LSHSS

Research Article

# Classification Accuracy of Teacher Ratings When Screening Nonmainstream English-Speaking Kindergartners for Language Impairment in the Rural South

Kyomi D. Gregory[a] and Janna B. Oetting[b]

**Purpose:** We compared teacher ratings as measured by the Teacher Rating of Oral Language and Literacy (TROLL; Dickinson, McCabe, & Sprague, 2001, 2003) and Children's Communication Checklist–Second Edition (CCC-2; Bishop, 2006) to 2 established screeners, the Part II of the Diagnostic Evaluation of Language Variation–Screening Test (DELV-ST-II; Seymour, Roeper, & de Villiers, 2003) and Dynamic Indicators of Basic Early Literacy Skills–Next (DIBELS; Good, Gruba, & Kaminski, 2009), and then examined whether teacher ratings alone or when combined with the DELV-ST-II or DIBELS accurately classify nonmainstream English-speaking kindergartners by their clinical status.

**Method:** Data came from 98 children who lived in the rural South; 47 spoke African American English, and 51 spoke Southern White English. Using the syntax subtest of the Diagnostic Evaluation of Language Variation–Norm Referenced (Seymour, Roeper, & de Villiers, 2005) as the reference standard, 43 were language impaired and 55 were typically developing. Analyses included analysis of variance, correlations, and discriminant function with sensitivity and specificity indices.

**Results:** The TROLL, CCC-2, DELV-ST-II, and DIBELS showed clinical status but not dialect effects, and they correlated with each other, the Diagnostic Evaluation of Language Variation–Norm Referenced, and other language measures. Classification accuracies of all 4 tools were too low for screening purposes; however, empirically derived cut scores improved the results, and a discriminant function selected the TROLL and DELV-ST-II as optimal for determining who should be referred for an evaluation, with the TROLL yielding the highest level of sensitivity (77%).

**Conclusion:** Findings support teacher ratings as measured by the TROLL when screening nonmainstream English-speaking kindergartners for language impairment in the rural South, while also calling for additional development and study of teacher rating tools and other screening instruments.

**Supplemental Material:** https://doi.org/10.23641/asha.6007712

School-based speech-language pathologists often use screeners to determine if an evaluation is warranted, and teacher involvement is typically advocated as part of the screening process (Fujiki & Brinton, 1984; Marsh, Pane, & Hamilton, 2006; Whitworth, Davies, Stokes, & Blain, 1993). Teachers observe children in their classrooms individually and collectively during a wide range of academic and nonacademic activities, and teacher ratings are efficient in terms of time and cost. Teacher referral is also a strong predictor of eligibility for special education services. Gottlieb, Alter, Gottlieb, and Wishner (1994) reported from 10 years of work that 73%–90% of children referred by teachers were found eligible for special education services; however, these authors also noted that most of the referrals were for children performing academically at the lowest levels.

Many children with developmental language impairments (LIs) present receptive and expressive language deficits in the absence of cognitive, sensory, social, or developmental disabilities (Leonard, 2014). Without more conspicuous conditions, the language weaknesses of children with LI may go undetected by teachers, especially in the early school years. Indeed, in Tomblin et al.'s (1997) large epidemiological study, 71% of the kindergartners who presented with a test profile consistent with an LI diagnosis had never been referred to a speech-language pathologist (for similar concerns regarding the identification of child LI in primary care settings, see Wallace et al., 2015).

[a]Salus University, Elkins Park, PA
[b]Louisiana State University, Baton Rouge

Correspondence to Kyomi D. Gregory: kgregory@salus.edu

Editor: Shelley Gray

Received May 18, 2017
Revision received August 16, 2017
Accepted September 15, 2017

https://doi.org/10.1044/2017_LSHSS-17-0045

**Publisher Note:** This article is part of the Clinical Forum: Toward Accurate Identification of Developmental Language Disorder Within Linguistically Diverse Schools.

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

Rurality may further complicate the referral process for teachers, especially if their schools are located in the South, where rates of child poverty are higher than other regions of the United States. According to the U.S. Department of Agriculture (2017), one fourth of children in rural areas are poor compared with one fifth of urban children, and many of the 48 counties with child poverty rates above 50% are in the South. In addition, many nonmainstream dialects of English are spoken in the rural South, and teachers may be hesitant to refer children who speak these dialects in fear of misinterpreting a dialect difference as LI.

In a series of studies, Oetting and colleagues examined children's use of African American English (AAE) and Southern White English (SWE), two nonmainstream dialects of English that are spoken in rural Louisiana (Cleveland & Oetting, 2013; Oetting, 2015; Oetting & McDonald, 2002; Oetting & Newkirk, 2011; Roy, Oetting, & Moland, 2013). Repeatedly, these studies found child AAE and SWE to be perceptually and linguistically distinct, although these dialects shared many of the same nonmainstream patterns (e.g., zero copula and auxiliary BE, zero regular and irregular verbal *–s,* zero auxiliary DO, subject–verb agreement with BE and DO, multiple negation, and copula and auxiliary *ain't*). Child speakers of AAE also produced higher rates (i.e., densities) of nonmainstream patterns than did child speakers of SWE; however, within both dialects, individual variability existed, a finding that was first documented for child AAE by Washington and Craig (1994; see also Terry, Connor, Thomas-Tate, & Love, 2010).

In the current study and using an existing data set, teacher ratings were evaluated for screening purposes. Specifically, we asked if teacher ratings were valid measures of AAE- and SWE-speaking kindergartners' language abilities and whether teacher ratings could be used alone or in combination with one or more tools to determine who should be referred for a language evaluation. Screening every child in kindergarten by a speech-language pathologist is expensive. If valid, teacher ratings may offer a cost-effective alternative for determining who should be referred.

Teacher ratings were collected using two tools, the Teacher Rating of Oral Language and Literacy (TROLL; Dickinson, McCabe, & Sprague, 2001, 2003) and the Children's Communication Checklist–Second Edition (CCC-2; Bishop, 2006). Both tools were first compared with each other and two established screeners, Part II of the Diagnostic Evaluation of Language Variation–Screening Test (DELV-ST-II; Seymour, Roeper, & de Villiers, 2003) and the Dynamic Indicators of Basic Early Literacy Skills–Next (DIBELS; Good, Gruba, & Kaminski, 2009). Then, all four tools (TROLL, CCC-2, DELV-ST-II, and DIBELS) were evaluated for how well they classified the children's clinical status (LI vs. typically developing [TD]), which was based on the syntax subtest of the Diagnostic Evaluation of Language Variation–Norm Referenced (DELV-NR; Seymour, Roeper, & de Villiers, 2005). Within diagnostic accuracy studies, this analysis treated the four tools as index tests and the DELV-NR as the clinical reference standard (Bossuyt et al., 2015).

## Studies of Teacher Ratings

Multiple studies have examined teacher ratings of children's speech, language, and/or literacy abilities (Bates & Nettlebeck, 2001; Bedore, Pena, Joyner, & Macken, 2011; Botting, Conti-Ramsden, & Crutchley, 1997; Cabell, Justice, Zucker, & Kilday, 2009; Feinberg & Shapiro, 2009; Gijsel, Bosman, & Verhoeven, 2006; Gilmore & Vance, 2007; Gray et al., 2017; Hauerwas & Stone, 2000; Jessup, Ward, Cahill, & Keating, 2008; Martin & Shapiro, 2011; Pua, Lee, & Liow, 2017; Williams, 2006). These studies have collected data from teachers using semistructured interviews, short questionnaires created by researchers, unpublished questionnaires used within clinical practice, and published rating scales such as those examined here. Although findings from these studies support the use of teacher ratings, the level of support varies as a function of the analysis. The strongest support for teacher ratings comes from group comparison and correlational studies. For example, Cabell et al. (2009) examined teacher ratings using 12 items from the Clinical Evaluation of Language Fundamentals Preschool–Second Edition Pre-Literacy Rating Scale (Wiig, Secord, & Semel, 2004). Children rated by the teachers as at risk for reading impairment compared with those rated as not at risk scored lower on direct assessments of print-concept, alphabet, and emergent writing, and correlations between the teacher ratings and the direct assessments ranged from .43 to .60.

