ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

# Investigating Narrative Performance in Children With Developmental Language Disorder: A Systematic Review and Meta-Analysis

Katherine L. Winters,[a] Javier Jasso,[a,b] James E. Pustejovsky,[c] and Courtney T. Byrd[a]

[a]The University of Texas at Austin [b]Widener University, Chester, PA [c]University of Wisconsin–Madison

ARTICLE INFO

Article History:
Received January 11, 2022
Revision received April 26, 2022
Accepted June 22, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Sarah Elizabeth Wallace

https://doi.org/10.1044/2022_JSLHR-22-00017

ABSTRACT

**Purpose:** Narrative assessment is one potentially underutilized and inconsistent method speech-language pathologists may use when considering a diagnosis of developmental language disorder (DLD). However, narration research encompasses many varied methodologies. This systematic review and meta-analysis aimed to (a) investigate how various narrative assessment types (e.g., macrostructure, microstructure, and internal state language) differentiate children with typical development (TD) from children with DLD, (b) identify specific narrative assessment measures that result in greater group differences, and (c) evaluate participant and sample characteristics that may influence performance differences.
**Method:** Electronic databases (PsycINFO, ERIC, and PubMed) and ASHAWire were searched on July 30, 2019, to locate studies that reported oral narrative language measures for both DLD and TD groups between ages 4 and 12 years; studies focusing on written narration or other developmental disorders only were excluded. We extracted data related to sample participants, narrative task(s) and assessment measures, and research design. Group differences were quantified using standardized mean differences. Analyses used mixed-effects meta-regression with robust variance estimation to account for effect size dependencies.
**Results:** Searches identified 37 eligible studies published between 1987 and 2019, including 382 effect sizes. Overall meta-analysis showed that children with DLD had decreased narrative performance relative to TD peers, with an overall average effect of −0.82 *SD*, 95% confidence interval [−0.99, −0.66]. Effect sizes showed significant heterogeneity both between and within studies, even after accounting for effect size–, sample-, and study-level predictors. Across model specifications, grammatical accuracy (microstructure) and story grammar (macrostructure) yielded the most consistent evidence of TD–DLD group differences.
**Conclusions:** Present findings suggest some narrative assessment measures yield significantly different performance between children with and without DLD. However, researchers need to improve consistency of inclusionary criteria, descriptions of sample characteristics, and reporting of correlations between measures to determine which assessment measures reliably distinguish between groups.
**Supplemental Material:** https://doi.org/10.23641/asha.21200380

Developmental language disorder (DLD) is a common neurodevelopmental communication disorder characterized by significant deficits in language learning, comprehension, and expression that impacts approximately 7.5% of children, or two school-age students per every classroom of 30 (Norbury et al., 2016; Tomblin et al., 1997). Children with DLD may demonstrate a variety of impairments across the domains of phonology, morphology/syntax, semantics, and pragmatics that persist into adulthood (Dollaghan, 2004; Dollaghan & Horner, 2011; Goffman & Leonard, 2000; McGregor et al., 2020; Rudolph et al., 2019). Despite the prevalence and far-reaching effects of DLD, there is relatively low awareness of the disorder and

Correspondence to Katherine L. Winters: katiewinters@utexas.edu.
***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

children are often not identified for speech-language services (McGregor, 2020; Wittke & Spaulding, 2018). In addition, at least two critical issues limit generalization of research findings to present understanding of the disorder: (a) evolving disorder criteria and terminology (Bishop et al., 2016, 2017) and (b) inconsistently applied and substandard inclusionary criteria for DLD in research studies (Nitido & Plante, 2020). This study seeks to improve understanding of DLD by synthesizing previous research specific to narration, one potentially inconsistent and underutilized measure of language assessment (e.g., Kemp & Klee, 1997; Pavelko et al., 2016), in order to evaluate its potential contribution to the diagnostic process.

Narrative assessment, from a broad-based perspective, includes analysis of macrostructure (e.g., use of story grammar), microstructure (e.g., syntactic complexity), and internal state language (e.g., characters' thoughts or feelings). Narratives effectively elicit discourse-level language including syntactically complex utterances with causal and temporal details, often from a known context, and require the speaker to plan and share their story in real time (Channell et al., 2018; Southwood & Russell, 2004; Stein & Glenn, 1979; Westby, 1984; Westerveld & Moran, 2013). Thus, narrative performance has been recommended as a tool for diagnosis and as a method for determining intervention goals and classroom recommendations (Gallagher & Hoover, 2020; Newman & McGregor, 2006; Pico et al., 2021). Furthermore, although language learning contexts differ globally, the use of narrative assessment likely extends beyond monolingual White Mainstream English–speaking children (commonly referred to in the field as "Mainstream American English"), as demonstrated by both assessment of dialect-neutral narrative elements (e.g., use of mental state verbs such as "wanted" or "thought" and conjunctions that tie sentences together such as "while" and "when"; Burns et al., 2012) and data with multilingual and non-English speakers (Gagarina et al., 2012; Govindarajan & Paradis, 2019; Pesco & Kay-Raining Bird, 2016).

Over the nearly 30 years of research completed to date, researchers have reported DLD performance in narrative tasks using various methodologies and outcome measures. This diversity makes it difficult to determine optimal narrative elicitation procedures and analyses for speech-language pathologists (SLPs) to use when assessing narrative performance of a child with suspected DLD. In fact, a majority of SLPs reportedly do not collect a narrative sample or use language sample analysis, in part due to a lack of time and lack of training on how to complete this analysis (Pavelko et al., 2016). Promisingly, recent research suggests SLPs can measure elements of narrative microstructure (e.g., number of different words [NDW] per minute, mean length of utterance [MLU], and percent grammatical utterances [PGU]) in language samples of 3–7 min (Wilder & Redmond, 2022). Furthermore, SLPs may elicit these skills through structured protocols, including the Edmonton Narrative Norms Instrument (ENNI; see Guo et al., 2019) and the Test of Narrative Language–Second Edition (TNL-2; see Magimairaj et al., 2022). The purpose of this study was to meta-analyze the available evidence from primary research studies, in order to elucidate narrative assessment methods and analyses that differentiate children with DLD from TD peers for both clinical SLPs and researchers.

## Narratives of Children With DLD

Previous research has reported narrative deficits in DLD across measures of macrostructure, microstructure, and internal state language. As a group, children with DLD produce narratives that are less syntactically complex (i.e., microstructure). Specifically, they produce narratives of reduced length (e.g., lower MLU), lexical diversity (e.g., fewer NDW), grammatical accuracy (e.g., increased number of syntactic errors), and fluency (e.g., more mazes) compared to the narratives of TD peers (e.g., Govindarajan & Paradis, 2019). Children with DLD also produce narratives with less sophisticated story grammar and referencing (i.e., macrostructure; Fichman & Altman, 2019; Hao et al., 2018; Norbury & Bishop, 2003; Peña et al., 2006; Reilly et al., 2004). In addition, there is some evidence that narratives of children with DLD also contain less internal state language (Altman et al., 2016; Boerma et al., 2016; Burns et al., 2012).

### The Role of Assessment Type

Although weaknesses in narrative macrostructure, microstructure, and internal state language have been reported in multiple studies, there are mixed findings regarding which aspects of narrative performance are impaired. For example, in a sample of 5-year-old English–Hebrew bilingual children, Altman et al. (2016) reported distinct differences in narrative microstructure but similar performance in narrative macrostructure and internal state language between children with DLD and TD peers. In contrast, Duinmeijer et al. (2012) observed deficits in both microstructure and macrostructure in monolingual Dutch-speaking children ages 6–9 years. Using a sample of monolingual English-speaking school-age children in the United States and Canada, Colozzo et al. (2011) provided evidence that children with language deficits may present with different patterns of deficit, such as relative strengths in narrative content (macrostructure) and relative weaknesses in narrative form (microstructure).

### The Role of Participant Characteristics

In addition to heterogeneity in the presentation of DLD within and across previous studies, researchers also

note that individual participant characteristics (i.e., sample-level variation) such as age and dialect or language background may uniquely influence narrative performance. For example, Guo and Schneider (2016) reported percent grammatical C-units yielded greater diagnostic accuracy for 6-year-old children compared to 8-year-old children, indicating smaller TD–DLD group differences as children age. Similarly, both age and monolingual language experience appear to influence TD–DLD differences in lexical–semantic performance across narrative and standardized language tasks (Charest et al., 2020; Jasso et al., 2020). Given these relevant participant characteristics, the present meta-analysis assessed differences in narrative performance across narrative assessment types and within assessment type while accounting for participant characteristics.

### The Role of Study Characteristics

Certain study-level characteristics may also influence differences in narrative performance between children with DLD and their TD peers, such as sample size, participant recruitment methods, and the narrative task implemented. For example, primary research studies with clinical populations often have smaller sample sizes and may be underpowered, which may influence the ability to draw accurate inferences about group differences (Gaeta & Brydges, 2020). In addition, to our knowledge, no study on narrative performance has specifically investigated the effect of recruitment method (i.e., population or community sample vs. referral from an SLP) or inclusionary criteria on DLD group narrative performance. However, previous work has shown many children who meet DLD or specific language impairment (SLI) criteria for research studies are not receiving speech-language intervention (Wittke & Spaulding, 2018). Furthermore, Dollaghan and Horner (2011) have reported spectrum bias, or diagnostic accuracy that does not reflect the full spectrum of clinical diagnosis, in research with DLD. Thus, it is possible that comparisons between children with DLD recruited via a community sample and TD peers may not yield the same difference in performance as comparisons between children with DLD who are referred by an SLP and TD peers.

The specific narrative task implemented by researchers may also influence a child with DLD's performance. For example, narrative format (i.e., tell vs. retell) and level of support provided (e.g., use of visual aids) have been shown to influence the sophistication of a child's narration (Hayward, 2003; Peña et al., 2006; Schneider & Dubé, 2005). Finally, there is variability between and within studies as to how to calculate commonly used language analyses such as MLU (Duinmeijer et al., 2012; Thordardottir et al., 2011). Thus, this meta-analysis sought to evaluate if these distinct methodological characteristics influenced differences in performance between children with DLD and their TD peers.

## This Study

Due to the variability across studies with respect to participant characteristics (e.g., age), narrative characteristics (e.g., narrative task and measure type), and study characteristics (e.g., DLD inclusionary criteria), it is not surprising that there is no universally adopted narrative assessment and subsequent analysis that yields high true-positive identification of DLD and TD status. Presently, SLPs and researchers have many methods to consider when evaluating narration but do not have a synthesis of this primary research to inform clinical practice. Fortunately, recent advances in statistical methods such as multilevel meta-analysis and robust variance estimation (RVE) allow researchers to effectively summarize effect sizes across studies while accounting for both multiple narrative measures and multiple samples within primary research studies, both of which lead to dependence in effect size estimates. These methods have been adopted in a few previous meta-analyses in communication disorders research (Hopkins-Rossabi et al., 2019; Lancaster et al., 2020; Sandbank et al., 2020). This study provides further insight into the statistical, clinical, and practical significance of this methodological approach as well as a much-needed pathway for understanding differences in outcomes across the available studies that have analyzed narration in children with DLD.

To our knowledge, no meta-analysis has previously synthesized the available evidence regarding differences in narrative performance between TD children and children with DLD, nor has any study specifically investigated the influence of DLD recruitment and inclusionary criteria. Therefore, the purpose of this study was to conduct a systematic review and quantitative meta-analysis in order to effectively summarize the research findings of the past three decades. We asked the following research questions.

