ASHA American Speech-Language-Hearing Association

JSLHR

Research Note

Predicting Language Performance From Narrative Language Samples

Kimberly A. Murphy,[a] Alisha P. Springle,[b] Mollee J. Sultani,[c] Autumn McIlraith,[d] and Language and Reading Research Consortium (LARRC)

[a] Department of Communication Disorders and Special Education, Old Dominion University, Norfolk, VA [b] Department of Rehabilitation Sciences, Indiana University South Bend [c] Department of Speech-Language-Hearing: Sciences & Disorders, College of Liberal Arts & Sciences, The University of Kansas, Lawrence [d] Independent Researcher, Houston, TX

ARTICLE INFO

Article History:
Received May 10, 2021
Revision received September 11, 2021
Accepted September 22, 2021

Editor-in-Chief: Stephen M. Camarata
Editor: Krystal L. Werfel

https://doi.org/10.1044/2021_JSLHR-21-00262

ABSTRACT

Purpose: Analysis of narrative language samples is a recommended clinical practice in the assessment of children's language skills, but we know little about how results from such analyses relate to overall oral language ability across the early school years. We examined the relations between language sample metrics from a short narrative retell, collected in kindergarten, and an oral language factor in grades kindergarten through 3. Our specific questions were to determine the extent to which metrics from narrative language sample analysis are concurrently related to language in kindergarten and predict language through Grade 3.
Method: Participants were a sample of 284 children who were administered a narrative retell task in kindergarten and a battery of vocabulary and grammar measures in kindergarten through Grade 3. Language samples were analyzed for number of different words, mean length of utterance, and a relatively new metric, percent grammatical utterances (PGUs). Structural equation models were used to estimate the concurrent and longitudinal relationships.
Results: The narrative language sample metrics were consistently correlated with the individual vocabulary and grammar measures as well as the language factor in each grade, and also consistently and uniquely predicted the language factor in each grade. Standardized path estimates in the structural equation models ranged from 0.20 to 0.39.
Conclusions: This study found narrative language sample metrics to be predictive, concurrently and longitudinally, of a latent factor of language from kindergarten through Grade 3. These results further validate the importance of collecting and analyzing narrative language samples, to include PGU along with more traditional metrics, and point to directions for future research.
Supplemental Material: https://doi.org/10.23641/asha.17700980

Language sample analysis (LSA) is considered a gold standard for assessing children's language production in everyday communicative contexts (Costanza-Smith, 2010). LSA measures demonstrate strong associations with norm-referenced tests of oral language, particularly for younger children (Ebert & Scott, 2014). Another major advantage of language samples is that they can be analyzed in multiple, descriptive ways that can lead directly to functional treatment goals. However, a barrier to speech-language pathologists' (SLP) use of LSA is the time and expertise needed to collect, transcribe, and analyze recorded samples of children's speech (Pavelko et al., 2016). To add to the clinical challenge, research has not yet made clear how LSA measures predict children's performance over time, making it difficult for SLPs to interpret LSA results. Individual variability in development has made prediction of language outcomes difficult, even when using early measures with good sensitivity and specificity (Armstrong et al., 2018). Such prediction challenges occur with standardized tests for vocabulary, grammar, and comprehensive language measurement (Snowling

Correspondence to Kimberly A. Murphy: kamurphy@odu.edu. ***Disclosure:*** *Kimberly A. Murphy received doctoral student funding from LARRC. The other authors have declared that no other competing financial or nonfinancial interests existed at the time of publication.*

et al., 2016) and result in only moderate predictive value across analytic method (Armstrong et al., 2018). The purpose of this study is to investigate how language measures from a short narrative sample relate to the same children's language performance on formal language measures longitudinally.

LSA represents valid assessment of real-world communication that may be elicited through conversational, narrative, or expository sampling procedures. Narrative production and retell samples have several clinical advantages. Children from the ages of 4–21 years produce representative and ecologically valid narrative samples (Channell et al., 2018). Narrative samples can be appropriately used across cultures and languages (Heilmann & Westerveld, 2013) and result in multiple reliable language measurements (Tilstra & McMaster, 2007). Compared with conversation, narrative retell may also result in more complex language use (Nippold et al., 2014), thus better reflecting the literate language required in school settings. Retelling a story from a model is likely to be easier than generating a new story, which makes story retell a useful context for younger children with less storytelling experience (Spencer & Petersen, 2020).

There is no agreed-upon guideline for the length of language samples, with adequate reliability noted in samples as short as 3 min in length (Heilmann et al., 2010) to as long as 20 min in length (Gavin & Giles, 1996). Logically, the longer the elicited sample, the more time clinicians will need to invest to transcribe, analyze, and interpret the results. Fortunately, there is some evidence that analysis of shorter samples may be clinically appropriate. Studies such as those of Heilmann et al. (2010) suggest that the LSA measures of mean length of utterance (MLU), total number of words (TNWs), and number of different words (NDWs) can be reliable with short samples. Eisenberg and Guo (2015) found that percent grammatical utterance (PGU) can be measured reliably in samples averaging approximately 30 utterances. Using robust sampling methods, Pavelko et al. (2020) reported no significant difference between 25- and 50-utterance conversational samples in TNW, MLU_{SUGAR} (an MLU calculation specific to Pavelko and colleagues' coding procedures), words per sentence, and clauses per sentence (CPS). However, story retelling can be a difficult task for young children (Westerveld et al., 2004), resulting in short samples of fewer than 20 utterances, on average, while existing data from narrative LSA measures are based on sample sizes of 30, 50, or even 100 utterances achieved by combining multiple retells.

