Tell or Retell? The Role of Task and Language in Spanish-English Narrative Microstructure Performance

Mary Claire Wofford,[1] Jessica Cano,[1] J. Marc Goodrich,[2] and Lisa Fitton[3]

[1]Department of Communication Sciences and Disorders, Western Carolina University

[2]Department of Teaching, Learning, & Culture, Texas A&M University

[3]Communication Sciences and Disorders Department, University of South Carolina

Author Note

Mary Claire Wofford https://orcid.org/0000-0001-6243-7214

J. Marc Goodrich https://orcid.org/0000-0003-1072-8305

Lisa Fitton https://orcid.org/0000-0003-0524-7339

The research reported in this paper was supported by a New Investigators Research Grant awarded to Lisa Fitton and J. Marc Goodrich by the American Speech-Language-Hearing Foundation.

Correspondence concerning this article should be addressed to Lisa Fitton, Communication Sciences & Disorders Dept., 1705 College Street, Columbia, SC, 29208. Email: fittonl@mailbox.sc.edu Phone: 517-614-7264

Abstract

Purpose: This study examined performance of dual language learners (DLLs) on Spanish- and English-language narrative story retells and unique tells. Transcription and analysis focused on comparisons of common microstructural language sample measures in Spanish and English across tasks. Each language sample measure was evaluated for its possible convergence with norm-referenced standardized assessments for DLL children.

Method: Spanish-English DLLs (n = 133) enrolled in English-only kindergarten or first grade classrooms completed two language sample tasks (one in each language), which were transcribed and analyzed using Systematic Analysis of Language Transcripts (Miller & Iglesias, 2017) for measures of syntactic complexity (MLU in words), lexical diversity (NDW), and grammaticality (percent grammatical utterances; PGU). Students also completed a norm-referenced sentence repetition task (Peña et al., 2014) and expressive vocabulary assessment (Martin, 2013).

Results: Comparison of story retells and unique stories revealed similar performance on MLU, NDW, and PGU across elicitation techniques, with one exception: NDW in Spanish was higher in the story retell condition. Predictive models revealed several differences in the relations between the microstructure measures and norm-referenced language measures by elicitation technique, though neither context demonstrated a consistent advantage across all metrics.

Conclusions: Measures derived from story retells and unique tells offer practical findings for SLPs and other educators to use in assessment of early-grade DLLs. This work increases knowledge of procedural differences across narrative assessments and their influence on language variables, supporting school based SLPs in making assessment decisions for DLLs on their caseload.

Keywords: narrative, bilingual, language sample analysis

Tell or Retell? The Role of Task and Language in Spanish-English Narrative Microstructure Performance

Dual language learners (DLLs) are a group of children characterized by numerous unique demographic characteristics that are tied to their language development. These variable characteristics include but are not limited to home language, heritage language learner status, race/ethnicity, nativity, age of exposure, socioeconomic status, and current community (Committee on Fostering School Success for English Learners, 2017). While the number of total DLLs served by the education system is difficult to estimate (Capps et al., 2015), there is consensus around the continued growth in the number of DLLs (Hemphill & Vanneman, 2011) and consequently in the number of DLLs requiring specialized classification, assessment, and intervention/modification for language-related disabilities (Abedi, 2008) in the United States.

Assessing DLLs in both the native language (L1) and the second language (L2) is necessary for comprehensive language evaluation because DLLs' language-specific skills are distributed across languages based on the level of exposure to each language (Quiroz et al., 2010). Because many young DLLs in the U.S. begin formal schooling in a language that is not their L1, the early elementary years often produce dynamic levels of relative proficiency in DLLs' two languages (Castilla-Earls et al., 2019; Rojas & Iglesias, 2013). When DLLs are assessed in only one of their languages, only a portion of their knowledge and skills are evaluated. From this partial view of a child's language ability, low proficiency may be mistaken for language impairment or other learning issues (Bedore & Peña, 2008; Kohnert, 2010). Using multiple methods of assessment allows for identification of converging evidence of language difficulty or disorder for more accurate diagnostic decision-making (Castilla-Earls et al., 2020).

The American Speech Language Hearing Association (ASHA, 2021) highlights language sampling as a valid, evidence-based assessment approach for assessing children who speak more than one language. Language sampling is considered a culturally responsive form of assessment for DLLs because it offers a wealth of information in a highly naturalistic, ecologically valid clinical task (Cleave et al., 2010; Gutiérrez-Clellen & Simon-Cereijido, 2009; Restrepo, 1998). Language sample analysis (LSA) is commonly used in evaluations and progress monitoring to examine key linguistic elements produced by DLLs in their two languages (Gutiérrez-Clellen, 2002; Gutiérrez-Clellen et al., 2000). Linguistic information including microstructural (e.g., lexical diversity, grammatical accuracy, syntactic complexity) and macrostructural elements of language (e.g., story structure, organization, coherence) have diagnostic utility in language assessment and are easily obtained from language sampling tasks (Méndez et al., 2018; Miller et al., 2006). Narrative microstructure and macrostructure represent distinct constructs that underlie narrative ability and contribute unique information to clinical assessment (Westerveld & Gillon, 2010). Among DLLs, measures of macrostructure generally appear to be associated across children's languages, while measures of microstructure do not (Boerma et al., 2016; Méndez et al., 2018; Squires et al., 2014). This suggests that narrative macrostructure may reflect more language-independent, transferable underlying language skills, whereas microstructure is likely more specific to each of DLLs' distinct linguistic systems.

Among early elementary DLLs, a common language sampling technique discussed in empirical evidence is story retell using a wordless picture book, in which the examiner tells the child a story and then asks the child to retell that same story. When provided an initial model, the examiner can track specific linguistic elements in the child's story in subsequent analysis using transcription software (Miller & Iglesias, 2017) or in real-time (Justice et al., 2010). Another common elicitation technique is spontaneous storytelling, also known as a unique story tell or story generation, in which a child produces a narrative without an examiner model. Unique story tell is an option that assumes familiarity with the story retell task (J. F. Miller et al., 2019, p. 302) and requires the child to spontaneously generate a story in conjunction with a single picture, a series of pictures, or wordless picture book stimuli. Spontaneous storytelling and story retelling tasks are considered distinct from other elicitation techniques (i.e., conversational, play) (Bliss & McCabe, 2006). Narratives additionally offer a more structured opportunity for children to produce language than open-response tasks (Govindarajan & Paradis, 2019), particularly when elicited in conjunction with pictures. Pictorial support reduces memory load and provides organizational guidance for narrative storytelling (Bliss & McCabe, 2006; Kapantzoglou et al., 2017).

In the present study, we focused on microstructural measures derived from two narrative language tasks: unique story generation and story retells. A substantial body of evidence supports the use of narrative language samples both for their clinical utility in detecting developmental changes in typical language growth (Bedore et al., 2010; Lucero, 2018; Orizaba et al., 2020) and in differentiating typical development from language impairment across monolingual and bilingual children (Hipfner-Boucher et al., 2015). However, additional work is needed to develop more precise understanding of the strengths and limitations of elicited narrative measures, particularly for DLLs who use distinct microstructure in their two languages. DLLs' performance in microstructure appears to vary across their two languages in the early elementary grades when language dominance is likely to shift, particularly for children with language disorders. In studies of DLLs with language learning difficulties matched to children with typical language development, children with language disorders tended to produce less complex microstructure in narrative storytelling than children with typical development (Kapantzoglou et al., 2017; Squires et al., 2014).

Our study focused on DLLs' performance on three summary measures of microstructure—number of different words (NDW), mean length of utterance in words (MLU), and percentage of grammatical utterances (PGU). We focus on these three measures given their clinical utility in identification of language impairment (Bedore et al., 2010; Kapantzoglou et al., 2017) and their suitability to facilitating normative comparison, as they are summary measures of narrative microstructure. Although summary statistics do not replace individual line-by-line error analysis, they are frequently used in clinical practice (Ebert & Scott, 2014).

In the following literature review, we describe current evidence of how elicitation approach may influence the microstructure of children's language sample productions, discuss how microstructural measures complement and expand upon norm-referenced test scores, and provide rationale for the approaches used in the present work. Specifically, we aim to demonstrate the need for research examining microstructural measures elicited from Spanish-English dual language learners, and how these measures converge and diverge from children's scores on norm-referenced language assessments.

LSA Microstructural Differences by Elicitation Technique

Evidence suggests that the elicitation technique and context influence child narrative performance (Channell et al., 2018; Miles et al., 2006). In an effort to standardize elicitation approaches for normative comparison or progress monitoring, visual supports and language models are commonly used to provide structure for children's responses (Heilmann et al., 2016; Rojas & Iglesia, 2009). These supports are frequently applied both in research and clinical contexts by professionals evaluating narrative language performance of DLLs (Heilmann, Miller, & Nockerts, 2010; Rojas & Iglesias, 2013).

A common form of contextual support that has been used in narrative elicitation is language modeling. Language models are a feature of story retells which may be pre-recorded, live, or embedded in video (Gazella & Stockman, 2003; Klop & Engelbrecht, 2013). These models are then shared with the individual before asking them to tell the story back to the examiner. Increasing complexity of the language model also may lead to more complex output, as was observed in a study with a narrated video model (Holloway, 1986). At the level of a single sentence, DLLs tend to produce more adult-like phonological representations when provided a verbal model (Goldstein et al., 2004), which may explain the broader advantage of a language model on microstructural outcomes in creating a representation of the story during the initial model and then subsequent retelling.

Given the extra support that language models provide, there is evidence that monolingual children (Merritt & Liles, 1998; Westerveld & Gillon, 2010) and DLLs (Duinmeijer et al., 2012; Sheng et al., 2020) benefit from the language model provided in a story retell. Duinmeijer and colleagues (2012) observed higher microstructural complexity (e.g., embedded sentences, overall MLU in words) in story retelling compared to story generation tasks among a sample of Dutch-English speaking children with typical development (n = 38) and language disorders (n = 34). Grammaticality was not influenced by elicitation technique. In a sample of 75 Polish-English speaking children, participants produced greater complexity across both languages in story retells compared to unique stories, though no significant differences were observed lexical diversity or syntactic complexity (e.g., Type-Token Ratio, MLU; Otwinowska et al., 2018).

