JSLHR

Research Article

Seeking a Valid Gold Standard for an Innovative, Dialect-Neutral Language Test

Barbara Zurer Pearson,[a] Janice E. Jackson,[b] and Haotian Wu[a]

Purpose: In this study, the authors explored alternative gold standards to validate an innovative, dialect-neutral language assessment.
Method: Participants were 78 African American children, ages 5;0 (years;months) to 6;11. Twenty participants had previously been identified as having language impairment. The Diagnostic Evaluation of Language Variation—Norm Referenced (DELV–NR; Seymour, Roeper, & J. de Villiers, 2005) was administered, and concurrent language samples (LSs) were collected. Using LS profiles as the gold standard, sensitivity, specificity, and other measures of diagnostic accuracy were compared for diagnoses made from the DELV–NR and participants' clinical status prior to recruitment. In a second analysis, the authors used results from the first analysis to make evidence-based adjustments in the estimates of DELV–NR diagnostic accuracy.
Results: Accuracy of the DELV–NR relative to LS profiles was greater than that of prior diagnoses, indicating that the DELV–NR was an improvement over preexisting diagnoses for this group. Specificity met conventional standards, but sensitivity was somewhat low. Reanalysis using the positive and negative predictive power of the preexisting diagnosis in a discrepant-resolution procedure revealed that estimates for sensitivity and specificity for the DELV–NR were .85 and .93, respectively.
Conclusion: The authors found that, even after making allowances for the imperfection of available gold standards, clinical decisions made with the DELV–NR achieved high values on conventional measures of diagnostic accuracy.

Key Words: nonbiased testing, test validation, diagnostic accuracy, language sampling, cultural and linguistic diversity, African American English speakers

It has been recognized since the 1960s that standardized language tests in which scoring is based on the average performance of General American English (GAE) speakers are not generally appropriate for children who speak a different dialect of English. Taylor (1969) was among the first to articulate this view within the field of communication disorders, and it was subsequently highlighted in a special issue of *Language, Speech, and Hearing Services in Schools* (Taylor, 1972) on "Language and the Black Urban Child." This perspective has been kept in the forefront by researcher–activists such as Craig and Washington (2006), Seymour (2004; Seymour, Bland, & Green, 1998; Seymour & Seymour, 1979), and Stockman (1986, 1996, 2010), among many others. The caution has been reiterated in strong position statements from the American Speech-Language-Hearing Association (1983, 2003), which reaffirm that differences based on cultural dialects are not to be considered a basis for referral to speech or language services.

In brief, unless a test normed specifically for culturally and linguistically different (CLD) children is used, respondents may score below their aptitude because of "difference, not deficit." Thus, misdiagnosis may contribute to the overrepresentation of CLD children reported for speech-language and other special education services (Losen & Orfield, 2002; Robinson & Norton, 2012). At the same time, underrepresentation is also a concern: If diagnostic outcomes are disregarded for CLD populations, as they sometimes are, children who exhibit both difference and disorder will not be identified for services that could benefit them (Robinson & Norton, 2012; Seymour et al., 1998).

Rationale for an Innovative Test of Language

The need for a nonbiased test normed specifically for speakers of African American English (AAE) formed the basis of a National Institutes of Health (NIH) funding program issued in 1996. In response to the NIH mandate, a pair of dialect-neutral tests were created that avoided—to the extent

[a]University of Massachusetts Amherst
[b]DeKalb County Public Schools, GA
Correspondence to Barbara Zurer Pearson: bpearson@research.umass.edu

Editor: Janna Oetting
Associate Editor: Ron Gillam
Received April 17, 2012
Revision received January 12, 2013
Accepted August 9, 2013
DOI: 10.1044/2013_JSLHR-L-12-0126

Disclosure: *The first and second authors worked on the DELV project. They have no financial interests related to their work on the DELV.*

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

possible—elements of the language that had been shown to be different, or *contrastive,* between GAE and AAE. Rather than contrastive items, such as past tense verb endings and number agreement, whose usage patterns among AAE speakers are known to differ from mainstream expectations (Craig & Washington, 2006; Seymour et al., 1998), the new tests used only elements that were noncontrastive and that were theorized to be essential for adequate language development. These elements derive from universal principles of grammar (Chomsky, 1965; Roeper, 2007) and are realized the same way in both AAE and GAE. Thus, the new tests avoided many of the pitfalls of previous tests. Extensive piloting demonstrated that the diagnostic items of the new tests were indeed noncontrastive—that is, they elicited the same response patterns from both GAE first-dialect speakers and AAE-background children learning GAE as a second dialect (henceforth, *GAE-first* and *AAE-first,* respectively).

These tests, called the Diagnostic Evaluation of Language Variation—Norm Referenced (DELV–NR; Seymour, Roeper, & J. de Villiers, 2005) and the Diagnostic Evaluation of Language Variation—Screening Test (DELV–ST; Seymour, Roeper, & J. de Villiers, 2003b), did not test unique aspects of AAE. Except for a small dialect-identifier portion in the DELV–ST, the DELV tests focused on a carefully selected subset of structures common to both GAE and AAE. It is crucial to note that the tests were normed on a nationwide sample of 100% African American children, so their scoring would represent age-graded proficiency patterns for that population. The DELV–NR was subsequently re-normed on a general U.S. population. The two sets of norms were the same, and so the test was demonstrated to be appropriate for both AAE-first and GAE-first speakers.

Divergence of Scores

It could have happened that the dialect-neutral test would validate the use of standard diagnostic procedures for AAE-first children. If the level of agreement of diagnoses derived from existing practice in the field and diagnoses from the DELV–NR were high, the very exercise of creating the DELV tests would show them to be unnecessary. That, however, was not the case. Consistent with the motivation for its creation, the noncontrastive DELV–NR and standard diagnostic practice—by which children in the norming sample had been selected for speech-language services before recruitment—picked out somewhat different sets of children as showing language impairment (LI). According to a study of the DELV–NR norming sample (Jackson & Pearson, 2010), the discrepancy between clinical status at recruitment (CSaR) and the assignment of clinical status based on the DELV–NR was estimated at around 25% overall. There was less disagreement by percentage in identifying typical development (22%) than LI, for which the two methods (i.e., the DELV–NR and CSaR) diverged in 45% of the cases. Disagreement, though, tells nothing about whether the innovative test is more accurate than the established tests, or vice versa. To decide about the accuracy of the two diagnoses, one must measure them against an independent, trusted basis of comparison, a *reference standard*, or a so-called gold standard (Dollaghan & Horner, 2011).

From a practical point of view, however, the uniqueness of the innovative test—that intentionally differed from previous assessment instruments—increases the difficulty of finding a valid gold standard for it. Prior tests that were considered inappropriate or potentially biased for evaluating CLD *children* would also be inappropriate references for evaluating a new *test* for CLD children. Clearly, it would be counterproductive to attempt to establish the new test's concurrent validity and diagnostic accuracy by reference only to diagnoses made by the same tests of the same skills that had already been deemed inappropriate for the population.

Options for Establishing Test Validity in the Absence of a Satisfactory Gold Standard

How, then, does one establish test validity without a satisfactory gold standard? One alternative is to use a criterion-referenced approach, as indeed the DELV authors did in preparing for the task of creating the tests. In a series of technical papers and talks on the theme of "What Every 5- [or 3- or 7-] Year-Old Should Know," P. de Villiers, J. de Villiers, and Roeper (e.g., P. de Villiers, Roeper, & J. de Villiers, 1999; J. de Villiers & Roeper, 2001) inventoried the language acquisition and theoretical linguistics literatures for syntactic, semantic, and pragmatic constructions that represented appropriate language competence at each age between 4 and 9 years. Examples of such skills are giving exhaustive answers to double *wh*- questions (e.g., "Who bought what?") or deriving the meaning of a sentence containing a nonsense verb from its morphosyntax and argument structure. (See a fuller description in Seymour & Pearson, 2004.) The DELV authors also looked at children's language skills in context, noting especially skills that are necessary for success in school, such as how to ask pertinent questions or tell a cohesive story. A large set of candidate linguistic structures that addressed age expectations and academic requirements were extensively field tested on both AAE-first and GAE-first children. Items that showed the clearest development, discrimination, and dialect neutrality were incorporated into the final editions of the DELV–ST and DELV–NR (P. de Villiers & J. de Villiers, 2010) and were normed first on a 100% African American sample, as mentioned above, and then on a sample representing the general U.S. population. Thus, the test provided tools for clinicians to discover which essential age-appropriate skills children could demonstrate and which skills appeared to be absent from their repertories.

