AJSLP

Research Article

Comparison of the Diagnostic Evaluation of Language Variation–Screening Test Risk Subtest to Two Other Screeners for Low-Income Prekindergartners Who Speak African American English and Live in the Urban South

Christy Wynn Moland[a] and Janna B. Oetting[a]

Purpose: We compared the Risk subtest of the Diagnostic Evaluation of Language Variation–Screening Test (DELV–Screening Test Risk) with two other screeners when administered to low-income prekindergartners (pre-K) who spoke African American English (AAE) in the urban South.
Method: Participants were 73 children (six with a communication disorder and 67 without) enrolled in Head Start or a publicly funded pre-K in an urban Southern city. All children completed the DELV–Screening Test Risk, the Fluharty Preschool Speech and Language Screening Test–Second Edition (FLUHARTY-2), and the Washington and Craig Language Screener (WCLS). Test order was counterbalanced across participants.
Results: DELV–Screening Test Risk error scores were higher than those reported for its standardization sample, and scores on the other screeners were lower than their respective standardization/testing samples. The 52% fail rate of the DELV–Screening Test Risk did not differ significantly from the 48% rate of the WCLS. Fail rates of the FLUHARTY-2 ranged from 34% to 75%, depending on the quotient considered and whether scoring was modified for dialect. Although items and subtests assumed to measure similar constructs were correlated to each other, the three screeners led to inconsistent pass/fail outcomes for 44% of the children.
Conclusions: Like other screeners, the DELV–Screening Test Risk subtest may lead to high fail rates for low-income pre-K children who speak AAE in the urban South. Inconsistent outcomes across screeners underscore the critical need for more study and development of screeners within the field.

Screenings often serve as the primary means by which children are referred for a speech and language evaluation or provided response to intervention within multitiered educational systems (Hall-Mills, 2019). Unfortunately, almost all speech and language screeners within the field have been designed for mainstream dialects of English and not for others, such as African American English (AAE). The Diagnostic Evaluation of Language Variation–Screening Test (DELV–Screening Test; Seymour et al., 2003b) is an exception. This screener was designed for a variety of dialects including AAE.

The DELV–Screening Test includes a Dialect subtest (DELV–Screening Test Dialect) that can be used to classify a child's dialect as mainstream American English (MAE), presenting some variation from MAE, or presenting strong variation from MAE, and a Risk subtest (DELV–Screening Test Risk) that can be used to classify a child's risk of language impairment (LI) as low, low to medium, medium to high, or high. The purpose of this study was to evaluate the concurrent validity of the DELV–Screening Test Risk subtest for low-income prekindergartners (pre-K) who speak AAE and live in the urban South. To do this, we compared its scores and outcomes with those of two other screeners, the FLUHARTY-2 (Fluharty, 2001a) and the Washington and Craig Language Screener (henceforth referred to as WCLS; Washington & Craig, 2004). The FLUHARTY-2 is a well-established screener with a history of being evaluated for its appropriateness for African American (AA)

[a]Louisiana State University, Baton Rouge
Correspondence to Christy Wynn Moland: cwynnmoland@gmail.com
Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Stacy Betz
Received September 8, 2020
Revision received February 5, 2021
Accepted June 6, 2021
https://doi.org/10.1044/2021_AJSLP-20-00270

Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

children; this screener also provides modified scoring recommendations for children who speak AAE. While experimental, the WCLS was specifically developed for AA children, with consideration of both the cultural appropriateness of the materials and the dialect of AAE. The WCLS is also the only screener of the three that has been examined for its classification accuracy using a battery of clinical reference measures.

Empirical studies of screeners are needed to make sure referral decisions are valid and optimal for low-income AA children. This is especially true in large school districts with historically high screening fail rates such as the one studied here, where children, unless they are referred by a teacher or enter with a clinical diagnosis, must fail two screeners before a speech and language evaluation is scheduled. Moreover, various third parties complete the first screening using their choice of tool; children who fail are screened a second time by school clinicians who dedicate 1 day a month to screening. This screening approach reduces the number of children referred for an evaluation (and hopefully reduces the number of false positives), but it can also lead to delays in children's receipt of services.

Disproportionate representation of AA children in special education has been a concern of the U.S. Department of Education for over 4 decades, and many studies have found overrepresentation of AA children in various special education disability categories (e.g., Chinn & Hughes, 1987; Horner et al., 1986), with at least one study also showing higher percentages of AA children receiving stigmatized disability labels, multiple disability labels, and segregated educational placements compared with White children (De Valenzuela et al., 2006). Results have been mixed for the category of speech and LI. Some studies have found overrepresentation of AA children (e.g., Skiba et al., 2016; Sullivan & Bal, 2013), and others have found underrepresentation (e.g., De Valenzuela et al., 2006; Morgan et al., 2015, 2017; Skiba et al., 2006). Robinson and Norton (2019) calculated risk ratios for AA and non-AA children with a speech and LI disability label in the schools. Their data spanned 10 years, beginning in 2004. Results averaged across years varied by state, with 62% of states showing underrepresentation of AA children, 14% showing overrepresentation, and 22% showing proportionate representation. Moreover, states with the highest densities of AA children enrolled in schools tended to show underrepresentation of AA children, and states with the lowest densities of AA children tended to show overrepresentation.

Beyond national and state school statistics, studies of speech and language screenings of low-income AA children have resulted in extremely high fail rates and inconsistent outcomes. In a study of 173 AA kindergarteners, Rhyner et al. (1999) reported a 51% fail rate for the screening version of the Bankson Language Test (Bankson, 1990) and 58% for the Structured Photographic Expressive Language Test–Preschool (SPELT-P; Werner & Kresheck, 1983), which was also viewed as a screener by the authors. Of those who failed one screener, only 49% failed both. At the same time, low-income children experience disparities in access to services by speech-language pathologists (Bishop & McDonald, 2009; Keegstra et al., 2007; Morgan et al., 2016). Wittke and Spaulding (2018) documented this disparity with fifty-four 3- to 5-year-olds classified as LI; 23 were receiving services and 31 were not. Whereas the two LI groups did not differ in gender representation or language test scores, differences were found for their maternal education levels ($d = 0.797$), which served as a proxy for the children's socioeconomic status. Mothers of children receiving treatment reported more levels of education than mothers of children who were not receiving treatment.

Together, these findings underscore the complexity, potential subjectivity, and context-specific nature of disparate practices across the United States. One way speech-language pathologists can help reduce disparities is to make sure we select the very best screeners when making referrals. To do this, screeners need to be evaluated not only at the national level but also for specific groups, such as low-income AA children who are at risk for being over- and underidentified for services (for others who recognize the need to explore how tools behave for well-specified groups, see Kilgus et al., 2014).

DELV–Screening Test Risk Subtest

The DELV–Screening Test Risk subtest includes 11 items focused on children's production of grammar and six on their repetition of nonwords. Five items target the copular or auxiliary BE past form *was.* In various dialects of English, including many (but not all) dialects of AAE, the overt form *was* is produced over 90% of the time within singular BE past tense contexts (Oetting et al., 2019; Roy et al., 2013; for a counter AAE dialect example, see Berry & Oetting, 2017). The invariant nature of *was* in this context makes it ideal for a screener because a response with a zero form (e.g., *It Ø windy*) or a response that does not answer a question can be interpreted as reflecting a dialect weakness rather than a dialect difference.

The DELV–Screening Test Risk subtest also targets children's productions of regular past tense forms with the verb *play* and the possessive pronouns *his* and *theirs.* Again, these structures were selected because in many dialects of English, including many (but not all) dialects of AAE, children who are typically developing (TD) produce similar types of overt forms. For example, in a study of 30 AAE-speaking TD children who varied by grade (pre-K vs. kindergartner [K]) and socioeconomic status (low vs. middle), overt forms for past tense were produced 89%–98% of the time when the verb ended in a vowel, such as *play* (Pruitt & Oetting, 2009; for a review of other studies, see Lee & Oetting, 2014). Similarly, Brown (2017) examined the pronoun productions of 96 AAE-speaking kindergartners, and for those classified as TD, genitive case marking (e.g., *his, theirs, hers, my*) was produced 98% of the time when the pronoun expressed possession.

