Research Article

Language Assessment With Children Who Speak Nonmainstream Dialects: Examining the Effects of Scoring Modifications in Norm-Referenced Assessment

Alison Eisel Hendricks^a and Suzanne M. Adlof^a

Purpose: We compared outcomes from 2 measures of language ability in children who displayed a range of dialect variation: 1 using features that do not contrast between mainstream American English (MAE) and nonmainstream dialects (NMAE), and 1 using contrastive features. We investigated how modified scoring procedures affected the diagnostic accuracy of the measure with contrastive features.

Method: Second-grade students ($N = 299$; 167 White, 106 African American, 26 other) completed measures of language variation and ability (the Diagnostic Evaluation of Language Variation–Screening Test and the Clinical Evaluation of Language Fundamentals–Fourth Edition [CELF-4]). The CELF-4 was scored with and without the recommended scoring modifications for children who spoke African American English.

Results: Partial correlations controlling for socioeconomic status revealed small to moderate correlations between measures of language ability and the use of NMAE features. Modified scoring yielded higher scores for children who spoke African American English and a reduced association between the use of NMAE features and CELF-4 scores. Modified scoring also affected the diagnostic accuracy of the CELF-4, resulting in a lower positive likelihood ratio and a higher negative likelihood ratio.

Conclusions: The decision to apply scoring modifications affects both the false positive and false negative rates. Implications for language assessment for children who speak NMAE dialects are discussed, including the need for further investigation.

Identifying language impairment (LI) in children who speak nonmainstream American English (NMAE) dialects is difficult (Oetting, Lee, & Porter, 2013; Seymour, 2004; Seymour, Bland-Stewart, & Green, 1998). Clinical diagnostic decisions are complicated by a number of factors, including persistent bias in standardized language assessments (Qi, Kaiser, Milan, & Hancock, 2006; Stockman, 2000, 2010; Thomas-Tate, Washington, & Edwards, 2004; Washington & Craig, 1999; Wyatt, 2012) and the fact that many of the features that have been demonstrated to be clinical markers of language impairment in speakers of mainstream American English (MAE) may be produced variably in NMAE (Lee & Oetting, 2014; Oetting & Horohov, 1997; Oetting & McDonald, 2001). This overlap between grammatical

uses of variable NMAE features and ungrammatical uses of MAE makes it difficult to determine whether a particular instance reflects LI, use of an NMAE feature, or both.

Historically, some researchers and clinicians historically conflated use of NMAE features and language ability, and there was little recognition that language differences were not indicators of language deficits (Seymour, 2004). As a consequence, children who speak NMAE dialects, such as African American English (AAE), and have typical language skills may have been erroneously considered to have LI on the basis of the dialect they speak. Since the 1970s and 1980s, there has been an increased interest in improving the accuracy of LI diagnoses in children who speak NMAE dialects. In 2003, the American Speech-Language-Hearing Association published a technical report on American English dialects that stated, “An essential step toward making accurate assessments of communication disorders is to distinguish between those aspects of linguistic variation that represent regular patterns in the speaker’s dialect and those that represent true disorders in speech and language.” The technical

^aUniversity of South Carolina, Columbia

Correspondence to Alison Eisel Hendricks, who is now at the State University of New York (SUNY) at Buffalo: ahendric@buffalo.edu

Editor: Shelley Gray

Associate Editor: Monique Mills

Received July 27, 2016

Revision received December 23, 2016

Accepted April 14, 2017

https://doi.org/10.1044/2017_LSHSS-16-0060

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

report further outlined the required competencies for speech-language pathologists (SLPs) to accurately distinguish between grammatical uses of NMAE and communicative disorders, including “recognizing all American English dialects as rule-governed linguistic systems, understanding the rules and linguistic features of American English dialect(s) represented by their clientele and being familiar with non-discriminatory testing and dynamic assessment procedures.”

Thus, it is important to understand the similarities and differences between dialects of English, such as AAE and MAE. For example, in AAE regular past tense marking is produced variably, meaning that sometimes it is overtly produced (e.g., *She kissed him*) and other times it is not overtly produced (e.g., *He walk away*; Oetting & McDonald, 2002). Although cross-dialect differences in morphosyntax are most commonly associated with the difficulty in identifying LI, it is important to note that there are differences in the domains of phonology (e.g., final consonant-cluster reduction, as in MAE *cold* /kold/ to AAE *cold* /kol/; Green, 2002), vocabulary (Smitherman, 1998), discourse and pragmatics (e.g., different frequency of use of certain discourse forms; Morgan, 1998; Smitherman, 1977), and narrative structures (Champion, 2003). Similar to past tense marking, many of these features are also produced variably.

Diagnostic decisions may be further complicated by factors, such as socioeconomic status (SES), which are correlated with the use of NMAE features. It is well established cross-linguistically that speakers from lower SES backgrounds tend to use a larger range of nonmainstream forms and to use nonmainstream forms more frequently than speakers from higher SES backgrounds. This has been well documented in adult speakers in the United States (Labov, 1966, 1990) and internationally (Trudgill, 1972). Likewise, children with typical development follow similar patterns of use of nonmainstream dialect features (Labov, 1989; Miller, 2013; Roberts, 1994; Smith, Durham, & Fortune, 2007, 2009). Although speakers from lower SES backgrounds tend to use nonmainstream forms more frequently, it is important to note that children who speak NMAE dialects do not come from exclusively low SES backgrounds (Horton-Ikard & Miller, 2004) and that not all children from low SES backgrounds speak NMAE dialects. On the other hand, regardless of which dialect they speak, children of low SES frequently demonstrate lower mean performance on measures of language ability (e.g., Dollaghan et al., 1999; Hart & Risley, 1995; Qi et al., 2006). These differences are most commonly attributed to differences in maternal language input (Hart & Risley, 1995; Hoff, 2003). However, children of poverty also have a greater incidence of low birth weight, malnutrition, and reduced access to quality health care, as well as other factors placing them at higher risk for LI and other developmental disabilities (Morgan et al., 2017). Thus, clinicians are faced with a complex task when determining whether children who speak NMAE dialects exhibit LI.

During the diagnostic process, SLPs must make at least two decisions: (i) Is the child speaking a mainstream or nonmainstream dialect, and (ii) does the child exhibit LI? There are several methods for answering the first

question (Horton & Apel, 2014; Oetting & McDonald, 2002). For example, language sampling is one method of examining children’s use of NMAE features (Garrity & Oetting, 2010; Pruitt & Oetting, 2009). There are two common analyses of language samples to determine whether a child speaks a mainstream or nonmainstream dialect: dialect density analysis, which determines how frequently a child uses a predefined set of NMAE features (Washington & Craig, 2002), and listener judgments (Garrity & Oetting, 2010; Oetting & Garrity, 2006; Oetting & McDonald, 2002; Pruitt, Oetting, & Hegarty, 2011). However, it is important to note that how often a speaker uses NMAE features in language samples is influenced by a number of sociolinguistic variables, such as the formality of the context in which the language sample is collected and the type of language sample collected (e.g., narrative versus expository; Mills, Watkins, & Washington, 2013; Thompson, Craig, & Washington, 2004; Washington, Craig, & Kushmaul, 1998), SES (e.g., Horton-Ikard & Miller, 2004), and gender (Horton-Ikard & Miller, 2004; Washington et al., 1998). Another approach to examining the use of NMAE features involves using a standardized measure of NMAE dialect use, in which a standardized set of tasks is used to elicit a variety of potential NMAE dialect features. With this approach, the examination of which dialect a child speaks occurs in the same formal context as the probing of language ability. The first subtest within the Diagnostic Evaluation of Language Variation–Screening Test (DELV-S; Seymour, Roeper, deVilliers, & deVilliers, 2003) is an example of this approach.