1. How large are TD–DLD differences across narrative assessment type (macrostructure, microstructure, and internal state language)?
2. Are TD–DLD differences in macrostructure moderated by narrative assessment subtype (e.g., story grammar)?
3. Are TD–DLD differences in microstructure moderated by narrative assessment subtype (e.g., lexical diversity)?
4. Across narrative assessment types, are TD–DLD differences moderated by DLD recruitment method or inclusionary criteria?
5. Across narrative assessment types, are TD–DLD differences moderated by task type (e.g., tell vs. retell) or level of support (e.g., verbal cues)?
6. Across narrative assessment types, which common narrative assessment measures (e.g., MLU) yield large TD–DLD differences?

## Method

The procedures for this systematic review and meta-analysis were in accordance with relevant best practice guidelines as outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher et al., 2009; Page et al., 2020). The supplemental materials include a completed PRISMA checklist. The review was not preregistered.

### Study Selection

#### Inclusion Criteria

We included studies in this synthesis if they reported the following: (a) oral narrative language performance for a group of children with DLD, specific/primary language impairment, or language impairment (i.e., DLD group) and a group of age-matched peers with typical language (i.e., TD group); (b) group membership assignment (DLD or TD) determined through standardized assessment, non-standardized assessment, and/or clinical judgment by a certified SLP; (c) participant age range between 4 and 12 years (or, if age was not reported, grades preschool to sixth grade); and (d) sufficient quantitative results (e.g., sample sizes, means, standard deviations, and relevant test statistics) to calculate one or more effect size estimates and associated sampling variances. Because of the range of clinical terminology and diagnostic criteria used across studies (Bishop et al., 2016, 2017; Nitido & Plante, 2020), we made the decision to use broad inclusion criteria for group assignment (DLD or TD) and carefully document differences in terminology and DLD inclusionary criteria in our coding procedures, which are described later.

#### Selection Procedure

Identification of eligible studies comprised a three-step process: (a) an electronic database search, (b) an abstract screening, and (c) a full-text article review (see Figure 1; Haddaway & McGuinness, 2020). A comprehensive electronic search was conducted on July 30, 2019, using three electronic databases: PsycINFO, ERIC, and PubMed. The following search terms were used across databases: ("narrative skill*" or "narrative abilit*" or narration) AND (assess* or evaluat* or perform*) AND ("language impair*" or "language delay*" or "language disorder*" or "specific language impairment" or "developmental language disorder") NOT ("adult" or "motor" or "spell*" or "writ*" or "hear" or "read*"). To maximize search coverage, truncations, indicated by an asterisk, were used to search all word variations (i.e., word roots with different derivational or inflectional suffixes); for example, "evaluat*" searched "evaluate," "evaluation," "evaluating," and so on. The PubMed search made use of relevant Medical Subject Headings ("narration," "child," "disease"). In addition to these databases, ASHAWire was used to search all American Speech-Language-Hearing Association journals, using the following search terms: ("narrative skill*" or "narrative abilit*" or narration) AND (assess* or evaluat* or perform*). Across databases, studies were not restricted to recent publications or peer-reviewed articles; therefore, dissertations and poster abstracts were not excluded from consideration. This database search resulted in 1,172 studies, which were then screened for eligibility based on title and abstract (Screening Step 1) and full-text article review (Screening Step 2).

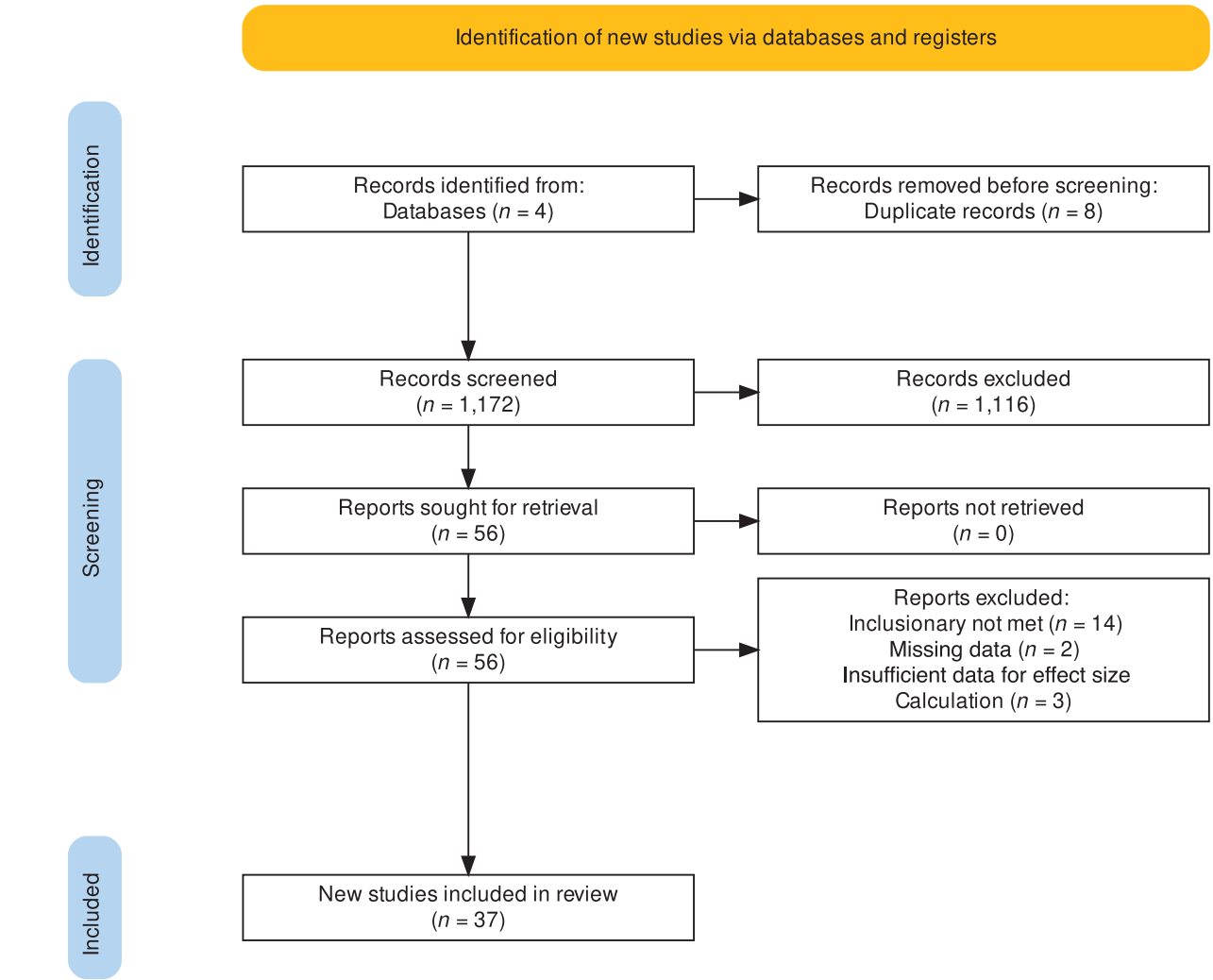
The first author completed the title and abstract screening process by evaluating study abstracts from all 1,172 studies with respect to the four inclusionary criteria. Duplicate studies ($n$ = 8) were removed during this screening process. Studies that met all four criteria advanced to the full article review screening stage. Studies where inclusion could not be determined solely by the abstract (e.g., articles that did not expand upon DLD or TD group assignment but met all other criteria) also advanced to the full article review stage. Interrater reliability between the first and second authors on a randomly selected 20% of abstracts ($n$ = 236) showed 96% reliability. This screening process resulted in a total of 56 studies that advanced to the full-text article review phase.

The second author reviewed the full text of these 56 studies to determine whether they met all inclusionary criteria and were eligible for inclusion in the present analysis. Studies with missing effect size data or missing demographic information were included at this stage. Of the 56 studies evaluated, 14 did not meet inclusionary criteria, three were excluded due to missing data, and three were excluded due to reported but insufficient data for effect size calculation (i.e., authors reported standard deviations of zero). In cases of missing or insufficient data ($n$ = 6), the first author contacted the corresponding author of each study; all responded, and one provided sufficient data for effect size calculation and was re-introduced into the sample (i.e., studies excluded due to missing data were reduced from $n$ = 3 to $n$ = 2). Interrater reliability on 20% of the full-text articles ($n$ = 11) showed 91% reliability between the second and first authors. In total, this full-text article review stage yielded a total of 37 studies included in the present meta-analysis.

#### Coding Procedure

Data from each study that passed the full article review (final $n$ = 37) were entered into a Microsoft Excel workbook by the first and second authors and a trained research assistant, using a detailed codebook. The complete dataset is available in the replication materials (https://osf.io/ybdx7/). We extracted information related to each study's sample(s), as well as both descriptive and quantitative results summarizing group performances. Specifically, we

**Figure 1.** Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart.




recorded descriptive data for each study related to sample characteristics, such as average age, average socioeconomic status (SES) level, and determination of group assignment. For studies that reported data on multiple, nonoverlapping samples, we extracted information for each sample meeting inclusion criteria. In a separate spreadsheet, we coded information regarding group performance on standardized language assessments, with one row per assessment per sample per study. In a third spreadsheet, we coded descriptive data related to the narrative task(s) administered and quantitative data related to group performances in order to calculate effect sizes. This resulted in multiple rows of data for studies reporting multiple narrative assessment tasks (i.e., multiple effect sizes were extracted for each sample). Detailed information regarding all coding procedures is outlined below.

The second author and a trained research assistant coded study data on demographic information or data relevant to the planned analyses. This included participant characteristics for each group (age, sex, language performance, nonverbal IQ, SES, bilingual/bidialectal status, language[s] spoken, and country tested), task-level data (task language, task material, elicitation type, and level of support), and information regarding the narrative measure (specific index used across macrostructure, microstructure, or internal state language). The first author subsequently checked and confirmed these values.

*Participant sample-level data.* For TD and DLD groups, summary data (including means and standard deviations) were extracted for age, sex, language performance, nonverbal IQ, SES, bilingual/bidialectal status, language(s) spoken, country tested, DLD/TD inclusionary criteria, DLD recruitment, and diagnostic term used (e.g., DLD or SLI). Age values were transformed to age in months.

Language ability status determination (TD vs. DLD) differed across studies, consistent with previous

reports in the literature (Bishop et al., 2016, 2017; Nitido & Plante, 2020). For example, some studies used criteria that aligned with district or state eligibility requirements, whereas others relied on stricter SLI criteria. To capture these differences in recruitment and determination of ability status, studies were coded for both whether DLD children were recruited via a clinical or population-based sample and how DLD status was determined. DLD inclusionary criteria were classified as (a) SLP identified, (b) two or more standardized language subtests/tests used, or (c) both.

For studies of bilingual/bidialectal populations (e.g., Spanish–English bilingual children), in addition to the task language/dialect and country variables that were coded in monolingual groups, the specific languages spoken by the child were also coded (e.g., Russian [first language (L1)] and Hebrew [second language (L2)]). Given the wide variability in descriptions of participants' bilingual/bidialectal experience (e.g., input/output, age of acquisition), more fine-grained coding was not possible (see Surrain & Luk, 2019, for a synthesis on bilingual sample reporting). Language performance on each standardized assessment measure reported in primary studies (i.e., to describe the sample and/or determine group assignment) was coded for each language reported.