Studies have also shown that AAE-speaking children with LI produce overt forms of *was*, regular past tense, and genitive case marking on possessive pronouns less often

than their same dialect-speaking TD peers (Brown, 2017; Hendricks & Adlof, 2020; Oetting et al., 2019; for past tense only, see Seymour et al., 1998). These studies provide independent evidence to support the inclusion of these grammar structures within the DELV–Screening Test Risk subtest.

A second appealing feature of the DELV–Screening Test Risk subtest is its use of questions to elicit descriptions from children. Peña and Quinn (1997) showed that descriptive tasks are less biased against children from minority backgrounds than tasks that elicit labels. Their participants were 127 children classified as either AA or Puerto Rican, and the tasks came from commercially available, norm-referenced tests. Both groups earned higher standardized scores on the descriptive task than on the labeling task and those classified as LI scored lower than those classified as TD on the descriptive task.

Finally, the DELV–Screening Test Risk subtest includes six nonwords. Nonword repetition tasks are also considered less biased against children from minority backgrounds than other standardized assessments, and they have been found to differentiate children with and without LI in various mainstream dialects of English (Campbell et al., 1997; Dollaghan & Campbell, 1998; Graf Estes et al., 2007). Four studies have examined the nonword repetition abilities of AAE-speaking children. Although in two of these studies, the children's nonword repetitions were found to vary by the density of their nonmainstream English forms (McDonald & Oetting, 2019; Moyle et al., 2014), in all studies that included a clinical group, those with LI earned lower scores than those without LI (Oetting et al., 2008; Rodekohr & Haynes, 2001), even when the children's nonmainstream form densities served as a covariate (McDonald & Oetting, 2019). It is also worth noting that the nonwords included in the studies that found nonmainstream form density effects did not contain consonant clusters and were created with early developing phonemes in multiple dialects of English, including AAE (Dollaghan & Campbell,1998).

The standardization sample of the DELV–Screening Test is also appropriate for AA children who speak AAE and live in the South because it included 1,258 children, with 65% classified as AA, 63% classified as speakers of AAE, and 58% classified as living in the South. Although the children's socioeconomic status and school enrollment were not documented, the standardization sample included a higher percentage of children whose primary caregivers completed less than high school (21%) and no more than high school (58%), as compared with the U.S. population, which was listed as 18% and 37%, respectively (Seymour et al., 2003c, pp. 41–42).

Concurrent validity was examined by comparing the DELV–Screening Test Risk error scores with the domain scores from the Diagnostic Evaluation of Language Variation–Criterion Referenced (DELV–Criterion Referenced; Seymour et al., 2003a, 2003c). Correlations were highest when the errors scores were compared with the syntax domain of the DELV–Criterion Referenced ($r = -.70$) and lowest when compared with the DELV–Criterion Referenced phonological domain ($r = -.23$). Also, correlations between DELV–Screening Test Risk error scores for items targeting grammar and the DELV–Criterion Referenced were higher than those targeting nonword repetition (r ranged from $-.47$ to $-.66$ vs. $r = -.29$). Finally, classification accuracy of the screener was evaluated using a priori clinical classifications of 708 children, 217 of which were 4 years of age (i.e., 169 classified as nonclinical and 48 classified as clinical, with diagnoses of LI and in some cases also with attention-deficit/hyperactivity disorder or developmental delay). The error scores of the 4-year-old nonclinical and clinical groups averaged 8.85 ($SD = 4.50$) and 14.42 ($SD = 4.73$), respectively. Using this age group and the highest risk category to determine a failed screening, sensitivity and specificity values were .70 and .76, respectively. While classification accuracy values of .90 and higher are recommended for diagnostic tests, values of .70 or .80 are often accepted for screeners, especially for indices of specificity, to ensure that children who may need services are not missed (Kilgus et al., 2014; Youngstrom, 2014).

Ciolli and Seymour (2004) examined the interexaminer reliability of the DELV–Screening Test Risk subtest by asking two examiners to administer the screener to 23 children. Eighty-four percent of the children received either the same risk classification by both examiners or classifications differing by no more than one risk category. Finally, Petscher et al. (2012) and Terry et al. (2017) examined the factor structure and measurement invariance of the DELV–Screening Test Risk subtest using data from children in pre-K to second grade. These authors found support for the screener, although they recommended a two-factor structure to evaluate the grammar and nonword repetition items separately, the use of norm-referenced scores, and the development of different norms for children in pre-K as compared with K–2.

Two Other Speech and Language Screeners

Two other screeners were selected to examine the concurrent validity of the DELV–Screening Test Risk subtest. As shown in Table 1, these screeners differ from the DELV–Screening Test Risk subtest in the type of content assessed and the number of items administered. Their standardization/test samples and the types of evidence collected to support their use also differ from those of the DELV–Screening Test Risk subtest.

FLUHARTY-2

The FLUHARTY was published in 1978 and revised in 2001. This screener currently has 53 items and five subtests, with one focused on articulation and four focused on language (i.e., Repeating Sentences, Following Directives and Answering Questions, Describing Actions, Sequencing Events). Sturner et al. (1993) examined the initial version of the FLUHARTY using data from 700 children, aged 4–5 years (75% White, 25% AA, and approximately 1% other/not reported). The children were divided into two cohorts, which were described by the authors as varying in the number of children from low socioeconomic backgrounds. Fail rates for the two cohorts were 12% and 24%, with the higher

Table 1. Comparison of items on the three screeners.

DELV–Screening Test Risk	FLUHARTY-2	WCLS
1 subtest	5 subtests	3 subtests
Grammatical structures: 11 items	Articulation: 15 items, with 30 sounds targeted	PPVT: 192 items, with testing discontinued when child reaches a ceiling
Nonword repetition: 6 items	Repeating Sentences: 10 items	*Wh*-Question Task: 24 items
	Following Directives: 15 items	CMLU-words: Calculated from the child's three longest utterances produced during picture descriptions
	Describing Actions: 10 items	
	Sequencing Events: 4 items	

Note. DELV–Screening Test Risk = Diagnostic Evaluation of Language Variation–Screening Test Risk subtest; FLUHARTY-2 = Fluharty Preschool Speech and Language Screening Test–Second Edition; WCLS = Washington and Craig Language Screener; PPVT =Peabody Picture Vocabulary Test; CMLU = mean length of communication units in morphemes.

fail rate obtained for the cohort with more children from low socioeconomic backgrounds.

Revision of the FLUHARTY included new pictures and objects, three new subtests, and a new normative sample of 705 children. Although the socioeconomic status of the normative sample was not described, 15% were AA; 37% lived in the South; and 18% presented either a learning disability, speech impairment, mental retardation, or other educational disability (Fluharty, 2001b, p. 26). As part of the revision, the developers also examined the children's scores by their race/ethnicity. Using differential item functioning analyses and delta subgroup comparison scores, correlations between those classified as AA and not AA were found to be moderate (r = .76) for the Articulation subtest and high (r ranged from .92 to .97) for the language subtests. Also, the AA group earned standard scores that were within −1 *SD* of the normative means, and three forms of reliability (i.e., content, time, and scorer) were found to be adequate, with correlations between scores at or above .95. Finally, the manual provides subtest scores for 77 children with speech disorders and 12 with learning disabilities. Although these a priori clinical groups were not used to examine the screener's classification accuracy, the average score of the former group was below −1 *SD* of the normative mean for the Articulation subtest and the average score for the latter group was below −1 *SD* for the Following Directives and Answering Questions subtest.

As mentioned earlier, the FLUHARTY-2 includes guidelines for modifying the scoring of a child's responses to allow for nonmainstream forms of English. For example, on the Describing Actions subtest, examiners are encouraged to count as correct "deletion of auxiliary BE verbs that are common in AAE." Unfortunately, the manual does not specify all child responses that should be classified as representing dialectal variants. Instead, this scoring decision is left to the examiner's discretion. It is also not clear if modified scoring was used to create the normative data.