Early studies to test the validity of diagnoses made by the DELV–NR used concurrent language samples (LSs) for the purpose. Analysis of variance was used to compare DELV–NR and CSaR LI and typically developing (TD) groupings on the LS measures (P. de Villiers & J. de Villiers, 2010; Magaziner, Sunderland, Pearson, & P. de Villiers, 2008). The data demonstrated that the difference between TD and LI groups on LS measures was greater when the clinical status groups were determined by the DELV–NR scores than when the preexisting TD and LI (CSaR) groups

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

were compared. These analyses pointed to superior psychometric properties of the DELV–NR. However, if one follows the logic of Dollaghan and Horner's (2011, p. 1078) review of the process, those "pre-accuracy" studies were not sufficient for test validation. A true diagnostic accuracy study, Dollaghan and Horner argued, required case-by-case comparison to a reference standard.

Rationale for Using LSs for a Dialect-Neutral Reference Standard

The rationale for adopting LSs as the reference standard to establish validity and diagnostic accuracy in the absence of a conventional gold standard is strong. For CLD children, LSs have long been recommended as alternatives to using potentially biased standardized language tests (Stockman, 1996; Wyatt, 2002). In Stockman's (2010) words, LSs are more authentic, more readily accessible, and "implicitly [more] sensitive to linguistic differences [than standardized tests]" (p. 30). Typical LS protocols are designed to let the speakers choose their own words, but the protocols can be organized to provide opportunities that encourage specific pragmatic and even syntactic behaviors that are less likely to be evidenced completely spontaneously (Hadley, 1998). Overall, as articulated by Stockman (1996, 2010), LSs provide multidimensional, relatively unbiased information about a child's productive vocabulary, syntax, phonology, and pragmatics.

However, as also pointed out by Stockman (1996), LSs are not perfect. They show what a child chooses to say, not what she or he *can* say. Furthermore, they represent a static moment in a dynamic process. They are especially limited in not being able to capture rare structures that are unlikely to occur in a half-hour of speech. Even samples of 4 or 5 hr will not reliably tell about complex syntactic phenomena, such as long-distance movement of *wh*- elements nor semantic issues of implicatures about exhaustivity. These are among the structures which are tested in the DELV–NR and which give important information about the extent to which a child's grammar is developing normally. Another issue is that LSs have no comprehension component, and so they will not be expected to align with diagnoses for children whose major problems involve comprehension. The lack of a receptive component in LSs also limits the opportunity for children with severe articulation deficits, but not language or conceptual problems, to demonstrate their intact abilities in complex receptive syntax or the language of theory of mind that they can display on the DELV–NR. Indeed, precisely because they are not standardized, it is not immediately obvious how LSs can be used as a yardstick either to measure small increments in proficiency or to calculate diagnostic accuracy. Strategies to turn holistic impressions of the child's discourse into objective measures involve subjective decisions that may differ from practitioner to practitioner and according to the needs of the situation.

Nevertheless, despite LSs' restriction to productive measures and the relatively unpredictable, nonstandardized display of a child's language that they provide, proposals have been advanced to find diagnostic potential in them. Moreover, in recent years, new technology has facilitated the collection and storage of LSs, and new initiatives have encouraged open access to data collected for other purposes (Justice, Breit-Smith, & Rogers, 2010), such as the large volume of samples collected over many years by the Systematic Analysis of Language Transcripts (SALT) project (Heilmann, Miller, & Nockerts, 2010; Miller et al., 2005). In Justice et al.'s (2010) Clinical Forum, Heilmann and colleagues (2010) demonstrated how one can tailor a comparison group suited to the background of the children one needs to test. Using the metadata noted with each LS, one selects a subset from the large database matched to the target children in age, gender, or other relevant background variables, such as socioeconomic status or experience with a second language or second dialect. Heilmann et al. demonstrated their approach by collecting new LSs from 244 children with LI—children who had been identified in their schools as being on the caseloads of speech-language pathologists (SLPs). Then, they matched them on age and school type with 244 TD children from the database. With discriminant analysis, they showed that a selection of 10 LS measures generated by the SALT programs—such as mean length of utterance (MLU), number of different words (NDW), words per minute, and number of maze words—could identify the children's clinical status with about 80% agreement with the children's preexisting diagnosis, a level consistent with conventional standards for concurrence (Dollaghan & Horner, 2011).

Using LSs to Evaluate Diagnostic Accuracy

As mentioned above, one can use Miller et al.'s (2005) strategy for translating LS performance into a small set of consistent measures. A rank ordering of variables from a set of comparison LSs becomes the basis on which adequate or deficient levels of performance—passing or failing—can be determined for a particular group. In addition to local averages calculated in this way for specific groups of children, one can turn to published guidelines from authors who have used the same or similar measures. For example, Rice et al. (2010) presented average MLUs in words and morphemes for TD children and for those with LI at 6-month intervals from 3 to 9 years. Hewitt, Hammer, Yonte, and Tomblin (2005) contributed averages for NDW in 50 utterances and the Inventory of Productive Syntax (IPSyn; Scarborough, 1990) for 54 kindergarteners from Tomblin et al.'s (1997) corpus. Similarly, Blake, Quartaro and Onorati (1993) argued for a measure of complex syntax, which they calibrated to MLU levels. To find an appropriate reference group, one can use the database samples, as Heilmann et al. (2010) did, to compare the LSs of two groups.

Once a pass–fail designation has been assigned, the LS can also be used in comparisons with a range of other measures in within-subject analyses. With a criterion for testing "positive" as opposed to "negative" for LI, LSs can be used in the framework of the conventional cross-table for determining diagnostic accuracy. Given the limitations of LSs described at the outset, one cannot expect them to be an "untarnished gold standard" (P. de Villiers & J. de Villiers,

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

2010, p. 240), so we will not expect to agree with 100% of their diagnoses. Still, they can provide useful descriptive information about diagnoses for a particular child, and they can be used to compare relative diagnostic accuracy measures for different instruments used with the same group.

Furthermore, the field of epidemiology provides proposals to compensate for LSs' limits as a reference standard. For medical decision making, where the disease status of individuals is seldom certain and available reference standards are often imperfect, several strategies have been proposed to combine data from two reference standards, using a second imperfect test as a "resolver test" in ways that attempt to compensate for potential biases in each of them (Green, Black, & Johnson, 1998; Hawkins, Garrett, & Stephenson, 2001).

Standard Diagnostic Accuracy Measures

The most common metrics for describing a test's case-by-case agreement with the reference standard are *sensitivity* and *specificity* (Dollaghan & Horner, 2011; Oetting, Cleveland, & Cope, 2008; Plante & Vance, 1994). In comparing outcomes of the new test to the reference provided by the gold standard, one asks whether the new assessment will be sensitive enough to pick out all of the children in the sample the gold standard labels as having LI but also specific enough to avoid false positives—that is, children who are identified by the test as having LI but are TD according to the gold standard. In the case of CLD children, there have historically been problems with both false positives (i.e., identifying TD children as having LI) and false negatives (identifying children with LI as TD), but the more crucial issue has been a larger proportion of false positives than among mainstream children (Losen & Orfield, 2002). With a test specifically designed for CLD children, one hopes to reduce the number of false positives without increasing the number of false negatives.