Overall, these studies suggest that elicitation approach does influence children’s narrative productions. The studies used distinct narrative elicitation materials and focused on microstructural measures (Duinmeijer et al., 2012), as well as distinct sequencing of tasks when compared to the current study (Otwinowska et al., 2018). However, across this work, evidence suggests the use of retell tasks may support narrative productions with greater microstructural outcomes compared to unique story generation, though there is variability across methods and measures. The current study addresses the need for additional examination of procedural differences for Spanish-English speaking DLLs across a continuum of language abilities. Furthermore, the current study addresses the influence of elicitation approach on microstructural indices obtained for narratives produced in both of their languages, not just the majority language.

Convergence with Standardized, Norm-Referenced Measures

Summary measures of microstructure are commonly used by SLPs when conducting LSA (Gutiérrez-Clellen & Simon-Cereijido, 2009). However, to interpret these measures accurately, there is a need to understand how they converge with, compliment, and diverge from currently available norm-referenced measures of bilingual language. We need to examine both how elicitation technique may influence children's MLU, NDW, and PGU and consider if elicitation technique potentially alters the constructs that MLU, NDW, and PGU purportedly reflect. To address this need, the criterion validity of summary microstructural measures has been explored relative to standardized, norm-referenced measures appropriate for DLLs. Kapantzoglou et al. (2017) observed classification accuracy of LSA microstructural measures from story retells and unique story tells in native language samples from DLLs with both typical development and diagnosed language disorder based on performance on the *Bilingual English-Spanish Assessment* (BESA; Peña et al., 2014) and teacher report. Classification accuracy was greatest in the story retell condition with grammaticality and lexical diversity as significant predictors. Moreover, classification accuracy was acceptable in the unique story tell condition with grammaticality and syntactic complexity as significant predictors. The current study extends this work both by considering microstructural measures in the DLLs' two languages and by evaluating the convergence and divergence of these metrics with norm-referenced measures in a large sample of DLLs. SLPs who use best practices will incorporate language sampling methods and understand their complementariness with norm-referenced measures (Ebert & Pham, 2017).

In a study of 170 kindergarten age Spanish-English speaking children (Bedore et al., 2010), microstructural measures derived from unique story tells were correlated with a norm-referenced measure of language ability (Peña et al., 2014). The microstructural measures that accounted for significant variance in norm-referenced language scores were MLU in English, grammatical utterances in English, and grammatical utterances in Spanish. A unique contribution of the study was its use of a composite variable to account for ability in both languages (Bedore et al., 2010). The current study builds upon this research both by examining differences in microstructural measures by elicitation technique and by evaluating the concurrent validity of the elicited microstructural measures for predicting grammar and vocabulary measured separately. This separation of language skills is important given evidence that suggests that more than one factor underlies language ability (Language and Reading Research Consortium (LARRC) et al., 2018; Lonigan & Milburn, 2017).

Importantly, LSA contains less bias than norm-referenced measures do when assessing DLLs. In a study observing norm-referenced assessment performance and narrative language measures in monolingual children and DLLs with specific language impairment, narrative language measures revealed similar performance across groups in microstructure (e.g., grammaticality, verb accuracy) on retell and spontaneous tasks, while norm-referenced assessment disadvantaged DLLs (Cleave et al., 2010). The authors cautioned that exclusive use of standardized norm-referenced measures of expressive morphosyntax may lead to difficulty in interpreting DLLs' expressive ability and that combined observation with LSA is recommended. Converging evidence demonstrates the importance of microstructural measures derived from narrative tasks as a differentiating metric for children with language learning difficulty (Liles et al., 1995) among children with both monolingual and bilingual language backgrounds (Kapantzoglou et al., 2017). It is critical to recognize that narrative microstructural measures complement but do not fully overlap with performance on norm-referenced measures (Bedore et al., 2010; Rojas & Iglesias, 2009) but can provide converging evidence of language ability and enhance bilingual language assessment. In sum, there is evidence that norm-referenced assessment and specific microstructural outcomes may be considered jointly to assist SLPs in clinical decision-making with DLLs (Ebert & Pham, 2017; Ebert & Scott, 2014).

Despite the evident utility of LSA, its integration as a staple of language evaluation protocols poses a challenge in the field. A survey of school-based SLPs (*n* = 1,399) indicated that most clinicians rely on brief, real-time analysis of conversation rather than full transcription when evaluating language samples (Pavelko et al., 2016). SLPs' responses overwhelmingly indicated (78%) that evidence-based procedures for LSA were used infrequently due to the length of time required to transcribe and analyze samples (Pavelko et al., 2016). One possible explanation is a lack of information about what is gained from LSA. Currently, evidence shows that the type of procedure chosen for administration can influence performance on certain LSA measures (Scott & Windsor, 2000), though relatively scant literature discusses the nuances of procedural techniques in LSA specifically among DLLs (Kapantzoglou et al., 2017). Increased knowledge of the relations between LSA measures and norm-referenced assessments may increase evidence-based usage of LSA techniques in practice. Greater understanding of the differences between LSA tasks could clarify and illuminate rationale for its use. A working knowledge of procedural differences between story retell and unique story tell tasks and their influence on variables of interest in typically developing children will aid SLPs in their choice between these tasks as well as illuminate the implications of their choice.

Rationale for the Current Study

The purpose of this exploratory study was to examine if differences exist between unique story tells and story retells produced by Spanish-English speaking children enrolled in **kindergarten and first grade** across several common LSA measures that indicate lexical diversity (NDW), syntactic complexity (MLU in words), and grammaticality (PGU). Additionally, we examined the relations between microstructural language sample measures and children's scores on norm-referenced language outcomes. This builds on previous research examining task differences and convergence with norm-referenced measures in bilingual children (Bedore et al., 2010; Kapantzoglou et al., 2017). We explored differences in these LSA metrics separately in each language, while controlling for child age and sample length. Results of this study will add to existing evidence describing DLLs' performance on distinct language sampling tasks during a critical period of shifting language dominance when DLLs' language systems can appear to be in flux (Castilla-Earls et al., 2019). Knowledge about the task type and the language of administration will better inform SLPs about procedural differences in Spanish and English language samples and what distinct tasks offer the evaluating clinician. Based on prior evidence, we predicted that NDW and MLU would be greater in retell vs. unique tell, and there would be no difference in grammaticality across tasks (Duinmeijer et al., 2012; Fiestas & Peña, 2004 Kapantzoglou et al., 2017; Otwinowska et al., 2018).

Furthermore, we sought to observe the relations between DLLs' scores obtained from the

LSA measures and those obtained from language-specific norm-referenced assessments designed for Spanish-English bilingual children. We focused on norm-referenced assessments measuring Spanish vocabulary, Spanish morphosyntax, English vocabulary, and English morphosyntax for these comparisons. Specifically, we aimed to explore the possible influence of narrative elicitation technique on the relations between LSA microstructure measures and children's scores on norm-referenced language assessments.

For both MLU and NDW, we expected a positive association between the microstructural measure and the same-language norm-referenced language measures with an interaction between task type and microstructural measure. Some prior evidence suggests children may produce greater MLU following a language model (Duinmeijer et al., 2012), which would be in close alignment with current norm-referenced measures of bilingual morphosyntax (e.g., BESA Sentence Repetition). For NDW, we expected children to generate fewer different words during unique story generation compared to the story retell (Lucero & Uchikoshi, 2019), acknowledging that NDW from the unique story may better align with current norm-referenced measures of expressive language (Bedore et al., 2010). Finally, we predicted a positive association between PGU and norm-referenced language with no interaction by task type, given that PGU has been observed to be fairly stable across elicitation approaches (Kapantzoglou et al., 2017). We expected all hypothesized patterns to appear both in Spanish and English. The research aims, which were addressed separately in Spanish and English, were:

1. Are there differences in measures of microstructure (NDW, MLU, and PGU) on unique story tells and story retells produced by Spanish-English speaking children enrolled in kindergarten and first grade?
2. Do the relations between DLLs' narrative microstructure and norm-referenced

assessment performance differ based on elicitation technique (unique story tell vs. story retell)?

Method

Participants

Participants included 133 Spanish-English DLLs recruited as part of a larger study examining bilingual language and reading development. Children ranged in age from 5 years, 2 months to 7 years, 10 months (*M* = 6.34 years, *SD* = 0.68) and were in kindergarten (*n* = 86) or first grade (*n* = 47) at the time of participation. The children were enrolled in eleven different elementary schools, one located in South Carolina and ten in Nebraska, all of which provided English-only instruction. A total of 91 participants were recruited from the South Carolina school, and 42 participants were recruited from the ten Nebraska schools. Differences in recruitment rates are likely attributable to (a) the greater density of Spanish-speakers in the Midlands of South Carolina compared to southeast Nebraska, and (b) consent procedures governing each site, as passive consent procedures were used in South Carolina (consistent with Institutional Review Board approvals at the University of South Carolina) and active consent procedures were used in Nebraska (consistent with Institutional Review Board approvals at the University of Nebraska-Lincoln).

All students identified as having at least some Spanish exposure at home according to parent and/or teacher report were invited to participate in the study. All children enrolled in the participating schools were recruited to participate, regardless of developmental language status or eligibility classification(s). This approach was used to obtain a participant sample including students with a broad range of Spanish and English proficiencies, consistent with the heterogeneity observed in the larger Spanish-English speaking population in the United States.

Consent to participate was obtained from students' guardians. All procedures used were consistent with site-specific Institutional Review Board approvals at the University of South Carolina and University of Nebraska-Lincoln.

Procedure

Participants completed a battery of Spanish-English bilingual language measures including the Bilingual English-Spanish Assessment (BESA) Sentence Repetition task (Peña et al., 2014), the Expressive One-Word Picture Vocabulary Test-4: Spanish Bilingual Edition (EOWPVT-4 SBE; Martin, 2013), and narrative language samples during the middle of the kindergarten or first grade year. These assessments are psychometrically sound, age-appropriate, and specifically designed for Spanish-English speaking children. All assessments were administered in both Spanish and English by trained undergraduate and graduate research assistants. Children completed the full assessment battery within a two-week window.

