JSLHR

Research Article

Bayesian Generalized Linear Mixed-Model Analysis of Language Samples: Detecting Patterns in Expository and Narrative Discourse of Adolescents With Traumatic Brain Injury

Gavin Collins,[a] Jennifer P. Lundine,[b,c] and Eloise Kaizar[a]

Purpose: Generalized linear mixed-model (GLMM) and Bayesian methods together provide a framework capable of handling a wide variety of complex data commonly encountered across the communication sciences. Using language sample analysis, we demonstrate the utility of these methods in answering specific questions regarding the differences between discourse patterns of children who have experienced a traumatic brain injury (TBI), as compared to those with typical development.
Method: Language samples were collected from 55 adolescents ages 13–18 years, five of whom had experienced a TBI. We describe parameters relating to the productivity, syntactic complexity, and lexical diversity of language samples. A Bayesian GLMM is developed for each parameter of interest, relating these parameters to age, sex, prior history (TBI or typical development), and socioeconomic status, as well as the type of discourse sample (compare–contrast, cause–effect, or narrative). Statistical models are thoroughly described.
Results: Comparing the discourse of adolescents with TBI to those with typical development, substantial differences are detected in productivity and lexical diversity, while differences in syntactic complexity are more moderate. Female adolescents exhibited greater syntactic complexity, while male adolescents exhibited greater productivity and lexical diversity. Generally, our models suggest more advanced discourse among adolescents who are older or who have indicators of higher socioeconomic status. Differences relating to lecture type were also detected.
Conclusions: Bayesian and GLMM methods yield more informative and intuitive results than traditional statistical analyses, with a greater degree of confidence in model assumptions. We recommend that these methods be used more widely in language sample analysis.
Supplemental Material: https://doi.org/10.23641/asha.14226959

Language sampling is a method to elicit verbal or written discourse. It provides a natural, contextualized view of an individual's ability to combine multiple utterances to converse, tell a story, persuade a listener, or explain a scientific phenomenon (Nippold, 2014). Language sampling has strong ecological validity, offering students the opportunity to produce multi-utterance discourse passages as they might be expected to do in a conversational or classroom-based task. Language sample analysis (LSA) is the evaluation of the language produced in verbal or written discourse samples. LSA is used to supplement findings from standardized language tests, especially to reduce the potential for cultural/linguistic biases that can exist in these tests (Betz et al., 2013; McCabe & Champion, 2010). LSA also offers valuable guidance for interventions (Nippold, 2014). LSA was included in approximately 25% of all child-focused studies published in American Speech-Language-Hearing Association journals between 2000 and 2011 (Finestack et al., 2014).

Among the many outcome variables that can be extracted from an LSA, we focus on those most commonly studied: microstructural variables evaluating productivity, lexical diversity, and syntactic complexity (Nippold, 2014)

[a]Department of Statistics, The Ohio State University, Columbus
[b]Department of Speech & Hearing Science, The Ohio State University, Columbus
[c]Division of Clinical Therapies and Inpatient Rehabilitation Program, Nationwide Children's Hospital, Columbus, OH
Correspondence to Jennifer P. Lundine: lundine.4@osu.edu

Editor-in-Chief: Stephen M. Camarata
Editor: Julius Fridriksson
Received August 10, 2020
Revision received October 16, 2020
Accepted December 14, 2020
https://doi.org/10.1044/2020_JSLHR-20-00471

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

—each of which will be carefully defined in the methods section. Because of the varied methods of elicitation and large number of outcome variables typically collected (Finestack et al., 2014), LSA provides an excellent example of data for which model-based analyses may provide more accurate research conclusions, compared to traditional statistical methodology. Largely, studies of LSA data have relied heavily on traditional statistical approaches such as *t* tests or analysis of variance (ANOVA) to compare performance between groups of participants (e.g., Chapman et al., 2001; Hay & Moran, 2005; Moran et al., 2012; Scott & Windsor, 2000; Westby et al., 2010). Unfortunately, these methods rely on tenuous assumptions about the data.

LSA data are often counts, which are sometimes transformed to ratios, rates, and proportions, for which discrete probability distributions are more natural than the ANOVA-assumed normal distribution. Continuous numerical covariates such as age and standardized test score are often of interest as they relate to variables collected in LSA, but traditional ANOVA fails to consider such relationships (although simple linear regression [SLR] extensions allow continuous covariates as described by Oleson et al., 2019). Furthermore, in many studies, multiple language samples are taken from each subject, yielding correlated within-subject observations. But analyses based on these basic statistical methods commonly ignore within-subject correlation and miss the opportunity to improve power and interpretability that may better support clinical decision making (Perry & Kucker, 2019). Data produced in LSA effectively demonstrate the utility of generalized linear mixed-model (GLMM) and Bayesian methods in the speech, language, and hearing sciences. Analysis of language samples has traditionally focused on comparing average microstructural summaries across groups, but moving to a model-based analysis frees us to examine a broader class of parameters, which we describe further below.

A Flexible Approach to Statistical Modeling: GLMM and Bayes

Our analytic approach combines three components: regression-style models appropriate for outcomes that are counts of events (rather than continuous measures) called *generalized linear models* (GLMs; Nelder & Wedderburn, 1972), mixed (multilevel) models to account for correlation among multiple outcomes collected from the same individual, and Bayesian inference to make both complex analyses practically feasible and our results more practically interpretable.

GLMs

ANOVA and SLR techniques are naturally used to analyze normally distributed outcomes. GLMs accommodate outcomes with many other probability distributions while preserving the central idea of regression: Variation in observed outcomes is described by a known family of probability distributions, with means (i.e., expected values) that follow a pattern controlled by the covariates. As opposed to SLR, GLM accommodates many probability distributions, including Poisson and Bernoulli distributions for count and binary outcomes. Because these distributions' means may be restricted to certain ranges (e.g., Bernoulli means [probabilities] must be between 0 and 1), linear associations between covariates and means may not make sense. Thus, GLMs specify linear associations between covariates and some function of means, called a *link function*. Logistic regression is a special case of GLM with a binary outcome variable, which has a Bernoulli distribution with mean p and logit link function, that is, $\log\left(\frac{p}{1-p}\right)$.

Mixed Models

The independent observation requirement of ANOVA- and SLR-based statistical inference is often violated in data from language studies (e.g., multiple discourse samples from each participant). Including random effects among the covariates naturally extends SLR to apply to repeated measures on the same set of individuals, as Gordon (2019) clearly describes. Models that incorporate random effects into GLMs are called GLMMs.

Bayesian Inference

We utilize the Bayesian view of probability to directly and naturally learn about our parameters of interest. As opposed to the "long-run frequency" view of classical statistics, the essential tenant of Bayesian statistics is that probability is a degree of belief regarding the value of a population parameter, which is continuously updated as new data are observed. Thus, results of a Bayesian analysis are typically summaries of these updated parameter beliefs, called *posterior distributions*. We refer interested readers to McMillan and Cannon's (2019) excellent introduction for a more thorough explanation, but we briefly note three practical advantages of Bayesian inference. First, it allows researchers to incorporate prior knowledge into their analysis, which gives them the chance to build upon past research in a transparent and intuitive manner. Second, the quantification of post-study beliefs is more natural and informative than p values and confidence intervals. Third, it often facilitates easy estimation of quantities that would be relatively difficult to estimate under the classical paradigm.

Putting all of these ideas together, Bayesian inference based on GLMMs is well suited for LSA. The GLM component allows us to flexibly model data with distributions appropriate for count and categorical variables—and incorporate both continuous explanatory variables such as age, income, and standardized test score, and categorical factors like sex, education level, and race. The mixed-model component provides the opportunity to model correlated observations, seamlessly handle longitudinal data with missing values, and describe underlying patterns via intuitive parameters (Gordon, 2019). Although GLMMs can be used within the traditional frequentist paradigm, the advantages of Bayesian inference are particularly important for the complex and interrelated variables often encountered in speech, language, and hearing research. Recent articles described analyses of speech data that relied on pairs of these ideas (Gordon, 2019; Nalborczyk et al., 2019). We demonstrate

the utility of Bayesian inference with a set of related GLMMs designed to separately model counts, ratios of counts, and proportions. When taken together, our model provides a full picture of the relationship among language constructs and their association with individual characteristics.

Past LSA Research in Youth With Traumatic Brain Injury

A typical goal of LSA is to compare discourse samples produced by a group of children with typical development to those produced by a group of children who are suspected of or diagnosed as having a language or learning difficulty, such as developmental language disorder or traumatic brain injury (TBI). Discourse samples of at-risk groups of individuals may show subtle but clinically important differences when compared to peers with typical development. For example, students with language and learning disorders may exhibit decreased use of complex syntax and vocabulary or use less language to converse or tell a story than age-matched peers with typical development (e.g., Hay & Moran, 2005; Nippold et al., 2008; Scott & Windsor, 2000; Ward-Lonergan et al., 1999), but these findings have not held across all studies (e.g., Moran et al., 2012; Turkstra & Holland, 1998). Children with TBI often perform well on standardized tests of language but struggle in real-world contexts where they must incorporate higher level language and cognitive skills into a discourse sample (Coelho et al., 1991). Youth with TBI exhibit a large amount of within-subject and between-subjects variability in their performance on cognitive and language tasks, which frequently limits the generalizability of research findings. Between-subjects variability is likely due to a combination of many factors, including, but not limited to, socioeconomic status (SES), severity of injury, age at time of injury, time since injury, and preinjury cognitive-communication characteristics (e.g., Anderson et al., 2000; Catroppa et al., 2016; Durber et al., 2017; Rashid et al., 2014; Ryan et al., 2014; Yeates et al., 2002). Although it is clear that more advanced statistical analyses may help to identify subtle differences between discourse produced by youth with and without TBI, *t* tests and ANOVA are the most commonly used statistical methodologies employed in past studies examining language samples from children and adolescents with TBI (e.g., Aguilar et al., 2018; Chapman et al., 1992; Hay & Moran, 2005; Hemphill et al., 1994; Jordan & Murdoch, 1994; Turkstra & Holland, 1998; Walz et al., 2012).