To describe language performance at the group level, a single proxy for language performance was used for each language reported. In order to select one representative assessment, we established four decision rules. First, all single-word vocabulary measures (e.g., Peabody Picture Vocabulary Test; Dunn, 2019) were excluded, given their poor identification of DLD in monolingual/bilingual populations (Anaya et al., 2018; Gray et al., 1999; Shahmahmood et al., 2016; Spaulding et al., 2006). Second, all parent report measures of language performance were excluded in order to avoid confounding parent report and behavioral measures, as most studies reported behavioral assessments. Third, if only one nonvocabulary behavioral measure was reported, this was used. Finally, if multiple measures were included, we used the following hierarchy: a composite score of a standardized battery with specified cutoffs and adequate specificity/sensitivity (Nitido & Plante, 2020), a subtest of a battery (if only one was administered) or a composite score of a battery that did not have adequate psychometrics to be used alone, and an average of two or three subtest scores from a standardized battery. In order to compare language performance across a heterogeneous list of assessments ($n$ = 22; a complete list of reported assessments is available in the dataset in replication materials at https://osf.io/ybdx7/), we collected normative data for these selected assessments. These normative means and standard deviations were used to transform sample means and standard deviations into $z$ scores.

*Task-level data.* We also extracted information about the narrative tasks used. This included the task material (e.g., frog story; *TNL*, Gillam & Pearson, 2017), the language of the task, and the task context (e.g., story retell with sequential pictures), as described in the primary studies. In order to represent similarities and differences across task contexts, we classified each task along two dimensions: task type (retell, tell, or both retell and tell) and level of support (visual, auditory only, or a combination). Interrater reliability on 20% of these data between the first and second authors was greater than 90% (task type = 97%, task support = 91%).

*Categorization of narrative assessment measures.* Data related to written language or oral–written composites were excluded from the analysis. Although our focus was on narrative production tasks, production measures (e.g., narrative retell) that were administered as part of a comprehension–production composite task (e.g., a narrative retell and comprehension questions) were retained. For intervention studies that reported both pre- and posttest measures, only pretest measures were retained, given our interest in overall performance, not gains after a period of intervention. For studies reporting multiple TD control groups (e.g., age matched, language matched), only the performance of age-matched controls was entered, as our research questions sought to assess performance differences in same-age children with and without DLD. Finally, to ensure interpretability of effect size comparisons and pooling across studies, measures with negative valence (i.e., ones where a higher score equals more of a negative construct such as "ungrammaticality," "errors," and "maze production") were reversed.

Each eligible outcome measure was first coded by (a) the narrative measure type (macrostructure, microstructure, or internal state language), in line with previous work (Baixauli et al., 2016), and (b) the specific narrative assessment measure reported by the study (e.g., MLU in morphemes or $MLU_m$, percent correct adverbs). Given the large number of unique measures extracted across studies (macrostructure: $n$ = 84, microstructure: $n$ = 134, internal state language: $n$ = 13, mixed: $n$ = 10), we used thematic coding that was theoretically grounded to further classify those measures into a narrative assessment subtype. For example, MLU was coded within the microstructure narrative assessment subtype of "length," whereas measures that were identified by the primary study authors as measuring either feature-specific (e.g., verb accuracy) or global sentence accuracy (e.g., PGU) were coded as "accuracy." Table 1 provides examples of this categorization process. Reliability for these intermediate measure constructs between the first and second authors was 92%.

*Effect size calculations.* Standardized mean differences, which quantified the difference in performance between TD and DLD groups, were the effect sizes metric. Effect sizes were coded so that a negative value

**Table 1.** Categorization of narrative measures.

| Narrative assessment type | Narrative assessment subtype | Example narrative measure | Study |
|---|---|---|---|
| Macrostructure | Events | Total events recalled | Dodwell & Bavin (2008) |
| | Events | Complete episodes | Liles (1987) |
| | Story grammar | Story components | Reilly et al. (2004) |
| | Story grammar | Internal response | Hao et al. (2018) |
| | Referencing/cohesion | Number of first mentions | Rezzonico et al. (2015) |
| | Referencing/cohesion | Pronoun maintenance | Fichman & Altman (2019) |
| | Combination | Narrative Language Measure story retell | Petersen et al. (2017) |
| Microstructure | Accuracy | Errors per C-unit | Colozzo et al. (2011) |
| | Accuracy | Grammaticality (%) | Duinmeijer et al. (2012) |
| | Diversity | Number of different words | Guo & Schneider (2016) |
| | Diversity | Different verbs per C-unit | Reuterskiold et al. (2011) |
| | Fluency | Mazes per C-unit | Newman & McGregor (2006) |
| | Fluency | C-units with mazes | Newman & McGregor (2006) |
| | Length | MLU in morphemes | Thordardottir (2008) |
| | Length | Subordination index | Pham et al. (2019) |
| Internal state language | — | Motivational verbs | Altman et al. (2016) |
| | — | Perceptual verbs | Altman et al. (2016) |
| Mixed | — | DSLT short narrative subtest | Burns et al. (2012) |
| | — | TNL composite | Colozzo et al. (2011) |

*Note.* "Internal state language" and "mixed" categories were not coded further. MLU = mean length of utterance; DSLT = Dialect Sensitive Language Test; TNL = Test of Narrative Language.

indicated that children with DLD performed worse than TD peers and a positive value indicated the opposite pattern. For measures where a higher score represented a less desirable performance (e.g., "ungrammaticality," "errors," or "maze production"), the direction of the effect size was reversed to ensure interpretability when pooling across studies. Standardized mean differences were calculated using the Hedges' *g* estimator (Hedges, 1981).

## Planned Analyses

Data were analyzed using R 4.0.2 (R Core Team, 2020) in RStudio (RStudio Team, 2015), using two packages appropriate for meta-analysis: *metafor* (Version 2.4.0; Viechtbauer, 2010) and *clubSandwich* (Version 0.5.2; Pustejovsky, 2020). R scripts used to conduct all reported analyses are available in the replication materials at https://osf.io/ybdx7/.

### Outliers

Before conducting analysis, we examined the distribution of effect size estimates for outliers, defined as an effect size estimate falling more than 3 times the interquartile range below the first quartile or more than 3 times the interquartile range above the third quartile (Tukey, 1977). Based on this definition, there were no outlying effect sizes, so we conducted analysis using all eligible effect size estimates.

### Effect Size Dependency

The set of effect size estimates included in our analyses had two features that necessitated use of advanced meta-analytic methods. First, many of the eligible samples included effect size data for multiple measures of narrative performance. Effect size estimates calculated from a common sample of participants will have correlated sampling errors and so should not be treated as independent. However, the information needed to assess the degree of correlation was seldom provided in primary study reports. Second, some studies reported data for multiple participant samples. Effect sizes estimated from different samples within the same study may be similar if researchers use the same measurement procedures, recruitment strategies, or other study procedures across the samples. Thus, the effect size data exhibited a hierarchical dependence structure, with individual effect sizes nested within samples and samples nested within studies.

To synthesize the effect size estimates while accounting for these two features, we used a meta-analytic method known as RVE (Hedges et al., 2010; Pustejovsky & Tipton, 2022). RVE involves use of a working model that describes tentative assumptions about the degree of correlation and other forms of dependence in the effect size estimates and that is used to determine an appropriate weight assigned to each effect size and study. However, the standard errors, hypothesis tests, and confidence intervals (CIs) based on RVE remain valid even if some of the assumptions of the working model are wrong (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). In addition, we used RVE with the "CR2" small-sample correction method (Tipton, 2015; Tipton & Pustejovsky, 2015) to ensure that tests and confidence intervals are well calibrated (i.e., have near-nominal Type I error rates or

coverage rates) even for analyses involving a small number of studies.

## Summary Meta-Analysis and Sensitivity Analysis

As a preliminary summary of the effect size distribution, we estimated an overall average effect size across all included samples. For the summary meta-analysis, we used RVE with a working model that assumed that effect size estimates calculated from a common sample would be correlated at $\rho = .4$ and that there would be heterogeneity at three distinct levels: across effect sizes nested within samples, across samples nested within studies, and across studies. With this working model, heterogeneity is characterized in terms of estimated standard deviations at each level (effect level, sample level, and study level). Larger standard deviations indicate that there is a higher degree of variability in the effect sizes, after accounting for the variation expected due purely to sampling error. Because we used RVE, estimates of overall average effect size should be valid even if our assumption about the degree of correlation was inaccurate. Nonetheless, we conducted sensitivity analyses to assess the extent to which our estimates size was sensitive to this assumption, by repeatedly re-estimating the model using correlations ranging from $\rho = .0$ to $\rho = .9$. As a further sensitivity analysis, we also repeatedly re-estimated the summary meta-analysis while excluding each study in turn, in order to gauge whether results were influenced by inclusion of a single, specific study.

## Publication Bias

Publication bias and other forms of selective reporting of study results are common in many scientific disciplines (Rothstein et al., 2006; Vevea et al., 2019), including communication disorders, where publication bias is often acknowledged but rarely evaluated (e.g., Chow, 2018). Selective reporting occurs if specific study results (or even full studies) are more likely to be reported—and thus available for inclusion in a meta-analysis—when they are statistically significant compared to when they are not statistically significant. This can lead to inflation of effect size estimates and distortion of meta-analytic results. To assess this possibility in our data, we created a contour-enhanced funnel plot (Peters et al., 2008). Asymmetry in the distribution of effect sizes within the funnel plot is a warning sign of potential selective reporting (although other factors can also lead to asymmetry). We also formally tested for asymmetry using a modified form of Egger's regression test, which used RVE to account for effect size dependency (Rodgers & Pustejovsky, 2020).

## Meta-Regression Analysis

To address each of the first three research questions, as well as the sixth research question, we estimated a series of three meta-regression models using RVE with the same working model as in the summary meta-analysis. For each research question, we used meta-regression models with indicator variables for each level of the focal moderator (e.g., each measure type for Research Question [RQ] 1, each macrostructure narrative construct for RQ2), along with one or more additional predictors to account for variation in effect size magnitude due to factors other than the focal moderator. Model A adjusted for average age of the DLD group. Model B adjusted for average age of the DLD group, an indicator for non-English tasks, an indicator for bilingual samples, and an indicator for outcomes assessed in a second language. Finally, Model C adjusted for all of the predictors included in Model B, along with indicators for each level of the focal moderator that were centered at the mean of each unique sample. These sample mean-centered indicator variables account for sample-level differences in composition of the focal moderator. As a result, differences between levels of the focal moderator are estimated based only on *within-sample* comparisons—all sample-level factors are held constant. Model C therefore provides the most stringent test of differences in effect size between levels of the focal moderator.

The fourth and fifth research questions involved more than one focal moderator. To address them, we specified meta-regression models that included indicator variables for each of the four narrative measure types, along with indicators for the levels of the focal moderators. For RQ4, the recruitment strategy indicators were compared to the reference level of clinical recruitment; the inclusionary criteria indicators were compared to the reference level of SLP only. For RQ5, the task-type indicators were compared to the reference level of retell tasks; the task support indicators were compared to the reference level of auditory support.

