The overall prevalence of language disorder within this *a priori* sample was ~20%, which falls within the broad range reported by large scale screening studies in preschool and kindergarten populations (Tomblin et al., 1997). Demographic data for participants by group assignment (TD vs. DLD) is also presented in Table 5.

3.1.4 Procedure

Each participant was tested individually at the child's school or clinic. Test administrators held a minimum of a Bachelor's degree and included research assistants, laboratory coordinators, graduate student clinicians, and postdoctoral fellows. For the QUILS, the administrator sat beside the child at a small table and asked the child to practice touching the computer screen before beginning the assessment to ensure that the child knew the appropriate way to press the screen so that each response would be registered by the program. Once the training items had begun, the administrator did not provide any additional feedback; if the child requested information about a test item or failed to select a response, the administrator simply encouraged the child to continue playing or make their best guess. The PLS-5 was administered

within a two-week period of the QUILS; standard administration and scoring procedures were followed for all participants.

3.2 Planned Analyses

Children's performance on the QUILS was examined descriptively and inferentially. Signal detection analyses were used to evaluate the QUILS as a screening instrument for preschool children's language skills. ROC analyses were conducted in SPSS (SPSS Version 26.0; IBM Corp., 2019). Each QUILS index was entered as a unique predictor (Vocabulary + Syntax; Process; Overall; Minimum). Children's standard score on the PLS-5 AC was entered as the dependent measure for all ROC analyses. We hypothesized that the MIN Score would yield the highest classification accuracy because it captures variability in profiles for children at risk for DLD. We compared classification accuracy at the norm-referenced cut point (<25th percentile) with two empirically derived cut points using discriminant analyses: a balanced score that maximized sensitivity and specificity as well as a weighted score that prioritized sensitivity and set the minimal acceptable criterion for sensitivity >.80 and specificity > .70 to maximize the potential of the screener to detect children who should be referred for more comprehensive language evaluation.

3.3 Results

Descriptively, mean percentile scores on the four QUILS indices showed substantial variability in both the TD and the DLD groups (Table 6). For all indices, scores for the DLD group are below those of the TD group, but – as in Study 1 – there was also noteworthy overlap in the score distribution. Consistent with findings from Study 1, one-way ANOVAs revealed that QUILS percentile scores across all indices were significantly lower for children in the DLD group than children in the TD group, Fs (1, 124) = 22.1–38.9, all ps = .000. Moreover, children's

PLS AC standard scores also varied by QUILS risk profile (0 = no risk, 1 = minimal risk, 2 = moderate risk, 3 = high risk), $F(3, 122) = 30.0$, $p = .000$. Children with "no risk" on the basis of the QUILS indices had the highest PLS standard scores ($M = 108.3$; $SD = 12.2$) whereas children with "high risk" had the lowest average PLS scores ($M = 85.6$; $SD = 12.3$). For the participants who were classified as DLD on the QUILS and the PLS-5 ($n = 17$), 3 were below 25 on a single index; 2 were below 25 on two indices; and 12 were below 25 on all three indices. It is notable that a majority of the participants accurately classified as DLD on the QUILS performed below the cutoff on all three QUILS indices. In this study, children in the low-SES group had minimum percentile scores that were significantly lower than children in the mid-to-high SES group, $F(1, 124) = 12.2$, $p = .001$, which is consistent with evidence from the full normative sample (Levine et al., 2018). However, SES-based differences in QUILS scores did not result in over-identification of low-SES children into the DLD group; rather, group membership reflected the overall distribution of the sample.

As a first step in our discriminant analysis, ROC curves were calculated for all four QUILS index scores. Overall classification accuracy for each index, measured by the area under the ROC curve (AUC), is presented in Table 7. The AUC for all dependent measures ranged from .783 to .863, indicating fair to good overall discriminability in this sample. As predicted, the AUC was largest when MIN Score was entered as the dependent variable (.863; 95% CI = .79 to .94). The ROC plot using the QUILS MIN Score for all participants is displayed in Figure 5. Next, we calculated sensitivity, specificity, and likelihood ratios for each QUILS index using the recommended 25th percentile cut score (also presented in Table 7). Although specificity values were fair to high on all indices (.77 to .90), only the MIN Score yielded a sensitivity value that approached the recommended minimum for screeners (.76). Using the MIN Score at the 25th

percentile cutoff, 23 children with typical language development on the PLS were identified as at risk on the QUILS (i.e., false positives) and 6 children in the DLD group based on PLS scores were identified as typically developing on the QUILS (i.e., false negatives).

Next, we ran sensitivity and specificity analyses without specifying the clinical cutoff of the QUILS. With this method, an empirically derived cutoff score that maximized the classification accuracy of the QUILS was generated, which yielded identical results: the recommended QUILS cutoff score below the 25th percentile (which reflects a standard score of 90), was the optimal cutoff to balance and maximize sensitivity and specificity. We also empirically derived a weighted cut point along the curve that set sensitivity at a minimum of .80 and specificity at a minimum of .70 when predicting group membership. Results showed that increasing the cut point to the 32nd percentile met these criteria, yielding a sensitivity value of .84, with a corresponding decrease in specificity to the minimal acceptable value of .70. The LR+ and LR– at this cutoff were both modestly informative (2.8 and .22 respectively).

Children who were misclassified at the 32nd percentile included 31 without impairment (i.e., false positives) and 4 with impairment (i.e., false negatives). Comparing this model to the 25th percentile cut point yielded an increase in sensitivity from .76 to .84 and a corresponding decrease in the false negative rate from 24% to 16%. Improvement to sensitivity while still maintaining the minimal acceptable value for the tool's specificity, suggests that the clinical utility of the QUILS – while acceptable at the 25th percentile cut point – could be improved by screening at the 32nd percentile to maximize sensitivity.