This article utilizes GLMM and Bayesian methods to learn about the productivity, lexical diversity, and syntactic complexity of expository and narrative discourse produced by adolescents with TBI compared to their peers with typical development. In a reanalysis of results previously published by Lundine and Barron (2019), this article analyzes similar questions as the earlier article but with useful advancements in statistical methodology. In the Method section, we first describe our data collection and then outline three statistical models, which separately address patterns relating to the productivity, syntactic complexity, and lexical diversity of collected discourse samples. These models include parameters that relate these microstructural discourse characteristics to certain demographic characteristics and to the type of lecture being summarized. Productivity, syntactic complexity, and lexical diversity are examined using count data, ratios of counts, and proportions, respectively. In the Results section, we summarize our findings and comment on the outcomes, including a comparison to results reported in the original Lundine and Barron (2019) article. Finally, we discuss the effect of using the prescribed methods and relate our findings to the broader literature.

Method

Data Collection

Data for this study were collected as part of a larger study, previously described by Lundine, Harnish, McCauley, Blackett, et al. (2018) and Lundine, Harnish, McCauley, Zezinka, et al. (2018). All necessary review boards approved the study before the time of data collection, and subjects and/or their parents signed assent/consent forms before participating.

Briefly, participants consisted of 55 adolescents, ages 13–18 years (average 15.5), 52% female, of varying SES (as measured by census-tract data), each of whom spoke English as the primary language in their home. Five adolescents had experienced a moderate-to-severe closed head injury (< 12 on the Glasgow Coma Scale; Teasdale & Jennett, 1974) at the age of 9 years or older, after completing the fourth grade, and at least 9 months prior to participation. Adolescents with TBI were excluded if child abuse was documented as the cause of the injury, if there was any history of developmental delay, autism, or substantial neurological disorder prior to the injury, or if there existed any substantial motor, language, or speech impairments that would prohibit successful completion of the required tasks. The remaining 50 subjects had a history of typical development (as documented by parent report and tests of cognition and expressive syntax). The primary goal of the larger study was to discover how three types of discourse summaries (i.e., compare–contrast [CC], cause–effect [CE], narrative [N]) differed by developmental history (TBI or typical development), sex, SES, and age. For our analyses, we centered continuous explanatory variables at their mean and coded binary variables with $\pm$ 0.5, where female sex and typical development were designated the comparison groups (coded −0.5).

Subjects verbally summarized three short lectures about the fictional nation of "Lifeland." The three lectures were presented by video in a random order on a computer monitor. Immediately following each lecture, participants were asked to summarize the information they heard and then participate in standardized cognitive/expressive syntax tests as part of the larger protocol (data from these tests are not analyzed here; for details, see Lundine, Harnish, McCauley, Blackett, et al., 2018; Lundine, Harnish, McCauley, Zezinka, et al., 2018). Each of the three lecture stimuli contained approximately the same number of words, sentences, and main and supporting ideas and were written

at approximately the same reading level. Each lecture was read by the same speaker, in front of the same neutral background. Two of the lectures (CC and CE) were expository, while the third (N) was narrative. Secondary interest lies in identifying differences in microstructural discourse patterns across the different lecture types.

Each of the 165 total discourse samples (55 adolescents × 3 summaries) was audio- and video-recorded and was later transcribed using Systematic Analysis of Language Transcripts software (Miller & Iglesius, 2010), which was also used to extract the microstructural discourse variables of interest. Following the initial coding of transcripts, coders reanalyzed 20% of the transcripts for intra- and interrater reliability checks. Transcription reliability was > 95% for point-to-point comparisons across all three types of summaries.

As depicted in Figure 1, each discourse summary contains a number of utterances, each utterance consists of an independent clause and its accompanying dependent clauses (Hunt, 1965), and each clause is made up of a number of words. With this structure in mind, for each discourse (j = CC, CE, N) delivered by each participant (i = 1, …, 55), we recorded four statistics that together provide an effective summary of the microstructural characteristics of the discourse: the total number of utterances (denoted U_{ij}), clauses (denoted C_{ij}), words (denoted W_{ij}), and distinct words (denoted D_{ij}) spoken in the discourse summary. These summary statistics, along with demographic data for each of the subjects in this study, can be found on Dryad at https://doi.org/10.5061/dryad.v15dv41v8. We use these four summary statistics to characterize patterns relating to the productivity, syntactic complexity, and lexical diversity of the study participants. Corresponding features for the population (adolescents living in central Ohio) are described by parameters. To help shift analytical thinking from the familiar data summaries to a direct focus on relevant model parameters, we list our proposed parameter (described in more detail below) for each of the corresponding language constructs in Table 1. While respecting the interconnectedness of the four microstructural variables of interest, we separately consider appropriate models for each in the following subsections.

Productivity (Analyzing Count Data)

Broadly defined, productivity is the amount of language produced by a subject during a discourse task, usually measured in words or utterances. Note that productivity is *not* necessarily a measure of the amount of *information* provided by the discourse; for example, one might repeat ideas, words, or phrases several times without conveying additional information, but a discourse sample is considered to be highly productive if a copious amount of language is produced, regardless of the content of that language.

The average number of utterances per discourse sample has often been used to characterize productivity (e.g., Hay & Moran, 2005; Scott & Windsor, 2000; Ward-Lonergan et al., 1999). By averaging the number of utterances within groups, one can estimate the group-specific population mean. We define a parallel parameter, termed the *rate of utterances per discourse* (RUD), which is the expected number (or rate) of utterances per discourse sample. We use RUD_{ij} to represent this true (unknown) mean for discourse samples given by subject i summarizing a lecture of type j. Note that each data value U_{ij} is an estimate of the true productivity, RUD_{ij}, but is measured with some variability. Rather than averaging U_{ij} within groups, we use a GLMM to describe patterns of RUD.

Because U_{ij} is a count type variable, it would be inaccurate to assume it follows a normal distribution with mean RUD_{ij}. Instead, it is natural to assume a Poisson distribution, shifted to account for the fact that each discourse

Figure 1. Microstructure for an example hypothetical discourse summary recorded for some generic participant i in relation to lecture type j. The hypothetical numbers of characteristics (utterances, clauses, and words) were chosen to ease visualization and are smaller than typical discourse summaries in our sample.

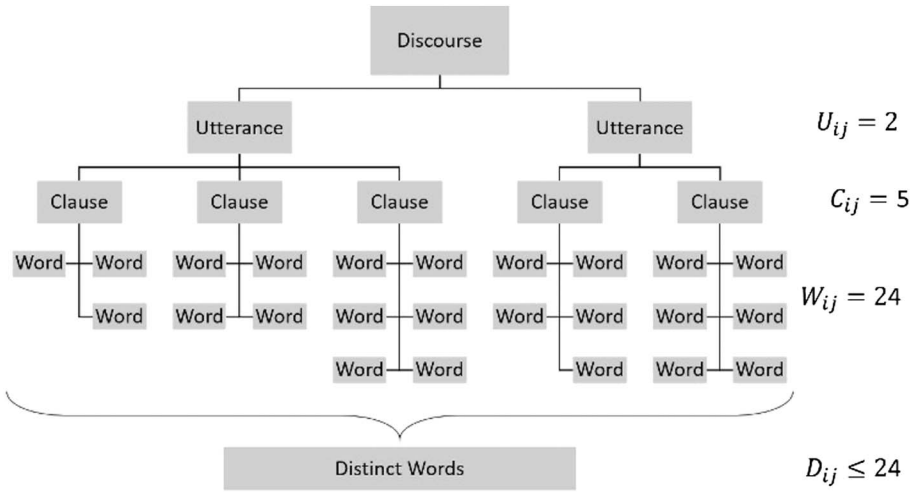


Table 1. Corresponding concepts, summary statistics, and model parameters.

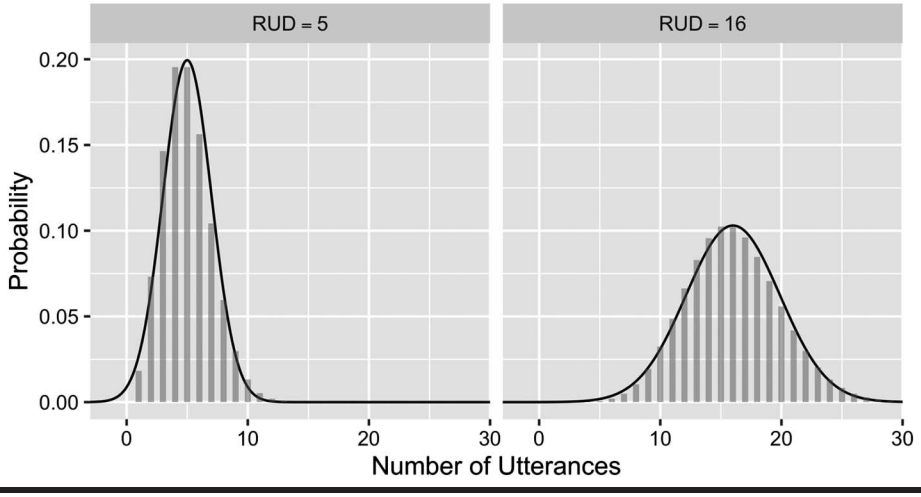
Language construct	Traditional summary statistic	Model parameter
Productivity	Total number of utterances U_{ij}	Rate of utterances per discourse (RUD) RUD_{ij}
Syntactic complexity	Subordination index (SI) $\frac{C_{ij}}{U_{ij}}$	Rate of subordination (RS) RS_{ij}
	Mean length of utterance (MLU) $\frac{W_{ij}}{U_{ij}}$	Rate of words per utterance (RWU) RWU_{ij}
Lexical diversity	Type token ratio (TTR) $\frac{D_{ij}}{W_{ij}}$	Probability of a distinct word (PDW) PDW_{ij}
	Distinct words per utterance $\frac{D_{ij}}{U_{ij}}$	Rate of distinct words per utterance (DWU) $\mathrm{DWU}_{ij} = \mathrm{RWU}_{ij} \times \mathrm{PDW}_{ij}$

summary must contain at least one utterance. That is, we consider the number of utterances beyond the first one ($U_{ij} - 1$) or supplemental utterances to follow a Poisson distribution with mean $\mathrm{RUD}_{ij} - 1$; we term this adjusted mean the rate of supplemental utterances per discourse (sRUD). This implies that $U_{ij} - 1$ has mean $\mathrm{RUD}_{ij} - 1$, and U_{ij} has mean RUD_{ij}, as desired.

