2015

Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects

Gerasimos Fergadiotis
Portland State University, gf3@pdx.edu

Heather Harris Wright
East Carolina University

Samuel B. Green
Arizona State University

Citation Details

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, 1-13.

Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects

Gerasimos Fergadiotis[1], **Heather Harris Wright**[2], and **Samuel B. Green**[3]

[1]Portland State University, Portland, OR

[2]East Carolina University, Greenville, NC

[3]Arizona State University

Abstract

Purpose—Several novel techniques have been developed recently to assess the breadth of speaker's vocabulary exhibited in a language sample. The specific aim of this study was to increase our understanding of the validity of the scores generated by different lexical diversity (LD) estimation techniques. Four techniques were explored: *D*, Maas, Measure of Textual Lexical Diversity (MTLD), and the Moving Average Type Token Ratio (MATTR).

Method—Four LD indices were estimated for language samples on four discourse tasks (procedures, eventcasts, story retell, and recounts) from 442 neurologically intact adults. The resulting data were analyzed using structural equation modeling.

Results—The scores on the MATTR and MTLD estimation techniques were stronger indicators of the LD of the language samples. The results for the other two techniques were consistent with the presence of method factors representing construct-irrelevant sources.

Conclusions—These findings offer a deeper understanding of the relative validity of the four estimation techniques and should assist clinicians and researchers in the selection of LD measures of language samples that minimize construct-irrelevant sources.

Lexical diversity (LD) has been used in a wide range of areas, producing a rich history in speech-language pathology. For example, LD has been used to understand the relationship between phonological processing and the development of reading skills (e.g., Smith, 2009); to differentiate typically developing children from children with specific language impairment (e.g., Thordardottir & Namazi, 2007); to screen and identify bilingual children with language deficits (Kapantzoglou, Fergadiotis, & Restrepo, 2010); to evaluate the progress of children after cochlear implantation (Ertmer, Strong, & Sadagopan, 2002); to differentiate individuals with aphasia from neurologically intact adults (e.g., Lind et al., 2009); to capture the differences between individuals with fluent and nonfluent aphasia (Fergadiotis & Wright, 2011); to inform models of language processing in aphasia (e.g., Gordon, 2008); to assess the efficiency of therapeutic approaches (Rider, Wright, Marshall,

Contact Information: Gerasimos Fergadiotis, Ph.D., CCC-SLP, Portland State University, Department of Speech and Hearing Sciences, PO Box 751, Portland, OR 97201, gfergadiotis@pdx.edu.

& Page, 2008); and to use as a possible early marker of dementia (Bucks, Singh, Cuerden, & Wilcock, 2000).

Both researchers and clinicians have utilized a variety of measures to capture the lexical diversity of language samples (LD). However, identifying a robust approach to measure LD has been challenging because typical measures of LD reflect not only the construct of interest (i.e. LD), but also construct-irrelevant variation, such as the length of samples. As a result, the score interpretation of these measures is problematic. For example, the type toke ratio (TTR; Chotlos, 1944; Templin, 1957), which is one of the most used indices in speech-language pathology, is inherently flawed because it varies as a function of sample length (Heap, 1978). Therefore, comparisons across language samples of different speakers, or even across different samples produced by the same speaker, may be confounded by sample length. Standardizing sample size is not a satisfactory solution to this problem because researchers would have to agree on the number of tokens to estimate TTR--an unrealistic expectation given the multitude of applications. Further, discarding text may reduce a language sample's integrity and lead to spurious results because the introduction of novel vocabulary in the truncated language sample would not necessarily be uniformly spread throughout the sample (Youmans, 1991).

Recently, a new generation of tools has emerged from the field of computational linguistics that can be used to measure the LD of a language sample. All four measures described here claim to be relatively robust to length variation, a feature that allows for comparisons of discourse samples within and between participants as well as across studies without requiring truncation of language samples. Further, because they are estimated from the whole transcript, they should be less prone to measurement error due to clustering of novel content words. With few exceptions, these measures have been used in limited applications in the field of speech-language pathology. The purpose of the current study is to explore the validity of these four LD measures. Before delineating the goals of our study in greater detail, we describe how these four measures are computed and the literature addressing their validity.

Computation of Four LD Measures

D

The measure D (Malvern & Richards, 1997; McKee, Malvern, & Richards, 2000) reflects the LD of a language sample by capturing how fast TTR decreases in a language sample. Computationally, estimating *D* involves a series of random text samplings to plot an empirical TTR versus number-of-tokens curve for a sample. Initially, 35 tokens are randomly drawn from the text without replacement and the TTR is estimated. This process is repeated 100 times and the average TTR for 35 tokens is estimated and plotted. The same routine is then repeated for subsamples from 36 to 50 tokens. The average TTR for each subsample of increasing token size is subsequently plotted to form the empirical curve. Then, a theoretical curve is produced that maximizes its fit to the empirical TTR curve using the least squares approach, with TTR calculated as follows (McKee et al., 2000):

$$TTR = \frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]. \quad (1)$$

Lower D values reflect steeper theoretical curves. Because D is the product of a stochastic process, its value varies each time the program is run. For that reason the whole process is repeated three times and the final D value is the average of the three runs.

Maas index

Another approach researchers have used to minimize the effect of sample length is linearizing. Conceptually, this approach is based on the assumption that the TTR curve can be fit relatively well by a logarithmic curve. In theory, if one could transform the relationship between N and TTR to achieve linearity, it is straightforward to use regression analysis to estimate the slope of that linear relationship. This approach was recently advocated by McCarthy and Jarvis (2007) and first described by Maas (a^2; 1972):

$$a^2 = \frac{(logTokens - logTypes)}{log^2 Tokens}. \quad (2)$$

Measure of textual lexical diversity

Another tool that has been proposed recently for estimating LD is the measure of textual lexical diversity (MTLD; McCarthy, 2005). Conceptually, for any given sample, MTLD reflects the average number of consecutive words for which a certain TTR is maintained. For language samples with high LD, multiple words may be produced before TTR drops below a certain cutoff score because there is a lower propensity of the same words being repeated. During the estimation process, TTR is estimated after each word. For example, “I” (TTR = 1.00) “am” (TTR= 1.00) “tired” (TTR = 1.00) “but” (TTR = 1.00) “I” (TTR = .800) “am” (TTR = .667) “also” (TTR = .714) and so forth. However, when the default TTR factor size value (i.e., .720) is reached, a factor count increases by a value of 1, and the TTR evaluations are reset. The same process is repeated until the last token of the language sample has been added and the TTR has been estimated. Then, the total number of words in the text is divided by the total factor count. For example, if the text has 300 tokens and the factor count is 4.1, then the MTLD value is 73.17. Subsequently, the whole text in the language sample is reversed and another score of MTLD is estimated. The forward and the reversed MTLD scores are averaged to provide the final MTLD estimate.