It is more difficult to answer the second question, whether a child exhibits LI. Even after a clinician has established that a child speaks a nonmainstream dialect, it is not always clear whether a particular utterance that differs from the mainstream grammar represents dialectal variation or a marker of LI. Several approaches have been proposed for developing dialect-neutral assessments to identify children with LI. For example, researchers have examined the use of processing-based measures, such as nonword repetition (e.g., Campbell, Dollaghan, Needleman, & Janosky, 1997; Dollaghan & Campbell, 1998; but see Moyle, Heilmann, & Finneran, 2014; Oetting, et al., 2013; Oetting & Cleveland, 2006) and dynamic assessment methods (e.g., Peña et al., 2016). These methods place the focus of assessment on the child’s perceived potential for acquiring language versus the current level of achievement. Although these measures have shown promise in various preliminary investigations, at this time there is limited normative information available to provide a generalized standard of reference for making diagnostic decisions.

Another approach separates linguistic features into *contrastive features*, which differ between MAE and NMAE dialects, and *noncontrastive features*, which are the same across MAE and NMAE dialects (Craig & Washington, 2006; Seymour et al., 1998). Under this approach, difficulty with noncontrastive features could provide evidence of LI for speakers of both MAE and NMAE. The DELV-S and the Diagnostic Evaluation of Language Variation–Norm

Referenced (DELV-NR; Seymour et al., 2005) follow this approach. The DELV-S and DELV-NR are normed for ages 4;0 (years;months) to 9;11. According to the manual, the diagnostic accuracy of the DELV-NR was examined in a clinical sample of 88 children aged 4;0–9;11 with diagnosed language disorders, and a control group of 88 children from the norming sample who were matched on age, parental education, region, and sex. Using a standard score cut-point of 85, the DELV-NR was determined to provide excellent discrimination, with 95% sensitivity and 93% specificity.

A third approach is to use an assessment developed for MAE speakers and modify the scoring so as not to penalize responses which are grammatical within the child's dialect. For example, when applying scoring modifications such as those suggested by the Clinical Evaluation of Language Fundamentals—Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003) and Fifth Edition (Wiig, Semel, & Secord, 2013), clinicians are encouraged to give children credit for responses that are grammatical within their dialects but ungrammatical in MAE (e.g., in response to “Here the bird____,” *flies* and *fly* are both scored as correct). Although this approach is suggested in test manuals, as well as the 2003 technical report from the American Speech-Language-Hearing Association, empirical data regarding the validity of scoring modifications are limited. In a study of 99 African American preschoolers aged 4;0–5;11, Rhyner, Kelly, Brantley, & Krueger (1999) found that the application of scoring modifications led to slight increases in raw scores and small to moderate reductions in observed screening failures. Those researchers did not have a gold-standard measure against which to compare modified versus unmodified scoring approaches. J. M. Terry, Jackson, Evangelou, and Smith (2010) examined the influence of AAE features on the performance of preschool children with typical development on a sentence-repetition task. They found that the influence of AAE features was lower for the modified scores, but scoring modifications were not able to fully correct for potential test bias. Oetting, McDonald, Seidel, and Hegarty (2016) also investigated the impact of dialect-specific scoring procedures on an experimental sentence-recall task, similar to the CELF-4 Recalling Sentences subtest, and found that the application of strategic scoring modifications provided high sensitivity and specificity. Modifications were made for three features that were highly frequent in NMAE but have not been demonstrated to be clinical markers in children who speak MAE or NMAE dialects. However, the researchers found that scoring sentences on the basis of exact repetition accuracy (i.e., without modifications) yielded similarly high sensitivity and specificity. Taken together, these preliminary findings raise questions about how to implement scoring modifications and whether scoring modifications lead to better diagnostic decisions.

NMAE and Diagnostic Decisions

A recent survey of SLPs by Betz, Eickhoff, and Sullivan (2013) similarly raises questions about the extent to which tests developed for MAE are being used with

speakers of NMAE dialects. An analysis of survey responses from 364 SLPs in the United States indicated that clinicians preferred to use assessments with a long track record, which have recently updated norms available (i.e., tests that were on their second or later version). The CELF-4 was the most frequently used assessment, with 67% of SLPs reporting they used it at least sometimes in their practice. In contrast, clinicians rarely reported using the DELV-NR or DELV-S; their average reported frequencies of use were just over 1.0 on a scale of 1–5 (where 1 corresponded with *never* and 5 corresponded with *always*). The survey did not specifically address clinician's approach to diagnosis of culturally and linguistically diverse populations, or their techniques beyond standardized assessments, so it is not clear whether clinicians tended to use alternative assessment approaches (e.g., dynamic assessment), scoring modifications, or some other strategy with these groups. A recent survey of SLPs working with bilingual children similarly found that although SLPs reported increased use of alternative assessment procedures, including dynamic assessment, they also continue to use standardized measures of language ability, including some assessments not specifically designed for bilingual populations (Arias & Friberg, 2017).

Although the application of scoring modifications may appear straightforward, accurate application of scoring modifications is complicated by a number of factors. One is that scoring modifications apply only to language-production tasks and not to language-comprehension tasks. That is, if a child produces a form that includes an NMAE feature that is ungrammatical in MAE, a clinician can modify the scoring; however, there is no adjustment to the subtests that involve language comprehension. Second, the number of items potentially affected by scoring modifications differs across subtests but theoretically may be as high as 100% for some, such as those that require sentence generation. Further, because children have flexibility in selecting the forms they produce on sentence-generation tasks, the number of total items affected by scoring modifications will vary across children. Studies to date have not examined the extent to which scoring modifications affect different types of assessments on omnibus measures of language ability, such as the CELF-4.

In addition, it is unclear whether to apply all possible scoring modifications for any given child. That is, if a clinician determines that a child speaks an NMAE dialect, this does not mean that it is always appropriate to apply scoring modifications to all items. Speakers differ in how often they use NMAE features, with some children speaking dialects that vary strongly from MAE and others speaking dialects with less variation from MAE (Washington & Craig, 1994). The general trend is that children who speak NMAE upon school entry tend to use NMAE features less frequently over time, but there is substantial variability between children as to how their use of NMAE features changes over time (Van Hofwegen & Wolfram, 2010). In addition, use of NMAE features differs depending on the context, with NMAE features more common in informal than formal contexts (Labov, 1966, 1990), as well as the

type of discourse. Certain contexts, such as narratives, provide more opportunities for NMAE feature use than expository contexts. Thus, contextual changes and changes over development complicate decision making regarding how and when to apply scoring modifications for children who speak NMAE dialects.

Further, not all NMAE features are equally frequent in NMAE dialects. For example, Washington and Craig (1994) found that although nearly all children who speak AAE use zero copula (e.g., *My mom o nice.*), only children who speak dialects with the strongest variation from mainstream use zero marked plural (e.g., *Dogo are fun.*). In a similar manner, some features, such as zero marking on regular third person singular verb agreement (e.g., *He walko*), are common whereas other features, such as zero marked past tense (e.g., *Yesterday, she walko the dog*), were used less frequently (Lee & Oetting, 2014; Oetting & McDonald, 2001). This finding holds true both for children with typical development and children with LI, although children with LI use NMAE features more frequently than age- and language-matched peers with typical development (Oetting & McDonald, 2001). Because NMAE features differ in how likely they are to be produced, a clinician must decide whether to apply all possible scoring modifications or limit them in some way, for example, only modifying scores for forms that are produced more frequently in the NMAE dialect.

Study Purpose

To date, the validity of modified scoring procedures for language assessment with speakers of NMAE dialects has not been well investigated. The current study's purpose was twofold: Our first objective was to compare the association between frequency of NMAE use and two measures of language ability in school-age children—one designed for children who speak NMAE dialects (DELV-S) and one not specifically designed for children who speak NMAE dialects (CELF-4). Our second objective was to investigate how the use of modified scoring procedures affected the diagnostic accuracy of the CELF-4. On the basis of data from extant studies (Connor & Craig, 2016; N. P. Terry & Connor, 2012; N. P. Terry, Connor, Petscher, & Conlin, 2012), we hypothesized that the frequency of NMAE feature use would be positively correlated with continuous measures of language ability. However, our question about diagnostic accuracy was exploratory. The study involved a secondary analysis of a data set from a larger project investigating subtypes of language and reading disorders in second-grade students (Adlof, Scoggins, Brazendale, Babb, & Petscher, 2017). The specific focus on second-grade students was advantageous in that they had 2 years of formal schooling but were still primarily acquiring language skills (vocabulary, complex syntax) on the basis of oral input.