For ease of interpretation, we centered most of the additional predictor variables near their median values and we centered publication year at 2020, given that most studies were published in the recent past. Average effect size estimates can therefore be interpreted as applying to a contemporary population where DLD samples have an average age of 7.0 years, an average nonverbal IQ of 95, and an average language test performance $z$ score of $-1.64$; are monolingual; use English-language tasks; and use tasks in L1 (rather than L2 for bilingual speakers). For each research question, we tested for differences across categories of the focal moderator using robust Wald tests with small-sample corrections (Tipton & Pustejovsky, 2015), which are similar to $F$ tests but with denominator degrees of freedom that can take noninteger values.

## Results

### Participant and Study Characteristics

The final meta-analysis summarized a total of 57 samples from 37 studies. Studies were published between 1987 and 2019, and each study reported between one and six samples (mean = 1.5 samples per study). Table S1 in the Supplemental Material includes all descriptive statistics related to sample size, participant characteristics (e.g., age, language status), and sample characteristics (e.g., method of DLD recruitment). Sample sizes ranged from 20 to 435 total participants. Only two studies had samples greater than 100 participants (Burns et al., 2012; Kapantzoglou et al., 2019). The average sample size was 55.63 participants.

Table 2 reports summaries of participant characteristics for the included samples; Table 3 summarizes the distribution of task characteristics. Across samples, the average age of participating children with DLD ranged from 4.3 to 11.4 years (mean = 7.1 years). The average age of participating TD children ranged from 4.2 to 11.5 years (mean = 7.1 years). Participants represented Australia, Canada, China, Iceland, Israel, Italy, Navajo Nation, the Netherlands, Sweden, Vietnam, and the United States. Forty-eight samples included monolingual speakers, and nine samples included bilingual speakers; language experience included Dutch, English, French, Hebrew, Icelandic, Italian, Mandarin, Russian, Spanish, Swedish, and Vietnamese. Children with DLD were recruited by population sampling ($n$ = 6 samples), by being clinically referred by an SLP ($n$ = 48 samples), or through unspecified means ($n$ = 3 samples). Inclusionary criteria involved a referral from an SLP ($n$ = 3 samples), standardized testing performance ($n$ = 17 samples), or both ($n$ = 37 samples). Additional sample characteristics (i.e., language test performance, nonverbal IQ, SES) were available for some but not all studies (see Table 2 for a summary of missing data and Supplemental Material, Table S1, for descriptive statistics for these variables). In general, DLD and TD groups performed similarly on measures of nonverbal IQ, TD groups earned higher scores on standardized language testing, and both groups reported similar SES across samples.

A total of 382 effect sizes were obtained across all 57 samples and 37 studies. Individual studies reported between 1 and 42 effect sizes. The simple average effect size estimate was −0.75 ($SD$ = 0.73), with TD children outperforming children with DLD by 0.75 $SD$s, on average. Fourteen studies included samples with one or more positive effect sizes, indicating tasks where children with DLD outperformed TD children, albeit marginally. See Figure 2 for a forest plot depicting these referenced effect size estimates from each included sample. In the figure, each row (along the vertical axis) represents data from one unique sample, with effect size estimates (i.e., standardized mean differences for TD–DLD differences) from that sample represented along the horizontal axis. Samples are ordered based on average effect size magnitude, from most positive (at the top) to most negative (at the bottom). The majority of estimates are negative, corresponding to superior TD performance.

### Meta-Analysis

An overall summary meta-analysis found an average TD–DLD group difference of −0.822 $SD$s, $t(33.7)$ = −10.1, $p < .0001$, 95% CI [−0.987, −0.657]. There was substantial heterogeneity across effect sizes, $Q(381)$ = 1865.1, $p < .0001$, with variance component estimates (reported as standard deviations) of 0.38 at the study level, 0.11 at the sample level, and 0.49 at the effect size level. The overall average effect size was not sensitive to our assumption about correlation between effect size

**Table 2.** Distribution of participant characteristics for $k$ = 57 samples.

| Sample characteristic | Group | % Missing | Mean | *SD* | Min | Max |
|---|---|---|---|---|---|---|
| Sample size | DLD | 0.0 | 20.7 | 15.2 | 9.0 | 100.0 |
| | TD | 0.0 | 35.0 | 45.0 | 9.0 | 335.0 |
| Age (months) | DLD | 0.0 | 85.6 | 23.8 | 51.7 | 137.0 |
| | TD | 0.0 | 85.5 | 23.4 | 50.6 | 138.0 |
| Nonverbal IQ | DLD | 61.4 | 95.3 | 5.6 | 87.0 | 107.1 |
| | TD | 66.7 | 103.9 | 5.8 | 93.6 | 116.0 |
| SES | DLD | 56.1 | 18.5 | 18.2 | 2.3 | 48.7 |
| | TD | 56.1 | 20.0 | 19.4 | 1.9 | 48.8 |
| Language test performance | DLD | 56.1 | –1.9 | 0.9 | –3.9 | –0.6 |
| | TD | 68.4 | 0.0 | 0.5 | –1.2 | 0.7 |
| TD–DLD difference in language test | | 68.4 | –2.0 | 0.8 | –4.0 | –0.6 |

*Note.* Standardized test performance statistics reported in original studies were transformed to $z$ scores. DLD = developmental language disorder; TD = typically developing; SES = socioeconomic status.

**Table 3.** Number of studies, samples, and effect sizes by task characteristics.

| Category | | Studies (*m*) | Samples (*n*) | Effect sizes (*k*) |
|---|---|---|---|---|
| Assessment type | | | | |
| Macrostructure | | 26 | 39 | 111 |
| | Referencing/cohesion | 8 | 12 | 45 |
| | Story grammar | 14 | 24 | 43 |
| | Events | 10 | 11 | 21 |
| | Combination | 1 | 1 | 2 |
| Microstructure | | 31 | 49 | 235 |
| | Length | 27 | 38 | 119 |
| | Accuracy | 21 | 33 | 68 |
| | Diversity | 15 | 25 | 35 |
| | Fluency | 4 | 4 | 13 |
| ISL | | 3 | 4 | 20 |
| Mixed | | 6 | 13 | 16 |
| DLD recruitment | | | | |
| Clinical | | 29 | 48 | 288 |
| Population | | 6 | 6 | 36 |
| Not specified | | 2 | 3 | 58 |
| DLD inclusionary criteria | | | | |
| Standardized assessment | | 13 | 17 | 189 |
| SLP clinical judgment | | 3 | 3 | 18 |
| Standardized assessment and SLP clinical judgment | | 21 | 37 | 175 |
| Task type | | | | |
| Tell | | 16 | 31 | 191 |
| Retell | | 20 | 23 | 174 |
| Retell and tell | | 3 | 5 | 17 |
| Task support | | | | |
| Visual | | 22 | 38 | 240 |
| Auditory | | 3 | 3 | 27 |
| Visual and auditory | | 14 | 18 | 115 |
| Task language | | | | |
| English | | 28 | 44 | 264 |
| Other language | | 12 | 15 | 118 |
| Language status | | | | |
| Monolingual | | 31 | 48 | 289 |
| Bilingual | | 9 | 9 | 93 |

*Note.* ISL = internal state language; DLD = developmental language disorder; SLP = speech-language pathologist.

estimates, with estimates varying from −0.850, 95% CI [−1.016, −0.685], assuming ρ = .0, to −0.794, 95% CI [−0.963, −0.624], assuming ρ = .9. The estimated total variation in true effect sizes was similarly insensitive, with a total *SD* ranging from 0.626 to 0.659, although individual variance component estimates were more sensitive to the assumed ρ (see Supplemental Material, Figure S3 and Table S5). The overall average effect size and total variation in true effect sizes were not strongly influenced by any single study (see Supplemental Material, Figure S4). A cluster-robust Egger's regression test provided no clear indication of publication bias (see Supplemental Material, Section S4, for a contour-enhanced funnel plot and results of Egger's regression).

Our overall multilevel meta-analysis model assessed TD–DLD differences across all $k = 382$ effect sizes while accounting for nesting within samples and studies; however, it did not account for pertinent characteristics such as the measure type (e.g., macrostructure vs. microstructure) or age of the participants. Therefore, in order to answer each of our research questions, we assessed three meta-regression models with different study-, sample-, and effect size–level characteristics.

## RQ1: Are There TD–DLD Differences Across Narrative Assessment Type (Macrostructure, Microstructure, and Internal State Language)?

We assessed differences across narrative assessment types using Models A, B, and C; results are reported in Table 4. Based on Model A, there were significant differences between microstructure, macrostructure, internal state language, and mixed narrative assessment type, $F(3, 2.8) = 17.7$, $p = .024$. Average group differences were statistically distinct from zero for each narrative assessment type, with estimated effects of −1.042 *SD*, 95% CI [−1.915, −0.169], for mixed types; −0.853, 95% CI

**Figure 2.** Forest plot of TD–DLD group differences in narrative performance (standardized mean differences, 95% confidence interval). TD = typically developing; DLD = developmental language disorder.