Moving average type token ratio

Conceptually, the moving-average type-token ratio (MATTR; Covington & McFall, 2010) calculates the LD of a sample using a moving window that estimates TTR’s for each successive window of fixed length. Initially, a window length is selected (e.g., ten words) and the TTR for words 1–10 is estimated. Then, the TTR is estimated for words 2–11, then 3–12, and so on to the end of the text. For the final score, the estimated TTR’s are averaged.

Validity of LD score Interpretations

The validity of LD scores would be threatened to the extent that construct-irrelevant sources of variance were found to influence the observed LD scores. In other words, in the face of evidence that observed scores varied systematically as a function of a variable other than the construct of interest, the viability of the claims could be compromised (e.g., Kane, 2009). That would be the case if, for example, scores obtained from a specific approach were found to reflect more than just LD. In this case, it would be difficult to argue in favor of such an approach and the conclusions reached when using this approach.

The validity of D has been investigated in several studies (e.g., Malvern & Richards, 1997; Durán, Malvern, Richards, & Chipere, 2004; Richards & Malvern, 1997) in which estimates of D were found to correlate strongly with well-validated measures of language and also developmental and demographic variables. However, other studies have suggested problems with this technique. For example, Owen and Leonard (2002) found that mean LD, as estimated by D, varied as a function of sample length. Samples that were truncated to 250 words had a significantly lower mean D score compared to samples that were truncated to 500 words. Based on these findings, Owen and Leonard concluded that “it appears that D does not entirely avoid the problem of sample size influence” (p. 935). In addition, Fergadiotis, Wright, and West (2013) evaluated four measures of LD to determine how effective they were at measuring LD for people with aphasia. They applied MTLD, MATTR, D, and the Hypergeometric Distribution-based D (McCarthy & Jarvis, 2007). Results from a confirmatory factor analysis suggested that MTLD and MATTR were relatively pure measures of LD, whereas D and the Hypergeometric Distribution-based D covaried with TTR, even after accounting for the variance that was attributed to LD. This relationship suggested that the Hypergeometric Distribution-based D and D were affected by language sample size. Finally, Koizumi and In’nami (2012) investigated D and MTLD in Japanese second language learners of English and reached similar conclusions regarding the validity of these two indices.

With respect to MTLD, McCarthy and Jarvis (2010) explored its convergent and discriminant validity. The authors used three indices of LD, which included Maas, D, and Yule’s K (a probabilistic index; Yule, 1944) to investigate its convergent validity. They found that MTLD was moderately to strongly correlated with all three indices (.84, .69, and .85, respectively). The high coefficients were evidence of its convergent validity. Further, MTLD did not correlate strongly with TTR (r = .32), a measure known to be influenced by length. This finding was evidence of its discriminant validity Koizumi and In’nami (2012) reported additional supporting evidence for the validity of MTLD in their study. They found that MTLD was minimally related to length, especially in samples that were longer than 100 tokens.

Goals of the Study

D, Maas index, MTLD, and MATTR are four novel techniques that have been developed to assess the breadth of one’s vocabulary exhibited in a language sample. Though all of the techniques claim to measure LD, each takes a distinctly different computational approach

based on its own set of theoretical assumptions. Therefore, it is unclear whether these techniques measure the same construct and to what extent they produce valid and reliable scores. The current study explored how the scores of different techniques for estimating LD related to each other. Further, the research questions of this study were cast in a modern psychometric framework and investigated using structural equation modeling. The specific aim of this study was to increase our understanding of the validity of the scores generated by different LD estimation techniques. The following research questions were addressed:

i. Do all the techniques generate scores that are manifestations of the same latent variable (i.e., LD)?

ii. Is there a single latent variable determining performance for each estimation technique or are there specific method factors that jointly determine the scores?

iii. What are the magnitude and the nature of the relationships among the observed scores, the LD, and the method factors?

iv. Is there evidence that supports the use of a specific estimation technique?

Method

Participants

Four-hundred and forty-two adults met the following inclusion criteria for participation in the study: (a) no history of stroke, head injury, or neurogenic disorder based on self-report; (b) aided or unaided hearing acuity within normal limits; (c) normal or corrected visual acuity; (d) monolingual speakers of English; (e) normal cognitive functioning, as indicated by a score 24 (out of 30) or higher on the Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 2002; Lopez, Charter, Mostafavi, Nibut, & Smith, 2005); and (f) no signs of depression, as indicated by a passing score (0–4) on the Geriatric Depression Scale (Brink et al., 1982). Participants' demographic information can be seen in Table 1.

Discourse Elicitation: Stimuli and Instructions

Tasks were designed to elicit four discourse types from participants. The first task was developed to elicit procedural discourse. The other three tasks were created to elicit three types of narrative discourse: eventcasts, storytelling, and recounts. Each discourse task was introduced with a warm-up task.

For the procedural discourse task, the examiner demonstrated the steps required to make a pot of coffee. Then, to obtain two procedural language samples, participants were asked to provide the steps (a) to make a peanut butter and jelly sandwich and (b) to plant a flower in a garden.

For the eventcasts discourse task, the warm-up task included a brief narrative provided by the examiner describing the Picnic Scene from the Western Aphasia Battery-Revised (WAB-R; Kertesz, 2007). Next the participants were asked to produce a story for the Cookie Theft picture from the Boston Diagnostic Aphasia Examination-3 (BDAE-3; Goodglass, Kaplan, & Barresi, 2001). As part of the warm-up task, feedback was provided to avoid eliciting a simple description of objects, characters and/or their physical characteristics.

Then to elicit the eventcast language sample, participants were presented with the Nicholas and Brookshire (1993) single and sequential pictures and were asked to produce a story that was based on temporal sequencing ("*Take a minute to look at this picture; when you are ready, tell me a story that has a beginning, middle and end*").

For the storytelling discourse task, the warm-up task involved the examiner providing an example of how to tell a story using a wordless picture book (The Great Ape; Krahn, 1978). To obtain two storytelling language samples, participants were given an unlimited amount of time to view and become familiar with the stories depicted in the wordless picture books *Picnic* (McCully, 1984) and *Good Dog Carl* (Day, 1985). Then, they were asked to "Tell a story that goes with the pictures" for each of the books.

For the recounts discourse task, the examiner initially modeled the task by providing a brief personal narrative about a trip to San Diego, California. To elicit the recounts language sample, each participant was asked to recall and share three past experiences; what they did (a) last weekend, (b) during their last holiday, and (c) during their last vacation. If the participant stopped after 15 seconds or more, he/she was prompted with "Can you tell me more?"

Transcription

Samples were digitally recorded and then orthographically transcribed in the computerized language analysis format (CLAN; MacWhinney, 2000) by trained research assistants. Approximately 10% of the samples were randomly selected and transcribed again for reliability purposes. Intra- and inter-rater word-by-word transcription reliability were above 85%. Nonwords, hesitations, revisions, repetitions, and onomatopoeia were coded via transcription codes in CLAN and were excluded from further analysis. Next, for each participant, four separate files were created that represented the four discourse types. Procedures included statements about "How to make a peanut butter and jelly sandwich" and "How to plant a flower in a garden"; eventcasts involved the four picture descriptions; stories included narratives about Picnic and Good Dog Carl; and recounts comprised the participant's accounts of three previous experiences.