LSA Measurements

LSA measures capture snapshots of different aspects of children's language. Currently, quantitative LSA measures are typically thought of as measuring either vocabulary or grammar.

Vocabulary

The simplest measure of vocabulary derived from language samples is TNW, which is the child's word count for the sample. Although TNW can vary in samples of different lengths and contexts, there is emergent evidence that relatively short samples in both narrative and conversational contexts generate consistent TNW values (Heilmann et al., 2010; Tilstra & McMaster, 2007). Another measure, NDW, measures breadth of vocabulary by counting unique words used within a sample (Malvern & Richards, 2012). NDW also demonstrated stability across sample lengths and contexts (Heilmann et al., 2010); however, compared with TNW, NDW had a higher test–retest reliability (Gavin & Giles, 1996) and higher correlations with standardized tests of vocabulary (Ukrainetz & Blomquist, 2002).

Grammar

Children's grammatical ability is stable over time and useful for differentiating those with and without language impairment (Bedore & Leonard, 1998). MLU is a measure of the average number of morphemes (MLUm) or words (MLUw) in each utterance in a sample. MLUm and MLUw increase over time and differentiate children with and without language impairment (Rice et al., 2010). MLU broadly captures children's ability to combine words and morphemes to increase utterance length. PGU is an overall language measure that, unlike MLU, specifically captures grammatical accuracy (Eisenberg & Guo, 2013). PGU is calculated by excluding utterances that do not require a subject and verb, marking utterances that contain at least one grammatical error, and then dividing the number of grammatical (or unmarked) utterances by the total number of included utterances. For example, if a sample contained 15 utterances, five of which were excluded and three of which contained grammatical errors, the PGU would be 70% (seven grammatical utterances divided by 10 included utterances = 0.70 or 70%). PGU can be reliably applied to narrative samples of children between 4 and 9 years of age (Guo et al., 2019). It adequately differentiates children with and without language impairment (Guo et al., 2019; Guo & Eisenberg, 2014; Guo & Schneider, 2016) and demonstrates reliability over time (Guo et al., 2019; Leonard et al., 2017).

Relationships Between LSA Metrics and Standardized Language Assessments

Limited research has examined concurrent relationships between LSA measures and standardized assessment data. While studies have identified predictable age-related changes in LSA metrics between children of different ages (Guo et al., 2019; Pavelko & Owens, 2017; Tilstra & McMaster, 2007), others have found that the correlations

with standardized test results may vary with the age of the child (Ebert & Scott, 2014). Ukrainetz and Blomquist (2002) compared several standardized vocabulary assessments to combined conversation and narrative LSA measures for children aged 3;11–6;0 (years;months). They reported weak-to-moderate positive correlations between LSA measures and NDW and weak correlations between LSA measures and both TNW and MLU. Tilstra and McMaster (2007) noted a moderate correlation between TNW and number of grammatical errors (similar to PGU) from narrative samples and standardized language composite scores for children aged 5–9 years, with more significant relationships for verbal fluency measures, for example, TNW per minute. This study did not compute MLU. Ebert and Scott (2014) found significant correlations between MLU, NDW, and multiple norm-referenced test scores with measures of both vocabulary and grammar for younger children ages 6;0–9;0. They also found significant correlations between TNW and scores on a norm-referenced test of reading comprehension and NDW with measures of both vocabulary and grammar for children ages 9;1–12;8; however, MLU was not significantly correlated with any LSA measure for older children. Owens and Pavelko (2017) compared conversational measures of MLU, TNW, CPS, and words per sentence from children ages 3;7–7;6 to their performance on norm-referenced subtests of vocabulary, grammar, and pragmatic language. They found MLU, TNW, and words per sentence to be significantly correlated with all norm-referenced subtest results. Guo et al. (2019) have also demonstrated correlations of PGU and percent grammatical responses with a standardized test of grammar.

Although it is clear that language sample metrics are associated with standardized test results, this study expands on prior research in important ways. Existing studies of LSA metrics have used cross-sectional samples of varying sizes (Ebert & Scott, 2014; Guo et al., 2019; Guo & Schneider, 2016; Pavelko & Owens, 2017, 2019; Tilstra & McMaster, 2007). Our data are longitudinal, allowing a preliminary investigation of the predictive nature of LSA metrics over time. Like Guo et al. (2019) and Pavelko and Owens (2017, 2019), we have a large sample of children. In combination with our wide variety of language measures, this large, longitudinal sample enables us to use structural equation modeling and latent factors for language.

Research Questions

The research questions that we asked in this study are as follows:

1. What is the relationship between LSA metrics, from a brief story retelling, and an oral language factor construct, both measured in kindergarten?
2. What is the longitudinal relationship between the kindergarten LSA metrics and an oral language factor measured in Grades 1, 2, and 3?