Method

Study procedures were approved by the University of South Carolina Institutional Review Board. Study

invitations were distributed by classroom teachers to parents of second-grade students. Parents provided informed consent and completed a demographic questionnaire containing information about their child's medical and educational history, languages spoken by the child, and race and ethnicity. Child participants completed a battery of norm-referenced and experimental language, reading, and cognitive assessments in a quiet room of their school in one to three brief sessions. Data-collection sessions were video- and audio-recorded for offline scoring and reliability checking.

Participants

All children included in the study were recruited from one school district in South Carolina and were monolingual, without hearing loss, motor disorder, or other diagnosed physical or medical problems that would interfere with speech or language development. On the basis of the location and demographics of the study, we assume that children in this sample speak several dialects of English, including MAE, AAE, and Southern White English. Due to the larger project's focus on language and reading disorders, children at risk for language or reading difficulties were invited to participate more often than children who were not at risk. Therefore, the current sample has a larger percentage of children with LI than the general population. The analyses in this study involved 299 second-grade students (M age = 8;0; SD = 0;5; range = 7;3–9;4) who had complete data for two primary measures of interest (DELV-S and CELF-4; see Measures section). The sample comprised 135 (45.2%) boys and 158 (52.8%) girls, with gender information not reported for six (2.0%) participants. Free- and reduced-lunch status was obtained via school report: 56.5% of the participants qualified for free or reduced lunch, 29.1% for paid lunch, and lunch status was not available for 14.4% of the participants. According to the parent report, 106 (35.5%) participants were Black/African American, 167 (55.9%) were White, one (0.3%) was Asian, one (0.3%) was American Indian, 11 (3.7%) were two or more races, and 13 (4.3%) did not provide race information.

Measures

DELV-S

All participants completed the DELV-S, which has two parts. Part I, Language Variation Status, is intended to provide an estimate of the extent to which the child's dialect varies from MAE. Items in Part I assess phoneme production and morphosyntactic structures within sentence contexts. Raw scores from Part I are used to derive one of three classifications of language variation status: strong variation from MAE, some variation from MAE, and MAE. We also derived a variable for percentage of NMAE feature use by dividing the number of NMAE responses by the total number of scoreable responses (see N. P. Terry & Connor, 2012). Part II, Diagnostic Risk Status, measures morphosyntactic structures, understanding of complex *wh*-questions,

and nonword repetition to assess a child's risk for language disorder. The selected structures included noncontrastive features—that is, features that are stable across dialects but for which production accuracy may vary across speakers with different LI status. Raw scores from Part II are summed to generate the Diagnostic Error score (range = 0–23), with higher scores indicating a higher risk for LI. The Diagnostic Error score is then used to derive one of four classifications of diagnostic risk: highest risk, medium-high risk, low-medium risk, and lowest risk. The DELV-S manual reports reliability in terms of interexaminer decision consistency. For Part I, 92% of participants were classified exactly the same or within one category on both screening occasions. For Part II, 84% of participants were classified exactly the same or within one category on both screening occasions.

CELF-4

Children completed four subtests to derive the Core Language standard score from the CELF-4. These were Concepts and Following Directions, which measures the child's ability to understand and follow verbal instructions of increasing length and complexity; Word Structure, which assesses morphology and pronoun use; Recalling Sentences, which assesses the child's ability to listen to spoken sentences of increasing length and complexity and repeat them aloud verbatim; and Formulated Sentences, which assesses the child's ability to generate a spoken sentence from a given word. A few children ($n = 10$) in the sample were 9 years of age. These children did not complete the Word Structure subtest but instead completed the Word Classes subtest, which assesses a child's ability to understand and explain semantic relationships between words. According to the test manual, the internal consistency reliability of the Core Language standard score ranges from .94 to .95 for the age groups represented in this study (7;3–9;4).

Because of our interest in examining outcomes from modified versus unmodified scoring, all participants were first scored according to the guidelines provided in the CELF-4 manual for MAE speakers. Then participants who spoke AAE were rescored according to the guidelines for scoring modifications provided by the CELF-4 manual and with reference to AAE features listed by Washington and Craig (1994) and Oetting and McDonald (2001). A participant was considered to speak AAE if (i) the child spoke a dialect with some or strong variation from MAE on the basis of the DELV-S Part I, and (ii) parents reported that the child was Black/African American on the demographic questionnaire. Note that 33 White participants and 10 participants of other racial/ethnic groups (including not reported) met the DELV-S criteria for some or strong variation from MAE, but these participants were not included in modified-scoring analyses. Modified scoring involved rescoring three subtests: Word Structure, Recalling Sentences, and Formulated Sentences. No scoring modifications were applied to the Concepts and Following Directions or Word Classes subtests, because these tasks involve receptive language and semantic relationships that should not be affected by speaking an NMAE dialect. During administration,

assessors continued to administer items until they were confident that a ceiling would have been achieved according to both modified and unmodified scoring systems. However, during off-line scoring it was determined that several participants ($n = 6$) failed to achieve a modified-scoring ceiling, and these participants were removed from the modified-scoring analysis. Thus the analyses comparing modified versus unmodified scoring contain 77 participants who spoke AAE and 173 who spoke MAE.

TONI-4

The Test of Nonverbal Intelligence–Fourth Edition (TONI-4; Brown, Sherbenou, & Johnsen, 2010), was used to assess participants' nonverbal cognitive skills. Although nonverbal cognition factors into the inclusionary criteria for many studies of LI, in this study the TONI-4 was used for purely descriptive purposes. No decisions for participant inclusion or exclusion were made on the basis of this test, but the mean nonverbal IQ score was near the population mean ($M = 102.02$; $SD = 10.04$; range = 75–155). Twelve participants (4%) earned scores more than 1 SD below the mean on the TONI-4. Of these participants, six were Black/African American and six were White. These 12 participants demonstrated a range of NMAE feature use: Five spoke MAE, five spoke a dialect with strong variation from MAE, and two spoke a dialect with some variation from MAE. Five of these children met the criteria for medium-high risk of LI, and seven met the criteria for highest risk.

Scoring Reliability

Data were collected over two academic years, but each participant completed testing in only one year. In Year 1 (2013–2014 academic year), the DELV-S, TONI-4, and unmodified CELF-4 were double-scored by trained research assistants to ensure reliability, and all disagreements were reconciled through discussion. In Year 2 (2014–2015 academic year), all scorers were required to pass a scoring test before they were allowed to score protocols independently. Scorers kept a log of tests that they had scored, and a random sample of at least 20% of each scorer's list was double-scored to assess reliability. Reliability scorers for all individually administered assessments used blank protocols and video/audio recordings of the assessments and were unaware of initial scores. Reliability was assessed as by-item agreement for each of the individually administered measures and was 92% for DELV-S, 99.3% for the TONI-4, and 92.8% for the unmodified CELF-4. The rescoring of the CELF-4 subtests to apply scoring modifications was completed in fall 2016 by the first author and three trained research assistants. Similar to the unmodified scoring, 20% of the protocols for each subtest were double-scored to assess reliability. By-item agreement was 95% for Word Structure, 93% for Recalling Sentences, and 94% for Formulated Sentences.

Results

Preliminary analyses examined the distribution of performance for the measures of language variation and

language ability in our participant sample. Across all participants, there was a range of NMAE feature use ($M = 27.7\%$; $SD = 26.2\%$; range = 0%–100%) as measured by the DELV-S Part I. Of the 299 participants, on Part I of the DELV-S, 173 (57.9%) met the criteria for MAE, 58 (19.4%) met the criteria for some variation from MAE, and 68 (22.7%) met the criteria for strong variation from MAE. Of the 106 Black/African American participants, 51 (48.1%) met the criteria for strong variation, 30 (28.3%) met the criteria for some variation, and 25 (23.6%) met the criteria for MAE. In contrast, of the 167 White participants, 10 (6%) met the criteria for strong variation, 23 (13.8%) met the criteria for some variation, and 134 (80.2%) met the criteria for MAE.