Estimating Lexical Diversity

For each of the four input files provided by a participant, LD was estimated using each estimation technique. *D* was estimated using the voc-*D* program in CLAN. Maas and MTLD were computed using Gramulator 5 (McCarthy, Watanabe, & Lamkin, 2012). Finally, MATTR was calculated using the computer software developed by Covington (2007). Based on the results of these programs, a dataset was created that included 16 LD variables (4 discourse types x 4 estimation methods) for each individual.

Modeling Approach

We investigated the structure of the 16 variables using a multitrait-multimethod (MTMM; Eid & Diener, 2006; Campbell & Fiske, 1959) approach with structural equation modeling (SEM). The MTMM approach refers to the analysis of relationships among measures that are a function of multiple traits assessed using multiple methods (Geiser, 2008). Casting the

MTMM approach within an SEM framework allows researchers to build SEM models that can decompose the variance that is due to traits, variance that is due to methods, and unique variance associated with each observed indicator. The MTMM matrices were analyzed using the correlated traits-correlated methods parameterization (Lance, Noble, & Scullen, 2002; Widaman, 1985).

In the context of our study, the multiple traits of MTMM were represented by the four types of discourse, and the methods were represented by the four lexical diversity methods. The discourse types were not viewed as traits in the traditional psychological sense, but were in recognition that the lexical diversity of language samples is likely to differ as a function of these types. In other words, variance due to discourse type is conceptually meaningful. On the other hand, for a particular discourse type, lexical diversity should not differ among samples as a function of estimation method. Thus, variance due to estimation method is viewed as problematic.

To answer the research questions of this study using SEM, a series of nested models were specified on an a priori basis. Each model was evaluated in terms of a variety of global fit indices. Overall, the model was evaluated using the robust χ^2 statistic (Muthén, & Muthén, 2011), which allows for the non-normality of the data and missing data. However, it is important to note that this χ^2 test is sensitive to sample size, and with large enough sizes it is likely to lead to rejection of the null hypothesis, even when the model fits well (Kline, 2005). χ^2 difference tests also were used to evaluate differences between nested models. Three additional fit indices were computed to evaluate models: comparative fit index (CFI; Bentler, 1990); root-mean square error of approximation (RMSEA; Steiger & Lind, 1980); and, standardized root mean residual (SRMR; Hu & Bentler, 1999). Based on published guidelines, good fit is indicated by a CFI value close to or higher than .95, an RMSEA value below .08, with the upper bound of the 95% confidence interval below .10, and an SRMR value less than or close to .08, although cutoffs are somewhat model dependent (Brown, 2006; Hu & Bentler, 1999; Kline, 2005).

We used the following steps in our model evaluation process:

Step 1—The first model (Model 1) specified a single factor to underlie all 16 combinations of discourse types and LD estimation techniques, with no covariances among the residual scores. This factor could be interpreted as general LD characterizing speakers.

Step 2—The second model (Model 2) specified four factors differentiating the four types of discourses: procedural, eventcasts, storytelling, and recounts (i.e., LD_{Pro}, LD_{Eve}, LD_{Sto}, and LD_{Rec}, respectively). LD scores for these different discourse types were seen to be independent of each other. In other words, lexical diversity for one discourse type did not predict lexical diversity for other discourse types. For this model there was an assumption that the results for the different methods converged (i.e., the orderings of the language samples were comparable across estimation methods for a particular discourse type).

Step 3—The model in this step was the same as the model in the previous step except the discourse type factors were allowed to be correlated. This model was probably more realistic

in that individuals who have higher scores in one discourse type are likely to have higher scores in other discourse types. Again, comparable to Model 2, the assumption was that the various methods converged.

Step 4—Model 4 included not only the four correlated factors representing the discourse types (LD_{Pro}, LD_{Eve}, LD_{Sto}, and LD_{Rec}) as did Model 3, but also four additional factors representing the four estimation techniques or methods (i.e., M_{D}, $M_{Maas,}$ M_{MTLD}, and M_{MATTR}). Accordingly, each variable was a function of one discourse-type factor and one estimation method factor (as well as a residual source). The method factors were specified as uncorrelated. If Model 4 increased fit over and above Model 3, the inclusion of method factors improved fit, suggesting that the estimation methods produced divergent results for the various discourse types.

Step 5—The last step included an exploration of the relationship between the method factors. Specifically, it allowed for the method factors to be correlated that would indicate they were similar in nature.

Preliminary Analyses

Descriptive statistics for the number of types, tokens and type-token ratios can be found in Table 2. Overall, these descriptive analyses are within expected ranges.

The amount of missing data was minimal for the type-method variables. The percentage of missing data ranged from .23% to 2.71%, with the mean percentage being 1.02%. Reasons for missing data included recording equipment failures, video format conversion difficulties, and testers' oversights during task administration and data collection.

Data were screened for outliers univariately (scores > 3.3 *SD*'s beyond the mean) and multivariately with Mahalanobis distance (*p* values < .001) using SPSS. No scores were identified as outliers multivariately. Univariately, .39% of the scores across all variables were identified as univariate outliers and were removed from the dataset. The language sample of each outlier was inspected, and for the majority of individuals, outliers were due to participants not following task instructions. For example, some participants produced a mixture of expository, narrative and procedural discourse when asked to produce procedural discourse.

After the outliers were eliminated, descriptive statistics were computed, as shown in Table 3. The assumption of normality of the linguistic complexity measures was evaluated by examining skewness and kurtosis statistics in Table 3 and by visually inspecting graphs of them. Overall, none of the variables demonstrated extreme skewness or kurtosis. However, to be prudent, the models were estimated with Mplus 6.1 (Muthén, & Muthén, 2011) using the MLR estimator, which estimates parameters using maximum likelihood with standard errors and a chi-square test statistic that are robust to non-normality. MLR accommodates missing data in the analyses under the assumption of missingness at random (cf. Enders, 2010).

Results

Main SEM Analyses

A series of hypotheses using nested model comparisons were tested followed by a substantive evaluation of model parameters if the model fit adequately. To perform the nested model comparisons, the scaled difference χ^2 test statistic (Muthén, & Muthén, 2011) was used. A summary of the fit statistics for the various fitted models can be seen in Table 4.

Steps 1 & 2—The parsimonious one-factor model demonstrated very poor fit to the data and failed to satisfy any of the published criteria for acceptable fit discussed earlier. Model 2 that specified four, uncorrelated discourse-type factors resulted in a large increase in model fit compared to Model 1. However, despite the considerable improvement, Model 2 also did not demonstrate acceptable fit to the data.

Step 3—Model 3 was comparable to Model 2, but specified with freely estimated covariances among the discourse-type factors. Model 3 resulted in further improvement of model fit compared to Model 2 based on the indices reported in Table 4. Further, using the scaled difference χ^2 test statistic, Model 3 was found to fit the data significantly better than Model 2, $\Delta \chi^2 (6) = 370.29$, $p < .001$. However, despite the improvement in overall fit, Model 3 failed to demonstrate acceptable fit to the data.