Method

Participants

Participants were selected from a sample that was originally recruited in preschool as part of a larger longitudinal study by the Language and Reading Research Consortium (LARRC; for more details on participants and methods, see LARRC et al., 2016). Recruitment took place at four research sites in different regions of the United States. Children were recruited from the classrooms of participating teachers, and there were no requirements for presence or absence of language difficulties; thus, it can be assumed that most children were typically developing in language. The number of kindergarten participants in the second year of the larger study was 379. Of those, 356 had LSA data and 284 had a sufficient number of utterances ($\geq$ 5) in their language sample, which is recommended in PGU coding (detailed below). Thus, the sample size for this study was 284. The average age of the children at the time of testing in kindergarten was 6;1. Nearly 10% of the children were reported by caregivers to have an Individualized Education Program. Nonverbal intelligence, measured in preschool using the matrices subtest of the Kaufman Brief Intelligence Test, was within the average range (standard score $M = 102$, $SD = 11.71$).

Demographic information obtained from caregiver questionnaires at the beginning of the larger study showed that English was the primary language of the home for 98% of families, caregiver race was 92% White, the majority (88.6%) of female caregivers had at least some college education, and 15% of children qualified for free and reduced-price lunch. For household income, 34.7% of the sample reported $\leq$ \$60,000; 25.6% between \$60,001 and \$85,000; and 39.8% were > \$85,000.

Procedures

In the LARRC study, a battery of measures was administered by trained research staff in the spring of each year in a quiet room in the school, local research site, community center, or home. For this study, we used the vocabulary and grammar measures administered in kindergarten to Grade 3 and a narrative language measure administered in kindergarten.

Measures

Vocabulary and Grammar

Three measures of vocabulary and four measures of grammar were administered in all grades, with an additional

three measures of grammar in kindergarten. Raw scores were used in analyses except where otherwise specified. Reliabilities (coefficient alpha) reported by LARRC in the first year of the study ranged from 0.91 to 0.96 for the vocabulary measures and from 0.78 to 0.92 for the grammar measures, except for Morphological Lexical Judgment (MLJ), which was 0.68.

Vocabulary. Receptive vocabulary was measured using Form A of the Peabody Picture Vocabulary Test–Fourth Edition (PPVT; Dunn & Dunn, 2007), which requires students to select a picture from a choice of four that represents the meaning of a word spoken by the examiner. Expressive vocabulary was measured using the Expressive Vocabulary Test–Second Edition (EVT; Williams, 2007), which requires students to name a picture or provide a synonym. The Word Classes (WC) subtest of the Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4; Semel et al., 2003) assessed students' ability to recognize relationships between words and the ability to explain those relations. WC 1 was given to all students up to Grade 2, and WC 2 was given to all students in Grade 3, regardless of age. This is a modification from standardized procedures that specify WC 2 is administered for age 9 years and up. We used a sum of the WC receptive and expressive raw scores for analyses.

Grammar. Three measures were administered in kindergarten only—two probes from the Test of Early Grammatical Impairment (TEGI; Rice & Wexler, 2001) and MLJ (Duncan et al., 2009). The TEGI Past Tense probe (TEGT) assessed production of regular and irregular past tense verbs, and the Third Person Singular probe (TEGS) assessed production of present tense verb forms with singular subjects. The MLJ task is a measure of morphological awareness that assessed knowledge of morphological relations between words. Children were asked whether two phonologically similar words come from the same morphological family (e.g., heat heater) or not (e.g., ham-hammer). All 20 test items were given to all children.

The remaining measures were administered in all grades. The Test for Reception of Grammar–Version 2 (TROG; Bishop, 2003) assessed comprehension of 20 grammatical structures using a multiple-choice format. Children were asked to choose the picture, among an array of four, that depicted the sentence. Each structure is tested in a block of four items, and all four must be correct to pass the block. The number of blocks passed was the score used in analyses. The Word Structure (WS) subtest of the CELF-4 assessed the ability to use morphemes for marking inflections, derivations, comparison, and pronoun use. The morphological derivation task (MDT; Wagner, n.d.) assessed morphological derivation knowledge. Children were asked to complete a sentence by deriving a word from a given word. An example test item is "Sudden. The bus stopped _____." (correct response: suddenly). The test contains 32 items and has a ceiling of eight incorrect items.

Narrative Language Sample

The Dragon Story from the Test of Narrative Language (Gillam & Pearson, 2004) was administered in kindergarten to elicit a narrative retell. The examiner read the story aloud and then asked the child to retell the story in their own words. One examiner prompt was provided at the end of the retell to elicit more details (e.g., "Is there anything else you remember about the story?"). The retell was audio-recorded and later transcribed by research assistants.

Data Transcription, Coding, and Computation

Analysis of the language samples was conducted in two stages. First, the samples were transcribed following procedures from the larger LARRC study. Each retell was transcribed by a research assistant into a series of communication units (C-units). An additional 20% of retells were randomly selected and transcribed by a second research assistant; as noted in previous research (LARRC et al., 2016), reliability was over 95% across 3 years.

Additional procedures were conducted for this study, which involved Systematic Analysis of Language Transcripts (SALT; Miller et al., 2019) coding for morphemes, and coding for PGU. Only complete and intelligible (C&I) utterances/C-units were included in the analysis set. We chose to exclude C-units that were not related to the task of the narrative retell (e.g., a comment about a buzzer going off). Total utterances (TotUtt), MLU, and NDW were calculated.