Using a sample of sixty-two 3-year-olds, Eisenberg et al. (2019) compared outcomes of the FLUHARTY-2 with those from the Structured Photographic Expressive Language Test (SPELT; Dawson et al., 2005) and measures from language samples (i.e., mean length of utterances in morphemes [MLU-m], percentage of overtly marked finite verbs [FVMC], and the Index of Productive Syntax [IPSyn]; Scarborough, 1990). The FLUHARTY-2 was scored without modifications made for nonmainstream English forms. Although there were relatively high levels of agreement between the children who passed the FLUHARTY-2 and three of the measures (i.e., 100% agreement for the SPELT, 97% for MLU-m, and 81% for IPSyn), very low levels of agreement were found for the children who failed the FLUHARTY-2 and these measures (i.e., 7% agreement for the SPELT, 7% for MLU-m, 68% for FVMC, and 68% IPSyn). Moreover, fail rates of the FLUHARTY-2 varied by the race and ethnicity of the children, with 43% of the children classified as White failing the screener, compared with 71% classified as AA, 67% classified as Asian, and 57% classified as Hispanic.

WCLS

The WCLS was specifically designed for AA children living in an urban setting (Washington & Craig, 2004). It includes the Peabody Picture Vocabulary Test–III (PPVT-III; Dunn & Dunn, 1997), a *Wh*-Question Task, and calculation of a child's average utterance length. The PPVT-III is a standardized test of single-word receptive vocabulary knowledge. The *Wh*-Question Task includes 24 questions that relate to two pictures from the Bankson Language Test (Bankson, 1990), and the items range in difficulty from asking for a label (e.g., "What is this?") to asking for a description or interpretation (e.g., "When is this happening?"). The children's average utterance length is calculated from their longest three utterances as measured by C-units in words (MCLUw) and as elicited through picture descriptions.

Washington and Craig (2004) evaluated the WCLS with a community sample of 196 AA children enrolled in pre-K or K in Detroit, MI. Using data from 41% of the participants, 59% of the sample was classified as low income and 39% was classified as middle income. In the first phase of the study, the overall fail rate of the WCLS was 18% (pre-K = 23% and K = 7%). In the second phase, 81 children (25 who failed and 56 who passed) completed a comprehensive language assessment battery that included a longer language sample, a nonword repetition task, and a sentence comprehension task. The measures within this battery are noteworthy because they were drawn from studies of childhood LI, and they were not found to vary by

the gender or socioeconomic status of the children studied. Using clinical classifications based on the comprehensive battery, the sensitivity and specificity values of the WCLS screener were .60 and .93, respectively.

Summary and Research Questions

In summary, the DELV–Screening Test Risk subtest was carefully designed for children who speak various dialects of English, including AAE, and there is independent evidence from studies to support its content for the dialect of AAE, as well as some information about its validity and reliability. The standardization sample of the DELV–Screening Test Risk subtest also appears appropriate for the current sample because it oversampled low-income AA children who lived in the south. The FLUHARTY-2 and the WCLS are two other tools that have been recommended for screening AA children, and both have some information about their reliability and validity. Compared with the DELV–Screening Test Risk subtest, the standardization sample of the FLUHARTY-2 includes less representation of low-income AA children who live in the South and a significant percentage of children with various clinical/educational exceptionalities. Although the WCLS is not a published screener, it was tested with a community sample of low- and middle-income AA children who lived in an urban setting, and both its overall fail rate of 18% and specificity index of .93 are impressive relative to the other tools reviewed. Finally, compared with the DELV–Screening Test Risk subtest, the FLUHARTY-2 and the WCLS include more items and subtests, which might mean that they provide a more comprehensive screen of children. On the other hand, more items do not necessarily ensure higher levels of accuracy.

Using data from low-income AA children who speak AAE in the urban South, this study was designed to compare the DELV–Screening Test Risk subtest with the FLUHARTY-2 and the WCLS. The questions driving the study were as follows: (a) What scores do the children earn on the DELV–Screening Test Risk, the FLUHARTY-2, and the WCLS relative to the screeners' standardization/testing samples? (b) What percentage of children fail the DELV–Screening Test Risk, the FLUHARTY-2, and the WCLS, and do those who fail include children previously diagnosed with a communication disorder? (c) How well do the pass/fail outcomes of the DELV–Screening Test Risk agree with those of the FLUHARTY-2 and the WCLS? (d) How well do the various items on the DELV–Screening Test Risk and subtests from the other screeners correlate to each other?

We expected the children's scores and fail rates to be most like those previously reported for the DELV–Screening Test Risk subtest and the WCLS because these two tools were specifically designed for AAE-speaking children. Their standardization/testing samples were also more like the current study sample than those of the FLUHARTY-2. Differences in items and subtests across the screeners led us to expect differences in their fail rates, yet we also expected the screeners to rank children by ability in relatively the same order and be correlated to each other, especially if we selected items and subtests assumed to measure similar constructs. If true, then we hoped to optimize the use of one or more of these screeners by recommending a greater focus on them or more study of the items and subtests within them.

Method

Participants

The participants were 73 AA (38 male and 35 female) children whose mean age was 54.59 months (SD = 4.86, range: 48–64; 58 were 4 years of age and 15 were 5 years of age). They lived in an urban Southern city with 55% of the residents reporting their race as AA; other city demographics at the time of data collection included a population of 225,362, a poverty rate of 25.2%, and an unemployment rate of 11.8%, which was higher than the national average of approximately 15% and approximately 10%, respectively (U.S. Census Bureau, 2010). Based on the Robinson and Norton (2019) results, the children lived in a state that underrepresents AA children in the disability category of speech and LI.

The children attended either a Head Start (n = 45) or a publicly funded pre-K (n = 28) with 98%–100% enrollment of AA children. In the two schools for which consent forms were sent to all children, the rate of return with caregiver consent was 35% and 38%. The children's families were classified as low income based on school enrollment, with 100% of the children in Head Start and 95% in pre-K receiving free lunch and an additional 4% in pre-K receiving lunch at a reduced rate. Per report from 61 primary caregivers (57 women and four men), the children's mean level of caregiver education was 12.63 (SD = 1.77; range: 9–16, with 12 = high school), with 10 caregivers not completing high school, 32 completing high school, and 19 completing more than high school. On the DELV–Screening Test Dialect subtest, all children produced at least one nonmainstream English response, with two (3%) children classified as producing a dialect consistent with MAE, five (7%) classified as producing a dialect with some variation from MAE, and 66 (90%) classified as presenting a dialect with strong variation from MAE. We also calculated the children's percentage of nonmainstream responses out of their scored responses on this subtest. The children's nonmainstream forms averaged 95% (SD = 10%, range: 50%–100%), which is higher than typically reported in AAE studies that have used this metric (cf. Berry & Oetting, 2017; Horton-Ikard & Apel, 2014; Maher et al., 2021; Terry et al., 2010).

At the time of data collection, all children either passed a hearing screening at 25 dB at 1000, 2000, and 4000 Hz (n = 66) or had a recent hearing screening documented in their school records (n = 7). Although we were not granted access to the audiological records of these seven children, their performance on the screeners paralleled those of the other children, and the results did not change significantly

when they were excluded. Per school records, 8% (n = 6) received services by a school speech-language pathologist (two for speech, one for fluency, one for language, two for speech and language) and 92% (n = 67) were developing speech and language typically.

Materials

DELV–Screening Test Risk Subtest

The children's responses were scored according to the manual (Seymour et al., 2003c). Based on the children's error scores, the children's risk for LI were classified as low, low to medium, medium to high, or high. For children 4 and 5 years of age, the cutoff for highest risk is an error score of 13 and 9, respectively; these scores were used to determine a failed screening.

FLUHARTY-2

Given that the children were speakers of AAE, the FLUHARTY-2 was scored twice, once without scoring modifications for the children's use of nonmainstream English forms and once with modifications. According to the manual, individual raw scores were summed and converted to four quotients (Articulation, Receptive, Expressive, and General Language), which have a normative mean of 100 (SD = 15). A failed screening was determined when a child earned a standard score of ≤ 89. Although a composite score of 89 is higher than −1 SD of the normative mean, according to the manual, additional testing is warranted for children who earn a score at or below this cutoff (Fluharty, 2001b, p. 23). The manual also recommends additional testing if any quotient falls ≤ 89. However, the DELV–Screening Test Risk and the WCLS do not screen for articulation. Given this, we examined the FLUHARTY-2 as recommended in the manual using all quotient scores and then with only the General Language Quotient. Results from the General Language Quotient were also used whenever the screeners were compared with each other. This quotient does not include the Articulation subtest, and it generally shows higher reliability coefficients than the Receptive and Expressive Language quotients used in isolation (Fluharty, 2001b, pp. 30–31).