Norm-Referenced Standardized Language Measures. Participants completed the BESA Sentence Repetition task separately in Spanish and English. For this task, children are asked to repeat sentences verbatim. Current evidence suggests that children's performance on sentence repetition tasks primarily reflects their morphosyntactic skill (Kapantzoglou et al., 2016; Polišenská et al., 2015; Rujas et al., 2021), though additional abilities including working memory and vocabulary may also contribute to DLLs' performance (Pratt et al., 2020). Raw and norm-referenced scores were obtained for each language, following BESA standardization guidelines (Peña et al., 2014). The BESA sentence repetition task is well-vetted, with evidence supporting it as a functionally unidimensional tool with good reliability (Fitton et al., 2019). Internal consistency is $\alpha = 0.96$ for Spanish and $\alpha = 0.95$ for English (Peña et al., 2014). The manual for the BESA reports strong evidence of construct validity for the morphosyntax subtest through differences in performance between children with and without language impairment, correlations with other norm-referenced language measures (*rs* range from .35 to .72), and high sensitivity and specificity for classifying language impairment.

The EOWPVT-4 SBE was administered separately in Spanish and English, consistent with evidence and recommendations provided by Anaya et al., 2018 and Gross et al., 2014. This work suggests that EOWPVT-4 SBE prompts should be explicitly provided in both languages to quantify bilingual expressive vocabulary accurately. For this assessment, participants are asked to name pictures they are shown. Based on participants' responses, three separate scores were derived. First, Spanish-only and English-only raw and norm-referenced scores were obtained. Then a conceptual vocabulary score was computed with participants receiving credit for responding correctly either in Spanish or English for each item. The EOWPVT-4 SBE also has good internal consistency reliability ($\alpha = 0.95$). The manual for the EOWPVT-4 SBE (Martin, 2013) reports strong correlations with other measures of vocabulary knowledge (*r*s range from .66 to .90), indicating strong construct validity. Additionally, the manual reports that performance on the EOWPVT-4 SBE differs significantly across individuals with and without disabilities, providing evidence of criterion validity.

Language Sample Tasks

Random Assignment. One Spanish language sample and one English language sample was elicited from each child. In adherence with SALT recommendations (Miller, Andriacchi, & Nockerts, 2019, p. 302-303), students always completed the story retell using *Frog Where Are You?* (Mayer, 1969) first to ensure that they had at least an initial exposure to the storytelling schema for the wordless picture books. Unique story tells were always completed with *One Frog Too Many* (Mayer, 1975). A large sample (*n* = 831) of Spanish-English bilingual children performed similarly across different titles in the wordless picture book series from Mayer on standard language sample measures (Heilmann et al., 2016). To assess the potential influence of how initial elicitation language may influence language sampling results, students were randomly assigned to either Spanish-first or English-first elicitation. Students assigned to Spanish-first completed the Spanish story retell and then the English unique story. Students assigned to English-first completed the English story retell and then the Spanish unique story. Randomization occurred within each research site (South Carolina vs. Nebraska), with students randomly assigned to condition upon enrollment.

For both task types, the administration in the current study followed the elicitation protocol for story retells provided in the SALT reference book (Miller et al., 2019). During the story retell, the examiner modeled the story for the child loosely following a script provided by SALT. The child was then asked to tell the story back to the examiner in the same language that the examiner told the story. Administration of the story tell occurred on a different day from the story retell and followed the elicitation protocol for unique story tells provided in the SALT reference book(Miller et al., 2019). In both scenarios, the examiner only provided minimal open-ended prompts (i.e., prompts that “do not provide the child with answers or vocabulary”, p. 272) to guide the child’s retelling of the story.

Spanish-language stories were administered by trained research assistants with native or near-native Spanish proficiency, and English-language stories were administered by a research assistant with native or near-native English proficiency. If significant code-switching occurred during the sample, the examiner prompted the child to use the target language with minimal interruption of the story, consistent with SALT administration guidelines.

Transcription. Recorded audio files of children’s language samples were transcribed by trained, experienced transcribers through Systematic Analysis of Language Transcripts (SALT) Transcription Services. Files were transcribed using standard SALT transcriptions and conventions, including code-switching at the utterance level. All transcripts were reviewed by a second, independent transcriber who corrected any spelling or convention errors. Additionally, 20% of the samples were double transcribed by an independent transcriber for reliability. To assess transcription reliability, the original and second versions of these transcripts were compared. Reliability was computed by dividing the number of matching units by the total number of units for each child utterance. For c-units segmented, percent agreement was 99.27%. For morphemes segmented, agreement was 99.13%. For words transcribed, agreement was 97.82%. For error codes identified, agreement was 96.84%.

Microstructure Measures. Formatted transcripts were loaded into SALT 18 Research Version 18.3.14 (Miller & Iglesias, 2017) for analysis. Metrics from the Standard Measures Report, including MLU in words, number of different words (NDW), and percent utterances with errors (PGU), were extracted for each transcript. We also obtained counts of the number of utterances including code-switching and the number of error codes (e.g., omitted words, omitted bound morphemes). All measures were examined descriptively. To compute PGU (Guo et al., 2019), the percent utterances with at least one grammatical error was subtracted from 100.

Exclusionary criteria (Code-switching)

To allow for comparison of how elicitation approach might influence narrative language in Spanish and English, some samples were excluded due to code-switching. Samples were excluded if more than 30% of the child's words were produced in the non-target language, similar to SALT Software (SALT Software LLC, 2020) protocols, which use a criterion of 20%. We elected to use a slightly less strict exclusion level for two primary reasons. First, much of the code-switching observed in our sample was restricted to single word substitutions rather than multiple words, which would minimally influence standard measures such as MLU, NDW, and PGU (as children were not penalized for grammatically-correct code switches). Second, unlike the SALT bilingual databases, our participant sample was not restricted to children being educated in English language learner classrooms. We included children with a wide range of Spanish and English proficiency, but all of whom were receiving English-only instruction. These environmental differences may influence bilingual children's linguistic development in a way that could influence word borrowing across languages (Byers-Heinlein, 2013).

Missing Data: COVID-19

Both recruitment and data collection were ongoing when schools closed due to the COVID-19 pandemic in March of 2020, resulting in missing data within the sample. At the time of school closures, 182 children were enrolled in the larger study and had been randomly assigned to Spanish-first or English-first elicitation of narrative language samples. In considering how to appropriately address this missing data, several points were relevant (Logan, 2020). First, 133 children had started testing, and most of these children had complete data. Of the Spanish assessments scheduled to be administered to these 133 children, 96% had been completed, whereas 91% of the scheduled English assessments had been completed. Second, school closures equally impacted all children enrolled in the study. All participation ended when schools closed, resulting in an equal likelihood for any enrolled child to have missing data. Third, the timing of assessment for any individual child depended on several external factors, such as individual classroom teacher timing preference, availability of assessors to complete assessments, and school schedule. We did not observe any patterns in the missing data across participants, sites, tasks, or languages. Consequently, data were treated as missing at random (MAR).

Analytic Approach

All analyses were conducted separately for Spanish and English. To examine differences in MLU, NDW, and PGU by elicitation approach, we used linear mixed models. This approach was taken to examine differences across story type after accounting for child age and total utterances produced, and to incorporate nesting of participants within different states (South Carolina and Nebraska). Although children were randomly assigned, small differences in age and utterances produced were observed by group (see Supplementary Table S1). Because child age and narrative productivity can influence standard measures of LSA, we elected to account for these variables in the analyses as covariates. To assist with interpretation of findings, Hedge's *g* values are provided as a metric of the standardized mean differences in MLU, NDW, and PGU by elicitation approach. Hedge's *g* is similar to Cohen's *d*, as it is based on Cohen's *d* effect sizes but includes a correction factor to address potential bias associated with the sample size (Hedges, 1981). Because interpretation of these effect sizes is field- and context-specific (Lakens, 2013; Thompson, 2007), we offer recommendations for considering the magnitude of the obtained effect sizes within the results and discussion sections.

To address the second aim of the study, we again used mixed effects modeling, but focused on the individual contribution of each LSA measure to two standardized and norm-referenced measures of language: sentence repetition and expressive vocabulary raw scores (examined separately). Age and total number of utterances were again included as covariates. Site was included as a random effect and task (retell versus unique story) as a fixed effect. To determine if task type influenced (i.e., moderated) the relation between any of the LSA measures and the norm-referenced measures, we examined interactions between task type and each LSA measure.

All analyses were conducted in R Version 3.6.3 (R Core Team, 2020) using the lme4 package (Bates et al., 2015). Restricted estimation maximum likelihood was used to limit bias in the estimation of variance parameters, given the relatively small sample size. For each model, residual values were plotted and examined for consistency with assumptions of residual independence, normality, and homogeneity of variance.

Results

From the full sample of 133 participating children, a total of 108 narrative language samples were elicited in Spanish and 111 language samples were elicited in English. Examination of code-switching revealed that 15 of these recordings included responses with more than 30% words produced in the non-target language (12 elicited in Spanish and 3 elicited in English). Six participants exhibited code-switching above 30% in both languages. Elimination of these samples resulted in a final participant sample of 127 students and an analytic dataset including 96 Spanish language samples and 108 English language samples. Within this dataset of 127 students, 77 participants produced samples in both Spanish and English.

The mean total number of utterances produced was similar across languages, with 24.72 (*SD* = 13.02) utterances produced on average in Spanish and 24.02 (*SD* = 15.11) utterances on average in English (see Table S1). The Spanish samples included 95% (*SD* = 0.08) intelligible words, similar to that observed within the English samples (95%, *SD* = 0.10). A mean of 9.90 (*SD* = 9.16) grammatical errors appeared in the Spanish samples. A mean of 7.69 (*SD* = 7.81) errors appeared in the English samples. Descriptive statistics and correlations among the LSA measures of primary interest, as well as the standardized scores obtained from the EOWPVT-4 SBE and the BESA Sentence Repetition task, are provided in Tables 1 (Spanish) and 2 (English).

To provide metrics of general underlying language abilities within the sample, we examined participating children's best language norm-referenced scores on the BESA Sentence Repetition, taking the highest score in either Spanish or English as recommended in the BESA Manual (Peña et al., 2014). We also report their conceptual vocabulary norm-referenced scores on the EOWPVT-4 SBE. Within the sample of participants who completed the English narratives, n = 6 participants had best language scores below 80, n = 8 scored between 80 and 85, and n = 59 scored 90 or above. An additional 7 participants who only completed the sentence repetition task in one language scored 85 or above. Overall, participants scored an average of 99.52 (SD = 13.90) in their best language and an average of 103.35 (SD = 15.56) for conceptual vocabulary. Within the sample of participants who completed the Spanish narratives, n = 4 participants had best language scores below 80, n = 8 scored between 80 and 85, and n = 61 scored 90 or above. An additional 3 participants who only completed the sentence repetition task in one language scored 85 or above. Overall, participants scored an average of 100.68 (SD = 12.67) in their best language and an average of 102.94 (SD = 15.58) for conceptual vocabulary.