Several steps were followed to calculate PGU in accordance with procedures outlined in previous research (e.g., Eisenberg et al., 2012; Guo & Schneider, 2016; & Guo et al., 2019). First, C-units that did not contain a verb (except those where the missing verb was a copula), were excluded from analyses using the code (XU). Next, any utterances with grammatical or semantic errors were coded with (EU). Judgments about grammar were made based on descriptive, not prescriptive, and grammar (e.g., C-units ending with a preposition were not counted as errors). Finally, PGU was calculated as countable utterances (TotUtt minus XU) with no errors divided by the number of countable utterances; in other words, total number of grammatical utterances divided by the total number of utterances included for analysis (# of C&I utterances − # of XU − # of EU)/(# of C&I utterances − # of XU). Participants who had fewer than five countable utterances (n = 72) were excluded from analyses (Guo & Spencer, 2017).

Coding for SALT and PGU was conducted by six research assistants under the direction of the first author. The lead coder was a graduate SLP student, and the other

five assistants were undergraduate and graduate students. Each research assistant completed the SALT Transcription Self-Paced Online Course modules to learn SALT coding conventions (Miller et al., 2019) and participated in a training session led by the first author to learn PGU coding conventions (Eisenberg & Guo, 2013; Guo & Schneider, 2016). Next, 10 practice transcripts were coded by each research assistant. Results were compared, and any areas of difficulty were further explained. Following completion of training, the transcripts were evenly distributed among research assistants for coding. Regular meetings were held throughout the coding period where the first author answered questions and clarified any areas of confusion. In addition, the lead coder did spot checks, randomly selecting a few completed transcripts from each other coder to check for accuracy. The lead coder offered explanations if errors were observed, and the two coders resolved any discrepancies. One coder required some additional training for coding error utterances (e.g., the coder sometimes missed subtle grammatical errors such as shift in verb tense). Once initial coding of all transcripts was completed, 65 transcripts (18%) were randomly selected for double coding by either the lead coder or one other coder. Discrepancies were discussed and resolved. The interrater reliability, measured with intraclass correlation, was > .98 for each of the SALT metrics and .92 for utterance-level errors (EU).

Data Analyses

Univariate descriptive statistics were calculated for the LSA metrics (see Table 1), and bivariate correlations were calculated among LSA metrics and language measures (see Supplemental Materials S1–S4). A series of structural equation models were estimated in Mplus Version 8.3 (Muthén & Muthén, 2017). In structural equation modeling, individual measures known as *observed variables* are considered indirect measures of underlying constructs known as *latent factors*. The variance of the observed variables is partitioned into (a) shared variance, which is the variance that can be explained by the latent factor, and (b) residual variance, which is the variance unique to an individual-observed variable, also called error variance. Latent factor scores are therefore more reliable than any one measure, because they do not contain the error variance from the individual observed scores (Kline, 2011, p. 9). Whereas individual measures of language have their unique sources of error, a parent factor combines information across measures and removes the individual error components. In this study, structural equation models were used to examine the extent to which LSA metrics measured in kindergarten were associated with oral language in kindergarten and later in Grades 1, 2, and 3. Latent factors for oral language were created using the observed variables CELF-WS, TROG, MDT, PPVT, EVT, and CELF-WC measured in each year, and TEGS, TEGT, and MLJ measured in kindergarten only. Four separate models were estimated, each including the oral language factor for one grade and the LSA metrics. Fit of the structural equation models was evaluated using the root-mean-square error of approximation, standardized root-mean-square residual, comparative fit index, and Tucker-Lewis index (Kline, 2011).

Table 1. Descriptive statistics for language sample analysis variables.

Metrics	*M*	*SD*	Minimum	Maximum
Total number of utterances	11.14	6.45	5	70
Mean length of utterance	7.53	1.67	2	13.4
Number of different words	44.82	19.75	9	143
Percent grammatical utterances	0.8	0.17	0.14	1

Results

Descriptive Statistics

Means, standard deviations, and ranges for the LSA metrics are reported in Table 1. Bivariate correlations between all variables are reported in Supplemental Materials S1–S4. Among the LSA metrics, total number of utterances was highly correlated with NDW (r = .87), as expected based on the calculations for these metrics. Given the high correlation, we included only one of the variables in the structural equation models; NDW was selected because of its higher correlations with the language measures.

Correlations were also calculated between the individual observed language measures in each year and the kindergarten LSA metrics (MLU, NDW, and PGU). In kindergarten, the correlations ranged from .13 to .36. In Grade 1, there was one very low correlation (.03 between PGU and CELF-WC) and the rest ranged from .25 to .34. In Grade 2, the range was .12–.33, and in Grade 3, it was .16–.32. There was little change in these correlations over time.

Structural Equation Models

For the four models estimated for Grades K–3, model fit was good, with fit statistics within accepted ranges (see Table 2). Standardized parameters for these models are reported in Table 3. All observed language

Table 2. Fit statistics for structural equation models.