WCLS

The original WCLS included the PPVT-III, a *Wh*-Question Task, and calculation of the child's MCLUw. Although the PPVT-4 (Dunn & Dunn, 2007) was used in this study, all other aspects of administration and scoring were based on Washington and Craig (2004). The normative mean of the PPVT-4 was 100, with an SD deviation of 15. Failure on the PPVT-4 was a standard score of ≤ 85.

The *Wh*-Question Task included 24 questions, and each question could receive up to 3 points for a maximum of 72 points. A score of 3 reflected the target answer, a score of 2 reflected a nonspecific answer or a misnamed referent within the answer, a score of 1 reflected a response that answered a different question, and a score of 0 reflected an answer unrelated to the question. Failure on this task was a raw score of ≤ 49.

For MCLUw, each child was shown a picture of a mother and two children at a grocery store (Arwood, 1985). The child's responses were transcribed into C-units, and then the number of words within the child's three longest C-units were averaged. A C-unit was defined as an independent clause plus modifiers, individual responses to adult questions, and acknowledgment by the child of a previous adult comment. Failure on this measure was an MCLUw of ≤ 5.00.

Procedure

This study was approved by the Louisiana State University Institutional Review Board prior to data collection. Caregiver consent and child assent were also obtained, and each child participated in two 30-min sessions. The first author, who was an AA female, native resident of the community, and licensed and certified speech-language pathologist with 7 years of clinical experience, administered all three screeners to the children. To control for practice effects, the screeners were administered to the children in one of six orders to counterbalance their orders and the order of the screeners that came before and after each one.

Reliability

Twenty percent (n = 15) of the children's test data were randomly selected to assess reliability of the data coding and scoring. These tests were independently scored by student clinicians trained by the first author. Agreement between the two sets of scores for each screener ranged from 91% to 100%.

Results

DELV–Screening Test Risk

The children's DELV–Screening Test Risk error scores averaged 11.27 (SD = 4.78, range: 2–21). This error score is higher than the 8.85 (SD = 4.50) score earned by the nonclinical group of 4-year-olds who contributed to the standardization of the tool. Using these scores, nine (12%) children were classified as lowest risk for impairment, 11 (15%) were classified as low to medium risk, 15 (21%) were classified as medium to high risk, and 38 (52%) were classified as high risk, which was the category that corresponded to a failed screen. Table 2 provides the children's average error scores and fail rates for the DELV–Screening Test Risk subtest by diagnosis. The two children with a speech disorder passed the DELV–Screening Test Risk subtest, and the other four with a communication disorder (i.e., one fluency, one language, and two speech and language) failed. Using these a priori clinical classifications of the children (i.e., six with communication disorders and 67 without), the DELV–Screening Test Risk yielded manually calculated sensitivity and specificity values of .67 and .49, respectively. If the two children with a speech

Table 2. Diagnostic Evaluation of Language Variation–Screening Test (DELV–Screening Test) Risk error scores and fail rate by diagnosis.

Diagnosis	DELV–Screening Test Risk	Fail rate
No diagnosis	11.16 (4.66) 2.00–19.00	34/67 = 51%
Language and articulation	16.33 (4.04) 14.00–21.00	3/3 = 100%
Articulation only	5.00 (1.41) 4.00–6.00	0/2 = 0%
Fluency only	16.00 (0.00)	1/1 = 100%

Note. Means reported first, followed by standard deviations in parentheses and ranges when available.

disorder were considered typical for language (and moved to the nonclinical group), then sensitivity and specificity values increased to 1.0 and .51, respectively.

FLUHARTY-2

Recall that this screener was scored twice, once with and once without modification of the children's nonmainstream English forms. Two types of forms, zero auxiliary *is* (e.g., *The boy Ø drinking*) and dialect-specific subject pronouns (e.g., *Him Ø drinking*), were produced by the children during administration of the FLUHARTY-2. No other word or morpheme productions were considered nonmainstream using a list of previous forms produced by southern AAE-speaking children within language samples (Oetting & McDonald, 2001). The number and percentage of children who produced at least one of these response types were as follows: zero auxiliary *is* only: $n = 31$, 42%; dialect-specific subject pronoun only: $n = 3$, 4%; and zero auxiliary *is* and dialect-specific pronoun: $n = 16$, 22%.

Table 3 provides the means of the children's quotients with and without modified scoring. Modified scoring was only applied to the Expressive Language Quotient and General Language Quotient. Regardless of scoring approach, the children's average quotients were lower than the normative mean of 100 ($SD = 15$). Quotients without scoring modifications were also significantly lower than those with modifications: Expressive Language Quotient, $F(1, 72) = 82.52$, $p < .001$, $\eta^2 = .534$, and General Language Quotient, $F(1, 72) = 75.45$, $p < .001$, $\eta^2 = .512$. Without scoring modifications, fail rates of the individual quotients ranged from 18% to 66%; with modifications, they ranged from 18% to 34%. Using the manual's recommended criterion of failing one or more quotients, the fail rate was 75% without scoring modifications and 56% with modifications. Using the General Language Quotient to determine a screening failure, the fail rate was 56% without scoring modifications and 34% with them. In other words, the scores and fail rates of the FLUHARTY-2 varied considerably, depending upon the number and type of quotients considered and how the children's responses were scored.

Both children with a speech disorder failed the Articulation Quotient, and one of these children also failed the General Language Quotient, even when modified scoring was applied. With modified scoring, the child with a fluency disorder passed the Articulation Quotient but failed the General Language Quotient. The child with a language disorder passed the Articulation Quotient but failed the General Language Quotient as did one of the children with a speech and language disorder. The other child with a speech and language disorder failed the Articulation Quotient and passed the Language Quotient. Using the children's a priori clinical classifications (i.e., six with communication disorders and 67 without), modified scoring, and the General Language Quotient to determine a screening failure, sensitivity and specificity values were .67 (4/6) and .69 (46/67), respectively. If the two children with speech disorders were considered typical for language (and moved to the nonclinical group), then sensitivity and specificity values increased to .75 (3/4) and .68 (47/69).

WCLS

Table 4 lists the children's scores for each subtest of the WCLS, along with those reported by Washington and Craig (2004). As shown, there was a relatively wide range of scores for each subtest; however, the children's mean scores for the PPVT-4 and the *Wh*-Question Task were lower than those reported by Washington and Craig for both the pre-K and K groups, and the children's mean score on the latter was also lower than the 49 cutoff used to indicate failure on this measure. In contrast, the children's mean MCLUw was consistent with those reported by Washington and Craig.

Using Washington and Craig's (2004) criterion of failing two or more subtests, 35 (48%) children failed the WCLS; by subtest, 37 (51%) failed the PPVT-4, 46 (63%) failed the *Wh*-Question Task, and 16 (22%) failed the MCLUw. These fail rates are higher than Washington and Craig's overall fail rate of 18% and their pre-K and K fail rates of 23% and 7%, respectively. However, like Washington and Craig's participants, when children failed the WCLS, they were most likely to fail the PPVT-4 and the *Wh*-Question Task.

Four of the six children with a communication disorder failed the WCLS; the two who passed included one child with a speech disorder and one child with a language disorder. Using the a priori clinical classifications of the children (i.e., six with communication disorders and 67 without), the WCLS yielded sensitivity and specificity values of .67 and .54, respectively. If the two children with a speech disorder were considered typical for language (and moved to the nonclinical group), then sensitivity increased to .75 and specificity remained the same at .54.

Performance Across Screeners

Across screeners, 52% of the children failed the DELV–Screening Test Risk, 34%–75% failed the FLUHARTY-2

Table 3. Fluharty Preschool Speech and Language Screening Test–Second Edition scores, number of children who failed, and fail rate by type of scoring.