Participants also exhibited a range of language ability as measured by the DELV-S Part II. Eighty participants (26.8%) met the criteria for highest risk for LI, 101 (33.8%) met the criteria for medium-high risk, 48 (16.1%) met the criteria for low-medium risk, and 70 (23.4%) met the criteria for lowest risk. Of the 106 Black/African American participants, 38 (35.8%) met the criteria for highest risk, 38 (35.8%) met the criteria for medium-high risk, 12 (11.3%) met the criteria for low-medium risk, and 18 (17%) met the criteria for lowest risk. Of the 167 White participants, 35 (21.0 %) met the criteria for highest risk, 52 (31.1%) met the criteria for medium-high risk, 31 (18.6%) met the criteria for low-medium risk, and 49 (29.3%) met the criteria for lowest risk. Thus, both Black/African American and White participants were represented across all risk categories for the dialect-neutral screen of language ability.

The mean unmodified CELF-4 standard score for the full sample was 88.80 ($SD = 14.82$; range = 46–124). Table 1 presents the unmodified and modified scaled scores of the CELF-4 and subtests for children who spoke AAE ($n = 77$) and MAE ($n = 173$). Recall that participants who spoke an NMAE dialect but were not African American were excluded from modified-scoring analyses because it was not possible to confirm what dialect they spoke and which scoring modifications should be applied. For children who spoke AAE, the modified CELF-4 score ($M = 85.22$; $SD = 11.78$; range = 50–111), was approximately

6 standard-score points higher on average than the unmodified CELF-4 score ($M = 79.29$; $SD = 11.71$; range = 46–108). Scoring modifications had a greater impact on Word Structure than other subtests, with modified scores an average of 2.73 scaled-score points higher than the unmodified scores. Modified scores for the Recalling Sentences and Formulated Sentences subtests were on average less than 1 scaled-score point higher than the unmodified scores. Overall, these descriptive data show a range of performance on measures of language variation and language ability across the full sample, as well as the subgroups of children speaking AAE or MAE.

Correlations

Although the measures of interest were well distributed across the full sample, visual inspection of the data indicated that NMAE feature use and lunch status (a proxy for SES) were distributed differently across racial groups. To be specific, the distribution of NMAE feature use was positively skewed for White participants but less so for African American participants. Also, White participants showed similar proportions of paid versus free or reduced lunch, whereas nearly all of the African American participants received free or reduced lunch (see Figures 1 and 2). Our first research objective was to examine the association between the frequency of NMAE feature use and performance on measures of language ability with and without contrastive features and scoring modifications. Given these observed patterns, we examined patterns of association separately for African American and White participants. Small sample sizes (1–13 participants) precluded separate analysis of other racial groups.

Because previous studies have found that SES is associated with NMAE feature use and with measures of language ability (see N. P. Terry, Connor, Thomas-Tate, & Love, 2010), we first examined point-biserial correlations between school-lunch status (as a proxy for SES) and our measures of language variation (i.e., percentage of NMAE feature use on Part I of the DELV-S) and language ability (i.e., Diagnostic Error score on Part II of the

Table 1. Descriptive statistics for modified and unmodified standard scores on the CELF-4 for participants who spoke MAE ($n = 173$) and AAE ($n = 77$).

Measure	Unmodified MAE ^a	Unmodified AAE ^b <i>M</i> (<i>SD</i>)	Modified AAE <i>M</i> (<i>SD</i>)
CELF-4	95.18 (13.08)	79.29 (11.71)	85.22 (11.78)
Concepts and Following Directions	10.21 (5.78)	6.91 (2.65)	
Word Classes ^c	n/a	5.33 (2.42)	
Word Structure ^d	8.52 (4.06)	4.90 (2.46)	7.63 (2.68)
Recalling Sentences	10.17 (7.97)	7.06 (2.50)	7.94 (2.45)
Formulated Sentences	10.67 (4.96)	7.18 (2.54)	7.84 (2.41)

Note. CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; MAE = mainstream American English; AAE = African American English; DELV-S = Diagnostic Evaluation of Language Variation–Screening Test.

^aIncludes children of all races who met the criteria for MAE on Part I of the DELV-S. ^bAAE includes children who were African American and met the criteria for some or strong variation on Part I of the DELV-S. ^cAdministered for 9-year-old participants (AAE: $n = 6$; MAE: $n = 0$).

^dAdministered for 7- and 8-year-old participants (AAE: $n = 71$; MAE: $n = 173$).

Figure 1. Percentage use of nonmainstream American English (NMAE) and Diagnostic Evaluation of Language Variation–Screening Test (DELV-S) Diagnostic Error score, by race. Higher Diagnostic Error scores represent higher risk for language impairment.

DELV-S and the CELF-4 standard score). Within the African American sample, no significant associations were detected between lunch status and measures of NMAE feature use ($r = .07$, $n = 89$, $p = .542$) or DELV-S Diagnostic

Error score ($r = -.08$, $n = 89$, $p = .487$). A small but significant correlation was observed between school-lunch status and unmodified CELF-4 scores ($r = -.25$, $n = 89$, $p = .021$), whereas the correlation between lunch status

Figure 2. Percentage use of nonmainstream American English (NMAE) and unmodified Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4) scores, by race. Lower standard scores represent lower language ability.

and modified CELF-4 score did not reach significance ($r = -.19, n = 86, p = .79$). Within the White sample, the association between lunch status and NMAE feature use did not reach significance ($r = .16, n = 144, p = .062$). The correlation between lunch status and DELV-S Diagnostic Error score was also nonsignificant ($r = .12, n = 144, p = .147$). The correlation between lunch status and unmodified CELF-4 score was small but significant ($r = -.27, n = 144, p < .001$).

Overall, the relationships between lunch status and measures of language variation and language ability were small in this study. However, because some were statistically significant, to address our first research objective we computed partial correlations between language ability and NMAE feature use, controlling for lunch status. Partial correlations were examined separately for White and African American children. Figures 1 and 2 present scatter plots of the percentage of NMAE feature use and two measures of language ability: the Diagnostic Error score from Part II of the DELV-S and the unmodified CELF-4 score. African American and White participants are plotted separately, and markers within the scatter plots indicate school-lunch status (as a proxy for SES). For the African American participants, partial correlations controlling for lunch status revealed significant, moderate associations between percentage of NMAE feature use and measures of language ability (DELV-S: $r = .46, p < .001$; unmodified CELF-4: $r = -.46, p < .001$; modified CELF-4: $r = -.341, p < .001$). Note that the difference in the direction of the correlations for measures of language ability is due to the fact that higher DELV-S Diagnostic Error scores represent more errors, whereas higher scores on the CELF-4 indicate higher language ability. For the White participants, partial correlations controlling for lunch status revealed significant, moderate associations between percentage of NMAE feature use and measures of language ability (DELV-S: $r = .51, p < .001$; unmodified CELF-4: $r = -.43, p < .001$).

Overall, these results indicate that there was a significant association between NMAE feature use and language-ability scores, the associations were of similar magnitudes for both DELV-S and unmodified CELF-4 standard scores, and this was true for both African American and White children. For African American children who spoke AAE, the use of modified scoring resulted in less variability in their CELF-4 standard scores, and a reduced correlation between frequency of NMAE feature use and modified CELF-4 scores. Despite significant correlations, considerable variability in language ability scores remained throughout most of the distribution of NMAE feature use. Overall, correlational analyses are useful for examining patterns across the range of ability, but they provide little information regarding diagnostic accuracy.