Fit statistic	Model 1	Model 2	Model 3	Model 4
Chi-Square (*df*, *p*-value)	124.12 (51, < .01)	41.95 (24, .01)	23.67 (24, .48)	35.98 (24, .06)
RMSEA	0.071	0.051	< 0.01	0.042
CFI	0.936	0.979	1	0.984
TLI	0.921	0.971	1	0.978
SRMR	0.047	0.032	0.024	0.035

Note. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root-mean-square residual.

variables loaded significantly onto the latent factors, with standardized factor loadings ranging from 0.50 to 0.87 (values of 0.5 or greater are desired). Bivariate correlations were estimated between the LSA metrics and latent language factors (see Table 3); these ranged from .30 to .39.

Regarding the prediction of language from the LSA metrics, PGU, MLU, and NDW all significantly and uniquely predicted concurrent (kindergarten) and later (Grades 1, 2, and 3) language. Standardized path estimates ranged from 0.36 to 0.39 for PGU, 0.20 to 0.26 for MLU, and 0.23 to 0.30 for NDW. Values for standardized path estimates can range from −1.0 to +1.0 and can be interpreted similarly to standardized coefficients from multiple regression. They represent the unique predictive relation between an LSA metric and the language factor.

Discussion

This study adds to extant literature related to the validity of narrative LSAs. Our results were similar to those found previously in that LSA metrics explain a small yet significant amount of variance in language ability. Our results add to existing research in that we used latent variables for language, which removes the measurement error associated with individual observed language measures, and we examined the relationship longitudinally. Specifically, metrics from a narrative language sample collected in kindergarten uniquely and significantly predicted children's language ability each year from kindergarten to Grade 3. This occurred even with language samples that were very brief (average length of 11 utterances). These findings support the value of narrative LSA for practitioners in the field as well as for researchers interested in predicting oral language abilities. Narrative LSA is an assessment method that closely resembles everyday academic language use and thus provides functional information about a child's language and, when intervention is needed, informs the selection of intervention targets. We acknowledge that our sample overrepresented White children in upper middle-class homes, and therefore, our results cannot be assumed to apply to the broader U.S. population.

The longitudinal nature of this study allowed examination of both concurrent (in kindergarten) and predictive relations between LSA metrics and language ability. The importance of each of the LSA metrics was found to be relatively consistent across grades, suggesting that with a stable measure of language ability such as that provided by the latent construct, measures of PGU, MLU, and NDW in kindergarten remain informative predictors of language ability through Grade 3. Latent constructs of language ability at each grade allow the relations between LSA measures and language to be estimated independent of the measurement error that would be associated with any single measure of language ability.

In our study, PGU was found to have the strongest predictive validity with the oral language construct. This study is one of the first to examine the use of PGU in a longitudinal sample of primary grade students. Considered with the growing body of other research on PGU and its psychometric properties (e.g., Guo et al., 2019), the evidence suggests that PGU may be a useful clinical tool in the assessment of children's language skills; additional research is needed to further elucidate its clinical utility. We do note, however, that thorough training procedures will be essential for SLPs to reliably code for PGU. Although we did obtain high interrater reliability in our study, our graduate student coders had somewhat more difficulty becoming reliable on PGU than the other metrics. One potential reason is the newness of this type of scoring/assessment. It is possible that while graduate students have much experience coding language transcripts using conventional metrics such as MLU, they do not have much experience making a more global judgment of accuracy. Additionally, our training materials and procedures for PGU, as compared with these more common LSA metrics, were not as established.

Past research has shown that the task type used to elicit a language sample can impact results of LSA analyses (e.g., Evans & Craig, 1992; Nippold et al., 2005). Our results were obtained using a narrative retell procedure. A

Table 3. Standardized parameter estimates from Models 1 to 4.

Variable	Model 1 (Kdg Language)			Model 2 (Gr1 Language)			Model 3 (Gr2 Language)			Model 4 (Gr3 Language)		
	Est.	*SE*	*p* value	Est.	*SE*	*p* value	Est.	*SE*	*p* value	Est.	*SE*	*p* value
Language latent factor loadings												
CELF-WS	0.69	0.04	< .01	0.70	0.04	< .001	0.60	0.05	< .001	0.52	0.05	< .001
TROG	0.71	0.03	< .01	0.71	0.04	< .001	0.73	0.03	< .001	0.72	0.04	< .001
MDT	0.64	0.04	< .01	0.77	0.03	< .001	0.76	0.03	< .001	0.78	0.03	< .001
PPVT	0.78	0.03	< .01	0.84	0.02	< .001	0.86	0.02	< .001	0.83	0.03	< .001
EVT	0.86	0.02	< .01	0.86	0.02	< .001	0.87	0.02	< .001	0.86	0.02	< .001
CELF-WC	0.68	0.04	< .01	0.55	0.05	< .001	0.58	0.05	< .001	0.80	0.03	< .001
TEGS	0.52	0.05	< .01									
TEGT	0.50	0.05	< .01									
MLJ	0.59	0.04	< .01									
Path estimates to latent language factor												
PGU	0.39	0.05	< .001	0.37	0.05	<.001	0.36	0.06	< .001	0.36	0.06	<.001
MLU	0.24	0.06	< .001	0.24	0.06	< .001	0.26	0.07	< .001	0.20	0.07	0.004
NDW	0.30	0.06	< .001	0.28	0.06	< .001	0.23	0.06	< .001	0.23	0.07	< .001
Correlations among observed variables												
PGU WITH MLU	.05	.06	.42	.05	.06	.42	.05	.06	.42	.05	.06	.42
PGU WITH NDW	–.07	.06	.28	–.07	.06	.28	–.07	.06	.28	–.07	.06	.28
MLU WITH NDW	.46	.05	< .001	.46	.05	< .001	.46	.05	< .001	.46	.05	< .001
Correlations with latent variables												
	With KdgLANG			With Gr1LANG			With Gr2LANG			With Gr3LANG		
MLU	.39			.38			.38			.32		
NDW	.38			.37			.33			.30		
PGU	.38			.36			.36			.36		