	Without modified scoring			With modified scoring		
Quotient	**Standard scores**	**Number who failed**	**Fail rate**	**Standard scores**	**Number who failed**	**Fail rate**
Articulation	95.63 (8.86) 70.00–110.00	13	18%	—	—	—
Receptive Language	91.86 (10.45) 70.00–112.00	29	40%	—	—	—
Expressive Language	87.47 (10.24) 70.00–112.00	48	66%	95.27 (9.44) 73.00–112.00	20	27%
General Language Quotient	88.56 (10.18) 68.00–110.00	41	56%	92.86 (9.74) 72.00–110.00	25	34%
	Fail rate based on one or more quotients = 75% Fail rate based on General Language Quotient = 56%			Fail rate based on one or more quotients = 56% Fail rate based on General Language Quotient = 34%		

Note. Means reported first, followed by standard deviations in parentheses and ranges when available. Fail rate is the percentage of children who failed. An em dash (—) indicates that modified scoring was not applied to these subtests.

depending on the number and type of quotients considered and whether the children's scores were modified, and 48% failed the WCLS. The DELV–Screening Test Risk fail rate was not significantly different from that of WCLS when tested with a McNemar test, $p = .68$. The 56% fail rate of the FLUHARTY-2 using the General Language Quotient and without modified scoring also did not differ from the fail rates of either the DELV–Screening Test Risk, $p = .65$, or WCLS, $p = .21$, but with modified scoring, the 34% fail rate of the FLUHARTY-2 was lower: FLUHARTY-2 modified versus DELV–Screening Test Risk, $p = .007$, and Fluharty modified versus WCLS, $p = .04$.

Table 5 lists the number of children who passed all three screeners or failed one or more of them. Again, the FLUHARTY-2 data were based on the modified scoring results and the General Language Quotient. As shown, 22 (30%) children passed all three screeners, 19 (26%) failed all three screeners, 23 (32%) failed one, and nine (12%) failed two. In other words, the screeners led to identical clinical outcomes for the 41 (56%) children who either passed or failed all three screeners, but they led to inconsistent outcomes for the remaining 32 (44%).

Correlations

Table 6 lists correlations between the various items and subtests on the three screeners. Recall that we hoped that even if the screeners did not yield consistent fail rates, they would rank children in relatively similar orders and be correlated to each other if we examined items and subtests assumed to measure similar constructs. When examining the DELV–Screening Test Risk subtest, items were

Table 4. Washington and Craig Language Screener scores and fail rates.

		Washington and Craig's participants	
Current participants *N* = 73		**Pre-K *N* = 140**	**K *N* = 56**
PPVT-4	86.00 (10.67) 65.00–113.00	93.27 (11.17) —	96.22 (11.09) —
Wh-Question Task	46.33 (9.44) 21.00–65.00	49.36 (9.25) —	55.88 (7.52) —
MCLUw	7.33 (2.49) 4.00–15.67	6.61 (2.01) —	7.42 (2.14) —
Fail rate	48%	23%	7%

Note. Means reported first, followed by standard deviations in parentheses and ranges when available. Fail rate is the percentage of children who failed. Pre-K = prekindergartners; K = kindergartners; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; MCLUw = utterances as measured by C-units in words.

Table 5. Number of participants, percentages, and outcomes across screeners (N = 73).

Screeners	No. of participants	Percentages	Outcomes
Passed all three screeners	22	30.1%	Agreement = 56%
Failed all three screeners	19	26.0%	
Failed DELV–Screening Test Risk only	11	15.1%	Disagreement = 32%
Failed FLUHARTY-2 only	3	4.1%	
Failed WCLS only	9	12.3%	
Failed DELV–Screening Test Risk and FLUHARTY-2	2	2.7%	Disagreement = 12%
Failed DELV–Screening Test Risk and WCLS	6	8.2%	
Failed FLUHARTY-2 and WCLS	1	1.4%	

Note. Outcomes reported with rounding. DELV–Screening Test Risk = Diagnostic Evaluation of Language Variation–Screening Test Risk subtest; FLUHARTY-2 = Fluharty Preschool Speech and Language Screening Test–Second Edition; WCLS = Washington and Craig Language Screener.

separated into two categories, those focused on grammar and those focused on nonword repetition. For the FLUHARTY-2, subtest scores rather than quotients were examined. Given the number of correlations examined and concern over making a Type I error, we focused on those that were significant at the .001 level.

As expected, correlations were highest for the DELV–Screening Test Risk grammar items and three of the language subtests of the FLUHARTY-2, the PPVT-V, and the *Wh*-Question Task. The same three FLUHARTY-2 language subtests also yielded higher correlations among themselves and the PPVT-4 and the *Wh*-Question Task than they did with the FLUHARTY-2 Articulation subtest. The Describing Actions subtest of the FLUHARTY-2, which received modified scoring for the children's nonmainstream English forms, yielded low correlations with the other language subtests of the FLUHARTY-2 and with all other measures collected. The PPVT-4 and the *Wh*-Question Task also correlated with each other, whereas MCLUw was not correlated with these subtests or any of the other subtests or items collected. Based on the highest correlations (i.e., $r \geq .60$), the items and subtests most closely related to each other were the grammar items from the DELV–Screening Test Risk, the FLUHARTY-2 Repeating Sentences and Following Directions subtests, and the WCLS *Wh*-Question Task.

A Post Hoc Analysis of Practice Effects

Twenty-three children failed one screener, and nine failed two. Although the screeners were counterbalanced for their order, given the high fail rates of the DELV–Screening Test Risk and the WCLS and the inconsistent outcomes across screeners, we wondered whether these children benefited from screening practice. The 32 children failed 41 screeners. Sixteen of their fails occurred with the first screener administered, 17 occurred with the second, and eight occurred with the third. The number of children who failed the second screener was not significantly lower than the number who failed the first, but the number who failed the third was significantly lower than the number who failed the second, $\chi^2(1, 32) = 7.07$, $p = .008$. These results indicate that test practice positively affected some children who performed inconsistently across the screeners.

Discussion

Disparities exist in the speech and language services provided to children, and AA children have been found to be both over- and underrepresented on speech and language caseloads. Children who speak AAE and reside in low-income homes are especially vulnerable to disparate

Table 6. Correlations by screeners' items and subtests.

Measure	1	2	3	4	5	6	7	8	9
1 DELV–Screening Test Risk Grammar									
2 DELV–Screening Test Risk Nonwords	–.07								
3 FLUHARTY-2 Articulation	.28*	–.28*							
4 FLUHARTY-2 Repeating Sentences	.62**	–.07	.38**						
5 FLUHARTY-2 Following Directives	.50**	<.01	.32**	.62**					
6 FLUHARTY-2 Describing Actions	.23*	.01	.07	.19	.45**				
7 FLUHARTY-2 Sequencing Events	.45**	–.01	.38**	.53**	.46**	.23			
8 WCLS PPVT-4	.40**	–.16	.30**	.55*	.42**	.17	.52**		
9 WCLS *Wh*-Questions	.61**	–.05	.20	.63**	.62**	.38**	.49**	.45**	
10 WCLS MCLUw	.01	–.04	.12	.06	.19	.19	.18	.06	.17

Note. DELV–Screening Test Risk = Diagnostic Evaluation of Language Variation–Screening Test Risk subtest; FLUHARTY-2 = Fluharty Preschool Speech and Language Screening Test–Second Edition; WCLS = Washington and Craig Language Screener; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; MCLUw = utterances as measured by C-units in words.

*Correlation is significant at the .05 level. **Correlation is significant at the .001 level.

practices when screeners yield high fail rates and inconsistent outcomes. High fail rates and inconsistent outcomes can tax a school system if all children who fail a screening are referred for speech and language evaluations. Children from low-income households may also have less access or delayed access to speech and language services if they do not have families who can advocate for services (Wittke & Spaulding, 2018) or they are required to wait for a second failed screening before being referred for an evaluation, as is done in the school district that participated in the current study.