Less support for teacher ratings comes from analyses examining the accuracy at which teachers classify children with and without impairments. Findings from these studies are often reported in terms of overall accuracy (i.e., the proportion of children correctly classified as impaired or typical), sensitivity (i.e., the proportion of children classified as impaired who are impaired), and specificity (i.e., the proportion of children classified as typical who are typical). Although sensitivity and specificity values ≥ 90% are recommended for diagnostic tests, slightly lower values (i.e., sensitivity ≥ 80% and specificity ≥ 70%) are often recommended for screeners to ensure that children with impairments are not underreferred (Kilgus, Methe, Maggin, & Tomasula, 2014). Unfortunately, previous studies have not always reported high levels of sensitivity for teacher ratings. For example, Cabell et al. (2009) reported a sensitivity level of 52% for teacher ratings, and Jessup et al. (2008) reported an even lower level (15%) for teacher ratings that had been collected via a developmental checklist. In both studies, specificity levels were higher (88% and 97%, respectively), but low levels of sensitivity indicated that the teachers (or the tools used to collect the teacher ratings) were not sensitive to the language weaknesses of many children with LI or at risk for reading impairment.

## TROLL and CCC-2

In 2010, we selected the TROLL (Dickinson et al., 2001, 2003) for a larger study of kindergartners living in the rural South, because it focused on teacher ratings of children's oral language, reading, and writing skills (for other studies that utilized the TROLL, see Cunningham, 2009; McCabe, Boccia, Bennet, Lyman, & Hagen, 2010). Although the TROLL was designed for preschoolers, we considered the tool appropriate for exploratory purposes for our sample, because within the published norms, 5-year-olds at the end of preschool did not reach ceiling. Although we could not find classification accuracy indices for the TROLL, Dickinson et al. (2001) reported Cronbach alpha estimates of internal consistency ranging from .77 to .92 for separate subscales, and these values indicate strong internal consistency. In addition, in a sample of over 400 children, TROLL total scores were moderately correlated (*r* ranged from .42 to .47) with direct measures of children's vocabulary, early phonemic awareness, and emergent literacy skills.

Unfortunately, Rodriguez and Guiberson (2011) reported negative findings for the TROLL using a sample of 353 English-speaking, Spanish-speaking, and English–Spanish bilingual children, aged 4 years. Teachers were bilingual or had bilingual educational assistants, and the direct measure of the children's English language abilities was the Preschool Language Scale–Fourth Edition (PLS-4; Zimmerman, Steiner, & Pond, 2002). Correlations between the teachers' TROLL oral language ratings and the children's PLS-4 scores were lower than expected, and they varied by the children's language group. TROLL scores for the English-speaking children weakly correlated with PLS-4 receptive and expressive subtests ($r = .22$ and .20, respectively), TROLL scores for the Spanish-speaking children weakly correlated with only the PLS-4 receptive subtest ($r = .22$), and TROLL scores for the bilingual children were not correlated with either subtest ($r < .10$). In addition, teacher ratings were higher for the English-speaking group than for the others, and a smaller proportion of the English-speaking group failed the TROLL as compared with the others (20% vs. 32%–35%). From these findings, the authors cautioned against using the TROLL with culturally and linguistically diverse groups of children. One limitation of the study was that the teachers completed their ratings within the first 6 weeks of school when they may not have known the children well. Nevertheless, findings from the study were alarming given the dialects of our study sample.

In 2013, we added the CCC-2 (Bishop, 2006) as a teacher rating tool to our larger study of kindergartners. The CCC-2 is designed to help identify children, aged 4–16 years, with various communication disorders including LI and autism. Within the test manual, Bishop reports that, in a sample of 108 children, a composite score of 85 was 78% accurate at differentiating children with and without LI; sensitivity was 70%, and specificity was 85%. The CCC-2 or the original version of this tool has been included in numerous studies (e.g., Antoniazzi, Snow, & Dickson-Swift, 2010; Bishop, 1998; Bishop & Baird, 2001; Bishop, Laws, Adams, & Norbury, 2006; Bishop & McDonald, 2009; Bishop & Norbury, 2002; Botting, 2004; Norbury, Nash, Baird, & Bishop, 2004; Philofsky, Fidler, & Hepburn, 2007; Timler, 2014; Vaisanen, Loukusa, Moilanen, & Yliherva, 2014). Most relevant for the current work are results by Antoniazzi et al. (2010), whose sample included preschoolers at low and high risks for LI, and Timler (2014), whose sample included children, aged 5–8 years, with attention-deficit/hyperactivity disorder (ADHD) and LI, ADHD without LI, or TD. Whereas Antoniazzi et al. found low levels of sensitivity (41%) and specificity (73%), Timler found high levels (i.e., sensitivity was 100%, and specificity was 85%, with the misclassified children coming from the ADHD-without-LI group).

### DELV-ST-II, DIBELS, and DELV-NR

The DELV-ST-II screens children's risk for LI and includes items appropriate for different dialects of English, including AAE and SWE (Seymour et al., 2003). Within the test manual and using the high-risk category as the cut score, Seymour et al. report that, in a sample of 266 5-year-olds, sensitivity was 73% and specificity was 82%. Use of the DELV-ST-II for clinical practice also has been supported by two other studies, although in both, local norms were established and recommended rather than the criterion-based scores from the test manual (Petscher, Connor, & Al Otaiba, 2012; Terry, Petscher, & Rhodes, 2016).

The DIBELS is a curriculum-based screener of children's reading abilities (Good et al., 2009). Multiple studies have investigated the reliability and validity of the DIBELS (Cummings, Park, & Bauer Schaper, 2013; Dewey, Kaminski, & Good, 2014; Dewey, Powell-Smith, Good, & Kaminski, 2015; Elliot, Lee, & Tollefson, 2001; Shaw & Shaw, 2002). In a recent meta-analysis, which set sensitivity levels at ≥ 80% and specificity at ≥ 70%, the DIBELS was supported for schoolwide screening purposes, although a key finding of the study was the need for different cut scores to optimize sensitivity and specificity values across samples (Kilgus et al., 2014).

The DELV-NR is the only dialect-neutral, norm-referenced language test within the field of speech-language pathology, and 63% of its normative sample was speakers of nonmainstream English (Seymour et al., 2005). Within the test manual, Seymour et al. report that, in a sample of 176 children, a cut score of −1 *SD* from the normative mean was 94% accurate in classifying children as LI or TD; sensitivity was 95%, and specificity was 93%. Use of the DELV-NR for clinical practice has been supported by two additional studies (Pearson, de Villiers, Magaziner, Perisho, & Sutherland, 2005; Pearson, Jackson, & Wu, 2014). In a previous study with some of the children studied here, DELV-NR syntax scores also correlated with scores on a sentence recall task (Oetting, McDonald, Seidel, & Hegarty, 2016).

### Research Objectives

As stated earlier, the goal of the study was to evaluate teacher ratings as measured by the TROLL and CCC-2 for screening purposes when working with nonmainstream English-speaking kindergartners in the rural South. We first conducted analyses of variance (ANOVAs) and correlational analyses with the TROLL, CCC-2, DELV-ST-II, and

DIBELS. These analyses provided an examination of the convergent validity of the TROLL and CCC-2 and allowed for a comparison of the findings with previous studies. Kilgus et al. (2014) further described these analyses as necessary (but not sufficient) in studies of diagnostic accuracy. Then, we examined the classification accuracies of the TROLL, CCC-2, DELV-ST-II, and DIBELS using the syntax subtest of the DELV-NR. All analyses were conducted with the dialect groups combined and separated to examine whether the results varied by the children's cultural and linguistic backgrounds. Questions guiding the study were the following:

1. Do teacher ratings as measured by the TROLL and CCC-2 and scores on the DELV-ST-II and DIBELS differ by the children's clinical group (LI vs. TD) and dialect (AAE vs. SWE)?
2. Do teacher ratings as measured by the TROLL and CCC-2 and scores on the DELV-ST-II and DIBELS correlate to each other and other measures of language ability? Are the correlations comparable for the two dialects?
3. What are the classification accuracies of the TROLL, CCC-2, DELV-ST-II, and DIBELS? Are classification accuracies comparable for the two dialects? Can classification accuracies be improved by using different cut scores or by considering scores from more than one tool?

### *Research Hypotheses*

If teacher ratings are valid measures of kindergartners' language abilities, we expected results from the TROLL and CCC-2 to be similar to those from the DELV-ST-II and DIBELS, with all four tools showing effects for the children's clinical status but not dialect and all four tools correlating with each other and other measures of language ability. Although we did not expect the TROLL and CCC-2 to show high levels of classification accuracy (i.e., sensitivity), we hoped to determine if one teacher rating instrument was better than the other, explore different cut scores, and examine whether one or both teacher rating instruments could be combined with the DELV-ST-II and/or DIBELS for screening purposes.