Furthermore, epidemiologists rarely rely solely on sensitivity and specificity (see, e.g., Pewsner et al., 2004). They also apply the framework that they used to calculate sensitivity and specificity for a test to derive its positive predictive power (PPP) and negative predictive power (NPP) and positive and negative likelihood ratios (LR+ and LR–). These are both measures that give more precise guidance about how useful a particular diagnosis based on a given test result will be in practice (Dollaghan & Horner, 2011; Sonis, 1999).

Cautions for Measures of Diagnostic Accuracy

Heilmann and colleagues (2010) cautioned, however, that measures of diagnostic accuracy, such as sensitivity and specificity calculations, risk being circular—depending on one's gold standard. They pointed out that if one is validating a test that examines only a narrow range of skills (e.g., productive morphosyntax), with a gold standard that had ranked children according to their performance on the same skills in the first place, the validity of the reported sensitivity is limited and may not be informative about the child's language status with respect to other language skills. Therefore, according to Heilmann et al., some tests identified as having high sensitivity and specificity (Plante & Vance, 1994; Spaulding, Plante, & Farinella, 2006) may have achieved high levels in a research study without showing their success with the variety of cases one sees in treatment in a typical school or clinic. In Heilmann et al.'s words, "the [sensitivity and specificity] results cannot be generalized to a heterogeneous group of children with LI" (p. 92), many of whom will have disorders in domains other than common markers of LI, such as productive morphosyntax. Heilmann et al. proposed that LSs are precisely the multidimensional assessment needed, and these authors reported a study that evaluated clinical decisions based on the LS measured against a heterogeneous reference standard. As mentioned above, they found that their LS analyses agreed with their gold standard—in this case, prior diagnoses—in about 80% of their clinical decisions: They identified about 85% of the TD children (specificity) and 77% of those with impairment (sensitivity; p. 90), levels of agreement that are close to conventional standards for convergence, often set at .80 (cf. Spaulding et al., 2006).

However, Heilmann et al.'s (2010) own gold standard, "children currently receiving services," is also somewhat questionable. Tomblin et al.'s (1997) epidemiological study, for example, tells us that "receiving services" as a yardstick has low sensitivity. In fact, 73% of the children in Tomblin et al.'s study who were identified as having LI were not receiving services (i.e., were false negatives; pp. 1245, 1256). Their study does not tell us how many children receiving services were determined to be TD (false positives), so it is silent on whether "provision of services" has reasonable specificity. On the other hand, Heilmann et al.'s decision to use children on SLPs' caseloads as a reference standard for LS is not so different from techniques used in other diagnostic accuracy studies. One prominent example of such studies by Plante and Vance (1994) used the categorization of children "referred from clinics and schools … made on the basis of performance on standardized tests and clinical judgment of impaired language … [and who] were all receiving services" (p. 18) as "the true categorization" (p. 20), whereas, in fact, there was no independent validation of the clinical categories they used. As we have already seen, these existing diagnoses will be especially limited for CLD children if the reference standard is based primarily on clinical markers that are known to be ambiguous and/or ineffective for speakers of nonmainstream varieties (P. de Villiers & J. de Villiers, 2010; Seymour et al., 1998; Stockman, 2010).

The Present Study

The goal of this study was to evaluate the validity of an innovative, dialect-neutral test using alternative approaches for establishing diagnostic accuracy. One approach was to use a comprehensive but somewhat less than "gold" standard (here, LS) to find relative levels of convergence. We proposed to first compare the diagnostic accuracy of the dialect-neutral test and of prior diagnoses for the same CLD children in order to determine whether the DELV–NR was an improvement over prior methods of assessment for these children. A second approach was to strengthen the validity of the accuracy measures derived from LSs by using LSs in combination with another, also imperfect reference standard. Specifically, we tested the following hypotheses:

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Hypothesis 1

Using benchmarks based on levels of variation in measures derived from the children's own LS (e.g., MLU, NDW, a pragmatics composite, etc.), we predicted that, for this group, standard diagnostic accuracy measures using LS profiles as the gold standard would be higher for clinical status assigned by the DELV–NR than for clinical status at recruitment (i.e., CSaR).

Hypothesis 2

Stronger evidence of diagnostic accuracy will be attained if two measures are used together to strengthen the reference standard, using the second reference standard as a resolver, that is, to resolve discrepancies between the first reference standard and the test being evaluated. For this purpose, we proposed to use the results of the first phase of the analysis to pinpoint where a candidate resolver, CSaR diagnoses, were likely to be more accurate and less accurate, and thus where they might be a reasonable source of corroboration for the test being evaluated, referred to as the *index test*. We hypothesized that when more of the available evidence was used for the analysis, even from an imperfect reference standard, sensitivity and specificity measures for the DELV–NR would be more accurate than from one measure alone.

Finally, we suggest that if the DELV–NR shows improved diagnostic accuracy compared with the CSaR, this study would help validate the innovations in the nontraditional test. It would also add to evidence supporting the identification of LI among CLD children using assessments based on a diverse set of language elements (Kohnert, 2008; Owen & Leonard, 2006; Seymour & Pearson, 2004; Tager-Flusberg & Cooper, 1999).

Method

Participants

Participants were 80 children, ages 5;0 (years;months) to 6;11, who were a subset of the preliminary African American standardization sample for the DELV–NR. Participants were all African American, and they lived in predominantly African American communities according to the U.S. Bureau of the Census (2000). They exhibited a range of language variation as measured by the DELV–ST, Part 1: Language Variation Status (Seymour et al., 2003b). The DELV–ST identifies three levels of variation relative to GAE in the responses of the test-takers: (a) no difference from GAE, (b) some difference from GAE, or (c) strong difference from GAE. Three-quarters of the current sample were identified as showing some or strong difference from GAE, whereas the others' DELV–ST Language Variation Status responses were more similar to responses given to those items in the standardization phase by GAE speakers.

Unlike the full standardization sample, which carefully represented census data about the population, children with LI were overrepresented in this subset so that there would be sufficient variability in the LI group for statistical comparisons. Therefore, the cohort had, by design, 25% children with LI. The diagnosis of LI for inclusion in the study was made by certified SLPs on the basis of performance on standardized tests and clinical judgment of impaired language corroborated by the classroom teacher. The standardized tests used were those in most common use—for example, the Clinical Evaluation of Language Fundamentals, Oral and Written Language Scales, Test of Language Development, and so forth.[1] All children with LI were receiving services for impaired language at the time of the study; five were also receiving speech services, but children who previously had been identified with only phonological impairment were not included in this study. Sixty participants were considered to be TD by their parents, teachers, and clinicians. No children had conditions such as hearing impairment or autism that might have contributed to a diagnosis of LI.

Audiotapes of two TD 5-year-olds were not sufficiently audible, and so there were 78 children in the final sample. Children were from all regions of the United States in approximately the same proportions as in the U.S. census of African Americans (U.S. Bureau of the Census, 2000): 50% from the South, 34% from the North Central region, 12% from the Northeast, and 4% from the West. Five levels of parent education (PED) were also matched to the proportions of African Americans in the census. The demographics of the sample for whom DELV–NR and LSs were analyzed are shown in Table 1.

Gender was somewhat unbalanced. Although it was carefully balanced for the TD children, there were more boys than girls among those previously identified as having LI (cf. Leonard, 1998) and more girls than boys among the younger children. Therefore, we used a multivariate analysis of variance (MANOVA) to test the effects of gender, Gender × Age interactions, and Gender × Clinical Status interactions.

Materials

The DELV–NR

The DELV–NR[2] (Seymour et al., 2005) is a norm-referenced test consisting of 125 items in three sections: Syntax, Semantics, and Pragmatics.[3] All of the items avoided constructions known to differ between GAE and AAE—so,

[1]During piloting, the DELV authors were able to confirm that each diagnosis of LI was supported by scores from a well-known standardized test. The current authors no longer have access to the Psychological Corporation records of the tests and, thus, cannot list the editions used. However, the authors can attest that the list represents the versions of the tests that were most current in the early 2000s, when the participants were recruited.