One way speech-language pathologists can help reduce disparities within clinical practice is to examine how well tools perform with groups of children most vulnerable to clinical error. In this study, we did this by examining the DELV–Screening Test Risk subtest as a language screener for low-income pre-K children who speak AAE in the urban South. The evaluation involved a comparison of scores and outcomes from the DELV–Screening Test Risk subtest with those obtained from the FLUHARTY-2 and the WCLS. The DELV–Screening Test Risk subtest was designed for children who speak a variety of dialects, including AAE, and the standardization sample oversampled AA children, children who lived in the South, and low-income children. The FLUHARTY-2 was not designed for AAE-speaking children, but it is often recommended because AA children who were included within the standardization sample earned scores that were, on average, within 1 *SD* of the normative means. The WCLS was specifically designed for AA children, and it was tested with an AA sample that was community based and drawn from an urban city. Compared with the other screeners, the WCLS is also the only one that has been evaluated against a battery of clinical reference measures.

Unfortunately, results showed that the children's DELV–Screening Test Risk error scores were higher than those of its standardization sample, and their scores on the other two screeners were lower than their respective standardization/testing samples. The 52% fail rate of the DELV–Screening Test Risk did not differ significantly from the 48% rate of the WCLS. Fail rates of the FLUHARTY-2 ranged from 34% to 75%, depending on the quotients considered and whether scoring was modified for the children's use of nonmainstream English forms. Although items and subtests assumed to measure similar constructs were moderately correlated to each other, the screeners led to inconsistent pass/fail outcomes for 44% of the children. Using the children's a priori clinical classifications and considering the two children with a speech disorder as typical for language, sensitivity values for the DELV–Screening Test Risk, the FLUHARTY-2, and the WCLS were 1.0, .75, and .75, respectively, and specificity values were .51, .68, and .54, respectively. These findings indicate that the screeners were better able to classify children with speech and language disorders as impaired than classifying children without speech and language disorders as typical. Finally, for some children with inconsistent screening outcomes, there was some evidence that they benefited from practice as fewer failed the third screener administered compared with the first or second.

Previous Studies

The fail rates of the DELV–Screening Test Risk subtest and the WCLS are consistent with the 51% and 58% fail rates reported by Rhyner et al. (1999) for the Bankson Language Test (Bankson, 1990) and the SPELT-P (Werner & Kresheck, 1983), but they are higher than the 18% overall and 23% pre-K fail rates reported by Washington and Craig (2004) for the WCLS. They are also higher than fail rates reported for two other screening studies not reviewed earlier. Tomblin et al.'s (1997) study of childhood LI yielded a screening fail rate of 26.8% overall and 26.2% for monolingual English speakers. Their screener included 40 items drawn from the Test of Language Development–Primary: Second Edition (Newcomer & Hammill, 1988), and their participants were 7,218 kindergartners of various races and ethnicities who lived in the Midwest, were stratified by community (rural, urban, and suburban), and were described as varying in socioeconomic status. Weiler et al. (2018) reported a 16.48%–28.64% screening fail rate in their study of 148 kindergartners who lived in the rural South. Their screener involved two subtests from the Test of Early Grammatical Impairment (Rice & Wexler, 2001), and their participants were described as White and non-Hispanic. The poverty rate of their community was 15.9%, which was slightly higher than the national average, but much lower than the 25.2% poverty rate of the current community.

With modified scoring, the fail rate of the FLUHARTY-2 was 34% utilizing the General Language Quotient only and 56% utilizing one or more quotients. Without modified scoring, fail rates were 56% utilizing the General Language Quotient and 75% utilizing one or more quotients. This wide range of fail rates makes it difficult to know which one should be used to compare results from this screener with those obtained from previous studies. We also cannot directly compare the current FLUHARTY-2 results with Eisenberg et al. (2019) because that study involved a comparison of preselected groups based on their screening outcomes.

Clinical Implications

The findings of this study show that, like other screeners, the DELV–Screening Test Risk subtest can lead to higher error scores than its standardization sample and high fail rates when administered to low-income children who speak AAE and live in the urban South. The DELV–Screening Test Risk, the FLUHARTY-2, and the WCLS—which are all screeners that have been recommended for AA children—may also lead to inconsistent screening outcomes. These findings indicate that clinicians should be cautious when screening AA children who present with a similar sociolinguistic profile as those studied here, even when they are using tools recommended for AA children. These findings also lend support for the participating school district's practice of requiring children to fail two screeners before

referring a child for a full evaluation by a speech-language pathologist.

Although it is tempting to recommend the FLUHARTY-2 over the DELV–Screening Test Risk subtest or the WCLS given its 34% fail rate, this outcome occurred only when using the General Language Quotient and modified scoring. The FLUHARTY-2 manual recommends further testing for any child who fails one or more of the subtests. With all subtests considered as the manual recommends, the fail rate was extremely high at 75% without modified scoring and comparable to the rates of the other screeners at 56% with modified scoring. Also, the Describing Actions subtest did not correlate with the other language measures when the scores were modified, so it is unclear what this subtest is measuring.

The wide range of scores obtained on the FLUHARTY-2 is also worrisome as not all clinicians will know the AAE dialects of their communities or have the resources and school support to modify scores or make interpretive adjustments of a published tool. Recall that Eisenberg et al. (2019) evaluated the concurrent validity of FLUHARTY-2 using the SPELT and three language sample measures. Although these authors call into question the use of the FLUHARTY-2 for clinical practice, the evidence supporting their conclusion was based on the low level of agreement found between those who failed the screener and the other measures. Of less concern was the appropriateness of the FLUHARTY-2 for the children's dialects, and scoring was not modified, even though fail rates varied by the children's race. We raise this issue not to criticize this study because measures of a child's dialect are relatively new to the field as are studies of scoring modifications. Instead, the study is highlighted to call for a greater focus on the sociolinguistic profiles of study participants, including direct measures of their dialects, and the effects of modified scoring approaches within screening studies. This type of work is urgently needed because in three recent studies, modified scoring systems have led to increased rates of underidentification of AA children with LIs (Hendricks & Adolf, 2017; Oetting et al., 2019, 2021).

Defining characteristics of the current AA sample were the children's low-income status, age of approximately 4 years, enrollment in Head Start or a public pre-K in a large school district within the urban South, and very high percentages (M = 95%) of nonmainstream form use as measured by the DELV–Screening Test Dialect subtest. Future studies of screeners in different types of communities and with different types of participants will likely yield different fail rates. Specifying the characteristics of future participants, including the dialect(s) they speak and the densities of their nonmainstream English forms, as well as reporting results by participant characteristics should help move the field forward and lead to a better understanding of how tools work for well-defined groups of children within and across communities.

The findings also raise the question as to whether it is reasonable to expect any single tool or any single set of norms to be used to screen children's language abilities in the United States and elsewhere. Use of the DELV–Screening Test Risk subtest allows clinicians to rule out dialect mismatches as a reason for a child's higher-than-expected error scores, but as was shown here, there are other variables, such as low socioeconomic status, that can contribute to children's screening outcomes. Given this, the field may not need yet another screener or another set of nation-wide norms. Instead, what might be needed is refinement of existing screeners and the collection of multiple normative data sets for making decisions about children's abilities. If additional screeners or refinement of screeners are pursued, then content from the items and subtests that correlated with each other in this study should be considered. This content included the DELV–Screening Test Risk grammar items, the FLUHARTY-2 Repeating Sentences and Following Directions subtests, and the *Wh*-Question Task. Future efforts should also explore adding testing practice to a screening program as some children benefited from practice in the current study.

Future studies of screeners may also want to consider adding caregiver/teacher ratings and risk factors for LI into the decision-making process. There is growing evidence to support the use of caregiver/teacher ratings for monolingual and bilingual speakers of English (e.g., Ebert et al., 2020; Gregory & Oetting, 2018; Pua et al., 2017), although recent work by Hendricks and Jimenez (2021) indicates that teachers may need additional training about dialects when children are speakers of AAE. There is also an accumulating literature on factors that increase a child's risk for LI. In a review of 16 studies, Nelson et al. (2006) found the most consistently reported risk factors to be a positive family history of speech and language difficulties, biological male gender, and perinatal factors. Wallace et al. (2015) found a similar finding from a review of 23 studies but added caregiver education to the list. Rudolph (2017) reviewed eight studies and found 11 risk factors, with maternal education, 5-min Apgar score, birth order, and biological male gender identified as clinically significant and as predictive of LI as late talker status.