Figure 2 illustrates some key advantages of the proposed model, as compared to traditional ANOVA assumptions. The gray bars represent the shifted Poisson distribution for $\mathrm{RUD}_{ij} = 5$ and 16 (approximately the 10th and 90th percentiles, respectively, for typically developing [TD] subjects in our study). For example, according to the Poisson model with $\mathrm{RUD}_{ij} = 5$, the probability of actually observing five utterances is about 20%, and it would be impossible to observe 0 utterances. The solid black lines are normal distribution approximations to the same shifted Poisson distributions. While the shape of the normal distribution is similar, it implies that fractions of utterances, zero utterances, and even negative numbers of utterances are possible. The difference in the spread of the distributions across the two panels highlights an additional inconsistency: The usual ANOVA assumption is that the variance is constant regardless of the mean value, but in reality, the variance often increases with the mean for count data, which is captured by the Poisson distribution.

Our goal is not simply to estimate each participant- and lecture-specific productivity RUD_{ij}, but rather to discover broader patterns concerning the manner in which RUD_{ij} differs across demographic backgrounds, lecture type, and language/learning profile. This goal may be accomplished using Poisson regression: a GLMM for count data that includes both fixed and random effects (see Agresti & Kateri, 2011). In Poisson regression, we use a linear function, similar to an ordinary regression model that relates true mean productivity RUD_{ij} to the fixed effects: discourse type and demographic characteristics. Because we recorded multiple measures on individual subjects, we also include a random intercept for each one. Finally, to reflect the impossibility that the mean number of utterances (i.e., RUD_{ij}) is less than 1, we set the log transform of $\mathrm{RUD}_{ij} - 1$ equal to a linear

Figure 2. Distribution of utterances with a mean of 5, the 10th percentile for the control participants, and a mean of 16, the 90th percentile for the control participants. The gray bars represent the probability of observing that number of utterances if the shifted Poisson model is correct; the solid curved line is the corresponding approximate normal distribution. RUD = rate of utterances per discourse.



combination of explanatory variables. This completes our regression equation. The log transformation results in a slightly more complex interpretation of the regression coefficients, similar to logistic regression, but our use of Bayesian methods allows us to transform these coefficients into the more familiar additive effects. See Supplemental Material S1 for further details.

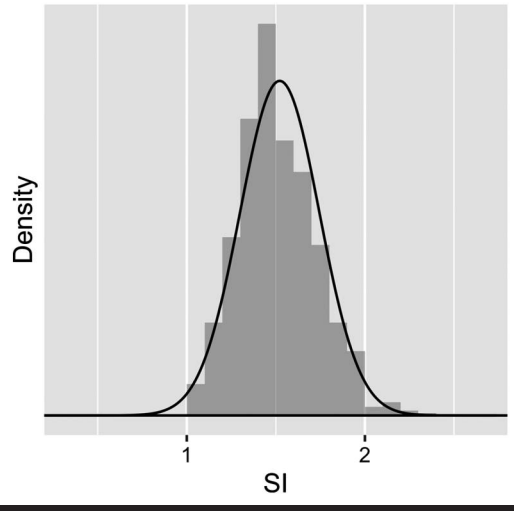
Compared to a traditional ANOVA approach, the GLMM framework allows us to be far more flexible, accurate, and complete in describing our data, while allowing us to answer important questions related to interesting parameters. In the next subsection, we build upon the methods just described, utilizing Poisson regression to model ratios of counts.

Syntactic Complexity (Analyzing Ratios of Counts)

Syntactic complexity is a measure of the complexity of individual utterances in a discourse passage. In past research, it has typically been measured using subordination index (SI) and/or mean length of utterance. We first construct a model to help us describe subordination and then adapt it to model mean length of utterance. The SI (see Loban, 1976) characterizes complexity by dividing the total number of clauses by the total number of utterances, that is, $SI_{ij} = C_{ij}/U_{ij}$. We define a corresponding population's rate of subordination (RS) to be the expected (i.e., mean) number of clauses in a single utterance, denoted RS_{ij} for subject i summarizing lecture j. In parallel to our productivity parameters, we are also interested in the expected number of supplemental clauses past the first one, or the supplemental RS, denoted $RS_{ij} - 1$.

Unfortunately, SI is also unlikely to follow a normal distribution. First, because each utterance contains at least one clause, we have the constraint that $C_{ij} \geq U_{ij}$, implying that $SI_{ij} = \frac{C_{ij}}{U_{ij}} \geq 1$, while the normal distribution incorporates no such restriction. Second, a ratio of two positive counts typically has a skewed distribution—particularly if the numerator and denominator are correlated—while the normal distribution is symmetrical. In Figure 3, we visualize one possible distribution of SI for rates of utterances and clauses that are typical for our data set. Note that the normal distribution approximation does not capture the lower bound or slight left skew of the true distribution. Analyses related to the second, length-based method of characterizing syntactic complexity (mean length of utterance) also typically rely on a calculated ratio of words per utterance, which suffers from the same normal approximation limitations.

Figure 3. Distribution of subordination index (SI), based on uncorrelated rates of 11.5 utterances per discourse and 1.52 clauses per utterance (gray histogram, based on 1 million simulated discourse summaries) and a normal approximation (black line, based on the simulated sample mean and sample standard deviation).



Recognizing our goal to describe the underlying true RS, we model the total count of clauses in each discourse, again via Poisson regression, and use the parameters in this model to describe the relationship between RS and the explanatory variables. We again enforce the restriction that each utterance has at least one clause by modeling the total additional clauses after the first in each utterance according to a Poisson distribution with mean $U_i \times (RS_{ij} - 1)$. We can then complete our model with a regression-like equation to describe the association between the rate of additional clauses, $RS_{ij} - 1$, and the covariates of interest, again including a random intercept. Further details can be found in Supplemental Material S1.

Mean length of utterance, that is, the average number of words per utterance (W_{ij}/U_{ij}), is also used to characterize syntactic complexity. A natural corresponding population quantity is the expected number or rate of words per utterance (RWU), denoted $RWU_{ij}|$. We estimate $RWU_{ij}|$ using a model similar to our model for RS_{ij}, except we model the count of supplemental words instead of the count of supplemental clauses.

Lexical Diversity (Analyzing Proportions)

Lexical diversity is the variety of different words used in a discourse sample. Type–token ratio (TTR; a ratio of the number of unique words to total words) is one of the most commonly used measures of lexical diversity, but its usefulness is flawed because of its dependence on sample length (Fergadiotis et al., 2015; Richards, 1987). Because the raw total number of different words (D_{ij}) would clearly be highly correlated with the length of the discourse sample, researchers have proposed other approaches to measure lexical diversity that control for the length of the sample in different ways. Proposed methods include the D statistic (e.g., Jacobson & Walden, 2013; Owen & Leonard, 2002), moving-average TTR (e.g., Charest et al., 2020), and number of different word ratio (e.g., Greenhalgh & Strong, 2001; Lundine & Barron, 2019; Mills et al., 2013), that is, the sample ratio $\frac{D_{ij}}{U_{ij}}$. We believe that adjusting by the number of utterances conflates lexical diversity and syntactic complexity, since it directly depends on the number of observed words per utterance via the equation $\frac{D_{ij}}{U_{ij}} = \frac{D_{ij}}{W_{ij}} \frac{W_{ij}}{U_{ij}}$. Thus, we prefer to

quantify lexical diversity similarly to TTR and define the population probability that any word used in a discourse is distinct, which we term the *probability of a distinct word* (PDW) and denote PDW_{ij}. If we wish, we can mirror the traditional construct by taking the product of the PDW and RWU since $\frac{D_{ij}}{U_{ij}}$ is simply an estimate of $\text{PDW}_{ij} \times \text{RWU}_{ij}$, observed with some measurement variability.

Our next challenge is to describe how PDW_{ij} varies for different lecture types, for people of different demographic backgrounds, and for lectures of varying lengths. To do so, we again implement a GLMM, but instead of Poisson regression, we use binomial regression, which is another type of GLMM. Although D_{ij} is a count variable, it differs from the other count variables we have modeled thus far because D_{ij} may never be larger than the total word count W_{ij} (although because words are not chosen independently, we recognize the binomial distribution is only an approximation). In such a case, the binomial distribution is more appropriate than the Poisson distribution. In order to put PDW_{ij} on the proper scale for binomial regression, we transform it using the logit function, $\text{logit}(\text{PDW}) = \log\left(\frac{\text{PDW}}{1-\text{PDW}}\right)$, similar to how the log transformation is used in Poisson regression. Parameter interpretation for our model is identical to logistic regression, which also relies on the logit link function but models strictly binary outcomes (i.e., a count with a maximum of 1; see Agresti & Kateri, 2011, for more details on logistic regression).

To adjust for the dependence of PDW_{ij} on sample length, we also include a term involving W_{ij} in the regression equation. See Supplemental Material S1 for further details.