[2]More information about the makeup and rationale for the DELV–NR is available in Seymour and Pearson (2004). Note also that the DELV–NR should be distinguished from the Diagnostic Evaluation of Language Variation—Criterion Referenced (Seymour, Roeper, & J. de Villiers, 2003a), with which it is sometimes confused. The items are the same, but the Criterion Referenced version was not normed with a norming sample and, as the name implies, gives scores only in broad categories, not in standard scores.

[3]There is also a Phonology subtest, which is scored separately. Because there was no focus on phonology in the current study, results for phonology are not discussed here.

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 1. Participant characteristics by clinical status.

Group	Age (*SD*) (yrs;mos)	Gender		Parent Education Level[a] (*SD*)	Some or strong difference from GAE (per DELV–ST LVS) (%)
		Males (%)	Females (%)		
Total Sample (*N* = 78)	6;0 (.5)	55	45	3.4 (.1)	78
TD (*n* = 57)	6;0 (.5)	50	50	3.4 (.1)	72
LI (*n* = 21)	5;11 (.5)	67	33	3.4 (.1)	95

[a]1 = *fewer than 8 years*; 2 = *fewer than 12 years*; 3 = *high school degree*; 4 = *some college*; 5 = *college degree*. GAE = General American English; DELV–ST = Diagnostic Evaluation of Language Variation—Screening Test, Part 1, Language Variation Status; TD = typically developing; LI = language impairment.

for example, there were no contrastive morphosyntax items on the DELV–NR. The items were approximately evenly divided between comprehension and production. Comprehension items evaluated knowledge of structures like complex *wh*- clauses, passives, and quantifiers, and a dynamic process, called *fast-mapping*, that tested vocabulary. Production items tested *wh*- question asking, article usage, short narrative, and communicative role-taking. Each section of the test gives a scaled score (M = 10), and there is an overall standard score (M = 100, SD = 15).[4]

LSs

Care was taken to create a multifaceted LS, so that there would be several dimensions (e.g., complex syntax, pragmatics, and lexical ability, as well as Miller et al.'s, 2005, standard measures such as MLU) along which to evaluate clinical decisions from the nontraditional assessment. The protocol for the LS was designed by the DELV authors, incorporating best practices for collecting LSs (Hadley, 1998; Paul, Tetnowski, & Reuler, 2007). The target was 100 child utterances, elicited through an informal script that included 5–10 min of conversation as well as prompts for a short personal narrative, a story based on a four-picture sequence, and some exposition. The instruction packet also contained a set of eight pictures of people representing professions (teacher, pilot, etc.) for eliciting the habitual present, plus a box of crayons with one color missing, and three pictures for the child to color according to the examiner's directions, some of which could not be fulfilled (i.e., "color something red," when no red crayon was provided).

Procedure

Collecting the Data

The DELV–NR and the DELV–ST unpublished standardization editions were administered by certified SLPs according to directions now published in the test manuals. However, examiners did not score the tests (because the scoring guidelines had not yet been developed). Children were tested individually in a quiet room in their schools. Thirty-two children had African American examiners, and 46 had non–African American examiners. The same examiners also audio-recorded LSs from the children within 3 weeks of the DELV testing. Although Dollaghan and Horner (2011) recommended against having the same person administer both the reference test and the comparison test, the de facto blinding of the examiners avoided the potential problems against which they cautioned.

Creating the Transcripts

The audiotaped LSs were transcribed in SALT Student Version 7.0 (Miller & Chapman, 2002), and from there, they were imported into Computerized Profiling (Long, Fey, & Channell, 2003). For each child, there was a hand-annotated transcript in spreadsheet format, a SALT file, a Computerized Profiling file, and the associated reports as described below.

Scoring the Tests

The DELV–NR tests were scored by the researchers according to the guidelines on the record forms and supplementary administration booklet (Seymour et al., 2003b, 2005). The criterion for LI for these analyses was a DELV–NR score that was <1.5 *SD* below the mean. Although the DELV–NR technical manual recommends that –1 *SD* be used as the most sensitive cutoff, that recommendation was made on the basis of prior diagnoses (CSaR). Even though there was no more than 15% overlap between the participants in the study reported in the manual and those in the current study, the methods of participant selection were very similar. Furthermore, because the CSaR was one of the index measures evaluated in the current study, we did not use the results of that study for our cutoff, and we redid the analyses using LSs as the gold standard. At –1 *SD,* sensitivity and specificity at 0.69 were unacceptably low, whereas at –1.5 *SD,* a better balance was exhibited between the measures, as shown in the Results section.

Coding the LS measures

Two sets of variables were coded. The first set included the following four variables from SALT and Computerized Profiling: number of utterances in the transcript, NDW in the first 50 utterances (NDW-50), MLU in words (MLU_w), and syntactic complexity (Blake et al., 1993).

[4]To avoid circularity, scoring was based on published norms from the standardization sample for the general American population, not the exclusively African American sample used for the NIH contract, which included many of these participants.

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

The second set included the following two variables from hand-coding: (a) IPSyn Sentence Structure subtotal and (b) pragmatics coding (see the two subsections that follow).

IPSyn Sentence Structure subtotal. Although the IPSyn (Scarborough, 1990) appears to have limited diagnostic value for 6-year-olds, who are older than the age range for which it was designed (Oetting et al., 2010), we included the Sentence Structure subscale to help confirm that our protocol was successful in eliciting a broad range of syntactic forms.

Pragmatics coding. Pragmatics coding was based on elements from Westby (1999; also used by Burns, P. de Villiers, Pearson, & Champion, 2012) and incorporated measures of discourse complexity and fluency. The pragmatics composite totaled 18 points and consisted of the discourse score and the literate language score. The *discourse score* (up to 8 points) coded fluency in terms of the following criteria: (a) how much prompting the child required in two specific segments of the transcript (ranging from 3 [no prompts] to 0 [two or more prompts]); (b) referential clarity (e.g., using adjectives or adjective phrases, relative clauses, or proper names) while telling a story from the pictures provided (0–1 point); and (c) references to the mental states of characters in the stories (0–1 point). A *literate language* score (up to 10 points) was built from five specific syntactic/semantic constructions based on Westby (1999): elaborated noun phrase, temporal adverbial clause, causal adverbial clause, indirect speech, and a mental state verb with a complement clause. Following the scoring principle from the IPSyn (Scarborough, 1990), one occurrence of the construction was awarded 1 point, and two or more occurrences were awarded 2 points.

Establishing Criterion Measures: "Standardizing" the LS Measures

Although there are normative guidelines for many of the measures used in this study—for example, Rice et al. (2010) for MLU, and Hewitt et al. (2005) for NDW-50 and IPSyn sentences—norms were not on the same scale, and they did not represent the average performance of this particular set of children. Therefore, before the LS measures could be used as a reference standard against which to measure the results from the DELV–NR and CSaR for this group, LS quality was operationalized. We calculated a *z* score for each variable for each child based on the mean and *SD* for the group. The *z* scores for the individual variables and three composite scores—for the syntax and semantics measures, for the pragmatics measures, and for scores overall—were tabled, and LS profiles for individual children were made from the tables.

To use the LS scores for diagnostic accuracy analyses, we gave each LS profile the category assignment of "passing" if *z* scores for all variables were >1 *SD* below the mean and "failing" if at least one *z* score was below –1. For 16 of the LSs who were assigned to the "failing" category, all of their LS variables were below the cutoff. However, seven children with very low scores in only one variable were included in the "failing" category following the observation by ourselves and consultants that children could be receiving services based on a low performance on one skill and an average performance on other skills.

Reliability

Two listeners transcribed 10% of the tapes independently. Counting each word that differed as the numerator and the total number of words as the denominator, level of agreement for the transcriptions averaged 88% (range = 85% to 92%). Disagreements were resolved by a consensus reached between the first author and the supervising transcriber, who was an advanced PhD candidate and a native speaker of AAE.

