Modeling Lexical Diversity

Across Language Sampling and Estimation Techniques

by

Gerasimos Fergadiotis

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2011 by the
Graduate Supervisory Committee:

Heather Wright, Chair
Richard Katz
Samuel Green

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

Lexical diversity (LD) has been used in a wide range of applications, producing a rich history in the field of speech-language pathology. However, for clinicians and researchers identifying a robust measure to quantify LD has been challenging. Recently, sophisticated techniques have been developed that assert to measure LD. Each one is based on its own theoretical assumptions and employs different computational machineries. Therefore, it is not clear to what extent these techniques produce valid scores and how they relate to each other. Further, in the field of speech-language pathology, researchers and clinicians often use different methods to elicit various types of discourse and it is an empirical question whether the inferences drawn from analyzing one type of discourse relate and generalize to other types.

The current study examined a corpus of four types of discourse (procedures, eventcasts, storytelling, recounts) from 442 adults. Using four techniques (D; Maas; Measure of textual lexical diversity, MTLD; Moving average type token ratio, MATTR), LD scores were estimated for each type. Subsequently, data were modeled using structural equation modeling to uncover their latent structure.

Results indicated that two estimation techniques (MATTR and MTLD) generated scores that were stronger indicators of the LD of the language samples. For the other two techniques, results were consistent with the presence of method factors that represented construct-irrelevant sources. A hierarchical factor analytic model indicated that a common factor underlay all combinations of types of

discourse and estimation techniques and was interpreted as a general construct of LD. Two discourse types (storytelling and eventcasts) were significantly stronger indicators of the underlying trait.

These findings supplement our understanding regarding the validity of scores generated by different estimation techniques. Further, they enhance our knowledge about how productive vocabulary manifests itself across different types of discourse that impose different cognitive and linguistic demands. They also offer clinicians and researchers a point of reference in terms of techniques that measure the LD of a language sample and little of anything else and also types of discourse that might be the most informative for measuring the LD of individuals.

To Maria, our unborn son, and, my family and friends in Greece.

ACKNOWLEDGMENTS

Writing a dissertation can be a lonely and isolating experience, yet it is obviously not possible without the personal and practical support of numerous people. So, first and foremost, I would like to thank my advisor, Heather Harris Wright, for her endless guidance and support during my Ph.D. study and research. I appreciate all her contributions of time, ideas, and funding which made my Ph.D. experience productive and stimulating. She has been a coach and a cheerleader, believing that I would finish even when I doubted. It has been an honor to be her student. In this same vein, I would like to thank both of my committee members, Sam Green and Richard Katz for their encouragement, insightful comments, and high expectations. Also, I wish to acknowledge and give my appreciation to all the people of the Aging and Adult Language Disorders Lab, without whom, this work as it stands, would not have been possible.

Thank you.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Chapter 1

Literature Review

Discourse is a naturally occurring form of communication that entails the activation and interaction of multiple interconnected cognitive and linguistic subsystems. Because of this, discourse analysis offers an opportunity to observe complex cognitive/linguistic behaviors. And further, it carries the potential of allowing clinicians and researchers to conduct a wide variety of analyses to understand the nature of cognitive-communicative deficits and age-related changes.

It is not surprising then that eliciting and analyzing language samples has been gaining prominence among clinicians and researchers. Language sample analysis has been used as a clinical tool for differential diagnosis (e.g., Fleming & Harris, 2008; Murray, 2009), a key indicator for determining the efficacy of treatment approaches for individuals with aphasia (e.g., Cameron, Wambaugh, Wright, & Nessler, 2006; del Toro, Altmann, Raymer, Leon, Blonder, & Rothi, 2008; Rider, Wright, Marshall, & Page, 2008) as well as an indicator of social validity (e.g., Ballard & Thompson, 1999).

Various content analyses are often used to evaluate the microlinguistic processes that give rise to specific discourse features. Examples include assessments of informativeness and efficiency of a speaker's production. The focus of this paper is on one of the most illuminative predictors of oral performance, *lexical diversity* (LD). LD has been defined broadly as '…something [related to] the range of vocabulary displayed' in different

instantiations of discourse (Durán, Malvern, Richards, & Chipere, 2004; pp. 220). LD has been linked to a wide variety of variables, such as vocabulary knowledge, writing quality, school success, and general characteristics of verbal competence (Avent & Austermann, 2003; Carrell & Monroe, 1993; Grela, 2002; Ransdell & Wengelin, 2003; Verhallen & Scoonen, 1998).

