JSLHR

Research Article

Diagnostic Accuracy of Sentence Recall and Past Tense Measures for Identifying Children's Language Impairments

Sean M. Redmond,[a] Andrea C. Ash,[a] Tyler T. Christopulos,[a] and Theresa Pfaff[a]

Purpose: Measures of linguistic processing and grammatical proficiency represent strong candidates for adaptation into language screeners for early elementary students. One key barrier, however, has been the lack of consensus around the preferred reference standard for assigning affected status. Diagnostic accuracies associated with sentence recall and past tense marking index measures were examined relative to 5 different reference standards of language impairment: receipt of language services, clinically significant levels of parental concern, low performance on language measures, a composite requiring at least 2 of these indicators, and a composite requiring convergence across all indicators.
Method: One thousand sixty grade K–3 students participated in school-based language screenings. All students who failed the screenings and a random sampling of those who passed were invited to participate in confirmatory assessments. The community-based sample was supplemented by a clinical sample of 58 students receiving services for language impairment. Two hundred fifty-four students participated in confirmatory testing. Examiners were naive to participants' status.
Results: Diagnostic accuracies for the sentence recall and past tense marking index measures varied across the different reference standards (areas under receiver operating characteristic curves: .67–.95). Higher levels of convergence occurred with reference standards based on behavioral measures. When affected status was defined by receipt of services and/or parental ratings, cases presented with higher levels of performance on the language measures than when affected status was based on behavioral criteria.
Conclusion: These results provide additional support for the adaptation of sentence recall and past tense marking to screen for language impairments in early elementary students.
Supplemental Material: https://doi.org/10.23641/asha.8285786

The consensus across three decades of population-based estimates is that roughly 7%–8% of children will enter kindergarten each year with the considerable disadvantage of linguistic deficits that cannot be attributed to concomitant sensory impairments, intellectual limitations, motor deficits, or other neurodevelopmental conditions (Beitchman et al., 1986; Norbury et al., 2016; Tomblin et al., 1997). These epidemiological studies confirm further that this particular profile of idiopathic language disorder, or, as it is often referred to in the literature, specific language impairment (SLI), represents the majority (70%–75%) of all cases of developmental language disorder (DLD). Investigations tracking kindergarteners affected by SLI through their compulsory education into young adulthood suggest that the majority of affected individuals do not "catch up" to their peers in language performance. Language deficits present at school entry have also been predictive of cumulative risk for later academic and socio-emotional difficulties (Conti-Ramsden, 2008; Johnson et al., 1999; Tomblin & Nippold, 2014).

Although a prevalence rate of 7%–8% suggests that the average classroom will contain two students with a profile characteristic of SLI, evidence available on the issue indicates that only a minority of children with SLI will receive services to address their limitations and that ascertainment biases within preschool and school-age services are probably systemic. Factors shown to increase the likelihood of receipt of school-based language services after controlling for the severity of children's language impairments include male sex, White race, mothers with

[a]Department of Communication Sciences and Disorders, University of Utah, Salt Lake City
Correspondence to Sean M. Redmond: sean.redmond@health.utah.edu

Editor-in-Chief: Julie Liss
Editor: Maria Grigos
Received September 21, 2018
Revision received March 5, 2019
Accepted March 22, 2019
https://doi.org/10.1044/2019_JSLHR-L-18-0388

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

postsecondary education, and the presence of concomitant conditions such as speech sound disorder and attention-deficit/hyperactivity disorder (ADHD; Morgan et al., 2015, 2016; Sciberras et al., 2014; Wittke & Spaulding, 2018; Zhang & Tomblin, 2000). Increasing identification rates of children from different backgrounds without concomitant conditions will likely require adoption of school-based screeners targeting language impairments. Using emergent literacy, articulation, or other screeners as proxy measures for children's linguistic vulnerabilities appears to be inadequate. For example, Weiler, Schuele, Feldman, and Krimm (2018) found that, within their sample of 274 kindergartners, the majority of children who failed their language screeners passed their articulation and emergent literacy screeners.

Despite the ongoing challenges to providing children with SLIs and other DLDs adequate access to services they are entitled to, the last three decades have also been associated with advances in our understating of the psycholinguistic phenotype associated with SLI. By focusing efforts on the SLI behavioral phenotype rather than on the broader DLD phenotype, rapid progress has been made in genetic investigations of this condition (Rice, Smith, & Gayán, 2009; Rice, Zubrick, Taylor, Hoffman, & Gayán, 2018). Presently, the most promising pathognomonic markers of language impairment for English-speaking students in early elementary grades (K–3) consist of measures of linguistic processing (e.g., nonword repetition [NWR] and sentence recall) and grammatical proficiency. Tense marking deficits have been shown to be particularly emblematic of affected children's grammatical limitations. Another reason to focus efforts on these particular metrics over alternatives is that they have been shown to successfully differentiate cases of language impairment from cases of ADHD (Oram, Fine, Okamoto, & Tannock, 1999; Parriger, 2012; Redmond, Thompson, & Goldstein, 2011), a condition where poor performance on individual language tasks could potentially reflect children's difficulties with sustained attention, distractibility, or planning rather than result from underlying linguistic deficits.

In a meta-analysis of diagnostic accuracy studies on these particular phenotypic markers, however, Pawlowska (2014) identified limitations with the evidence base that curb the translation of these markers into areas of routine clinical practice, such as universal or targeted language screenings. These limitations include the common practice across diagnostic accuracy studies in this area of recruiting affected cases exclusively from practitioner caseloads and unaffected cases primarily through public announcements/community bulletins. As Pawlowska pointed out, this ascertainment practice can lead to a spectrum bias in study samples resulting in overestimations of diagnostic accuracy because only the most severe/complex cases of language impairment have been compared to the most robust cases of average/above-average language ability. Borderline cases and cases that represent profiles of spared language abilities within the context of other clinical conditions (e.g., individuals who have age-appropriate language skills but who have autism, ADHD, or low nonverbal skills) have been effectively filtered out of consideration. Another significant weakness has been the limited use of blinding procedures where examiners are naive to children's clinical status. Adoption of more rigorous designs that incorporate blinding procedures and recruit participants from a common source irrespective of their language status would address these limitations (Pawlowska, 2014).

The adaptation of clinical markers into either universal or targeted screeners for language impairments will eventually require consideration of additional obstacles to their implementation. For example, protocols should ideally be brief and port easily into school and other clinical settings in order to scale into mass screenings. They should also strive to minimize burdens on available resources. This includes the amount of personnel training required to administer and score these metrics to proficiency. Protocols should also demonstrate enough temporal stability to accommodate for potentially extended waiting periods between the screenings for language impairments and when follow-up confirmatory assessments and determinations of eligibility can take place.

One key barrier to advancing phenotypically aligned screening protocols for the routine identification of language impairments though has been the lack of an agreed-upon reference standard for affected status. What constitutes proof of children's language impairment status? Variation in how language impairment status is confirmed across diagnostic accuracy studies makes synthesis challenging (Pawlowska, 2014). It is difficult to arrive at conclusions regarding the relative strengths of different clinical markers when they have been directed at different reference targets. Furthermore, whether these distinctions make a difference is unclear. The extent to which the accuracy of clinical markers and their cutoffs varies as a function of different reference standards of language impairment has not been systematically examined within the same study sample.

Reference standards of language impairment used in diagnostic accuracy studies frequently consist of composite test scores taken from well-regarded omnibus standardized language tests such as the Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003) or the Test of Language Development–Primary: Fourth Edition (Newcomer & Hammill, 2008). Sometimes, investigators rely on conventional cutoffs (e.g., 1.0 or 1.25 *SD*s below the mean) to assign performance thresholds for both their reference standard and their index measures (Archibald & Joanisse, 2009; Bedore & Leonard, 1998; Jones Moyle, Karasinski, Ellis Weismer, & Gorman, 2011). Other times, performance thresholds on the index measures have been optimized against their reference standard targets using Youden's *J* statistic or other metrics derived from receiver operating characteristic (ROC) curve analyses (Greensdale, Plante, & Vance, 2009; Poll, Betz, & Miller, 2010; Redmond et al., 2011). Some studies have utilized more flexible criteria, placing children into the language impairment group on the basis of poor performance across two or more individual subtests, or have added other clinical measures into their formulas (e.g., Ellis Weismer

et al., 2000; Poll et al., 2010). This approach allows for more heterogeneity in the scope and severity of children's presenting language symptoms, and rather than interpreted as potentially problematic, it has often been considered a strength because it aligns with widely endorsed views that individuals affected by language impairments constitute a highly heterogeneous group (e.g., Bishop, Snowling, Thompson, Greenhalgh, & CATALISE Consortium, 2016).

Using receipt of services as the benchmark for evaluating the screening potential of clinical markers provides an alternative to relying on standardized test performance (e.g., Conti-Ramsden, Botting, & Faragher, 2001; Dollaghan & Campbell, 1998) that has the distinct advantage of aligning with linguistic—as well as nonlinguistic—symptoms that teachers and other referral sources find worrying. Receipt of services criteria represents one way of recognizing the contributions of social values to the perceived urgency of addressing different clinical symptoms (see Tomblin, 2006). However, as the results of Morgan et al. (2015, 2016), Sciberras et al. (2014), Wittke and Spaulding (2018), and Zhang and Tomblin (2000) demonstrate, receipt of services appears to be associated with troublesome inequalities in access. Reproducibility and generalizability of receipt of services as the standard for language impairment status is limited further by the presence of variability in diagnostic and eligibility criteria across clinical settings. Even when clinical judgments are made in the same setting by the same practitioners, they are also not necessarily stable over time due to disruptive realignments brought in by changing federal, state, local, and health care mandates.