Analyses

Descriptive Statistics and Control Analyses

Means and *SD*s were derived for the full group for all LS measures and, for comparison purposes, for the TD and LI groups. To ensure that no major demographic factors confounded the results, we conducted a MANOVA in which gender, region, age in years, and PED were control variables, and the LS measures and DELV–NR scores were the dependent variables. We also included CSaR as an independent variable for the LS measures; in doing so, we were able to look for interactions of clinical status with age or gender (as discussed above in the Participants section). In addition, because examiner ethnicity could potentially influence children's volubility (Pearson, Velleman, Bryant, & Charko, 2009), we included examiner ethnicity as well as sample length (less than or greater than 100 utterances) among the control variables.

Study of Diagnostic Accuracy

For Hypothesis 1, we compared clinical status based on the profile of LS *z* scores for each individual, first to DELV–NR outcomes and then to CSaRs, in order to identify true and false positives and true and false negatives among the individuals classified as "passing" and those classified as "failing." Sensitivity, specificity, PPP and NPP, and LR+ and LR– were calculated according to the standard procedure (EpiMax, n.d.). The diagnostic accuracy measures relative to LSs for the DELV–NR and CSaR were compared statistically. For Hypothesis 2, we created a third cross-table to explore the available evidence that would permit us to resolve discrepancies empirically between the LSs and DELV–NR.

Results

Descriptive Results

Dependent Variables

As noted in the Method section, the measures that summarized the information in the LSs were lexical, syntactic, and pragmatic: sample length, NDW-50, MLU_w, IPSyn sentences, syntactic complexity, and a pragmatics composite made up of the discourse and literate language scores. Means and *SD*s for the LS measures, and the DELV–NR

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

standard scores for the whole group and for the TD and LI subgroups, are shown in Table 2.

Control Variables

Region, PED level, age, gender, examiner ethnicity, and sample length were entered into four MANOVAs because there were limited degrees of freedom for the 256 cells of an omnibus 4 (Region) × 4 (PED Level) × 2 (Age in Years) × 2 (Gender) × 2 (Examiner Ethnicity) × 2 (Sample Length). Because there were four MANOVAs, a Bonferroni correction set the significance level at .01 (.05/4). There were no substantive results for any of the potential confounds.

Hypothesis 1: Relative Diagnostic Accuracy

First-Pass Diagnostic Accuracy of the DELV–NR

Overall, 55 children had LS profiles within the average range or above (>–1 *SD*), and 23 were categorized as "failing" (i.e., those who had *z* scores below average [<–1 *SD*]). Table 3 is the standard contingency table that shows the concordance of diagnostic categories of the 78 LS profiles and DELV–NR scores.

Using the LS profiles as a gold standard, specificity of the DELV–NR was 0.89, and sensitivity was 0.65—that is, one would expect the new test to identify as TD 89% of those identified by the LS as TD (*true negatives*) and 65% of those the LS identified as LI (*true positives*). In reading the table row-wise to determine predictive power, one can see that 15 of the 21 children identified as positives (LI) by the DELV–NR were also positive according to the LS profile (i.e., true positives, not false positives), giving a PPP of 0.71. Predicting true negatives as opposed to false negatives, NPP would be 0.86. The LR+ was 6.0, and the LR– was 0.4. When interpreted according to standard guidelines (Sonis, 1999), the LR+ confirmed that for an individual who tested positive for LI on the DELV–NR, the probability of having LI increased from 29% (prevalence of LI in the sample) to approximately 65%.

Diagnostic Accuracy of CSaR

The analogous information for the CSaR is shown in Table 4. Specificity of the CSaR relative to the LS profiles was similar (0.83) to that of the DELV–NR, but sensitivity was a good bit lower (0.48) than for the DELV–NR scores. For children who had been identified as TD (and were not receiving services), the probability of a true negative (NPP) was 0.79. However, for PPP, only 11 of 20, or 55%, of the CSaR LI children were true positives, and 45% were false positives.

The LR ratios (LR+ = 2.9 and LR– = 0.6) were closer to 1—that is, "uninformative" (Sonis, 1999)—than were the analogous ratios for the DELV–NR. According to the guidelines given by Sonis (1999), the LR+ of 2.9 changed the CSaR prediction of LI from 29% (prevalence) to 50%, about the same as the toss of a coin.

Comparing Diagnostic Accuracy of the DELV–NR Relative to CSaR

The 95% confidence intervals (CIs) in Tables 3 and 4 show that the outcomes of the two analyses were not statistically distinct at the .05 level. Specificity, NPP, and LR– were similar for the DELV–NR and CSaR. By contrast, accuracy measures that relate to the identification of LI (sensitivity, PPP, and LR+) were substantially higher for the DELV–NR than for the CSaR, but they were not outside the 95% CIs: CSaR measures were at the lower bound of the 95% CIs for the DELV–NR, and the DELV–NR measures were at the upper bound of the CSaR 95% CIs.

When the DELV–NR and CSaR were compared to each other case by case, there were 14 discrepant cases, or 18% disagreement. LSs were concordant with the DELV–NR and not the CSaR in 11 of the 14 cases. A two-tailed binomial test suggested that 11 of 14 was not a chance result ($p = .057$), although it also just missed significance at the .05 level.

Hypothesis 2: Using CSaR PPP as a Resolver

In Table 3, we evaluate DELV–NR diagnoses using LS as the reference standard. We have acknowledged that LSs are not a perfect gold standard, but we have no gold standard against which to compare LSs, so their own sensitivity and specificity are unknown. Thus, we have a relatively high degree of confidence in the conclusions of the LSs, but we know that some of the LS diagnoses were not "true." How, then,

Table 2. Language sample measures, *M*s, and *SD*s for the total sample and by clinical status (per the Diagnostics Evaluation of Language Variation—Norm Referenced [DELV–NR]).

	Total sample *N* = 78		TD group *n* = 57		LI group *n* = 78	
	M	*SD*	*M*	*SD*	*M*	*SD*
Total utterances	175	62	168	61	196	61
NDW-50	100	25.7	107	24.8	82	18
MLU_w	4.2	1.1	4.5	1.1	3.5	0.8
IPSyn Sentences	28.4	5.8	29.3	5.4	25.9	6.3
Syntactic complexity	3.1	0.8	3.2	0.9	2.8	0.6
Discourse (of 8)	5.4	2.0	6	1.5	3.8	2.1
"Literate language" (of 10)	6.2	2.7	6.8	2.6	4.7	2.4
DELV–NR SS	89.6	16.5	97	12.6	70	6.4

NDW-50 = number of different words in 50 utterances; MLU_w = mean length of utterance in words; IPSyn Sentences = Index of Productive Syntax Sentence Structure Scale; SS = standard score.

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 3. First-pass diagnostic accuracy cross-table for the DELV–NR, using language sample profiles as the reference standard.

Variable	Language sample gold standard LI (*n* = 23)	Language sample gold standard TD (*n* = 55)	Predictive power [95% CI]
DELV–NR LI group (*n* = 21)	a. True positives: 15	b. False positives: 6	PPP = a/(a + b) = 0.71 [0.52, 0.86]
DELV–NR TD group (*n* = 57)	c. False negatives: 8	d. True negatives: 49	NPP = d/(d + c) = 0.86 [0.79, 0.91]
Sens. or Spec. ratio [95% CI]	Sens. = a/(a + c) = 0.65 [0.46, 0.85]	Spec. = d/(b + d) = 0.89 [0.81, 0.97]	
LR [95% CI]	LR+ sens./1 − spec. = 6.0 [2.70, 13.40]	LR− spec./1 − sens. = 0.4 [0.22, 0.69]	

Note. *N* = 78 for total sample. PPP = positive predictive power; NPP = negative predictive power; Sens. = sensitivity; Spec. = specificity; CI = confidence interval [lower bound, upper bound]; LR+ = positive likelihood ratio; LR− = negative likelihood ratio.

should we evaluate cases in which the LSs and the DELV–NR disagreed? According to the figures in Green et al. (1998, p. 378), to the extent that LS sensitivity would be less than 100 (if we could measure it), accepting all of the LS diagnoses as truth risks biasing the DELV–NR diagnostic accuracy measures downward. But how strong is the likelihood that the DELV–NR is the true diagnosis, even when it disagrees with the LS? Another measure available for all of the children was CSaR, which, although flawed as a diagnostic, may nonetheless provide some guidance as a resolver in LS/DELV–NR discrepant cases. Step 1 of the current analysis extended our knowledge of the validity of the CSaR, showing us that the NPP of the original diagnoses was adequate (0.8), but PPP (0.55) for this CLD sample was very low (see Table 4). Thus, a refinement of the CSaR's diagnostic accuracy could give an indication of how much more confidence in DELV–NR scores is warranted when they agreed with CSaR in general, including the discrepant cases in which the DELV–NR disagreed with the LS, a procedure called *discrepant resolution* (Hawkins et al., 2001).