Within the domain of speech-language pathology specifically, LD has been used to ask a wide range of questions in various populations. In the following section, I provide some illustrative examples of how LD has been used and why. Emphasis is placed on applications that focus on the study of language samples for research or clinical purposes within the field of communication disorders. Then, LD will be more formally defined for the purposes of this paper.

Lexical Diversity Research in Communication Disorders

Several studies have focused on whether LD can be used to help differentiate typically developing children (TD) from children with specific language impairment (SLI) (e.g. Kapantzoglou, Fergadiotis, & Restrepo, 2010; Klee, 1992; Klee, Stokes, Wong, Fletcher & Gavin, 2004; Owen & Leonard, 2002; Thordardottir & Namazi, 2007; Watkins, Kelly, Harbers, & Hollis, 1995). For example, Owen and Leonard (2002) analyzed spontaneous language samples from play interactions and found that younger and older children with SLI differed from their age matched peers in terms of how lexically diverse their language samples were. Klee et al. (2004) found similar results when they assessed whether Cantonese-speaking children (27-68 months old) with and

without SLI differed in terms of LD. Based on their findings they concluded that LD could be used to accurately differentiate the two groups. This finding was replicated by Klee, Gavin and Stokes (2007).

Kapantzoglou et al. (2010) used two different tasks to elicit language samples in predominately Spanish-speaking children with and without SLI and explored the classification accuracy based on LD. They compared performance on spontaneous and retell story tasks and found that the type of language elicitation procedure influenced LD scores, which in turn may influence classification accuracy. Also, children with SLI demonstrated low LD scores regardless of the type of the task but TD children performed significantly better than the SLI group on the retell story task only.

In research with children with hearing impairment, LD has often been used as a criterion for evaluating the development of expressive language skills and their improvement after cochlear implantation. Ertmer, Strong, and Sadagopan (2002) examined the language progress of a young, profoundly hearing-impaired girl who had been fitted with a cochlear implant when she was 20 months old. Ertmer et al. used LD measures to quantify the participant's vocabulary growth and compared her spoken output with that of normally developing children. Ertmer et al. were able to document to some extent the developmental trend of vocabulary growth after activating the cochlear implant and pointed out the need for "…longitudinal studies and age-at-implantation comparisons […] to increase understanding of the effects of early implantation on oral language development" (p. 338).

Dillon and Pisoni (2003) explored whether lexicon size, as reflected in LD scores, mediates the relationship between non-word repetition tasks and reading skills in children with cochlear implants. Their study was based on the hypothesis that children's ability to represent phonological units separately from words emerges as a consequence of vocabulary growth. In a sample of 76 children with cochlear implants, non-word repetition significantly correlated with several reading outcome measures (partial correlations ranged from .41 to .55, controlling for age and IQ). However, when LD scores were introduced in the model, the partial correlations were reduced substantially in magnitude (.15 to .32) and were no longer significant. Based on these findings, Dillon and Pisoni argued that as children's LD increases, the robustness of phonological representations is strengthened; which in turn, influences the development of reading skills.

Further, Maner-Idrissi et al. (2009) investigated how several variables such as age at implantation, communication mode before implantation, and school integration level influenced the development of language skills including LD. By videotaping and analyzing language samples from a one year period, Maner-Idrissi et al. found that in a sample of 38 children averaging (3, 66 years) only school integration in a hearing environment impacted LD; a finding that was attributed to "peer pressure" to use spoken language. The authors concluded that certain school environments might be more conducive to language development than others.

In addition, Geers, Spehar, and Sedey (2002) investigated the development of speech and language skills of children who were enrolled in total

communication programs, which make use of multiple modes of communication, after receiving cochlear implantation. Emphasis was placed on identifying predictors of spoken language proficiency because they have been related to children's educational placement in mainstream classes after cochlear implantation. Language samples were obtained that included both sign communication and spoken language. Participants, who were identified as using more spoken language as opposed to sign language, were found to have significantly higher LD compared to children who used more sign language during their interactions. Further, the former group was more likely to be placed in mainstream classrooms.