Fitting the Model (Bayesian Analysis)

As previously described, there are several key advantages of using a Bayesian approach. One is that Bayesian methods allow researchers to incorporate their prestudy beliefs about reasonable parameter values into the analysis, including beliefs that have been obtained from past studies and literature reviews, for example (see McMillan & Cannon, 2019). Of course, researchers need to be very thoughtful about their choice of such prior distributions and need to communicate the reasoning behind their choices openly and clearly. Choice of prior distributions should not be influenced by the data. Because this is a reanalysis of data originally investigated by one of the authors, her choice of subjective priors at this point may be subconsciously swayed by the previous analysis. Instead, we elected to use diffuse priors, which represent a very wide range of possible values for the parameters of interest, mimicking a naïve reader's opinions (McMillan & Cannon, 2019). These prior intervals implied the prior distributions for each of the coefficients relating the microstructural language constructs to the demographic variables of interest and also to the overall mean for each construct. For a detailed list of these prior distributions, see Supplemental Material S1.

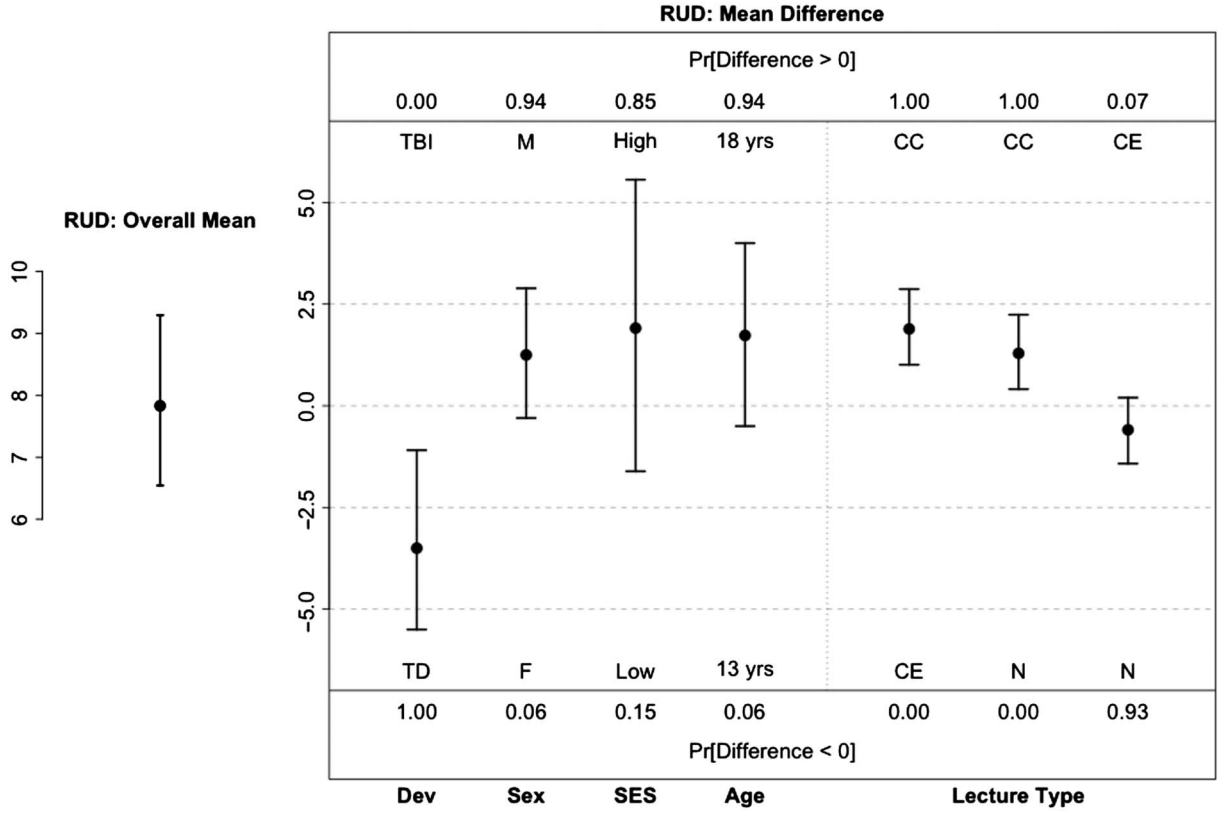
Using a Markov chain Monte Carlo algorithm implemented via STAN software (Stan Development Team, 2018b) accessed via the RStudio interface (R Core Team, 2019; R Studio Team, 2020; Stan Development Team, 2018a), we obtained three independent chains of 10,000 samples each (for a total of 30,000 samples) from the posterior distributions of all parameters in our models. In particular, we display estimates and credible intervals for each parameter of interest, including the overall mean of the core parameters RUD, RS, RWU, and PDW, along with mean differences between demographic groups corresponding to development type (TBI vs. TD), sex (male vs. female), SES (high vs. low), and age (18 vs. 13 years) and between discourses on the three different lecture types (CC, CE, and N) for each of these core parameters. Finally, we calculate a Bayesian p value, which is the estimated probability that a parameter exceeds a value of interest—in our case, 0—for each of the mean difference parameters. These posterior summaries are based on the available data, combined with our diffuse prior beliefs and the model assumptions described in this section. They are valid measures of our posterior beliefs insomuch as the model assumptions are met and insomuch as data have been collected from the desired population of interest. As far as we can discern, this is the case with our data.

Results

Figures 4–7 display posterior estimates, 95% credible intervals, and Bayesian p values for all RUD-, RS-, RWU-, and PDW-related parameters, respectively. Posterior estimates, along with prior and posterior credible intervals for the demographic variables, are also listed in Table 2. Note that the prior intervals are considerably wider than the posterior intervals, indicating that the data have substantially increased our certainty about the parameters of interest. Additionally, Table 3 lists estimates and posterior credible intervals for the parameters relating to lecture type.

Based on our data, combined with our prior beliefs and the model described in the preceding section, we estimate that the overall mean number of utterances per discourse (RUD), across all lecture types and demographic groups for 13- to 18-year-old adolescents in the central Ohio area, is about 7.83 and that there is a 95% probability that overall mean RUD is truly between 6.54 and 9.29 utterances (see Figure 4 and Table 2). The average difference between the RUD of individuals with TBI and TD individuals is estimated to be about −3.50 utterances, with a 95% credible interval ranging from −5.50 to −1.09 utterances (see Figure 4 and Table 2). The Bayesian p value for this association is approximately 0 (see Figure 4), which implies an almost certain posterior belief that mean RUD for individuals with TBI is lower than the mean for TD individuals. We also estimate important differences in RUD production for other demographic categorizations. We estimate that mean RUD is 1.25 higher for male adolescents than for female adolescents; 1.91 higher for those with the highest SES, as compared to those with the lowest SES; and 1.73 higher for 18-year-olds than for 13-year-olds. However, much weaker evidence for these associations in the data results in relatively wide credible intervals that indicate nonnegligible posterior probability that the true effects are reversed to

Figure 4. Estimate (dot) and 95% credible interval for the mean rate of utterances per discourse (RUD), summary (left), and for the mean difference between the RUD of contrasting demographic groups and of various discourse summaries for each of the three lecture types (right). Note that *p* values are rounded estimates and are never exactly 0 or 1. Dev = development; SES = socioeconomic status; TBI = traumatic brain injury; TD = typical development; M = male; F = female; CC = compare–contrast; CE = cause–effect; N = narrative.

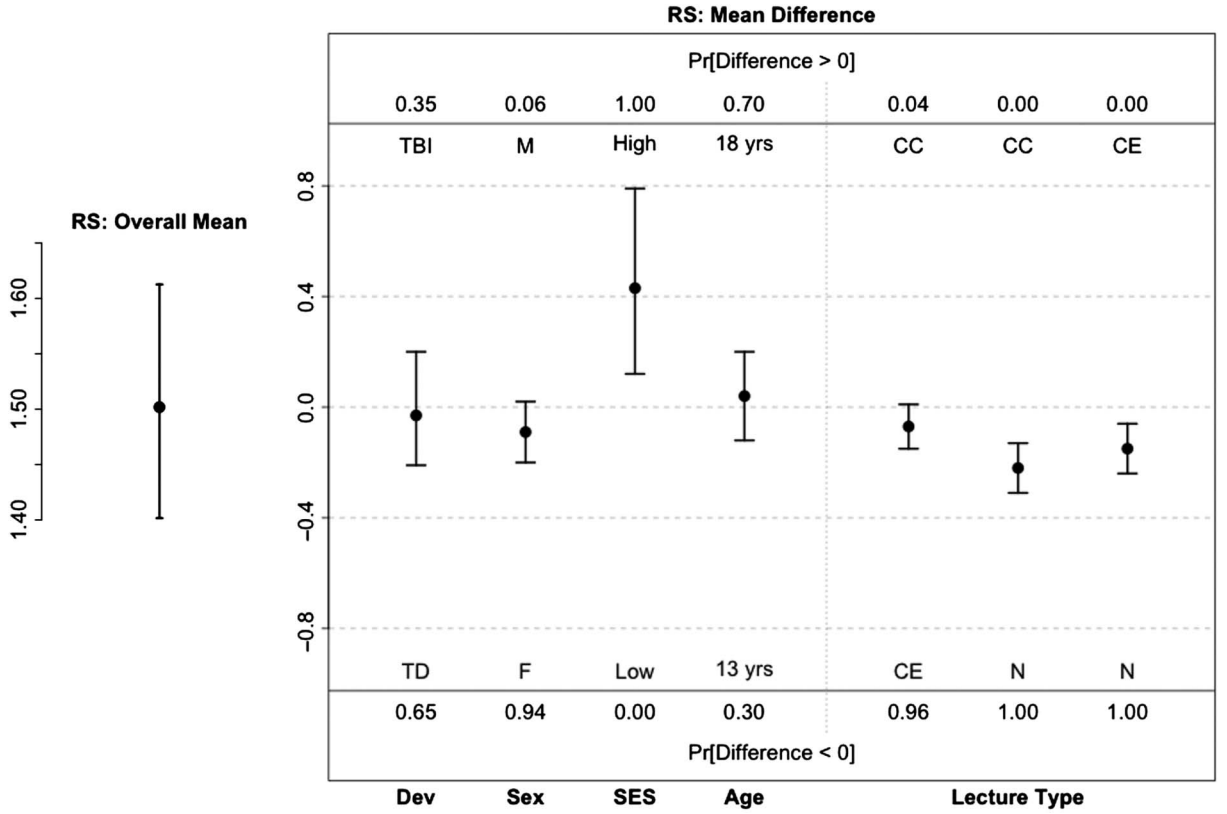


some degree or are actually stronger than estimated (see Figure 4 and Table 2). Based on a Bayesian *p* value of nearly zero (see Figure 4), we also conclude, with a high degree of confidence, that summaries of the CC lecture tend to include the most utterances per lecture, with 1.89 (95% CI [1.01, 2.87]) more utterances than summaries of the CE lecture and 1.29 (95% CI [0.41, 2.24]) more than summaries of the N lecture (as shown in Figure 4 and reported in Table 3).