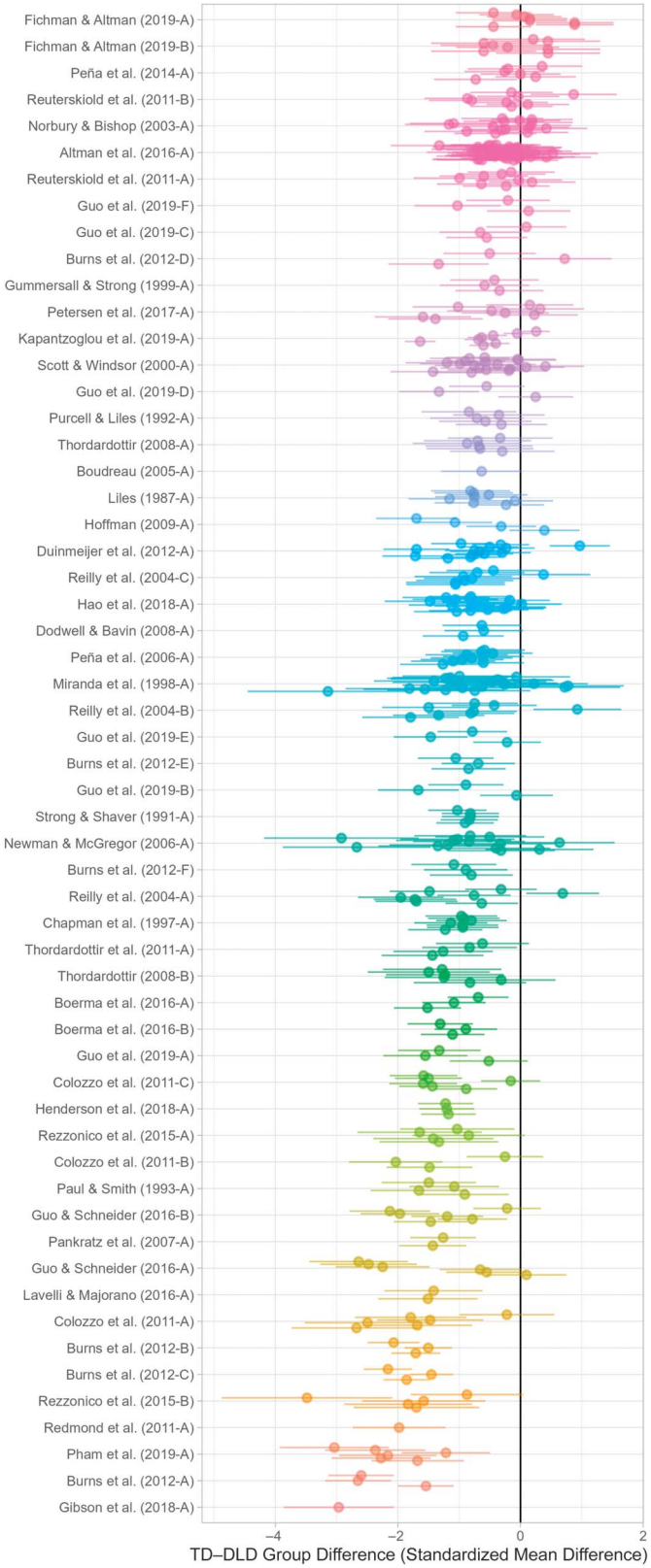


**Table 4.** Average TD–DLD group differences by narrative assessment type.

| Term | Model A | | | Model B | | | Model C | | |
|---|---|---|---|---|---|---|---|---|---|
| | Est. (*SE*) | 95% CI | *p* value | Est. (*SE*) | 95% CI | *p* value | Est. (*SE*) | 95% CI | *p* value |
| Average effect size | | | | | | | | | |
| Macrostructure | −0.782 (0.113) | [−1.014, −0.551] | < .001 | −0.855 (0.124) | [−1.112, −0.599] | < .001 | −0.832 (0.118) | [−1.077, −0.586] | < .001 |
| Microstructure | −0.853 (0.090) | [−1.036, −0.670] | < .001 | −0.924 (0.093) | [−1.114, −0.733] | < .001 | −0.897 (0.086) | [−1.076, −0.718] | < .001 |
| ISL | −0.608 (0.080) | [−0.930, −0.286] | .014 | −0.690 (0.093) | [−1.022, −0.358] | .009 | −0.677 (0.077) | [−0.964, −0.390] | .007 |
| Mixed | −1.042 (0.284) | [−1.915, −0.169] | .031 | −1.132 (0.286) | [−1.991, −0.274] | .023 | −0.789 (0.423) | [−2.249, 0.672] | .172 |
| Wald tests | | | | | | | | | |
| Assessment type | $F(3, 2.8) = 17.7$ | | .024 | $F(3, 2.9) = 8.5$ | | .061 | $F(3, 2.2) = 15.3$ | | .052 |
| Variance components | | | | | | | | | |
| Study-level *SD* | 0.356 | | | 0.321 | | | 0.269 | | |
| Sample-level *SD* | 0.000 | | | 0.000 | | | 0.000 | | |
| Effect-level *SD* | 0.486 | | | 0.489 | | | 0.482 | | |
| Total *SD* | 0.602 | | | 0.585 | | | 0.553 | | |

*Note.* Model A controls for the average age of the participants with DLD. Model B controls for the average age of the participants with DLD, task language, bilingual samples, and outcomes assessed in a second language. Model C controls for all predictors included in Model B, along with sample mean-centered indicators for each level of the focal moderator. TD = typically developing; DLD = developmental language disorder; Est. = estimate; *SE* = standard error; CI = confidence interval; ISL = internal state language.

[−1.036, −0.670], for microstructure; −0.782, 95% CI [−1.014, −0.551], for macrostructure; and −0.608, 95% CI [−0.930, −0.286], for internal state language. However, differences in average effects between macrostructure, microstructure, and internal state language were not as large, or as statistically distinct, when controlling for additional participant characteristics (Model B). Differences between narrative assessment types were further reduced when controlling for the composition of narrative assessments used in each study (Model C). Macrostructure, microstructure, and internal state language assessments held differences statistically distinct from zero across all models. Notably, substantial heterogeneity at the study and effect size levels remained despite the addition of sample- and task-related predictors.

## RQ2: Are TD–DLD Group Differences in Macrostructure Moderated by Narrative Assessment Subtype (e.g., Story Grammar)?

Table 5 reports estimated TD–DLD group differences for each macrostructure subtype, based on models estimated for the subset of effect sizes from macrostructure assessments. There was not strong evidence of differences between macrostructure subtypes. Descriptively, the macrostructure subtype that emerged as having the largest effect sizes across models was story grammar (average effect size estimates ranging from −1.059 to −0.852 across Models A, B, and C), followed by events (−0.872 to −0.802), referencing (−0.780 to −0.603), and a combination of these constructs (−0.734 to −0.406). Across all three models, variance component estimates indicated considerable heterogeneity both between and within studies.

## RQ3: Are TD–DLD Differences in Microstructure Moderated by Narrative Assessment Subtype (e.g., Lexical Diversity)?

Table 6 reports estimated TD–DLD differences for each microstructure subtype, based on the subset of effect sizes from microstructure assessments. Across all three models, there was clear and consistent evidence of differences between microstructure assessment subtypes. Accuracy measures emerged as having the largest average effect size, showing a TD–DLD difference of −1.135, 95% CI [−1.587, −0.683], in Model A and similar levels in Models B and C. Fluency had the lowest average effect size across models, showing a TD–DLD difference of −0.039, 95% CI [−0.627, 0.550], in Model A and similar levels in Models B and C. Similar to the analysis of macrostructure subtypes, there was substantial remaining heterogeneity between and within studies.

## RQ4: Across Narrative Assessment Type, Are TD–DLD Differences Moderated by DLD Recruitment Method or Inclusionary Criteria?

There was no strong evidence that DLD recruitment method (clinical vs. population) or inclusionary criteria (SLP determination only, standardized assessment only, or a combination of SLP determination and standardized assessment) influenced TD–DLD differences. Supplemental Material, Table S6, reports results from this analysis. Differences in recruitment method were not statistically significant in any model. Descriptively, studies that determined inclusion by standardized assessment (alone or in combination with SLP judgment) exhibited larger TD–DLD differences than studies that used SLP determination only; however, differences between inclusionary criteria were imprecisely estimated and not statistically distinguishable. Of note, there was an uneven number of studies across comparison groups, resulting in limited power to detect a true difference, if one existed. For example, very few studies reported population-based samples as opposed to samples recruited from clinical practice (see Table 3).

## RQ5: Across Narrative Assessment Type, Are TD–DLD Group Differences Moderated by Task Type (e.g., Tell vs. Retell) or Level of Support (e.g., Verbal Cues)?

We examined task type and level of support within the same model specifications, which also controlled for narrative assessment type (see Supplemental Material, Table S7). There was insufficient evidence to rule out the possibility that task type was unassociated with TD–DLD group differences in narrative performance. Descriptively, retell-and-tell tasks had larger group differences compared to narrative retell-only tasks, but this difference was imprecisely estimated and not statistically distinct from zero. Task level of support explained little variation in TD–DLD group differences and was not statistically significant in any of the models.

## RQ6: Across Narrative Assessment Type, Are TD–DLD Group Differences Moderated by Common Narrative Assessment Measures (e.g., MLU)?

For this research question, we identified specific narrative assessment measures across macrostructure, microstructure, and internal state language assessment types that were reported most often in primary studies (i.e., reported by at least three primary studies). We condensed related but slightly different measures (e.g., MLU in morphemes compared to MLU in words) to capture as many measures as possible in this analysis. This resulted in 12

**Table 5.** Average TD–DLD group differences by macrostructure narrative assessment subtype.

| Term | Model A | | | Model B | | | Model C | | |
|---|---|---|---|---|---|---|---|---|---|
| | Est. (*SE*) | 95% CI | *p* value | Est. (*SE*) | 95% CI | *p* value | Est. (*SE*) | 95% CI | *p* value |
| Average effect size | | | | | | | | | |
| Combination | –0.406 (0.024) | [–0.465, –0.347] | < .001 | –0.734 (0.163) | [–1.124, –0.345] | .003 | –0.428 (0.140) | [–0.754, –0.102] | .017 |
| Events | –0.802 (0.116) | [–1.064, –0.540] | < .001 | –0.874 (0.119) | [–1.138, –0.609] | < .001 | –0.831 (0.170) | [–1.316, –0.346] | .010 |
| Referencing/cohesion | –0.603 (0.168) | [–1.004, –0.203] | .010 | –0.679 (0.134) | [–0.988, –0.371] | < .001 | –0.780 (0.126) | [–1.104, –0.455] | .002 |
| Story grammar | –0.959 (0.139) | [–1.267, –0.650] | < .001 | –1.059 (0.172) | [–1.445, –0.672] | < .001 | –0.852 (0.122) | [–1.221, –0.483] | .004 |
| Wald tests | | | | | | | | | |
| Subtype | $F(2, 7.6) = 1.4$ | | .299 | $F(2, 7.8) = 1.6$ | | .257 | $F(2, 1.5) = 1.6$ | | .424 |
| Variance components | | | | | | | | | |
| Study-level *SD* | 0.343 | | | 0.319 | | | 0.286 | | |
| Sample-level *SD* | 0.000 | | | 0.000 | | | 0.000 | | |
| Effect-level *SD* | 0.432 | | | 0.439 | | | 0.441 | | |
| Total *SD* | 0.552 | | | 0.543 | | | 0.526 | | |

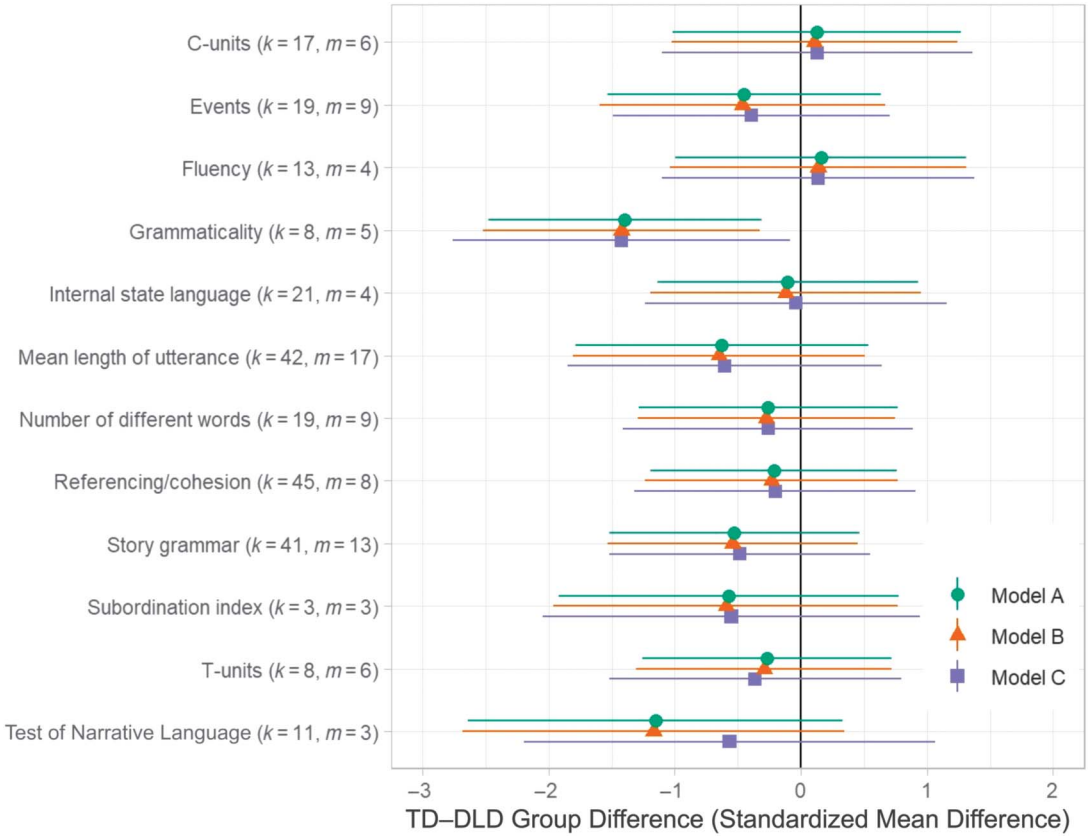
*Note.* Model A controls for average age of the participants with DLD. Model B controls for average age of the participants with DLD, task language, bilingual samples, and outcomes assessed in a second language. Model C controls for all predictors included in Model B, along with sample mean-centered indicators for each level of the focal moderator. TD = typically developing; DLD = developmental language disorder; Est. = estimate; *SE* = standard error; CI = confidence interval.