Note. Kdg = kindergarten; Gr1 = Grade 1; Gr2 = Grade 2; Gr3 = Grade 3; Est. = estimate; *SE* = standard error; CELF-WS = Clinical Evaluation of Language Fundamentals–Fourth Edition, Word Structure subtest; TROG = The Test for Reception of Grammar–Version 2; MDT = Morphological Derivation Task; PPVT = Peabody Picture Vocabulary Test–Fourth Edition; EVT = Expressive Vocabulary Test–Second Edition; CELF-WC = Clinical Evaluation of Language Fundamentals–Fourth Edition, Word Classes subtest; TEGS = Test of Early Grammatical Impairment, Third Person Singular probe; TEGT = Test of Early Grammatical Impairment, Past Tense probe; MLJ = Morphological Lexical Judgment; PGU = percent grammatical utterances; MLU = mean length of utterance; NDW = number of different words; LANG = latent language factor.

retell might constrain children's vocabulary use, which might affect the usefulness of a metric such as the NDW. For example, asking children to retell a simple story with limited complex vocabulary might result in a retell that is similarly limited, even with a child with average or above-average language abilities. Also, this task may not be the best choice with an older child; it may be that a self-generated story or an expository retell might be a more appropriate task for upper elementary students. Although this study showed that a single retell is useful in predicting later language, further research should investigate the effects of task type on children across the age span, including at what ages various tasks or metrics are most useful. Additionally, research has shown that clinicians are more likely to collect conversational language samples rather than narrative samples (Pavelko et al., 2016), which furthers the case for extending this line of research to examine task type, and perhaps more importantly, if the same LSA metrics across tasks would be the most predictive of overall language construct.

Many children on an SLP caseload have developmental language disorders (DLDs), and research clearly shows that many of these students will manifest reading comprehension difficulties as they progress through elementary school (Catts et al., 2002). Further research could examine the usefulness of LSA metrics to predict reading ability, particularly reading comprehension. Converging evidence has shown that higher level language skills, such as text comprehension and inferencing, are predictive of later reading comprehension (Oakhill & Cain, 2012). However, researchers continue to identify single measures that can be efficiently administered and scored, which would be able to predict risk for later reading disabilities. If a brief language sample, as was obtained in this study, can be shown to do this, then it could be used as a type of screening measure. Because school-aged children continue to evidence grammatical errors beyond verb tense marking (e.g., Scott & Windsor, 2000), PGU might be a particularly relevant metric to capture risk of language disorder and, thus, potential concomitant reading disabilities. This would be an important contribution, as many screening measures predict word reading risk but not as many have shown good predictive validity for later reading comprehension (e.g., Adlof & Hogan, 2019; Elwér et al., 2013; but see Catts et al., 2016).

In conclusion, LSA has long been recommended as an ecologically valid way to supplement standardized test results, determine treatment goals, and monitor progress

(Costanza-Smith, 2010), and our results provide support for including narrative LSA to supplement standardized language assessments. We found not just concurrent but also predictive relations between narrative LSA and a standardized oral language construct. PGU in particular showed promise as a clinically feasible metric of children's grammatical ability. Further work should continue to explore the use of LSA across task types, ages, and populations (e.g., children with DLD or who are dual language learners) in order to advise clinicians on best practices for assessment. Longitudinal research could also provide clinical guidance on whether specific LSA metrics can identify children at risk for DLD.

Acknowledgments

This article was prepared by a Language and Reading Research Consortium (LARRC) task force consisting of Mindy Bridges, Kimberly Murphy (convenor), Alisha P. Springle, Mollee Sultani, and Autumn McIlraith. LARRC project sites and investigators are as follows:

- Ohio State University (Columbus, OH): Laura M. Justice (Site PI), Richard Lomax, Ann O'Connell, Jill Pentimonti (now at University of Notre Dame, Notre Dame, IN), Stephen A. Petrill (coinvestigator from 2010 to 2013), and Shayne B. Piasta
- Arizona State University (Tempe, AZ): Shelley Gray (Site PI) and Maria Adelaida Restrepo
- Lancaster University (Lancaster, UK): Kate Cain (Site PI)
- University of Kansas (Lawrence, KS): Hugh Catts (Site PI, now at Florida State University, Tallahassee, FL), Mindy Bridges (now at University of Kansas Medical Center, Kansas, KS), and Diane Nielsen
- University of Nebraska-Lincoln (Lincoln, NE): Tiffany Hogan (Site PI), Jim Bovaird, and J. Ron Nelson (coinvestigator from 2010 to 2012)
- MGH Institute of Health Professions (Boston, MA): Tiffany Hogan (Site PI).