4. General Discussion

The present research evaluated the clinical utility of the QUILS computerized language screener for identifying preschool children for DLD in a high-prevalence clinical sample (Study

1) and a sample that approximated population levels of DLD (Study 2). Examination of group mean differences on the QUILS in both studies indicated that the screener significantly differentiated children identified as typically developing from those who were diagnosed with developmental language disorder. Classification accuracy of the language screener was promising when applied to the lower prevalence sample (Study 2). Severity profiles provided converging evidence for children's level of risk. Together, this research suggests that the QUILS can be used with modest to moderate confidence to screen preschool-aged children for DLD.

4.1 Describing receptive language in clinical samples with the QUILS

Our first research question in both studies asked whether the QUILS was a valid measure of children's receptive language skills in two independent clinical samples. Overall, QUILS index percentile scores (Vocabulary + Syntax; Process; Overall; and Minimum) successfully differentiated between children assigned to TD compared with DLD groups. Moreover, risk profiles based on the number of QUILS index scores that fell below the recommended criterion ($<25^{th}$ percentile) converged with evidence from scores across all standardized assessments (PLS-5; DELV-NR; SPELT-P2; and SPELT 3) for describing children who may be at mild, moderate, or high risk for DLD. It is important to emphasize that children should be referred for more comprehensive diagnostic evaluation regardless of combined risk; that is, a score below the 25^{th} percentile on *any* of the QUILS index scores warrants further assessment. These findings suggest that the QUILS can be used to effectively describe language comprehension skills within clinical samples that vary in prevalence of disorder.

Although children with language impairment tend to score somewhat lower on average than their typically developing peers on language tests, their scores also show overlap with the normal distribution (Peña et al., 2006; Spaulding et al., 2006). Results from the current research

align with existing evidence: the mean group differences between scores of children classified as DLD and those of their typically developing peers on QUILS components were discriminant, although groups showed substantial variability and overlap between distributions in both samples. Identifying children who present with mild impairment on language screeners is important because even mild impairment is known to have significant impact on children's daily academic functioning and quality of life indicators (Eadie et al., 2018). However, differentiating children with mild impairments from typically developing children on the lower end of the normal distribution is a challenge in practice as well as an ongoing issue for theoretical debate (Spaulding et al., 2006; Dockrell & Marshall, 2015). Recent research has suggested that DLD may be better conceptualized as a continuum disorder so that treatment can address severity rather than perseverating on identification and treatment of distinct subtypes (Lancaster & Camarata, 2019). QUILS risk profiles may provide a practical solution to describing severity. Future research should investigate whether risk profiles are useful for guiding diagnostic evaluation or intervention decisions.

4.2 Classification Accuracy of the QUILS

Our second research question compared classification accuracy using QUILS-recommended versus empirically-derived cutoffs in two clinical samples. Children's minimum score (i.e., MIN score) was the optimal index for identification in both studies. A key reason that the MIN score outperformed individual indices is that it accounted for heterogeneity in children's language skills (LARRC, 2015; Lonigan & Milburn, 2017; Tambyraja et al., 2015). For instance, a preschool child with average vocabulary comprehension skills may still have difficulty understanding finite verb morphology, irregular past tense, or other syntactic structures (Rice & Wexler, 1996). Moreover, a child with receptive vocabulary and syntax skills in the

typically developing range may still have difficulty when they encounter a new vocabulary word or morphological form for the first time (Gordon et al., 2021). Thus, relying on a qualifying minimum score across indices strengthens the QUILS' ability to accurately identify children who should be referred to a speech-language pathologist for further evaluation.

Obtained values of sensitivity and specificity from Study 2 were promising for a brief screener that has the potential to be implemented widely in a classroom or community context. When the QUILS recommended cutoff (<25th percentile) was applied, specificity was high in Study 1 (92.3%) and fair in Study 2 (77%), reflecting false positive rates within the acceptable range. This finding is clinically relevant because it means that children with typically developing language were unlikely to be referred for further evaluation based on their QUILS scores. Further, these findings emerged despite the present sample's diversity. This is notable given exceedingly high rates of overidentification in children who speak dialects of English that differ from the established standard in prior research (Craig & Washington, 2004).

Sensitivity was somewhat lower, ranging from unacceptable in Study 1 (.65) to adequate in Study 2 (.76) when the 25th percentile cut score was applied. Raising the cut point to the empirically-derived 32nd percentile increased sensitivity above the minimum acceptable value to .84, although this led to a decrease in specificity from 77% to 70% (i.e., 8 additional participants with otherwise typically developing language recommended for further evaluation). This is an unavoidable tradeoff, which requires careful consideration about the costs and benefits of misclassification given the screening context and population. Children who fall within the 25th to 32nd percentile range could be good candidates for Tier 2 Response to Intervention (RTI) approaches implemented in the classroom context that provide supplemental small-group

instruction to students who may benefit from additional strategies or supports but may not meet the need for intensive intervention.

Taken together, it can be concluded with some confidence that a cut point between the 25th and 32nd percentile is likely to correctly identify 76 to 84 percent of children at risk for language impairment to be referred for additional evaluation. Findings from Study 2 support the current standard of best practice, which recommends the use of empirically derived cutoff scores to ensure accurate identification of children at risk for language impairment. Classification accuracy of the QUILS falls within the acceptable range when compared with other commercially available preschool language screeners (Lugo-Neris et al., 2015; Hendricks et al., 2018). Together, these findings suggest that the QUILS is an appropriate instrument to implement as part of a screening protocol within a community sample.