Step 4—Model 4 included not only correlated discourse-type factors, but also orthogonal estimation method factors. In addition, the discourse-type factors were specified to be uncorrelated with the estimation method factors. Although the global fit indices generally indicated relatively adequate fit for Model 4, two parameter yielded out-of-bound estimates (i.e., Heywood cases). Specifically, the unstandardized residual variance of the observed variable corresponding to MATTR scores based on storytelling was negative (−.418), and the standardized loading on this measure with the MATTR method factor, which should be a correlation coefficient, was estimated to be larger than one (1.16).

Further inspection of estimates for the solution suggested that the MATTR method factor was ill-defined, as evidenced by the very low standardized loadings with the remaining MATTR estimated variables (i.e., all loading were less than .01 in absolute value). Also the standardized loadings for the MTLD method factor were very low (i.e., all standardized loadings were less than .10 in absolute value).

Given the two method factors appeared to be poorly defined with Model 4, a new model (Model 4a) was specified with only two orthogonal method factors, D and Maas factors. Model 4a is consistent with the hypothesis that only the LD variables estimated using the D and Maas indices were jointly determined by both content and method factors. The model converged to a proper solution and yielded good fit (Table 4). It should be noted that the standardized parameter values for this model were similar in value for those of Model 4.

Step 5—The estimation method factors were specified to be uncorrelated in Step 4 to avoid improper solutions (Kenny & Kashy, 1992). In this step, Model 5 was specified that allowed

the D and Maas method factors (i.e., M_D and M_{Maas}) to be correlated. The model converged to a solution, it yielded no out-of-bound estimates and demonstrated good fit to the data. Model 4a and 5 were compared using a χ^2 test statistic to explore whether freeing the covariance between the two method factors resulted in improved model fit. Allowing for a correlation between the two method factors did show improved fit, $\Delta\chi^2(1) = 55.49$, $p < .001$. The path diagram for Model 5 is presented in Figure 1 and the variances of the lexical diversity measures can be seen in Table 5. In Figure 1, consistent with SEM conventions for graphical model representation, squares indicate observed variables (e.g., D based on the procedural discourse language samples). Circles represent unobserved latent variables that are estimated as part of the model (e.g., LD_{Pro} that represents an error- and bias-free estimate of LD of a procedural sample). In this model, straight arrows are standardized factor loadings and represent the effect of latent variables on observed variables. The curved double-headed arrows between latent variables represent the correlations between these variables. If latent variables are not connected by double-headed arrows, they are uncorrelated. Because any observed variable is affected only by uncorrelated latent variables, the standardized factor loadings are correlations between the observed and latent variables.

As shown in Figure 1 and Table 5, a number of interesting patterns emerged across the lexical diversity measures. The factor loadings for the discourse-type factors (i.e. LD_{Pro}, LD_{Eve}, LD_{Sto}, and $LD_{Rec)}$ generally were large in magnitude (median = .91; range of absolute values = .46 – .98). In comparison, the loadings for the two estimation method factors were relatively small, but nevertheless substantial. For the D factor (M_D), the loadings ranged from .19 to .38 in absolute value, with a median value of .35. For the Maas factor (M_{Maas}), the loadings tended to be larger, ranging in absolute value from .27 to .62, with a median value of .45. It is interesting note that the factor loadings for the discourse-type factors were highest for the Story language samples when they were evaluated using MTLD and MATTR, but highest for the Procedure language samples when they were evaluated using D and Maas.

In Table 6, we present the decomposition of the variances associated with all combinations of discourse types and LD estimation techniques into the proportions attributed to the various factors and the residual source. Provocatively, the lexical diversity measures assessed using D and Maas had not only a portion of theirs variances account for by method factors, but also greater proportions of their variances unaccounted for by all factors in the model relative to other measures.

Follow-Up Analyses

A series of follow-up analyses were performed to investigate the nature of the method factors that were found previously. In these previous analyses, complete language samples were analyzed that varied naturally in length. Based on prior literature, it was hypothesized that these method factors were related to length effects. To assess whether different lengths of text produce different outcomes on measures of lexical complexity, we manipulated the language samples to equate them in terms of length. If the method factors in the previous

analyses were due to length, then experimentally removing the length variation should eliminate the need to include method factors in the models.

The language samples were truncated to 75 words, which was the length of the shortest language sample in the database. By truncating to 75 words, we were able to retain the language samples for all 442 participants. Two models were fit to the data for the truncated samples: Model 3′, which specified four correlated discourse-type factor, but no method factors, and Model 5b′, which specified two correlated method factors in addition to four correlated discourse-type factors. These models were identical to Model 3 and 5b, but were estimated using the truncated dataset.

The fit for Models 3′ and 5b′ are presented at the bottom of Table 4. Model 3′ yielded very good fit without specifying additional method factors. On the other hand, Model 5b′ had similar fit, but yielded an inappropriate solution (i.e., Heywood cases). In other words, when language samples were equated in terms of length, method factors were no longer necessary to model the data. Overall, these results, in combination with the previous findings, suggest that the method factors were required in the analyses for the language samples that varied in length because D and Maas were sensitive to sample length.

Discussion

The goal of this study was to investigate the validity of four computational tools for quantifying LD. We examined a corpus of four types of discourse (procedures, eventcasts, storytelling, recounts) from 442 adults. Language samples varied with respect to their length. Using four techniques (D; Maas; MTLD; MATTR), LD scores were estimated for each type. Subsequently, data were modeled using SEM to uncover their latent structure. The modeling yielded several important findings.

LD Indices

D & Maas—Regarding D, the parameters of Model 5 suggested that the validity of its score interpretations with respect to LD was not as strong as for MTLD and MATTR. First, the average loading of the content factors was .83 for D-based scores across discourse types. In addition, the average proportion of variance attributed to the content factors was approximately 70%. Even though these estimates are high and suggest a strong relationship with appropriate factors, they are considerably lower than the respective parameter values of MTLD (i.e., 90%) and MATTR (i.e., 93%). Further, when language samples were measured using D, scores reflected the effects of a D-specific method factor in addition to the content factors, Even though on average the proportion of variance that was accounted for by this method factor was not excessive (= 11%), it constitutes a second dimension along which D scores varied systematically.

The modeling in this study provides a coherent explanation for the contradictory conclusions reached by McCarthy and Jarvis (2010) regarding D, Maas, and MTLD. In their paper, they reported high correlations among the three tasks that were interpreted as evidence of convergent validity. However, they also reported a discriminant analysis in which scores generated by these three techniques were used to differentiate among the types of discourse.

The results of this analysis suggested that D, Maas, and MTLD each contributed unique information to the prediction model. Based on these results, the authors concluded that “at least three of the sophisticated LD indices used in this study do not appear to assess exactly the same latent trait.” (p. 390–391). Based on Model 5, these results could be explained on the basis of the different sources of covariation in the data. Observed LD scores based on D and Maas are influenced by secondary method factors that “enrich” their scores and allow them to contribute additional explanatory information in the prediction model of the discriminant analysis.