Model Fit Considerations

Although intraclass correlation coefficients suggested some site-specific variation (see Tables 3-8), values ranged from 0 - 0.38. In some instances, it was not necessary to account for site-specific clustering of scores (e.g., NDW predicting vocabulary). In these cases, model results were nearly identical to those obtained from OLS regression.

Several outliers were identified in examining descriptive statistics and model fit diagnostics. Outliers are not surprising, given the variable and open-ended nature of narratives. The outliers represented children that simply produced long, complex samples. However, these outliers did seem to have disproportionate influence on the results. Rather than remove these representative cases from the dataset, we elected to bound the values at 1.5 times the interquartile range and re-run all analyses. This adjustment resolved concerns observed within the model diagnostics and did not substantially impact the primary results, nor their interpretation.

Aim 1 - Differences by Elicitation Approach

Results revealed significant differences between elicitation approaches in only one of the Spanish LSA measures, after accounting for child age, total utterances, and site. Children produced a slightly higher NDW (Hedge's $g = 0.23$, $SE = 0.21$, $p = .027$) in the story retell context compared to the unique story. Approximately 5.31 fewer different words were produced in the Spanish unique stories compared to the story retells. No significant differences were observed for MLU in words (Hedge's $g = 0.10$, $SE = 0.20$, $p = .539$) or PGU (Hedge's $g = 0.10$, $SE = 0.20$, $p = .647$) in the Spanish samples. Full model results are provided in Table S2.

Similar results were observed for the English LSA measures, although the difference in NDW by elicitation approach was smaller and did not meet conventional criterion for significance: Hedge's $g = 0.19$, $SE = 0.20$, $p = .051$. No significant differences were observed for MLU in words (Hedge's $g = 0.01$, $SE = 0.20$, $p = .984$) or PGU (Hedge's $g = 0.08$ $SE = 0.20$, $p = .722$) when age and total utterances were held constant. Full results are available in Table S3.

Aim 2 - Concurrent Criterion: LSA Predicting Language Measures

To maximize readability, results from statistical models including interaction terms are provided only in text throughout this section. These interaction terms provided an overall test of differences in the predictive relations between the LSA measures and the language measures by elicitation technique (i.e., did LSA measures elicited from the unique story more strongly predict outcomes than those elicited from the retell?). The main effects models with estimates separated out by elicitation approach are reported fully in Tables 3-8. Standardized estimates based on z-scored predictors and outcomes are provided in Table S4 for all predictive models.

Spanish Measures

MLU - Spanish. Models examining the predictive relations between MLU and Spanish sentence repetition favored the story retell approach, evidenced by a significant interaction between MLU and elicitation technique: -2.20, 95% CI [-4.23, -0.17], p = .033. As shown on the left half of Table 3, Spanish MLU in words predicted Spanish sentence repetition to a lesser degree when elicited in the unique story context compared to the story retell, with age and total number of utterances (TNU) held constant. Specifically, a 1-word increase in MLU elicited from the unique story context corresponded with a 1.84 (95% CI [0.32, 3.36], p = .018) increase in participants' raw Spanish sentence repetition scores, whereas a 1-word increase in MLU from the story retell corresponded with a 3.19 (95% CI [1.29, 5.09], p = .001) increase in sentence repetition scores. See Table S4, lines 1-2, for estimates based on the z-scored measures.

The predictive relations between MLU and vocabulary, however, were stable across the elicitation approaches. Interactions between MLU and story type were not statistically significant in predicting Spanish vocabulary: -3.39, 95% CI [-7.35, 0.58], p = .094. As shown on the right side of Table 3, a 1-word increase in MLU corresponded with either a 5.29 (95% CI [2.18, 8.40], p = .001) or a 5.41 (95% CI [2.21, 8.60], p = .001) increase in participants' raw Spanish vocabulary scores, whether elicited from the unique story or retell context, respectively. See Table S4, lines 1-2 on the right, for estimates based on the z-scored measures.

NDW - Spanish. No significant differences in the relations between NDW and either of the language measures were observed by elicitation approach, with interaction terms of -0.05 (95% CI [-0.16, 0.05], p = .331) for predicting sentence repetition, and -0.09 (95% CI [-0.31, 0.13], p = .397) for predicting vocabulary. Holding age and TNU constant, children's NDW in Spanish predicted sentence repetition and vocabulary consistently across the two elicitation approaches. A 1-word increase in NDW elicited from the unique story corresponded with a 0.33 (95% CI [0.18, 0.48], $p < .001$) increase in raw sentence repetition score, whereas a 1-word increase in story retell NDW corresponded with a 0.46 (95% CI [-0.16, 0.05], $p < .001$) increase in sentence repetition. Similar findings were observed for predicting Spanish vocabulary, with estimates of 0.75 (95% CI [0.43, 1.07], $p < .001$) obtained for unique story NDW and 0.64 (95% CI [0.36, 0.92], $p < .001$) for retell NDW (see Table 4). Results from the models based on z-scored measures are provided in Table S4, lines 3-4.

PGU - Spanish. No significant differences were observed for PGU as a predictor of vocabulary or sentence repetition by elicitation approach. For predicting Spanish sentence repetition, the interaction term by story = -0.82 (95% CI [-18.28, 16.65], $p = .927$). Predicting Spanish vocabulary, the interaction by story = 10.31 (95% CI [-23.13, 43.75], $p = .546$). Holding age and total utterances constant, participants' PGU in Spanish predicted sentence repetition consistently across the two elicitation approaches. As shown in Table 5, a 1.0% increase in unique story PGU corresponded with a 0.16 (95% CI [0.06, 0.26], $p = .001$) increase in Spanish sentence repetition score. Similarly, a 1.0% increase in story retell PGU corresponded with a 0.17 (95% CI [0.02, 0.32], $p = .024$) increase in sentence repetition.

PGU did not significantly contribute to predicting Spanish vocabulary above and beyond children's age and total number of utterances, regardless of elicitation context (see right side of Table 5). Although participants' PGU elicited from the unique story generally trended toward a positive association with Spanish vocabulary (0.19, 95% CI [-0.02, 0.41], $p = .076$), PGU elicited from the story retell did not (0.10, 95% CI [-0.15, 0.35], $p = .439$). Results from the models based on z-scored measures are provided in Table S4, lines 5-6.

English Measures

MLU - English. Models examining the predictive relations between MLU and the English language measures revealed no significant differences by elicitation approach, as evidenced by no significant interactions in predicting sentence repetition (-0.95, 95% CI [-2.55, 0.66], p = .249) or vocabulary (-2.02, 95% CI [-5.71, 1.67], p = .284). Children's MLU consistently contributed to predicting sentence repetition and vocabulary across the two elicitation approaches (see Table 6). A 1-word increase in MLU from the unique story corresponded with a 2.19 (95% CI [0.48, 3.90], p = .012) increase in English sentence repetition raw score. Similarly, a 1-word increase in story retell MLU corresponded with a 2.88 (95% CI [1.69, 4.06], p < .001) increase in sentence repetition. For predicting English vocabulary, a 1-word increase in unique story MLU corresponded with a 4.86 (95% CI [1.17, 8.55], p = .010) increase in vocabulary, similar to the 5.06 (95% CI [2.37, 7.75], p < .001) increase corresponding with a 1-word increase in retell MLU. See Table S4, lines 7-8, for results for z-scored measures.

NDW - English. Participants produced highly variable NDWs in English, particularly when elicited from the unique story context. Consequently, unique story NDW did not meet criteria for statistical significance in predicting English sentence repetition after accounting for age and TNU, though a modest positive trend was observed (0.14, 95% CI [-0.001, 0.28], p = .052). By contrast, story retell NDW did meet criteria for statistical significance as a predictor of English sentence repetition: 0.36, 95% CI [0.25, 0.48], p < .001. However, results from the interaction model were not statistically significant (-0.08, 95% CI [-0.17, 0.02], p = .119), suggesting that unique story NDW did not substantially differ from retell NDW in predicting sentence repetition. Taken together, these results indicate a modest positive association between English NDW and English sentence repetition, above and beyond age and TNU, regardless of elicitation context (see Table 7).

Similar complexity was evident in the interaction between NDW and story type for predicting vocabulary, favoring the NDW elicited from the story retell: -0.20, 95% CI [-0.40, 0.01], $p = .048$. A 1-word increase in unique story NDW corresponded with a 0.48 (95% CI [0.19, 0.77], $p = .001$) increase in raw vocabulary scores, whereas a 1-word increase in story retell NDW corresponded with a 0.77 (95% CI [0.52, 1.03], $p < .001$) increase in vocabulary (Table 7). Results based on z-scored measures are provided in Table S4, lines 9-10.

PGU - English. Participants' English PGU only predicted sentence repetition significantly when elicited from the unique story context (0.21, 95% CI [0.13, 0.29], $p < .001$). Both the interaction term (0.13, 95% CI [0.01, 0.25], $p = .040$) and main effect estimate indicated a significant difference in PGU predicting sentence repetition by elicitation context, with no significant relation observed between story retell PGU and sentence repetition (0.02, 95% CI [-0.07, 0.10], $p = .703$). A similar pattern was observed for PGU predicting vocabulary, with a generally positive association between unique story PGU and raw English vocabulary scores. However, unique story PGU did not meet criteria for statistical significance in predicting scores (0.21, 95% CI [-0.02, 0.44], $p = .068$), holding age and TNU constant. Retell PGU did not predict vocabulary: 0.03, 95% CI [-0.14, 0.21], $p = .701$. Results from the models based on z-scored measures are provided in Table S4, lines 11-12.

Discussion

The purpose of this study was to determine whether microstructural measures derived from narrative language assessments in Spanish and English vary by elicitation methods. An additional purpose of this study was to evaluate the relations between these measures of narrative microstructure and norm-referenced measures of language commonly used with DLLs, including vocabulary and sentence repetition tasks.