Classification Accuracy Statistics

Our second study objective was to investigate the impact of scoring modifications on diagnostic classifications, using classification-accuracy statistics. In the examination of classification accuracy, an index measure (the measure

being evaluated) is compared to a gold standard or reference standard, which is the best available evidence regarding the presence or absence of the condition (see Dollaghan, 2007). The diagnostic accuracy of the index measure can then be described in terms of how well it can accurately identify both individuals who have an impairment as having the impairment (true positives) and those who do not have the impairment as not having the impairment (true negatives). Diagnostic errors include erroneously classifying individuals without impairment as having impairment (false positives) and failing to identify individuals with impairment as having impairment (false negatives).

It is important to note that this study is a post hoc analysis of an existing data set, which did not contain a gold-standard assessment of LI for children who spoke an NMAE dialect, such as the DELV-NR. Rather, for the purposes of this analysis we treated the risk status of the DELV-S as the reference standard or “true” status for each child. To be specific, children who were classified as highest risk on the DELV-S were considered true positives (having LI), whereas children who were classified as lowest risk, low-medium risk, and medium-high risk were considered true negatives (having typical language skills). This decision was made in order to have the highest confidence that individuals identified by the DELV-S were true cases of children with LI; however, the overall pattern of results is similar when the highest risk and medium-high risk categories are treated as the reference measures of LI (see Appendix A). We then applied a cut score of 85 for the unmodified and modified CELF-4 standard scores. This cut score was selected because the test manual indicates that it resulted in 100% sensitivity and 82% specificity of classification for individuals represented by the normative sample (see Appendix B for results when a cut score of 77, or 1.5 *SD* below the mean, was applied). Tables 2 and 3 present the number of children with modified and unmodified CELF-4 scores above or below an 85 standard-score cutoff split by their level of dialect variation and risk status.

Applying this cutoff, 125 of 299 participants in the total sample (41.8%) would be considered to have LI, and 174 (58.2%) would be considered to have typical language skills on the basis of their unmodified CELF-4 scores. Of the 173 participants who spoke MAE, 42 (24%) met the criteria for LI. Of the 77 children who spoke AAE, 52 (66%) met the criteria for LI with unmodified CELF-4 scores, and 37 (48%) met the criteria for LI with modified CELF-4 scores. Thus, the application of scoring modifications reduced the proportion of participants who spoke AAE who met the criteria for LI on the CELF-4. Table 4 presents the classification statistics for the CELF-4 scores with unmodified scoring for all participants ($N = 299$), unmodified scoring for children who spoke AAE ($n = 77$), and modified scoring for children who spoke AAE ($n = 77$).

We first considered how well each of the CELF-4 scoring systems (modified vs. unmodified) accurately classified participants as having LI (true positives) or typical language (true negatives). The true positive rate, or sensitivity, for the unmodified scoring was 76% for all participants

Table 2. Unmodified scoring results split by language variation and risk of LI for the full sample.

DELV-S variation status	CELF-4 unmodified impairment status	DELV-S risk status				
		Lowest	Low-medium	Medium-high	Highest	Total
Strong variation	TL	2	2	10	4	18
	LI	1	2	13	34	50
Some variation	TL	8	6	8	3	25
	LI	1	4	12	16	33
MAE	TL	52	27	40	12	131
	LI	6	7	18	11	42
Total		70	48	101	80	299

Note. DELV-S = Diagnostic Evaluation of Language Variation–Screening Test; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; TL = typical language; LI = language impairment; MAE = mainstream American English.

and 88% when only participants who spoke AAE were considered. The modified scoring system resulted in relatively lower sensitivity (63%) for participants who spoke AAE. The true negative rate, or specificity, was 71% with unmodified scoring for all participants and 48% for unmodified scoring for children who spoke AAE. The modified scoring system resulted in relatively higher specificity (63%) for participants who spoke AAE.

We then considered the number of participants who were potentially misidentified by each of the scoring systems. The false positive and false negative rates were similar for the unmodified scoring for all participants (false positive: 29%; false negative: 24%).¹ For the participants who spoke AAE, the false positive rate decreased when scoring modifications were applied (unmodified: 52%; modified: 36%), whereas the false negative rate increased when scoring modifications were applied (unmodified: 12%; modified: 36%).

We also calculated positive and negative likelihood ratios for each scoring system. Likelihood ratios consider sensitivity and specificity simultaneously, and are useful when examining diagnostic classifications for conditions, such as LI, which have low incidence in the general population. The positive likelihood ratio (LR+) reflects the probability that an individual who is identified as having impairment in fact does, whereas the negative likelihood ratio (LR–) reflects the probability that an individual who is classified as not having impairment in fact does not. LR+ and LR– values of 1 indicate that the test gives no reliable information about impairment status, whereas higher LR+ values increase confidence in identifying true cases of impairment and lower LR– values increase confidence in classifications of typical language abilities. Positive and negative likelihood ratios are on different scales. For LR+, values greater than 10 are considered optimal and values greater than 3 are considered moderately accurate. For LR–, values less than .1 are considered optimal and values less than .3 are considered moderately accurate (Dollaghan, 2007).

¹None of the participants who were potential false positives had TONI-4 scores more than 1 *SD* below the mean.

The LR+ for unmodified scoring was 2.61 for the full sample and 1.68 for participants who spoke AAE. Using modified scoring resulted in a small LR+ improvement for participants who spoke AAE (1.75), suggesting that using scoring modifications with this subgroup resulted in a slight improvement in the probability that children who were identified as having LI on the basis of CELF-4 scores below 85 actually had LI according to the criterion of DELV-S highest risk status. The LR– for unmodified scoring was .33 for all participants and .25 for participants who spoke AAE. Using scoring modifications resulted in an LR– increase (.57), indicating that use of modified scoring resulted in a moderate increase in the probability that true cases of LI were missed.

Discussion

This study had two main objectives. The first research objective was to compare the relationship between frequency of NMAE feature use and two measures of language ability: the DELV-S, which was designed for children who speak NMAE dialects (primarily AAE), and the CELF-4, which was not specifically designed for children who speak NMAE dialects. The CELF-4 includes items that test features that contrast between MAE and many NMAE dialects, whereas the DELV-S uses features that do not contrast between MAE and NMAE (primarily AAE). In a diverse sample of second-grade students, we found that use of NMAE features showed a significant moderate correlation with both language ability measures, after controlling for SES. This finding held true in both White and African American participants, across both language ability measures, and in the African American children with both modified and unmodified scoring systems for the CELF-4.

The second research objective was to examine the agreement between the two assessments when scoring modifications for dialect variation were and were not applied to the diagnostic test. Results indicated that when scoring modifications were not applied, the LR+ was lower for children who spoke AAE than for the overall sample, which included speakers with a range of NMAE feature use. This underscores the difficulty of accurately assessing LI

Table 3. Modified scoring results split by language variation and CELF-4 LI status for children who spoke AAE and MAE.

DELV-S variation status	CELF-4 modified impairment status	DELV-S risk status				
		Lowest	Low-medium	Medium-high	Highest	Total
Strong variation (AAE)	TL	2	2	12	8	24
	LI	1	1	6	16	24
Some variation (AAE)	TL	5	2	5	4	16
	LI	0	2	6	5	13
MAE ^a	TL	52	27	40	12	131
	LI	6	7	18	11	42
Total		66	41	87	56	250

Note. CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; LI = language impairment; AAE = African American English; MAE = mainstream American English; DELV-S = Diagnostic Evaluation of Language Variation–Screening Test; TL = typical language.

^aScores were not modified for children who spoke MAE.

in children who speak NMAE when using a measure that assesses contrastive as well as noncontrastive features. Results further showed that diagnostic classification accuracy varied depending on whether modified or unmodified scoring was used. Whereas scoring modifications led to a slight reduction in the false positive rate for participants who spoke AAE, they led to a larger corresponding increase in the false negative rate. When likelihood ratios were examined, the negative effect on LR– was greater than the positive effect on LR+. Here we must bring a strong note of caution to the interpretation of these results—the results can only be generalized to the extent that the reference standard is accurate. Our study is limited by the fact that the reference standard was a screen, not a norm-referenced assessment. To increase confidence in the LI determination from the DELV-S, we considered only the children at highest risk as having LI. As shown in Appendix A, when DELV-S medium-high and highest risk groups were combined as the reference standard for LI, the performance of the unmodified CELF-4 with children who speak AAE yielded a somewhat improved LR+ and a slightly worsened LR–. Yet scoring modifications still produced worse performance overall than unmodified scoring.