Limitations

We did not examine the classification accuracy of the DELV–Screening Test Risk subtest or the other screeners using a comprehensive battery of clinical reference measures. Instead, we compared the children's scores with those from each screener's standardization/testing samples and evaluated their pass/fail outcomes relative to each other's fail rates and the children's a priori clinical classifications. The study sample, which included six children with a diagnosed communication disorder, may have included some children with undiagnosed communication disorders, especially since the children lived in a state found to underrepresent AA children in the disability category of speech and LI. Even so, these potentially undiagnosed children would not explain the extremely high fail rates of the DELV–Screening Test Risk subtest and the WCLS, the wide range of fail rates obtained for the FLUHARTY-2, or the inconsistent outcomes across screeners. Finally, the study was not conducted with blinding. The first author administered all of screeners. Unblinded studies typically generate inflated

results, which in this study would have led to higher-than-justified levels of consistency across the three screeners' outcomes. Inflated results are less of a concern in this study because the screening outcomes were inconsistent for almost half of the children.

Conclusions

Like other screeners including the FLUHARTY-2 and the WCLS, the DELV–Screening Test Risk subtest may lead to high fail rates and inconsistent outcomes for low-income pre-K children who speak AAE in the urban South. This finding is noteworthy as the DELV–Screening Test Risk subtest, the FLUHARTY-2, and the WCLS have all been recommended for AA children. This finding also underscores the critical need for more development and study of screeners within the field. Recommendations for future studies include focusing on content such as the DELV–Screening Test Risk grammar items, the FLUHARTY-2 Repeating Sentences and Following Directions subtests, and the WCLS *Wh*-Question Task; considering testing practice as part of the screening process; and incorporating caregiver/teacher ratings and risk factors into the decision-making process. These recommendations need to be tested for their feasibility, effectiveness, economic efficiency, equality, and equity in many different types of communities and with well-specified groups of children, especially those who are most vulnerable to clinical error.

Acknowledgments

During data collection, the first author was supported by a Huel Perkins Fellowship from Louisiana State University, and the second author was supported by National Institute on Deafness and Other Communication Disorders Grant RO1DC009811. Although scoring and analyses were updated, the data were collected as part of the first author's dissertation. Appreciation is extended to Jessica Richardson Berry, Karmen Porter, Andy Rivière, Brittany Harris, Missy Monaghan, Danielle Morrill, and Kaitlyn Rodrigue for help with data scoring and Julie Washington for help with the scoring of the *Wh*-Question Task. The authors would also like to thank the schools, Head Starts, families, and children who participated in the study.

References

Arwood, E. L. (1985). *Apricot I language kit*. Apricot.

Bankson, N. W. (1990). *Bankson Language Test-2*. Pro-Ed.

Berry, J. R., & Oetting, J. B. (2017). Dialect variation of copula and auxiliary verb BE: African American English–speaking children with and without Gullah/Geechee heritage. *Journal of Speech, Language, and Hearing Research, 60*(9), 2557–2568. https://doi.org/10.1044/2017_JSLHR-L-16-0120

Bishop, D. V. M., & McDonald, D. (2009). Identifying language impairment in children: Combining language test scores with parental report. *International Journal of Communication Disorders, 44,* 600–615. https://doi.org/10.1080/13682820802259662

Brown, G. R. (2017). *Pronoun marking in African American English–speaking children with and without specific language impairment* [Master's thesis, Louisiana State University].

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment. *Journal of Speech, Language, and Hearing Research, 40*(3), 519–525. https://doi.org/10.1044/jslhr.4003.519

Chinn, P. C., & Hughes, S. (1987). Representation of minority students in special education classes. *Remedial and Special Education, 8*(4), 41–46. https://doi.org/10.1177/074193258700800406

Ciolli, L., & Seymour, H. (2004). Dialect identification versus evaluation of risk in language screening. *Seminars in Speech and Language, 25*(1), 33–40. https://doi.org/10.1055/s-2004-824824

Dawson, J., Stout, C., Eyer, J., Tattersall, P., Fonkalsrud, J., & Croley, K. (2005). *Structured Photographic Expressive Language Test–Preschool 2*. Janelle Publications.

De Valenzuela, J. S., Copeland, S. R., Qi, C. H., & Park, M. (2006). Examining educational equity: Revisiting the disproportionate representation of minority students in special education. *Exceptional Children, 72,* 425–441. https://doi.org/10.1177/001440290607200403

Dollaghan, C., & Campbell, T. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136–1146. https://doi.org/10.1044/jslhr.4105.1136

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–III*. Pearson.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition*. Pearson.

Ebert, K., Ochoa-Lubinoff, C., & Holmes, M. (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

Eisenberg, S., Victorino, K., & Murray, S. (2019). Concurrent validity of the Fluharty Preschool Speech and Language Screening Test–Second Edition at age 3: Comparison with four diagnostic measures. *Language, Speech, and Hearing Services in Schools, 50*(4), 673–682. https://doi.org/10.1044/2019_LSHSS-18-0099

Fluharty, N. B. (2001a). *Fluharty Preschool Speech and Language Screening Test–Second Edition*. Pro-Ed.

Fluharty, N. B. (2001b). *Fluharty Preschool Speech and Language Screening Test–Second Edition: Examiner's manual*. Pro-Ed.

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Gregory, K. D., & Oetting, J. B. (2018). Classification accuracy of teacher ratings when screening nonmainstream English-speaking kindergartners for language impairment in the rural South. *Language, Speech, and Hearing Services in Schools, 49*(2), 218–231. https://doi.org/10.1044/2017_LSHSS-17-0045

Hall-Mills, S. (2019). A comparison of the prevalence rates of language impairment before and after response-to-intervention implementation. *Language, Speech, and Hearing Services in Schools, 50*(4), 703–709. https://doi.org/10.1044/2019_LSHSS-18-0144

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. (2020). Production of morphosyntax within and across different dialects of American English. *Journal of Speech, Language, and Hearing Research, 63*(7), 2322–2333. https://doi.org/10.1044/2020_JSLHR-19-00244

Hendricks, A. E., & Jimenez, C. (2021). Teacher report of students' dialect use and language ability. *Language, Speech, and Hearing Services in Schools, 52*(1), 131–138. https://doi.org/10.1044/2020_LSHSS-19-00113

Horner, C. M., Maddux, C. D., & Green, C. (1986). Minority students and special education: Is overrepresentation possible? *NASSP Bulletin, 70*(492), 89–93. https://doi.org/10.1177/019263658607049219

Horton-Ikard, R., & Apel, K. (2014). Examining the use of spoken dialect indices with African American children in the Southern United States. *American Journal of Speech-Language Pathology, 23*(3), 448–460. https://doi.org/10.1044/2014_AJSLP-13-0028

Keegstra, A. L., Knijff, W. A., Post, W. J., & Goorjuis-Brouwer, S. M. (2007). Children with language problems in a speech and hearing clinic: Background variables and extent of language problems. *International Journal of Pediatric Otorhinolaryngology, 71*(5), 815–821. https://doi.org/10.1016/j.ijporl.2007.02.001

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52,* 377–405. https://doi.org/10.1016/j.jsp.2014.06.002

Lee, R., & Oetting, J. B. (2014). Zero marking of past tense in child African American English. *SIG 1 Language Learning and Education, 21*(4), 173–181. https://doi.org/10.1044/lle21.4.173

Maher, Z. K., Erskine, M. E., Byrd, A. S., Harring, J. R., & Edwards, J. R. (2021). African American English and early literacy: A comparison of approaches to quantifying nonmainstream dialect use. *Language, Speech, and Hearing Services in Schools, 52*(1), 118–130. https://doi.org/10.1044/2020_LSHSS-19-00115

McDonald, J., & Oetting, J. B. (2019). Nonword repetition across two dialects of English: Effects of specific language impairment and nonmainstream form density. *Journal of Speech, Language, and Hearing Research, 62*(5), 1381–1391. https://doi.org/10.1044/2018_JSLHR-L-18-0253

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Macuga, S. (2017). Replicated evidence of racial and ethnic disparities in disability identification in U.S. schools. *Educational Researcher, 46*(6), 305–322. https://doi.org/10.3102/0013189X17726282

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Macuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher, 44*(5), 278–292. https://doi.org/10.3102/0013189X15591157

Morgan, P. L., Hammer, C. S., Farkas, G., Hillemeier, M. M., Maczuga, S., Cook, M., & Morano, S. (2016). Who receives speech/language services by 5 years of age in the United States? *American Journal of Speech-Language Pathology, 25*(2), 183–199. https://doi.org/10.1044/2015_AJSLP-14-0201

Moyle, M. J., Heilmann, J. J., & Finneran, D. A. (2014). The role of dialect density in nonword repetition performance: An examination with at-risk African American preschool children. *Clinical Linguistics & Phonetics, 28*(9), 682–696. https://doi.org/10.3109/02699206.2014.882990

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: Systematic evidence review for the U.S. preventive services task force. *Pediatrics, 117*(2), e298–e319. https://doi.org/10.1542/peds.2005-1467

Newcomer, P., & Hammill, D. (1988). *Test of Language Development–Primary: Second Edition*. Pro-Ed.