Differences in Microstructural Measures across Elicitation Approaches

Overall, results suggested that, for DLLs enrolled in English-only kindergarten and first grade classrooms, microstructural indices derived from language samples did not differ substantially across elicitation approaches in either Spanish or English. This finding has important implications for practicing clinicians, as it suggests that decisions to use story retells versus unique story tells when collecting a narrative sample largely does not dramatically influence DLLs' performance on microstructure summary measures. Typically, story retells are completed before a unique tell, to ensure that children have familiarity with the process of telling a story using a wordless picture book (Miller & Iglesias, 2017). Given evidence that microstructure scores derived from narrative language samples are sensitive to change among DLLs (e.g., Bedore et al., 2010; Orizaba et al., 2020) and can be used for progress monitoring purposes (Gorman et al., 2016), school based SLPs and clinicians may be interested in using narrative language sampling frequently to track progress with DLLs' language acquisition and development. Evidence that elicitation approach does not strongly influence children's microstructural performance can inform assessors in making decisions about how to elicit a narrative sample. Further, unique story tells have less potential for test-retest effects, given the absence of a model that could be memorized over repeated exposures. Importantly, these findings are limited to overall microstructural performance in narratives, and macrostructural analysis should be considered in tandem with microstructure.

Despite the overall non-significant differences by elicitation technique, subtle differences were observed. There was a small advantage in lexical diversity produced in the context of story retells when compared to story tells (*g*s ranging from .19 to .23). This finding was not surprising, as children hear the examiner tell the story in the context of the retell, which may prompt children to use certain words or structures during their own retell that they would not otherwise have used in a unique story tell. This priming effect may affect NDW most among the microstructural indices because pictorial support facilitates recall of highly imageable nouns, rather than morphosyntactic elements. Further, NDW is not calculated as an average as are the other microstructural indices. Consistent with our findings, prior evidence indicates that both monolingual and bilingual children included more content in their stories when retelling a story versus telling a unique story from pictures (Lucero & Uchikoshi, 2019; Schneider & Dubé, 2005). Differences in elicitation techniques did not result in differences in MLU or PGU in our sample which was consistent with past literature (Duinmeijer et al., 2012; Otwinowska et al., 2018). This suggests that clinicians should exercise caution when comparing microstructural indices of lexical diversity, such as NDW, across tell and retell formats.

Which Narrative Language Scores Predict DLLs' Language Outcomes on Norm-Referenced Measures?

Spanish

Regardless of elicitation technique, for Spanish language skills, NDW in Spanish narratives was the strongest predictor of norm-referenced measures of vocabulary and morphosyntax. Consequently, when assessing children's narrative skills in their home language, specifically in contexts in which the predominant language used at school is English, lexical diversity may be a key microstructural measure for clinicians to evaluate across children; however, additional research is needed to determine whether indices of lexical diversity such as NDW are strong clinical markers for language disorder among DLLs. Some prior research does indicate significant differences in NDW across children with and without language disorder produced in narrative language samples (Hewitt et al., 2005; Mills, 2015). Kapantzoglou et al.

(2017) reported that lexical diversity was a strong indicator of underlying language ability of DLLs when elicited via a story retell (but not a story tell) in children's home language. Our results converge with these prior findings, while also suggesting that lexical diversity may be a strong indicator of language ability in DLLs' two languages, regardless of elicitation approach.

MLU in Spanish narratives was also a consistent predictor of Spanish vocabulary and morphosyntax outcomes on norm-referenced measures, although to a lesser degree than lexical diversity. Consistent with our expectations, we did observe an interaction for the relation between MLU and Spanish morphosyntax outcomes, with a stronger predictive relation for the story retell than for the unique story tell. Children may have used working memory resources to retain and recall information presented in the story retell scenario that they were not able to draw upon during the unique story tell. Given that the morphosyntax task used in this study required children to retain sentences in memory and repeat them to the examiner, this may explain stronger links between MLU and morphosyntax in the story retell context. PGU did not consistently predict performance on norm-referenced language outcomes.

English

Like the Spanish language outcomes, results indicated that lexical diversity was generally the strongest predictor of performance on English-language norm-referenced measures. Generally, findings were consistent with our hypothesis that lexical diversity would be more strongly related to English language outcomes in the story retell context. In a previous study, NDW in English in a story retell offered significant positive associations to a norm-referenced vocabulary measure in a sample of 145 kindergarten and first-grade DLLs (Wood et al., 2018). Examining lexical diversity of English narrative language samples appears to be a good indicator of overall language ability (Bedore et al., 2010) and overall story quality (Heilmann, Miller, Nockerts, et al., 2010). MLU elicited from English language samples also appears to be a consistent indicator of language ability on norm-referenced measures in English. Percent grammaticality did not consistently predict performance on norm-referenced English language outcomes.

Limitations and Future Directions

In considering the findings from this work, contextualization is essential. Specifically, this work was conducted in school settings that centered English language use. Anecdotally, limited day-to-day support for Spanish was observed by research assistants conducting assessments in the school settings. Students being educated in settings in which both languages are supported may produce different language samples than those observed in this work.

Additionally, the participants ranged in age from 5-7 years and were assessed during the middle of either their kindergarten or first grade year. Although this approach allowed for broad examination of language sampling with strong statistical power, it is possible that subgroup analyses by age may reveal differences. As demonstrated by Castilla-Earls et al. (2019), DLLs being educated in English-dominant educational settings tend to experience a proficiency shift during the early school years. During this proficiency shift, DLLs may temporarily exhibit low grammaticality in both languages (Castilla-Earls et al., 2019). This may have contributed to the finding that there were not consistent associations between PGU and the norm-referenced measures of language in Spanish. We also acknowledge that the elicitation protocol did not include counterbalancing tasks which would have strengthened our methodology. Lack of counterbalancing may have created a practice effect which could have increased story tell outcomes.

It is also important to interpret this work as a relatively exploratory contribution to the literature. Dual language development is rich and complex, not easily distilled to single summary metrics. There is ongoing need for research to continue to evaluate the validity and reliability of assessment tools used to quantify the language abilities of bilingual children, both for diagnosis of language disorder and for general evaluation of language development. This work provides a small contribution and requires both careful contextualization and consideration of limitations in the current knowledge base regarding bilingual language development in the U.S.

Conclusions and Practical Implications

This study yielded two key conclusions that have practical implications for assessment of DLLs' language skills by school-based SLPs. First, microstructural summary indices of narrative language ability did not differ substantially across story tells and retells. Differences were more subtle and require careful consideration in clinical application. Unique story tells may be particularly useful for school-based clinicians seeking to monitor student progress, as they often require less time to collect (as the examiner does not need to spend time reading the story script to the child). Furthermore, story retell elicitation approaches provide children with a language model they can refer to when retelling the story. Consequently, individual differences in story retell performance may not reflect a pure indicator of narrative language ability, as children may be able to utilize other cognitive resources (e.g., working memory) when retelling the story. . However, narrative retells may provide students with opportunities to demonstrate more complex language skills given the linguistic model.

Second, regardless of the language of elicitation, microstructural indices derived from narrative language samples were significantly related to children's performance on norm-referenced language assessments. More specifically, lexical diversity was the strongest predictor of children's performance on norm-referenced language measures, regardless of language. This suggests some overlap in the abilities reflected by NDW compared to currently available norm-referenced measures, whereas the skills measured by MLU and PGU may be more distinct. (e.g., Bedore et al., 2010, Kapantzaglou et al., 2017). Future research should continue to consider the predictive validity of lexical diversity for differentiating students with and without language disorder. Such evidence would provide information on key skills to screen for prior to conducting lengthy diagnostic language assessment. Overall, findings from this study support the use of narrative language sampling for young DLLs as having strong validity across languages and elicitation approaches.

Acknowledgements

The research reported in this paper was supported by a New Investigators Research Grant awarded to Lisa Fitton and J. Marc Goodrich by the American Speech-Language-Hearing Foundation. Views expressed herein are those of the authors and have neither been reviewed nor approved by the granting agency.

References

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice*, *27*(3), 17–31. https://doi.org/10.1111/j.1745-3992.2008.00125.x

Anaya, J. B., Peña, E. D., & Bedore, L. M. (2018). Conceptual scoring and classification accuracy of vocabulary testing in bilingual children. *Language, Speech, and Hearing Services in Schools*, *49*(1), 85–97. https://doi.org/10.1044/2017_LSHSS-16-0081

ASHA. (2021). *Bilingual service delivery*. ASHA Practice Portal. https://www.asha.org/practice-portal/professional-issues/bilingual-service-delivery/

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism*, *11*(1), 1–29. https://doi.org/10.2167/beb392.0

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, *43*(6), 498–510. https://doi.org/10.1016/j.jcomdis.2010.05.002

Bliss, L. S., & McCabe, A. (2006). Comparison of discourse genres: Clinical implications. *Contemporary Issues in Communication Science and Disorders*, *33*(Fall), 126–167. https://doi.org/10.1044/cicsd_33_f_126

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children’s vocabulary size. *Bilingualism*, *16*(1), 32–48.

https://doi.org/10.1017/S1366728912000120

Capps, R., Newland, K., Fratzke, S., Groves, S., Auclair, G., Fix, M., & McHugh, M. (2015). Integrating refugees in the United States: The successes and challenges of resettlement in a Global Context. In *Statistical Journal of the IAOS* (Vol. 31, Issue 3, pp. 341–367). IOS Press. https://doi.org/10.3233/SJI-150918

Castilla-Earls, A., Bedore, L., Rojas, R., Fabiano-Smith, L., Pruitt-Lord, S., Restrepo, M. A., & Peña, E. (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. *American Journal of Speech-Language Pathology*, *29*, 1116–1132. https://doi.org/10.1044/2020_AJSLP-19-00179

Castilla-Earls, A., Francis, D., Iglesias, A., & Davidson, K. (2019). The impact of the Spanish-to-English proficiency shift on the grammaticality of English learners. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1739–1754. https://doi.org/10.1044/2018_JSLHR-L-18-0324

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

Cleave, P. L., Girolametto, L. E., Chen, X., & Johnson, C. J. (2010). Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders*, *43*(6), 511–522. https://doi.org/10.1016/j.jcomdis.2010.05.005

Committee on Fostering School Success for English Learners. (2017). Promoting the educational success of children and youth learning English. In R. Takanishi & S. Le Menestrel (Eds.), *Promoting the Educational Success of Children and Youth Learning English*. National Academies Press. https://doi.org/10.17226/24677