## Methods

### *Research Sites, Recruitment Design, and Retrospective Nature of Analysis*

Data were collected during the 2013–2014 school year as part of a multiyear study. Four public schools in one rural southeastern Louisiana school district were selected as research sites. One additional public school in the district enrolled kindergartners; although this school agreed to participate in the study, space limitations at the school prevented data collection. According to the U.S. Department of Agriculture, the district is located in a parish (i.e., county) with persistent child poverty, which is defined as a child poverty rate of 20% or more over 3 decades. With high percentages of children historically receiving free and reduced lunches, all four targeted schools also were eligible for the federally funded Community Eligibility Provision program, and they joined this program in 2017. This program allows the nation's highest-poverty schools to serve breakfast and lunch at no cost to all enrolled students without the burden of collecting household applications.

Recruitment for the study at the participating schools most closely resembled a one-gate design, because all consent forms were sent home via the children's backpacks. Specifically, speech-language pathologists at the schools were provided consent forms for children on their caseloads, and they gave the forms (often with a note of endorsement or approval to the parents) to the teachers to disseminate along with consent forms for all others. During the 2013–2014 school year, 213 children were enrolled in the participating kindergartens, and 172 (81%) returned a consent form. Of the 172 children, 98 were excluded from the current analyses for the following reasons: Spanish–English bilingual status (*n* = 4), placement in a French immersion program (*n* = 23), documented developmental disabilities or communication disorders other than LI (*n* = 10), grade repetition (*n* = 5), inconsistent attendance (*n* = 1), school transfer (*n* = 7), twin of another participant (*n* = 2), incomplete or missing CCC-2 data (*n* = 2), and insufficient time in the school year to complete data collection (*n* = 44; most of these children received at least 1 day of testing, and results suggested TD status).

Although a one-gate design was employed for recruitment, the goal of the multiyear study was to identify LI and TD groups of AAE and SWE speakers who were matched on dialect, age, nonverbal cognition, and, when possible, maternal education; these children also earned nonverbal cognitive quotients and articulation scores within normal limits (≥ −1.2 and ≥ −1 *SD*s of the normative mean, respectively). As shown in Table 1, of the 98 participants included in the current study, 24 met the criteria for the matched-group study (and were included in Oetting et al., 2016). The other 74 children were recruited for the matched-group study, but they were not included because an LI or TD match was unavailable or they did not present with nonverbal cognitive and/or articulation abilities ≥ −1.2 *SD*s of the normative mean.

### *Participants*

#### Children

Fifty children were male, and 48 were female; their ages in months averaged 65.41 (*SD* = 3.97, range = 52–75). Race and maternal education information was not reported by the caregivers for four children. For the others, 40 (41%) were Black, 50 (51%) were White, and four (4%) were classified as "other" (one Native American and three with more than one race); their average maternal educational level was 12.48 years (*SD* = 2.40 years, range = 6–17+ years). The children's gender and race distributions were similar to

**Table 1.** Participant pool, study criteria, and participant overlap across studies.

| Data available for the current analysis 2013–2014 | Data collected for multiyear study 2010–2014 |
|---|---|
| Children enrolled in target kindergartens, *N* = 213 | Children enrolled in target kindergartens, *N* = 834 |
| Children who returned a consent form, *N* = 172 | Children who returned a consent form, *N* = 669 |
| Children with TROLL and CCC-2 data, *N* = 98 | Oetting et al. (2016), *N* = 106 |
| Unmatched group with LI and TD group; hearing screen, DELV-NR syntax to determine clinical status; PTONI, GFTA-2, PPVT-4 free to vary | Matched group with LI and TD group; hearing screen, DELV-NR syntax, PTONI, and GFTA-2 to determine clinical status; PPVT-4 free to vary |
| Children with TROLL and CCC-2; also in Oetting et al., *N* = 24 ↔ | Children with TROLL and CCC-2; also in the study, *N* = 24 |
| Children with TROLL and CCC-2; not in Oetting et al., *N* = 74 | Children without TROLL and CCC-2, *N* = 10 |
|  | Children with TROLL only (used as independent test sample), *N* = 72 |

*Note.* CCC-2 = Children's Communicative Checklist–Second Edition; DELV-NR = Diagnostic Evaluation of Language Variation–Norm Referenced; GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition; LI = language impairment; PTONI = Primary Test of Nonverbal Intelligence; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; TD = typically developing; TROLL = Teacher Rating of Oral Language and Literacy.

those of the schoolwide kindergarten rosters (gender: male = 47%, female = 53%; race: Black = 35%, White = 61%, and other = 4%). Their maternal education profile was similar to the profile of the multiyear sample (*M* = 12.57, *SD* = 2.70, range = 6–17+) studied by Oetting et al. (2016).

Forty-seven children spoke AAE and 51 spoke SWE based on the children's race and informal listener judgments. In addition, nonmainstream features consistent with AAE and/or SWE were identified for 50 children who had available language samples, and all children completed Part I of the DELV-ST (DELV-ST-I; Seymour et al., 2003). The DELV-ST-I focuses on the children's use of mainstream and nonmainstream English patterns when imitating sentences, completing sentence close items, and answering questions. As expected, the AAE group's percentage (*M* = 0.85, *SD* = 0.17) of nonmainstream patterns on the DELV-ST-I was higher than the SWE group's percentage (*M* = 0.51, *SD* = 0.28), *F*(1, 96) = 51.63, $p < .001$, $\eta^2 = .35$.

Using the syntax subtest of the DELV-NR, 43 children (28 AAE and 15 SWE) were classified as LI, and 55 (19 AAE and 36 SWE) were classified as TD. The syntax subtest includes three types of items (i.e., comprehension of complex *wh*-questions, comprehension of passives and alternative by-phrase constructions, and production of articles to assess comprehension of new vs. old information); it has a normative mean of 10 (*SD* = 3). All children classified as LI earned a standard score of ≤ 7 on this subtest, and all children classified as TD earned a score > 7. Six (16%) children in the group with LI (four AAE and two SWE) and none in the TD group received services by a speech-language pathologist. Although the caregivers of many children did not provide family history information, 12 (28%) in the group with LI (seven AAE and five SWE) and 10 (18%) in the TD group (five AAE and five SWE) reported a positive family history for speech and/or LI. These findings are consistent with those of others who have reported low identification rates of LI in kindergarten (Tomblin et al., 1997) and higher rates of a positive family history of impairment by children with LI than by TD controls (Leonard, 2014).

All children passed a school-administered hearing screening and completed the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2; Goldman & Fristoe, 2000), Primary Test of Nonverbal Intelligence (PTONI; Ehrler & McGhee, 2008), and Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4; Dunn & Dunn, 2007); these tests have a normative mean of 100 (*SD* = 15). All but one child in the group with LI (who earned a 70) earned a standard score > 85 on the GFTA-2. Although none of the children had been identified by their schools as presenting low nonverbal cognitive abilities, 24 earned scores lower than 85 on the PTONI—17 were from the group with LI (13 AAE and four SWE), and seven were from the TD group (two AAE and five SWE). Fifteen children in the group with LI (11 AAE and four SWE) and one AAE child in the TD group earned a standard score below 85 on the PPVT-4.

Table 2 lists the children's test scores by their dialect and clinical status. Two-way ANOVAs indicated that the children's scores on all four tests varied by clinical status: DELV-NR, *F*(1, 94) = 169.4, $p < .001$, $\eta_p^2 = .64$; GFTA-2, *F*(1, 94) = 7.02, $p = .009$, $\eta_p^2 = .07$; PTONI, *F*(1, 94) = 16.87, $p < .001$, $\eta_p^2 = .15$; and PPVT-4, *F*(1, 94) = 41.82, $p < .001$, $\eta_p^2 = .31$. Although the DELV-NR was the only tool used to classify the children's clinical status, the scores of the group with LI on all four tools were lower than the TD group's scores. The ANOVA results also indicated that the children's PPVT-4 scores varied by dialect, *F*(1, 94) = 6.62, $p = .012$, $\eta_p^2 = .07$, with the AAE group mean lower than the SWE group mean (*M* = 93.96, *SD* = 13.94 vs. *M* = 104.22, *SD* = 11.49, respectively). Others who have found children's vocabularies to vary by cultural and linguistic backgrounds include Campbell, Dollaghan, Needleman, and Janosky (1997); Ellis Weismer et al. (2000); and Qi, Kaiser, Milan, Yzquierdo, and Hancock (2003).

### Teachers

All regular-education kindergarten teachers (eight women: one Black and seven White) employed by the children's schools agreed to participate. All were lead teachers, and they completed the TROLL and CCC-2 ratings for the participating children enrolled in their classes (*M* per teacher = 12.25, *SD* = 4.80, range = 6–19). Four teachers reported teaching for 5 years or less, and four reported teaching for 15 years or more.