The interpretation of a Maas score as a clear indication of the LD may not warranted. Maas demonstrated the lowest average loading across discourse types (= −.56), and it was influenced heavily by a method factor. Maas loadings were estimated to be negative because unlike the other three indices for which higher scores indicate higher levels of the latent trait, higher Maas scores are associated with lower levels of the latent trait because of how Maas is estimated. This inverse relationship is reflected in the negative sign. The average proportion of variance that was attributed to the method factor was approximately 22%, more than twice as large compared to D. In fact, when Maas scores were derived based on eventcasts, the influence of the method factor was stronger than the influence of the content factor. Given that Maas scores were so strongly influenced by a variable other than the construct of interest, they should not be interpreted as valid LD indicators.

MTLD & MATTR—Variables associated with MTLD and MATTR were found to be very strong indicators of the LD of a language sample in absolute terms, and stronger in comparison with D and Maas. Content factors (i.e., discourse-type factors) accounted for, on average, 91% of the variances of the variables associated with these techniques across all discourse types. Importantly, MATTR and MTLD differed in another way from the other two indices in the study: scores that were generated using MATTR and MTLD were influenced only by the content factors and did not demonstrate systematic effects from construct-irrelevant sources. Therefore, holding all else constant, MTLD and MATTR provide a more accurate reflection of the LD for the samples of the discourse type elicited in this study

The highest average proportion of variance shared between measures and content factors across discourse types was associated with MATTR-generated variables. Content factors accounted for, on average, 93% of the variances of these variables across all discourse types. This finding highlights the potential of this new measure to produce estimates that are valid and reliable indicators of the LD of language samples. Further, this finding is important because, with the exception of Fergadiotis et al. (2013), currently there is very limited research on this measure.

Regarding MTLD, our findings lend support for its use and add to the results of previous studies reported in the literature. MTLD was found to correlate strongly with a number of LD indices including D and Maas, leading researchers to argue in favor of MTLD’s validity (e.g., McCarthy, 2005). In our study, variables that were estimated using D, Maas, and MTLD loaded strongly factors that represented discourse-specific LD (Model 5). This model may provide a more accurate account of why variables that are generated based on

different estimated techniques relate and can be used to explain the high correlations that have been observed across other studies in the literature.

The Nature of the Method Factors

The results from the follow up analyses provided evidence that both method factors were associated directly or indirectly with length. In the first set of analyses, language samples varied with respect to their length. Subsequently, using truncated language samples, a comparison of Models 3, $3_{(2)}$, 5b and $5b_{(2)}$ revealed that when the observed scores were based on language samples that varied in length, the inclusion of method factors was necessary to achieve adequate model fit. However, when data were truncated, the method factors were no longer required to obtain acceptable fit. This pattern of results suggested that relating the method factors to length was justified.

With respect to D, the nature of the method factor might not be directly related to length, but rather to the number of topics or themes a speaker produces. Owen and Leonard (2002) found that D scores were *higher* for 500-word samples than for 250-word samples. These results were similar both for typically developing as well as language impaired children. Further analysis showed that the distributions of D scores (for each length size) overlapped by approximately 70%. This finding was attributed to the possibility of new themes appearing in the second half of the samples that had not appeared in the first half. Owen and Leonard argued that the introduction of new themes "would, by necessity, introduce new content words, altering the D scores of the children towards a slightly higher score" (p. 936).

If this was indeed the case, it could also explain why the correlation between the Maas and D method factors in our study was moderately high (i.e. .50). If D and Maas were directly related to length, one might have expected a higher correlation between these method factors. However, if what increases D estimates were only indirectly related to length and mediated by the number of themes, elaborations, or episodes it would be expected that the two method factors would still be correlated, but not as strongly.

Even though it is not clear how length relates to D, the results from Owen and Leonard's study are consistent with the configuration and the parameter estimates of Model 5. For example, if the loadings from discourse specific LD factors to D-estimated variables were equal to 1, this would suggest that D reflected purely the LD of the language samples and nothing else. In that hypothetical scenario, one would expect the two distributions from Owen and Leonard's study to overlap completely because manipulating other factors (such as length) would have no direct effects on the scores. However, results of this study indicated that the loadings were not perfect and a second factor, which could be associated directly or indirectly with length, was found to influence D scores in Model 5. Therefore, relating these findings back to Owen and Leonard's results, scores for the 500 word language samples would be expected to differ from language samples of 250 words. Also, whereas TTR would be expected to drop systematically as a function of sample size, D would be expected to increase for both groups when longer samples were analyzed.

With respect to Maas, similar findings have been reported in the literature. For example, Tweedie and Baayen (1998) used Monte Carlo simulation and demonstrated that a

notational variant of the Maas index was monotonically influenced by length. Cossette (1994) reached similar conclusions and argued that Maas was not length invariant.

However, our findings are not consistent with the conclusions reached by McCarthy and Jarvis (2007), who argued that influence of length on Maas was minimal. In their study, they used data from 23 previously published corpora; 16 were written corpora and 7 were spoken. To examine the effects of length, they partitioned the texts to assess whether the smaller parts can project the score of the whole text when reconstituted. The main idea was that if a text is divided into two parts, their average LD should approximate the LD of the original text. By partitioning the text into smaller parts, the trend of the average LD scores would reveal whether there was a relationship between the LD measurement and text size. They argued that "the more the mean LD score of the section sizes correlates with the mean token size of the section size, the *less* the LD measure is able to satisfactorily project" (p. 478). Using this approach they concluded that all LD measures were a function of length, but Maas shared only 2% of variance with the mean token size across all corpora.

One possible explanation for the strong method effects in our data set is that Maas may correlate with length much more in spoken tasks than written tasks. McCarthy and Jarvis (2007) reported the percentage of variance Maas shared with the mean token size *across all corpora.* However, an examination of Table 7 (p. 480) indicates that the correlations between Maas and mean token size for written and spoken corpora was .12 and .32, respectively. The squared average correlation across all corpora would indeed suggest that Maas and mean token size share only 2% of variance. However, the squared correlations of Maas scores and the mean token size for written and spoken discourse tasks separately would be equal to 1% and 10%, respectively. This would suggest that one would expect a much stronger length effect in data that are based on spoken discourse.

Clinical and research implications

There are several situations in which the technique one uses to estimate LD might play a significant role in the interpretation of the LD scores. For example, it is often the case in the field of speech-language pathology that within clinical populations there are subtypes that demonstrate unique clusters of symptoms. For example, individuals with Broca's and Wernicke's aphasia differ characteristically in terms of the volume of their verbal output. Comparisons across these two groups could be quite misleading, depending on the estimation technique one would choose to utilize. For example, given that individuals with Wernicke's aphasia produce longer language samples, D would probably overestimate their LD. That, in turn, could result in artificially inflating the possibility of detecting differences between the language samples of the two groups. In contrast, using Maas could potentially lead to reaching the opposite conclusion; that is, given that people with Wernicke's aphasia produce longer samples, their Maas scores would suggest lower LD because of the length effects. Subsequently, the use of Maas scores could minimize or mask any differences between the language samples of the two groups.