4.3 Factors that contributed to misclassification

Current best practice for screening preschool children for language disorders generally involves a combination of norm-referenced assessments, parent or teacher report, informal probes, and informed clinical judgement (Gallagher et al., 2019; Sim et al., 2019). An important feature of the present research was the use of different gold-standard instruments for group assignment (TD vs. DLD) across studies. Because each study relied upon different commercially available tools, the QUILS screener was cross-validated across two independent samples with different, evidence-based classification criteria. In Study 1, sensitivity did not reach clinically useful levels (i.e., a false negative rate of 33%). This result may be partially explained by the reference instruments used to assign group membership in the high prevalence sample. Specifically, the QUILS is a measure of receptive language, but *a priori* groups were determined in part by scores on measures of expressive language (DELV-NR; SPELT-P2; SPELT-3).

Although there is continued debate about the clinical utility of differentiating expressive from mixed (expressive-receptive) language disorders (Dockrell & Marshall, 2015; Leonard, 2009), it is possible that children who were not identified on the QUILS could be characterized as having language comprehension skills that fell within normal limits in combination with significant impairment in expressive grammar (Deevy & Leonard, 2018; Yarian et al., 2021). Alternatively, it could be possible that this group of misclassified children had learned other compensatory strategies through intervention or educational support that contribute to their performance on receptive language measures such as the QUILS. This interpretation is supported by the finding that this group of children also had significantly higher vocabulary scores than the children who were accurately classified by the QUILS into the DLD group.

These results highlight a tradeoff between the efficiency and ease of an automated computerized screener that can quickly screen a large sample and more comprehensive, detailed screening of language skills across domains and modalities that may result in improved classification accuracy. Testing the QUILS classification accuracy against different commercially available instruments was a strength of this research because it reflects the diversity of diagnostic decision-making that exists in the field. However, low rates of classification accuracy in Study 1 compared with Study 2 suggest that although the QUILS may be used to describe receptive language in both high and low prevalence clinical samples, it should only be applied as a screening instrument in contexts that approximate population levels of the disorder. These results also underscore the importance of using multiple converging measures as components of a comprehensive evaluation to accurately identify children with DLD and not relying on screening measures alone for diagnosis. One way to increase the likelihood of accurate identification would be to retest borderline children 6 months or a year later.

4.4 Clinical implications and future directions

The QUILS joins a growing list of innovative tools and approaches to screening for preschool children developed in response to calls for preventive, universal measures (Greenwood et al., 2011; Kaiser et al., 2022; Sim et al., 2019). The present research suggests that the QUILS can be used as an effective tool to describe receptive language development in children with and without language disorders and may be particularly useful for screening in contexts with limited access to clinical providers (e.g., SLPs). As best practice for early identification involves the use of multiple converging measures (Conti-Ramsden & Durkin, 2012), the QUILS could be incorporated into a screening protocol that includes a brief developmental screener that is already widely implemented (e.g., Ages and Stages Questionnaire; Bricker et al., 1999) or a history of parental or teacher concern. Findings from the current study have implications for screening recommendations and practices, which must strive to meet the greatest public need and should be widely available at low or no cost to families with young children. Addressing the need for increased access and implementation of universal screening will also rely on continued efforts to engage parents and child care providers as collaborators, administer tools in diverse settings, and link screening efforts with effective intervention (Missall et al., 2021).

Additional evidence is needed to evaluate the feasibility of widespread implementation of the QUILS or other language screeners in preschool classrooms, speech and language clinics, or primary healthcare settings. Future research should also investigate additional child-level variables that could result in misclassification (e.g., nonverbal cognition) and extend the current research to children with language disorders secondary to other diagnoses such as Autism Spectrum Disorders. Efforts to develop screening tools that are appropriate for multilingual learners need to be redoubled. There now exists a bilingual Spanish-English version of the

QUILS (QUILS: ES; de Villiers et al., 2021) and similar studies should also examine its classification accuracy in dual language learners acquiring Spanish and English. In future research with the QUILS, it will be important to compare the screener's predictive validity with that of other commercially available screeners and cross-validate the empirically derived cut off scores from this study. Because language development at kindergarten entry is a key predictor of children's later academic and social achievement (Pace et al., 2019), this research may be relevant for efforts by educators and clinicians to increase early detection of children at risk for DLD during the preschool period.

References

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.

Bishop, D. V. (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.

Bricker, D., Squires, J., Mounts, L., Potter, L., Nickel, R., Twombly, E., & Farrell, J. (1999). Ages and Stages Questionnaire. *Baltimore, MD: Paul H. Brookes*.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research*, *40*(3), 519-525.

Conti-Ramsden, G., & Durkin, K. (2012). Language development and assessment in the preschool period. *Neuropsychology Review*, *22*(4), 384-401.

Conti-Ramsden, G., Durkin, K., Toseeb, U., Botting, N., & Pickles, A. (2018). Education and employment outcomes of young adults with a history of developmental language disorder. *International Journal of Language & Communication Disorders*, *53*(2), 237-255.

Dawson, J., Stout, C., & Eyer, J. (2003). Structured Photographic Expressive Language Test - 3rd Edition. DeKalb, IL: Janelle Publications.

Dawson, J. I., Stout, C., Eyer, J., Tattersall, P. J., Fonkalsrud, J., Croley, K., & Janelle Publications (Firm). (2005). *SPELT-P2: Structured Photographic Expressive Language Test*. DeKalb, IL: Janelle Publications.