**Table 2.** Participant mean test scores by dialect and clinical status.

| Group | DELV-NR | GFTA-2 | PTONI | PPVT-4 |
|---|---|---|---|---|
| AAE | | | | |
| LI (*n* = 28) | 5.11 (1.47) | 104.54 (9.50) | 88.96 (14.57) | 87.71 (12.78) |
| TD (*n* = 19) | 9.74 (1.20) | 109.32 (3.82) | 102.79 (13.62) | 103.16 (10.05) |
| SWE | | | | |
| LI (*n* = 15) | 5.47 (1.36) | 107.60 (3.85) | 92.73 (10.55) | 94.27 (12.70) |
| TD (*n* = 36) | 9.72 (1.92) | 109.83 (4.46) | 103.97 (15.60) | 108.36 (8.00) |
| All participants | | | | |
| LI (*N* = 43) | 5.23 (1.43) | 105.60 (8.07) | 90.28 (13.30) | 90.00 (13.00) |
| TD (*N* = 55) | 9.73 (1.69) | 109.65 (4.22) | 103.56 (14.83) | 106.56 (9.02) |

*Note.* AAE = African American English; DELV-NR = Diagnostic Evaluation of Language Variation–Norm Referenced Test (*M* = 10, *SD* = 3); GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition (*M* = 100, *SD* = 15); LI = language impairment; PTONI = Primary Test of Nonverbal Intelligence (*M* = 100, *SD* = 15); PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition (*M* = 100, *SD* = 15); SWE = Southern White English; TD = typically developing.

## *Materials*

### TROLL

The TROLL includes 25 items: eight focus on the child's use of language (e.g., willingness to start a conversation), 11 focus on reading (e.g., frequency/quality of emergent reading), and six focus on writing (e.g., frequency/quality of emergent writing). Twenty-four items are scored with a 4-point scale, and one is scored as either 1 or 2. For some items, the 4-point scale involves "never," "rarely," "sometimes," and "often." For other items, more elaborate choices are offered (e.g., child almost never begins a conversation…; if unsuccessful at first, child sometimes…; if initial efforts fail, child will sometimes…; if initial efforts fail, child will…). The maximum score possible is 98 (4 points for 24 items; 2 points for one item). The initial cut score for failing was set at 65 or the 25th percentile in the spring for 5-year-olds. Recall that the TROLL was designed for preschoolers, so this cut score was not expected to be optimal for the kindergartners studied here. Instead, it was selected for the initial analysis, with the expectation that other cut scores would be explored within the analyses.

### CCC-2

The CCC-2 includes 70 items that are divided into 10 scales (i.e., speech, syntax, semantics, coherence, initiation, scripted language, context, nonverbal communication, social relations, interests). Each scale contains seven items, with five addressing difficulties (e.g., being left out of activities) and two addressing strengths (e.g., being able to have an interesting conversation). Items are scored using a 4-point scale (e.g., less than once a week, at least once a week but not every day); raw scores were converted to standard scores. The CCC-2 composite score has a normative mean of 100 (*SD* = 15). The initial cut score for failing this screener was set at 85 or the 16th percentile.

### DELV-ST-II

The DELV-ST-II includes 17 items: seven target morphology, four target *wh*-question comprehension, and six require the child to repeat a nonword. The children's incorrect responses are used to calculate an error score. The cut score for failing was set at nine errors, which corresponded to the high-risk category for 5-year-olds. Although the ages of three children fell outside the 5-year-old age range (one was 52 months old, and two were 75 months old) and different cut scores are recommended for these ages, interpretation of these children's error scores did not change with the cut score of nine errors.

### DIBELS

The DIBELS was administered by the school three times (beginning, middle, and end) during the year. The fall DIBELS scores were included in the analysis, because all children had these scores, and the findings did not differ as a function of which score was analyzed. Administration of the DIBELS in the fall included two subtests: first sound fluency and letter naming fluency. For the first sound fluency subtest, children heard 30 words and produced the initial sound of each word. Two points were awarded for the correct initial sound, and 1 point was awarded for the correct initial blend or syllable. Children who could not complete the first five items earned a zero, and testing was discontinued. For the letter naming fluency subtest, children saw a set of uppercase and lowercase letters and named as many letters as possible. Children who could not name the first 10 letters earned a zero, and testing was discontinued. The two subtests formed a composite raw score that was converted to a percentile rank. The initial cut score was set at a composite raw score of 11 or the 25th percentile.

## *Procedures*

Caregiver and teacher consent was secured before data collection. Teachers completed the CCC-2 and TROLL in the spring or immediately after the school year ended; forms were returned via post, and teachers received $20.00 for each form returned. The children's DIBELS scores were collected from the schools at the end of the school year. Twelve trained student examiners administered the remaining measures to the children individually during the school year.

All children completed the same set of standardized tests, with the DELV-ST-I and DELV-ST-II administered in the first session. The examiners were not blind to the children's DELV-ST-I and DELV-ST-II test scores when administering the other tests, but they were blind to the teacher ratings and the DIBELS scores. The teachers were blind to the dialect and clinical groupings of the children and all data collected by the examiners, but they were not blind to the DIBELS scores.

### *Reliability*

A second examiner independently scored 20% of the TROLL and CCC-2 forms. There were 840 (21 children × 40 scores that came from various subtests of the tools) opportunities for agreement, and the rate of agreement was 97% (814/840). A second examiner also checked the scoring of all other examiner-administered test protocols, with disagreements resolved through consensus. Data entry was checked by comparing manual and computer-generated totals and subtotals within the databases. Although scoring of the DIBELS could not be examined, data entry was checked in the same manner as the other tools.

## Results

### *Preliminary Analyses*

One-way ANOVAs were completed to examine whether the children's TROLL and CCC-2 scores differed by the teachers' years of experience (≤ 5 vs. ≥ 15); differences were not detected (TROLL: $p = .37$, CCC-2: $p = .97$). We also examined correlations between the TROLL subtest scores and total and the CCC-2 subtest scores and composite. TROLL subtest scores were highly correlated ($r$ ranged from .90 to .96) with the TROLL total score. CCC-2 subtest scores for speech, syntax, semantics, scripted language, context, and nonverbal communication were highly correlated ($r$ ranged from .80 to .91) with the CCC-2 composite score, with the remaining subtest scores moderately to highly correlated ($r$ ranged from .64 to .71). Finally, the TROLL total and CCC-2 composite yielded the highest correlation ($r = .71$) between the two instruments as compared with the individual subtest scores ($r$ ranged from .33 to .70). These results led to the use of the TROLL total and CCC-2 composite within the analyses.

### *Group Differences*

Table 3 presents the teacher ratings from the TROLL and CCC-2 and the children's scores on the DELV-ST-II and DIBELS. Four 2-way ANOVAs were completed to examine the children's scores by clinical status and dialect. There were significant differences for each tool by clinical status: TROLL, $F(1, 94) = 21.17$, $p < .001$, $\eta_p^2 = .18$; CCC-2, $F(1, 94) = 20.46$, $p < .001$, $\eta_p^2 = .18$; DELV-ST-II, $F(1, 94) = 15.57$, $p < .001$, $\eta_p^2 = .14$; and DIBELS, $F(1, 94) = 11.42$, $p < .001$, $\eta_p^2 = .11$. The scores of the group with LI for the TROLL, CCC-2, and DIBELS were lower than the TD group's scores. The DELV-ST-II scores reflected the number of errors the children made, so here, the scores of the group with LI were higher than the TD group's scores. There were no significant differences by dialect for the tools (TROLL: $p = .65$, CCC-2: $p = .87$, DELV-ST-II: $p = .94$, DIBELS: $p = .23$), and effect sizes accompanying these null effects were negligible (TROLL: $\eta_p^2 = .002$, CCC-2: $\eta_p^2 < .001$, DELV-ST-II: $\eta_p^2 < .001$, DIBELS: $\eta_p^2 = .015$).

### *Correlations*

Correlations between the various tools are reported in Table 4 for the dialects combined and in Table 5 for the AAE and SWE dialects separately. With the dialects combined, the teacher ratings from the TROLL and CCC-2 moderately to highly correlated with each other ($r = .71$) and moderately correlated with the DIBELS ($r$s = .59 and .43, respectively); the TROLL, but not the CCC-2, moderately correlated with the DELV-ST-II ($r = −.30$). All four tools also correlated with the other tests in similar ways; across tools, the highest correlations were with the DELV-NR ($r$ range = ±.31 to .47) and PPVT-4 ($r$ range = ±.33 to .45), and the lowest were with the GFTA-2 ($r$ range = ±.19 to .35).