Parental reports of general concerns about their children's communicative competence represent another diagnostic target that shares with the receipt of services reference standard the advantage of being derived from functional deficits. Furthermore, parents are uniquely positioned to view the adaptability of their children's communication skills across a variety of settings. Standardized rating protocols for parental concerns are available (e.g., Bishop, 2006), offering a balance between age-referenced criteria and recognition of clinically important variation across individuals in the translation of their underlying linguistic vulnerabilities into functional limitations. The results of Sciberras et al. (2014), however, encourage some caution with relying on parental ratings exclusively to identify children at risk for language impairment. In that study, parents of children with language impairments and concomitant ADHD were much more likely to initiate speech and language evaluations than parents of children with SLI, even though the levels of language impairment were comparable between these two groups.

Combining different reference standards of language impairment represents a sensible accommodation. For example, assignment of language impairment status for the purposes of evaluating clinical markers could require consistency among at least two different sources of evidence (multiple standardized tests/subtests, parental report, or receipt of services). Convergence across a minimum of two language measures represents a common eligibility criterion in clinical settings, although in practice the selections of specific measures and their cutoffs are usually left to individual practitioner judgment and influenced by the availability of resources. The chief drawback, however, to combining different reference standards is the potential unintended consequence of compounding error rather than reducing it as multiple measurements are brought into the decision process. The well-known problem of familywise error rates brought in when multiple statistical comparisons are applied to experimental data has an analog in the clinical context when multiple clinical metrics with various psychometric properties are applied to diagnostic decisions. Furthermore, the preferred process for resolving divergent results when they inevitably happen is unclear. Even if there were an accepted process, its implementation would likely vary considerably across practitioners and settings. Finally, there is an untested assumption that different sources of information about children's language impairment status should be given equal weight in clinical decision making. It is more likely the case that some sources of information provide more diagnostically relevant information than others.

A more stringent composite-based standard for the evaluation of language screeners would be the assignment of language impairment status to only those cases within the study sample where multiple behavioral measures coincide with both parental concerns and with children's service status. To increase confidence that only cases with unambiguous language impairment will be identified, we could also set more stringent standard score cutoffs. We would expect, relative to the other reference standards under consideration, significantly fewer children would qualify as affected by this operational definition. As a result, the value of screening measures designed around this reference standard for identifying children with less severe symptoms, who are nonetheless genuinely at risk for poor academic and social outcomes, would be compromised. Although perhaps too inflexible to serve as the basis for research studies of SLI and other DLDs, this presumably unequivocal version of language impairment represents an important benchmark for considering the relative tradeoffs associated with adopting different phenotypic markers.

The impact of multiple reference standards on estimates of diagnostic accuracies of clinical markers to identify children at risk for language impairments has been relatively unexplored. Convergence, where a clinical marker's observed consistency across different standards turns out to be so high that, in practice, it would not matter much which cutoffs were used, is desired but probably unlikely. Discrepancies between standardized, evidence-based criteria for language impairment and language impairments identified through clinician judgment and/or conventional eligibility guidelines have been reported in other aspects of practice. For example, Schmitt, Justice, Logan, Schatschneider, and Bartlett (2014) examined levels of alignment between Individualized Education Program treatment goals of 99 students receiving services for language impairments and their performance on norm-referenced measures of grammar, vocabulary, listening comprehension, and literacy skills.

These investigators found very limited alignment between objective measures of students' linguistic symptoms (e.g., standard vocabulary and grammar scores) and whether these areas received intervention. Similarly, Greensdale et al. (2009) found that using an empirically validated but potentially more generous cutoff standard score of 87 maximized classification accuracy of their preschool cases of SLI from cases of typical development for the Structured Photographic Expressive Language Test–Preschool: Second Edition. Greensdale et al. used as their reference for assigning language impairment status a convergence between receipt of services and standardized language testing.

In the current study, we examined accuracy levels of sentence recall and past tense marking indices when they were directed at five different methods for assigning affected status: (a) receipt of school-based and/or other language services, (b) standardized parental ratings of general communicative competence using two different composite cutoff standard scores (85 and 80), (c) language measures using two different cutoff standard scores (85 and 80), (d) a broad-based composite requiring consistency across at least two indicators using a standard score cutoff of 85, and (e) a restrictive composite requiring consistency across all indicators and using the more conservative threshold of 80 or lower.

The selection of the specific index measures used in the current study (Redmond, 2005; Rice & Wexler, 2001) was guided by the outcomes of Archibald and Joanisse (2009) and Redmond et al. (2011). Archibald and Joanisse reported that the sentence recall measure used in Redmond (2005) was more effective at identifying cases of language impairment (viz. CELF-4 Core Language standard scores thresholds at either 85 or 80) in a community ascertained sample of 400 grade K–3 students than an NWR task. Redmond et al. found further that, as a set, the Redmond sentence recall task and the screening test score from the Rice and Wexler (2001) Test of Early Grammatical Impairment (TEGI; consisting of the instrument's past tense and third-person singular probes) were as effective as a longer test battery at differentiating known cases of SLI from cases of typical development and from cases of ADHD that included Dollaghan and Campbell's (1998) NWR task and the Test of Narrative Language (Gillam & Pearson, 2004). Redmond et al. reported that, on average, administration of the past tense and regular third-person singular present tense probes required 9.5 min and the sentence recall measure required 3.5 min, suggesting their suitability as either universal or targeted screening instruments. Because we observed a high level of consistency within the Redmond et al. study sample in children's performances across the past tense and regular third-person singular present tense probes, we elected to focus on the past tense probe in the current study in the interest of further reducing administration time.

In Redmond et al. (2011), children with SLI were recruited through practitioner caseloads, which introduced the possibility of spectrum and other ascertainment biases, limiting the potential value of accuracy estimates to universal or targeted screenings. Because our focus in that study was on the issue of differential diagnosis of SLI and ADHD, we excluded cases of co-occurring language impairments and ADHD. We also excluded other types of DLD that would not have met criteria for SLI and children from other clinical groups with spared language abilities. The impact of the results of Redmond et al. was limited further by the absence of blinded evaluations (cf. Pawlowska, 2014). In the current study, we addressed these limitations. Cases of language impairment and typical language ability were drawn from both community-based (n = 1,060) and clinical (n = 58) samples of grade K–3 students. By combining samples drawn from both sources, we ensured that our study sample would include cases of language impairment with and without concomitant clinical conditions. By extending our community recruitment to include children receiving services for autism, ADHD, emotional/behavioral disorders, reading deficits, and other learning disabilities, we further ensured that cases of undiagnosed/misdiagnosed language impairments and cases of spared language abilities would also be included in our catchment. Separate teams of examiners administered our screening protocols and conducted the confirmatory assessments. All examiners were naive to children's clinical status.

Research Questions

To examine further the potential for sentence recall and past tense marking to screen for SLI and other DLDs, we addressed the following questions:

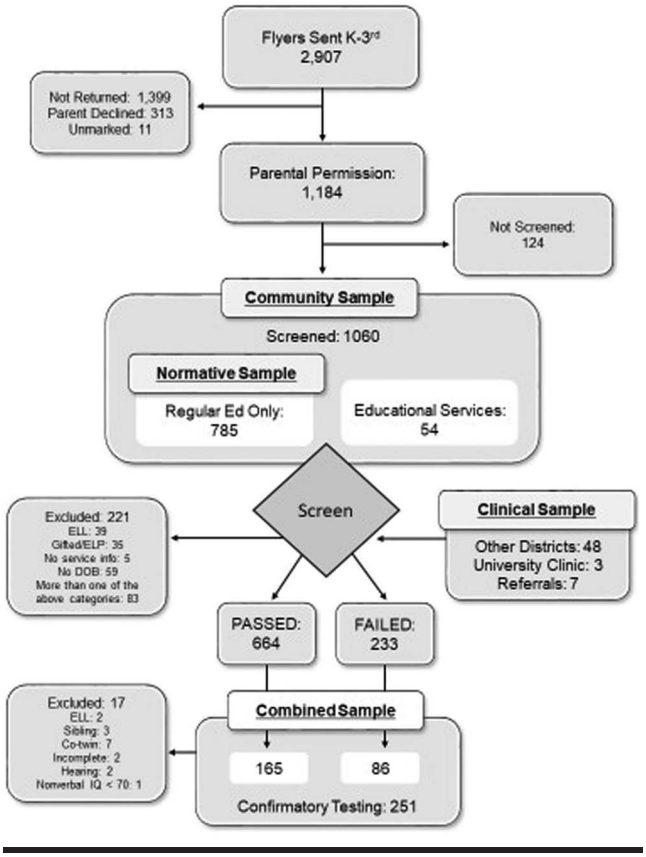
1. How do estimates of diagnostic accuracy of sentence recall and past tense marking index measures vary as a function of being directed at different reference standards for language impairment?
2. How consistent are the optimized cutoff values for sentence recall and past tense marking index measures across different reference standards for language impairment?
3. Are symptom severity levels associated with different reference standards of language impairment comparable?