De Villiers, J., Iglesias, A., Golinkoff, R., Hirsh-Pasek, K., Wilson, M. S., & Nandakumar, R. (2021). Assessing dual language learners of Spanish and English: Development of the QUILS: ES. *Revista de Logopedia, Foniatría y Audiología*, *41*(4), 183-196.

Deevy, P., & Leonard, L. B. (2018). Sensitivity to morphosyntactic information in preschool children with and without developmental language disorder: A follow-up study. *Journal of Speech, Language, and Hearing Research*, *61*(12), 3064-3074.

Dockrell, J. E., & Marshall, C. R. (2015). Measurement issues: Assessing language skills in young children. *Child and Adolescent Mental Health*, *20*, 116–125. doi:10.1111/camh.12072

Dollaghan, C. (1985). Child meets word: "Fast Mapping" in preschool children. *Journal of Speech, Language, and Hearing Research*, *28*(3), 449-454.

Dollaghan, C. A. (2007). *The handbook of evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Dollaghan, C., & Campbell, T. F. (1998). Nonword Repetition and Child Language Impairment. *Journal of Speech, Language and Hearing Research*, *41*(5), 1136-1146.

Elliott, C. D., Salerno, J. D., Dumont, R., & Willis, J. O. (2018). The Differential Ability Scales—Second Edition. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 360–382). The Guilford Press.

Fisher, C. (2002). Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science*, *5*(1), 55–64. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2002-02740-016&site=ehost-live

Gallagher, A. L., Murphy, C. A., 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.

Golinkoff, R. M., de Villiers, J., Hirsh-Pasek, K., Iglesias, A., & Wilson, M. S. (2017). *User's manual for the Quick Interactive Language Screener (QUILSTM): A measure of vocabulary, syntax, and language acquisition skills in young children*. Baltimore, MD: Paul H. Brookes.

Golinkoff, R. M., Jacquet, R. C., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child Development*, *67*(6), 3101–3119.

Gordon, K. R., Storkel, H. L., Lowry, S. L., & Ohlmann, N. B. (2021). Word Learning by Preschool-Age Children With Developmental Language Disorder: Impaired Encoding and Robust Consolidation During Slow Mapping. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4250-4270.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools*, *30*(2), 196-206.

Greenslade, K. J., Plante, E., & Vance, R. (2009). The Diagnostic Accuracy and Construct Validity of the Structured Photographic Expressive Language Test—Preschool: Second Edition. *Language Speech and Hearing Services in Schools*, *40*(2), 150. https://doi.org/10.1044/0161-1461(2008/07-0049)

Greenwood, C. R., Carta, J. J., & McConnell, S. (2011). Advances in measurement for universal screening and individual progress monitoring of young children. *Journal of Early Intervention*, *33*(4), 254-267.

Guralnick, M. J. (2019). *Effective early intervention: The developmental systems approach.* Paul H. Brookes Publishing Co.

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

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*(1), 4.

IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.

Jones, S.M. & Lesaux, N.K., Co-PIs, *Early Learning Study at Harvard (ELS@H).* Saul Zaentz Charitable Foundation (2016-2021).

Kaiser, A. P., Chow, J. C., & Cunningham, J. E. (2022). A Case for Early Language and Behavior Screening: Implications for Policy and Child Development. *Policy Insights from the Behavioral and Brain Sciences*, *9*(1), 120-128.

Knottnerus, J. A., & Muris, J. W. (2003). Assessment of the accuracy of diagnostic tests: The cross-sectional study. *Journal of Clinical Epidemiology*, *56*(11), 1118-1128.

Komesidou, R., & Summy, R. (2020). Developmental Language Disorder: Considerations for Implementing School-Based Screenings. *Clinical Psychology and Special Education*, *9*(3), 34-47.

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.

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development*, *86*(6), 1948-1965.

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: Literacy, mental health, and employment outcomes. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1401–1416. https://doi.org/10.1044/1092-4388(2009/08-0142)

Leonard. L.B., (2009). Is expressive language disorder an accurate diagnostic category? *American Journal of Speech Language Pathology*, *18*(2), 115-123.

Levine, D., Pace, A., Luo, R., Hirsh-Pasek, K., Golinkoff, R. M., de Villiers, J., Iglesias, A., Wilson, M. S. (2020). Evaluating socioeconomic gaps in preschoolers' vocabulary, syntax and language process skills with the Quick Interactive Language Screener (QUILS). *Early Childhood Research Quarterly*, 50(1), 114–228. https://doi.org/10.1016/j.ecresq.2018.11.006

Li'el, N., Williams, C., & Kane, R. (2019). Identifying developmental language disorder in bilingual children from diverse linguistic backgrounds. *International Journal of Speech-Language Pathology*, *21*(6), 613-622.

Lipkin, P. H., Macias, M. M., Norwood, K. W., Brei, T. J., Davidson, L. F., Davis, B. E., ... & Voigt, R. G. (2020). Promoting optimal development: identifying infants and young children with developmental disorders through developmental surveillance and screening. *Pediatrics*, *145*(1). doi: 10.1542/peds.2019-3449

Lipkin, P. H., Macias, M. M., Chen, B. B., Coury, D., Gottschlich, E. A., Hyman, S. L., ... & Levy, S. E. (2020b). Trends in pediatricians' developmental screening: 2002–2016. *Pediatrics*, *145*(4).

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research*, *60*(8), 2185-2198.

Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., & Gillam, R. B. (2015). Utility of a Language Screening Measure for Predicting Risk for Language Impairment in Bilinguals. *AJSLP*, *24*(2), 426–437. https://doi.org/10.1044/2015

Mant, J. (1999). Studies assessing diagnostic tests. In M. Dawes, P. Davies, A. Gray, J. Mant, K. Seers, & R. Snowball. Evidence-based practice: A primer for health care professionals (pp. 67–78). New York: Churchill Livingston.

McGregor, K. K. (2020). How we fail children with developmental language disorder. *Language, Speech, and Hearing Services in Schools*, *51*(4), 981-992.

McGregor, K. K., Goffman, L., Van Horne, A. O., Hogan, T. P., & Finestack, L. H. (2020). Developmental language disorder: Applications for advocacy, research, and clinical service. *Perspectives of the ASHA Special Interest Groups*, *5*(1), 38-46.

Missall, K., Artman-Meeker, K., Roberts, C., & Ludeman, S. (2021). Implementing multitiered systems of support in preschool: Begin with universal screening. *Young Exceptional Children*, *24*(4), 213-224.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of child language*, *17*(2), 357-374.

Naigles, L., Gleitman, H., & Gleitman, L. (1993). Syntactic bootstrapping and verb acquisition. *Language and Cognition: A Developmental Perspective*, *5*, 104-140.

Oetting, J. B., Gregory, K. D., & Rivière, A. M. (2016). Changing how speech-language pathologists think and talk about dialect variation. *Perspectives of the ASHA Special Interest Groups*, *1*(16), 28-37.

Pace, A., Alper, R., Burchinal, M. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly*, *46*, 112-125.

Paul, R. (2020). Language disorders. In *Handbook of Clinical Neurology* (Vol. 174, pp. 21-35). Elsevier.

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology*, *15*, 247–254.

Perona, K., Plante, E. and Vance, R. (2005). Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition (SPELT-3). *Language, Speech, and Hearing Services in Schools*, *36*, 103 - 115.

Pruitt, S., & Oetting, J. (2009). Past tense marking by African American English–speaking children reared in poverty. *Journal of Speech, Language, and Hearing Research*, *52*(1), 2–15. https://doi.org/10.1044/1092-4388(2008/07-0176)

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.

Rice, M.L. (2020). Advances in specific language impairment research and intervention: an overview of five research symposium papers (PDF). *Journal of Speech, Language, and Hearing Research*, *63*, 3219–3223.

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2; 6 to 21 years of age. *Journal of Speech, Language, and Hearing Research*, *58*(2), 345-359.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1239-1257.

Schonwald, A., Horan, K., & Huntington, N. (2009). Developmental screening: Is there enough time?. *Clinical Pediatrics*, *48*(6), 648-655.

Seymour, H. N., Roeper, T. W., & deVilliers, J. (2003). Diagnostic Evaluation of Language Variation (DELV). San Antonio, TX: The Psychological Corporation.

Sim, F., Thompson, L., Marryat, L., Ramparsad, N., & Wilson, P. (2019). Predictive validity of preschool screening tools for language and behavioural difficulties: A PRISMA systematic review. *PLOS One*, *14*(2), e0211409.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*, *37*(1), 61–72. https://doi.org/10.1044/0161-1461(2006/007)

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

Tambyraja, S. R., Schmitt, M. B., Farquharson, K., & Justice, L. M. (2015). Stability of language and literacy profiles of children with language impairment in the public schools. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1167-1181.

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.

Tomblin, J. B., Zhang, X., Buckwalter, P., & Catts, H. (2000). The association of reading disability, behavioral disorders, and language impairment among second-grade children. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *41*(4), 473-482.

Treat, T. A., & Viken, R. J. (2012). Measuring test performance with signal detection theory techniques. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 723–744). American Psychological Association. https://doi.org/10.1037/13619-038

Umemneku Chikere, C. M., Wilson, K., Graziadio, S., Vale, L., & Allen, A. J. (2019). Diagnostic test evaluation methodology: A systematic review of methods employed to

evaluate diagnostic tests in the absence of gold standard–an update. *PLoS One*, *14*(10), e0223832.

Vanderheyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children*, *77*(3), 335-350.

Washington, J. A., & Seidenberg, M. S. (2022). Language and dialect of African American children. In *Handbook of Literacy in Diglossia and in Dialectal Contexts* (pp. 11-32). Springer, Cham.

Williams, K. T. (1997). Expressive Vocabulary Test (2nd ed.) [Measurement instrument].

Weiler, B., Schuele, C. M., Feldman, J. I., & Krimm, H. (2018). A multiyear population-based study of kindergarten language screening failure rates using the Rice Wexler Test of Early Grammatical Impairment. *Language, Speech, and Hearing Services in Schools*, *49*(2), 248-259.

Yarian, M., Washington, K. N., Spencer, C. E., Vannest, J., & Crowe, K. (2021). Exploring predictors of expressive grammar across different assessment tasks in preschoolers with or without DLD. *Communication Disorders Quarterly*, *42*(2), 111-121.

Zablotsky, B., Black, L. I., Maenner, M. J., Schieve, L. A., Danielson, M. L., Bitsko, R. H., ... & Boyle, C. A. (2019). Prevalence and trends of developmental disabilities among children in the United States: 2009–2017. *Pediatrics*, *144*(4).