LD has also been used as an indicator of lexical retrieval improvement during discourse production pre- and post-treatment. Again, the decision to use a problematic technique to estimate the LD of a sample could lead to erroneous conclusions due to length effects.

Longer samples could be produced as a result of testers establishing rapport with patients or patients getting familiar with the task demands. Therefore, it is possible that clinicians could be making judgments about the efficacy of a treatment by misinterpreting length effects as changes in LD.

From the four techniques examined in this study, D has been employed most often in the field of speech-language pathology. Several studies have reported results and have reached conclusions based on D scores. The interpretation of the findings has relied on the assumption that D is not a function of length and therefore language samples of different size could be compared meaningfully. However, the findings of this study suggest that when a language sample is evaluated using D, scores cannot be interpreted unequivocally as reflections of the language sample's LD without taking into account the samples' length and/or the possibility that the production of new themes might be contributing to score differences.

Given the estimation of LD can be contaminated by sample length, it would be very helpful to publish language sample characteristics, such as mean and range of types and tokens, when reporting results. Sharing this kind of information would allow readers to critically evaluate the findings of the study. This is the case even if authors decide to estimate LD using one of the methods that were found to be relatively free of systematic length effects in this paper. The first reason is that techniques such as MTLD and MATTR might perform differently in different types of discourse that have not been studied yet. Further, it is possible that what might be considered an unbiased technique today, might be proven biased in the future. The long history of LD indices that had claimed to be free of length effects, only to be found problematic later, makes a strong argument for this conclusion.

Overall, applying MTLD and MATTR to the language samples of 442 individuals provide evidence suggesting that these techniques generated scores that were strong indicators of the LD of language samples. However, an advantage of MTLD over MATTR is that, at this point, the former is actively researched and more evidence has been accumulated that favors the validity of MTLD. Alternatively, very little is known about MATTR, and future studies should be designed to explore its performance under different conditions. An advantage of MATTR is that it is equivalent to TTR, and thus fairly straightforward to grasp and explain. It does not require an understanding of frequency distributions, curve fitting, or the nature of stochastic processes in order to convey its meaning. Its simplicity increases the face validity of the technique (i.e., its potential to measure what it is supposed to measure at face value). Face validity is a very desirable property, especially for professionals who work with adults or children with speech and language disorders in clinical settings (e.g., Gordon, 2008; Lind, Kristoffersen, Moen, & Simonsen, 2009). If the involved parties believe that an index is nonsensical, absurd and/or a waste of time, they would be reluctant to use it or trust it. Yet, if the meaning of the scores is easily conveyed, it enables more meaningful communication between clinicians, patients, and their families.

Acknowledgments

This research was partially supported by the National Institute on Aging Grant R01AG029476. We are especially grateful to the study participants. We also thank the volunteers in the Aging and Adult Language Lab at Portland State University for assistance with language analyses.

References

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246. [PubMed: 2320703]

Bollen, KA. Structural equation modeling with latent variables. New York: Wiley; 1989.

Brink TL, Yesavage JA, Lum O, Heersema P, Adey MB, Rose TL. Screening tests for geriatric depression. Clinical Gerontologist. 1982; 1:37–44.

Brown, T. Confirmatory factor analysis for applied research. New York, NY: Guilford; 2006.

Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analyzing lexical performance. Aphasiology. 2000; 14(1):71–91.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitraitmultimethod matrix. Psychological Bulletin. 1959; 56:81–105. [PubMed: 13634291]

Chen F, Bollen KA, Paxton P, Curran PJ, Kirby JB. Improper solutions in structural equation models: Causes, consequences, and strategies. Sociological Methods and Research. 2001; 29:468–508.

Chotlos JW. Studies in language behavior. IV. A statistical and comparative analysis of individual written language samples. Psychological Monographs. 1944; 56:75–111.

Cossette, A. La Richesse Lexicale et sa Mesure. Paris: Slatkine-Champion, Geneva; 1994. Number 53 in Travaux de Linguistique Quantitative

Covington, MA. CASPR Research Report 2007–05. MATTR User Manual. Athens, Georgia: The University of Georgia; 2007.

Covington MA, McFall JD. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). Journal of Quantitative Linguistics. 2010; 17:94–100.

Day, A. Good dog Carl. New York, NY: Simon & Schuster; 1985.

Dugast D. Sur quoi se fonde la notion d'etendue theoretique du vocabulaire? Le Francais Moderne. 1978; 46:25–32.

Durán P, Malvern D, Richards B, Chipere N. Developmental trends in lexical diversity. Applied Linguistics. 2004; 25(2):220–242.10.1093/applin/25.2.220

Eid, M.; Diener, E. Handbook of multimethod measurement in psychology. Washington, DC: American Psychological Association; 2006.

Enders, CK. Applied missing data analysis. New York, NY: Guilford Press; 2010.

Ertmer DJ, Strong LM, Sadagopan N. Beginning to communicate after cochlear implantation: oral language development in a young child. Journal of Speech, Language, and Hearing Research. 2002; 46:328–40.

Fergadiotis G, Wright HH. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology. 2011; 25(11):1414–1430. [PubMed: 23125474]

Fergadiotis G, Wright hH, West TM. Measuring Lexical Diversity in Narrative Discourse of People With Aphasia. American Journal of Speech-Language Pathology. 2013; 22:S397–S408. [PubMed: 23695912]

Folstein MF, Folstein SE, McHugh PR. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12(3):189–198.10.1016/0022-3956(75)90026-6 [PubMed: 1202204]

Geiser, C. Doctoral Dissertation. 2008. Structural Equation Modeling of Multitrait-Multimethod-Multioccasion Data.

Goodglass, H.; Kaplan, E.; Barresi, B. The Boston diagnostic aphasia examination; the assessment of aphasia and related disorders. 3. Baltimore: Lippincott Williams & Wilkins; 2001.

Gordon JK. Measuring the lexical semantics of picture description in aphasia. Aphasiology. 2008; 22(7–8):839–852.10.1080/02687030701820063 [PubMed: 22399832]

Heap, HS. Information retrieval – computational and theoretical aspects. New York: Academic Press; 1978.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55.

Jarvis S. Short texts, best-fitting curves and new measures of lexical diversity. Language Testing. 2002; 19(1):57–84.

Kane, M. Validating the interpretations and uses of test scores. In: Lissitz, RW., editor. The concept of validity. Charlotte, NC: Information Age Publishing; 2009. p. 939-64.

Kapantzoglou, M.; Fergadiotis, G.; Restrepo, MA. Lexical diversity and language sample elicitation effects in Spanish-speaking children with and without language impairment. Paper session at the 28th World Congress of the International Association of Logopedics and Phoniatrics (IALP); Athens, Greece.. Aug. 2010

Kenny DA, Kashy DA. Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. Psychological Bulletin. 1992; 112:165– 172.

Kertesz, A. Western aphasia battery-revised. New York: Grune and Stratton; 2007.

Kline, RB. Principles and practice of structural equation modeling. 2. New York, NY: Guilford Press; 2005.