Similar results were found when the AAE and SWE dialects were examined separately. The TROLL and CCC-2 were moderately to highly correlated with each other (AAE: $r = .77$, SWE: $r = .67$) and moderately correlated with the DIBELS (AAE: $r$s = .57 and 44, SWE: $r$s = .62 and .43). In addition, the highest correlations between the four tools and the other tests were with the DELV-NR (AAE: $r$ range = ±.28 to .48, SWE: $r$ range = ±.37 to .55) and PPVT-4 (AAE: $r$ range = ±.30 to .47, SWE: $r$ range = ±.37 to .44), and the lowest correlations were with the GFTA-2 (AAE: $r$ range = ±.20 to .35, SWE: $r$ range = ±.17 to .36). The only result that did not replicate when the dialects were separated involved the correlation between the TROLL and the DELV-ST-II for the AAE group (AAE: $r = −.25$, $p > .05$, compared with dialects combined: $r = −.30$, $p < .001$), but the magnitude of the correlational difference was minimal.

### *Classification Accuracy*

In the final set of analyses, we examined the accuracy at which the TROLL, CCC-2, DELV-ST-II, and DIBELS classified the children based on their clinical status (LI vs. TD). We first completed this analysis using the cut scores recommended by the test developers and as specified in the Methods section above. Then, we completed a series of discriminant function analyses to identify the optimal cut score for each tool and the tool or combination of tools that maximized the differences between the group with LI and the TD group.

#### Recommended Cut Scores

The recommended cut scores led to low classification accuracies: TROLL = 59%, CCC-2 = 68%, DELV-ST-II = 66%, and DIBELS = 62% (see Table 6). The low levels of

**Table 3.** Participant scores on tools by dialect and clinical status.

| Group | TROLL | CCC-2 | DELV-ST-II | DIBELS |
|---|---|---|---|---|
| AAE | | | | |
| LI | 82.29 (11.75) | 94.43 (16.65) | 9.57 (3.85) | 55.93 (28.29) |
| TD | 89.63 (8.34) | 107.16 (15.21) | 5.84 (2.77) | 65.47 (25.43) |
| SWE | | | | |
| LI | 77.67 (15.25) | 91.60 (18.21) | 9.00 (3.19) | 41.20 (24.96) |
| TD | 92.08 (9.67) | 108.92 (13.53) | 6.53 (4.10) | 69.56 (21.14) |
| All participants | | | | |
| LI | 80.67 (13.08) | 93.44 (17.04) | 9.37 (3.61) | 50.79 (27.79) |
| TD | 91.24 (9.23) | 108.31 (14.01) | 6.29 (3.69) | 68.13 (22.56) |

*Note.* Means reported first, followed by standard deviations in parentheses. AAE = African American English; CCC-2 = Children's Communicative Checklist–Second Edition (*M* = 100, *SD* = 15); DELV-ST-II = Part II of the Diagnostic Evaluation of Language Variation–Screening Test (number of errors reported, high risk = 9 errors); DIBELS = Dynamic Indicators of Basic Early Literacy Skills–Next (percentiles reported, referral ≤ 24th percentile); LI = language impairment; SWE = Southern White English; TROLL = Teacher Rating of Oral Language and Literacy (maximum = 98); TD = typically developing.

accuracy were accompanied by extremely low levels of sensitivity (range = 9%–56%), which indicated that many children classified as LI scored above the cut score. Corresponding levels of specificity (range = 75%–98%) were higher, but this was not surprising given the low sensitivity levels. Table 6 also lists the sensitivity and specificity levels of each tool using the recommended cut scores for the AAE and SWE dialects separately. All four tools led to unacceptably low levels of sensitivity (range = 7%–57%), regardless of the dialect group examined. Specificity was higher than sensitivity (range = 72%–100%) for both dialects, but again, those values were tied to the unacceptably low levels of sensitivity. In other words, had these tools and cut scores been used to decide who should be referred for a language evaluation, many children in the group with LI across both dialects would have been missed.

### Empirically Derived Cut Scores

Four discriminant function analyses, one for each tool, were completed to identify the cut score that best separated children in the group with LI and the TD group. As part of these analyses, the normality of the data was examined, and a negative skew was identified for the TROLL. Square root transformations with reflection were completed to normalize the distribution of the TROLL data, and the tools were examined for outliers by comparing each child's score with their group mean. One outlier (a child in the TD group with a very low score of 45) was identified for the TROLL. We removed this case to complete all discriminant function analyses involving the TROLL and then added the case back to calculate classification accuracies. We also tested the equality of the groups' covariance matrices using the Box's *M* statistic. The assumption of equality was met for each tool (TROLL: $p = .08$, CCC-2: $p = .18$, DELV-ST-II: $p = .88$, DIBELS: $p = .23$), and the log determinants for each group were similar (TROLL: 0.92 and 0.41, CCC-2: 1.82 and 1.80, DELV-ST-II: 2.57 and 2.61, DIBELS: 6.25 and 5.91).

Table 6 lists the empirically derived cut scores from the analyses and their corresponding levels of sensitivity and specificity (see also Supplemental Material S1 for raw data and other accuracy indices and confidence intervals). The

**Table 4.** Correlations between measures: all participants.

| Measure | TROLL | CCC-2 | DELV-ST-II | DIBELS | DELV-NR | GFTA-2 | PTONI |
|---|---|---|---|---|---|---|---|
| TROLL | — | | | | | | |
| CCC-2 | .71** | — | | | | | |
| DELV-ST-II | −.30** | −.16 | — | | | | |
| DIBELS | .59** | .43** | −.26** | — | | | |
| DELV-NR | .47** | .46** | −.45** | .31** | — | | |
| GFTA-2 | .26** | .35** | −.24* | .19 | .34** | — | |
| PTONI | .41** | .38** | −.21* | .35** | .49** | .22* | — |
| PPVT-4 | .45** | .43** | −.38** | .33** | .63** | .28** | .56** |

*Note.* CCC-2 = Children's Communicative Checklist–Second Edition; DELV-ST-II = Part II of the Diagnostic Evaluation of Language Variation–Screening Test; DIBELS = Dynamic Indicators of Basic Early Literacy Skills–Next; DELV-NR = Diagnostic Evaluation of Language Variation–Norm-Referenced Test; GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition; PTONI = Primary Test of Nonverbal Intelligence; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; TROLL = Teacher Rating of Oral Language and Literacy.

*$p < .05$. **$p < .001$.

**Table 5.** Correlations between measures: AAE and SWE dialects.

| | TROLL | CCC-2 | DELV-ST-II | DIBELS | DELV-NR | GFTA-2 | PTONI | PPVT-4 |
|---|---|---|---|---|---|---|---|---|
| TROLL | — | .67** | −.33* | .62** | .55** | .19 | .33** | .44** |
| CCC-2 | .77** | — | −.07 | .43** | .45** | .36* | .29* | .37** |
| DELV-ST-II | −.25 | −.23 | — | −.29* | −.40** | −.17 | −.24 | .42** |
| DIBELS | .57** | .44** | −.24 | — | .37** | .20 | .30* | .42** |
| DELV-NR | .35* | .44** | −.48** | .28 | — | .35* | .35* | .62** |
| GFTA-2 | .32* | .35* | −.27 | .20 | .29* | — | .02 | .36* |
| PTONI | .50** | .45** | −.14 | .41** | .57** | .29* | — | .55** |
| PPVT-4 | .46** | .47** | −.33* | .30* | .57** | .16 | .52** | — |

*Note.* Data for the AAE group are presented below axis; data for the SWE group are presented above axis. AAE = African American English; CCC–Second Edition = Children's Communicative Checklist-2; DELV-ST-II = Part II of the Diagnostic Evaluation of Language Variation–Screening Test; DIBELS = Dynamic Indicators of Basic Early Literacy Skills–Next; DELV-NR = Diagnostic Evaluation of Language Variation–Norm-Referenced Test; GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition; PTONI = Primary Test of Nonverbal Intelligence; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; SWE = Southern White English; TROLL = Teacher Rating of Oral Language and Literacy.

*$p$ < .05. **$p$ < .001.

derived cut scores were more stringent than the recommended cut scores. The TROLL total score increased from 65 to 89, the CCC-2 composite increased from 85 to 100, the DELV-ST-II error score decreased from 9 to 8, and the DIBELS total score increased from 11 to 38. The derived scores led to higher levels of sensitivity, with corresponding reductions in specificity, and these results were observed when the dialects were combined and separated. Across both dialect groups, the derived cut score for the TROLL led to the highest level of sensitivity (77%), and the DIBELS led to the lowest (63%), although the 65% sensitivity level for the CCC-2 and DELV-ST-II was not much higher than the level obtained for the DIBELS.