**Table 6.** Average TD–DLD group differences by microstructure narrative assessment subtype.

| | Model A | | | Model B | | | Model C | | |
|---|---|---|---|---|---|---|---|---|---|
| **Term** | **Est. (*SE*)** | **95% CI** | ***p* value** | **Est. (*SE*)** | **95% CI** | ***p* value** | **Est. (*SE*)** | **95% CI** | ***p* value** |
| Average effect size | | | | | | | | | |
| Accuracy | –1.135 (0.216) | [–1.587, –0.683] | < .001 | –1.223 (0.206) | [–1.656, –0.791] | < .001 | –1.254 (0.234) | [–1.741, –0.767] | < .001 |
| Diversity | –0.755 (0.136) | [–1.047, –0.464] | < .001 | –0.861 (0.151) | [–1.187, –0.535] | < .001 | –0.862 (0.146) | [–1.173, –0.552] | < .001 |
| Fluency | –0.039 (0.207) | [–0.627, 0.550] | .861 | –0.123 (0.213) | [–0.718, 0.472] | .595 | –0.075 (0.201) | [–0.613, 0.463] | .728 |
| Length | –0.794 (0.086) | [–0.970, –0.618] | < .001 | –0.885 (0.086) | [–1.065, –0.706] | < .001 | –0.877 (0.099) | [–1.084, –0.670] | < .001 |
| Wald tests | | | | | | | | | |
| Subtype | $F(3, 7.6) = 6.7$ | | .015 | $F(3, 7.7) = 6.5$ | | .016 | $F(3, 7.0) = 10.8$ | | .005 |
| Variance components | | | | | | | | | |
| Study-level *SD* | 0.337 | | | 0.270 | | | 0.292 | | |
| Sample-level *SD* | 0.001 | | | 0.000 | | | 0.000 | | |
| Effect-level *SD* | 0.448 | | | 0.451 | | | 0.452 | | |
| Total *SD* | 0.561 | | | 0.525 | | | 0.538 | | |

*Note.* Model A controls for average age of the participants with DLD. Model B controls for average age of the participants with DLD, task language, bilingual samples, and outcomes assessed in a second language. Model C controls for all predictors included in Model B, along with sample mean-centered indicators for each level of the focal moderator. TD = typically developing; DLD = developmental language disorder; Est. = estimate; *SE* = standard error; CI = confidence interval.

**Figure 3.** Average TD–DLD group differences on commonly used narrative assessment measures. TD = typically developing; DLD = developmental language disorder; *k* = number of effect size estimates; *m* = number of studies.




specific narrative analyses reported in three or more studies, within 3–24 samples and with 3–45 effect sizes. These analyses and their average effect size estimates are depicted in Figure 3. There was insufficient evidence to suggest TD–DLD group differences varied across the 12 narrative analyses (see Supplemental Material, Table S8). Although differences across measures were not statistically distinguishable, descriptively, grammaticality and tests of narrative language measures were associated with the largest effect size estimates. Only for grammaticality measures were the average effect sizes statistically distinct from zero.

## Discussion

Research over the past three decades has substantially increased our knowledge of narration and narrative assessment for children with DLD. This study meta-analyzed TD–DLD group differences in narration and revealed TD children demonstrate superior narrative performance compared to age-matched peers with DLD. In order to answer our research questions, we employed three distinct meta-regression models to assess potential moderating effects across narrative assessment type (i.e., macrostructure, microstructure, internal state language [RQ1]), assessment subtype (e.g., story grammar, macrostructure [RQ2]; grammatical accuracy, microstructure [RQ3]), study characteristics (i.e., recruitment method and inclusionary criteria [RQ4]), narrative task characteristics (i.e., task type and level of support [RQ5]), and commonly employed assessment measures (e.g., MLU [RQ6]). Considerable heterogeneity, an uneven number of effect sizes across comparison groups, and inconsistent reporting of pertinent demographic information (e.g., age) somewhat limited our ability to answer these questions. However, these limitations provide valuable insight for prospective investigations that aim to further delineate differences between TD children and children with DLD. Promisingly, our findings do provide compelling evidence for SLPs to incorporate outcome measures for story grammar and grammatical accuracy as part of their assessment protocol when considering a diagnosis of DLD. We discuss these findings in relation to the existing literature and clinical practice and then provide suggestions for future research and meta-analyses in speech, language, and hearing sciences.

## Assessment Type and Subtype: Evidence for Story Grammar and Grammatical Accuracy

Although the results of our overall meta-analysis revealed significant TD–DLD group differences across effect sizes, no single assessment type (i.e., macrostructure, microstructure, internal state language) yielded greater TD–DLD differences, even when controlling for participant- and study-level characteristics. This may be due, in part, to the heterogeneity observed within and between primary studies. For example, some researchers have found relative weaknesses in specific narrative skills (e.g., microstructure; Altman et al., 2016) or different profiles of strengths and weaknesses across children (e.g., Colozzo et al., 2011). With respect to assessment subtypes, story grammar (RQ2 and RQ6) and grammatical accuracy (RQ3 and RQ6) represented the greatest differences between TD children and children with DLD.

Within the macrostructure subtype, story grammar yielded the greatest average effect sizes (RQ2). Descriptively, story grammar also had one of the largest average effect sizes when compared to other commonly used analyses (e.g., MLU; RQ6). In our sample, story grammar was measured across studies using the ENNI story grammar (Schneider et al., 2006), the dynamic assessment and intervention total story score (Miller et al., 2001), and TNL content score (Gillam & Pearson, 2017), among other measurements of story components. Based on these findings, story grammar should be included in SLPs' assessment of narration for children with DLD. Furthermore, not surprisingly, recent research demonstrates direct instruction in story grammar is an effective method for narrative intervention for school-age children (e.g., Story Grammar Marker; Pico et al., 2021).

Grammatical accuracy emerged as the largest average effect size across models when compared to other subtypes of microstructure (i.e., diversity, fluency, and length; RQ3) and descriptively when compared to other commonly used analyses (e.g., MLU; RQ6). Accuracy measures typically considered children's morphosyntactic productions in obligatory contexts. Across the studies analyzed, the most common measures of grammatical accuracy were PGU, a global (sentence-level) metric, or accuracy of language-specific features (e.g., *ba* particle in Mandarin; Hao et al., 2018). Additional measurements of grammatical accuracy included percent morphological errors, verb accuracy, and TNL form score. Increasingly, research suggests grammatical accuracy likely interacts with other elements of microstructure. For example, in their analysis of Spanish–English bilingual language samples, Castilla-Earls et al. (2022) reported more errors of omission in longer utterances in both Spanish and English in addition to fewer errors of omission in shorter utterances. Thus, SLPs should include measures of grammatical accuracy that are appropriate for the child's language history in their assessment of children with suspected DLD. Furthermore, for intervention studies, researchers should consider one or more measures of grammatical accuracy (e.g., percentage of grammatically correct C-units) as an outcome measure(s), in addition to the frequently used measures of length (e.g., MLU) or diversity (e.g., NDW; Pico et al., 2021).

## Assessing Narrative Assessment Type and Subtype

Internal state language, which draws on both macrostructure and microstructure, did not arise as an area of difficulty for children with DLD compared to their TD peers (RQ1). One reason for this finding may be that few studies investigated this aspect of narration or provided a detailed analysis of this assessment. Unlike other neurodevelopmental communication disorders (e.g., autism) that may have a specific deficit in this area (Baixauli et al., 2016), relative impairment in internal state language may be influenced by a more global language impairment, or an interaction between macrostructure and microstructure, in children with DLD. Future research is needed to successfully meta-analyze this assessment type.

Additional aims of this study were to identify specific narrative assessment subtypes (e.g., story grammar, macrostructure [RQ2]; length, microstructure [RQ3]) and specific analyses (e.g., MLU, length, microstructure; RQ6) that yield the greatest TD–DLD differences to aid SLPs in assessing children with suspected DLD. Although our findings provided moderate support for evaluation of grammatical accuracy and story grammar, large variability and heterogeneity at the effect, sample, and study levels made it difficult to assess TD–DLD differences. Future research should report correlations between measures of interest, as well as correlations with other commonly used narrative analyses. Routine reporting of correlations would allow for a more nuanced analysis of variation in TD–DLD differences across assessment types and would improve the capacity to predict such differences on specific assessments (e.g., PGU) or for specific populations (e.g., children with DLD).

## Assessing Study Characteristics

For all research questions, each of the three statistical models included an increasing number of participant-level characteristics (e.g., chronological age, Model A). In RQ4, we specifically assessed two study-level characteristics as potential moderators of TD–DLD differences: DLD recruitment method (i.e., clinical vs. population) and inclusionary criteria (i.e., SLP determination only, standardized assessment

only, or a combination). Although these participant- and study-level characteristics were extracted from the primary studies included in this meta-analysis, limited power reduced our ability to detect differences across comparison groups with uneven numbers. For example, as noted in RQ4, very few studies reported population-based samples compared to clinical samples, and as a result, we were unable to assess if population-based samples yield smaller group differences between children with and without DLD compared to clinical samples (i.e., spectrum bias; Dollaghan & Horner, 2011).

One limitation of the primary studies included in this meta-analysis was the insufficient description of participant characteristics, including age, DLD inclusionary criteria, language/dialect history, language assessment scores, nonverbal IQ, SES, and treatment history. Many of these clinically relevant elements could not be investigated because this information was either not provided or not described in sufficient detail to allow secondary analysis. For example, some researchers noted a general age or grade range without reporting descriptive statistics. Other researchers provided a general description of SES by stating that participants were recruited from a middle-class community without reporting specific information for the participants in their sample. Future research should therefore provide more detailed descriptions of participants included in a sample, including descriptive statistics (e.g., means, standard deviations) for these variables of interest.

### Assessing Narrative Task Characteristics

An additional purpose of this study was to examine the role of task characteristics (i.e., narrative retell vs. tell) on TD–DLD differences. Our results did not provide compelling evidence regarding task-level characteristics. Although we included both narrative format (tell vs. retell) and support provided (e.g., visual, verbal) in our models, narrative retell was overrepresented, and there was no sufficient variation to assess whether this specific task characteristic moderated greater TD–DLD differences. Similarly, few tasks employed verbal- or auditory-only support compared to visual support or a combination of visual and verbal support. One recommendation from these findings would be for clinicians to use both narrative formats and individually tailored support for each child because it would allow SLPs to navigate variability in deficits that may be unique to each child.