Scoring modifications are intended to address the issue of potential overidentification of children who speak NMAE dialects when using tests developed for MAE speakers. However, as reviewed in the introduction, the

application of scoring modifications is not straightforward. Our results suggest that when considering whether or not to apply scoring modifications, clinicians should weigh the risk of false positives (erroneously classifying children who speak NMAE dialects as having LI) as well as the risk of false negatives (classifying children who speak NMAE dialects as having typical language when they indeed do have LI). Both errors are problematic. False positives may result in misallocation of intervention resources and potentially negative socioemotional consequences for children who are misdiagnosed. In contrast, false negatives mean that intervention services will not be supplied to children who need them. Other studies are necessary to evaluate the costs and benefits of potential false positives and false negatives.

The current study supports other research which has raised concern that the universal application of scoring modifications for all items and all children who speak NMAE dialects may lead to an increase in false negatives, and thus underidentification of LI. Therefore, there is a need for further research into strategic application of scoring modifications to maximize sensitivity and specificity. For example, Oetting, Gregory, and Rivière (2016) demonstrated that a sentence repetition task yielded high sensitivity and specificity when scoring modifications were limited to three features that are highly frequent in NMAE dialects and have not been demonstrated to be effective clinical markers of LI in those same dialects. On the other hand, their results

Table 4. Classification statistics for modified and unmodified scoring of the CELF-4 using the highest risk category on the DELV-S as the reference measure.

Statistic	Unmodified scores for all participants (N = 299)	Unmodified scores for AAE-speaking participants (n = 77)	Modified scores for AAE-speaking participants (n = 77)
True positive (sensitivity)	0.76	0.88	0.64
True negative (specificity)	0.71	0.48	0.64
False positive	0.29	0.52	0.36
False negative	0.24	0.12	0.36
LR+ (sensitivity/1-specificity)	2.61	1.68	1.75
LR– (1-sensitivity/specificity)	0.34	0.25	0.57

Note. CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; DELV-S = Diagnostic Evaluation of Language Variation–Screening Test; AAE = African American English; LR+ = positive likelihood ratio; LR– = negative likelihood ratio.

showed that similar diagnostic accuracy could be achieved in the same task by scoring exact repetition accuracy and not applying scoring modifications. Future work should explore when and how these types of dialect-strategic scoring modifications can be extended to widely used language assessments, such as the CELF-4 or CELF-5.

In this study, we observed that modified scoring procedures affected scaled scores to different degrees across the different subtests of the CELF-4. In the introduction, we highlighted that the number of items potentially affected by scoring modifications could depend on the type of task, and we noted that tasks that allow children to produce their own sentences could theoretically have as many as 100% of items affected by scoring modifications. However, our results indicated that the Formulated Sentences subtest was the least affected by scoring modifications, with modified scores an average of 0.66 scaled-score point higher than unmodified scores. The Recalling Sentences subtest was affected to a similar degree, whereas the Word Structure subtest was strongly affected by scoring modifications. This result may not be surprising, because the Word Structure subtest elicits specific morphosyntactic forms, some of which differ across dialects of American English. Thus, for this subtest, participants were required to produce forms that may have been produced variably in their dialect. Future research could explore whether diagnostic decisions that are based on the other subtests are more accurate than those that include the Word Structure subtest, with or without scoring modifications.

Diagnostic decision making involves many factors and should not be limited to scores on standardized assessments. Prior to conducting a standardized assessment, clinicians should collect a detailed case history, including information from parents about the child's language understanding and production at home and information from teachers about academic performance. A detailed case history can help clinicians determine the extent to which children in a standardized test's norming sample are representative of a particular child they are assessing. Ideally, clinicians should use standardized assessments with a norming sample that is representative of the client's cultural and linguistic background and also shows high diagnostic accuracy for that population. In cases where such tests are not available, standardized tests can be administered for descriptive purposes (e.g., to examine knowledge of particular linguistic structures without deriving standard scores), alongside other informal language-assessment techniques such as language sampling and dynamic assessment. The results of this study underscore the importance of considering multiple sources of information, which can provide complementary information and a more complete picture of a child's language ability, when making decisions about diagnosis and treatment.

The current study focused on second-grade students, but students' use of NMAE features changes over time (N. P. Terry et al., 2012; N. P. Terry & Connor, 2012). Although there are individual differences in children's use of NMAE dialect features (Van Hofwegen & Wolfram,

2010), the general trend is that use of NMAE features drops during the first few years of elementary school as students master code-switching strategies, and then increases during adolescence as students begin using NMAE dialects for more social reasons (N. P. Terry et al., 2010; Van Hofwegen & Wolfram, 2010). An important question for future research is how these changes influence diagnostic decisions and errors at different points across development.

In our second-grade sample, we observed a moderate association between use of NMAE features and language ability across the DELV-S Part II and both the unmodified and modified CELF-4 scores. This is in line with a growing body of research that has demonstrated a link between children's use of NMAE features in certain contexts and language and literacy outcomes. For example, Oetting and McDonald (2001) compared the frequency of use of NMAE and documented that children with LI use NMAE features more frequently than age- or language-matched peers with typical development. Jackson and Pearson (2010) showed that children's use of contrastive features decreases during elementary school, and that exclusive use of AAE forms in a formal testing environment may suggest that a child has LI. Recent studies by Connor, N. P. Terry, and colleagues have shown that children who reduce their frequency of NMAE feature use during second grade perform better on literacy tasks at the end of second grade than children who do not (Connor & Craig, 2016; N. P. Terry et al., 2012; N. P. Terry & Connor, 2012). N. P. Terry et al. (2012) have argued that these lower literacy outcomes may be attributed to a difficulty mastering code switching—the ability to switch dialects in different social contexts. Thus, the findings of the present study add to a growing body of research showing that under certain circumstances, NMAE dialect feature use may provide some information about language and literacy abilities. More research is needed to improve the use of this information in clinical decision making.

The use of an existing data set limited this study in several ways. First, as previously discussed, we did not have a gold-standard measure of language ability designed for children who spoke NMAE dialects, such as the DELV-NR. Second, although the overall sample was fairly large and there was a range of language ability and NMAE use in the available sample, the number of children who spoke AAE—particularly those who were at highest risk of LI according to the DELV-S—was relatively small, albeit comparable with the DELV-S standardization samples for 7- and 8-year-olds. Thus, the results demonstrate the need for future studies designed specifically to investigate the impact of scoring systems on diagnostic accuracy.

Conclusion

In sum, the results of this study indicate that there is a moderate association between second-grade students' language ability and their use of NMAE features in formal testing contexts, and that there is a risk of misdiagnosis of children who speak AAE both when scoring modifications

are and are not applied to measures of language ability that include structures that contrast between MAE and NMAE dialects. We emphasize that the association between language ability and NMAE use in this study should not be interpreted as an indication that NMAE feature use is related to LI in all circumstances. To date, much of the research has focused on the problem of over-identification of LI in children who speak NMAE dialects, and our results suggest that clinicians should remain aware of this risk. However, our results also support recent research showing that underidentification of LI is problematic as well. The current study highlights the urgent need for language assessments which can be used for all children regardless of language variation and which provide both high sensitivity and high specificity.

Acknowledgments

This research was supported in part by National Institutes of Health Grant R03DC013399, awarded to Suzanne Adlof. We are grateful to the participants, teachers, and schools who were involved in this study. We thank research assistants from the South Carolina Research on Language & Literacy Lab at the University of South Carolina for their help with data collection and processing, and Allison Brazendale and Joanna Scoggins for serving as project coordinators.