Researchers who study aphasia in adults have used measures of LD both as an index of general discourse ability and as a tool for hypothesis testing. First, LD has been used to differentiate individuals with aphasia (IWA) from neurologically intact adults (NIA). For example, Holmes and Singh (1996) analyzed conversational language samples from 100 participants, 70 IWA and 30 NIA, in terms of eight linguistic variables, including indices of LD. Their goal was to create a statistical method of assessing an individual's lexical ability that could differentiate the two groups. The results from a discriminant analysis showed that using these variables 88% of the subjects were classified accurately. Further, the analysis showed that LD was one of the most important variables in terms of discriminating power. Lind et al. (2009) also noted the significance of measuring selected aspects of semi-spontaneous discourse in IWA (i.e., not

conversation) and developed a battery of tools to capture clinically relevant aspects of noun and verb production.

Wright, Silverman, and Newhoff (2003) examined whether LD differed across adults with fluent and nonfluent aphasia. Wright et al. analyzed language samples from picture descriptions and manipulated length and LD estimation technique. When language sample length was not controlled, the participants with fluent aphasia yielded significantly higher LD for two out of three LD indices. When samples were truncated to be equal in length, groups differed significantly for all measures.

LD has also been used as an external criterion to investigate the validity of linguistic indices derived from different elicitation techniques. For example, McNeil et al. (2007) included measures that reflect LD and verbal productivity to explore the concurrent validity of the Story Retell Procedure (SRP; Doyle et al., 1998), a method designed to elicit language samples in IWA.

LD has also been used as a crucible for testing hypotheses. For example, Gordon (2008) explored the productive vocabulary of individuals with fluent and non-fluent aphasia in the context of the "division of labor" hypothesis (Gordon & Dell, 2003). According to this hypothesis, some words vary on the extent to which they rely on semantic or syntactic contributions for production. This gives rise to the different speech patterns fluent and non-fluent IWA demonstrate with regard to the number of function and content words they use. Based on the observed diversity of individual word classes, Gordon concluded that results added, at least, partial support to the division of labor hypothesis.

Further, Crepaldi et al. (2011) studied the disproportionate impairment of nouns and verbs in seven IWA in spontaneous speech to examine the functional damage underlying their grammatical-class-specific impairment. Using a similar approach to Gordon's, Crepaldi et al. concluded that their data were consistent with the idea that the noun–verb dissociation might not be as evident in spontaneous speech as it is in picture naming tasks. This finding reinforces the hypothesis that lexical access and retrieval during picture naming and discourse production might be based on different underlying processes with little things in common.

Also, LD has been used to measure the efficacy of treatment studies and generalization to discourse. For example, Rider, Wright, Marshall, and Page (2008) evaluated whether training lexical items using semantic feature analysis would improve the verbal output of individuals with non-fluent aphasia. Using a multiple probes approach, they found that even though their three participants improved in terms of their confrontational naming skills LD did not increase from pre- to post-treatment sessions.

Bucks, Singh, Cuerden, and Wilcock (2000) explored whether the lexical retrieval deficits in dementia are reflected in measures of overall range of vocabulary and whether they can be used to discriminate people with a diagnosis of probable dementia and age matched healthy adults. In their study they used several linguistic variables (including indices of LD) to assess conversational language samples from 24 participants (16 healthy adults). Based on the results, Bucks et al. concluded that the pattern observed suggested that it was possible to

measure lexical differences between the groups that could be used to reliably differentiate them.

Building on previous work from Garrard, Maloney, Hodges, and Patterson (2005), Velzen and Garrard (2008) tracked the gradual decline in LD in three books by Gerard Reve (1923–2006), an acclaimed Dutch author who wrote his last book shortly before being diagnosed with Alzheimer's disease. They split each book into first and second halves and estimated LD for each half. Then, using univariate analyses of variance they found a clear drop in LD that coincided chronologically with when the first reports of forgetfulness started.