Turning to syntactic complexity, average RS for individuals in the population of interest is around 1.50 clauses per utterance (95% CI [1.40, 1.61]). The analysis shows little evidence of a large difference in RS across development groups. We estimate that the RS of individuals with TBI is only 0.03 lower than their TD peers, on average, with a 95% credible interval ranging from −0.21 clauses, indicating a moderate reduction in RS for individuals with TBI, to 0.20 clauses, indicating the possibility that individuals with TBI may in reality produce more clauses per utterance than TD individuals. We also estimate small-to-null associations with sex and age. Specifically, we estimate that male adolescents produce 0.09 fewer clauses per utterance than female adolescents, with a Bayesian *p* value of .06 (see Figure 5) and credible interval of −0.20 to 0.02 clauses. We estimate that the mean RS for 18-year-olds is 0.04 higher than for 13-year-olds (Bayesian *p* value = .30). We estimate a much larger association, however, for SES. We estimate that those with the highest SES have a mean RS that is about 0.43 clauses higher than those with the lowest SES, and according to the credible interval, we believe with 95% probability that this difference is actually somewhere between 0.12 (relatively small) and 0.79 (substantial) clauses. As for the lecture type effects, based on the Bayesian *p* values in Figure 5, we estimate nearly 100% probability that the N lecture prompts the most clauses per utterance and 96% probability that the CC lecture prompts the fewest.

We estimate that mean RWU for this population is around 11.80 words per utterance (95% CI [10.66, 12.99]). We estimate that individuals with TBI have an RWU that is 1.60 lower than TD individuals, on average, but the 95% credible interval ranges from −3.67 to 0.68, indicating some probability that this difference may be even larger than estimated or even that individuals with TBI may produce slightly more words per utterance. The largest estimated effect is for SES, with high-SES individuals averaging an RWU that is about 4.23 words higher than low-SES individuals (95% CI [0.94, 7.62]). The patterns by sex and age are relatively moderate, but we estimate a 90% probability that female adolescents have higher RWU and an 80%

Figure 5. Estimate (dot) and 95% credible interval for the mean rate of subordination (RS; left) and for the mean difference between the RS of contrasting demographic groups and of various discourse summaries for each of the three lecture types (right). Note that *p* values are rounded estimates and are never exactly 0 or 1. Dev = development; SES = socioeconomic status; TBI = traumatic brain injury; TD = typically development; M = male; F = female; CC = compare–contrast; CE = cause–effect; N = narrative.



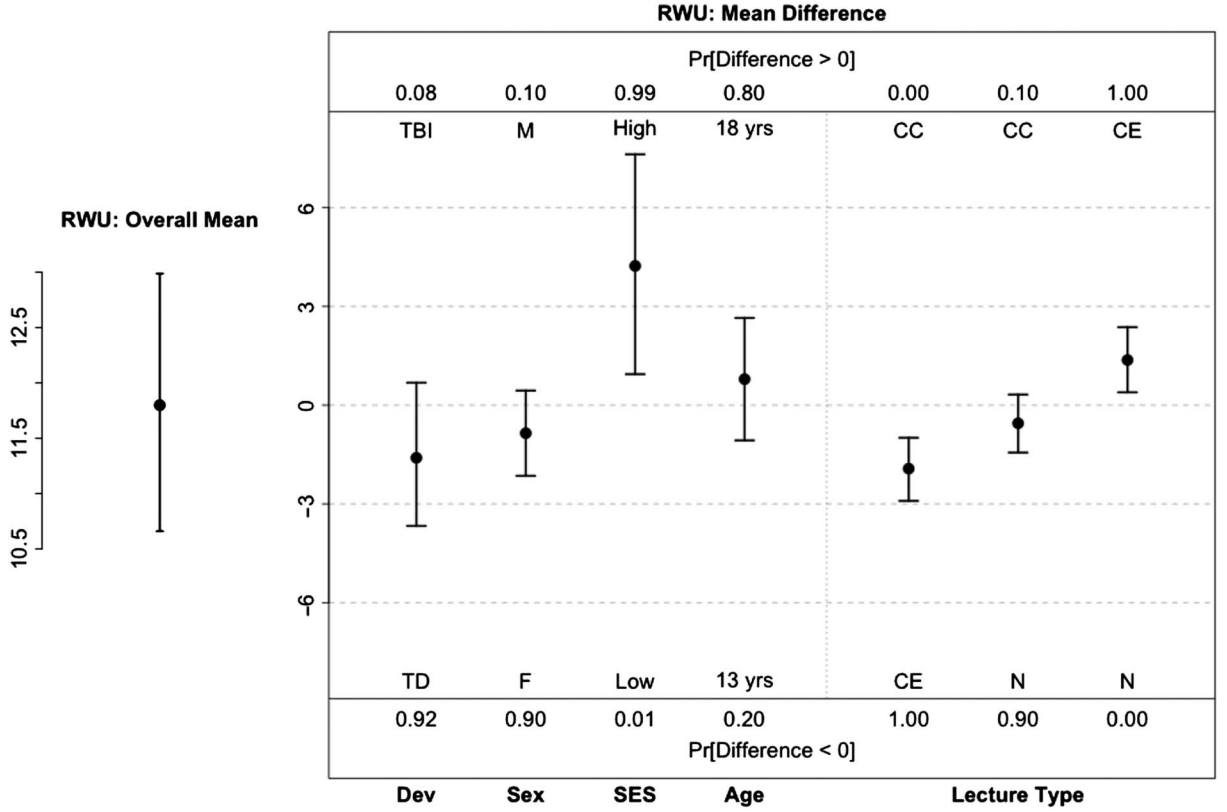
probability that older children have higher RWU, on average (see Table 2 and Figure 6). We also believe, with approximately 100% probability, that the CE lecture leads to the largest mean RWU and, with 90% probability, that the CC lecture leads to the smallest.

The final portion of our analysis deals with lexical diversity by means of the PDW parameter. As previously noted, because we expected PDW to decrease with increases in the number of words per discourse summary (W_i), we included the number of words as a covariate. Unsurprisingly, our adjusted models indicate that this effect is likely quite large. As a baseline, when $W_i = 100$, we estimate that the average proportion of words in a discourse summary that are distinct is about 0.58 (95% CI [0.56, 0.60]), but when comparing a very long discourse (e.g., $W_i = 500$) to a very short discourse (e.g., $W_i = 10$), we estimate that overall average PDW is about 0.32 (95% CI [0.28, 0.35]) for the long discourse and about 0.87 (95% CI [0.85, 0.89]) for the short discourse—a substantial difference. To illustrate the utility of including this effect in our model, Figure 8 displays the predicted average PDW across a large spectrum of W_i for individuals in four distinct subpopulations of 15-year-old female adolescents: low SES/TBI, high SES/TBI, low SES/TD, and high SES/TD. For each of these predictions, we assume that the subjects are giving a discourse summary of the CC lecture.

Demographic variables and lecture type also seemed to have a substantial association with trends of PDW (see Tables 2 and 3 and Figure 7). On average, we estimate that PDW is about 0.04 lower for individuals with TBI, as compared to their TD counterparts, with a 95% credible interval ranging from −0.08 to −0.01. We also estimate that the male adolescents have a PDW that is 0.03 higher than the female adolescents, on average (95% CI [0.01, 0.05]), and that those with the highest SES have an average PDW 0.07 higher than those with the lowest SES (95% CI [0.02, 0.11]). The estimated effect for age is 0.01, with a credible interval ranging from −0.02 (in favor of younger children having higher PDW) to 0.03 (in favor of older children having higher PDW). Finally, we estimate a 98% probability that PDW is highest for summaries of the CC lecture and a 79% probability that PDW is lowest for summaries of the CE lecture.

Results for lexical diversity would have been markedly different had we used the rate of distinct words per utterance (DWU) corresponding to the construct $\frac{D_i}{U_i}$, instead

Figure 6. Estimate (dot) and 95% credible interval for the mean rate of words per utterance (RWU; left) and for the mean difference between the RWU of contrasting demographic groups and of various discourse summaries for each of the three lecture types (right). Note that *p* values are rounded estimates and are never exactly 0 or 1. Dev = development; SES = socioeconomic status; TBI = traumatic brain injury; TD = typically development; M = male; F = female; CC = compare–contrast; CE = cause–effect; N = narrative.