Therefore, as shown in Table 5, we compared the DELV–NR outcomes to those of CSaR. Starting in Cell c, with DELV–NR false negatives (Dlv–/LS+), it turns out that, of eight cases, two CSaR positives disagreed with the Dlv–; therefore, in the face of two diagnoses of impairment, there was no support for calling those two DELV–NR negatives "unimpaired." However, six of the Dlv–/LS+ cases were also CSaR negatives (CSaR–). The NPP of CSaR (from Table 4) tells us that 80% of CSaR– cases are likely negatives, so we had some support for moving five of the six Dlv–/CSaR– cases to Cell d: true negatives. Likewise, in Cell b of Table 5 (Dlv+/LS–), CSaR disagreed with the Dlv+ in two of the cases, whereas it agreed with the DELV–NR+ in the other four cases. Because the PPP of the CSaR was 0.55, we had some support for moving half of the Dlv+/CS+ to Cell a: true positives. We grant, as argued by critics of discrepant resolution (summarized in Hawkins et al., 2001), that we were ignoring the cases of CSaR/DELV–NR disagreement in Cells a and d, but our knowledge of the CSaR's diagnostic accuracy did not give us confidence to have it override *two* opposite diagnoses.

Adding CSaR as a provisional reference standard resolver test yielded the results shown in Table 5. The reevaluated sensitivity of the DELV–NR was 0.85, and

Table 4. Diagnostic accuracy cross-table for the participants' clinical status at recruitment (CSaR), using language sample profiles as the reference standard (*N* = 78).

Variable	Language sample gold standard LI (*n* = 23)	Language sample gold standard TD (*n* = 55)	Predictive power [95% CI]
CSaR LI (*n* = 20)	a. True positives: 11	b. False positives: 9	PPP a/(a + b) = 0.55 [0.35, 0.73]
CSaR TD (*n* = 58)	c. False negatives: 12	d. True negatives: 46	NPP d/(d + c) = 0.79 [0.69, 0.90]
Sens. or Spec. ratio [95% CI]	Sens. = a/(a + c) = 0.48 [0.27, 0.68]	Spec. = d/(b + d) = 0.83 [0.74, 0.90]	
LR [95% CI]	LR+ Sens./1 − Spec. = 2.90 [1.40, 13.40])	LR− Spec./1 − Sens. = 0.60 [0.40, 0.94]	

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 5. Second-step diagnostic accuracy cross-table for the DELV–NR, using positive and negative predictive power of clinical status at recruitment as a resolver (N = 78).

Variable	Language sample gold standard LI (n = 23)	Language sample gold standard TD (n = 55)	n	Predictive power [95% CI]
DELV–NR LI	a. True positives: 15	b. False positives: 6		
CSaR resolver		2 CSaR negative 4 CSaR positive PPP: 0.55 × 4 = 2		
Revised DELV–NR LI	a_R. True positives: 15 + 2 = 17	b_R. False positives: 6 – 2 = 4	21	PPP = a/(a + b) = 0.81 [0.63, 0.91]
DELV–NR TD	c. False negatives: 8	d. True negatives: 49		
CSaR resolver	2 CSaR positives 6 CSaR negatives NPP: 0.8 × 6 = 5			
Revised DELV–NR TD	c_R. False negatives: 8 – 5 = 3	d_R. True negatives: 49 + 5 = 54	57	NPP = d/(d + c) = 0.95 [0.88, 0.98]
	Sens. = a/(a + c) = 0.85 [0.67, 0.95] LR+ Sens./1 – Spec. = 12.30 [4.70, 32.00]	Spec. = d/(b + d) = 0.93 [0.87, 0.97] LR– Spec./1 – sens. = 0.16 [0.06, 0.46]		

specificity was 0.93, both well within conventional standards for the measures. The LR+ of 12.3 also indicated that the prediction of LI from the DELV–NR increased from 29% (prevalence) to over 80%.

Discussion

Overall, our hypotheses were upheld. Diagnostic accuracy relative to LS profiles was substantially higher for the DELV–NR than for the CSaR, although statistical support for the difference just missed significance at the .05 level. With application of evidence supporting a percentage of diagnoses of the CSaR as capable of corroborating the DELV–NR diagnoses, the diagnostic accuracy contingency table was reanalyzed. Using the updated figures, sensitivity and specificity of the DELV–NR exhibited adequate to high levels for diagnostic accuracy.

Relationship to Previous Studies of LS Measures

Average values on the LS measures for this group were consistent with reports from other studies. As expected, scores for the entire group, which included 26%–29% children with LI, were low, but, as indicated in the second column of Table 2, averages for the TD group alone were comparable to published averages. The 4.5 average MLU for the TD 5- and 6-year-olds in this study matched Rice et al.'s (2010) MLU_w of 4.57 for TD children, ages 6;0–6;5. Similarly, the IPSyn Sentence Structure score mean of 29.3 for the TD group was almost identical to the 29.4 average for 6-year-olds in Hewitt et al.'s (2005) study, and the syntactic complexity and MLU measures were in the same relationship as for Blake et al.'s (1993) participants. Only NDW-50 was lower than published values (Hewitt et al., 2005), but the gap does not indicate that the present group was atypical because the literature (summarized in J. de Villiers, 2004) led us to expect vocabulary scores to be lower than norms for mainstream participants.

The comparison between TD participants versus the total group in Table 2 reminds us that this sample was not selected to be a representative one. The goal of the analysis was not to find LS values that would translate to other populations but to demonstrate that the DELV–NR, which gives such different information about a child's linguistic functioning than other tests of language, would be as good as or better than current diagnoses at matching the prediction of LSs within the group being analyzed.

Rationale for Using CSaR PPP and NPP to Reevaluate Accuracy Measures of the DELV–NR

If one devises an innovative method to make the determination of LI for CLD speakers—as, for example, the DELV–NR has done—the problem arises of how to judge the diagnostic accuracy of the nontraditional test without an independent, trusted basis of comparison whose dialect neutrality has been demonstrated. For this study, we enlisted LS measures as a basis of comparison. In doing so, we followed a growing practice of using local databases of transcribed LSs as reference standards (Heilmann et al., 2010; Justice et al., 2010).

However, the problem of having no access to true standards remains a thorny one, especially for diagnoses of LI. The 95% CI intervals around the original and the revised specificity and NPP figures for the DELV–NR in Tables 3 and 5 show that the true value is either high or very high (95% CI [0.81, 0.97]); thus, even taking potential error into account, one can have confidence that the diagnoses of

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

typical development made by the DELV–NR are likely to be true. By contrast, the ranges for sensitivity and PPP derived for the DELV–NR extend from high to uninformative (95% CI [0.95, 0.46]). How can we decide in which region of the CI the true value lies? What other evidence is available that could support lowering or raising our estimate? Application of other tests—that have also been validated against an inappropriate reference standard—will lead to the same quandary.

We have argued that CSaR can be applied as a provisional resolver (Green et al., 1998), used in accordance with its empirically derived measures of predictive power to provide an evidence-based method for corroborating some of the DELV–NR diagnoses that disagreed with the LS. In some forms of discrepancy resolution, as discussed by Green et al. (1998), one would reanalyze either the false negatives or the false positives—not both. However, that strategy is based on knowing which way the gold standard erred—toward lower specificity or lower sensitivity. It also is most useful when the resolver test has very high specificity and somewhat low sensitivity or vice versa, whereas Table 4 shows that the CSaR did not fit either profile.