Method

Recruitment Flow Into the Community, Normative, Clinical, and Combined Study Samples

Participants were recruited into this study from both community-based and clinical sources starting in the fall of 2011 and ending in Spring 2015. Confirmatory testing data from participants recruited through these two sources were then pooled into a combined sample to address our research questions (see Figure 1). The community sample was recruited through classroom notices sent home to children's parents asking them to indicate their permission to allow their children to participate in a school-sponsored language screening. During the first year of the study, children were recruited from six schools in the Salt Lake City School District that had been chosen by district personnel to provide

Figure 1. Recruitment flow into the community, normative, clinical, and combined study samples. ELL = English language learner.



the study with what they considered a representative catchment population for their district. An additional four schools and two school districts were added over the subsequent 3 years of data collection (school details are presented in Supplemental Material S1). Recruitment into language screenings was consecutive such that third graders were recruited into the study during the first year of data collection and each successive year of the project focused on an earlier grade eventually concluding with kindergarten. Over the 4 years of data collection, 2,907 flyers were distributed to families of children in grades K–3. Permission to administer the screening protocol was obtained from 1,184 parents. All children who had returned a consent form indicating parental permission to participate, who were in attendance at the day of the screening, and who had provided examiners with their assent were included in the community sample (n = 1,060; see Figure 1).

A subset of screening data collected on the community sample was used to create the normative sample (n = 782). The normative sample consisted of students who, according to educational records provided by school personnel, were not identified as an English language learner/limited English proficiency; were not receiving special education, remedial, or other support services at the time of the screenings (e.g., resource, reading support, speech-language); and were not participating in a district-provided gifted and talented program. Data were also excluded from those children whose date of birth or service status was unavailable. Data from children in the normative sample were used to create provisional age-referenced cutoff scores to identify children in the community sample who were performing significantly lower than their peers on the index measures. As demonstrated by Peña, Spaulding, and Plante (2006), when the goal is to identify potential cases of language impairment in children with unknown status, normative samples consisting exclusively of children who are not receiving clinical services are strongly preferred over samples that include clinical cases. The logic of working toward the construction of norms based on a "normal" range of performance would apply in the other direction as well, and this motivated our restriction of students enrolled in talented and gifted educational programs. Raw score conversions of the sentence recall and past tense measures into age-referenced standard scores and percentiles using the normative sample are available in Supplemental Material S2.

Two provisional screening criteria were created, one requiring age-referenced performance at the 10th percentile or lower on at least one of the two screening measures and the other requiring performance at the 15th percentile or lower on both screening measures. Formulating our provisional criteria in this way allowed us to recruit cases of both low/low-average and average/above-average performance and minimize threats of spectrum bias. The provisional thresholds do not, however, correspond to the optimized cutoffs associated with our index measures and the different reference standards we eventually vetted through ROC curve analyses using the Youden J index. These are presented later in this report. Once provisional cutoff scores for our index measures were established, participants were recruited for confirmatory testing during the second phase of the study.

Participants for the confirmatory testing phase of the study were recruited from a subset (n = 897) of the community sample. Children were not recruited into confirmatory testing sessions if they had been identified as receiving English language learner/limited English proficiency services. Children were recruited, however, regardless of special education/resource services, neurodevelopmental diagnoses, or psychiatric diagnoses. In cases of siblings recruited over the course of the project (including seven twinships), we randomly selected one child from each family to participate. All participants in the community sample who met these criteria and who performed below our provisional language screening criteria (n = 233) were invited to participate in the confirmatory testing phase of the study. Children in the community sample who performed at or above our criteria (n = 647) were assigned a random number (Haahr, 2006). A lottery system was used to invite participants who passed the screening or who were absent during the school screenings to participate in confirmatory testing. Those with smaller lottery numbers were contacted first (e.g., 5 before 14). We attempted to contact a total of 549 families from the community sample. Of the 549 families, 226 scheduled appointments

to participate in the confirmatory testing, 179 families were unreachable (did not answer their phone or their phone number was no longer in service), 109 families expressed initial interest in the study but did not follow-through, and 19 declined participation. Of the 19 participants whose families elected to participate in the screening portion of the study but then declined to participate in confirmatory testing, eight were male. Six of these 19 screening participants were receiving school services at the time of the study (two speech-language services only, two reading services only, and four speech-language and reading services).

To increase the representation of cases of language impairment (and potential cases of misdiagnoses) in our analyses, the community sample was supplemented by a clinical sample (n = 58; see Figure 1). Recruitment for the clinical sample was accomplished by targeting practitioner caseloads within school districts that were different than the one associated with the community sample. We also recruited through referrals from the University of Utah Speech-Language and Hearing Clinic. Speech-language pathologists from these settings were asked to distribute recruitment flyers to families of children on their caseloads being treated for language impairments over the 4 years of data collection. Adding cases recruited from the clinical sample to those successfully recruited from the community sample yielded the combined sample (n = 251; see Figure 1). Roughly a 2:1 ratio of students who performed above our provisional cutoffs to students who performed below completed confirmatory testing.

Demographic and educational service characteristics associated with the community, normative, clinical, and combined samples are presented in Table 1. On average, participants in the community and normative samples tended to be younger than those in the clinical and combined samples (7;0 compared to 7;6 [years;months]). Relatively balanced sex ratios were associated with community and normative samples, whereas the clinical and combined samples skewed toward higher representations of males. Relatively higher levels of representation of participants from low-income backgrounds were found in the community and normative samples than in the clinical and combined samples. This comparison, however, was complicated by the use of different metrics to estimate prevalence of low-income status (school level vs. individual familial census tract).

Measures

Index Measures

Screenings took place at children's schools in the libraries, gymnasiums, workrooms, and other spaces that were made available to the project. Multiple children were tested individually and simultaneously in these shared spaces by teams of two to seven examiners. Examiners were graduate and undergraduate students in the University of Utah Department of Communication Sciences and Disorders. The number of children participating in a given screening session varied from 10 to 65. Based on the volume of children being screened, sessions required 60–150 min of school time. Each individual's participation in the screening lasted between 15 and 25 min, based upon the length of time to walk from their classroom to the screening, complete the assent process, participate in the two screening measures, and then return to their classroom. Children's responses during the screenings were recorded on protocol forms by examiners and were also audio-recorded. These recordings were used by examiners to check their scoring accuracy after the screening sessions and adjust their initial scores if necessary. Recordings of children's responses were also used to estimate interrater reliability.

Sentence recall. Recordings of an adult female speaker presenting the stimuli from Redmond (2005) were administered to children during screenings via headphones attached to an MP3 player, and their responses were recorded by the examiner using digital audio recorders. Children were instructed by the speaker on the audio file to repeat exactly the sentences they heard ("Listen. I am going to say some sentences. After I have finished, I want you to say exactly what I have said. Say the same thing. Let's try a sentence"). The sentence set consisted of eight simple active declarative sentences and eight simple passive declarative sentences matched for word length (nine to 12 words). Following the scoring conventions of Archibald and Joanisse (2009), sentences received a score of 0 (*four or more errors*), 1 (*three or fewer errors*), or 2 (*no errors*). The maximum score is 32. Children's raw scores were transformed into z scores and then standard scores [(z * 15) + 100] using the means and standard deviations provided by the normative sample for different age bands.

Past tense marking. The past tense probe from the TEGI (Rice & Wexler, 2001) was included in our screening protocol. During administration of the past tense probe, children are shown two pictures where a human agent is depicted first engaging in an action (e.g., a boy painting a fence) followed by a scene indicating the agent's action had been completed (a boy standing next to a painted fence). The child's attention is directed to the second scene, and they are instructed to tell the examiner what the human agent did. The protocol consists of 10 high-frequency regular verbs and eight high-frequency irregular verbs. The past tense probe summary score is a percentage based on the total number of correctly produced regular, irregular, and overregularized irregular verbs divided by the total number of responses containing an obligatory context. One consequence of basing children's past tense probe scores on productions providing obligatory contexts rather than on whether they provided an expected response is that a score of zero does not mean children were unable to complete the probe or were providing unscorable responses. Unscorable responses (e.g., "I don't know," "He was painting," "I like this guy's shoes," or incomplete sentences) are completely removed from the calculation of the past tense probe summary score. Instead, a zero score indicates that, each time the child responded with a sentence containing a verb in a finite sentence site during the protocol, they provided a nonfinite form (e.g., "The boy paint a fence"). This design feature along with the requirement that a valid administration

Table 1. Demographic characteristics of community, normative, clinical, and combined study samples.

Sample	*N*	Age (years;months)	% Male	Ethnicity/race[a]	Low income	Educational services[b]
Community sample	1,060	*M* = 7;1 *SD* = 1;0 (5;0–9;7)	51.8	Hispanic = 9.7% Am Ind = 1.6% Asian = 4.2% Black = 2.5% Pac Isl. = 1.3% White = 89.2% Missing = 1.2%	32.18%[c]	ELL = 70 Gifted = 70 Speech = 31 Other = 66 Not provided = 18
Normative sample	782	*M* = 7;0 *SD* = 1;1 (5;0–9;4)	52.0	Hispanic = 6.0% Am Ind = 1.2% Asian = 3.6% Black = 2.3% Pac Isl. = 1.4% White = 91.0% Missing = 0.5%	28.90%[c]	ELL = 0 Gifted = 0 Speech = 0 Other = 0 Not provided = 0
Clinical sample	58	*M* = 7;5 *SD* = 1;3 (5;1–9;9)	67.3	Hispanic = 6.9% Am Ind = 0% Asian = 3.4% Black = 6.9% Pac Isl. = 0% White = 89.7% Missing = 0%	25.16%[d]	ELL = 0 Gifted = 0 Speech = 48 Other = 24 Not provided = 0
Combined sample	251	*M* = 7;6 *SD* = 1;2 (5;1–9;9)	61.0	Hispanic = 5.9% Am Ind = 0.8% Asian = 4.3% Black = 4.3% Pac Isl. = 1.6% White = 87.4% Missing = 1.6%	21.62%[d]	ELL = 0 Gifted = 0 Speech = 86 Other = 53 Not provided = 0

Note. Ethnicity was categorized as Hispanic or not Hispanic; racial categories included American Indian (Am Ind)/Alaskan, Asian, African American/Black, Pacific Islander (Pac Isl.)/Hawaiian Native, and White. ELL = English language learner.