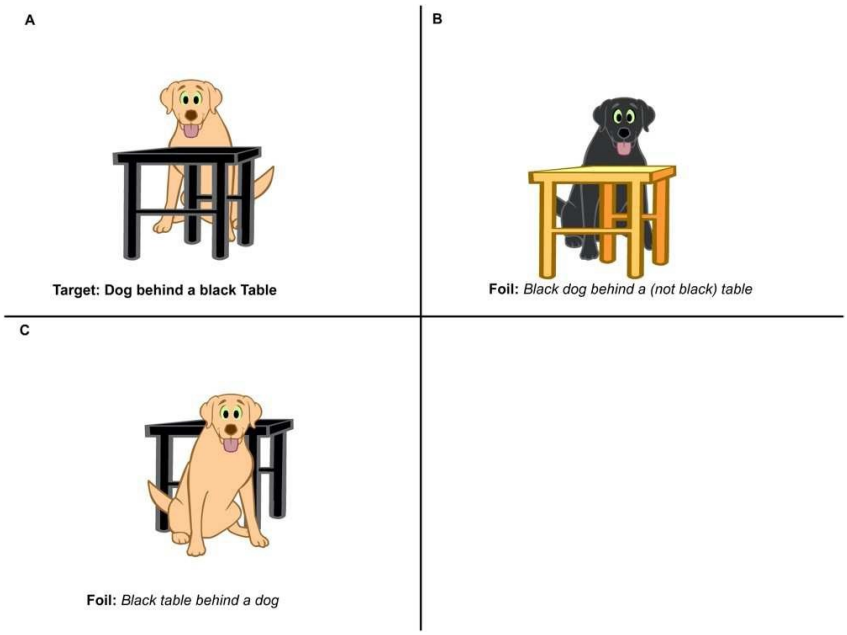

Figure 1. Sample item from the Preposition subtest of the Vocabulary Product component. Children hear, “Find a dog behind a black table.”

Figure 2. Sample item from the Past Tense subtest of the Syntax Product component. Children hear, “Where *was* the hat?” and the correct response is the yellow box framing the boy’s head.

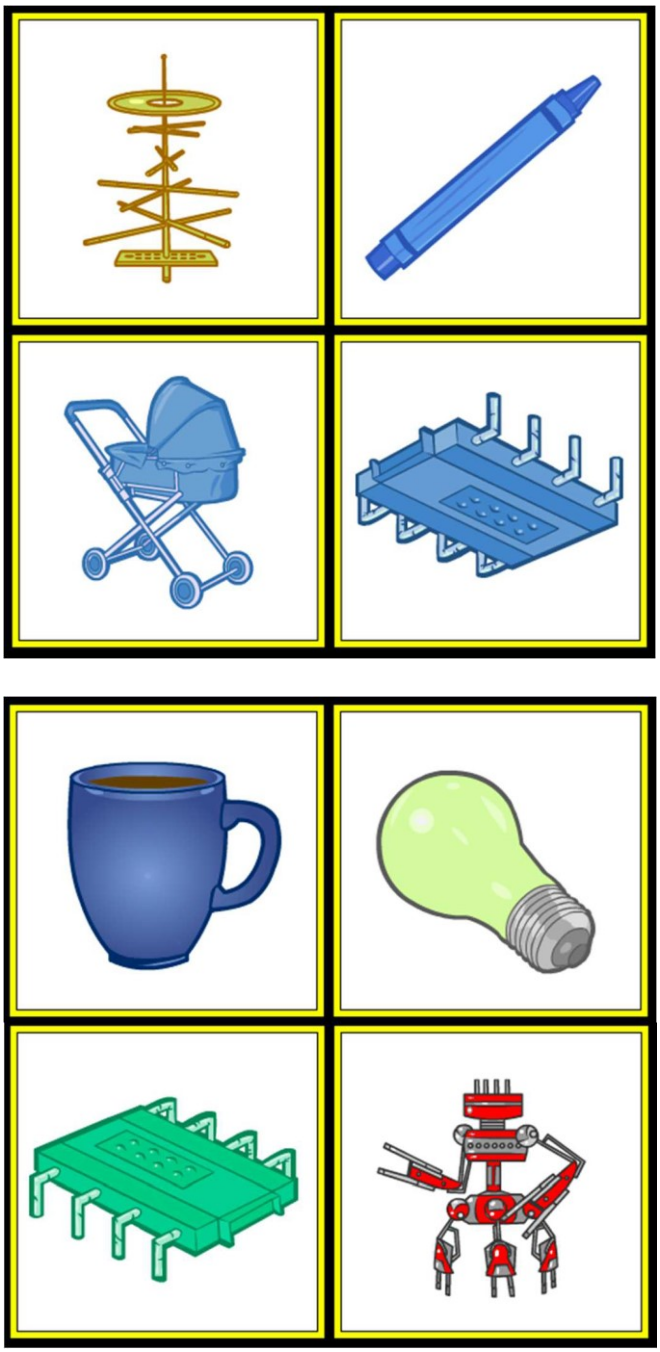

Figure 3. Sample item from the Noun Learning subtest of the Language Learning Process component. Children hear: (a, top image) “The fep is blue. Can you show me the blue fep?” and (b, bottom image) “Can you show me another fep?”.