Koizumi R, In'nami Y. Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. System. 2012; 40(4):554–564.

Krahn, F. The great ape; being the true version of the famous saga of friendship and adventure. New York: Viking Press; 1978.

Lance CE, Noble CL, Scullen SE. A Critique of the Correlated Trait–Correlated Method and Correlated Uniqueness Models for Multitrait-Multimethod Data. Psychological Methods. 2002; 7(2):228–244. [PubMed: 12090412]

Lind M, Kristoffersen KE, Moen I, Simonsen HG. Semi-spontaneous oral text production: Measurements in clinical practice. Clinical Linguistics & Phonetics. 2009; 23(12):872–886.10.3109/02699200903040051 [PubMed: 20001304]

Lopez MN, Charter RA, Mostafavi B, Nibut LP, Smith WE. Psychometric properties of the Folstein Mini-Mental State Examination. Assessement. 2005; 12(2):137–144.

Maas HD. Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. Zeitschrift für Literaturwissenschaft und Linguistik. 1972; 8:73–79.

MacWhinney, B. The CHILDES project: Tools for analyzing talk, Volume 1: Transcription format and programs. 3. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2000.

Malvern, DD.; Richards, BJ. A new measure of lexical diversity. In: Ryan, A.; Wray, A., editors. Evolving models of language. Clevedon, UK: Multilingual Matters; 1997. p. 58-71.

McCarthy, PM. Doctoral Dissertation. ProQuest Nursing & Allied Health Source database; 2005. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual lexical diversity.

McCarthy PM, Jarvis S. Vocd: A theoretical and empirical evaluation. Language Testing. 2007; 24(4): 459–488.10.1177/0265532207080767

McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods. 2010; 42(2):381–392. [PubMed: 20479170]

McCarthy, PM.; Watanabe, S.; Lamkin, TA. The gramulator: A tool to identify differential linguistic features of correlative text types. In: McCarthy, PM.; Boonthum-Denecke, C., editors. Applied natural language processing: Identification, investigation, and resolution. Hershey, PA: IGI Global; 2012. p. 312-333.

McCully, EA. Picnic. China: Harper Collins; 1984.

McCarthy P, Watanabe S, Lamkin TA. The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. in press.

McKee G, Malvern D, Richards B. Measuring vocabulary diversity using dedicated software. Literary and Linguistic Computing. 2000; 15(3):323–337. Retrieved from http://childes.psy.cmu.edu/manuals/vocd.doc.

Muthén, LK.; Muthén, BO. Mplus User's Guide. 6. Los Angeles, CA: Muthén & Muthén; 1998–2011.

Nicholas LE, Brookshire RH. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. Journal of Speech & Hearing Research. 1993; 36(2):338–350. [PubMed: 8487525]

Owen AJ, Leonard LB. Lexical diversity in the spontaneous speech of children with specific language impairment: Application of *D*. Journal of Speech, Language & Hearing Research. 2002; 45(5):927.

Richards, BJ.; Malvern, DD. Quantifying Lexical Diversity in the Study of Language Development. The University of Reading New Bulmershe Papers, Reading; 1997.

Rider JD, Wright HH, Marshall RC, Page JL. Using semantic feature analysis to improve contextual discourse in adults with aphasia. American Journal of Speech-Language Pathology. 2008; 17(2): 161–172.10.1044/1058-0360(2008/016) [PubMed: 18448603]

Satorra, A.; Bentler, PM. Corrections to test statistics and standard errors in covariance structure analysis. In: van Eye, A.; Clogg, CC., editors. Latent variable analysis in developmental research. Thousand Oaks, CA: Sage; 1994. p. 285-305.

Satorra A, Bentler PM. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika. 2001; 66:507–514.

Smith SL. Early phonological and lexical markers of reading disabilities. Reading and Writing. 2009; 22(1):25–40.

Steiger, JH.; Lind, JC. Statistically-based tests for the number of common factors. Paper presented at the annual spring meeting of the Psychometric Society; Iowa City, IA.. 1980 May.

Templin, M. Certain language skills in children. Minneapolis: University of Minneapolis Press; 1957.

Thordardottir ET, Namazi M. Specific language impairment in French-speaking children: Beyond grammatical morphology. Journal of Speech, Language, and Hearing Research. 2007; 50(3):698–714.

Tweedie FJ, Baayen RH. How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities. 1998; 32(5):323–352.

Widaman KF. Hierarchically nested covariance structures models for multitrait– multimethod data. Applied Psychological Measurement. 1985; 9:1–26.

Yule, GU. The statistical study of literary vocabulary. New York, NY, US: Cambridge University Press; 1944.

Youmans G. A new tool for discourse analysis: the vocabulary management profile. Language. 1991; 67:763–789.

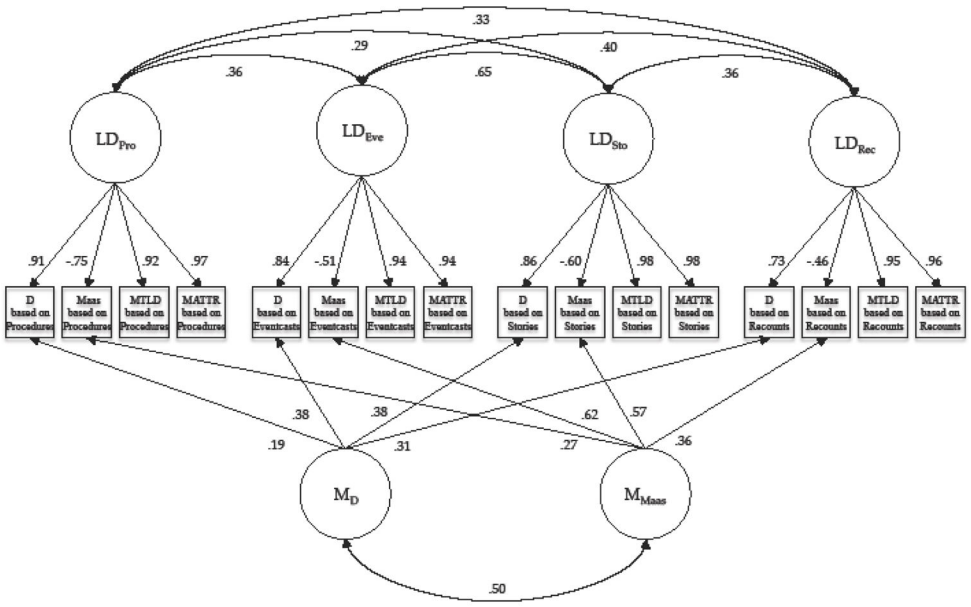

Figure 1. Model 5

LD factors are based on procedures, eventcasts, stories, and recounts. Method factors are based on D and Maas.