Should one then completely abandon CSaR as having no utility? We think not—at least, not completely. First, specificity and NPP were adequate, so one can have a fair amount of confidence in its diagnoses of typical development and will also expect between one quarter and two-thirds of the CSaR diagnoses of LI to be correct (95% CI around PPP [0.27, 0.68]). Given the large number of diagnoses expected to be incorrect, we would not have enough confidence in a CSaR diagnosis to override a DELV–NR diagnosis when there is no indication that the DELV–NR is wrong (as many critics of discrepancy resolution require; Hawkins et al., 2001). However, when the diagnosis of the DELV–NR has been put in doubt by an opposite diagnosis from the LS, we suggest that one is justified in accepting a CSaR diagnosis as a corroboration of the DELV–NR in proportion to the values of its predictive power calculated relative to LSs, as is illustrated in Table 5.

This strategy could be considered circular if we were using the LS-derived PPP and NPP figures to evaluate the LS. But we do not claim to have tools to evaluate the LS. Instead, we proposed using the LS-derived predictive power and LR of the CSaR to help evaluate the likelihood of the DELV–NR diagnosis being accurate. One might object that DR in general—and even more so, this modification of it—will always bias the analysis upward in favor of the DELV–NR, but the alternative is to accept the near-certain downward bias of the LS (Green et al., 1998). As Hawkins et al. (2001) stated,

> It is not automatically true that [the resolved sensitivity estimate] is more biased than the [original] figures. In different circumstances, either of these potentially biased estimates may be closer to the true sensitivity and neither is guaranteed to be the less biased. (p. 1991)

We reiterate Heilmann et al.'s (2010) caution that diagnostic accuracy measures are only as good as their gold standard. We propose that we have taken advantage of the best gold standard available, and we have also found a way to use empirically derived information to refine its predictions.

Threats to Validity of the Experimental Method

The experimental method used in the current study is also susceptible to threats to its validity, such as spectrum bias. If one is recruiting participants according to a preexisting plan, it is possible that only cases that are very far apart will be selected: They would not be hard to distinguish with any method and so would not be a stringent test of the new method. However, if the examiners indeed chose only very clear cases of typical development and LI for the LSs, the resulting spectrum bias should have made it easier for both the CSaR and the DELV–NR to agree with the LS profiles. In fact, it should favor the CSaR more than the DELV–NR because the confirmation of clinical status was made by the clinicians when they selected children to record.

The procedure in the current study was also vulnerable to subjective bias in that the reference standard and the test to be evaluated were administered concurrently in the same context by the same examiner. This procedure has the advantage of holding constant as many aspects of the administration as possible. However, as Dollaghan and Horner (2011) outlined, with this strategy there was a risk that the examiners were not blinded to the participants' clinical status and the purpose of the study. Those concerns were minimized for these analyses by the fact that the clinical status assigned by the DELV–NR had not been established at the time of the LS collection and transcription. No one, not even the test's authors or publishers, could have known what diagnostic decision the DELV–NR would make because the analyses of field-testing data had not been completed.

Furthermore, for the LSs in this study, the use of a range of measures meant there was opportunity for a child to show different levels of skill on different tasks, so the LSs did not always give a simple pass–fail answer. For children with widely varying scores on the different LS measures, we made the decision to count LSs with at least one very low score as "failing," even if the other scores were within the average range. We feel that the strategy is ecologically valid because children may be referred for services on the basis of weakness in only one area (e.g., expressive only or morphosyntax only). However, other researchers might make a different decision and find a different result. Similarly, we chose 1.5 *SD*s below the mean as a cutoff for failing despite the DELV–NR manual's report that sensitivity and specificity and PPP/NPP were observed to be higher at –1 *SD* (Seymour et al., 2005, pp. 140–141). This difference in recommendations is based on the use of different reference standards. The publisher used CSaR as the reference standard for this analysis, whereas we used LS. However, one must keep in mind that our sample had a 30% prevalence of LI, which is not representative of a broader population. Clinicians working with a sample with a

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

different rate of LI might find that another cutoff would be more appropriate for their population (Green et al., 1998).

Implications for Clinical Practice

Limitations of Language Sample Analysis

The limitations of LSs noted above remind us that all measures have weaknesses as well as strengths and that clinical decisions generally need to be made on the basis of more than one criterion. Using the procedures outlined in the current study, the tests of diagnostic accuracy validated the dialect-neutral DELV–NR, which is especially important for CLD children. Given the proposed sensitivity and specificity of the DELV–NR in Table 5, which are more than adequate according to current standards, clinicians can have a high degree of confidence in its diagnostic conclusions. When a second opinion is required, one recommendation would be to compare the DELV–NR outcomes with LS analysis, as is often done in practice. One's clinical judgment will also be called upon. Indeed, if clinical judgment were not required, the qualifications of the diagnostician would be irrelevant. No single test replaces clinical judgment.

The Phenotype of LI

Tager-Flusberg and Cooper (1999), Owen and Leonard (2006), and Kohnert (2008), among others, have proposed that a determination of LI be made on the basis of a comprehensive range of skills and item types. This is especially necessary for speakers of nonmainstream varieties of English whose typical language patterns are often confused for the most common clinical markers of disorder for mainstream speakers. One implication of the current findings is that a broad definition of LI, as embodied in the DELV–NR, was effective in distinguishing deficit from difference.

One might ask, what did LI look like for the children in the current study? The composite picture painted by the DELV–NR and the LS profiles included restricted vocabulary, lower than average sentence complexity, and ineffective pragmatics—that is, difficulties generating cohesive narratives or poor understanding of the listener's needs. Seymour (2004) suggested that using a combination of noncontrastive morphosyntax (e.g., past tense copula, which is not optional in AAE, as found in Part 2 of the DELV–ST) and other linguistic structures was effective for diagnosis of CLD AAE-first learners. The DELV–NR added information about comprehension of long-distance movement in complex sentences, articles, question asking, and fast mapping, among other constructions.

Conclusions

Overall, the results of this study strengthen the validation of the DELV–NR through its greater agreement with LS measures and its convergence with the proposed combination of reference standards, using LSs as the major reference and CSaR as a provisional resolver test. Because the DELV–NR—the index test being evaluated—was designed to differ from other tests, typical gold standards for determining concurrent validity and diagnostic accuracy were not appropriate for it. LSs met the requirement that a gold standard be dialect neutral and also multifaceted, but LSs were shown to have other limitations. Still, where there were discrepancies between the clinician's categorization of the children's preexisting clinical status and their status derived from the DELV–NR, LS analyses corroborated the DELV–NR test results three times more often than they supported the CSaR. We conclude that the DELV–NR provided a rich profile of language skills known to be linguistically significant in the 5- and 6-year range that was in keeping with the children's language in spontaneous speech. The higher agreement with LSs of the DELV–NR than LSs with CSaR showed that the DELV–NR can help improve the assessment of LI for AAE-first speakers and, on the basis of its re-norming on a general American population, for GAE-first speakers as well.

Acknowledgments

This work was supported, in part, by Contract N01 DC8-2104 from the National Institute on Deafness and Other Communication Disorders and Training Grant H029D30072-97 07 from the U.S. Department of Education to Harry Seymour, P.I. We are grateful to Harry Seymour and his collaborators, Tom Roeper and Jill de Villiers, for allowing us access to the language samples for this article, and to Peter de Villiers for significant contributions to earlier versions of this article.

References

American Speech-Language-Hearing Association. (1983, September). *Social dialects and implications of the position on social dialects* [Position paper]. Retrieved from www.asha.org/docs/html/PS1983–00115.html

American Speech-Language-Hearing Association. (2003). *American English dialects* [Technical report]. Retrieved from www.asha.org/docs/html/TR2003–00044.html

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language, 20,* 139–152.