Duinmeijer, I., De Jong, J., & Scheper, A. (2012). Narrative abilities, memory and attention in children with a specific language impairment. *International Journal of Language and Communication Disorders*, *47*(5), 542–555. https://doi.org/10.1111/j.1460-6984.2012.00164.x

Ebert, K. D., & Pham, G. (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. *Language, Speech, and Hearing Services in Schools*, *48*(1), 42–55. https://doi.org/10.1044/2016_LSHSS-16-0007

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

Fitton, L., Hoge, R., Petscher, Y., & Wood, C. (2019). Psychometric evaluation of the Bilingual English–Spanish Assessment sentence repetition task for clinical decision making. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1906–1922. https://doi.org/10.1044/2019_JSLHR-L-18-0354

Gazella, J., & Stockman, I. J. (2003). Children’s story retelling under different modality and task conditions: Implications for standardizing language sampling procedures. *American Journal of Speech-Language Pathology*, *12*(1), 61–72. https://doi.org/10.1044/1058-0360(2003/053)

Goldstein, B., Fabiano, L., & Iglesias, A. (2004). Spontaneous and imitated productions in Spanish-speaking children with phonological disorders. *Language, Speech, and Hearing Services in Schools, 35*(1), 5–15. https://doi.org/10.1044/0161-1461(2004/002)

Gorman, B. K., Bingham, G. E., Fiestas, C. E., & Terry, N. P. (2016). Assessing the narrative abilities of Spanish-speaking preschool children: A Spanish adaptation of the narrative assessment protocol. *Early Childhood Research Quarterly*, *36*, 307–317. https://doi.org/10.1016/j.ecresq.2015.12.025

Govindarajan, K., & Paradis, J. (2019). Narrative abilities of bilingual children with and without Developmental Language Disorder (SLI): Differentiation and the role of age and input factors. *Journal of Communication Disorders*, *77*, 1–16. https://doi.org/10.1016/j.jcomdis.2018.10.001

Gross, M., Buac, M., & Kaushanskaya, M. (2014). Conceptual scoring of receptive and expressive vocabulary measures in simultaneous and sequential bilingual children. *American Journal of Speech-Language Pathology*, *23*(4), 574–586. https://doi.org/10.1044/2014_AJSLP-13-0026

Guo, L. Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances between 4 and 9 years of age for 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

Gutiérrez-Clellen, V. F. (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education*, *13*(2), 175–197. https://doi.org/10.1016/S0898-5898(01)00061-4

Gutiérrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E., & Anderson, R. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools*, *31*(1), 88–98. https://doi.org/10.1044/0161-1461.3101.88

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. In *Seminars in Speech and Language* (Vol. 30, Issue 4, pp. 234–245). Semin Speech Lang. https://doi.org/10.1055/s-0029-1241722

Hedges, L. V. (1981). Distributional theory for Glass's estimator of effects size and related estimators. *Journal of Educational Statistics*, *6*(2), 107. https://doi.org/10.2307/1164588

Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools*, *41*(1), 84–95. https://doi.org/10.1044/0161-1461(2009/08-0075)

Heilmann, J. J., Miller, J. F., Nockerts, A., & Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-Age children. *American Journal of Speech-Language Pathology*, *19*, 154–166. https://doi.org/10.1044/1058-0360(2009/08-0024)

Heilmann, J. J., Rojas, R., Iglesias, A., & Miller, J. F. (2016). Clinical impact of wordless picture storybooks on bilingual narrative language production: A comparison of the 'Frog' stories. *International Journal of Language & Communication Disorders*, *51*(3), 339–345. https://doi.org/10.1111/1460-6984.12201

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress*. https://doi.org/NCES 2011-459

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*(3), 197–213.

https://doi.org/10.1016/j.jcomdis.2004.10.002

Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Narrative abilities in subgroups of English language learners and monolingual peers. *International Journal of Bilingualism*, *19*(6), 677–692. https://doi.org/10.1177/1367006914534330

Holloway, K. F. C. (1986). The effects of basal readers on oral language structures: A description of complexity. *Journal of Psycholinguistic Research 1986 15:2*, *15*(2), 141–151. https://doi.org/10.1007/BF01067519

Justice, L. M., Bowles, R., Pence, K., & Gosse, C. (2010). A scalable tool for assessing children's language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly*, *25*(2), 218–234. https://doi.org/10.1016/j.ecresq.2009.11.002

Kapantzoglou, M., Fergadiotis, G., & Restrepo, M. A. (2017). Language sample analysis and elicitation technique effects in bilingual children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2852–2864. https://doi.org/10.1044/2017_JSLHR-L-16-0335

Kapantzoglou, M., Thompson, M. S., Gray, S., & Restrepo, M. A. (2016). Assessing measurement invariance for Spanish sentence repetition and morphology elicitation tasks. *Journal of Speech, Language, and Hearing Research*, *59*(2), 254–266. https://doi.org/10.1044/2015_JSLHR-L-14-0319

Klop, D., & Engelbrecht, L. (2013). The effect of two different visual presentation modalities on the narratives of mainstream grade 3 children. *The South African Journal of Communication Disorders. Die Suid-Afrikaanse Tydskrif Vir Kommunikasieafwykings*, *60*, 21–26.

https://doi.org/10.7196/sajcd.242

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. In *Journal of Communication Disorders* (Vol. 43, Issue 6, pp. 456–473). https://doi.org/10.1016/j.jcomdis.2010.02.002

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *0*(NOV), 863. https://doi.org/10.3389/FPSYG.2013.00863

Language and Reading Research Consortium (LARRC), Yeomans-Maldonado, G., Bengochea, A., & Mesa, C. (2018). The dimensionality of oral language in kindergarten Spanish–English dual language learners. *Journal of Speech, Language, and Hearing Research*, *61*(11), 2779–2795. https://doi.org/10.1044/2018_JSLHR-L-17-0320

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 and Hearing Research*, *38*(2), 415–425. https://doi.org/10.1044/jshr.3802.415

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research*, *60*(8), 2185–2198. https://doi.org/10.1044/2017_JSLHR-L-15-0402

Lucero, A. (2018). Oral narrative retelling among emergent bilinguals in a dual language immersion program. *International Journal of Bilingual Education and Bilingualism*, *21*(2), 248–264. https://doi.org/10.1080/13670050.2016.1165181

Lucero, A., & Uchikoshi, Y. (2019). Narrative assessments with first grade Spanish-English emergent bilinguals: Spontaneous versus retell conditions. *Narrative Inquiry*, *29*(1), 137–

156. https://doi.org/10.1075/NI.18015.LUC

Martin, N. A. (2013). *Expressive One-Word Picture Vocabulary Test-4: Spanish Bilingual Edition*. Academic Therapy Publications.

Méndez, L. I., Perry, J., Holt, Y., Bian, H., & Fafulas, S. (2018). Same or different: Narrative retells in bilingual Latino kindergarten children. *Bilingual Research Journal*, *41*(2), 150–166. https://doi.org/10.1080/15235882.2018.1456984

Miles, S., Chapman, R., & Sindberg, H. (2006). Sampling context affects MLU in the language of adolescents with Down syndrome. *Journal of Speech, Language, and Hearing Research*, *49*(2), 325–337. https://doi.org/10.1044/1092-4388(2006/026)

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

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research and Practice*, *21*(1), 30–43. https://doi.org/10.1111/j.1540-5826.2006.00205.x

Miller, J. F., & Iglesias, A. (2017). *Systematic Analysis of Language Transcripts (SALT)* (Research Version 18). SALT Software, LLC.

Mills, M. T. (2015). The effects of visual stimuli on the spoken narrative performance of school-age African American children. *Language, Speech, and Hearing Services in Schools*, *46*(4), 337–351. https://doi.org/10.1044/2015_LSHSS-14-0070

Orizaba, L., Gorman, B. K., Fiestas, C. E., Bingham, G. E., & Terry, N. P. (2020). Examination of narrative language at microstructural and macrostructural levels in spanish-speaking preschoolers. *Language, Speech, and Hearing Services in Schools*, *51*(2), 428–440.

https://doi.org/10.1044/2019_LSHSS-19-00103

Otwinowska, A., Mieszkowska, K., Białecka-Pikul, M., Opacki, M., & Haman, E. (2018). Retelling a model story improves the narratives of Polish-English bilingual children. *Https://Doi.Org/10.1080/13670050.2018.1434124*, *23*(9), 1083–1107. https://doi.org/10.1080/13670050.2018.1434124

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

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B., & Bedore, L. M. (2014). *Bilingual English-Spanish Assessment (BESA)*. AR-Clinical publications.

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language and Communication Disorders*, *50*(1), 106–118. https://doi.org/10.1111/1460-6984.12126

Pratt, A. S., Peña, E. D., & Bedore, L. M. (2020). Sentence repetition with bilinguals with and without DLD: Differential effects of memory, vocabulary, and exposure. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/s1366728920000498

Quiroz, B. G., Snow, C. E., & Zhao, J. (2010). Vocabulary skills of Spanish—English bilinguals: impact of mother—child language interactions and home language and literacy support. *International Journal of Bilingualism*, *14*(4), 379–399. https://doi.org/10.1177/1367006910370919

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1398–1411.

https://doi.org/10.1044/jslhr.4106.1398

Rojas, R., & Iglesia, A. (2009). Making a case for language sampling: Assessment and intervention with (Spanish-English) second-language learners. In *ASHA Leader* (Vol. 14, Issue 3). American Speech-Language-Hearing Association. https://doi.org/10.1044/leader.ftr1.14032009.10

Rojas, R., & Iglesias, A. (2013). The language growth of Spanish-speaking English language learners. *Child Development*, *84*(2), 630–646. https://doi.org/10.1111/j.1467-8624.2012.01871.x

Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children*, *8*(7), 578. https://doi.org/10.3390/CHILDREN8070578

Schneider, P., & Dubé, R. V. (2005). Story presentation effects on children's retell content. *American Journal of Speech-Language Pathology*, *14*(1), 52–60. https://doi.org/10.1044/1058-0360(2005/007)

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.