Table 1

Participants' Demographic Information

Characteristic	Male (n=186)		Female (n=256)	
Ethnicity				
African-American	14	7.53%	20	7.81%
Hispanic	7	3.76%	7	2.73%
Other	7	3.76%	8	3.13%
White	158	84.95%	221	86.33%
Education level completed				
Some high school	3	1.61%	0	0.00%
12th grade	25	13.44%	30	11.72%
Some college	52	27.96%	80	31.25%
Bachelor's or higher	106	56.99%	146	57.03%
Age				
M (*SD*)	52.09 (20.19)		52.13 (18.60)	
Range	20–89		20–87	
MMSE	54.38		54.72	
GDS	1.36		1.11	

Note. MMSE = Mini Mental State Examination (Folstein, Folstein, & McHugh, 2002); GDS = Geriatric Depression Scale (Brink et al., 1982).

Table 2

Descriptive Statistics of the Number of Types, Tokens, and Type-Token Ratios for Each Type of Discourse

Lexical Diversity Index	*n*	*M*	SD	Range	Skewness	Kurtosis
Procedures						
Types	434	81.21	35.58	30 – 404	2.82	17.46
Tokens	434	192.43	109.33	53 – 1224	3.25	21.12
TTR[a]	434	0.45	0.08	0.25 – 0.66	0.3	0.01
Eventcasts						
Types	441	185.96	54.55	74 – 507	1.17	3.26
Tokens	441	448.56	192.5	118 – 2044	2.2	11.76
TTR[a]	441	0.44	0.06	0.25 – 0.63	0.35	0.41
Storytelling						
Types	441	310.18	80.83	115 – 639	0.87	1.19
Tokens	441	1048.45	398.97	295 – 2997	1.42	3.2
TTR[a]	441	0.31	0.04	0.17 – 0.47	0.09	0.27
Recounts						
Types	434	180.82	101.36	53 – 875	2.42	9.32
Tokens	434	472.4	432	85 – 3775	3.73	18.74
TTR[a]	434	0.44	0.09	0.2 – 0.71	0.02	0.01

[a]Type token ratio

Table 3

Descriptive Statistics of the Untransformed Major Study Variables after the Removal of Outliers

Lexical Diversity Index	*n*	*M*	*SD*	Range	Skewness	Kurtosis
Procedures						
D	430	37.03	11.70	14.82 – 75.41	0.71	0.29
Maas	434	60.91	10.05	37.05 – 89.34	0.16	−0.23
MTLD[a]	432	33.15	8.75	18.27 – 62.94	0.86	0.50
MATTR[b]	434	168.63	13.51	133 – 202.75	−0.03	−0.29
Eventcasts						
D	439	61.91	11.71	31.5 – 100.34	0.5	0.33
Maas	440	106.70	12.40	68.04 – 141.49	−0.25	0.15
MTLD[a]	438	50.30	10.72	25.45 – 86.17	0.64	0.40
MATTR[b]	441	300.02	11.78	264 – 331.2	−0.09	0.08
Storytelling						
D	439	57.53	13.36	22.78 – 101.63	0.77	0.62
Maas	437	133.72	11.06	101.3 – 161.12	0.11	−0.36
MTLD[a]	437	42.76	9.72	21.95 – 75.55	0.68	0.56
MATTR[b]	441	292.85	13.95	252.4 – 330.8	−0.10	0.01
Recounts						
D	433	64.40	14.54	28.44 – 112.15	0.30	0.07
Maas	430	105.25	16.60	56.06 – 155.77	0.04	−0.09
MTLD[a]	430	53.14	12.30	21.09 – 98.98	0.68	0.61
MATTR[b]	437	306.92	13.00	265.2 – 344.4	−0.03	0.04

[a]Measure of textual lexical diversity;

[b]Moving average type-token ratio.

Table 4

Goodness-of-Fit Indices for Structural Equation Models

Model	Model Description	χ^2	*df*	CFI	RMSEA (90% CI)	SRMR
Model 1	Single latent variable	3888.62*	104	0.4	.29 (.28 –.29)	0.21
Model 2	Four uncorrelated content factors	1112.63*	104	0.84	.15 (.14 – .16)	0.26
Model 3	Four correlated content factors	757.53*	98	0.89	.13 (.12 – .14)	0.08
Model 4[a]	Four correlated content factors; four uncorrelated method factors	302.63*	82	0.97	0.08 (.07 – .09)	0.07
Model 4a	Four correlated content factors; two uncorrelated method factors (M_D, M_{Maas})	327.69*	90	0.96	0.08 (.07 – .09)	0.07
Model 5	Four correlated content factors; two correlated method factors (M_D, M_{Maas})	274.86*	89	0.97	0.07 (.06 – .08)	0.07
Model 3′	Four correlated content factors; truncated language samples	300.56*	98	.98	0.07 (.06 – .08)	0.03
Model 5′[a]	Four correlated content factors; two uncorrelated method factors (M_D, M_{Maas}); truncated language samples	282.69*	89	.98	0.07 (.06 – .08)	0.03

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standard Root Mean Residual.

[a] Model converged to an inadmissible solution.

* $p < .001$.

Table 5

Standardized Factor Loadings and Residual Variances for Model 5

	Factor loadings						
Lexical diversity measures	**LD_{Pro}**	**LD_{Eve}**	**LD_{Sto}**	**LD_{Rec}**	**M_D**	**M_{Maas}**	**Residual Variance**
Procedures							
D	.91 **				.19 **		.14
Maas	−.75 **					.27 **	.37
MTLD[a]	.92 **						.16
MATTR[b]	.97 **						.06
Eventcast							
D		.84 **			.38 **		.16
Maas		−.51 **				.62 **	.35
MTLD[a]		.94 **					.12
MATTR[b]		.94 **					.12
Storytelling							
D			.86 **		.38 **		.12
Maas			−.60 **			.57 **	.32
MTLD[a]			.98 **				.05
MATTR[b]			.98 **				.04
Recounts							
D				.73 **	.31 **		.37
Maas				−.46 **		.36 **	.66
MTLD[a]				.95 **			.10
MATTR[b]				.96 **			.08

[a]Measure of textual lexical diversity;

[b]Moving average type-token ratio.

**p< .001.

Table 6

Proportion of Variance for Each Factor and the Residuals for Model 5

	Variance due to each factor						
Lexical diversity measures	**LD_{Pro}**	**LD_{Eve}**	**LD_{Sto}**	**LD_{Rec}**	**M_D**	**M_{Maas}**	**Residual Variance**
Procedures							
D	.83				.04		.14
Maas	.56					.07	.37
MTLD[a]	.85						.16
MATTR[b]	.94						.06
Eventcast							
D		.71			.14		.16
Maas		.26				.38	.35
MTLD[a]		.88					.12
MATTR[b]		.88					.12
Storytelling							
D			.74		.14		.12
Maas			.36			.32	.32
MTLD[a]			.96				.05
MATTR[b]			.96				.04
Recounts							
D				.53	.10		.37
Maas				.21		.13	.66
MTLD[a]				.90			.10
MATTR[b]				.92			.08

[a]Measure of textual lexical diversity;

[b]Moving average type-token ratio. For each combination of type of discourse and estimation technique, the sum of the variance attributed to different sources sums up to 1 (within rounding error).