Finally, we completed a stepwise discriminant analysis that included the TROLL, CCC-2, DELV-ST-II, and DIBELS as the predictor variables and the children's clinical status as the predetermined grouping variable. A stepwise discriminant analysis takes into consideration scores that are intercorrelated and selects the best score or set of scores (in this case, the best tool or set of tools) that maximizes differences between the group with LI and the TD group. Again, the assumption of equality for the groups' covariance matrices was met ($p$ = .389), and the groups' log determinants were similar (3.48 and 3.02) as tested by the Box's $M$ statistic. The stepwise discriminant function first selected the TROLL as the best tool for predicting the children's clinical status. The overall chi-square test of this discriminant function was significant (Wilk's $\lambda$ = 0.74, $\chi^2$ = 37.85, $df$ = 2, canonical correlation = .58, $p$ < .001), and the TROLL accounted for 33% of the variance between the group with LI and the TD group. As was found in the earlier analysis, classification accuracy was 76% (sensitivity = 77%, specificity = 75%).

The stepwise discriminant function analysis then selected the TROLL and DELV-ST-II as the best combination of tools for predicting the children's clinical status. The overall chi-square test of the discriminant function was again significant (Wilk's $\lambda$ = 0.67, $\chi^2$ = 37.85, $df$ = 2, canonical correlation = .58, $p$ < .001), and the TROLL and DELV-ST-II together accounted for the same amount

**Table 6.** Sensitivity (Se) and specificity (Sp) of tools.

| | | All participants | | AAE | | SWE | |
|---|---|---|---|---|---|---|---|
| **Test tool and cut scores** | | **Se** | **Sp** | **Se** | **Sp** | **Se** | **Sp** |
| Recommended clinical cut scores | | | | | | | |
| TROLL | 65 total score | .09 | .98 | .07 | 1.00 | .13 | .97 |
| CCC-2 | 85 composite score | .33 | .96 | .36 | 1.00 | .27 | .94 |
| DELV-ST-II | 9 error score | .56 | .75 | .57 | .79 | .53 | .72 |
| DIBELS | 11 total score | .21 | .96 | .18 | .95 | .27 | .94 |
| Empirically derived cut scores | | | | | | | |
| TROLL | 89 total score | .77 | .75 | .75 | .68 | .80 | .78 |
| CCC-2 | 100 composite score | .65 | .69 | .61 | .63 | .73 | .72 |
| DELV-ST-II | 8 error score | .65 | .62 | .64 | .74 | .67 | .56 |
| DIBELS | 38 total score | .63 | .62 | .57 | .58 | .73 | .64 |

*Note.* AAE = African American English; CCC-2 = Children's Communicative Checklist–Second Edition; DELV-ST-II = Part II of the Diagnostic Evaluation of Language Variation–Screening Test; DIBELS = Dynamic Indicators of Basic Early Literacy Skills–Next; SWE = Southern White English; TROLL = Teacher Rating of Oral Language and Literacy.

of variance (33%) between the group with LI and the TD group as the TROLL alone. The TROLL and DELV-ST-II together yielded a classification accuracy level of 79% (sensitivity = 74%, specificity = 82%). Although the TROLL and DELV-ST-II together correctly classified the most children (79% vs. 76%), the TROLL alone correctly classified the greatest proportion of children in the group with LI (77% vs. 74%).

## Discussion

The goal of the study was to examine the validity of teacher ratings when screening nonmainstream English-speaking kindergartners in the rural South. If valid, teacher ratings may provide a cost-effective alternative to language screenings conducted by speech-language pathologists in the schools. The children spoke one of two nonmainstream dialects, AAE or SWE, and teacher ratings were collected with the TROLL and CCC-2. The convergent validity of the TROLL and CCC-2 was examined with two established screeners, the DELV-ST-II and DIBELS, and then the classification accuracy of all four tools (i.e., TROLL, CCC-2, DELV-ST-II, and DIBELS) was examined using the syntax subtest of the DELV-NR.

Results indicated that both teacher rating tools as well as the DELV-ST-II and DIBELS yielded scores that differed by the children's clinical status but not their dialect. The TROLL and CCC-2 correlated with each other, the DELV-ST-II, and the DIBELS, and all four tools correlated with the DELV-NR and PPVT-4. The effect sizes of the clinical group differences and the magnitudes of the correlations were similar for the AAE and SWE speakers, which indicates that the four tools measured similar language ability constructs within the two dialects. Together, these findings provide support for teacher ratings as measured by the TROLL and CCC-2.

Recommended cut scores for the TROLL, CCC-2, DELV-ST-II, and DIBELS led to sensitivity levels that were too low for screening purposes. The results improved with discriminant function analyses, and a stepwise analysis identified the TROLL as the best tool of the four and the TROLL and DELV-ST-II as the best combination of tools for predicting the children's clinical status. The TROLL alone was 76% accurate in predicting the children's clinical status, and the TROLL and DELV-ST-II together were 79% accurate. Although the TROLL was less accurate than the TROLL and DELV-ST-II together, this tool alone yielded the highest level of sensitivity (77%). As noted in the literature review, tools with high levels of sensitivity ensure that children with LI are referred for evaluations.

### *Findings as Related to Previous Studies*

Previous studies have found teacher ratings to differ by children's clinical status and/or ability levels and to correlate at moderate levels to direct measures of children's language and/or literacy abilities. Results of the current study are consistent with these previous studies. Results from the current study are not consistent with Rodriguez and Guiberson (2011), who found the TROLL to be unrelated to a direct measure of children's language abilities and to vary by children's cultural and linguistic backgrounds. Rodriguez and Guiberson studied English- and Spanish-speaking children and bilinguals who were in preschool, and teachers completed their ratings at the beginning of the school year. In contrast, the cultural and linguistic differences of the participants in the current study were dialectal, the participants were in kindergarten, and the teachers completed their ratings at the end of the year.

Previous studies have reported low levels of classification accuracy (i.e., sensitivity) for teacher ratings. In the current study, low levels of sensitivity for the TROLL (9%) and CCC-2 (33%) were also found when initial cut scores were examined, but these levels increased (TROLL = 77%, CCC-2 = 65%) with empirically derived cut scores. Although we were unable to find published accuracy indices for the TROLL, this information was available for the CCC-2, with Bishop (2006), Antoniazzi et al. (2010), and Timler (2014) reporting sensitivity as 70%, 41%, and 100%, respectively. The participants within these previous studies have varied in age and nature of impairment. In addition, Bishop and Timler's groups with LI were recruited from either a clinical caseload or a clinical research sample, whereas Antonioazzi et al.'s participants and most of the children in the group with LI studied here were identified with testing (for others who have used testing to identify children with LI, see Brumbach & Goffman, 2014; Noonan, Redmond, & Archibald, 2014; Poll, Betz, & Miller, 2010; Spaulding, 2010; Tomblin et al., 1997; Victorino & Schwartz, 2015).

In the current study, sensitivity levels for the DELV-ST-II (56%) and DIBELS (21%) were also lower than expected with recommended cut scores, but they too increased with empirically derived cut scores (DELV-ST-II = 65%, DIBELS = 63%), albeit not at levels published elsewhere. Recall that, for the DELV-ST-II, Seymour et al. (2003) reported a sensitivity level of 73% for 5-year-olds, and the DIBELS was supported by a meta-analysis that required a sensitivity level of ≥ 80%. Like Bishop's (2006) group with LI for the CCC-2, Seymour et al.'s group with LI for the DELV-ST-II was recruited from clinical caseloads (and the manual notes that 47% of these children also received services for other conditions, such as ADHD or developmental delay). Within the meta-analysis of the DIBELS, participant characteristics varied across study samples. Recall also that a key finding of the meta-analysis was the need for different cut scores across samples; this same finding was observed in our study.

As discussed by Kilgus et al. (2014); Rutjes, Reitsma, Vandenbroucke, Glas, and Bossuyt (2005); Whiting, Rutjes, Westwood, Mallett, & QUADAS-2 Steering Group (2013); and many others in the fields of education and medicine, the diagnostic accuracy of any given tool and cut score is heavily dependent on the conditions under which the tool is administered. The observation that diagnostic accuracy information is inherently variable has also led Kilgus et al. and others to view diagnostic accuracy information as less about the tools themselves and more about how tools behave for particular groups of individuals. From this perspective, the diagnostic accuracy information found here and elsewhere for the TROLL, CCC-2, DELV-ST-II, and DIBELS must be considered relevant to the particular samples studied. As discussed by Dollaghan and Horner (2011), important variables within diagnostic accuracy studies include not only the inclusionary and exclusionary criteria used to select the participants but also the characteristics of the participant pool, the base rate of the clinical condition within the pool, and the recruitment methods.