This work was supported by Grant R305F100002 of the Institute of Education Sciences' Reading for Understanding Initiative. We are deeply grateful to the numerous staff, research associates, school administrators, children, and families who participated. Key personnel at study sites included Crystle Alonzo, Lisa Baldwin-Skinner, Lauren Barnes, Garey Berry, Beau Bevens, Jennifer Bostic, Shara Brinkley, Janet Capps, Tracy Centanni, Beth Chandler, Lori Chleborad, Emmanuel Cortez, Willa Cree, Dawn Davis, Kelsey Dickerhoof, Jaclyn Dynia, Michel Eltschinger, Kelly Farquharson, Tamarine Foreman, Yvonne Fraser, Abraham Aldaco Gastelum, Rashaun Geter, Sara Gilliam, Alexandria Hamilton, Cindy Honnens, Miki Herman, Hui Jiang, Elaine Joy, Jaime Kubik, Trudy Kuo, Gustavo Lujan, Chi Luu, Junko Maekawa, Carol Mesa, Denise Meyer, Maria Moratto, Kimberly Murphy, Marcie Mutters, Amy Pratt, Trevor Rey, Lizeth Sanchez-Verduzco, Amber Sherman, Shannon Tierney, Stephanie Williams, and Gloria Yeomans-Maldonado. We would like to thank student research assistants at Old Dominion University—Victor Lugo (lead), Rachel Barry, Sara Dundas, Anna Greco, Claire Mann, and Jawhaina Tomawis—for conducting Systematic Analysis of Language Transcripts and percent grammatical utterance coding.

References

Adlof, S. M., & Hogan, T. P. (2019). If we don't look, we won't see: Measuring language development to inform literacy instruction. *Policy Insights From the Behavioral and Brain Sciences, 6*(2), 210–217. https://doi.org/10.1177/2372732219839075

Armstrong, R., Symons, M., Scott, J. G., Arnott, W. L., Copland, D. A., McMahon, K. L., & Whitehouse, A. J. O. (2018). Predicting language difficulties in middle childhood from early developmental milestones: A comparison of traditional regression and machine learning techniques. *Journal of Speech, Language, and Hearing Research, 61*(8), 1926–1944. https://doi.org/10.1044/2018_JSLHR-L-17-0210

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185–1192. https://doi.org/10.1044/jslhr.4105.1185

Bishop, D. (2003). *Test for Reception of Grammar–Version 2 (TROG-2)*. Pearson.

Catts, H. W., Fey, M. E., Tomblin, B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research, 45*(6), 1142–1157. https://doi.org/10.1044/1092-4388(2002/093)

Catts, H. W., Nielsen, D. C., Bridges, M. S., & Liu, Y. S. (2016). Early identification of reading comprehension difficulties. *Journal of Learning Disabilities, 49*(5), 451–465. https://doi.org/10.1177/0022219414556121

Channell, M. M., Loveall, S. J., Conners, F. A., Harvey, D. J., & Abbeduto, L. (2018). Narrative language sampling in typical development: Implications for clinical trials. *American Journal of Speech-Language Pathology, 27*(1), 123–135. https://doi.org/10.1044/2017_AJSLP-17-0046

Costanza-Smith, A. (2010). The clinical utility of language samples. *SIG 1 Perspectives on Language Learning and Education, 17*(1), 9–15. https://doi.org/10.1044/lle17.1.9

Duncan, L. G., Casalis, S., & Cole, P. (2009). Early metalinguistic awareness of derivational morphology: Observations from a comparison of English and French. *Applied Psycholinguistics, 30*(3), 405–440. https://doi.org/10.1017/S0142716409090213

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4)*. Pearson.

Ebert, K. D., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools, 45*(4), 337–350. https://doi.org/10.1044/2014_LSHSS-14-0034

Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language impairment based on grammaticality.

Language, Speech, and Hearing Services in Schools, 44(1), 20–31. https://doi.org/10.1044/0161-1461(2012/11-0089)

Eisenberg, S. L., & Guo, L.-Y. (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools, 46*(2), 81–93. https://doi.org/10.1044/2015_LSHSS-14-0049

Eisenberg, S. L., Guo, L.-Y., & Germezia, M. (2012). How grammatical are 3-year-olds? *Language, Speech, and Hearing Services in Schools, 43*(1), 36–52. https://doi.org/10.1044/0161-1461(2011/10-0093)

Elwér, Å., Keenan, J. M., Olson, R. K., Byrne, B., & Samuelsson, S. (2013). Longitudinal stability and predictors of poor oral comprehenders and poor decoders. *Journal of Experimental Child Psychology, 115,* 497–516. https://doi.org/10.1016/j.jecp.2012.12.001

Evans, J. L., & Craig, H. K. (1992). Language sample collection and analysis: Interview compared to freeplay assessment contexts. *Journal of Speech and Hearing Research, 35*(2), 343–353. https://doi.org/10.1044/jshr.3502.343

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research, 39*(6), 1258–1262. https://doi.org/10.1044/jshr.3906.1258