[a]Percentage of parents who elected not to provide racial/ethnic category information across different samples ranged from 0% to 1.6%. [b]Some students were receiving more than one educational service. [c]Weighted based upon the number of participants from each school and the percent low income from the individual schools. [d]Percentage of families in poverty based upon census tract.

of the past tense probe requires demonstration that children consistently produce *–s*, *–z*, *–t*, and *–d* in word-final positions (e.g., *mouse, nose, bed, boat*) makes the TEGI protocol relatively unique among available standardized language tests in its level of control for confounding factors. TEGI past tense probe summary scores are typically reported as percentages, which provide a straightforward interpretation of children's levels of proficiency within obligatory contexts (e.g., a value of 75 on the past tense summary score indicates that children provided correctly marked finite verbs 75% of the time pooled across the obligatory contexts they produced for regular and irregular past tense verbs). However, in this study, because we needed a common metric to permit comparisons across index measures, we also transformed children's TEGI past tense summary scores into age-referenced standard scores. We did this using means and standard deviations calculated on participants in the normative sample.

Reference Standards for Language Impairment

Confirmatory testing sessions at the University of Utah were arranged with families of children from the community sample within 5 months of their child's participation in the school screenings (M = 19 weeks, range: 1–40 weeks). Children from the clinical sample completed both screening and confirmatory sessions at the University of Utah. Children's performances on measures associated with the different reference standards of language impairment and their performances on exclusionary measures were collected during the confirmatory sessions. The number of cases of language impairment identified by the different reference standards within the combined sample varied from 14 to 98 (see Table 2).

Reference standard I: Receipt of services. School records provided by district personnel indicating children's receipt of language services during the time of the study were combined with parental reports of these services received outside the school setting to identify children within the combined sample who met the receipt of services criteria for language impairment. A total of 86 participants in the combined sample met this criterion.

Reference standard II: Clinically significant levels of parental concern. The presence of clinically significant levels of parental concern across participants' communication skills was documented using the Children's Communication Checklist–Second Edition (CCC-2; Bishop, 2006). The CCC-2 represents one of the few age-referenced parent rating instruments available that incorporates proficiencies

Table 2. Participants identified with language impairment from the combined sample (n = 251) as a function of different reference standards for language impairment status.

Reference standard	Positive	Negative
I.		
Receiving SLP services	86 (34.5%)	165 (65.5%)
II.		
CCC-2 ≤ 85	52 (20.7%)	199 (79.3%)
CCC-2 ≤ 80	35 (13.9%)	216 (86.1%)
III.		
CELF-4 ≤ 85	63 (25.1%)	188 (74.9%)
TEGI ≤ 85	93 (37.1%)	158 (62.9%)
NWR ≤ 85	91 (36.3%)	160 (63.7%)
CELF-4 ≤ 80	48 (19.1%)	203 (80.9%)
TEGI ≤ 80	87 (34.7%)	164 (65.3%)
NWR ≤ 80	72 (28.7%)	179 (71.3%)
IV.		
At least two of the following: Receiving SLP services, CELF ≤ 85, TEGI ≤ 85, NWR ≤ 85, CCC-2 ≤ 85	97 (39.1%)	154 (60.9%)
V.		
All of the following: Receiving SLP services, CCC-2 ≤ 80, CELF ≤ 80, TEGI ≤ 80, NWR ≤ 80	14 (5.6%)	237 (94.4%)

Note. SLP = speech-language pathology; CCC-2 = Children's Communication Checklist–Second Edition; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; TEGI = Test of Early Grammatical Impairment; NWR = Dollaghan and Campbell's (1998) nonword repetition task.

across multiple language domains. General Communication Composite standard scores were calculated for each participant in the combined sample following procedures presented in the CCC-2 manual. The General Communication Composite represents the standardization of the summed scale scores across the Speech, Semantics, Coherence, Initiation, Scripted Language, Context, and Nonverbal Communication subscales. It does not include the Social Relations and Interests subscales designed to target potential symptoms of autism. Fifty-two combined sample participants met the 85 standard score threshold, and 35 met the 80 standard score threshold.

Reference standard III: Behavioral tests and measures. Three behavioral measures of children's linguistic proficiencies were collected during confirmatory testing (the CELF-4, the TEGI, and Dollaghan and Campbell's [1998] NWR task).

The CELF-4 has demonstrated strong psychometric properties (Spaulding, Plante, & Farinella, 2006) and has been used as the reference standard in several diagnostic accuracy studies (Pawlowska, 2014), allowing for direct comparisons between results obtained with our study sample and those from previous investigations. The CELF-4 also represents one of the most frequently used omnibus language tests in clinical practice (Betz, Eickhoff, & Sullivan, 2013; Finestack & Satterlund, 2018). Subtests associated with the instrument's Core Language score were administered to each participant in the combined sample according to their age. Core Language standard scores were calculated according to procedures presented in the CELF-4 manual. Sixty-three participants in the combined sample met the CELF-4 Core Language 85 standard score threshold, and 48 met the 80 standard score threshold.

The TEGI consists of four elicitation probes targeting children's productions of English finite forms (third-person singular present tense, past tense, auxiliary and copula BE, auxiliary and main verb DO) in simple declarative sentences and questions. High levels of concordance between affected and unaffected children's productions of these forms within spontaneous language samples and their performance on the TEGI probes over the course of their development have been established (Rice, Wexler, & Hershberger, 1998). This made the TEGI a particularly attractive reference standard to include in our evaluation given the long-standing and widely recognized ecological validity assigned to language sample measures (e.g., Miller, 1996). The TEGI has also demonstrated strong psychometric properties (Spaulding et al., 2006). Using age-referenced means and standard deviations provided in the TEGI manual, we transformed raw score on the Elicited Grammar Composite into standard scores for each participant in the combined sample. The presence of extremely low TEGI scores relative to age expectations (more than 6 *SD*s below the mean) within the combined sample (n = 16) required us to truncate TEGI standard scores to prevent negative standard scores. Derived standard scores less than 1 were replaced with 1. Because the Elicited Grammar Composite, like the past tense probe summary score, is based on the obligatory contexts children provide, low scores do not indicate a general inability to complete the protocol. Rather, low scores reflect children's elevated use of nonfinite verb forms in sentence positions licensed for finite verb forms across a variety of grammatical targets. Ninety-three participants met the TEGI Elicited Grammar Composite 85 standard score threshold, and 87 met the 80 standard score threshold.

Several reports have indicated that children with language impairments frequently demonstrate weaknesses in NWR, a task that taps into children's working memory capacities (Graf Estes, Evans, & Else-Quest, 2007). We administered Dollaghan and Campbell's (1998) NWR protocol to participants in the combined sample. Each phoneme was scored as either correct or incorrect, as outlined by Dollaghan and Campbell. We converted children's percentage of total number of phonemes correct into age-referenced standard scores based upon the percentages of 132 participants within the combined sample. We applied the same restrictions used to generate our screening norms (i.e., typically developing monolingual English students enrolled in regular education). NWR norms were established for 12-month increments covering ages 5;0–9;11 (see Supplemental Material S3). Based upon these norms, 91 participants met the 85 standard score threshold and 72 met the 80 standard score criteria.

Reference standard IV: Broad-based composite. A heterogeneous view of underlying language impairments and their overt symptoms motivates a reference standard that can be met in a variety of ways. The criteria of two or more indicators of language impairment across the service receipt, CCC-2 at or below 85, CELF-4 at or below 85, TEGI at or below 85, and NWR at or below 85 were applied to the combined sample. Ninety-seven children within the combined sample presented with at least two items from the set of five possible clinical indicators. Twenty-one of these children met criteria on all five items, representing the largest subgroup within the broad-based composite reference standard. Proportionally, this subgroup was followed by a subgroup of children who met criteria on service receipt, CELF-4, TEGI, and NWR indicators ($n = 12$); a subgroup of children who met criteria on the CELF-4, TEGI, and NWR ($n = 10$); a subgroup who met criteria on the TEGI and NWR ($n = 8$); and a subgroup who met criteria on service receipt, TEGI, and NWR ($n = 5$). All other possible indicator combinations yielded four or fewer cases. In the aggregate, profiles consisting of weaknesses on both the TEGI and NWR measures characterized the majority of cases captured by a reference standard for language impairment that allowed criteria to be met in multiple ways.

Reference standard V: Stringent composite. A restrictive view of language impairments and symptoms would require convergence across all five indicators and would set standard score criteria at 80 or below. When applied to the combined sample, 14 participants met the stringent composite criteria.

Exclusionary Measures

During confirmatory testing sessions, all potential participants were administered a hearing screening, the phonological screening subtest from the TEGI, and completed a standardized test of their nonverbal abilities (Naglieri, 2003). Children who failed the hearing screening ($n = 2$) were excluded from the combined sample. An additional participant with minimal verbal abilities was excluded from the study because they were unable to complete the phonological screening and the nonverbal test.

Training and Characteristics of Examiners

Prior to participating in training, examiners completed the Human Subjects in Research training through the Collaborative Institutional Training Initiative. Examiners who participated in the language screenings attended two 2-hr training sessions that covered the administration and recording of the sentence recall and past tense measures. Undergraduate ($n = 30$) and graduate ($n = 12$) student volunteers, as well as graduate research assistants ($n = 18$), conducted the language screenings in the schools. All volunteers and assistants met transcription reliability at 85% or higher ($M = 93$, $SD = 4$, range: 85–100).

Most onsite screenings were conducted by student volunteers with supervision by the project manager and graduate research assistants. Graduate research assistants completed all reference standard testing. When graduate research assistants participated in screenings, the project manager assigned them children they had not screened. This ensured that graduate research assistants were naive to the screening outcomes of the children they were testing.

Reliability

Reliability of our index measures was considered in three different ways: interrater scoring consistency, short-term stability, and long-term stability. Levels of interrater scoring consistency of the past tense and sentence recall screening were calculated using 50 recordings from the community sample. Scored responses between the initial (sentence recall: $M = 22.10$, $SD = 6.01$; past tense marking: $M = 95.54$, $SD = 8.31$) and second (sentence recall: $M = 22.18$, $SD = 6.02$; past tense marking: $M = 96.02$, $SD = 8.56$) independent transcription and scoring were not significantly different for either measure (past tense: $t(49) = -1.01$, $p = .32$; sentence recall: $t(49) = -0.663$, $p = .51$). Correlations indicated further high levels of interrater scoring consistency (sentence recall: $r(50) = .99$, $p < .001$; past tense marking: $r(50) = .92$, $p < .001$).

We also examined the stability of the index measures across two intervals using 133 recordings from the confirmatory sample. Thirty-seven children were administered the index measures twice within a 4-week period ($M = 1.03$ weeks, $SD = 1$ week, range: 1–4 weeks). Pearson correlations were robust: $r = .95$ ($p < .01$) for the sentence recall and $r = .90$ ($p < .01$) for the past tense marking index measures. An additional 82 children were administered the index measures twice over a longer period ($M = 14.01$ weeks, $SD = 3.03$ weeks, range: 10–20 weeks). Pearson correlations for both measures over the longer period were also robust: sentence recall, $r = .86$ ($p < .001$); past tense, $r = .86$ ($p < .001$). Results from both time intervals indicated sufficient measurement stability for our index measures.

Twenty percent of the reference standard measure data ($n = 51$) collected over the course of the confirmatory testing sessions were rescored from video by a second

examiner in order to calculate interrater reliability. We calculated Pearson correlation coefficients between the original testing and the second scoring for the CELF-4 (r = .99, $p < .001$), NWR (r = .89, $p < .001$), and TEGI (r = .99, $p < .001$), indicating high levels of interrater scoring consistency for our index measures.

Analytic Approach

ROC curves were generated using SPSS v.25 to address our research questions. ROC curves plot sensitivity (Se) as a function of specificity (Sp) and display the discriminatory accuracy associated with different cutoffs to classify cases into two groups (Fluss, Faraggi, & Reiser, 2005). The area under the ROC curve (AUC) provides an overall estimate of an index measure's accuracy. Swets, Dawes, and Monahan (2000) have offered the following benchmarks for interpreting AUCs: .90–1.0, "excellent"; .80–.90, "good"; .70–.80, "fair"; and lower than .70, "poor" (see also Carter, Pan, Rai, & Galandiuk, 2016). Others, however, have pointed out that these benchmarks are probably unrealistic given the measurement complexities associated with neurodevelopmental and mental health disorders and seem to be more appropriate for engineering and biomedical applications. For example, Youngstrom (2014) pointed out that, for children's mental health disorders, the best performing checklists and inventories available provide AUC estimates in the .70–.80 range. For this reason, Youngstrom suggested values above .70 probably provide clinically "adequate" levels of accuracy. We used the Youden index (J) to identify optimal cutoff points on the ROC curves for both index measures across each of the five reference standards (Youden, 1950). The Youden index (J) is defined as $J = \max\{Se + Sp - 1\}$, such that a value of $J = 1$ would provide complete separation of affected and unaffected groups, whereas a $J = 0$ would indicate complete overlap. J represents the value for which $Se + Sp - 1$ is maximized. Once the optimal cutoff points were identified, the sensitivity, specificity, and positive and negative likelihood ratios were calculated for each cutoff score.

Results

Complete data were available for the index and reference standard measures on all children in the combined sample. No adverse events occurred during screening or confirmatory testing sessions.

The means, standard deviations, and ranges of raw scores from children on the index measures from the normative sample are provided in Table 3. The normative sample was divided into nine age bands based upon 6-month intervals (range: 5;0–9;5). Raw score means and standard deviations for each age band were used to create z scores and calculate individual standard scores for each index measure. The standard score conversion tables for the sentence recall and past tense marking index measures are provided in Supplemental Material S1.

Table 3. Means, standard deviations, and ranges of raw scores on the sentence recall and past tense marking index measures for the normative sample across different age ranges.

		Sentence recall			Past tense marking		
Age	*n*	*M*	*SD*	Range	*M*	*SD*	Range
5;0–5;5	65	16.06	7.16	0–29	90.58	15.90	0–100
5;6–5;11	92	16.71	6.70	0–29	90.10	15.67	19–100
6;0–6;5	106	19.47	7.11	1–32	91.76	13.76	13–100
6;6–6;11	141	21.65	5.58	0–32	94.83	7.64	43–100
7;0–7;5	114	21.91	7.15	1–32	94.54	10.70	19–100
7;6–7;11	81	23.65	5.19	7-31	95.30	7.76	50–100
8;0–8;5	84	25.62	4.78	8–32	97.46	5.74	67–100
8;6–8;11	73	25.70	4.76	10–32	97.74	3.57	88–100
9;0–9;5	26	26.96	2.54	22–31	99.12	2.63	89–100

Note. Redmond (2005) sentence recall and the past tense probe from the Test of Early Grammatical Impairment.

Correlations between the index standard scores and the standard scores for continuous language variables involved in the reference standards (CELF-4, TEGI, NWR, and CCC-2) are presented in Table 4. All correlations were significant at $p < .001$ and ranged in strength from "weak" to "very strong" (r = .35–.82). The weakest associations were between the CCC-2 and the index measures, and the strongest were between the index measures and the CELF-4 and TEGI. Given that the CELF-4 Core Language score represents a composite that includes its own Sentence Recall subtest and that the past tense marking index measure we used represents one of the probes on the TEGI, the observation of strong associations between sentence recall and CELF-4 and between past tense marking and the TEGI was probably not that surprising. What was not necessarily expected, however, was the observed strength of the associations between past tense marking and the CELF-4 (r = .68) and between sentence recall and the TEGI (r = .73).

Table 5 provides the AUCs, optimal cutoffs, sensitivity values, specificity values, positive likelihood ratios, and negative likelihood ratios for the sentence recall index

Table 4. Intercorrelations between the index measures and the behavioral measures (combined sample n = 251).

Measure[a]	1	2	3	4	5	6
1. Past tense marking	—					
2. Sentence recall	.63	—				
3. CELF-4	.68	.84	—			
4. TEGI	.82	.73	.75	—		
5. NWR	.57	.73	.75	.69	—	
6. CCC-2	.35	.45	.48	.40	.38	—

Note. Past tense marking = past tense probe from the TEGI; Sentence recall = sentence recall task from Redmond (2005); CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; TEGI = Test of Early Grammatical Impairment; NWR = Dollaghan and Campbell's (1998) nonword repetition task; CCC-2 = Children's Communication Checklist–Second Edition.

[a]All coefficients are significant at $p < .001$.

Table 5. Diagnostic accuracy of the sentence recall index measure as a function of different reference standards.

Reference standard	Area under the curve[a]	Optimal cutoff[b]	Sensitivity	Specificity	Positive likelihood ratio[c]	Negative likelihood ratio[d]
I. Receiving SLP services	.722	91.50	.667	.733	2.50	.454
II. Parental ratings						
CCC-2 ≤ 85	.773	94.50	.792	.643	2.21	.323
CCC-2 ≤ 80	.761	93.00	.800	.641	2.22	.312
III. Standardized tests and measures						
CELF-4 ≤ 85	.952	85.50	.891	.846	5.78	.129
TEGI ≤ 85	.851	93.00	.766	.785	3.56	.298
NWR ≤ 85	.872	94.50	.826	.769	3.57	.226
CELF-4 ≤ 80	.950	78.50	.878	.887	7.77	.138
TEGI ≤ 80	.865	93.00	.795	.780	3.61	.263
NWR ≤ 80	.870	93.00	.849	.754	3.45	.200
IV. At least two of the following:						
Receiving SLP services, CCC-2 ≤ 85, CELF ≤ 85, TEGI ≤ 85, NWR ≤ 85	.905	94.50	.857	.812	4.55	.176
V. All of the following:						
Receiving SLP services, CCC-2 ≤ 80, CELF ≤ 80, TEGI ≤ 80, NWR ≤ 80	.898	85.50	1.0	.700	3.33	0

Note. SLP = speech-language pathology; CCC-2 = Children's Communication Checklist–Second Edition; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; TEGI = Test of Early Grammatical Impairment; NWR = Dollaghan and Campbell's (1998) nonword repetition task.

[a]All areas under the receiver operating characteristic curve were significant at $p < .001$. [b]Determined using Youden index (J) where J = maximum {Sensitivity + Specificity – 1} [c]Positive likelihood ratio = Sensitivity / (1 – Specificity): values of 1 = "neutral," 3 = "moderately positive," ≥ 10 = "very positive." [d]Negative likelihood ratio = (1 – Sensitivity) / Specificity: values of 1 = "neutral," ≤ 0.30 = "moderately negative," ≤ 0.10 = "extremely negative."

measure across the five reference standards. AUCs for sentence recall and the various reference standards were all statistically significant at $p < .001$. AUCs ranged from .72 to .95, indicating "fair" to "excellent" levels of diagnostic accuracy across the different reference standards, using stringent benchmarks. Following Youngstrom's suggestion, these values would all be characterized as "adequate" for clinical use. The different reference standards, however, were associated with a wide range of optimal cutoff standard scores (79–95), suggesting only modest levels of consistency. This implies that diagnostic accuracy for a given cutoff score on the sentence recall index measure was, to a large extent, a function of the kind of language impairment being targeted.

Table 6 provides the AUCs, optimal cutoffs, sensitivity values, specificity values, positive likelihood ratios, and negative likelihood ratios for the past tense marking index measure across the five reference standards. AUCs for past tense marking were all statistically significant at $p < .001$ and ranged from .67 to .91, indicating "fair" to "excellent" levels of diagnostic accuracy across the different reference standards. With the exception of the receiving services reference standard, these values would all be characterized as "adequate" for clinical use. A smaller range of optimal cutoff scores was observed for the past tense marking index measure (89–97), suggesting that cutoff standard scores on the past tense marking index measure across the different reference standards were more consistent than those on the sentence recall. This was likely a consequence of the relatively sharper peak of the frequency distribution curve (kurtosis) associated with children's performances on the past tense measure across the different age bands (past tense kurtosis range: 0.365–23.88, sentence recall kurtosis range: –846 to 0.209). The frequency distribution of our past tense marking index measure reflects a well-established empirical finding regarding the status of tense marking as a potential clinical marker for SLI. For typically developing school-age children, reports consistently document the presence of very little variability in their capacities to provide finite verbs, and deviations from mastery levels of performance are generally not expected. In contrast, substantial variability in finite verb use has been associated with study samples of SLI and, in some cases, well into adolescence and beyond (see Ash & Redmond, 2014, for a review). This creates a situation where the presence of finiteness marking deficits in school-age children appears to be sufficient to assign language impairment status, but it is not always necessary. This is particularly true for older affected children where many consistently use finite verb forms correctly in conversation and during elicitation tasks but continue to meet experimental criteria for SLI (Rice et al., 1998). The presence of co-occurring low nonverbal ability (i.e., "nonspecific language impairment") in children with DLD has been associated with slower, more linear growth in finite verb marking (Rice, Tomblin, Hoffman, Richman, & Marquis, 2004).

Table 7 provides means, standard deviations, and ranges for participants' standard scores on the CCC-2, CELF-4, TEGI, and NWR measures segregated into affected and unaffected as defined by each of the five reference standards. One way of examining convergence across different reference standards is to consider whether severity

Table 6. Diagnostic accuracy of the past tense marking index measure as a function of different reference standards.

Reference standard	Area under the curve[a]	Optimal cutoff[b]	Sensitivity	Specificity	Positive likelihood ratio[c]	Negative likelihood ratio[d]
I. Receiving SLP services	.671	95.50	.529	.776	2.36	.607
II. Parental ratings						
CCC-2 ≤ 85	.752	89.00	.585	.829	3.42	.501
CCC-2 ≤ 80	.736	89.00	.657	.806	3.39	.426
III. Standardized tests and measures						
CELF-4 ≤ 85	.840	92.50	.750	.835	4.55	.299
TEGI ≤ 85	.855	96.50	.713	.880	5.94	.326
NWR ≤ 85	.757	96.50	.652	.837	4.00	.416
CELF-4 ≤ 80	.855	92.50	.796	.803	4.04	.254
TEGI ≤ 80	.843	96.50	.716	.860	5.11	.330
NWR ≤ 80	.736	96.50	.658	.788	3.10	.434
IV. At least two of the following:						
Receiving SLP services, CCC-2 ≤ 85, CELF ≤ 85, TEGI ≤ 85, NWR ≤ 85	.803	96.50	.684	.877	5.56	.360
V. All of the following:						
Receiving SLP services, CCC-2 ≤ 80, CELF ≤ 80, TEGI ≤ 80, NWR ≤ 80	.909	89.00	1.0	.789	4.74	0

Note. SLP = speech-language pathology; CCC-2 = Children's Communication Checklist–Second Edition; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; TEGI = Test of Early Grammatical Impairment; NWR = Dollaghan and Campbell's (1998) nonword repetition task.

[a]All areas under the receiver operating characteristic curve were significant at $p < .001$. [b]Determined using Youden index (J) where J = maximum {Sensitivity + Specificity – 1} [c]Positive likelihood ratio = Sensitivity / (1 – Specificity): values of 1 = "neutral," 3 = "moderately positive," ≥ 10 = "very positive." [d]Negative likelihood ratio = (1 – Sensitivity) / Specificity: values of 1 = "neutral," ≤ 0.30 = "moderately negative," ≤ 0.10 = "extremely negative."

levels captured by different reference standards align with each other. As indicated in Table 7, affected group mean standard scores on the CCC-2, CELF-4, TEGI, and NWR measures for Reference Standard V (stringent composite) were consistently lower than the other reference standards. This outcome was expected and represents a function of the more restrictive criteria associated with this reference standard. Means for the affected groups based on Reference Standards I (service receipt) and II (parental ratings) across the behavioral measures were similar to each other. Means for these two reference standards also tended to be higher than those associated with the other reference standards. In other words, identifying children as having a language impairment on the basis of either their service status or the levels of parental concern yielded a group of children with less severe behavioral symptoms than cases based on behavioral criteria.

Even though there was considerable overlap in cases identified as affected by the different reference standards, there were some cases that were only identified by one of the reference standards. Of those receiving speech-language pathology services, 22% (n = 19) were only identified by this reference standard. Approximately 15% (14.28%–15.27%) of the participants who scored either below 80 or 85 on the TEGI (SS ≤ 85 = 14, SS ≤ 80 = 13) and the NWR (SS ≤ 85 = 16, SS ≤ 80 = 11) were only identified as affected by these reference standards. For the parental judgment reference standard, 11.53% (SS ≤ 85 = 6) and 8.57% (SS ≤ 80 = 3) were identified as affected by only their CCC-2 scores. In contrast, the CELF-4 reference standard demonstrated the highest level of overlap with the other reference standards. Only 3.17% (SS ≤ 85 = 2) and 2.08% (SS ≤ 80 = 1) were identified as affected by their CELF-4 scores only.

Discussion

In a community-based sample of early elementary monolingual English-speaking students that had been supplemented by cases gathered from an independent clinical sample, we investigated the extent to which sentence recall and past tense marking indices could be used to identify children at risk for language impairment. Because an agreed-upon reference standard for affected language status does not presently exist, we considered diagnostic accuracy in light of five different ways of operationally defining language impairment.

As expected, concurrent validity associated with our index measures varied depending on how reference standards defined "language impairment." In other words, our data suggest various reference standards used to estimate diagnostic accuracy are probably not interchangeable. Overall, our index measures aligned reasonably well with those reference standards of language impairment that incorporated behavioral measures into their criteria. AUCs associated with the optimal cutoff scores identified for the sentence recall index measure ranged from .870 to .952 for Reference Standards III (standardized tests and measures), IV (broad-based composite), and V (stringent composite), indicating consistently excellent levels of diagnostic accuracy. As a comparison, our AUC values easily exceed those associated

Table 7. The means, standard deviations, and ranges of performance on the confirmatory language measures using Reference Standards I–V.