### Conducting Meta-Analyses

Results of this meta-analysis are robust to heterogeneity and variation across primary studies due to the use of analytic methods based on RVE, which account for such heterogeneity without relying on strong modeling assumptions. One benefit of the current methodology was our ability to assess a variety of narrative assessments across the areas of macrostructure, microstructure, and internal state language. This was particularly useful given that researchers provided extensive information related to a diverse set of theory-informed narrative analyses. However, our methods also have some limitations that are important to note. Although PRISMA recommends reporting assessments of risk of bias for included studies (Page et al., 2020), we did not include this because of lack of established tools for assessing risk of bias for the descriptive research studies included in the review. In addition, we used single-author screening for 80% of the possible titles and abstracts. Although this may have introduced some degree of error, our high reliability (> 90%) across this and other single-author procedures increases our confidence that these decisions did not meaningfully influence the findings. Finally, this study was not preregistered. Although we did not find any preregistered reviews with overlapping research questions, future studies would benefit from preregistration to mitigate risk of duplicative research efforts and to delineate confirmatory, prespecified analyses from exploratory analyses.

Future meta-analysis investigations should incorporate methodology that is suitable not only for transparency and reproducibility (Chow et al., 2021) but also for the small sample sizes and heterogenous data reported by primary studies. Although previous meta-analyses in the field have included moderator analyses, only a few prior syntheses (e.g., Sandbank et al., 2020) have incorporated these advanced methods. Collaboration between SLPs, researchers, and methodologists can yield mutually beneficial projects that allow for the advancement both of speech, language, and hearing sciences and of methodology for research syntheses.

## Conclusions

This study implemented multilevel meta-analysis and RVE to assess differences in narration between TD children and peers with DLD across assessment type, assessment measure, and participant and sample characteristics. Although all assessment types yielded TD–DLD differences significantly different from zero, there was substantial heterogeneity within and between studies. Grammatical accuracy and story grammar analyses yielded the most compelling evidence for identifying significant TD–DLD differences. The available participant and sample characteristics did not appear to influence difference in narration; however, consistent reporting of inclusionary criteria, sample characteristics, and correlations of assessment measures may assist in future meta-analyses. Findings from this study provide valuable insight into future primary research studies and meta-analyses in speech, language, and hearing sciences that will further elucidate TD children from children with DLD.

## Author Contributions

**Katherine L. Winters:** Conceptualization (Equal), Data curation (Equal), Formal analysis (Supporting), Methodology (Equal), Project administration (Lead), Writing – original draft (Equal), Writing – review & editing (Equal). **Javier Jasso:** Conceptualization (Equal), Data curation (Equal), Formal analysis (Supporting), Methodology (Equal), Writing – original draft (Equal), Writing – review & editing (Equal). **James E. Pustejovsky:** Data curation (Supporting), Formal analysis (Lead), Methodology (Equal), Software (Lead), Visualization (Lead), Writing – original draft (Supporting), Writing – review & editing (Equal). **Courtney T. Byrd:** Supervision (Lead), Writing – original draft (Supporting), Writing – review & editing (Equal).

## Data Availability

Additional supplemental materials, complete raw data, and R code for replicating all reported analyses are available on the Open Science Framework at https://osf.io/dgz7m/ (supplemental material) and https://osf.io/ybdx7/ (replication materials).

## Acknowledgments

The corresponding author received an ASHA Convention Student Research Travel Award to present preliminary findings at the ASHA 2019 Convention. The authors would like to extend their sincere gratitude to Tasha Beretvas for her mentorship of Katherine L. Winters and Javier Jasso and for her support during the early stages of this study. They would also like to thank Kerry Lyster for her assistance with data coding.

## References

References marked with an asterisk (*) indicate studies included in the meta-analysis.

*Altman, C., Armon-Lotem, S., Fichman, S., & Walters, J. (2016). Macrostructure, microstructure, and mental state terms in the narratives of English–Hebrew bilingual preschool children with and without specific language impairment. *Applied Psycholinguistics, 37*(1), 165–193. https://doi.org/10.1017/S0142716415000466

**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

**Baixauli, I., Colomer, C., Roselló, B., & Miranda, A.** (2016). Narratives of children with high-functioning autism spectrum disorder: A meta-analysis. *Research in Developmental Disabilities, 59,* 234–254. https://doi.org/10.1016/j.ridd.2016.09.007

**Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE Consortium.** (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS ONE, 11*(7), Article e0158753. https://doi.org/10.1371/journal.pone.0158753

**Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE-2 Consortium.** (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *The Journal of Child Psychology and Psychiatry, 58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

*Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: Implications for clinical practice. *International Journal of Language & Communication Disorders, 51*(6), 626–638. https://doi.org/10.1111/1460-6984.12234

*Boudreau, D. (2005). Use of a parent questionnaire in emergent and early literacy assessment of preschool children. *Language, Speech, and Hearing Services in Schools, 36*(1), 33–47. https://doi.org/10.1044/0161-1461(2005/004)

*Burns, F. A., de Villiers, P. A., Pearson, B. Z., & Champion, T. B. (2012). Dialect-neutral indices of narrative cohesion and evaluation. *Language, Speech, and Hearing Services in Schools, 43*(2), 132–152. https://doi.org/10.1044/0161-1461(2011/10-0101)

**Castilla-Earls, A., Francis, D. J., & Iglesias, A.** (2022). The complex role of utterance length on grammaticality: Multivariate multilevel analysis of English and Spanish utterances of first-grade English learners. *Journal of Speech, Language, and Hearing Research, 65*(1), 238–252. https://doi.org/10.1044/2021_JSLHR-20-00464

**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

*Chapman, S. B., Watkins, R., Gustafson, C., Moore, S., Levin, H. S., & Kufera, J. A. (1997). Narrative discourse in children with closed head injury, children with language impairment, and typically developing children. *American Journal of Speech-Language Pathology, 6*(2), 66–76. https://doi.org/10.1044/1058-0360.0602.66

**Charest, M., Skoczylas, M. J., & Schneider, P.** (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology, 29*(4), 1866–1882. https://doi.org/10.1044/2020_AJSLP-19-00176

**Chow, J. C.** (2018). Prevalence of publication bias tests in speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research, 61*(12), 3055–3063. https://doi.org/10.1044/2018_JSLHR-L-18-0098

**Chow, J. C., Sjogren, A. L., & Zhao, H.** (2021). Reporting and reproducibility of meta-analysis in speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research, 64*(7), 2786–2793. https://doi.org/10.1044/2021_JSLHR-21-00047

*Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54*(6), 1609–1627. https://doi.org/10.1044/1092-4388(2011/10-0247)

*Dodwell, K., & Bavin, E. L. (2008). Children with specific language impairment: An investigation of their narratives and memory. *International Journal of Language & Communication Disorders, 43*(2), 201–218. https://doi.org/10.1080/13682820701366147

Dollaghan, C. A. (2004). Taxometric analyses of specific language impairment in 3- and 4-year-old children. *Journal of Speech, Language, and Hearing Research, 47*(2), 464–475. https://doi.org/10.1044/1092-4388(2004/037)

Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54*(4), 1077–1088. https://doi.org/10.1044/1092-4388(2010/10-0093)

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

Dunn, D. M. (2019). *Peabody Picture Vocabulary Test* (5th ed.). NCS Pearson.

*Fichman, S., & Altman, C. (2019). Referential cohesion in the narratives of bilingual and monolingual children with typically developing language and with specific language impairment. *Journal of Speech, Language, and Hearing Research, 62*(1), 123–142. https://doi.org/10.1044/2018_JSLHR-L-18-0054

Gaeta, L., & Brydges, C. R. (2020). An examination of effect sizes and statistical power in speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research, 63*(5), 1572–1580. https://doi.org/10.1044/2020_JSLHR-19-00299

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2012). MAIN: Multilingual assessment instrument for narratives. *ZAS Papers in Linguistics, 56,* 155. https://doi.org/10.21248/zaspil.56.2019.414

Gallagher, J. F., & Hoover, J. R. (2020). Measure what you treat: Using language sample analysis for grammatical outcome measures in children with developmental language disorder. *Perspectives of the ASHA Special Interest Groups, 5*(2), 350–363. https://doi.org/10.1044/2019_PERSP-19-00100

*Gibson, T. A., Peña, E. D., & Bedore, L. M. (2018). The receptive–expressive gap in English narratives of Spanish–English bilingual children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 61*(6), 1381–1392. https://doi.org/10.1044/2018_JSLHR-L-16-0432

Gillam, R. B., & Pearson, N. A. (2017). *Test of Narrative Language–Second Edition*. Pro-Ed.

Goffman, L., & Leonard, J. (2000). Growth of language skills in preschool children with specific language impairment: Implications for assessment and intervention. *American Journal of Speech-Language Pathology, 9*(2), 151–161. https://doi.org/10.1044/1058-0360.0902.151

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

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*(2), 196–206. https://doi.org/10.1044/0161-1461.3002.196

*Gummersall, D. M., & Strong, C. J. (1999). Assessment of complex sentence production in a narrative context. *Language, Speech, and Hearing Services in Schools, 30*(2), 152–164. https://doi.org/10.1044/0161-1461.3002.152

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

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

Haddaway, N. R., & McGuinness, L. A. (2020). *PRISMA2020: R Package and ShinyApp for producing PRISMA 2020 compliant flow diagrams (0.0.1)*. Zenodo. https://doi.org/10.5281/zenodo.4287835

*Hao, Y., Sheng, L., Zhang, Y., Jiang, F., de Villiers, J., Lee, W., & Liu, X. L. (2018). A narrative evaluation of Mandarin-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 61*(2), 345–359. https://doi.org/10.1044/2017_jslhr-l-16-0367

Hayward, M. (2003). Critiques of narrative therapy: A personal response. *Australian and New Zealand Journal of Family Therapy, 24*(4), 183–189. https://doi.org/10.1002/j.1467-8438.2003.tb00558.x

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5

*Henderson, D. E., Restrepo, M. A., & Aiken, L. S. (2018). Dynamic assessment of narratives among Navajo preschoolers. *Journal of Speech, Language, and Hearing Research, 61*(10), 2547–2560. https://doi.org/10.1044/2018_JSLHR-L-17-0313

*Hoffman, L. M. (2009). The utility of school-age narrative microstructure indices: INMIS and the proportion of restricted utterances. *Language, Speech, and Hearing Services in Schools, 40*(4), 365–375. https://doi.org/10.1044/0161-1461(2009/08-0017)

Hopkins-Rossabi, T., Curtis, P., Temenak, M., Miller, C., & Martin-Harris, B. (2019). Respiratory phase and lung volume patterns during swallowing in healthy adults: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 62*(4), 868–882. https://doi.org/10.1044/2018_JSLHR-S-18-0323

Jasso, J., McMillen, S., Anaya, J. B., Bedore, L. M., & Peña, E. D. (2020). The utility of an English semantics measure for identifying developmental language disorder in Spanish–English bilinguals. *American Journal of Speech-Language Pathology, 29*(2), 776–788. https://doi.org/10.1044/2020_AJSLP-19-00202

*Kapantzoglou, M., Fergadiotis, G., & Auza Buenavides, A. (2019). Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research, 62*(1), 70–83. https://doi.org/10.1044/2018_JSLHR-L-18-0110

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13*(2), 161–176. https://doi.org/10.1177/026565909701300204

Lancaster, H. S., Lien, K. M., Chow, J. C., Frey, J. R., Scherer, N. J., & Kaiser, A. P. (2020). Early speech and language development in children with nonsyndromic cleft lip and/or palate: A meta-analysis. *Journal of Speech, Language, and*

*Hearing Research, 63*(1), 14–31. https://doi.org/10.1044/2019_JSLHR-19-00162

***Lavelli, M., & Majorano, M.** (2016). Spontaneous gesture production and lexical abilities in children with specific language impairment in a naming task. *Journal of Speech, Language, and Hearing Research, 59*(4), 784–796. https://doi.org/10.1044/2016_JSLHR-L-14-0356

***Liles, B. Z.** (1987). Episode organization and cohesive conjunctives in narratives of children with and without language disorder. *Journal of Speech and Hearing Research, 30*(2), 185–196. https://doi.org/10.1044/jshr.3002.185

**Magimairaj, B. M., Capin, P., Gillam, S. L., Vaughn, S., Roberts, G., Fall, A. M., & Gillam, R. B.** (2022). Online administration of the Test of Narrative Language–Second Edition: Psychometrics and considerations for remote assessment. *Language, Speech, and Hearing Services in Schools, 53*(2), 404–416. https://doi.org/10.1044/2021_LSHSS-21-00129

**McGregor, K. K.** (2020). How we fail children with developmental language disorder. *Language, Speech, and Hearing Services in Schools, 51*(4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003

**McGregor, K. K., Arbisi-Kelm, T., Eden, N., & Oleson, J.** (2020). The word learning profile of adults with developmental language disorder. *Autism & Developmental Language Impairments, 5,* 1–19. https://doi.org/10.1177/2396941519899311

***Merritt, D. D., & Liles, B. Z.** (1989). Narrative analysis. *Journal of Speech and Hearing Disorders, 54*(3), 438–447. https://doi.org/10.1044/jshd.5403.438

**Miller, L., Gillam, R. B., & Peña, E.** (2001). *Dynamic assessment and intervention: Improving children's narrative abilities*. Pro-Ed.

***Miranda, E. A., McCabe, A., & Bliss, L. S.** (1998). Jumping around and leaving things out: A profile of the narrative abilities of children with specific language impairment. *Applied Psycholinguistics, 19*(4), 647–667. https://doi.org/10.1017/S0142716400010407

**Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group.** (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLOS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

***Newman, R. M., & McGregor, K. K.** (2006). Teachers and laypersons discern quality differences between narratives produced by children with or without SLI. *Journal of Speech, Language, and Hearing Research, 49*(5), 1022–1036. https://doi.org/10.1044/1092-4388(2006/073)

**Nitido, H., & Plante, E.** (2020). Diagnosis of developmental language disorder in research studies. *Journal of Speech, Language, and Hearing Research, 63*(8), 2777–2788. https://doi.org/10.1044/2020_jslhr-20-00091

***Norbury, C. F., & Bishop, D. V. M.** (2003). Narrative skills of children with communication impairments. *International Journal of Language & Communication Disorders, 38*(3), 287–313. https://doi.org/10.1080/136820310000108133

**Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A.** (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *The Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

**Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., … Moher, D.** (2020). *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews*. MetaArXiv. https://doi.org/10.31222/osf.io/v7gm2

***Pankratz, M. E., Plante, E., Vance, R., & Insalaco, D. M.** (2007). The diagnostic and predictive validity of the Renfrew Bus Story. *Language, Speech, and Hearing Services in Schools, 38*(4), 390–399. https://doi.org/10.1044/0161-1461(2007/040)

***Paul, R., & Smith, R. L.** (1993). Narrative skills in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech and Hearing Research, 36*(3), 592–598. https://doi.org/10.1044/jshr.3603.592

**Pavelko, S. L., Owens, R. E., Jr., 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., Gillam, R. B., & Bedore, L. M.** (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*(6), 2208–2220. https://doi.org/10.1044/2014_JSLHR-L-13-0151

***Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T.** (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49*(5), 1037–1057. https://doi.org/10.1044/1092-4388(2006/074)

**Pesco, D., & Kay-Raining Bird, E.** (2016). Perspectives on bilingual children's narratives elicited with the Multilingual Assessment Instrument for Narratives. *Applied Psycholinguistics, 37*(1), 1–9. https://doi.org/10.1017/S0142716415000387

**Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L.** (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

***Petersen, D. B., Chanthongthip, H., Ukrainetz, T. A., Spencer, T. D., & Steeve, R. W.** (2017). Dynamic assessment of narratives: Efficient, accurate identification of language impairment in bilingual students. *Journal of Speech, Language, and Hearing Research, 60*(4), 983–998. https://doi.org/10.1044/2016_JSLHR-L-15-0426

***Pham, G. T., Pruitt-Lord, S., Snow, C. E., Nguyen, Y. H. T., Pham, B., Dao, T. B. T., Tran, N. B. T., Pham, L. T., Hoang, H. T., & Dam, Q. D.** (2019). Identifying developmental language disorder in Vietnamese children. *Journal of Speech, Language, and Hearing Research, 62*(5), 1452–1467. https://doi.org/10.1044/2019_JSLHR-L-18-0305

**Pico, D. L., Prahl, A. H., Biel, C. H., Peterson, A. K., Biel, E. J., Woods, C., & Contesse, V. A.** (2021). Interventions designed to improve narrative language in school-age children: A systematic review with meta-analyses. *Language, Speech, and Hearing Services in Schools, 52*(4), 1109–1126. https://doi.org/10.1044/2021_LSHSS-20-00160

***Purcell, S. L., & Liles, B. Z.** (1992). Cohesion repairs in the narratives of normal-language and language-disordered school-age children. *Journal of Speech and Hearing Research, 35*(2), 354–362. https://doi.org/10.1044/jshr.3502.354

**Pustejovsky, J. E.** (2020). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package Version 0.5.2*. https://CRAN.R-project.org/package=clubSandwich

**Pustejovsky, J. E., & Tipton, E.** (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science, 23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

**R Core Team.** (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

***Redmond, S. M., Thompson, H. L., & Goldstein, S.** (2011). Psycholinguistic profiling differentiates specific language impairment from typical development and from attention-deficit/hyperactivity disorder. *Journal of Speech, Language, and Hearing Research, 54*(1), 99–117. https://doi.org/10.1044/1092-4388(2010/10-0010)

***Reilly, J., Losh, M., Bellugi, U., & Wulfeck, B.** (2004). "Frog, Where Are You?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88*(2), 229–247. https://doi.org/10.1016/s0093-934x(03)00101-9

***Reuterskiold, C., Hansson, K., & Sahlen, B.** (2011). Narrative skills in Swedish children with language impairment. *Journal of Communication Disorders, 44*(6), 733–744. https://doi.org/10.1016/j.jcomdis.2011.04.010

***Rezzonico, S., Chen, X., Cleave, P. L., Greenberg, J., Hipfner-Boucher, K., Johnson, C. J., Milburn, T., Pelletier, J., Weitzman, E., & Girolametto, L.** (2015). Oral narratives in monolingual and bilingual preschoolers with SLI. *International Journal of Language & Communication Disorders, 50*(6), 830–841. https://doi.org/10.1111/1460-6984.12179

**Rodgers, M. A., & Pustejovsky, J. E.** (2020). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods, 26*(2), 141–160. https://doi.org/10.1037/met0000300

**Rothstein, H. R., Sutton, A. J., & Borenstein, M.** (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

**RStudio Team.** (2015). *RStudio: Integrated development for R*. http://www.rstudio.com

**Rudolph, J. M., Dollaghan, C. A., & Crotteau, S.** (2019). The finite verb morphology composite: Values from a community sample. *Journal of Speech, Language, and Hearing Research, 62*(6), 1813–1822. https://doi.org/10.1044/2019_JSLHR-L-18-0437

**Sandbank, M., Bottema-Beutel, K., Crowley, S., Cassidy, M., Feldman, J. I., Canihuante, M., & Woynaroski, T.** (2020). Intervention effects on language in children with autism: A project AIM meta-analysis. *Journal of Speech, Language, and Hearing Research, 63*(5), 1537–1560. https://doi.org/10.1044/2020_JSLHR-19-00167

**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)

**Schneider, P., Hayward, D., & Dubé, R. V.** (2006). Storytelling from pictures using the Edmonton Narrative Norms Instrument. *Canadian Journal of Speech–Language Pathology and Audiology, 30*(4), 224–238.

***Scott, C. M., & Windsor, J.** (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43*(2), 324–339. https://doi.org/10.1044/jslhr.4302.324

**Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P.** (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences, 21*(1), 67. https://doi.org/10.4103/1735-1995.189648

**Southwood, F., & Russell, A. F.** (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research, 47*(2), 366–376. https://doi.org/10.1044/1092-4388(2004/030)

**Spaulding, T. J., Plante, E., & Farinella, K. A.** (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate. *Language, Speech, and Hearing Services in Schools, 37*(1), 61–72. https://doi.org/10.1044/0161-1461(2006/007)

**Stein, N., & Glenn, C.** (1979). *An analysis of story comprehension in elementary children* (Vol. 2). Ablex.

***Strong, C. J., & Shaver, J. P.** (1991). Stability of cohesion in the spoken narratives of language-impaired and normally developing school-aged children. *Journal of Speech and Hearing Research, 34*(1), 95–111. https://doi.org/10.1044/jshr.3401.95

**Surrain, S., & Luk, G.** (2019). Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition, 22*(2), 401–415. https://doi.org/10.1017/S1366728917000682

***Thordardottir, E.** (2008). Language-specific effects of task demands on the manifestation of specific language impairment: A comparison of English and Icelandic. *Journal of Speech, Language, and Hearing Research, 51*(4), 922–937. https://doi.org/10.1044/1092-4388(2008/068)

***Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., Trudeau, N., & Chilingaryan, G.** (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech, Language, and Hearing Research, 54*(2), 580–597. https://doi.org/10.1044/1092-4388(2010/09-0196)

**Tipton, E.** (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375–393. https://doi.org/10.1037/met0000011

**Tipton, E., & Pustejovsky, J. E.** (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634. https://doi.org/10.3102/1076998615606099

**Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M.** (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

**Tukey, J. W.** (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley Publishing Company.

**Vevea, J. L., Coburn, K., & Sutton, A.** (2019). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 383–432). Russell Sage Foundation. https://doi.org/10.7758/9781610448864.21

**Viechtbauer, W.** (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

**Westby, C.** (1984). Development of narrative language abilities. In G. Wallach & K. Butler (Eds.), *Language learning disabilities in school-age children* (pp. 103–127). Lippincott Williams & Wilkins.

**Westerveld, M. F., & Moran, C. A.** (2013). Spoken expository discourse of children and adolescents: Retelling versus generation. *Clinical Linguistics & Phonetics, 27*(9), 720–734. https://doi.org/10.3109/02699206.2013.802016

**Wilder, A., & Redmond, S. M.** (2022). The reliability of short conversational language sample measures in children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research, 65*(5), 1939–1955. https://doi.org/10.1044/2022_JSLHR-21-00628

**Wittke, K., & Spaulding, T. J.** (2018). Which preschool children with specific language impairment receive language intervention? *Language, Speech, and Hearing Services in Schools, 49*(1), 59–71. https://doi.org/10.1044/2017_LSHSS-17-0024