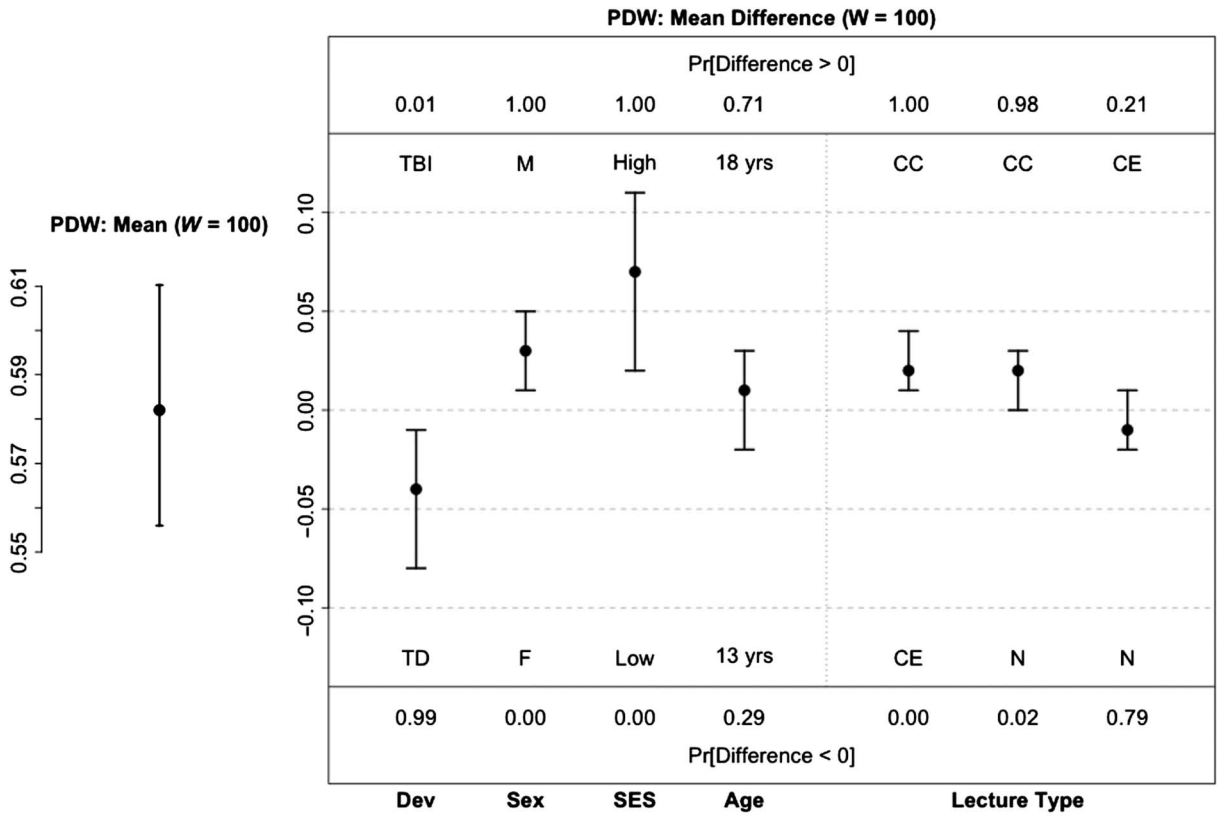
of using PDW as we did. Table 4 shows the probability that each of the demographic and lecture-type contrasts exceeds zero for relevant average RWU, DWU, and PDW. As noted previously, RWU is a measure of syntactic complexity, PDW is a measure of lexical diversity, and DWU conflates the two constructs. In some cases, this conflation still leads to similar conclusions for the DWU and PDW models, but in other cases, conclusions are much different. In particular, posterior probabilities are wildly different for the lecture-type effects, leading to completely opposite conclusions. For example, as compared to the N lecture, summaries of the CC lecture have clearly lower lexical diversity according to PDW and clearly higher syntactic complexity according to RWU, but when these two are conflated via DWU, there is no clear microstructural difference between the two discourse types. Similarly, the association of sex with PDW is strong and with RWU is moderate, but these opposite associations are neutralized in the conflated DWU; conclusions for the other effects happen to be largely unchanged for this example. Note that such construct comparisons would be much more difficult without the Bayesian approach to analysis.

We evaluated model fit via posterior predictive checks as explained in Supplemental Material S1 and found that our models largely fit the data well. The single exception is that we underestimate average mean length of utterance, likely due to inadequately modeling a few unusually large word counts. While we might marginally improve model fit via further fine-tuning of our model, for example, by including a greater number of covariates in the RWU model, we do not expect such adjustments to change our overall conclusions. We expect that traditional ANOVA of these data would suffer from similar issues, and we remain confident in our assertion that the Bayesian GLM is the preferred approach.

Discussion

The analyses introduced in this article offer a more accurate and flexible Bayesian and GLMM methodology as an alternative to the usual LSA methodology that employs ANOVA and *t* tests. The accuracy gained with the proposed methodology comes from making more realistic assumptions about the data, accounting for the fact that most data obtained from language samples are in the form of counts, recognizing the interrelated nature of these count-type variables, incorporating realistic subjective prior beliefs about the data, and also accounting for the case of repeated observations

Figure 7. Estimate (dot) and 95% credible interval for the mean probability of distinct word (PDW; left) and for the mean difference between the PDW of contrasting demographic groups and of various discourse summaries for each of the three lecture types (right). Note that *p* values are rounded estimates and are never exactly 0 or 1. Dev = development; SES = socioeconomic status; TBI = traumatic brain injury; TD = typically development; M = male; F = female; CC = compare–contrast; CE = cause–effect; N = narrative.



for a set of subjects. The models proposed are more flexible because they appropriately incorporate continuous covariates such as age, test scores, and continuous measures of SES. In addition, Bayesian methods allow us to estimate interpretable parameters that would be difficult to estimate (and compare) using classical methods, as well as posterior probabilities and credible intervals, which are more easily interpreted compared to their classical counterparts, *p* values, and confidence intervals.

We also introduce new terminology to describe some of the key parameters of language sampling. The purpose of this is to shift thinking away from sample statistics related to a single data set to population parameters. In this light, we believe researchers can begin to place greater value on flexible model-based statistical methods, enabling them to draw more accurate conclusions from their analyses and investigate a broader scope of scientific inquiry.

In this study, our primary interest is to learn how the language samples of adolescents with TBI in central Ohio tend to differ from the language samples of their peers with typical development. In summary, the main differences identified using the proposed analyses are that individuals with TBI seem to exhibit less productivity and less lexical diversity, on average. We estimate that the mean difference between the RUD of adolescents with TBI and TD adolescents is about 3.50 (95% CI [1.09, 5.50]) utterances per discourse and that the mean difference for PDW is about 4% (95% CI [1%, 8%]). We also estimate a very moderate 65% probability that RS is lower for individuals with TBI and a stronger 92% probability that RWU is lower, indicating some probability that syntactic complexity may also be lower for the group of adolescents with TBI, but more data are needed to support this hypothesis.

The between-groups differences found using the proposed methodology confirm all of the low *p*-value microstructural results obtained by Lundine and Barron (2019), who analyzed the same data set using a matched-pairs approach. In particular, the analyses presented in this article reach the common conclusion that adolescents with TBI exhibit less productivity and syntactic complexity (i.e., mean length of C-unit) for summaries of some discourse types (CC and CE for productivity and N for syntactic complexity). But the Lundine and Barron analysis did not detect the substantial differences in lexical diversity revealed by the current analysis. Lundine and Barron use a sound traditional approach, but we propose that these traditional results differ from those in this article for several reasons. First, because the analysis in this article favored a model-based covariate

Table 2. Model estimates along with 95% prior and posterior credible intervals for the overall mean rate of utterances per discourse (RUD), rate of subordination (RS), rate of words per utterance (RWU), and probability of a distinct word (PDW) and for each of the corresponding demographic parameters of interest, including effects for group, sex, socioeconomic status (SES), and age.

Language construct	Model parameters	Estimate	Prior ➔ posterior 95% credible interval
Productivity: RUD	Overall mean	7.83	[1.5, 48] ➔ [6.54, 9.29]
	Development: TBI vs. TD	–3.5	[–13, 13] ➔ [–5.50, –1.09]
	Sex: male vs. female	1.25	[–13, 13] ➔ [–0.30, 2.89]
	SES: highest vs. lowest	1.91	[–14, 14] ➔ [–1.61, 5.56]
	Age: 18 vs. 13 years	1.73	[–14, 14] ➔ [–0.50, 4.00]
Syntactic complexity: RS	Overall mean	1.5	[1, 5] ➔ [1.40, 1.61]
	Development: TBI vs. TD	–0.03	[–6, 6] ➔ [–0.21, 0.20]
	Sex: male vs. female	–0.09	[–6, 6] ➔ [–0.20, 0.02]
	SES: highest vs. lowest	0.43	[–6, 6] ➔ [0.12, 0.79]
	Age: 18 vs. 13 years	0.04	[–6, 6] ➔ [–0.12, 0.20]
Syntactic complexity: RWU	Overall mean	11.8	[5, 20] ➔ [10.66, 12.99]
	Development: TBI vs. TD	–1.6	[–20, 20] ➔ [–3.67, 0.68]
	Sex: male vs. female	–0.85	[–22, 22] ➔ [–2.15, 0.44]
	SES: highest vs. lowest	4.23	[–20, 20] ➔ [0.94, 7.62]
	Age: 18 vs. 13 years	0.79	[–20, 20] ➔ [–1.07, 2.65]
Lexical diversity: PDW	Overall mean (W = 100)	0.58	[0.1, 0.9] ➔ [0.56, 0.60]
	Development: TBI vs. TD	–0.04	[–0.7, 0.7] ➔ [–0.08, –0.01]
	Sex: male vs. female	0.03	[–0.8, 0.8] ➔ [0.01, 0.05]
	SES: highest vs. lowest	0.07	[–0.7, 0.7] ➔ [0.02, 0.11]
	Age: 18 vs. 13 years	0.01	[–0.8, 0.8] ➔ [–0.02, 0.03]
	Word count: 500 vs. 10	–0.56	[–1.00, 0.4] ➔ [–0.60, –0.51]

Note. TBI = traumatic brain injury; TD = typical development.

adjustment approach to control potential confounding (as opposed to a matched-pairs approach), we were able to use the full data set. Some of the differences in the inference on lexical diversity are due to the fact that the measure of lexical diversity used in similar studies (e.g., Greenhalgh & Strong, 2001; Mills et al., 2013), which Lundine and Barron use in their analysis, is actually confounded with syntactic complexity, whereas PDW is not. Additionally, the approach proposed in this article enables us to combine all data into one "superanalysis," whereas Lundine and Barron were confined to three separate analyses for summaries from each of the three different lecture types. In addition, differences between the conclusions of the two methods may be due to the fact that some of the assumptions of the t test—namely, normality and independence of the response variable—simply do not hold, as is typically the case for any traditional LSA.