Burns, F. A., de Villiers, P. A., Pearson, B. Z., & Champion, T. B. (2012). Dialect-neutral indices of narrative cohesion and evaluation. *Language, Speech, and Hearing Services in Schools, 43,* 132–152.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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.

de Villiers, J. G. (2004). Cultural and linguistic fairness in the assessment of semantics. *Seminars in Speech and Language, 25,* 73–90.

de Villiers, P. A., & de Villiers, J. G. (2010). Assessment of language acquisition. *Cognitive Science, 1,* 230–244. doi:10.1002/wcs.30

de Villers, J. G., & Roeper, T (2001, November). *What every 3-year-old should know*. Paper presented to the American Speech-Language-Hearing Association Convention, New Orleans, LA.

de Villiers, P. A., Roeper, T., & de Villiers, J. G. (1999, November). *What every 5-year-old should know: Syntax, pragmatics, and semantics*. Paper presented at the American Speech-Language-Hearing Association Convention, San Francisco, CA.

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

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

EpiMax. (n.d.). *Table calculator: Epidemiology and lab statistics from study counts*. Princeton, NJ: Clinical and Economic Software. Available from www.healthstrategy.com/epiperl/epiperl.htm

Green, T. A., Black, C., & Johnson, R. (1998). Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *Journal of Clinical Microbiology, 36,* 375–381.

Hadley, P. A. (1998). Language sampling protocols for eliciting text-level discourse. *Language, Speech, and Hearing Services in Schools, 29,* 132–147.

Hawkins, D. M., Garrett, J., & Stephenson, B. (2001). Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine, 20,* 1987–2001.

Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools, 41,* 84–95.

Hewitt, L., Hammer, C. S., Yonte, K., & Tomblin, B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSyn, and NDW. *Journal of Communication Disorders, 38,* 197–213.

Jackson, J. E., & Pearson, B. Z. (2010, November). *Non-contrastive versus traditional language assessment in General American and African American English speakers*. Paper presented at the American Speech-Language-Hearing Association Convention, Philadelphia, PA.

Justice, L. M., Breit-Smith, A., & Rogers, M. (2010). Clinical Forum, Prologue: Data recycling: Using existing databases to increase research capacity in speech-language development and disorders. *Language, Speech, and Hearing Services in Schools, 41,* 39–43.

Kohnert, K. (2008). *Language disorders in bilingual children and adults*. San Diego, CA: Plural.

Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Long, S., Fey, M., & Channell, R. (2003). Computerized Profiling [Computer software]. Retrieved from www.computerizedprofiling.com

Losen, D. J., & Orfield, G. (Eds.). (2002). *Racial inequity in special education: The Civil Rights Project at Harvard University*. Cambridge, MA: Harvard Education Press.

Magaziner, K., Sunderland, K., Pearson, B. Z., & de Villiers, P. (2008, July). *Assessing language growth and delay between ages 5 and 7*. Poster presented at the International Association for the Study of Child Language, University of Edinburgh, United Kingdom.

Miller, J. F., & Chapman, R. S. (2002). Systematic Analysis of Language Transcripts (Version 7.0) [Computer software]. Madison, WI: Waisman Center.

Miller, J. F., Long, S., McKinley, N., Thormann, S., Jones, M., & Nockerts, A. (2005). *Language Sample Analysis II: The Wisconsin guide*. Madison: Wisconsin Department of Education.

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

Oetting, J. B., Newkirk, B. L., Hartfield, L. R., Wynn, C., Pruitt, S. L., & Garrity, A. W. (2010). Index of Productive Syntax for children who speak African American English. *Language, Speech, and Hearing Services in Schools, 41,* 328–339.

Owen, A., & Leonard, L. B. (2006). The production of finite and nonfinite complement clauses by children with specific language impairment and their typically developing peers. *Journal of Speech, Language, and Hearing Research, 49,* 548–571.

Paul, R., Tetnowski, J., & Reuler, E. (2007). Communication sampling procedures. In R. Paul & P. W. Cascella (Eds.), *Introduction to clinical methods in communication disorders* (pp. 111–154). Baltimore, MD: Brookes.

Pearson, B. Z., Velleman, S. L., Bryant, T. J., & Charko, T. (2009). Phonological milestones for African American English-speaking children learning Mainstream American English as a second dialect. *Language, Speech, and Hearing Services in Schools, 40,* 1–16.

Pewsner, D., Battaglia, M., Minder, C, Marx, A., Bucher, H., & Egger, M. (2004). Ruling a diagnosis in or out with "SpPIN" and "SnNOut": A note of caution. *British Medical Journal, 329*(4759), 209–213. doi:10.1136/bmj.329.7459.209

Plante, E., & Vance, R. (1994). Selection of pre-school language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25,* 15–24.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). MLU levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53,* 333–349.

Robinson, G., & Norton, P. (2012, November). *How does your state represent? African Americans on speech-language caseloads*. Paper presented at the American Speech-Language-Hearing Association Convention, Atlanta, GA.

Roeper, T. (2007). *Prism of grammar*. Cambridge, MA: MIT Press.

Scarborough, H. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11,* 1–22.

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.

Seymour, H. N., Bland, L., & Green, L. J. (1998). Difference versus deficit in child African-American English. *Language, Speech, and Hearing Services in Schools, 29,* 96–108.

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

Seymour, H. N., Roeper, T., & de Villiers, J. G. (2003a). *Diagnostic Evaluation of Language Variation—Criterion Referenced*. San Antonio, TX: The Psychological Corporation.

Seymour, H. N., Roeper, T., & de Villiers, J. G. (2003b). *Diagnostic Evaluation of Language Variation—Screening Test*. San Antonio, TX: The Psychological Corporation.

Seymour, H. N., Roeper, T., & de Villiers, J. G. (2005). *Diagnostic Evaluation of Language Variation—Norm Referenced*. San Antonio, TX: The Psychological Corporation.

Seymour, H., & Seymour, C. M. (1979). The symbolism of Black English: I'd rather switch than fight. *Journal of Black Studies, 9,* 397–410.

Sonis, J. (1999). How to use and interpret interval likelihood ratios. *Family Medicine, 31,* 432–437.

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

Stockman, I. J. (1986). Language acquisition in culturally diverse populations: The Black child as a case study. In O. Taylor (Ed.), *Nature of communication disorders in culturally and linguistically diverse populations* (pp. 117–155). San Diego, CA: College-Hill Press.

Stockman, I. J. (1996). The promise and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools, 27,* 355–366.

Stockman, I. J. (2010). A review of developmental and applied language research on African American children: From a deficit

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

to difference perspective on dialect differences. *Language, Speech, and Hearing Services in Schools, 41,* 23–38.

Tager-Flusberg, H., & Cooper, J. (1999). Present and future possibilities for defining a phenotype for specific language impairment. *Journal of Speech, Language, and Hearing Research, 42,* 1275–1278.

Taylor, O. (1969). Social and political involvement of the American Speech and Hearing Association. *ASHA, 11,* 216–218.

Taylor, O. (Ed.). (1972). Language and the Black urban child [Special issue]. *Language, Speech, and Hearing Services in Schools, 3*(4).

Tomblin, B., Records, N. L., Buckwalter, P. R., Zhang, X., 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. Bureau of the Census. (2000). Current Population Survey, October 2000: School Enrollment Supplemental File [CD-ROM]. Washington, DC: Author.

Westby, C. E. (1999). Assessing and facilitating text comprehension problems. In H. W. Catts & A. G. Kamhi (Eds.), *Language and reading disabilities* (pp. 154–221). Boston, MA: Allyn & Bacon.

Wyatt, T. A. (2002). Assessing the communicative abilities of clients from diverse cultural and language backgrounds. In D. E. Battle (Ed.), *Communication disorders in multicultural populations* (3rd ed., pp. 415–459). Boston, MA: Butterworth-Heinemann.

Downloaded from: https://pubs.asha.org Boston University on 12/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions