LSHSS

Research Article

Relationships Between Narrative Language Samples and Norm-Referenced Test Scores in Language Assessments of School-Age Children

Kerry Danahy Ebert[a] and Cheryl M. Scott[a]

Purpose: Both narrative language samples and norm-referenced language tests can be important components of language assessment for school-age children. The present study explored the relationship between these 2 tools within a group of children referred for language assessment.
Method: The study is a retrospective analysis of clinical records from 73 school-age children. Participants had completed an oral narrative language sample and at least one norm-referenced language test. Correlations between microstructural language sample measures and norm-referenced test scores were compared for younger (6- to 8-year-old) and older (9- to 12-year-old) children. Contingency tables were constructed to compare the 2 types of tools, at 2 different cutpoints, in terms of which children were identified as having a language disorder.
Results: Correlations between narrative language sample measures and norm-referenced tests were stronger for the younger group than the older group. Within the younger group, the level of language assessed by each measure contributed to associations among measures. Contingency analyses revealed moderate overlap in the children identified by each tool, with agreement affected by the cutpoint used.
Conclusions: Narrative language samples may complement norm-referenced tests well, but age combined with narrative task can be expected to influence the nature of the relationship.

Language assessment for school-age children with suspected language disorders may be viewed as a daunting task. Across the multiple dimensions of the complex construct of language, clinicians must distinguish disordered from typical development, characterize strengths and weaknesses, and derive treatment goals. To tackle this challenge, speech-language pathologists will need a range of assessment tools.

Two of the most commonly used language assessment tools, norm-referenced language tests and criterion-referenced narrative language samples, are the focus of the present work. By definition, *norm-referenced tests* provide the opportunity to compare children with peers, facilitating the identification of language disorders (Paul & Norbury, 2012). Norm-referenced tests exist to assess a variety of language skills across all language levels (i.e., word, sentence, and discourse) as well as a wide range of ages (see Spaulding, Plante, & Farinella, 2006, for a review). In contrast to norm-referenced tests, *criterion-referenced language assessments* measure how well a child has acquired a particular language skill, typically in reference to a criterion for adequate performance. Narrative language sampling is a commonly used criterion-referenced language assessment in which a child is asked to compose or retell a story. Like norm-referenced tests, narrative language sample analysis can span a variety of skills across all language levels.

It is important for clinical practitioners to understand how these two types of language assessment tools intersect in order to optimize their assessments of school-age children. In the present study, we explore the relationship between norm-referenced test scores and oral narrative language samples in a group of school-age children referred for language assessment. To frame this study, we first compare the features of each type of language assessment tool. We then review previous work considering the relation between norm-referenced language tests and language samples before introducing our purpose and hypotheses.

[a]Rush University, Chicago, IL
Correspondence to Kerry Danahy Ebert: Kerry_ebert@rush.edu
Editor: Marilyn Nippold
Associate Editor: Stacy Wagovich
Received March 23, 2014
Revision received June 19, 2014
Accepted August 6, 2014
DOI: 10.1044/2014_LSHSS-14-0034

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

Norm-Referenced Tests in Language Assessment

Norm-referenced tests have enjoyed a position of relative prominence in language assessment. They form the core of established criteria for diagnosing language disorders (EpiSLI; Tomblin, Records, & Zhang, 1996) and are standard practice in qualifying children for research studies on language disorders (see Spaulding et al., 2006). Clinically, norm-referenced tests are used widely, at least in part due to the common practice of requiring test scores below a particular cutoff in order to qualify for school-based speech-language services (Betz, Eickhoff, & Sullivan, 2013). Cutoff scores are also frequently required by third-party payers in clinical settings and have been used to determine the severity of language impairment (see Spaulding, Szulga, & Figueroa, 2012).

Finally, norm-referenced tests may offer the ability to quickly sample a variety of language levels and modalities. *Language level* refers to the size of the linguistic unit to be manipulated by an examinee. Some instruments test at the word level, as when a child is asked to identify a picture named by the examiner or pick two words that are related from a group of four. Other tests require comprehension or production at the sentence level, such as following a direction or making up a sentence using a word. Discourse skills are tested when a child is asked to listen to a short paragraph and then answer several questions. *Modality* refers to the distinction between oral and written language, and norm-referenced instruments exist to test both modalities. Oral language tests require only listening and speaking; written language tests may require a child to read real or nonsense words aloud, read text and answer questions about content, or write in response to a prompt or picture.

Norm-referenced tests have a number of limitations, however, and other assessment tools may be needed as complements or alternatives to these tests. First, norm-referenced tests lack ecological validity; they reflect performance in an artificial testing situation rather than a real-life situation. Many tasks are decontextualized; for example, a child may be unable to make up a grammatically correct sentence using the word *because* when asked to do so as an isolated task, but the same child uses *because* correctly in a personal story told to a friend. On comprehensive language tests with several subtests, many skills are probed but each is tested with only a few items, leading to broad rather than deep coverage. Due to these validity and coverage concerns, norm-referenced tests are said to be ill suited to the development of language treatment targets (Paul & Norbury, 2012).

Norm-referenced tests are also poorly suited to some populations. Valid administrations of norm-referenced tests have strict behavioral requirements for examinees, such as remaining seated at a table, engaging with pictured stimuli, responding only when prompted, and performing tasks until failure without feedback from the examiner. Many children have difficulty complying with these requirements (Costanza-Smith, 2010). Children from culturally and linguistically diverse backgrounds may also be ill suited to assessment with norm-referenced tests (De Lamo White & Jin, 2011). Valid use of test norms requires examinees to match the normative sample, an assumption that is frequently not met given the diversity of cultural and linguistic backgrounds both within and outside of the United States.

Narrative Language Samples in Language Assessment

Language samples, and particularly narrative language samples, may offer a valid complement or even alternative to norm-referenced testing. Language samples address many of the weaknesses of norm-referenced testing: They provide rich, in-depth information about a child's use of language in real-world situations (Costanza-Smith, 2010; Hewitt, Hammer, Yont, & Tomblin, 2005), resulting in strong ecological validity and the ability to derive language treatment targets; they place very few behavioral requirements on examinees, allowing for flexible use across children of diverse ages and types of impairment (Costanza-Smith, 2010); and they have been shown to be a valid assessment for diverse populations including bilingual children (Restrepo, 1998) and speakers of nonstandard dialects (Stockman, 1996).

Narrative language samples can provide a wealth of information, both at the word and sentence levels (termed *microstructural* information) and at the global level, across utterances (termed *macrostructural* information). Although macrostructural measures such as story grammar ratings provide important information, microstructural measures have been shown to better distinguish children with language disorders from typically developing peers (Liles, Duffy, Merritt, & Purcell, 1995). Microstructural measures can be used to reflect a variety of linguistic aspects of a narrative, including sentence length, typically measured by mean length of utterance (MLU); lexical diversity, typically measured by the number of different words (NDW) present in a sample; syntactic complexity, measured by the number of clauses present per utterance (the Subordination Index, or SI, also known as *clausal density*); overall productivity, measured by the total number of words (TNW) or utterances; and grammatical error rates, measured by indices such as the number of omitted bound morphemes or the number of erroneous or omitted words. Microstructural measures derived from narrative language samples have been shown to be sensitive for detecting and discriminating several types of language disorders, including autism spectrum disorder (ASD; Manolitsi & Botting, 2011) and specific language impairment (SLI; Norbury & Bishop, 2003).

Thus, narratives offer a flexible, ecologically valid, informative, and sensitive tool for language assessment. However, in comparison to norm-referenced tests, normative data for microstructural narrative measures may be more difficult for practicing clinicians to access. Existing normative data sets are also variable in size, coverage across ages, and language sample context, which may limit clinicians' ability to use them in assessments (Hewitt et al., 2005; Norbury & Bishop, 2003). The time needed to transcribe and analyze samples has also been cited as a barrier to the use of narrative language samples (Hewitt et al., 2005). Finally, narratives are not a "one size fits all" solution to language assessment across the wide range of age

groups served by practicing speech-language pathologists. Particular narrative tasks may have limited age applicability. A task that elicits a good narrative effort from a 7-year-old may not be the best task to use with an older child who is capable of understanding more complex character development and plots (Nippold, 2014). And, in the late-elementary years and beyond, expository and persuasive genres offer additional choices for language sampling (Nippold, Mansfield, Billow, & Tomblin, 2008; Scott & Windsor, 2000).

Language Assessment in Practice

Thus far, we have highlighted the strengths and weaknesses of norm-referenced tests and narrative language samples as assessment tools. We now consider how these tools are used in clinical practice. Surveys of clinical practitioners have indicated that although both norm-referenced tests and language samples (of any type) are used routinely in the assessment of language disorders, norm-referenced tests have more consistent use (Caesar & Kohler, 2009; Wilson, Blackmon, Hall, & Elcholtz, 1991). For example, Wilson et al. (1991) reported that only one of the 266 school-based clinicians they surveyed did *not* use norm-referenced tests. Language samples, in contrast, were used in an average of 75% of assessments, making them frequently but not consistently used.

However, when the focus is narrowed to school-age children and to narrative language samples, the frequency of language sampling in assessments appears to drop. In a recent survey of public school clinicians, Caesar and Kohler (2009) found that 94% of respondents reported using language samples *often* or *sometimes* in assessments. However, the reported use of language sampling was inversely related to the age of the child to be assessed; clinicians working with preschool children used language sampling most frequently and clinicians working with high school students used it least often. In a previous survey of school-based speech-language pathologists (Hux, Morris-Friehe, & Sanger, 1993), a similar decrease in the use of language samples for middle and high school students was reported. Furthermore, Hux et al. (1993) reported that conversational language samples were the most frequently collected type, followed by picture descriptions. Only about half the respondents used forms of narrative language sampling in assessments.

It is also important to note that the survey data suggest that when language sampling is used for assessment, it is used in conjunction with norm-referenced testing (Wilson et al., 1991). The strengths and weaknesses of these tools, as outlined above, suggest they may indeed be good complements to each other. However, information about the exact nature of the relationship between language samples, particularly narratives, and norm-referenced tests is needed to support this clinical practice.

Associations Between Narrative Language Samples and Norm-Referenced Tests

The relationship between norm-referenced tests and narrative language samples in school-age children has been considered, though previous studies vary substantially in the ages and diagnostic status of participants, in the norm-referenced tests administered, and in the narrative measures included in analyses. Perhaps as a result, reported associations between the two types of tools have varied. We are aware of at least three studies that have reported correlations between narrative language sample measures and norm-referenced tests in children with and without language disorders (Bishop & Donlan, 2005; Manolitsi & Botting, 2011; Norbury & Bishop, 2003). Manolitsi and Botting (2011) collected narrative retells from small samples of Greek-speaking children with SLI and ASD, ranging in age from 4 to 13 years. Overall ratings of microstructural narrative devices (which combined vocabulary, adverbial clauses, and pronoun referencing) were significantly correlated with receptive composite norm-referenced test scores for the ASD group ($r = .70$). However, for the children with SLI, neither receptive nor expressive norm-referenced composite scores were significantly correlated with narrative microstructure ($r = .04$ and $r = .26$, respectively). The authors suggested that the relatively weak receptive language skills in children with ASD may have led to different patterns between groups, although they also acknowledged that the wide age range and small sample size in the study may have influenced results. It may also be important to note that only composite measures of narrative and language test performance were considered, rather than measures that index a specific level of language.

In contrast, Bishop and Donlan (2005) considered correlations between more fine-grained measures; these measures included norm-referenced language tests tapping sentence repetition, word finding, and receptive grammar as well as two sentence-level narrative measures (MLU and number of dependent clauses). A total of 63 English-speaking 7- to 10-year-old children with SLI participated in the study. Despite methodological differences from Manolitsi and Botting (2011), results in Bishop and Donlan (2005) were somewhat similar; correlations between the narrative measures and norm-referenced language tests, though uniformly positive, did not reach significance. The correlations between the sentence-level narrative measures were stronger with the two norm-referenced sentence-level tests (sentence repetition and receptive grammar) than with the word-level test (assessing word finding).

Finally, Norbury and Bishop (2003) considered relations between narrative language samples (told to *Frog, Where Are You?*; Mayer, 1969) and norm-referenced assessments of receptive vocabulary and sentence repetition. Participants ranged in age from 6 to 10 years and included a group of 18 children with typical language development and a combined group of children with high-functioning autism, SLI, or pragmatic language impairment ($N = 40$). The group with typical language demonstrated no significant correlations between language sample measures and norm-referenced tests, whereas the group with language disorders demonstrated significant relationships between sentence repetition and the use of complex sentences in narratives ($r = .41$), between receptive vocabulary and complex sentences in narratives ($r = .37$), and between receptive vocabulary and the

amount of relevant semantic information in the narrative (r = .35).

The results of these three studies are somewhat variable and may be viewed as a starting point for additional work in this area. Two additional studies support the presence of significant associations between (nonnarrative) language samples and norm-referenced test scores. Condouris, Meyer, and Tager-Flusberg (2003) collected play-based language samples from children with ASD, ranging in age from 4 to 14 years, and compared them with scores on norm-referenced vocabulary and omnibus language assessments. Numerous positive correlations were found between language sample measures such as MLU and NDW and norm-referenced test scores. Finally, Nippold, Mansfield, Billow, and Tomblin (2009) analyzed expository language sample measures in a large sample (N = 426) of 10th-grade students with and without language disorders. A composite norm-referenced score of syntax was significantly related to several measures of syntax from the language samples, including mean length of T-unit and clausal density.

To our knowledge, only one study has used a method other than correlation analysis to consider the relation between language samples and norm-referenced tests. Condouris et al. (2003) also created contingency tables to consider the classification of individual children with ASD as impaired or typical on each type of language measure. Norm-referenced test scores were categorized as falling more than 2 standard deviations (*SD*s) below the mean, between 1 and 2 *SD*s below the mean, or within 1 *SD* of the mean; measures from language samples were similarly categorized, using comparisons to the reference databases accompanying the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012) software. Overall, the participants performed more poorly in comparison to norms on the language samples than on the norm-referenced tests, which the authors suggest may relate to the difficulty children with ASD exhibit in using language in less structured situations.

Study Purpose and Hypotheses

The associations between norm-referenced tests and narrative language samples reported in the literature to date have ranged substantially. Participants have varied in age, language spoken, the type of narrative task, and diagnostic status. The skills assessed by the norm-referenced tests included in studies have varied as well (including overall composites and specific measures of syntax, sentence recall, vocabulary, and word finding). Further work is needed to examine the effects of these variables.

In addition, not all modalities and levels of language have been tapped by the norm-referenced tests in the existing literature base. More specifically, the relationships between narrative language sample measures and norm-referenced assessments of written language, relational vocabulary, and discourse-level comprehension have not yet been reported. Similarly, a limited set of narrative analyses have been explored in the existing literature. Notably absent from prior work is an examination of grammatical error rates using measures such as omitted words, omitted bound morphemes, and word-level errors. Grammatical errors may be particularly sensitive indicators of a language disorder (e.g., Scott & Windsor, 2000), and their relationship with norm-referenced testing should be explored.

Consideration of the relationship between narratives and norm-referenced tests has also been largely limited to correlation analyses. Such analyses are useful for identifying group-level associations but do not provide information on individual children's performance or on the overall frequency with which each tool identifies disordered language. For practicing clinicians, whose first decision is often whether or not a language disorder is present, classification rates are likely to be even more important than correlations.

The purpose of this study was to explore the relationship between norm-referenced language tests and storybook narrative language samples within a diverse group of school-age children referred for language assessment. We extend prior work in this area in several ways. First, we explicitly consider the role of age in the relation between these two types of tools. We also include a broad range of norm-referenced assessments, including measures of written language and discourse-level comprehension, and a broad range of narrative measures, including grammatical errors. We consider the influence of the varying types of norm-referenced and narrative measures by classifying each tool according to the level of language measured (i.e., word, sentence, or discourse level) in addition to modality (oral or written). Finally, we extend the analysis of the relationship beyond correlations by comparing the classification of children into disordered versus typical categories using each type of tool at two different cutpoints.

The study is exploratory, allowing for a range of possible outcomes. However, we propose four main hypotheses, which are detailed in the subsections that follow.

Hypothesis 1. Age will play a significant role in the relationship between narratives and norm-referenced tests. More specifically, storybook narrative language samples will be more closely related to norm-referenced tests for younger school-age children than for older children, reflecting a potential interaction of age and narrative task.

Hypothesis 2. Associations between norm-referenced tests and narrative measures will be strongest on measures that index the same level of language. For example, MLU (a sentence-level measure) in the narrative language sample will be more highly correlated with sentence-level norm-referenced tests and subtests (e.g., formulating or recalling sentences) than word-level measures (such as receptive vocabulary). This hypothesis has its roots in the existing literature (Bishop & Donlan, 2005; Norbury & Bishop, 2003), though it has not yet been rigorously explored.

Hypothesis 3. We expect some significant associations between written language tests and narrative measures, but such associations may be weaker due to the difference in modality across the two measures. Due to the lack of past literature comparing written language tests to narrative language samples, stronger hypotheses in this area are not yet supported.

Hypothesis 4. We expect overlap but not perfect agreement between norm-referenced tests and narrative language samples in terms of which children they classify as having a language disorder. Because both tools assess the construct of language, we expect some children to perform poorly and others to perform well on both tools. However, it is unlikely that there will be perfect agreement because of the differences between the tools. That is, some children may have particular difficulty with the structured format of norm-referenced tests and perform poorly on them; others may have difficulty using language knowledge in an unstructured situation and perform more poorly on narrative measures. Or, children's reactions to format and task aside, it may be that the two types of tools tap different types of skills. We do not have an a priori reason to expect one type of tool to identify more children than the other. Also, although the classification agreement between the tools may be influenced by the cutpoint used for identifying language disorders, we do not have an a priori reason to expect better agreement at a particular cutpoint.

Method

This study is a retrospective analysis of clinical records from the speech-language clinic at a major urban academic medical center. Qualifying records from speech-language assessments that included both norm-referenced language testing and language sampling were located, and information from these assessments was extracted to a de-identified database.

Participants

All participants had been referred for speech-language evaluation within the last 10 years. Evaluations were scheduled on the basis of parental concern coupled with physician referral. Children were eligible for the study if (a) their age fell between 6;0 (years;months) and 17;11, (b) they had completed at least one measure of language during the speech-language evaluation, and (c) they had no evidence of intellectual disability. Possible indicators of intellectual disability that would preclude study inclusion were a recorded diagnosis of cognitive delay or of a syndrome that indicates intellectual disability such as Down syndrome or fragile X, recorded individual education plan eligibility under a category indicating intellectual disability, or recorded IQ scores outside the normal range. These criteria resulted in a total of 138 eligible clinical records. For 85 of these children, a language sample transcript was located.

For the present study, only those children with both a narrative language sample transcript, obtained using a wordless picture book, and at least one composite or subtest score on a norm-referenced language test were included. The resulting sample included a total of 73 children, ranging in age from 6;0 to 12;8. Because of this wide age range as well as our prediction that age would influence results, participants were divided into two groups on the basis of age. The group was split at age 9 years because this age aligns with shifts in standardized testing procedures (e.g., the appropriate subtests of the Clinical Evaluation of Language Fundamentals [CELF] change at age 9) and with previous investigations in which narrative language samples have been shown to separate children with and without language disorders (Botting, 2002; Norbury & Bishop, 2003).

The younger participant group consisted of 50 children (13 females, 37 males) aged 6;0–8;11. The older participant group consisted of 23 children (seven females, 16 males), aged 9;1–12;8. In both groups, the recorded language evaluations indicated that participants ranged in terms of the presence and severity of language impairment. On the basis of the summary statements made by the clinicians (i.e., the first and second authors) at the conclusion of all testing, 30% of children were deemed to have no oral language impairment, and the remaining 70% had a language impairment with severity assignments ranging from mild to severe.

Finally, the sample was linguistically diverse, with 17 participants having significant exposure to a language other than English and an additional 17 speaking African American English dialect. For all children, English was the dominant language, and reasonable procedures for assessing children from linguistically diverse backgrounds were followed by the first and second authors (e.g., adjusted scoring on norm-referenced tests to allow for dialectal patterns).

Table 1 summarizes characteristics of the participants in the younger and older groups in terms of age, norm-referenced test scores, and language sample measures. Values are displayed separately for the two groups to allow group comparisons. The number of participants who completed each test is also listed; these numbers vary as the tests administered were selected by the clinician at the time of the evaluation.

Participant standard scores on language tests ranged from very impaired to within the average range. On norm-referenced tests measuring oral language skills, the older and younger groups showed similar mean scores. On measures of written word identification (i.e., Word Identification and Word Attack from the Woodcock Reading Mastery Tests [WRMT; Woodcock, 1987]), the older group demonstrated lower mean scores than the younger group. These observations were confirmed using an independent samples t test comparing the mean score on each measure for the older versus the younger group. The groups differed significantly on the Word Identification subtest, $t(41) = 2.20$, $p = .034$, $d = 0.73$, and on the Word Attack subtest, $t(28) = 2.42$, $p = .023$, $d = 1.02$. For all remaining comparisons, $p > .10$.

On the language sample measures, scores would be expected to improve with age. The older group demonstrates better mean values on all measures. This advantage is statistically significant for three measures: MLU, $t(71) = -2.33$, $p = .023$, $d = -0.62$; word-level errors, $t(71) = 2.33$, $p = .023$, $d = 0.59$; and total errors, $t(71) = 2.00$, $p = .049$, $d = 0.50$. Variables on which the older and younger group showed significant differences are marked with an asterisk in Table 1.

Table 1. Participant characteristics for younger and older groups.

Measure	N		*M*		*SD*		Range	
	Younger	Older	Younger	Older	Younger	Older	Younger	Older
Age	50	23	7;5	10;1	0;10	1;0	6;0–8;11	9;1–12;8
CELF CFD	43	21	6.26	5.05	2.79	3.03	1–11	1–13
CELF WS[a]	36	—	5.72	—	2.65	—	1–12	—
CELF RS	40	19	6.10	6.00	3.51	3.07	1–18	3–13
CELF FS	38	19	6.82	6.68	3.59	2.75	1–15	3–12
CELF WC	15	19	7.87	7.32	3.29	1.49	3–13	4–10
CELF USP	15	10	7.40	7.20	2.99	4.02	2–12	3–15
PPVT	27	16	88.26	82.25	12.89	12.52	64–119	61–107
GORT FL	19	11	8.00	5.82	3.43	3.37	3–16	1–10
GORT Comp	19	10	7.21	7.10	2.18	2.23	2–11	4–10
WRMT WI*	30	13	93.37	83.38	13.73	13.57	66–117	62–109
WRMT WA*	20	10	94.25	78.90	18.64	10.25	52–121	63–98
MLU-W*	50	23	6.45	7.11	1.20	0.99	4.02–10.16	5.44–8.44
TNW	50	23	267.64	298.17	131.94	171.95	95–732	87–796
NDW	48	22	45.75	47.32	7.14	5.86	30–66	38–61
Sub. Index	50	23	1.21	1.27	0.17	0.12	0.93–1.65	1.01–1.45
Omit Bound Morph	50	23	2.14	1.26	2.71	1.91	0–10	0–7
Omit Wds	50	23	2.18	1.52	3.77	2.43	0–22	0–9
WL Error*	50	23	3.46	2.04	2.76	2.01	0–11	0–7
UL Error	50	23	1.14	1.35	1.21	1.70	0–4	0–7
Total Error*	50	23	8.92	6.17	7.19	5.52	1–31	0–18

Note. Table displays sample characteristics for both the younger and older groups of participants in terms of age, norm-referenced test scores, and language sample measures. Age is reported as years;month. Test scores are reported as standard scores (i.e., $M = 100$, $SD = 15$) for PPVT, WRMT WI, and WRMT WA. All remaining test scores are scaled scores ($M = 10$, $SD = 3$). CELF CFD = Clinical Evaluation of Language Fundamentals, Concepts and Following Directions subtest; CELF WS = CELF, Word Structure subtest; CELF RS = CELF, Recalling Sentences subtest; CELF FS = CELF, Formulated Sentences subtest; CELF WC = CELF, Word Classes—Receptive subtest; CELF USP = CELF, Understanding Spoken Paragraphs subtest; PPVT = Peabody Picture Vocabulary Test; GORT FL = Gray Oral Reading Test, Fluency index; GORT Comp = Gray Oral Reading Test, Comprehension index; WRMT WI = Woodcock Reading Mastery Tests, Word Identification subtest; WRMT WA = Woodcock Reading Mastery Tests, Word Attack subtest; MLU-W = MLU in words; TNW = total number of words; NDW = number of different words; Sub. Index = Subordination Index; Wds = words; WL = word level; UL = utterance level.

[a]CELF WS is administered to children age 9 years and older and is therefore not reported for the older group.

*Significant difference between the younger and older groups ($p < .05$).

Procedure

Study procedures were approved by the Institutional Review Board at Rush University Medical Center. All clinical reports generated by either of the authors within the past 10 years were located, along with associated language sample transcripts. Reports were reviewed for study eligibility by a trained research assistant. Eligible reports were assigned an identification code and relevant information was extracted into a de-identified database. Language sample transcripts were assigned the same identification code and also de-identified. Coding of language sample transcripts was conducted separately from coding of language tests.

Coding of assessment reports. Three main types of variables were extracted from qualifying assessment reports: background information, clinician diagnostic judgments, and scores from norm-referenced tests. Background information included chronological age, gender, relevant medical diagnoses (e.g., attention-deficit/hyperactivity disorder, hearing loss), and history of speech-language treatment services in educational and clinical settings. Clinician diagnostic judgments included the presence, severity, and type (expressive or mixed receptive-expressive) of oral language impairment, as well as the presence of impairment in reading, writing composition, phonological processing, spelling, and speech production. In addition, all test scores included in the assessment report were extracted to the database. A total of 209 different subtest and composite scores, derived from 23 distinct norm-referenced language tests, were found in the assessment reports.

For the present study, a subset of these norm-referenced language tests was used. Consistent with our purpose, we selected only tests of oral or written language skill (excluding assessments of areas such as phonological awareness, pragmatic language, and nonverbal intelligence). We also excluded norm-referenced tests that had been administered to fewer than 20 children in the study sample of 73 children. Finally, we combined scores from previous versions of tests with scores from the most current version (e.g., scores from the 20 children who completed the Gray Oral Reading Test—4th Edition [GORT] were combined with those of the 10 children who completed the 5th edition of the same test; see Condouris et al., 2003, for a similar procedure).

These procedures resulted in the inclusion of a total of 11 subtest scores from four distinct norm-referenced tests in our analyses. The norm-referenced tests included in the present study are the CELF (3rd ed., 4th ed., and Preschool

ed.; Semel, Wiig, & Secord, 1995, 2003, and 2004, respectively), including the Concepts & Following Directions, Word Structure, Recalling Sentences, Formulated Sentences, Word Classes—Receptive, and Understanding Spoken Paragraphs subtests; the Peabody Picture Vocabulary Test (PPVT; 3rd and 4th eds.; Dunn & Dunn, 1997, 2007, respectively); the GORT (4th and 5th eds.; Wiederholt & Bryant, 2001, 2012, respectively), including the Fluency and Comprehension scales; and the WRMT (Rev. ed. and 3rd ed.; Woodcock, 1987, 2011, respectively), including the Word Identification and Word Attack subtests.

Consistent with our purpose, we categorized each measure according to the level of language it taps. The word-level measures included Word Classes—Receptive, PPVT, and Word Identification and Word Attack subtests. The sentence-level measures included Concepts & Following Directions, Word Structure, Recalling Sentences, and Formulated Sentences subtests. The discourse-level measures included Understanding Spoken Paragraphs, GORT Fluency, and GORT Comprehension. We also categorized measures as oral (CELF and PPVT) or written (GORT and WRMT).

Language sample coding and analysis. As stated above, all language samples included in the present study were storybook narratives collected using four of the wordless *Frog* series (*Frog Where Are You*; *One Frog Too Many*; *A Boy, a Dog, and a Frog*; and *Frog Goes to Dinner*; Mayer, 1969). All language samples had been transcribed into SALT format from audio recordings at the time of the original assessment. After these files were de-identified and assigned an identification code by the first research assistant, a second trained research assistant reviewed each file and made adjustments in accordance with a transcription protocol created for the present study. The major adjustments included (a) assuring utterance segmentation as C-units; (b) counting the number of clauses, main and subordinate, in each utterance and inserting the corresponding subordination index code; and (c) standardizing error codes according to the study protocol.

Transcripts were then analyzed using SALT software (Miller & Iglesias, 2012). For each language sample, a total of eight measures were extracted:

1. MLU in words (MLU): a measure of sentence length.
2. Total number of words (TNW): the total words in a sample, providing a measure of overall productivity of the narrative.
3. Subordination Index (SI): the number of total clauses (main and subordinate) in a sample divided by the number of utterances (C-units), providing an index of average clausal density per utterance. The calculation followed published SALT guidelines (Miller & Iglesias, 2012).
4. Number of different words (NDW): the number of new (nonrepeated) words in a sample, interpreted as a measure of lexical diversity. To account for the well-documented effect of sample length on lexical diversity, NDW was calculated on the first 100 words of the sample (Watkins, Kelly, & Harbers, 1995).
5. Omitted bound morphemes: any omission of an obligatory inflection (e.g., third person singular *–s*; regular past tense *–ed*; plural *–s*, possessive *–s*). Errors (Measures 5 through 8) were coded following definitions and examples in the SALT software transcription instructions.
6. Omitted words: any word omission resulting in an agrammatic utterance (e.g., an omitted auxiliary verb, preposition, or obligatory subject or object).
7. Word-level errors: any word error except omitted bound morphemes (e.g., regular for irregular verb, tense overgeneralization, pronoun case error – *see*/*saw*, *goed*/*go*, *him*/*he*).
8. Utterance-level errors: any error not attributable to a single word (e.g., a word order error) resulting in an agrammatic utterance.

As noted above, 17 children were identified as speakers of African American English, either in their assessment reports or in their language sample transcripts. For these children, transcripts were reviewed to ensure that common dialectal features (as identified in Craig & Washington, 2004) were not counted as errors.

A value for each of the eight measures was entered into the database for use in the partial correlation analyses (described below). For the contingency analyses, values were compared with those of age peers using the SALT reference database (Miller & Iglesias, 2012). The database best suited to our samples was the narrative story retell. This database uses story retells based on four different wordless picture books (including *Frog Where Are You*) for Grades K through 5. The comparison storybook, chosen individually for each participant, was the one that afforded the greatest number of matches within a 6-month age range (plus or minus), using SALT software to truncate the comparison samples to the same length as the participant sample (in words). The *z* scores (i.e., number of *SD*s away from the peer group mean) resulting from these comparisons were then entered into the database.

Like the norm-referenced tests, these measures were also categorized according to the level of language they measure. The four error types (numbered 5 through 8) as well as MLU and SI can be regarded as sentence-level measures. NDW measures language primarily at the word level, whereas TNW may be regarded as a discourse-level measure.

Reliability

Reliability was conducted for the coding of assessment reports and for the language sample transcripts. Twenty-one of the 138 reports (15.2%) were randomly selected and recoded by the first author. Point-by-point agreement was 99.9% for test scores, 93.9% for clinician diagnostic judgments, and 94.5% for the coding of all background variables. Language sample transcripts (100%) were reviewed by the second author, who checked utterance segmentation, error codes, and subordination index codes and made any necessary corrections.

Analyses

The distributions of all variables were examined prior to conducting analyses. With the exception of NDW, measures taken from the language samples (particularly count variables such as number of omitted words) showed evidence of nonnormal distribution. Therefore, MLU, TNW, SI, and all error measures were submitted to square-root transformations to increase normality. After the transformations, the four error measures (listed as 5–8 above) were summed into a total errors variable for the correlation analyses in order to reduce the number of comparisons performed. Distributions of all resulting variables were reexamined to ensure they met the assumption of normality.

Partial correlation analyses, controlling for the effects of age, were conducted to examine the association between norm-referenced tests and language sample measures. Analyses were conducted separately for the younger and older participant groups. Age was controlled in these analyses because language sample measures would be expected to improve with development. To control for possible Type I error resulting from multiple comparisons, the false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995) was applied to results. The FDR offers an alternative to traditional Bonferroni corrections, controlling error while preserving more power. The FDR was controlled at 0.10.

We then explored the tools' ability to identify language disorders using contingency tables. Because of the results of the partial correlation analyses, we considered only the younger participant group in this analysis. We used composite scores from norm-referenced tests, rather than subtests, and created composites from the individual language sample measures, because we reasoned that practicing clinicians rarely identify language disorders based on an individual subtest score or a single language sample measure. Three norm-referenced test scores were available: the Core Language composite from the CELF, the overall standard score from the PPVT, and the Oral Reading Quotient (ORQ) from the GORT.

We then created two composites from the narrative language sample measures. The first was an Error Composite, created by averaging the *z* scores from the four microstructural error measures. The second was termed a *Classic Composite*, as it was created by averaging z scores on three commonly used language sample measures (Hux et al., 1993): MLU, NDW, and TNW. The SI was not included in this composite for two reasons. First, it is rarely used by clinicians (Hux et al., 1993) and thus does not meet the criterion of being a "classic" measure. Second, it measures similar skills as MLU (MLU and SI are correlated at $r = .58$), but MLU demonstrated stronger relations with norm-referenced tests in the first set of analyses.

For each measure, we explored two different cutoff scores (1 and 1.5 *SD*s below the mean) for the identification of language disorders. Thus, for the −1.5 *SD* cutoff, each participant was classified as having a language disorder if his or her score fell at or below 1.5 *SD*s below the age mean (i.e., standard score of 77 or less; *z* score of −1.5 or less after comparison to age peers using SALT). Similarly, standard scores below 85 and *z* scores of −1.0 or less were categorized as demonstrating a language disorder for the −1.0 *SD* cutoff analysis. Scores above these values were categorized as within normal limits. For each combination of norm-referenced test and language sample composite, the number of participants was then tabulated in each of four categories: demonstrating a language disorder on both measures; within normal limits on both measures; demonstrating a language disorder on the norm-referenced test but within normal limits on the language sample measure; and vice versa. We also calculated the number of participants in our sample who would qualify as having a language disorder according to each measure.

Results

The results of the partial correlations between five language sample measures (MLU, TNW, SI, NDW, and Total Errors) and 11 norm-referenced language measures are displayed in Table 2. Within the older group of children, only four correlations reached statistical significance: GORT comprehension with TNW, $r(7) = .99$, $p < .001$; SI with Understanding Spoken Paragraphs, $r(7) = .81$, $p = .008$; NDW with Formulated Sentences, $r(15) = .55$, $p = .021$; NDW with Word Classes—Receptive, $r(15) = .51$, $p < .038$. Only the relation between TNW and GORT Comprehension remained significant after multiple-comparison correction.

For the younger group of children, many more correlations were significant. MLU correlated significantly with seven norm-referenced test scores: Word Structure, $r(33) = .46$, $p = .006$; Recalling Sentences, $r(37) = .36$, $p = .024$; Formulated Sentences, $r(35) = .42$, $p = .009$; Word Classes—Receptive, $r(12) = .55$, $p = .040$; PPVT, $r(24) = .63$, $p = .001$; WRMT Word Identification, $r(27) = .38$, $p = .041$; and WRMT Word Attack, $r(17) = .50$, $p = .028$. Significance for the correlations with Word Structure, Formulated Sentences, and PPVT remained after multiple-comparison correction. TNW did not correlate significantly with any of the 11 test scores in this group. The SI correlated significantly with three norm-referenced test scores: Word Structure, $r(33) = .35$, $p = .040$; Recalling Sentences, $r(37) = .42$, $p = .008$; and GORT Fluency, $r(16) = .55$, $p = .018$. The correlations with Recalling Sentences and with GORT Fluency remained significant after multiple-comparison correction. NDW correlated significantly with three norm-referenced test scores: Word Structure, $r(31) = .48$, $p = .004$; Recalling Sentences, $r(36) = .57$, $p < .001$; Formulated Sentences, $r(34) = .41$, $p = .014$. The significance of all three correlations remained after multiple-comparison correction. Finally, five norm-referenced tests were significantly correlated with Total Errors in the narratives: Word Structure, $r(33) = -.55$, $p = .001$; Recalling Sentences, $r(37) = -.67$, $p < .001$; Formulated Sentences, $r(35) = -.55$, $p < .001$; PPVT, $r(24) = -.45$, $p = .021$; WRMT Word Identification, $r(27) = -.39$, $p = .039$. The significance

Table 2. Partial correlations (age removed) between norm-referenced tests and language sample measures.

	MLU-W		TNW		SI		NDW		Total errors	
Variable	Younger	Older	Younger	Older	Younger	Older	Younger	Older	Younger	Older
CFD	.28	−.31	.03	.16	.14	.13	.24	−.32	−.27	.33
WS	.46**	—	.12	—	.35*	—	.48**	—	−.55**	—
RS	.36*	−.16	.25	−.01	.42**	.34	.57**	.05	−.67**	.08
FS	.42**	−.19	.15	−.24	.29	.39	.41**	.55*	−.55**	−.21
WC	.55*	−.02	.19	−.16	.28	.05	.08	.51*	−.40	-.18
USP	.12	.22	.34	.21	.49	.81*	.34	—	.13	−.04
PPVT	.63**	.41	−.21	.50	.03	.18	.20	.30	−.45**	.31
GORT FL	.41	−.42	.34	−.37	.55**	.01	.18	−.27	−.29	−.39
GORT Comp	.43	.28	.20	.99**	.45	−.10	.35	—	−.34	.64
WRMT WI	.38*	−.07	.09	.51	.27	−.16	.22	.15	−.39*	.15
WRMT WA	.50*	.19	.37	.62	.35	−.03	−.03	—	−.19	.27

Note. Partial correlations controlling for age are displayed for the younger and older groups. *N* varies by cell (see Table 1). Correlations between NDW and USP, GORT Comp, and WRMT WA are not reported (see dashes) because the *N* for these cells fell below 10. MLU-W = MLU in words; TNW = total number of words; SI = Subordination Index; NDW = number of different words; CFD = Concepts and Following Directions; WS = Word Structure; RS = Recalling Sentences; FS = Formulated Sentences; WC = Word Classes—Receptive; USP = Understanding Spoken Paragraphs; PPVT = Peabody Picture Vocabulary Test; GORT FL = Gray Oral Reading Test, Fluency index; GORT Comp = Gray Oral Reading Test, Comprehension index; WRMT WI = Woodcock Reading Mastery Tests, Word Identification subtest; WRMT WA = Woodcock Reading Mastery Tests, Word Attack subtest.

*$p < .05$ without applying false discovery rate (FDR) procedure. **Correlation is significant after controlling FDR at 0.10 (Benjamini & Hochberg, 1995).

of the correlations with Word Structure, Recalling Sentences, Formulated Sentences, and PPVT remained after multiple-comparison corrections.

The second set of analyses classified children within the younger group as demonstrating normal language or a language disorder using two different cutoffs. Results are displayed in Table 3 (for the −1.5 *SD* cutoff) and Table 4 (for the −1.0 *SD* cutoff). The tables illustrate the distribution of children across the four possible categories of each comparison between a language sample composite and a norm-referenced test composite: Both measures agree the child has language within normal limits (WNL); both measures agree the child has a language disorder (LD); one measure categorizes the child as having a language disorder, whereas the other does not, and vice versa. There were 12 combinations examined (by comparing three norm-referenced test composites with two narrative language sample composites at two different cutoff scores). In 11 of 12 comparisons, children are distributed across the four possible categories of agreement (the sole exception is the comparison between the Classic Composite and the GORT at the −1.0 *SD* cutoff).

The rate of agreement between the tools can be calculated for each comparison as the number of cases receiving the same classification (both WNL or both LD) divided by the total number of cases. For the −1.5 *SD*

Table 3. Comparison of case classifications using language sample composites versus norm-referenced test scores using a –1.5 *SD* cutoff.

		Norm-referenced tests								
		CELF Core			PPVT			GORT ORQ		
Language sample measure		WNL	LD	Total	WNL	LD	Total	WNL	LD	Total
Error Composite	WNL	**12**	11	23	**17**	3	20	**16**	3	19
Error Composite	LD	4	**8**	12	5	**2**	7	3	**2**	5
Total		16	19	35	22	5	27	19	5	24
Classic Composite	WNL	**8**	12	20	**12**	2	14	**13**	1	14
Classic Composite	LLD	8	**7**	15	10	**3**	13	6	**4**	10
Total		16	19	35	22	5	27	19	5	24

Note. Table displays the number of cases within the younger group who were classified as demonstrating language within normal limits (WNL) and language disorder (LD), based on a score cutoff of 1.5 *SD*s below the mean. Bold text indicates cases in which language sample composites and norm-referenced tests result in the same classification (i.e., a case is classified as WNL by both tools or as LD by both tools). CELF = Clinical Evaluation of Language Fundamentals; PPVT = Peabody Picture Vocabulary Test; GORT = Gray Oral Reading Test; ORQ = Oral Reading Quotient.

Table 4. Comparison of case classifications using language sample composites versus norm-referenced test scores using a –1.0 *SD* cutoff.

Language sample measure		Norm-referenced tests								
		CELF Core			PPVT			GORT ORQ		
		WNL	LD	Total	WNL	LD	Total	WNL	LD	Total
Error Composite	WNL	**7**	13	20	**9**	8	17	**12**	6	18
Error Composite	LD	2	**13**	15	7	**3**	10	2	**4**	6
Total		9	26	35	16	11	27	14	10	24
Classic Composite	WNL	**6**	5	11	**4**	5	9	**8**	0	8
Classic Composite	LLD	3	**21**	24	12	**6**	18	6	**10**	16
Total		9	26	35	16	11	27	14	10	24

Note. Table displays the number of cases within the younger group who were classified as demonstrating language within normal limits (WNL) and language disorder (LD), based on a score cutoff of 1.0 *SD*s below the mean. Bold text indicates cases in which language sample composites and norm-referenced tests result in the same classification (i.e., a case is classified as WNL by both tools or as LD by both tools). CELF = Clinical Evaluation of Language Fundamentals; PPVT = Peabody Picture Vocabulary Test; GORT = Gray Oral Reading Test; ORQ = Oral Reading Quotient.

cutoff (see Table 3), the highest rates of agreement were exhibited by the Error Composite with the PPVT (19/27, or 70.3% agreement), the Error Composite with the GORT (18/24, or 75.0% agreement), and the Classic Composite with the GORT (17/24, or 70.8% agreement). Examination of Table 3 demonstrates that this agreement stems from relatively few children being classified as having a language disorder on these measures. The lowest rate of agreement was found between the CELF and the Classic Composite (15/35, or 42.8%), although greater numbers of children were identified as having a language disorder by each measure. The remaining comparisons showed intermediate rates of agreement (CELF with Error Composite at 57.1% agreement; PPVT with Classic Composite at 55.6%).

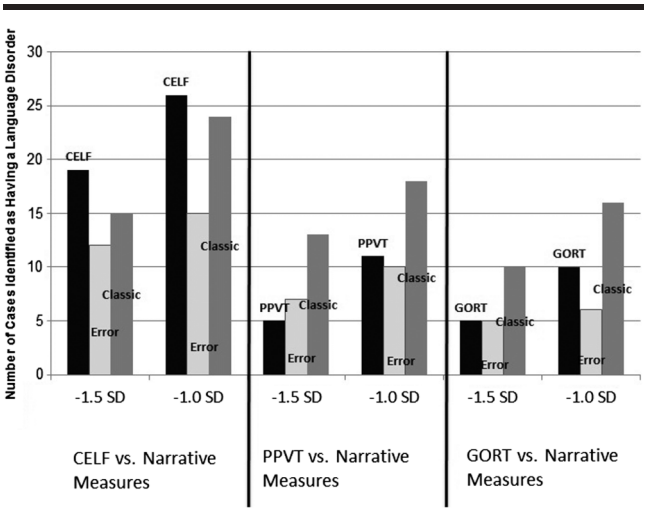
In contrast, more children are identified as having a language disorder using the −1.0 *SD* cutoff (see Table 4). For this cutoff point, the highest rates of agreement were exhibited by the Classic Composite with the CELF Core (27/35, or 77.1% agreement) and with the GORT (18/24, or 75.0% agreement). Rates of agreement between the Error Composite and the norm-referenced tests were either identical (for CELF Core, at 57.1%) or lower (for PPVT, 44.4%, and GORT, 66.7%) when using the −1.0 *SD* cutoff in comparison to the −1.5 *SD* cutoff.

Finally, Figure 1 displays the number of cases identified as having a language disorder by each measure, according to the data in Table 3 and Table 4. Of the children who completed both the CELF Core and narrative language sample, more children were identified as having a language disorder by the CELF than by either of the composite language sample measures (Classic or Error) regardless of the cutoff point used. The PPVT identified fewer children than the CELF, and also tended to identify fewer children than the composite language sample measures (with the exception of the Error Composite at the −1.0 *SD* cutoff). Finally, in comparison to the GORT Oral Reading Quotient (ORQ) the Classic Composite identified more children as having a language disorder, whereas the Error Composite identified the same number (at the −1.5 *SD* cutoff) or fewer (at the −1.0 *SD* cutoff).

Discussion

In this study, we analyzed clinical language assessment data from a clinically referred, culturally and linguistically diverse sample of school-age children. Results discussed here should be considered exploratory due to the retrospective study design and absence of tight controls on participant

Figure 1. Comparison of identification rates by assessment type and cutoff. The figure displays the number of children who received a score at or below the cutoff point for each assessment, within the subset of children who completed both the norm-references test and the language sample for that comparison. CELF = Clinical Evaluation of Language Fundamentals; PPVT = Peabody Picture Vocabulary Test; GORT = Gray Oral Reading Test.



eligibility or assessment procedures; however, they may reflect the realities of clinical practice, at least in an urban outpatient setting. We consider our results below in terms of the four main hypotheses posed in the introductory section.

The Role of Age

We hypothesized that younger children would demonstrate stronger associations between storybook narrative language sample measures and norm-referenced tests. This prediction was borne out by the correlation analyses, in which there were 12 correlations that reached significance (after correction for multiple correlations) for the younger children age 6;0–8;11, but only one correlation (corrected) for the older children age 9;1–12;8. Although there were fewer children in the older group than in the younger group, it is unlikely that the pattern of differences across age groups is driven entirely by the sample size. Not only was there a striking group difference in the number of significant correlations, but in several instances, correlations that were strong and significant for younger children were nonexistent for older children (e.g., Recalling Sentences and Total Errors at $r = -.67$ vs. $r = .08$; Recalling Sentences and NDW at $r = .57$ vs. $r = .05$). In the one significant correlation for the older children, the opposite pattern was observed. An almost perfect correlation between GORT Comprehension and TNW ($r = .99$) was low ($r = .20$) for the younger children.

Our pattern of results suggests that the norm-referenced tests measure several of the same language skills children use when they are telling a story, but only within a selected (younger) age range. For the older children, norm-referenced tests may not be measuring the same skills as the oral narrative language sample measures used in this study. Because the norm-referenced tests are designed to capture the effects of age, it is likely that the change in relationship between the two tools as children get older stems from the narrative language samples. We suspect that use of the *Frog* series wordless picture books may have artificially constrained the narrative abilities of the older children. The *Frog* stories convey a simple series of actions and were not originally designed for the purpose of eliciting children's narratives. A different, less constrained, narrative task, such as summarizing a narrative video with a more complicated plot, may have elicited a wider range of narrative abilities in this study. In particular, Nippold (2014) has argued that complex thoughts result in complex language. She and her colleagues have shown that asking eighth-grade students to retell fables generates language with sentences that are almost twice the length of those in a conversational sample (Nippold et al., 2014). Thus, clinicians may need to consider other narrative tasks, other modalities, or other genres of language sampling to accurately capture the abilities of children beyond the early elementary years. For older children, expository language samples are significantly correlated with norm-referenced test scores (Nippold et al., 2009), and written language samples are more sensitive to language disorders (Scott & Windsor, 2000).

Associations by Levels of Language

Our next question was whether measures that index the same level of language would be better related than those that index different levels of language. We focus this discussion on the younger group and on those correlations that were strong enough to maintain significance after the FDR procedure. Within the younger group, correlations between sentence-level measures were particularly prominent; they accounted for six of 12 correlations (Word Structure with MLU, Formulated Sentences with MLU, Recalling Sentences with SI, Word Structure with Total Errors, Recalling Sentences with Total Errors, and Formulated Sentences with Total Errors). In contrast, there were no correlations between word-level norm-referenced and word-level narrative measures. These findings point to the integrity and importance of the sentence across widely different language tasks in the assessment of school-age children. The number and strength of the correlations between the total errors variable, our composite index of grammatical errors in the language sample, and the various sentence-level CELF subtests was particularly notable. Grammatical errors have not been considered in past work on the relations between narrative language sample measures and norm-referenced tests (Bishop & Donlan, 2005; Manolitsi & Botting, 2011; Norbury & Bishop, 2003), though grammar plays a prominent role in the most common language disorders affecting school-age children (e.g., Eigsti, Bennetto, & Dadlani, 2007; Rice & Wexler, 1996). Our results suggest grammatical errors in language samples may be closely related to norm-referenced test results and that it may be worthwhile for clinicians to calculate grammatical error measures from language samples.

Five correlations crossed word and sentence levels (e.g., PPVT with MLU; Recalling Sentences with NDW). Correlations between lexical diversity in the narrative samples, indexed by NDW, and three different sentence-level subtests (Word Structure, Recalling Sentences, and Formulated Sentences) were particularly notable and point to an association between lexical and syntactic skills that transcends task. Strong word-level vocabulary skills may support even decontextualized sentence tasks, including repeating sentences and making up sentences.

Finally, although we considered discourse-level measures (including scores from the CELF's Understanding Spoken Paragraphs subtest and the GORT's Fluency and Comprehension scales, as well as the TNW measure from the narratives), these measures generated only one significant correlation (GORT Fluency and SI). It is possible that task factors play a larger role in children's scores on these measures than on the sentence- and word-level measures. For example, story memory and higher level skills such as inferencing and prediction may influence scores on the Understanding Spoken Paragraphs task but have a weak relationship with grammar and vocabulary skills. In addition, TNW, the traditional measure of overall productivity, may be influenced by the format of the narrative

task. The use of the wordless picture book may support the generation of relatively uniform narratives that can be compared (e.g., using SALT databases), but it may also constrain children's tendencies to tell longer or shorter stories. Finally, it is also possible that the variations in the number of children who completed each test—an inherent feature of our database—influenced our ability to detect significant relationships.

Written Language Measures

Due to the lack of past literature comparing written language test scores with oral language sample measures, our work in this area was exploratory. We hypothesized that significant associations would exist between written language tests and narratives but that these associations might be fewer or weaker than those between oral language tests and narratives due to the difference in modality. Of 40 possible correlations across both age groups, there were only two significant correlations (after correction), with one in each age group: GORT comprehension with TNW for the older group and GORT Fluency with SI for the younger group. Of interest, both correlations involved discourse-level reading skills (comprehension and discourse-level fluency) rather than single-word reading skills (reading real or pseudowords accurately). This finding is reasonable in light of the fact that storytelling is also a discourse-level task. The near-perfect correlation between reading comprehension and telling a robust (long) story in the older group underscores the important role of language level in this relationship.

It is possible that associations between written language tests and oral language samples would be stronger in older children, who are better able to map written words to oral ones; however, as discussed above, the storybook narrative language samples used in this study did not appear to index language skills well in these older children. Furthermore, we found the strongest relationships between sentence-level measures, but were not able to include a sentence-level test of written language. Thus, this study may be simply a starting point for future exploration of written language test scores and oral language sample measures.

Classification Rates

At the most general level, our hypothesis regarding classification rates was confirmed: There was overlap but imperfect agreement between the children identified as having a language disorder by our two types of tools. In all comparisons, there were children identified by the language sample measure and not by the norm-referenced test, and vice versa. Furthermore, it was not true that one type of tool consistently identified more children than the other (cf. Condouris et al., 2003). For clinicians seeking to accurately identify children with language disorders, our results point to the importance of including both types of tools in assessment; exclusive use of only norm-referenced tests, for example, might lead to underidentification of children who perform well on this type of tool but not as well on the less structured, more ecologically valid language sample task. This conclusion is, of course, consistent with previous literature recommending the use of language sampling in assessment (e.g., Costanza-Smith, 2010; Hewitt et al., 2005).

As would be expected, the use of a −1.0 *SD* cutpoint identifies more children than the −1.5 *SD* cutpoint across all measures. Of greater interest is the difference in the relationship between our two types of tools at the two cutpoints. In particular, the agreement between the CELF Core and the Classic Composite increased dramatically (from 43% to 77%) when the cutpoint was made less stringent. It is clear that a number of the children in our sample scored between −1.0 and −1.5 *SD*s from the mean on both the CELF Core and the Classic Composite; more specifically, there appear to be a number of children who scored below −1.5 *SD*s on one of these measures and between −1.0 and −1.5 on the other, resulting in relatively poor agreement at −1.5 *SD*s and much stronger agreement at −1.0 *SD*s. Thus, although −1.5 *SD*s (or even greater) is a common cutoff score for identifying language disorders in clinical practice (Betz et al., 2013), this stringent cutoff may result in different classification on commonly used clinical assessment tools.

In contrast, agreement with the Error Composite remained the same or dropped when we used the less stringent −1.0 *SD* cutpoint. Examination of Figure 1 suggests that few children fell between −1.0 and −1.5 *SD*s from the mean on the Error Composite, a pattern that likely reflects the skewed nature of the distribution of Error Composite scores. In other words, children either made few grammatical errors in their language samples or made frequent grammatical errors in their language samples. The skewed nature of tense and agreement error distributions in children with SLI has been emphasized by researchers in the past (see Rice, 2000).

Overall, the rates of agreement across tools and cutpoints were moderate at best, ranging from 37% to 77% agreement. We note two factors that may have played a role in these patterns, aside from differences in what the two types of language assessment tools measure. First, we used composite measures for identification of language disorder, as children are rarely identified on the basis of a single subtest or language sample measure. However, it is possible that some children showed particular weakness in some areas of language but not others (e.g., weakness in vocabulary but strength in grammar; see Colozzo, Gillam, Wood, Schnell, & Johnston, 2011, for evidence of this phenomenon). The use of composites might thus obscure language weaknesses. In addition, we note that the nature of the comparison sample differs between the two types of tools that we used. Norm-referenced tests have true psychometric norms, whereas the reference databases that accompany the SALT software have not undergone the same process of psychometric development (see also Condouris et al., 2003). It is possible that some of the differences in

classification that we observed relate to these differences in the source of the comparison.

Conclusions and Future Directions

We explored a number of aspects of the relationship between norm-referenced tests and storybook narrative language sample measures in a group of school-age children assessed for language disorder. Overall, our results show moderate levels of association and classification agreement between the two tools, within a restricted age range. Thus, we echo previous work suggesting that clinicians consider including both tools in their language assessments, though our results also support the need to carefully consider age in relation to the type of language sample included in the assessment. Another important message from these results is the importance of including sentence-level measures—such as grammatical errors from a language sample—in school-age language assessment. Finally, our results show that classification agreement may depend on the specific measures and cutpoints used; in the case of commonly used microstructural measures (MLU, TNW, and NDW) versus a commonly used oral language test (the CELF), use of a less stringent cutpoint dramatically improved agreement.

This exploratory study suggests a number of directions for future work. It will be important to explore the same relationships considered here using groups of older children (e.g., those at least 9 years old), using different types of language samples, to determine whether stronger relations may exist under these circumstances. Stronger relations between language sample measures and written language tests may also emerge under these circumstances. Continued exploration of the relationships between language sample measures and norm-referenced tests is important to optimizing language assessments for school-age children.

Acknowledgments

We thank Elizabeth Mikolajczyk, Jamie Schmidt, and L. Katherine Walters for their assistance with this project.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57,* 289–300.

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., & Donlan, C. (2005). The role of syntax in encoding and recall of pictorial narratives: Evidence from specific language impairment. *British Journal of Developmental Psychology, 23,* 25–46.

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy, 18,* 1–21.

Caesar, L. G., & Kohler, P. D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly, 30,* 226–236.

Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54,* 1609–1627.

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology, 12,* 349–358.

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning and Education, 17,* 9–15.

Craig, H., & Washington, J. A. (2004). Grade-related changes in the production of African American English. *Journal of Speech, Language, and Hearing Research, 47,* 450–463.

De Lamo White, C., & Jin, L. (2011). Evaluation of speech and language assessment approaches with bilingual children. *International Journal of Language & Communication Disorders, 46,* 613–627.

Dunn, L. M., & Dunn, D. M. (1997). *Peabody Picture Vocabulary Test—Third edition*. Minneapolis, MN: Pearson Assessments.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test—4th edition*. Minneapolis, MN: Pearson Assessments.

Eigsti, I. M., Bennetto, L., & Dadlani, M. B. (2007). Beyond pragmatics: Morphosyntactic development in autism. *Journal of Autism and Developmental Disorders, 37,* 1007–1023.

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. 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.

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24,* 84–91.

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech, Language, and Hearing Research, 38,* 415–425.

Manolitsi, M., & Botting, N. (2011). Language abilities in children with autism and language impairment: Using narrative as an additional source of clinical information. *Child Language Teaching and Therapy, 27,* 39–55.

Mayer, M. (1969). *Frog, where are you?* New York, NY: Dial Press.

Miller, J., & Iglesias, A. (2012). *SALT: Systematic Analysis of Language Transcripts* [Research version]. Middleton, WI: SALT Software.

Nippold, M. A. (2014). Language intervention at the middle school: Complex talk reflects complex thought. *Language, Speech, and Hearing Services in Schools, 45,* 153–156.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research, 57,* 876–886.

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2008). Expository discourse in adolescents with language impairments: Examining syntactic development. *American Journal of Speech-Language Pathology, 17,* 356–366.

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2009). Syntactic development in adolescents with a history of language impairments: A follow-up investigation. *American Journal of Speech-Language Pathology, 18,* 241–251.

Norbury, C. F., & Bishop, D. V. (2003). Narrative skills of children with communication impairments. *International Journal of Language & Communication Disorders, 38,* 287–313.

Paul, R., & Norbury, C. F. (2012). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing, and communicating* (4th ed.). St. Louis, MO: Elsevier.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1398–1411.

Rice, M. (2000). Grammatical symptoms of specific language impairment. In D.V. M. Bishop & L. Leonard (Eds.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome* (pp. 17–34). Hove, United Kingdom: Psychology Press.

Rice, M., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39,* 1239–1257.

Scott, C. M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43,* 324–339.

Semel, E., Wiig, E. H., & Secord, E. H. (1995). *Clinical Evaluation of Language Fundamentals—Third Edition*. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals—Fourth Edition*. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. A. (2004). *Clinical Evaluation of Language Fundamentals—Preschool Edition–2*. San Antonio, TX: The Psychological Corporation.

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.

Spaulding, T. J., Szulga, M. S., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between U.S. policy makers and test developers. *Language, Speech, and Hearing Services in Schools, 43,* 176–190.

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

Tomblin, J. B., Records, N., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39,* 1284–1294.

Watkins, R.V., Kelly, D. J., & Harbers, H. M. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38,* 1349–1355.

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Tests—4th edition*. Austin, TX: Pro-Ed.

Wiederholt, J. L., & Bryant, B. R. (2012). *Gray Oral Reading Tests—5th edition*. Austin, TX: Pro-Ed.

Wilson, K. S., Blackmon, R. C., Hall, R. E., & Elcholtz, G. E. (1991). Methods of language assessment: A survey of California public school clinicians. *Language, Speech, and Hearing Services in Schools, 22,* 236–241.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests, Revised*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (2011). *Woodcock Reading Mastery Tests, Third Edition*. San Antonio, TX: Pearson.