### *Limitations of the Study*

Diagnostic accuracy studies are susceptible to methodological biases (Dollaghan & Horner, 2011; Rutjes et al., 2005; Whiting et al., 2013), and three of these—subjectivity, spectrum, and incorporation—are of concern here. Subjectivity bias occurs when the same examiners administer the index tool(s) and the reference standard. In the current study, the same team of examiners administered all tests, but they were blind to the teacher ratings and the children's DIBELS scores. The location of the kindergartens (i.e., 45+ miles from the university), the kindergartners' availability (i.e., 1 hr per day), and the examiners' data collection schedules (i.e., twice weekly) reduced the likelihood that the same examiner administered all tests to the same child. Moreover, teacher ratings from the TROLL led to the highest level of sensitivity, and the teachers were blind to the examiners' data.

Spectrum bias occurs when the sample of participants used to calculate diagnostic accuracy indices does not represent the full spectrum of characteristics that would be encountered in real-world settings. In the current study, insufficient time in the school year precluded 44 children from participating, and an additional 54 were excluded for various reasons (e.g., bilingualism, participation in a French immersion program, other developmental disabilities), although some of these children received speech and language services in the schools. Beyond the exclusionary conditions, the participants included in the current study likely reflect the general population of kindergartners who live in the rural South, speak AAE or SWE, and attend public kindergartens. Recall that the return rate of consent forms was high at 81%, the sociodemographic profiles of the participants matched that of the larger school community, and the children with LI were identified with a dialect-appropriate referent standard.

Incorporation bias occurs when an index tool helps determine the participants' clinical status along with the reference standard. Although the children's scores on the index tools did not determine their clinical groupings, these scores served as the basis for the empirically derived cut scores. To test the generalization of the derived scores, an independent sample of participants is required. As shown in Table 1, TROLL scores were available for 72 other children (47 AAE and 25 SWE, 29 with LI and 43 TD) who participated in Oetting et al. (2016). These children attended the same kindergartens, spoke the same dialects, and presented the same psycholinguistic profiles as the current study sample, except that their data were collected between 2010 and 2013 and their PTONI and GFTA-2 scores were $\geq -1.2$ *SD*s of the normative mean. Using these 72 children as an independent test sample, the overall classification accuracy of the TROLL with a cut score of 89 was 65% (47/72 participants); sensitivity was 62% (18/29 participants), and specificity was 67% (29/43 participants). These findings demonstrate some generalizability of the 89 cut score for the TROLL, although sensitivity is lower than the 77% found in the study sample.

### *Conclusions*

Two conclusions can be drawn from the current study. First, the findings support the use of teacher ratings as measured by the TROLL for screening purposes when working with nonmainstream English-speaking kindergartners in the rural South. More specifically, clinicians who serve similar groups of children and who ask teachers to complete the TROLL at the end of kindergarten should consider using a cut score of 89 (or identify their own cut score using a local sample) to determine who should be referred for a language evaluation. This screening procedure will result in some misclassified cases, but the number of incorrect screens will likely not exceed those misclassified by the CCC-2, DELV-ST-II, or DIBELS. Second, the findings underscore the need for additional development and study of teacher rating instruments and other tools for screening purposes. Teacher ratings as measured by the TROLL correctly classified 77% of the children with LI studied here and 62% of the children in another group with LI as needing a language evaluation. These levels of sensitivity are not as high as the 80% recommended for screeners, and they need to be higher for clinicians who work in public schools. As demonstrated in the current work, future efforts in tool development should include detailed descriptions of the participant pools and recruiting strategies, analyses of multiple cut scores for different types of study samples, and methods of conduct that have been established to reduce reporting bias in diagnostic accuracy studies.

## Acknowledgments

This research was supported by Grants NIDCD RO1DC009811 (awarded to Janna Oetting, Janet McDonald, and Michael Hegarty) and RO1DC009811-03S1 (awarded to Janna Oetting). Appreciation is extended to Jessica Berry, Ryan Lee-James, Andy Rivière, Christy Seidel, Tina Villa, Sarah Williams, and other students who helped collect and input the data and the teachers, families, and children who agreed to participate in the study.

## References

**Antoniazzi, D., Snow, P., & Dickson-Swift, V.** (2010). Teacher identification for children at risk for language impairment. *International Journal of Speech-Language Pathology, 12,* 244–252.

**Bates, C., & Nettlebeck, T.** (2001). Primary school teachers' judgments of reading achievement. *Educational Psychology, 21,* 177–187.

**Bedore, L. M., Pena, E. D., Joyner, D., & Macken, C.** (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 14,* 489–511.

**Bishop, D. V. M.** (1998). Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology Psychiatry, 39,* 879–891.

**Bishop, D. V. M.** (2006). *CCC-2; Children's Communication Checklist–Second Edition*. San Antonio, TX: Pearson.

**Bishop, D. V. M., & Baird, G.** (2001). Parent and teacher report of pragmatic aspects of communication: Use of the Children's Communication Checklist in a clinical setting. *Developmental Medicine and Child Neurology, 43,* 809–818.

**Bishop, D. V. M., Laws, G., Adams, C., & Norbury, C.** (2006). High heritability of speech and language impairments in 6-year-old twins demonstrated using parent and teacher report. *Behavior Genetics, 36,* 173–184.

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

**Bishop, D. V. M., & Norbury, C. F.** (2002). Exploring the borderlands of autistic disorder and specific language impairment: A study using standardised diagnostic instruments. *Journal of Child Psychology and Psychiatry, 43,* 917–930.

**Bossuyt, P. M., Reitsma, J. B., Bruns, J. B., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . the STARD Group.** (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology, 277,* 826–832.

**Botting, N.** (2004). Children's Communication Checklist (CCC) scores in 11-year-old children with communication impairments. *International Journal of Language and Communication Disorders, 39,* 215–227.

**Botting, N., Conti-Ramsden, G., & Crutchley, A.** (1997). Concordance between teacher/therapist opinion and formal language assessment scores in children with language impairment. *European Journal of Disorders of Communication, 32,* 517–527.

**Brumbach, A. C., & Goffman, L.** (2014). Interaction of language processing and motor skill in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 57,* 158–171.

**Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R.** (2009). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language, Speech, and Hearing Services in Schools, 40,* 161–173.

**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.

**Cleveland, L. H., & Oetting, J. B.** (2013). Verbal –s marking by dialect and clinical status. *American Journal of Speech-Language Pathology, 22,* 604–614.

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

**Cunningham, D. D.** (2009). Relating preschool quality to literacy development. *Early Childhood Education Journal, 37,* 501–507.

**Dewey, E. N., Kaminski, R. A., & Good, R. H.** (2014). *DIBELS–Next national norms 2012–2013* (Technical report no. 17). Eugene, OR: Dynamic Measurement Group.

**Dewey, E. N., Powell-Smith, K. A., Good, R. H., & Kaminski, R. A.** (2015). *DIBELS–Next technical adequacy brief*. Eugene, OR: Dynamic Measurement Group, Inc.

**Dickinson, D. K., McCabe, A., & Sprague, K.** (2001). *Teacher Rating of Oral Language and Literacy (TROLL): A research-based tool*. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement, University of Michigan.

**Dickinson, D. K., McCabe, A. & Sprague, K.** (2003). Teacher Rating of Oral Language and Literacy (TROLL): Individualizing early literacy instruction with a standards-based rating tool. *The Reading Teacher, 56,* 554–569.

**Dollaghan, C. A., & Horner, E. A.** (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54,* 1077–1088.

**Dunn, L. M., & Dunn, D. M.** (2007). *Peabody Picture Vocabulary Test–Fourth Edition*. Bloomington, MN: PsychCorp.

**Ehrler, D. J., & McGhee, R. L.** (2008). *Primary Test of Nonverbal Intelligence*. San Antonio, TX: Pro-Ed.

**Elliot, J., Lee, S. W., & Tollefson, N.** (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Learning Skills–Modified. *School Psychology Review, 30,* 33–49.

**Ellis Weismer, S., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M.** (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 43,* 865–878.

**Feinberg, A. B., & Shapiro, E. S.** (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102,* 453–462.

**Fujiki, M., & Brinton, B.** (1984). Supplementing language therapy: Working with the classroom teacher. *Language, Speech, and Hearing Services in Schools, 15,* 98–109.