	Receiving SLP services		CCC-2		CELF-4		TEGI		NWR	
Reference standard	+	–	+	–	+	–	+	–	+	–
I. Receiving SLP services (n = 86)	100%	0%	89.17 (15.00) 59–130	103.78 (13.72) 51–132	83.51 (21.54) 40–126	101.06 (13.86) 52–132	60.28 (42.42) 1–117	94.24 (23.59) 1–124	77.80 (24.5) 40–118	97.62 (15.69) 53–130
II. Parental ratings										
CCC-2 ≤ 85 (n = 52)	73.1%	24.1%	76.77 (7.78) 51–85	104.53 (11.72) 86–132	78.79 (21.43) 40–118	99.30 (15.51) 46–132	57.89 (40.61) 1–114	89.06 (30.59) 1–124	75.23 (23.76) 40–123	94.91 (18.65) 40–130
CCC-2 ≤ 80 (n = 35)	77.1%	27.3%	73.43 (7.39) 51–80	102.88 (12.59) 81–132	77.69 (20.87) 40–115	97.86 (16.89) 40–132	53.67 (38.46) 1–110	87.29 (32.35) 1–124	75.04 (23.73) 40–118	87.29 (32.35) 1–124
III. Standardized tests and measures										
CELF-4 ≤ 85 (n = 63)	68.3%	22.9%	86.97 (13.77) 51–125	102.73 (14.36) 62–132	68.97 (13.84) 40–85	103.79 (10.16) 87–132	39.18 (35.31) 1–110	97.15 (19.82) 1–124	67.69 (19.23) 40–102	98.59 (15.62) 45–130
TEGI ≤ 85 (n = 93)	55.9%	21.5%	91.89 (16.64) 51–128	102.83 (13.73) 63–132	79.17 (18.26) 40–112	104.39 (11.42) 69–132	45.42 (31.37) 1–84	104.49 (9.64) 86–124	75.05 (20.37) 40–119	100.12 (15.68) 45–130
NWR ≤ 85 (n = 91)	54.9%	22.5%	91.80 (14.63) 59–128	102.74 (15.02) 51–132	79.42 (18.30) 40–112	103.94 (12.12) 60–132	57.12 (40.10) 1–116	97.10 (21.28) 1–124	67.26 (14.68) 40–84	104.24 (9.67) 86–130
CELF-4 ≤ 80 (n = 48)	68.8%	26.1%	85.96 (14.53) 51–125	101.81 (14.48) 62–132	64.75 (13.25) 40–79	102.21 (11.27) 81–132	34.01 (34.13) 1–107	94.09 (23.81) 1–124	65.19 (17.81) 40–102	96.90 (17.18) 40–130
TEGI ≤ 80 (n = 87)	58.6%	21.3%	91.07 (16.39) 51–128	102.87 (13.80) 63–132	78.07 (18.09) 40–111	104.05 (11.58) 69–132	42.79 (30.72) 1–80	103.72 (10.25) 82–124	74.03 (20.20) 40–119	99.74 (15.85) 45–130
NWR ≤ 80 (n = 72)	63.9%	22.3%	90.56 (14.43) 59–128	102.08 (15.07) 51–132	77.97 (19.14) 40–112	101.92 (13.60) 56–132	52.58 (41.06) 1–116	94.68 (23.55) 1–124	63.11 (13.75) 40–79	101.98 (11.26) 81–130
IV. At least two of the following: Receiving SLP services, CCC-2 ≤ 85, CELF ≤ 85, TEGI ≤ 85, NWR ≤ 85 (n = 97)	69.1%	12.3%	88.43 (14.93) 51–128	105.29 (12.46) 66–132	78.24 (17.22) 40–112	105.64 (9.98) 81–132	51.32 (36.89) 1–116	102.31 (12.54) 65–124	72.60 (19.22) 40–118	102.32 (12.93) 64–124
V. All of the following: Receiving SLP services, CCC-2 ≤ 80, CELF ≤ 80, TEGI ≤ 80, NWR ≤ 80 (n = 14)	100%	30.4%	72.93 (6.900) 59–80	100.30 (14.78) 51–132	58.86 (15.29) 40–78	97.19 (16.71) 40–132	25.83 (31.86) 1–78	85.96 (32.44) 1–124	56.10 (14.30) 40–78	92.89 (19.86) 40–131

Note. The plus sign (+) indicates affected, and the minus sign (–) indicates unaffected, as indicated by the reference standard (total n = 251) . SLP = speech-language pathology; CCC-2 = Children's Communication Checklist–Second Edition; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; TEGI = Test of Early Grammatical Impairment; NWR = Dollaghan and Campbell's (1998) nonword repetition task.

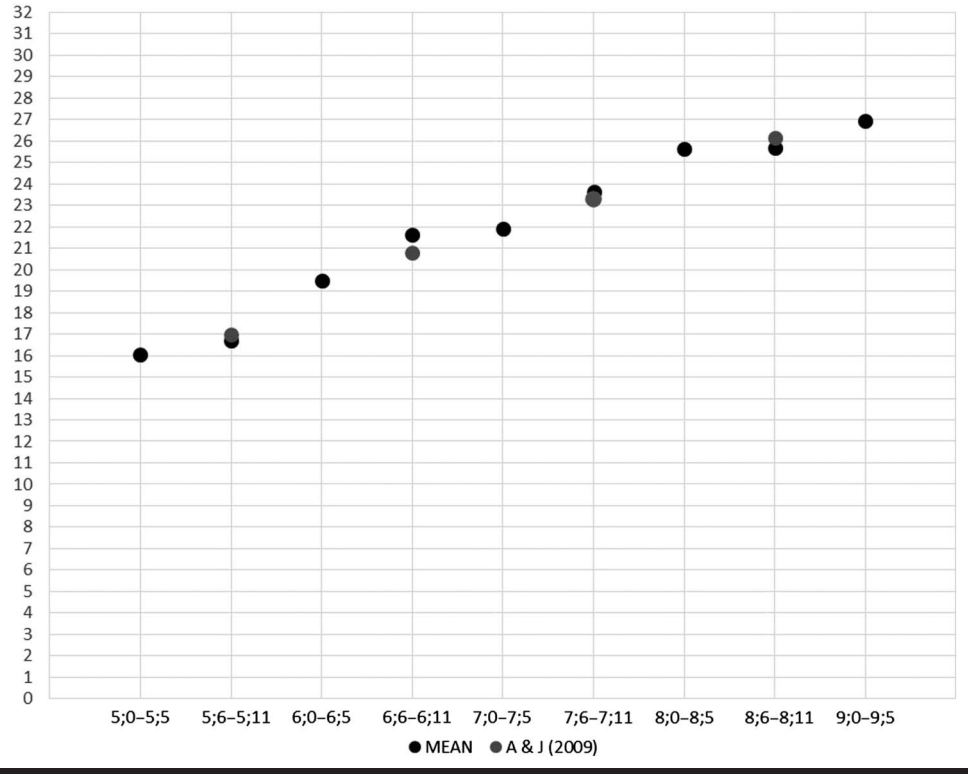
with the best performing behavioral checklists and inventories widely used in pediatric psychology (see Youngstrom, 2014). AUCs were similarly high for past tense marking—with notable exceptions in those instances when affected status was based on poor performance on the NWR measure (.843–.909 vs. .736–.757). This suggests that, even though profiles consisting of poor performance on both finite verb and NWR measures were common among children who met a variety of the reference standards, we considered—in fact, this represented the most common combination of indicators—these two measures appeared to tap into different dimensions of linguistic vulnerability. This would be consistent with characterizations that NWR may represent a stronger clinical marker for dyslexia status than language impairment status, even though these two conditions frequently co-occur (Bishop, McDonald, Bird, & Hayiou-Thomas, 2009; Catts, Adlof, Hogan, & Weismer, 2005; Catts, Adlof, & Weismer, 2006). In contrast, sentence recall appeared to align reasonably well with both the TEGI and NWR, suggesting sentence recall might be a stronger choice when screening for broadly based profiles of psycholinguistic vulnerability.

Our index measures were the least aligned with the reference standard defined by receipt of language services followed closely by the standard based on parental ratings of communicative concerns. As a group, children identified as having a language impairment by these reference standards presented with less severe behavioral symptoms than when language impairment status was determined using behavioral criteria. In practical terms, this means that a significant number of children who do not present with the kinds of language symptoms captured by the CELF-4, TEGI, or NWR might be overlooked by sentence recall and past tense marking measures.

The CELF-4 reference standard demonstrated the highest level of overlap with the other reference standards. At face value, this appears to provide an endorsement of using the CELF-4 as the reference standard for language impairment in future diagnostic accuracy studies. In contrast, the receipt of services reference standard was the least consistent with the other standards, with the largest portion of their cases identified by only that reference standard (22%). The observed lack of convergence in our study was generally consistent with previous investigations that reported less-than-optimal alignment between behavioral measures of children's language performance and service provision/parental concerns (Morgan et al., 2016; Schmitt et al., 2014; Zhang & Tomblin, 2000). The source of these discrepancies is open to speculation requiring additional investigations to reach resolution.

Our data provide important replications of previous studies that have used these particular index measures. For example, the obtained means for the sentence recall measure across different age groups aligned very closely with those reported in Archibald and Joanisse (see Figure 2). Likewise, obtained age-referenced means for the past tense marking measure were highly congruent with normative information provided in the TEGI manual (Rice & Wexler,

Figure 2. Observed sentence recall mean raw scores in the current study compared to values reported in Archibald and Joanisse (2009).



2001). These independent replications combined with high levels of interrater scoring consistency and stability observed in the current study sample suggest these particular index measures are reliable enough to be translated into clinical use as language screeners.

Our study has several limitations that encourage caution when interpreting results. Our results are limited to the particular reference standards we selected. Even though these reference standards were consistent with the different types of reference standards currently under consideration in diagnostic accuracy studies, they were not exhaustive of the range of empirical or clinical reference standards available. Additional research may reveal limitations when using sentence recall and past tense marking to screen for other reference standards of language impairment. Similarly, our results are limited to the age range we examined, and it is an open question whether our index measures would perform as well with older students. It is likely that different index measures would be more successful when screening older elementary, middle school, and high school students. Participation in our study samples required two stages of active parental consent and child assent rather than potentially more robust "opt-out" or negative consent procedures, as used in some studies (cf. Tomblin et al., 1997). This design element probably introduced selection biases into our sample. Participating school administrators were, understandably, protective of teacher time, and this prevented us from collecting data from teachers that could have complemented the standardized parent ratings of communicative competence. We were similarly restricted from collecting information regarding the type, frequency, duration, or goals associated with the interventions children were receiving. These details could have shed some light on the limited alignment between our index measures and the linguistic and nonlinguistic signs and symptoms entailed by children's enrollment in language services. Another limitation associated with our evaluation of sentence recall and past tense marking is our outcomes only apply to monolingual English-speaking children. Additional investigations are needed to address these limitations and to explore potential adaptations of these clinical markers for use with more diverse communities. Despite these limitations, however, the results of our investigation suggest that, when the reference standard for language impairment in early elementary students does eventually arrive in the field, sentence recall and past tense marking measures appear to be well positioned to screen for it.