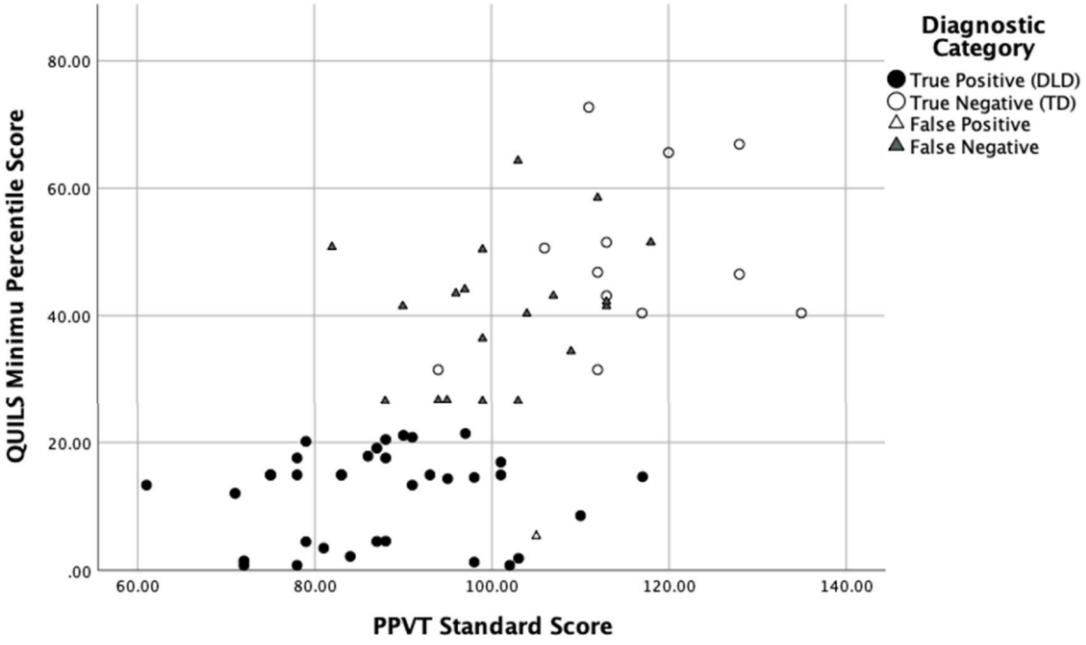

Figure 4. Grouped (TD, DLD) scatterplot of children's QUILS minimum percentile scores by PPVT-4 scores ($N = 67$).

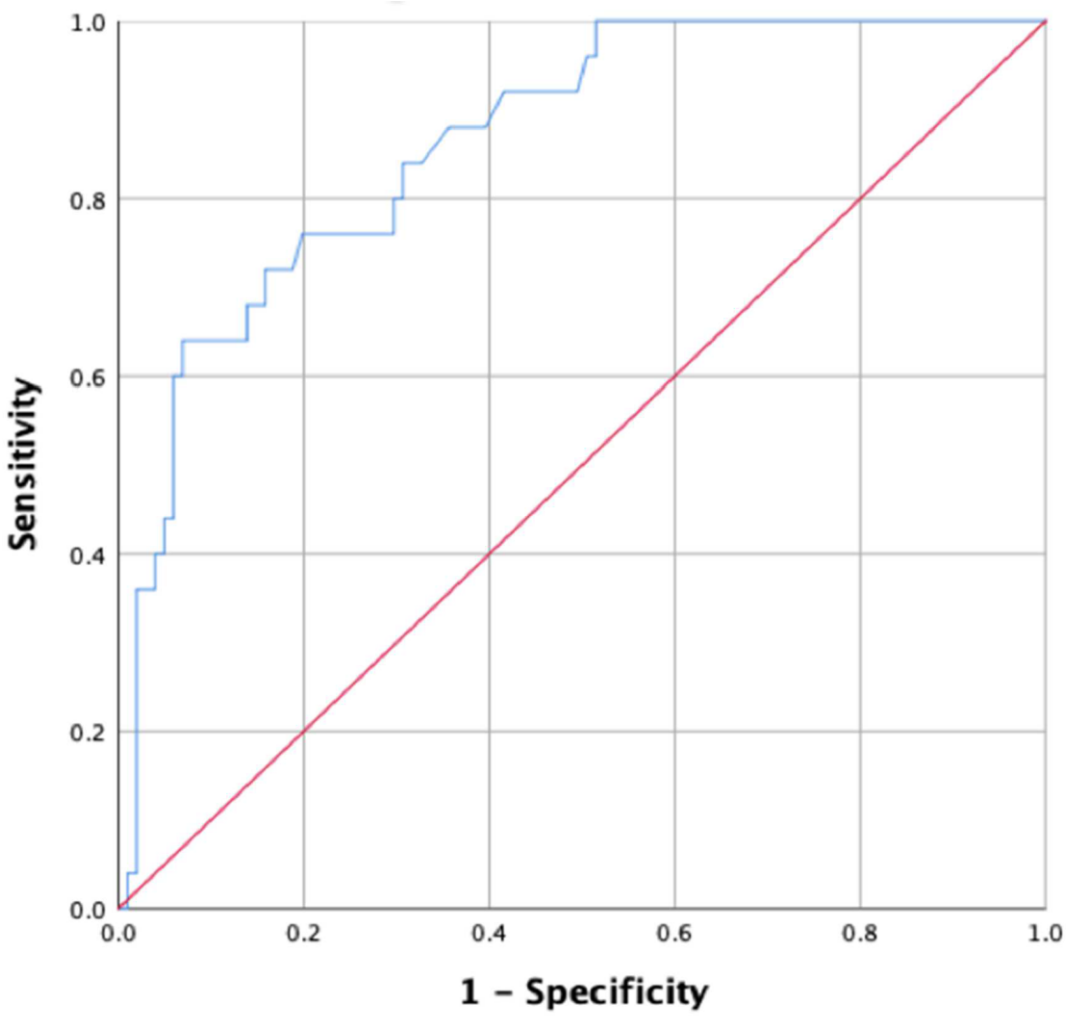

Figure 5. MIN Score ROC Curve for Study 2 (N = 126). AUC = .863.