Because of the Bayesian GLMM approach proposed in this article, we were also able to investigate other demographic patterns in the microstructural language constructs. Some of these include (a) highly probable differences in syntactic complexity and lexical diversity of discourse summaries

Table 3. Model estimates along with 95% posterior credible intervals for the overall mean rate of utterances per discourse (RUD), rate of subordination (RS), rate of words per utterance (RWU), and probability of a distinct word (PDW) and for the mean difference of these parameters for discourses on lecture types compare–contrast (CC), cause–effect (CE), and narrative (N).

Language constructs	Model parameters	Estimate	Posterior 95% CI
Productivity: RUD	Overall Mean	7.83	[6.54, 9.29]
	CC vs. CE	1.89	[1.01, 2.87]
	CC vs. N	1.29	[0.41, 2.24]
	CE vs. N	–0.59	[–1.42, 0.20]
Syntactic complexity: RS	Overall Mean	1.5	[1.40, 1.61]
	CC vs. CE	–0.07	[–0.15, 0.01]
	CC vs. N	–0.22	[–0.31, –0.13]
	CE vs. N	–0.15	[–0.24, –0.06]
Syntactic complexity: RWU	Overall Mean	11.8	[10.66, 12.99]
	CC vs. CE	–1.93	[–2.91, –0.99]
	CC vs. N	–0.55	[–1.44, 0.32]
	CE vs. N	1.37	[0.39, 2.37]
Lexical diversity: PDW	Overall mean (W = 100)	0.58	[0.56, 0.60]
	CC vs. CE	0.02	[0.01, 0.04]
	CC vs. N	0.02	[0.00, 0.03]
	CE vs. N	–0.01	[–0.02, 0.01]

Figure 8. Proportion of distinct words in a discourse summary plotted against number of total words in the discourse summary, for all discourse summaries in our data set. Discourse summaries on the CC lecture are represented by filled circles, while all other discourse summaries are represented by open circles. For 15-year-old female adolescents summarizing the CC lecture, the posterior mean probability of a distinct word (PDW) is represented by four separate curves, representing subpopulations of high/low SES and TD individuals/individuals with TBI. SES = socioeconomic status; TBI = traumatic brain injury; TD = typically development; CC = compare–contrast.

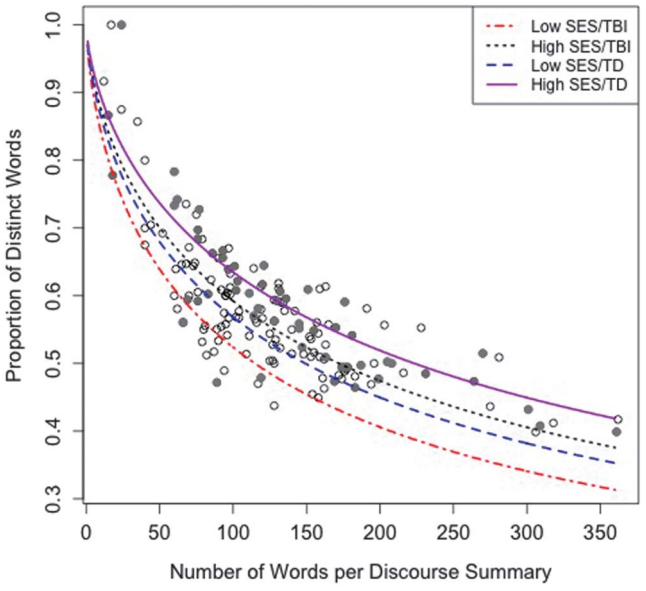


Table 4. Probability that the difference exceeds zero for each of the three parameters under examination: rate of words per utterance (RWU), measuring syntactic complexity; probability of a distinct word (PDW), measuring lexical diversity; distinct words per utterance (DWU), which conflates syntactic complexity with lexical diversity.

Pr[Difference > 0]	RWU	DWU	PDW
Group: TBI – TD	.08	.02	.01
Sex: Male – Female	.10	.34	1.00
SES: High – Low	.99	1.00	1.00
Age: 18 years – 13 years	.80	.83	.71
Lecture: CC – CE	.00	.00	1.00
Lecture: CC – N	.10	.31	.98
Lecture: CE – N	1.00	.99	.21

Note. Probabilities for DWU tend to lie between probabilities for RWU and PDW (except for the age difference). The probability that the association is in the opposite direction (i.e., Pr[Difference < 0]) may be obtained by subtracting the table entry from one. Note that table entries are estimates and true probabilities for these models are never exactly 0 or 1. TBI = traumatic brain injury; TD = typical development; SES = socioeconomic status; CC = compare–contrast; CE = cause–effect; N = narrative.

and moderately probable differences in productivity for those individuals with differing measures of SES, (b) a substantial difference in productivity and syntactic complexity for adolescents of differing ages, and (c) probable differences in all language constructs between male and female adolescents from this central Ohio population. In particular, results seem to indicate a high probability that male adolescents in this population exhibit greater productivity and lexical diversity, while female adolescents exhibit greater syntactic complexity. If replicated in larger studies, these differences are noteworthy, as the small body of literature examining sex differences in adolescent verbal discourse has been inconclusive thus far (e.g., Channell et al., 2018; Rice & Hoffman, 2015).

We also report differences in the microstructural patterns across discourse summaries for the three different lectures that prompted the language samples: CC, CE, and N. In summary, the CC lecture seemed to prompt the highest productivity and lexical diversity, but the lowest syntactic complexity. In comparing summaries produced for the N lecture to those given on the CE lecture, it seems that the N summaries led to greater productivity, lexical diversity, and syntactic complexity as measured by RS (corresponding to the traditional sample statistic: SI), while the CE summaries were associated with great RWU (corresponding to the sample statistic: mean length of C-unit). Few studies have examined the microstructural differences between different types of expository discourse productions (e.g., Scott & Windsor, 2000; Ward-Lonergan et al. 1999), and future work is needed to draw appropriate conclusions and determine if differences, indeed, result in clinically meaningful differences for individuals with differing language and learning abilities.

Limitations with this work should be considered as we attempt to move the statistical methodology of LSA forward. As always, results are subject to the validity of model assumptions, in full confidence that subjects and data were randomly sampled from the population of interest and that the data were sampled and analyzed in an unbiased manner. Results are also highly subject to sampling variability. The sample of adolescents included in this study is small, particularly the small number of adolescents with TBI who were included in this pilot work. Therefore, these results should not be applied to any larger population without great deliberation and should be applied to the population of interest with a healthy degree of caution, insomuch as replication studies have not yet been conducted.

Conclusions and Future Directions

In conclusion, the article proposes and describes the statistical justification for use of Bayesian GLMM analysis for LSA. Using an example data set that includes 50 adolescents with typical language development and a small group of five adolescents who sustained a TBI, the analyses identified microstructural differences between groups and between types of discourse summaries not identified in previous research using more traditional statistical methodology.

Incorporation of Bayesian GLMM analysis may enable a wider range of discoveries as researchers and clinicians incorporate LSA into their work. Indeed, the larger field of speech, language, and hearing science may be great benefactors of this approach as well, since Bayesian GLMM are applicable to a wide variety of data encountered in the discipline.

Author Contributions

Gavin Collins: Conceptualization (Equal), Data curation (Supporting), Formal analysis (Equal), Methodology (Equal), Writing – original draft (Lead), Writing – review & editing (Lead). **Jennifer P. Lundine:** Conceptualization (Lead), Formal analysis (Equal), Funding acquisition (Lead), Investigation (Lead), Methodology (Supporting), Writing – original draft (Equal), Writing – review & editing (Equal). **Eloise Kaizar:** Conceptualization (Supporting), Formal analysis (Supporting), Supervision (Lead), Writing – original draft (Supporting), Writing – review & editing (Supporting).

Acknowledgments

This research was supported in part by the Alumni Grant for Graduate Research and Scholarship and a laboratory start-up seed grant from The Ohio State University to the second author.