Sheng, L., Shi, H., Wang, D., Hao, Y., & Zheng, L. (2020). Narrative production in mandarin-speaking children: Effects of language ability and elicitation method. *Journal of Speech, Language, and Hearing Research*, *63*(3), 774–792. https://doi.org/10.1044/2019_JSLHR-19-00087

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/PITS.20234

Westerveld, M. F., & Gillon, G. T. (2010). Profiling oral narrative ability in young school-aged children. *International Journal of Speech-Language Pathology*, *12*(3), 178–189. https://doi.org/10.3109/17549500903194125

Wood, C., Wofford, M. C., & Schatschneider, C. (2018). Relationship between performance on oral narrative retells and vocabulary assessments for Spanish-English speaking children. *Communication Disorders Quarterly*, *39*(3), 402–414. https://doi.org/10.1177/1525740117722507

Table 1

Spanish: Means, standard deviations, and correlations

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Age (years)	6.28	0.68									
2. TNU	24.72	13.02	.02								
3. NDW	53.77	24.40	.07	.85**							
4. MLU (words)	5.96	1.50	.09	.51**	.66**						
5. PGU	0.68	0.19	-.03	-.06	.09	-.06					
6. English Vocab	97.15	17.80	.17	-.10	-.08	.13	-.30**				
7. Spanish Vocab	83.43	17.31	-.22*	.36**	.53**	.46**	.17	-.11			
8. Conceptual Vocab	102.94	15.58	.01	.11	.30**	.38**	-.02	.65**	.54**		
9. English SR	91.69	18.00	.23*	-.15	-.13	.17	-.19	.63**	-.28*	.32**	
10. Spanish SR	90.56	15.34	-.12	.22	.46**	.38**	.34**	-.17	.64**	.43**	.07

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. TNU = total number of utterances. NDW = number of different utterances. MLU = mean length of utterance in words. PGU = percent grammatical utterances. SR = Sentence repetition subtest of the Bilingual English-Spanish Assessment (Peña et al., 2014). All standardized assessment scores are norm referenced. The total sample size for participants who completed the Spanish narratives was $n = 96$.
* indicates $p < .05$. ** indicates $p < .01$.

Table 2

English: Means, standard deviations, and correlations

Variable	***M***	***SD***	1	2	3	4	5	6	7	8	9
1. Child Age (years)	6.32	0.69									
2. TNU	24.02	15.11	.36**								
3. NDW	54.45	30.43	.49**	.84**							
4. MLU (words)	6.43	1.96	.48**	.49**	.74**						
5. PGU	0.69	0.26	.37**	.24*	.28**	.12					
6. English Vocab	95.77	19.71	.11	.38**	.57**	.55**	.11				
7. Spanish Vocab	79.12	18.24	-.18	-.06	-.15	-.08	-.14	-.11			
8. Conceptual Vocab	103.14	15.98	-.08	.26*	.37**	.31**	.05	.76**	.39**		
9. English SR	92.88	17.00	.13	.34**	.53**	.59**	.19	.65**	-.19	.43**	
10. Spanish SR	87.46	16.69	-.18	.02	.05	.12	.08	-.07	.65**	.30**	.18

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. TNU = total number of utterances. NDW = number of different utterances. MLU = mean length of utterance in words. PGU = percent grammatical utterances. SR = Sentence repetition subtest of the Bilingual English-Spanish Assessment (Peña et al., 2014). All standardized assessment scores are norm referenced. The total sample size for participants who completed the English narratives was $n = 108$.
* indicates $p < .05$. ** indicates $p < .01$.

Table 3

MLU in Spanish Narratives Predicting Language Measures (raw scores)

	Spanish Sentence Repetition						**Spanish Vocabulary**					
	Unique Story			**Story Retell**			**Unique Story**			**Story Retell**		
Predictors	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*
Intercept	14.59	5.75 – 23.42	**.001**	-2.04	-13.36–9.28	.724	10.26	-6.41 – 26.93	.228	-9.15	-26.78–8.47	.309
MLU[1]	**1.84**	**0.32 – 3.36**	**.018**	**3.19**	**1.29 – 5.09**	**.001**	**5.29**	**2.18 – 8.40**	**.001**	**5.41**	**2.21 – 8.60**	**.001**
Age[2]	0.52	-2.41 – 3.45	.726	0.41	-3.52 – 4.33	.839	-1.40	-7.25 – 4.45	.640	0.34	-5.84 – 6.53	.914
TNU[3]	-0.07	-0.25 – 0.11	.430	0.12	-0.07 – 0.32	.218	-0.26	-0.63 – 0.12	.180	**0.42**	**0.09 – 0.75**	**.014**
Random Effects												
σ^2	43.00			59.38			185.79			183.52		
τ_{00}	10.79 Site			9.24 Site			17.73 Site					
ICC	0.20			0.13			0.09					
N	2 Site			2 Site			2 Site			2 Site		
Observations	43			45			45			48		
Marginal R^2	0.116			0.287			0.213			0.370		
Conditional R^2	0.294			0.383			0.281			NA		

[1]Mean Length of Utterance
[2]Centered at 6 years
[3]Total Number of Utterances

Table 4

NDW in Spanish Narratives Predicting Language Measures (raw scores)

	Spanish Sentence Repetition						**Spanish Vocabulary**					
	Unique Story			**Story Retell**			**Unique Story**			**Story Retell**		
Predictors	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*
Intercept	17.16	11.98 – 22.34	<.001	7.89	1.99 – 13.79	.009	20.99	11.62 – 30.37	<.001	9.19	0.59 – 17.79	.036
NDW[1]	**0.33**	**0.18 – 0.48**	**<.001**	**0.46**	**0.32 – 0.61**	**<.001**	**0.75**	**0.43 – 1.07**	**<.001**	**0.64**	**0.36 – 0.92**	**<.001**
Age[2]	0.38	-2.19 – 2.95	.770	0.34	-2.74 – 3.42	.829	-1.96	-7.27 – 3.35	.470	1.40	-4.19 – 6.99	.623
TNU[3]	**-0.42**	**-0.67 – -0.17**	**.001**	**-0.59**	**-0.89 – -0.29**	**<.001**	**-0.96**	**-1.50 – -0.42**	**.001**	-0.54	-1.12 – 0.05	.071
Random Effects												
σ^2	33.35			39.01			157.81			158.01		
τ_{00}	3.92 $_{Site}$			8.16 $_{Site}$			0.00 $_{Site}$			0.00 $_{Site}$		
ICC	0.11			0.17								
N	2 $_{Site}$			2 $_{Site}$			2 $_{Site}$			2 $_{Site}$		
Observations	43			45			45			48		
Marginal R^2	0.301			0.495			0.348			0.471		
Conditional R^2	0.375			0.582			NA			NA		

[1]Number of Different Words
[2]Centered at 6 years
[3]Total Number of Utterances

Table 5

PGU in Spanish Narratives Predicting Language Measures (raw scores)

	Spanish Sentence Repetition						**Spanish Vocabulary**					
	Unique Story			**Story Retell**			**Unique Story**			**Story Retell**		
Predictors	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*
Intercept	9.34	0.71 – 17.97	.034	2.52	-8.62 – 13.66	.657	17.42	-1.11 – 35.94	0.065	10.67	-7.93 – 29.27	.261
PGU[1]	**0.16**	**0.06 – 0.26**	**.001**	**0.17**	**0.02 – 0.32**	**.024**	0.19	-0.02 – 0.41	0.076	0.10	-0.15 – 0.35	.439
Age[2]	0.81	-1.96 – 3.58	.567	3.32	-0.61 – 7.25	.097	-1.10	-7.37 – 5.18	0.734	3.48	-3.21 – 10.16	.308
TNU[3]	0.10	-0.04 – 0.24	.151	0.19	-0.01 – 0.39	.057	0.18	-0.14 – 0.51	0.272	**0.59**	**0.23 – 0.94**	**.001**
Random Effects												
σ^2	40.45			67.20			220.84			226.41		
τ_{00}	0.26 $_{Site}$			2.84 $_{Site}$			0.00 $_{Site}$			0.00 $_{Site}$		
ICC	0.01			0.04								
N	2 $_{Site}$			2 $_{Site}$			2 $_{Site}$			2 $_{Site}$		
Observations	43			45			45			48		
Marginal R^2	0.216			0.223			0.082			0.231		
Conditional R^2	0.221			0.254			NA			NA		

[1]Percent Grammatical Utterances
[2]Centered at 6 years
[3]Total Number of Utterances

Table 6

MLU in English Narratives Predicting Language Measures (raw scores)

	English Sentence Repetition						**English Vocabulary**					
	Unique Story			**Story Retell**			**Unique Story**			**Story Retell**		
Predictors	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*
Intercept	10.64	1.11 – 20.16	.029	5.17	-1.60 – 11.93	.135	22.96	3.72 – 42.20	.019	5.07	-10.34 – 20.47	.519
MLU[1]	**2.19**	**0.48 – 3.90**	**.012**	**2.88**	**1.69 – 4.06**	**<.001**	**4.86**	**1.17 – 8.55**	**.010**	**5.06**	**2.37 – 7.75**	**<.001**
Age[2]	-0.77	-4.26 – 2.72	.666	1.14	-2.01 – 4.29	.480	4.04	-3.64 – 11.73	.302	4.79	-2.20 – 11.77	.179
TNU[3]	0.03	-0.14 – 0.20	.751	0.04	-0.09 – 0.18	.515	-0.06	-0.44 – 0.32	.756	**0.38**	**0.06 – 0.71**	**.021**
Random Effects												
σ^2	52.21			42.08			249.32			223.54		
τ_{00}	7.08 Site			0.34 Site			5.36 Site			0.00 Site		
ICC	0.12			0.01			0.02					
N	2 Site			2 Site			2 Site			2 Site		
Observations	40			53			39			55		
Marginal R^2	0.206			0.454			0.266			0.459		
Conditional R^2	0.301			0.458			0.281			NA		

[1]Mean Length of Utterance
[2]Centered at 6 years
[3]Total Number of Utterances

Table 7

NDW in English Narratives Predicting Language Measures (raw scores)