Table 1

Study 1 participant demographics.

	Total sample (n = 67)	TD (n = 13)	DLD (n = 54)
Characteristic		Frequency (%)	
Gender			
Female	16 (23.8)	6 (46.2)	10 (18.5)
Male	50 (74.6)	6 (46.2)	43 (79.6)
Missing	1 (1.5)	1 (7.7)	1 (1.9)
Race/Ethnicity			
African American/Black	23 (34.3)	1 (7.7)	22 (40.7)
American Indian/Alaskan Native	1 (1.5)	0 (0.0)	1 (1.9)
Asian American	3 (4.5)	0 (0.0)	3 (5.6)
Caucasian/White	36 (53.7)	12 (92.3)	24 (44.4)
Multi-racial	3 (4.5)	0 (0.0)	3 (5.6)
Hispanic			
Yes	3 (4.5)	1 (7.7)	2 (3.7)
No	41 (61.2)	6 (46.2)	35 (64.8)
Missing	23 (34.3)	6 (46.2)	17 (31.5)
SES			
Mid-to high SES	55 (82.1)	12 (92.3)	43 (79.6)
Low-SES	8 (11.9)	1 (7.7)	7 (13.0)
Missing	4 (5.9)	0 (0.0)	4 (7.4)

Table 2

Subtests of the Vocabulary Product, Syntax Product, and Language Learning Process component of the QUILS.

Vocabulary Product	Syntax Product	Language Learning Process
Nouns	WH-questions	Noun learning
Verbs	Past tense	Adjective learning
Prepositions	Prepositional phrases	Verb learning
Conjunctions	Embedded clauses	Converting active to passive

Table 3

Group performance on standardized assessment measures and QUILS Index percentiles (Study 1)

	TD				DLD			
	N	***M***	***SD***	***Range***	***N***	***M***	***SD***	***Range***
SPELT-3	0	-	-	-	15	78.1	9.6	63–93
SPELT-P2	11	100.2	7.6	91–111	10	74.5	7.9	66–85
DELV Syntax subtest	2	10.0	1.4	9–11	29	4.7	1.1	3–6
PPVT-4	13	114.9	10.9	94–135	54	92.2	12.8	61–118
EVT-2	13	117.9	10.8	101–136	52	93.8	10.9	64–120
QUILS Vocab + Syntax	13	56.4	13.8	31.9–74.2	54	31.3	17.2	4.55–74.8
QUILS Process	13	53.3	22.5	5.4–80.5	54	25.7	18.1	0.8–71.0
QUILS Overall	13	65.1	14.0	36.3–81.9	54	31.5	21.2	1.9–78.4
QUILS MIN Score	13	45.6	17.6	5.4–72.7	54	22.0	16.4	0.8–64.3

Note. SPELT-P2 and SPELT-III standard scores; DELV- Syntax Subtest scaled score; DAS t-score; QUILS percentile scores

Table 4

How the QUILS Indices predict to sensitivity and specificity using the 25th percentile cut score (Study 1, N = 67)

INDEX	SENS	SPEC	LR+	LR–
Vocab + Syntax	.46	1.00	Approach infinity	.54
Process	.50	.93	7.14	.53
Overall Composite	.43	1.00	Approach infinity	.57
MIN (lowest of all 3)	.65	.92	8.13	.38

Table 5

Study 2 participant demographics.

	Full Sample (n = 126)	PLS > 85 (TD; *n* = 101)	PLS ≤ 85 (DLD; *n* = 25)
Characteristic		Frequency (%)	
Gender			
Female	63 (50.0)	54 (53.5)	9 (36.0)
Male	63 (50.0)	47 (46.5)	16 (64.0)
Race/Ethnicity			
African American/Black	23 (18.3)	21 (20.8)	2 (8.0)
Caucasian/White	82 (65.1)	64 (63.4)	18 (72.0)
Multi-racial	13 (10.3)	10 (9.9)	3 (12.0)
Missing	8 (6.3)	6 (5.9)	2 (8.0)
Hispanic			
Yes	47 (37.3)	34 (33.7)	13 (52.0)
No	73 (57.9)	62 (61.4)	11 (44.0)
Missing	6 (4.8)	5 (5.0)	1 (4.0)
SES			
Mid-SES	45 (35.7)	38 (37.6)	7 (28.0)
Low-SES	81 (64.2)	63 (62.4)	18 (72.0)

Table 6

Group performance on standardized assessment measures and QUILS Index percentiles (Study 2)

	TD (n = 101)			DLD (n = 25)		
	M	***SD***	***Range***	***M***	***SD***	***Range***
PLS-5 AC SS	103.2	11.9	86–136	78.4	5.9	66–85
QUILS Vocab + Syntax	55.9	25.4	7.8–97.3	29.8	21.9	4.6–78.4
QUILS Process	56.1	26.5	1.3–99.9	21.3	16.9	.8–58.5
QUILS Overall	60.5	26.4	.6–98.4	25.9	23.7	2.2–70.5
QUILS MIN Score	47.1	25.1	.6–95.1	15.3	13.7	.8–43.5

Table 7

Area Under the Curve (AUC) for each QUILS Index with sensitivity and specificity calculations at the 25th percentile cut point (N = 126).

INDEX	AUC	95% CI	SENS	SPEC	LR+	LR–
Vocab + Syntax	.783	.68–.89	.56	.90	5.60	.48
Process	.859	.79–.93	.60	.86	4.28	.46
Composite	.827	.74–.91	.60	.90	6.00	.66
MIN (lowest of all 3)	.863	.79–.94	.76	.77	3.30	.31