References

Agresti, A., & Kateri, M. (2011). Categorical data analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science*. Springer. https://doi.org/10.1007/978-3-642-04898-2

Aguilar, J. M., Cassedy, A. E., Shultz, E. L., Kirkwood, M. W., Stancin, T., Yeates, K. O., Taylor, H. G., & Wade, S. L. (2018). A comparison of 2 online parent skills training interventions for early childhood brain injury: Improvements in internalizing and executive function behaviors. *Journal of Head Trauma Rehabilitation, 34*(2), 65–76. https://doi.org/10.1097/HTR.0000000000000443

Anderson, V., Catroppa, C., Morse, S., Haritou, F., & Rosenfeld, J. (2000). Recovery of intellectual ability following traumatic brain injury in childhood: Impact of injury severity and age at injury. *Pediatric Neurosurgery, 32*(6), 282–290. https://doi.org/10.1159/000028956

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44*(2), 133–146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Catroppa, C., Hearps, S. J. C., Crossley, L., Yeates, K. O., Beauchamp, M. H., Fussela, J., & Anderson, V. (2016). Social and behavioral outcomes following traumatic brain injury: What predicts outcome at 12 months post-insult? *Journal of Neurotrauma, 34*(7), 1439–1447. https://doi.org/10.1089/neu.2016.4594

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., Culhane, K. A., Levin, H. S., Harward, H., Mendelsohn, D., Ewing-Cobbs, L., Fletcher, J. M., & Bruce, D. (1992). Narrative discourse after closed head injury in children and adolescents. *Brain and Language, 43*(1), 42–65. https://doi.org/10.1016/0093-934X(92)90020-F

Chapman, S. B., McKinnon, L., Levin, H. S., Song, J., Meier, M. C., & Chiu, S. (2001). Longitudinal outcome of verbal discourse in children with traumatic brain injury: Three-year follow-up. *The Journal of Head Trauma Rehabilitation, 16*(5), 441–455. https://doi.org/10.1097/00001199-200110000-00004

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

Coelho, C. A., Liles, B. Z., & Duffy, R. J. (1991). The use of discourse analyses for the evaluation of higher level traumatically brain-injured adults. *Brain Injury, 5*(4), 381–392. https://doi.org/10.3109/02699059109008111

Durber, C. M., Yeates, K. O., Taylor, H. G., Walz, N. C., Stancin, T., & Wade, S. L. (2017). The family environment predicts long-term academic achievement and classroom behavior following traumatic brain injury in early childhood. *Neuropsychology, 31*(5), 499–507. https://doi.org/10.1037/neu0000351

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research, 58*(3), 840–852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

Finestack, L. H., Payesteh, B., Disher, J. R., & Julien, H. M. (2014). Reporting child language sampling procedures. *Journal of Speech, Language, and Hearing Research, 57*(6), 2274–2279. https://doi.org/10.1044/2014_JSLHR-L-14-0093

Gordon, K. R. (2019). How mixed-effects modeling can advance our understanding of learning and memory and improve clinical and educational practice. *Journal of Speech, Language, and Hearing Research, 62*(3), 507–524. https://doi.org/10.1044/2018_JSLHR-L-ASTM-18-0240

Greenhalgh, K. S., & Strong, C. J. (2001). Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools, 32*(2), 114–125. https://doi.org/10.1044/0161-1461(2001/010)

Hay, E., & Moran, C. (2005). Discourse formulation in children with closed head injury. *American Journal of Speech-Language Pathology, 14*(4), 324–336. https://doi.org/10.1044/1058-0360(2005/031)

Hemphill, L., Feldman, H. M., Camp, L., Griffin, T. M., Miranda, A. E., & Wolf, D. P. (1994). Developmental changes in narrative and non-narrative discourse in children with and without brain injury. *Journal of Communication Disorders, 27*(2), 107–133. https://doi.org/10.1016/0021-9924(94)90037-X

Hunt, K. W. (1965). *Grammatical structures written at three grade levels (Research Report No. 3)*. National Council of Teachers of English.

Jacobson, P. F., & Walden, P. R. (2013). Lexical diversity and omission errors as predictors of language ability in the narratives of sequential Spanish–English bilinguals: A cross-language comparison. *American Journal of Speech-Language Pathology, 22*(3), 554–565. https://doi.org/10.1044/1058-0360(2013/11-0055)

Jordan, F. M., & Murdoch, B. E. (1994). Severe closed-head injury in childhood: Linguistic outcomes into adulthood. *Brain Injury, 8*(6), 501–508. https://doi.org/10.3109/02699059409151002

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. National Council of Teachers of English.

Lundine, J. P., & Barron, H. D. (2019). Microstructural and fluency characteristics of narrative and expository discourse in

adolescents with traumatic brain injury. *American Journal of Speech-Language Pathology, 28*(4), 1638–1648. https://doi.org/10.1044/2019_AJSLP-19-0012

Lundine, J. P., Harnish, S. M., McCauley, R. J., Blackett, D. S., Zezinka, A., Chen, W., & Fox, R. A. (2018). Adolescent summaries of narrative and expository discourse: Differences and predictors. *Language, Speech, and Hearing Services in Schools, 49*(3), 551–568. https://doi.org/10.1044/2018_LSHSS-17-0105

Lundine, J. P., Harnish, S. M., McCauley, R. J., Zezinka, A. B., Blackett, D. S., & Fox, R. A. (2018). Exploring summarization differences for two types of expository discourse in adolescents with traumatic brain injury. *American Journal of Speech-Language Pathology, 27*(1), 247–257. https://doi.org/10.1044/2017_AJSLP-16-0131

McCabe, A., & Champion, T. B. (2010). A matter of vocabulary II: Low-income African American children's performance on the Expressive Vocabulary Test. *Communication Disorders Quarterly, 31*(3), 162–169. https://doi.org/10.1177/1525740109344218

McMillan, G. P., & Cannon, J. B. (2019). Bayesian applications in auditory research. *Journal of Speech, Language, and Hearing Research, 62*(3), 577–586. https://doi.org/10.1044/2018_JSLHR-H-ASTM-18-0228

Miller, J. F., & Iglesius, A. (2010). *Systematic Analysis of Language Transcripts (SALT), Research Version 2010*. SALT Software.

Mills, M. T., Watkins, R. V., & Washington, J. A. (2013). Structural and dialectal characteristics of the fictional and personal narratives of school-age African American children. *Language, Speech, and Hearing Services in Schools, 44*(2), 211–223. https://doi.org/10.1044/0161-1461(2012/12-0021)

Moran, C., Kirk, C., & Powell, E. (2012). Spoken persuasive discourse abilities of adolescents with acquired brain injury. *Language, Speech, and Hearing Services in Schools, 43*(3), 264–275. https://doi.org/10.1044/0161-1461(2011/10-0114)

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research, 62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370–384. https://doi.org/10.2307/2344614

Nippold, M. A. (2014). *Language sampling with adolescents: Implications for intervention* (2nd ed.). Plural.

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2008). Expository discourse in adolescents with language impairments: Examining syntactic development. *American Journal of Speech-Language Pathology, 17*(4), 356–366. https://doi.org/10.1044/1058-0360(2008/07-0049

Oleson, J. J., Brown, G. D., & McCreery, R. (2019). Essential statistical concepts for research in speech, language, and hearing sciences. *Journal of Speech, Language, and Hearing Research, 62*(3), 489–497. https://doi.org/10.1044/2018_JSLHR-S-ASTM-18-0239

Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech, Language, and Hearing Research, 45*(5), 927–937. https://doi.org/10.1044/1092-4388(2002/075)

Perry, L. K., & Kucker, S. C. (2019). The heterogeneity of word learning biases in late-talking children. *Journal of Speech, Language, and Hearing Research, 62*(3), 554–563. https://doi.org/10.1044/2019_JSLHR-L-ASTM-18-0234

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R Studio Team. (2020). *RStudio: Integrated development for R*. RStudio PBC. http://www.rstudio.com/

Rashid, M., Goez, H. R., Mabood, N., Damanhoury, S., Yager, J. Y., Joyce, A. S., & Newton, A. S. (2014). The impact of pediatric traumatic brain injury (TBI) on family functioning: A systematic review. *Journal of Pediatric Rehabilitation Medicine, 7*(3), 241–254. https://doi.org/10.3233/PRM-140293

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2;6 to 21 years of age. *Journal of Speech, Language, and Hearing Research, 58*(2), 345–359. https://doi.org/10.1044/2015_JSLHR-L-14-0150

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*(2), 201–209. https://doi.org/10.1017/S0305000900012885

Ryan, N. P., Anderson, V., Godfrey, C., Beauchamp, M. H., Coleman, L., Eren, S., Rosema, S., Taylor, K., & Catroppa, C. (2014). Predictors of very-long-term sociocognitive function after pediatric traumatic brain injury: Evidence for the vulnerability of the immature social brain. *Journal of Neurotrauma, 31*(7), 649–657. https://doi.org/10.1089/neu.2013.3153

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

Stan Development Team. (2018a). *RStan: The R interface to Stan* (2.17.3) [Computer software]. http://mc-stan.org

Stan Development Team. (2018b). *The Stan Core Library* (2.18.0) [Computer software]. http://mc-stan.org

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *The Lancet, 2*(7872), 81–84. https://doi.org/10.1016/S0140-6736(74)91639-0

Turkstra, L. S., & Holland, A. L. (1998). Assessment of syntax after adolescent brain injury: Effects of memory on test performance. *Journal of Speech, Language, and Hearing Research, 41*(1), 137–149. https://doi.org/10.1044/jslhr.4101.137

Walz, N. C., Yeates, K. O., Taylor, H. G., Stancin, T., & Wade, S. L. (2012). Emerging narrative discourse skills 18 months after traumatic brain injury in early childhood. *Journal of Neuropsychology, 6*(2), 143–160. https://doi.org/10.1111/j.1748-6653.2011.02020.x

Ward-Lonergan, J. M., Liles, B. Z., & Anderson, A. M. (1999). Verbal retelling abilities in adolescents with and without language-learning disabilities for social studies lectures. *Journal of Learning Disabilities, 32*(3), 213–223. https://doi.org/10.1177/002221949903200303

Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders, 30*(4), 275–287. https://doi.org/10.1097/TLD.0b013e3181ff5a88

Yeates, K. O., Taylor, H. G., Woodrome, S. E., Wade, S. L., Stancin, T., & Drotar, D. (2002). Race as a moderator of parent and family outcomes following pediatric traumatic brain injury. *Journal of Pediatric Psychology, 27*(4), 393–403. https://doi.org/10.1093/jpepsy/27.4.393