Acknowledgments

The National Institute on Deafness and Other Communication Disorders provided funding for this study (Grant R01DC011023 awarded to Sean M. Redmond). We are greatly indebted to the children and their families. We also acknowledge the school personnel who graciously extended to us their time and space: Deb Andrews, Earl Arnoldson, Francis Battle, Dan Bergman, Christine Bergquist, Mariann Broadhead, Adam Eskelson, Deb Luker, Shelley Halverson, Lisa Holmstead, Kenneth Limb, Catherine Kamphaus, Julie Miller, Kim Payne, Rebecca Pittam, April Reynolds, Ken Westwood, and Jared Wright. Pamela Mathy, Mary Foye, Laurie Fue, and Mark Cantor provided referrals to the project from the University of Utah Speech-Language-Hearing Clinic. Several graduate research assistants deserve recognition for their contributions: David Aamodt, Peter Behnke, Hannah Caron, Kimber Campbell, Jessica Carrizo, Faith Denzer, Olivia Erickson, Kristina Fassbender, Micah Foster, Elizabeth Hafen, Lyssandra Harker, Kristin Hatch, Nathan Lily, Amy Ludlow, Kristi Moon, Elie Muyankindi, McKenzie Rohde, Michelle Stettler, and Amy Wilder. We are also appreciative of the help provided by student volunteers: Aaron Allsop, Josh Anderson, Emily Barriochoa, Sadie Brayton, Mari Broadhead, Natalie Bryson, Tomas Chavez, Caroline Champougny, Amy Clark, Hannah Clements, Chantel Cook, Esperanza Cordero, Jessica Cox, Clint Curry, Jackie Dailey, Margaret Despres, Bailey Farnsworth, Jeff Glauser, Kayla Greenburg, Rachael Gorringe, Courtney Hammond, Lindzi Helgesen, Eliza Hintze, Brayden Jensen, Rosalyn Kirkendall, Dan Knodel, Jenna Madsen, Paul McGill, Madison McHaley, Molly Menzie, Madison Migacz, Jessica Miner, Callie Mortensen, Candice Paulsen, Callie Payne, Elizabeth Pinner, Melissa Phillips, Julie Platt, Nina Pryor, Mallory Puentes, Elizabeth Redding, Marisol Robinson, Courtney Robison, Kirsten Schermerhorn, Lauren Schreck, Dirk Schroeder, Alison Shimko, Natalie Smith, Sarah Symes, Chris Taylor, Katie Thompson, and Debbie Weaver.

References

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research, 52*(4), 899–914.

Ash, A. C., & Redmond, S. M. (2014). Using finiteness as a clinical marker to identify language impairment. *SIG 1 Perspectives on Language Learning and Education, 21,* 148–158.

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminate function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185–1192.

Beitchman, J. H., Nair, R., Clegg, M., Patel, P. G., Ferguson, B., Pressman, E., & Smith, A. (1986). Prevalence of speech and language disorders in 5-year old kindergarten children in the Ottawa-Carleton region. *Journal of Speech and Hearing Disorders, 51*(2), 98–110.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44,* 133–146.

Bishop, D. V. M. (2006). *The Children's Communication Checklist* (2nd ed.). San Antonio, TX: The Psychological Corporation.

Bishop, D. V. M., McDonald, D., Bird, S., & Hayiou-Thomas, M. E. (2009). Children who read words accurately despite language impairment: Who are they and how do they do it? *Child Development, 80*(2), 593–605.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS ONE, 11*(7), e0158753.

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery, 159*(6), 1638–1645.

Catts, H. W., Adlof, S. M., Hogan, T. P., & Weismer, S. E. (2005). Are specific language impairment and dyslexia distinct disorders? *Journal of Speech, Language, and Hearing Research, 48,* 1378–1396.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49,* 278–293.

Conti-Ramsden, G. (2008). Heterogeneity of specific language impairment in adolescent outcomes). In C. F. Norbury, J. B. Tomblin, & D. V. M. Bishop (Eds.), *Understanding developmental language disorders: From theory to practice* (pp. 117–130). New York, NY: Psychology Press.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(6), 714–748.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1136–2246.

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*(4), 865–878.

Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology, 27,* 1329–1351.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal, 47*(4), 458–472.

Gillam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language (TNL)*. Austin, TX: Pro-Ed.

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

Greensdale, K. J., Plante, E., & Vance, R. (2009). The diagnostic accuracy and construct validity of the Structured Photographic Expressive Language Test—Preschool: Second Edition. *Language, Speech, and Hearing Services in Schools, 40*(2), 150–160.

Haahr, M. (2006). *Random.org: True random number service*. Retrieved from http://www.random.org

Johnson, C., Beitchman, J. H., Young, A., Escobar, M., Atkinson, L., Wilson, B., … Wang, M. (1999). Fourteen-year follow-up of children with and without speech/language impairments: Speech/language stability and outcomes. *Journal of Speech, Language, and Hearing Research, 42*(3), 744–760.

Jones Moyle, M., Karasinski, C., Ellis Weismer, S., & Gorman, B. K. (2011). Grammatical morphology in school-age children with and without language impairment: A discriminate function analysis. *Journal of Speech, Language, and Hearing Research, 42*(4), 550–560.

Miller, J. F. (1996). The search for the phenotype of disordered language performance. In M. L. Rice (Ed.), *Toward a genetics of language* (pp. 297–314). Mahwah, NJ: Erlbaum.

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities disproportionally underrepresented in special education: Longitudinal evidence across five disability categories. *Education Research, 44*(5), 278–292.

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

Naglieri, J. A. (2003). *Naglieri Nonverbal Ability Test: Individual administration (NNAT-Individual)*. San Antonio, TX: Harcourt Assessment.

Newcomer, P. L., & Hammill, D. D. (2008). *Test of Language Development–Primary: Fourth Edition (TOLD-P:4)*. Austin, TX: Pro-Ed.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., … Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *The Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257.

Oram, J., Fine, J., Okamoto, C., & Tannock, R. (1999). Assessing the language of children with attention deficit hyperactivity disorder. *American Journal of Speech-Language Pathology, 8,* 72–80.

Parriger, E. (2012). *Language and executive functioning in children with ADHD* (Doctoral dissertation). the Netherlands: University of Amsterdam.

Pawlowska, M. (2014). Evaluation of three proposed markers for language impairment in English. A meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research, 57,* 2261–2273.

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15*(3), 247–254.

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.

Redmond, S. M. (2005). Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics & Phonetics, 19*(2), 109–127.

Redmond, S. M., Thompson, H. L., & Goldstein, S. (2011). Psycholinguistic profiling differentiates specific language impairment from typical development and from attention-deficit/hyperactivity disorder. *Journal of Speech, Language, and Hearing Research, 54*(1), 99–117.

Rice, M. L., Smith, S. D., & Gayán, J. (2009). Convergent genetic linkage and associations to language, speech and reading measures in families of probands with specific language impairment. *Journal of Neurodevelopmental Disorders, 1*(4), 264–282.

Rice, M. L., Tomblin, J. B., Hoffman, L., Richman, W. A., & Marquis, J. (2004). Grammatical tense deficits in children with SLI and nonspecific language impairment. *Journal of Speech, Language, and Hearing Research, 47,* 816–834.

Rice, M. L., & Wexler, K. (2001). *Test of Early Grammatical Impairment (TEGI)*. San Antonio, TX: The Psychological Corporation.

Rice, M. L., Wexler, K., & Hershberger, S. (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41*(6), 1412–1431.

Rice, M. L., Zubrick, S. R., Taylor, C. L., Hoffman, L., & Gayán, J. (2018). Longitudinal study of language and speech of twins at 4 and 6 years: Twinning effects decrease, zygosity effects disappear, and heritability increases. *Journal of Speech, Language, and Hearing Research, 61,* 79–93.

Schmitt, M. B., Justice, L. M., Logan, J. A., Schatschneider, C., & Bartlett, C. W. (2014). Do the symptoms of language disorder align with treatment goals? An exploratory study of primary-grade students' IEPs. *Journal of Communication Disorders, 52,* 99–110.

Sciberras, E., Mueller, K. L., Efron, D., Bisset, M., Anderson, V., Schilpzand, E. J., … Nicholson, J. M. (2014). Language

problems in children with ADHD: A community-based sample. *Pediatrics, 133*(5), 793–800.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. San Antonio, TX: Pearson Education Inc.

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

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26.

Tomblin, J. B. (2006). A normativist account of language-based learning disability. *Learning Disabilities Research and Practice, 21*(1), 8–18.

Tomblin, J. B., & Nippold, M. A. (2014). *Understanding individual differences in language development across the school years*. New York, NY: Psychology Press.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260.

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

Wittke, K., & Spaulding, T. J. (2018). Which preschool children with specific language impairment receive language intervention? *Language, Speech, and Hearing Services in Schools, 49*(1), 59–71.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3,* 32–35.

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39*(2), 204–221.

Zhang, X., & Tomblin, J. B. (2000). The association of intervention receipt with speech-language profiles and social-demographic variables. *American Journal of Speech-Language Pathology, 9*(4), 345–357.