Oetting, J. B., Berry, J. R., Gregory, K. D., Rivière, A. M., & McDonald, J. (2019). Specific language impairment in African American English and Southern White English: Measures of tense and agreement with dialect-informed probes and strategic scoring. *Journal of Speech, Language, and Hearing Research, 62*(9), 3443–3461. https://doi.org/10.1044/2019_JSLHR-L-19-0089

Oetting, J. B., Cleveland, L. H., & Cope, R. (2008). Empirically-derived combinations of tools and clinical cutoffs: An illustrative case with a sample of culturally/linguistically diverse children. *Language, Speech, and Hearing Services in Schools, 39*(1), 44–53. https://doi.org/10.1044/0161-1461(2008/005)

Oetting, J. B., & McDonald, J. (2001). Nonmainstream dialect use and specific language impairment. *Journal of Speech, Language, and Hearing Research, 44*(1), 207–223. https://doi.org/10.1044/1092-4388(2001/018)

Oetting, J. B., Rivière, A. M., Berry, J. R., Gregory, K. D., Villa, T. M., & McDonald, J. (2021). Marking of tense and agreement in language samples by children with and without specific language impairment in African American English and Southern White English: Evaluation of scoring approaches and cut scores across structures. *Journal of Speech, Language, and Hearing Research, 64*(2), 491–509. https://doi.org/10.1044/2020_JSLHR-20-00243

Peña, E. D., & Quinn, R. (1997). Task familiarity: Effects on the test performance of Puerto Rican and African American children. *Language, Speech, and Hearing Services in Schools, 28*(4), 323–332. https://doi.org/10.1044/0161-1461.2804.323

Petscher, Y., Connor, C. M., & Al'Otaiba, S. (2012). Psychometric analysis of the diagnostic evaluation of language variation assessment. *Assessment for Effective Intervention, 37*(4), 243–250. https://doi.org/10.1177/1534508411413760

Pruitt, S. L., & Oetting, J. B. (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)

Pua, E. P., Lee, M. L., & Liow, S. J. (2017). Screening bilingual preschoolers for language difficulties: Utility of teacher and parent reports. *Journal of Speech Language and Hearing Research, 60*(4), 950–968. https://doi.org/10.1044/2016_JSLHR-L-16-0122

Rhyner, P. M., Kelly, D. J., Brantley, A. L., & Krueger, D. M. (1999). Screening low-income African American children using the BLT-2S and the SPELT-P. *American Journal of Speech-Language Pathology, 8*(1), 44–52. https://doi.org/10.1044/1058-0360.0801.44

Rice, M., & Wexler, K. (2001). *Rice/Wexler Test of Early Grammatical Impairment*. University of Kansas. https://cldp.ku.edu/rice-wexler-tegi

Robinson, G., & Norton, P. (2019). A decade of disproportionality: A state-level analysis of African American students enrolled in the primary disability category of speech or language impairment. *Language, Speech, and Hearing Services in Schools, 50*(2), 267–282. https://doi.org/10.1044/2018_LSHSS-17-0149

Rodekohr, R. K., & Haynes, W. O. (2001). Differentiating dialect from disorder: A comparison of two processing tasks and a standardized language test. *Journal of Communication Disorders, 34*(3), 255–272. https://doi.org/10.1016/S0021-9924(01)00050-8

Roy, J., Oetting, J. B., & Wynn Moland, C. (2013). Linguistic constraints on children's overt marking of BE by dialect and age. *Journal of Speech, Language, and Hearing Research, 56*(3), 933–944. https://doi.org/10.1044/1092-4388(2012/12-0099)

Rudolph, J. M. (2017). Case history risk factors for specific language impairment: A systematic review and meta-analysis. *American Journal of Speech-Language Pathology, 26*(3), 991–1010. https://doi.org/10.1044/2016_AJSLP-15-0181

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11,* 1–22. https://doi.org/10.1017/S0142716400008262

Seymour, H. N., Bland-Stewart, L., & Green, L. J. (1998). Difference versus deficit in child African American English. *Language, Speech, and Hearing Services in Schools, 29*(2), 96–108. https://doi.org/10.1044/0161-1461.2902.96

Seymour, H. N., Roeper, T. W., & de Villiers, J. (2003a). *Diagnostic Evaluation of Language Variation–Criterion Referenced*. The Psychological Corporation.

Seymour, H. N., Roeper, T. W., & de Villiers, J. (2003b). *Diagnostic Evaluation of Language Variation–Screening Test*. The Psychological Corporation.

Seymour, H. N., Roeper, T. W., & de Villiers, J. (2003c). *Diagnostic Evaluation of Language Variation–Screening Test: Examiner's manual*. The Psychological Corporation.

Skiba, R. J., Artiles, A. J., Kozleski, E. B., Losen, D. J., & Harry, E. G. (2016). Risks and consequences of oversimplifying educational inequities. *Educational Researcher, 45,* 221–225. https://doi.org/10.3102/0013189X16644606

Skiba, R. J., Poloni-Staudinger, L., Gallini, S., Simmons, A. G., & Feggins-Azziz, R. (2006). Disparate access: The disproportionality of African American students with disabilities across educational environments. *Exceptional Children, 72*(4), 411–424. https://doi.org/10.1177/001440290607200402

Sturner, R. A., Heller, J. H., Funk, S. G., & Layton, T. L. (1993). The Fluharty Preschool Speech and Language Screening Test. *Journal of Speech and Hearing Research, 36*(4), 738–745. https://doi.org/10.1044/jshr.3604.738

Sullivan, A. L., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children, 79,* 475–494. https://doi.org/10.1177/001440291307900406

Terry, N. P., Connor, C. M., Thomas-Tate, S., & Love, M. (2010). Examining relationships among dialect variation, literacy skills, and school context in first grade. *Journal of Speech, Language, and Hearing Research, 53*(1), 126–145. https://doi.org/10.1044/1092-4388(2009/08-0058)

Terry, N. P., Petscher, Y., & Rhodes, K. (2017). Psychometric analysis of the Diagnostic Evaluation of Language Variation–Screening Test: Extension to low-income African American prekindergarteners. *Assessment for Effective Intervention, 42*(3), 176–185. https://doi.org/10.1177/1534508416679402

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

U.S. Census Bureau. (2010). *Baton Rouge City, Louisiana* [Data file]. https://data.census.gov/cedsci/profile?g=1600000US2205000&tid=ACSDP1Y2018.DP05

Wallace, I. F., Berkman, N. D., Watson, L. R., Coyne-Beasley, T., Wood, C. T., Cullen, K., & Lohr, K. N. (2015). Screening for speech and language delay in children 5 years old and younger: A systematic review. *Pediatrics, 136,* e448–e462. https://doi.org/10.1542/peds.2014-3889

Washington, J. A., & Craig, H. K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology, 13*(4), 329–340. https://doi.org/10.1044/1058-0360(2004/033)

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. https://doi.org/10.1044/2017_LSHSS-17-0071

Werner, E. O., & Kresheck, J. D. (1983). *Structured Photographic Expressive Language Test–Preschool*. Janelle Publications.

Wittke, K., & Spaulding, T. J. (2018). Which preschool children with specific language impairment receive language intervention. *Language, Speech, and Hearing Services in Schools, 49*(1), 59–71. https://doi.org/10.1044/2017_LSHSS-17-0024

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39,* 204–221. https://doi.org/10.1093/jpepsy/jst062