	English Sentence Repetition						**English Vocabulary**					
	Unique Story			**Story Retell**			**Unique Story**			**Story Retell**		
Predictors	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*
Intercept	18.50	11.93 – 25.07	<.001	13.95	10.04 – 17.85	<.001	37.47	26.12 – 48.83	<.001	18.79	11.19 – 26.40	<.001
NDW[1]	0.14	-0.00 – 0.28	.052	**0.36**	**0.25 – 0.48**	**<.001**	**0.48**	**0.19 – 0.77**	**.001**	**0.77**	**0.52 – 1.03**	**<.001**
Age[2]	-1.04	-4.77 – 2.70	.587	1.84	-0.83 – 4.51	.177	1.35	-6.33 – 9.04	.730	4.74	-0.99 – 10.47	.105
TNU[3]	-0.05	-0.29 – 0.20	.699	**-0.45**	**-0.67 – -0.23**	**<.001**	-0.44	-0.93 – 0.05	.082	**-0.72**	**-1.22 – -0.23**	**.004**
Random Effects												
σ^2	54.96			34.22			223.32			167.38		
τ_{00}	11.01 $_{\text{Site}}$			1.86 $_{\text{Site}}$			17.87 $_{\text{Site}}$			0.00 $_{\text{Site}}$		
ICC	0.17			0.05			0.07					
N	2 $_{\text{Site}}$			2 $_{\text{Site}}$			2 $_{\text{Site}}$			2 $_{\text{Site}}$		
Observations	40			53			39			55		
Marginal R^2	0.149			0.532			0.303			0.592		
Conditional R^2	0.291			0.556			0.354			NA		

[1]Number of Different Words
[2]Centered at 6 years
[3]Total Number of Utterances

Table 8

PGU in English Narratives Predicting Language Measures (raw scores)

	English Sentence Repetition						**English Vocabulary**					
	Unique Story			**Story Retell**			**Unique Story**			**Story Retell**		
Predictors	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*	*Est.*	*Conf. Int (95%)*	*p*
Intercept	8.49	-0.28–17.27	.058	18.63	11.72 – 25.54	<.001	32.50	13.58 – 51.42	.001	28.20	14.60 – 41.81	<.001
PGU[1]	**0.21**	**0.13 – 0.29**	**<.001**	0.02	-0.07 – 0.10	.703	0.21	-0.02 – 0.44	.068	0.03	-0.14 – 0.21	.701
Age[2]	-2.04	-4.96–0.88	.171	**4.35**	**0.90 – 7.80**	**.014**	3.68	-4.43 – 11.80	.374	**10.55**	**3.44 – 17.67**	**.004**
TNU[3]	0.12	-0.00 – 0.24	.057	0.13	-0.02 – 0.29	.092	0.18	-0.15 – 0.51	.285	**0.52**	**0.16 – 0.88**	**.004**
Random Effects												
σ^2	35.39			60.19			261.49			282.41		
τ_{00}	22.06 Site			3.85 Site			42.87 Site			0.00 Site		
ICC	0.38			0.06			0.14					
N	2 Site			2 Site			2 Site			2 Site		
Observations	40			53			39			55		
Marginal R^2	0.327			0.200			0.168			0.322		
Conditional R^2	0.585			0.248			0.285			NA		

[1]Percent Grammatical Utterances
[2]Centered at 6 years
[3]Total Number of Utterances

Table S1

Descriptive Information for the Sample by Language and Elicitation Technique

Descriptive Statistics for Children Completing Narratives in Spanish (*n* = 96)

	Spanish: Unique Story			Spanish: Story Retell		
	M	*SD*	*Min - Max*	*M*	*SD*	*Min - Max*
Age	6.31	0.70	5.17 - 7.37	6.26	0.66	5.25 - 7.83
TNU - Spanish	26.02	13.79	2 - 69	23.57	12.32	3 - 59
NDW - Spanish	52.56	23.13	3 - 94	54.84	25.66	12 - 149
MLU - Spanish	6.14	1.69	2.00 - 9.10	5.80	1.31	2.44 - 8.15
PGU - Spanish	0.67	0.21	0.13 - 1.00	0.69	0.17	0.29 - 1.00
English Vocab	95.64	20.05	55 - 135	98.72	15.18	55 - 126
Spanish Vocab	85.69	16.40	55 - 129	81.40	18.02	55 - 118
Conceptual Vocab	103.57	16.52	66 - 136	102.35	14.78	62 - 126
English SR	93.11	17.65	60 - 115	90.27	18.48	55 - 115
Spanish SR	94.59	12.93	70 - 120	87.09	16.52	55 - 115

Descriptive Statistics for Children Completing Narratives in English (*n* = 108)

	English: Unique Story			English: Story Retell		
	M	*SD*	*Min - Max*	*M*	*SD*	*Min - Max*
Age	6.35	0.68	5.17 - 7.42	6.28	0.72	5.25 - 7.83
TNU - English	23.43	14.42	1 - 77	24.84	16.15	29983.00
NDW - English	56.19	30.78	1 - 126	52.02	30.10	3 - 134
MLU - English	5.98	1.86	1 - 8.82	6.06	1.73	1.50 - 9.03
PGU - English	0.71	0.27	0 - 1.00	0.67	0.26	0.20 - 1.00
English Vocab	96.04	17.50	55 - 135	102.45	16.05	67 - 145
Spanish Vocab	81.98	16.98	55 - 129	77.00	16.91	55 - 111
Conceptual Vocab	101.53	16.09	64 - 136	105.82	14.64	74 - 145
English SR	93.78	15.92	60 - 115	93.14	17.49	60 - 115
Spanish SR	90.83	16.45	55 - 120	83.14	15.77	55 - 115

Note. M and *SD* are used to represent mean and standard deviation, respectively. TNU = total number of utterances. NDW = number of different utterances. MLU = mean length of utterance in words. PGU = proportion grammatical utterances. SR = Sentence repetition subtest of the Bilingual English-Spanish Assessment (Peña et al., 2014). All standardized assessment scores are norm referenced, but note that scores were not computed for children outside the normative age range.

Table S2

Unstandardized Differences by Elicitation Technique in Spanish: Controlling for Age & TNU

	Mean Length of Utterance			**Number of Different Words**			**Proportion Grammatical Utterances**		
Predictors	*Est.*	*Conf. Int (95%)*	*P-Value*	*Est.*	*Conf. Int (95%)*	*P-Value*	*Est.*	*Conf. Int (95%)*	*P-Value*
Intercept	2.93	0.39 – 5.47	0.024	2.91	-19.58 – 25.41	0.800	0.75	0.38 – 1.12	<0.001
Age in Years	0.24	-0.15 – 0.62	0.228	2.41	-1.08 – 5.90	0.177	-0.01	-0.06 – 0.05	0.813
Total Number of Utterances	**0.06**	**0.04 – 0.08**	**<0.001**	**1.53**	**1.34 – 1.71**	**<0.001**	-0.00	-0.00 – 0.00	0.604
Elicitation (Unique Story)	0.16	-0.35 – 0.68	0.539	**-5.31**	**-10.03 – -0.59**	**0.027**	-0.02	-0.10 – 0.06	0.647
Random Effects									
σ^2	1.63			137.29			0.04		
τ_{00}	0.16 $_{Site}$			0.59 $_{Site}$					
ICC	0.09			0.001					
N	2 $_{Site}$			2 $_{Site}$			2 $_{Site}$		
Observations	96			96			96		
Marginal R^2 / Conditional R^2	0.247 / 0.315			0.741 / 0.742			0.006 / NA		

Table S3

Unstandardized Differences by Elicitation Technique in English: Controlling for Age & TNU

	Mean Length of Utterance			**Number of Different Words**			**Proportion Grammatical Utterances**		
Predictors	*Est.*	*CI (95%)*	*P-Value*	*Est.*	*CI (95%)*	*P-Value*	*Est.*	*CI (95%)*	*P-Value*
Intercept	-0.13	-2.75 – 2.49	0.922	-37.22	-64.31 – -10.14	0.007	-0.23	-0.69 – 0.22	0.315
Age in Years	**0.83**	**0.41 – 1.25**	**<0.001**	**8.96**	**4.52 – 13.41**	**<0.001**	**0.14**	**0.07 – 0.21**	**<0.001**
Total Number of Utterances	**0.04**	**0.02 – 0.06**	**<0.001**	**1.56**	**1.35 – 1.76**	**<0.001**	0.00	-0.00 – 0.01	0.234
Elicitation (Unique Story)	-0.01	-0.56 – 0.55	0.984	-5.79	-11.60 – 0.01	0.051	-0.02	-0.11 – 0.07	0.722
Random Effects									
σ^2	2.04			228.65			0.05		
τ_{00}	0.09 $_{Site}$						0.01 $_{Site}$		
ICC	0.04						0.19		
N	2 $_{Site}$			2 $_{Site}$			2 $_{Site}$		
Observations	108			108			108		
Marginal R^2 / Conditional R^2	0.316 / 0.344			0.755 / NA			0.150 / 0.312		

Table S4

Estimates from Z-Scored LSA Predictors and Language Outcome Measures

		Spanish Measures			
Parallel Table	LSA Measure	Sentence Repetition		Vocabulary	
		Est.	95% CI	*Est.*	95% CI
Table 3	1. MLU – Unique	0.33*	0.06 – 0.60	0.50*	0.21 – 0.79
	2. MLU – Retell	0.57*	0.23 – 0.91	0.51*	0.21 – 0.81
Table 4	3. NDW – Unique	0.96*	0.53 – 1.40	1.15*	0.66 – 1.65
	4. NDW – Retell	1.35*	0.92 – 1.78	0.98*	0.55 – 1.42
Table 5	5. PGU – Unique	0.37*	0.15 – 0.60	0.23	-0.02 – 0.49
	6. PGU – Retell	0.38*	0.05 – 0.72	0.12	-0.18 – 0.42

		English Measures			
Parallel Table	LSA Measure	Sentence Repetition		Vocabulary	
		Est.	95% CI	*Est.*	95% CI
Table 6	7. MLU – Unique	0.45*	0.10 – 0.80	0.43*	0.10 – 0.76
	8. MLU – Retell	0.59*	0.34 – 0.83	0.45*	0.21 – 0.69
Table 7	9. NDW – Unique	0.50	-0.01 – 1.01	0.74*	0.29 – 1.19
	10. NDW – Retell	1.30*	0.89 – 1.72	1.20*	0.81 – 1.60
Table 8	11. PGU – Unique	0.64*	0.39 – 0.89	0.29	-0.02 – 0.59
	12. PGU – Retell	0.05	-0.21 – 0.31	0.05	-0.19 – 0.28

*Denotes $p < .05$. Specific p-values are provided in Tables 1-6.