References

- Adlof, S. M., Scoggins, J., Brazendale, A., Babb, S., & Petscher, Y. (2017). *Identifying children at risk for language impairment or dyslexia with group-administered classroom measures*. Manuscript submitted for publication.
- American Speech-Language-Hearing Association. (2003). *American English dialects* [Technical report]. Retrieved from <http://www.asha.org/policy>
- Arias, G., & Friberg, J. (2017). Bilingual language assessment: Contemporary versus recommended practice in American schools. *Language, Speech, and Hearing Services in Schools, 48*, 1–15. https://doi.org/10.1044/2016_LSHSS-15-0090
- Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44*, 133–146. [https://doi.org/10.1044/0161-1461\(2012/12-0093\)](https://doi.org/10.1044/0161-1461(2012/12-0093))
- Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2010). *Test of Non-verbal Intelligence—Fourth Edition*. Austin, TX: Pro-Ed.
- 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*, 519–525. <https://doi.org/10.1044/jslhr.4003.519>
- Champion, T. B. (2003). *Understanding storytelling among African American children: A journey from Africa to America*. Mahwah, NJ: Erlbaum.
- Connor, C. M., & Craig, H. K. (2016). African American preschoolers' language, emergent literacy skills, and use of African American English: A complex relation. *Journal of Speech, Language, and Hearing Research, 49*, 771–792. [https://doi.org/10.1044/1092-4388\(2006/055\)](https://doi.org/10.1044/1092-4388(2006/055))
- Craig, H. K., & Washington, J. A. (2006). *Malik goes to school: Examining the language skills of African American students from preschool–5th grade*. Mahwah, NJ: Erlbaum.
- Dollaghan, C. A. (2007). *The handbook for 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*, 1136–1146. <https://doi.org/10.1044/jslhr.4105.1136>
- Dollaghan, C. A., Campbell, T. F., Paradise, J. L., Feldman, H. M., Janosky, J. E., Pitcairn, D. N., & Kurs-Lasky, M. (1999). Maternal education and measures of early speech and language. *Journal of Speech, Language, and Hearing Research, 42*, 1432–1443. <https://doi.org/10.1044/jslhr.4206.1432>
- Garrity, A. W., & Oetting, J. B. (2010). Auxiliary BE production by African American English-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 53*, 1307–1320. [https://doi.org/10.1044/1092-4388\(2010/09-0016\)](https://doi.org/10.1044/1092-4388(2010/09-0016))
- Green, L. J. (2002). *African American English: A linguistic introduction*. Cambridge, United Kingdom: Cambridge University Press.
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.
- Hoff, E. (2003). The specificity of environmental influence: Socio-economic status affects early vocabulary development via maternal speech. *Child Development, 74*, 1368–1378. <https://doi.org/10.1111/1467-8624.00612>
- Horton, 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*, 448–460. https://doi.org/10.1044/2014_AJSLP-13-0028
- Horton-Ikard, R., & Miller, J. F. (2004). It is not just the poor kids: The use of AAE forms by African-American school-aged children from middle SES communities. *Journal of Communication Disorders, 37*, 467–487. <https://doi.org/10.1016/j.jcomdis.2004.02.001>
- Jackson, J. E., & Pearson, B. Z. (2010). Variable use of features associated with African American English by typically developing children. *Topics in Language Disorders, 30*(2), 135–144. <http://doi.org/10.1097/TLD.0b013e3181e03ff6>
- Labov, W. (1966). The effect of social mobility on linguistic behavior. *Sociological Inquiry, 36*, 186–203. <https://doi.org/10.1111/j.1475-682X.1966.tb00624.x>
- Labov, W. (1989). The child as linguistic historian. *Language Variation and Change, 1*, 85–97. <https://doi.org/10.1017/S0954394500000120>
- Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change, 2*, 205–254. <https://doi.org/10.1017/S0954394500000338>
- Lee, R., & Oetting, J. B. (2014). Zero marking of past tense in child African American English. *SIG 1 Perspectives on Language Learning and Education, 21*, 173–181. <https://doi.org/10.1044/1le21.4.173>
- Miller, K. (2013). Variable input: What Sarah reveals about non-agreeing *don't* and theories of root infinitives. *Language Acquisition, 20*, 305–324. <https://doi.org/10.1080/10489223.2013.828061>
- Mills, M. T., Watkins, R. V., & Washington, J. A. (2013). Structural and dialectal characteristics of the fictional and personal narratives of school-age African American children. *Language, Speech, and Hearing Services in Schools, 44*, 211–223. [https://doi.org/10.1044/0161-1461\(2012/12-0021\)](https://doi.org/10.1044/0161-1461(2012/12-0021))
- Morgan, M. (1998). More than a mood or an attitude: Discourse and verbal genres in African-American culture. In S. S. Mufwene,