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

Guo, L.-Y., & Eisenberg, S. (2014). The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. *American Journal of Speech-Language Pathology, 23*(2), 203–212. https://doi.org/10.1044/2013_AJSLP-13-0007

Guo, L.-Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances between 4 and 9 years of age for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *American Journal of Speech-Language Pathology, 28*(4), 1448–1462. https://doi.org/10.1044/2019_AJSLP-18-0228

Guo, L.-Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research, 59*(2), 317–329. https://doi.org/10.1044/2015_JSLHR-L-15-0066

Guo, L.-Y., & Spencer, L. J. (2017). Development of grammatical accuracy in English-speaking children with cochlear implants: A longitudinal study. *Journal of Speech, Language, and Hearing Research, 60*(4), 1062–1075. https://doi.org/10.1044/2016_JSLHR-H-16-0182

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools, 41*(4), 393–404. https://doi.org/10.1044/0161-1461(2009/09-0023)

Heilmann, J., & Westerveld, M. (2013). Bilingual language sample analysis: Considerations and technical advances. *Journal of Clinical Practice in Speech-Language Pathology, 15,* 87–93.

Kline, R. B. (2011). *Methodology in the social sciences: Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Language and Reading Research Consortium, Farquharson, K., & Murphy, K. A. (2016). Ten steps to conducting a large, multi-site, longitudinal investigation of language and reading in young children. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00419

Leonard, L. B., Haebig, E., Deevy, P., & Brown, B. (2017). Tracking the growth of tense and agreement in children with specific language impairment: Differences between measures of accuracy, diversity, and productivity. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3590–3600. https://doi.org/10.1044/2017_JSLHR-L-16-0427

Malvern, D., & Richards, B. (2012). Measures of lexical richness. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. https://doi.org/10.1002/9781405198431.wbeal0755

Miller, J. F., Andriacchi, K., & Nockerts, A. (2019). *Assessing language production using SALT software: A clinician's guide to language sample analysis* (3rd ed.). SALT Software.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus* (Version 8.3) [Computer software]. https://www.statmodel.com/index.shtml

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., 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*(3), 876–886. https://doi.org/10.1044/1092-4388(2013/13-0097)

Nippold, M. A., Hesketh, L. J., Duthie, J. K., & Mansfield, T. C. (2005). Conversational versus expository discourse. *Journal of Speech, Language, and Hearing Research, 48*(5), 1048–1064. https://doi.org/10.1044/1092-4388(2005/073)

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*(2), 91–121. https://doi.org/10.1080/10888438.2010.529219

Owens, R. E., & Pavelko, S. L. (2017). Relationships among conversational language samples and norm-referenced test scores. *Clinical Archives of Communication Disorders, 2*(1), 43–50. https://doi.org/10.21849/cacd.2017.00052

Pavelko, S. L., & Owens, R. E. (2017). Sampling Utterances and Grammatical Analysis Revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools, 48*(3), 197–215. https://doi.org/10.1044/2017_LSHSS-17-0022

Pavelko, S. L., & Owens, R. E. (2019). Diagnostic accuracy of the Sampling Utterances and Grammatical Analysis Revised (SUGAR) measures for identifying children with language impairment. *Language, Speech, and Hearing Services in Schools, 50*(2), 211–223. https://doi.org/10.1044/2018_LSHSS-18-0050

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools, 47*(3), 246–258. https://doi.org/10.1044/2016_LSHSS-15-0044

Pavelko, S. L., Price, L. R., & Owens, R. E. (2020). Revisiting reliability: Using Sampling Utterances and Grammatical Analysis Revised (SUGAR) to compare 25- and 50-utterance language samples. *Language, Speech, and Hearing Services in Schools, 51*(3), 778–794. https://doi.org/10.1044/2020_LSHSS-19-00026

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Rice, M. L., & Wexler, K. (2001). *Rice/Wexler Test of Early Grammatical Impairment*. The Psychological Corporation. https://cldp.ku.edu/rice-wexler-tegi

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*(2), 324–339. https://doi.org/10.1044/jslhr.4302.324

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

Snowling, M. J., Duff, F. J., Nash, H. M., & Hulme, C. (2016). Language profiles and literacy outcomes of children with resolving, emerging, or persisting language impairments. *The Journal of Child Psychology and Psychiatry, 57*(12), 1360–1369. https://doi.org/10.1111/jcpp.12497

Spencer, T. D., & Petersen, D. B. (2020). Narrative intervention: Principles to practice. *Language, Speech, and Hearing Services in Schools, 51*(4), 1081–1096. https://doi.org/10.1044/2020_LSHSS-20-00015

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency? *Communication Disorders Quarterly, 29*(1), 43–53. https://doi.org/10.1177/1525740108314866

Ukrainetz, T. A., & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy, 18*(1), 59–78. https://doi.org/10.1191/0265659002ct227oa

Wagner, R. K. (n.d.). *Morphological derivation task*. Florida State University.

Westerveld, M., Gillon, G., & Miller, J. (2004). Spoken language samples of New Zealand children in conversation and narration. *Advances in Speech-Language Pathology, 6,* 195–208. https://doi.org/10.1080/14417040400010140

Williams, K. T. (2007). *Expressive Vocabulary Test–Second Edition (EVT-2)*. Pearson.