**Gijsel, M. A. R., Bosman, A., & Verhoeven, L.** (2006). Kindergarten risk factors, cognitive factors and teacher judgments as predictors of early reading in Dutch. *Journal of Learning Disabilities, 39,* 558–772.

**Gilmore, J., & Vance, M.** (2007). Teacher ratings of children's listening difficulties. *Child Language Teaching and Therapy, 23,* 133–156.

**Goldman, R., & Fristoe, M.** (2000). *Goldman-Fristoe Test of Articulation–Second Edition*. Circle Pines, MN: American Guidance Service, Inc.

**Good, R. H., Gruba, J., & Kaminski, R. A.** (2009). *Dynamic Indicators of Basic Early Literacy Skills–Next*. Longmont, CO: Cambium Learning Group.

**Gottlieb, J., Alter, M., Gottlieb, B. W., & Wishner, J.** (1994). Special education in urban America: It's not justifiable for many. *Journal of Special Education, 27,* 453–465.

**Gray, S., Kvalsvig, A., O'Connor, M., O'Connor, E., Incledon, E., Tarasuik, J., & Goldfeld, S.** (2017). Can a teacher-reported indicator be used for population monitoring of oral language skills at school entry? *International Journal of Speech-Language Pathology*. Advance online publication. https://doi.org/10.1080/17549507.2017.1294200

**Hauerwas, L. B., & Stone, C.** (2000). Are parents of school-age children with specific language impairments accurate estimators of their child's language? *Child Language Teaching and Therapy, 16,* 73–86.

**Jessup, B., Ward, E., Cahill, L., & Keating, D.** (2008). Teacher identification of speech and language impairment in kindergarten students using the Kindergarten Development Check. *International Journal of Speech-Language Pathology, 10,* 449–459.

**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.

**Leonard, L. B.** (2014). *Children with specific language impairment*. Cambridge, MA: MIT Press.

**Marsh, J., Pane, J., & Hamilton, L.** (2006). *Making sense of data-driven decision-making in education*. Santa Monica, CA: Rand.

**Martin, S. D., & Shapiro, E. S.** (2011). Examining the accuracy of teachers' judgments of DIBELS performance. *Psychology in the Schools, 48,* 343–356.

**McCabe, A., Boccia, J., Bennet, M. B., Lyman, N., & Hagen, R.** (2010). Improving oral language and literacy skills in preschool children from disadvantaged backgrounds: Remember, writing, and reading (RWR). *Imagination, Cognition and Personality, 29,* 363–390.

**Noonan, N. B., Redmond, S. M., & Archibald, L. M.** (2014). Contributions of children's linguistic and working memory proficiencies to their judgments of grammaticality. *Journal of Speech, Language, and Hearing Research, 57,* 979–989.

**Norbury, C. F., Nash, M., Baird, G., & Bishop, D. V. M.** (2004). Using a parental checklist to identify diagnostic groups in children with communication impairment: A validation of the Children's Communication Checklist-2. *International Journal of Language and Communication Disorders, 39,* 345–364.

**Oetting, J. B.** (2015). Dialect differences between African American English and Southern White English in children. In S. Lanehart (Ed.), *Oxford handbook of African American language* (pp. 512–518). New York, NY: Oxford University Press.

**Oetting, J. B., McDonald, J., Seidel, C., & Hegarty, M.** (2016). Sentence recall by children with SLI across two nonmainstream dialects of English. *Journal of Speech, Language, and Hearing Research, 59,* 183–194.

**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.

**Oetting, J. B., & Newkirk, B. R.** (2011). Children's relative clause markers in two nonmainstream dialects of English. *Clinical Linguistics and Phonetics, 25,* 725–740.

**Pearson, B. Z., de Villiers, P. A., Magaziner, K., Perisho, N., & Sunderland, K.** (2005). *Validation of the DELV-NR by language sample analysis*. Poster presented to the American Speech-Language-Hearing Association Annual Meeting, San Diego, CA.

**Pearson, B. Z., Jackson, J. E., & Wu, H.** (2014). Seeking a gold standard for an innovative, dialect-neutral language test. *Journal of Speech, Language, and Hearing Research, 57,* 495–508.

**Petscher, Y., Connor, C. M., & Al Otaiba, S.** (2012). Psychometric analysis of the Diagnostic Evaluation of Language Variation Assessment. *Assessment for Effective Intervention, 37,* 243–250.

**Philofsky, A., Fidler, D. J., & Hepburn, S.** (2007). Pragmatic language profiles of school-age children with autism spectrum disorders and Williams syndrome. *American Journal of Speech-Language Pathology, 16,* 368–380.

**Poll, G. H., Betz, S. K., & Miller, C. A.** (2010). Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language, and Hearing Research, 53,* 414–429.

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

**Qi, C. H., Kaiser, A. P., Milan, S., Yzquierdo, Z., & Hancock, T.** (2003). The performance of low-income African American children on the Preschool Language Scales-3. *Journal of Speech, Language, and Hearing Research, 43,* 576–590.

**Rodriguez, B. L., & Guiberson, M.** (2011). Using a teacher rating scale of language and literacy skills with preschool children of English-speaking, Spanish-speaking, and bilingual backgrounds. *Early Childhood Education Journal, 39,* 303–311.

**Roy, J., Oetting, J. B., & Moland, C.** (2013). Linguistic constraints on children's overt marking of BE by dialect and age. *Journal of Speech, Language, and Hearing Research, 56,* 933–944.

**Rutjes, A. W. S., Reitsma, J. B., Vandenbroucke, J. P., Glas, A. S., & Bossuyt, P. M. M.** (2005). Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry, 51*(8), 1335–1341.

**Seymour, H. N., Roeper, T. W., & de Villiers, J.** (2003). *Diagnostic Evaluation of Language Variation–Screening Test*. San Antonio, TX: PsychCorp.

**Seymour, H. N., Roeper, T. W., & de Villiers, J.** (2005). *Diagnostic Evaluation of Language Variation–Norm Referenced*. San Antonio, TX: PsychCorp.

**Shaw, R., & Shaw, D.** (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado state assessment program (CSAP)* [technical report]. Eugene, OR: University of Oregon.

**Spaulding, T. J.** (2010). Investigating mechanisms of suppression in preschool children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 53,* 725–738.

**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.

**Terry, N. P., Petscher, Y., & Rhodes, K. T.** (2016). Psychometric analysis of the Diagnostic Evaluation of Language Variation–Screening Test: Extension to low-income African American pre-kindergarteners. *Assessment for Effective Intervention, 42,* 1–10.

**Timler, G. R.** (2014). Use of the Children's Communication Checklist-2 for classification of language impairment risk in young school-age children with attention-deficit/hyperactivity disorder. *American Journal of Speech-Language Pathology, 23,* 73–83.

**Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, Z., Smith, E., & O'Brien, M.** (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40,* 1245–1260.

**U.S. Department of Agriculture.** (2017). *Child poverty: Rural poverty and wellbeing*. Retrieved from https://www.ers.usda.gov/topics/rural-economy-population/rural-poverty-well-being/child-poverty.aspx

**Vaisanen, R., Loukusa, S., Moilanen, I., & Yliherva, A.** (2014). Language and pragmatic profile in children with ADHD as measured by Children's Communication Checklist–Second Edition. *Journal of Logopedics, Phoniatrics, Vocology, 39,* 179–187.

**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,* 448–462.

**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.

**Whiting, P. E., Rutjes, A. W. S., Westwood, M. E., Mallett, S., & QUADAS-2 Steering Group.** (2013). A systematic review classifies sources of bias and variation in diagnostic test accuracy. *Journal of Clinical Epidemiology, 66,* 1093–1104.

**Whitworth, A., Davies, C., Stokes, S., & Blain, T.** (1993). Identification of communication impairment in preschoolers: A comparison of parent and teacher success. *Australian Journal of Human Communication Disorders, 21,* 112–133.

**Wiig, E. H., Secord, W. A., & Semel, E.** (2004). *Clinical Evaluation of Language Fundamentals Preschool—Second Edition*. San Antonio, TX: Harcourt Assessment.

**Williams, C.** (2006). Teacher judgements of the language skills of children in the early years of schooling. *Child Language Teaching and Therapy, 22,* 135–154.

**Victorino, K. R., & Schwartz, R. G.** (2015). Control of auditory attention in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 58,* 1245–1257.

**Zimmerman, I. L., Steiner, V. G., & Pond, R. E.** (2002). *Preschool Language Scale–Fourth Edition*. San Antonio, TX: PsychCorp.