- J. R. Rickford, G. Bailey, & J. Baugh (Eds.), *African-American English: Structure, history and use* (pp. 251–281). London, United Kingdom: Routledge.
- Morgan, P. L., Farkas, G., Cook, M., Strassfeld, N. M., Hillemeier, M. M., Pun, W. H., & Schussler, D. L. (2017). Are Black children disproportionately overrepresented in special education? A best-evidence synthesis. *Exceptional Children*, 83, 181–198. <https://doi.org/10.1177/0014402916664042>
- 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, 682–696. <https://doi.org/10.3109/02699206.2014.882990>
- Oetting, J. B., & Cleveland, L. H. (2006). The clinical utility of nonword repetition for children living in the rural south of the US. *Clinical Linguistics & Phonetics*, 20(7–8), 553–561. <http://doi.org/10.1080/02699200500266455>
- Oetting, J. B., & Garrity, A. W. (2006). Variation within dialects: A case of Cajun/Creole influence within child SAAE and SWE. *Journal of Speech, Language, and Hearing Research*, 49, 16–26. [https://doi.org/10.1044/1092-4388\(2006/002\)](https://doi.org/10.1044/1092-4388(2006/002))
- 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. <https://doi.org/10.1044/perspl.SIG16.28>
- Oetting, J. B., & Horohov, J. E. (1997). Past-tense marking by children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, 40, 62–74. <https://doi.org/10.1044/jslhr.4001.62>
- Oetting, J. B., Lee, R., & Porter, K. L. (2013). Evaluating the grammars of children who speak nonmainstream dialects of English. *Topics in Language Disorders*, 33, 140–151. <https://doi.org/10.1097/tld.0b013e31828f509f>
- Oetting, J. B., & McDonald, J. L. (2001). Nonmainstream dialect use and specific language impairment. *Journal of Speech, Language, and Hearing Research*, 44, 207–223. [https://doi.org/10.1044/1092-4388\(2001/018\)](https://doi.org/10.1044/1092-4388(2001/018))
- Oetting, J. B., & McDonald, J. L. (2002). Methods for characterizing participants' nonmainstream dialect use in child language research. *Journal of Speech, Language, and Hearing Research*, 45, 505–518. [https://doi.org/10.1044/1092-4388\(2002/040\)](https://doi.org/10.1044/1092-4388(2002/040))
- Oetting, J. B., McDonald, J. L., Seidel, C. M., & Hegarty, M. (2016). Sentence recall by children with SLI across two non-mainstream dialects of English. *Journal of Speech, Language, and Hearing Research*, 59, 183–194. https://doi.org/10.1044/2015_JSLHR-L-15-0036
- Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2016). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research*, 49, 1037–1057. [https://doi.org/10.1044/1092-4388\(2006/074\)](https://doi.org/10.1044/1092-4388(2006/074))
- 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, 2–15. [https://doi.org/10.1044/1092-4388\(2008/07-0176\)](https://doi.org/10.1044/1092-4388(2008/07-0176))
- Pruitt, S. L., Oetting, J. B., & Hegarty, M. (2011). Passive participle marking by African American English-speaking children reared in poverty. *Journal of Speech, Language, and Hearing Research*, 54, 598–607. [https://doi.org/10.1044/1092-4388\(2010/09-0134\)](https://doi.org/10.1044/1092-4388(2010/09-0134))
- Qi, C. H., Kaiser, A. P., Milan, S., & Hancock, T. (2006). Language performance of low-income African American and European American preschool children on the PPVT-III. *Language, Speech, and Hearing Services in Schools*, 37, 5–16. [https://doi.org/10.1044/0161-1461\(2006/002\)](https://doi.org/10.1044/0161-1461(2006/002))
- 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, 44–52. <https://doi.org/10.1044/1058-0360.0801.44>
- Roberts, J. L. (1994). *Acquisition of variable rules: (-t,d) deletion and (ing) production in preschool children*. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1095&context=ircs_reports University of Pennsylvania, Philadelphia.
- Semel, E., Wiig, E., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals—Fourth Edition*. Toronto, Ontario, Canada: Pearson.
- Seymour, H. N. (2004). The challenge of language assessment for African American English-speaking children: A historical perspective. *Seminars in Speech and Language*, 25, 3–12. <https://doi.org/10.1055/s-2004-824821>
- Seymour, H. N., Bland-Stewart, L., & Green, L. J. (1998). Difference versus deficit in child African American English. *Language, Speech, and Hearing Research*, 29, 96–108. <https://doi.org/10.1044/0161-1461.2902.96>
- Seymour, H. N., Roeper, T. W., deVilliers, J., & deVilliers, P. A. (2003). *Diagnostic Evaluation of Language Variation—Screening Test*. Bloomington, MN: Pearson.
- Seymour, H. N., Roeper, T., deVilliers, J., & deVilliers, P. A. (2005). *Diagnostic Evaluation of Language Variation—Norm Referenced*. Bloomington, MN: Pearson.
- Smith, J., Durham, M., & Fortune, L. (2007). “Mam, my trousers is fa’in doon!”: Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation and Change*, 19, 63–99. <https://doi.org/10.1017/S0954394507070044>
- Smith, J., Durham, M., & Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and /d deletion. *Language Variation and Change*, 21, 69–95. <https://doi.org/10.1017/S0954394509000039>
- Smitherman, G. (1977). *Talkin and testifyin: The language of Black America*. Boston, MA: Houghton Mifflin.
- Smitherman, G. (1998). Word from the hood: The lexicon of African-American vernacular English. In S. S. Mufwene, J. R. Rickford, G. Bailey, & J. Baugh (Eds.), *African American English: Structure, history, and use* (pp. 203–225). London, United Kingdom: Routledge.
- Stockman, I. J. (2000). The New Peabody Picture Vocabulary Test—III: An illusion of unbiased assessment? *Language, Speech, and Hearing Services in Schools*, 31, 340–353. <https://doi.org/10.1044/0161-1461.3104.340>
- Stockman, I. J. (2010). A review of developmental and applied language research on African American children: From a deficit to difference perspective on dialect differences. *Language, Speech, and Hearing Services in Schools*, 41, 23–38. [https://doi.org/10.1044/0161-1461\(2009/08-0086\)](https://doi.org/10.1044/0161-1461(2009/08-0086))
- Terry, J. M., Jackson, S. C., Evangelou, E., & Smith, R. L. (2010). Expressive and receptive language effects of African American English on a sentence imitation task. *Topics in Language Disorders*, 30, 119–134. <https://doi.org/10.1097/TLD.0b013e3181e04148>
- Terry, N. P., & Connor, C. M. (2012). Changing nonmainstream American English use and early reading achievement from kindergarten to first grade. *American Journal of Speech-Language Pathology*, 21, 78–86. [https://doi.org/10.1044/1058-0360\(2011/10-0093\)](https://doi.org/10.1044/1058-0360(2011/10-0093))
- Terry, N. P., Connor, C. M., Petscher, Y., & Conlin, C. R. (2012). Dialect variation and reading: Is change in nonmainstream American English use related to reading achievement in first and second grades? *Journal of Speech, Language, and Hearing*

- Research*, 55, 55–69. [https://doi.org/10.1044/1092-4388\(2011/09-0257\)](https://doi.org/10.1044/1092-4388(2011/09-0257))
- 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, 126–145. [https://doi.org/10.1044/1092-4388\(2009/08-0058\)](https://doi.org/10.1044/1092-4388(2009/08-0058))
- Thomas-Tate, S., Washington, J., & Edwards, J.** (2004). Standardized assessment of phonological awareness skills in low-income African American first graders. *American Journal of Speech-Language Pathology*, 13, 182–190. [https://doi.org/10.1044/1058-0360\(2004/018\)](https://doi.org/10.1044/1058-0360(2004/018))
- Thompson, C. A., Craig, H. K., & Washington, J. A.** (2004). Variable production of African American English across oracy and literacy contexts. *Language, Speech, and Hearing Services in Schools*, 35, 269–282. [https://doi.org/10.1044/0161-1461\(2004/025\)](https://doi.org/10.1044/0161-1461(2004/025))
- Trudgill, P.** (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, 1, 179–195. <https://doi.org/10.1017/S0047404500000488>
- Van Hofwegen, J., & Wolfram, W.** (2010). Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*, 14, 427–455. <https://doi.org/10.1111/j.1467-9841.2010.00452.x>
- Washington, J. A., & Craig, H. K.** (1994). Dialectal forms during discourse of poor, urban, African American preschoolers. *Journal of Speech, Language, and Hearing Research*, 37, 816–823. <https://doi.org/10.1044/jshr.3704.816>
- Washington, J. A., & Craig, H. K.** (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test–III. *Language, Speech, and Hearing Services in Schools*, 30, 75–82. <https://doi.org/10.1044/0161-1461.3001.75>
- Washington, J. A., & Craig, H. K.** (2002). Morphosyntactic forms of African American English used by young children and their caregivers. *Applied Psycholinguistics*, 23, 209–231. <https://doi.org/10.1017/S0142716402002035>
- Washington, J. A., Craig, H. K., & Kushmaul, A. J.** (1998). Variable use of African American English across two language sampling contexts. *Journal of Speech, Language, and Hearing Research*, 41, 1115–1124. <https://doi.org/10.1044/jslhr.4105.1115>
- Wiig, E. H., Semel, E., & Secord, W. A.** (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition*. Bloomington, MN: Pearson.
- Wyatt, T.** (2012). Assessment of multicultural and international clients with communication disorders. In D. E. Battle (Ed.), *Communication disorders in multicultural and international populations* (4th ed., pp. 243–278). St. Louis, MO: Elsevier Mosby.

Appendix A

Classification statistics for modified and unmodified scoring of the CELF-4 using the highest risk and medium-high risk categories on the DELV-S as the reference measure.

Statistic	Unmodified scores for all participants (<i>N</i> = 299)	Unmodified scores for AAE-speaking participants (<i>n</i> = 77)	Modified scores for AAE-speaking participants (<i>n</i> = 77)
True positive (sensitivity)	.57	.76	.53
True negative (specificity)	.82	.67	.73
False positive	.18	.33	.27
False negative	.43	.24	.47
LR+ (sensitivity/1-specificity)	3.23	2.27	2.00
LR- (1-sensitivity/specificity)	.52	.36	.64

Note. CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; DELV-S = Diagnostic Evaluation of Language Variation–Screening Test; AAE = African American English; LR+ = positive likelihood ratio; LR- = negative likelihood ratio.

Appendix B

Unmodified scoring results split by language variation and CELF-4 LI status (1.5 SDs below the mean).

DELV-S variation status	CELF-4 impairment status ^a	DELV-S risk status				
		Lowest	Low-medium	Medium-high	Highest	Total
Strong variation	TL	2	4	16	12	34
	LI	1	0	7	26	34
Some variation	TL	9	7	18	11	45
	LI	0	3	2	8	13
MAE	TL	57	34	54	15	160
	LI	1	0	4	8	13
Total		70	48	101	80	299

Note. CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; LI = language impairment; DELV-S = Diagnostic Evaluation of Language Variation–Screening Test; TL = typical language; MAE = mainstream American English.

^aLI = CELF-4 ≤ 77; TL = CELF-4 > 77.