Defining Lexical Diversity

LD was defined earlier as '…something [related to] the range of vocabulary displayed' in different instantiations of discourse (Durán, Malvern, Richards, & Chipere, 2004; pp. 220). Durán et al. resorted to this definition in an attempt to reconcile decades of disagreement and confusion regarding the nomenclature and nature of LD. Part of this confusion stems from the fact that the term LD has been used in a wide range of scientific areas in which it has been conceptualized differently (e.g., forensic linguistics, stylometry, bilingualism, aphasia, assessment of first language speaking and writing; see Malvern et al., 2004, pp. 5-15 for a review). The picture is further distorted because researchers have used tools to quantify LD that focus on different aspects of it. As a result, analysis, synthesis, and generalization of findings across studies that could shed light on the nature of LD may often be problematic. Further, according to Yu

(2009), confusion arises because LD has been used to characterize a person's knowledge of vocabulary in some research areas; whereas, LD has been treated as a quality of a verbal or written product in other areas. Yu pointed out that the two could be related, in the sense that a product (e.g., a book) reflects its producer's (e.g., a writer's) vocabulary breadth.

For the purposes of this paper, LD will be defined within Chapelle's (1994) model of vocabulary knowledge. Chapelle, drawing from the area of applied linguistics and the work of Bachman (1990), proposed a model of vocabulary ability that consists of four dimensions. The first dimension is vocabulary size and denotes one's breadth of lexicon that is exhibited in a specific context. The second dimension is knowledge of word characteristics, i.e. aspects of a persons' word knowledge in terms of its phonology, semantics, syntactic properties etc. The third dimension is related to how lexical items are organized in the mental lexicon and how rich the semantic network is. The fourth dimension relates to the processes that are involved in lexical access and retrieval. These dimensions are not meant to be orthogonal nor static; they can vary as individuals develop or as a result of events such as a cerebrovascular accident.

Within this four-dimensional space, LD aligns more closely with vocabulary size and, under certain conditions, the state of the cognitive-linguistic mechanisms that support the access and retrieval of lexical items. The definition of LD offered by Durán et al. (2004) highlights the quantitative aspect of lexical knowledge. Indeed, most researchers agree that LD reflects one's breadth of vocabulary and thus it is more indicative of the lexical knowledge in terms of vocabulary size. LD

does not reflect primarily the depth or the complexity of vocabulary knowledge expressed in Chapelle's second and third dimensions, respectively[1]. In terms of the fourth dimension, exhibiting a range of vocabulary is contingent upon the fundamental processes associated with lexical processing. This is true both with respect to neurologically intact and neurologically impaired adults. Particularly, in the latter case, LD would also reflect the extent to which the cognitive system can support access and retrieval of target lexical items in a given context.

Following Chapelle's perspective, then, language performance is assumed to depend on both the implicit knowledge one possesses (e.g., size of lexicon) and the mechanisms that allow her/him to process it (e.g., access and retrieval). This is in agreement with the idea that *knowledge of* vocabulary and the *capacity to* demonstrate that knowledge cannot be equated (Chomsky, 1980). It is also consistent with clinical neuropsychological performance definitions of language disorders. For example, McNeil and Pratt (2001) de-emphasize the loss of language knowledge as the primary deficit in stroke-induced aphasia and instead recast it as an access deficit.

Based on these premises and for the purposes of this paper, LD_i will be defined as an individual's capacity to deploy a diverse vocabulary, by accessing and retrieving lexical items from a relatively intact knowledge base (i.e., lexicon) for the construction of higher linguistic units. The subscript *i* in LD_i stands for "individual". This definition remains consistent with Chapelle's, according to

[1] Even though it could be argued that it is likely that individuals that have extended vocabularies might also exhibit greater word sophistication and denser semantic networks

whom vocabulary ability reflects "both knowledge of language and the ability to put language to use in context" (Chapelle, 1994, p. 163; see also, Nation, 2007, p. 42 for a similar view). However, it is different in two ways. First, it is tailored to LD rather than vocabulary ability (the latter is considered a superordinate construct that subsumes the former).

Second, it recasts LD as a characteristic of the individual and differentiates it from LD_S which for the remainder of this paper will refer to LD of a given language sample that might take the form of a book, an essay, or telling Cinderella to a child (the subscript *S* denotes sample). Explicitly distinguishing LD_i from LD_S alleviates the confusion identified by Yu (2009) and allows LD_i to be conceptualized as an unobserved trait that characterizes individuals whereas LD_S is considered a quality of a sample. Also, henceforth, LD with no subscript will be used to denote either one when the distinction is not critical or an argument applies to both.

Estimating Lexical Diversity in Language Samples

"A review of the literature on quantifying vocabulary richness gives the sense of a quest for the Holy Grail" (Malvern et al., 2004, p. 3). In this section, why identifying a robust approach to measure LD_S has been challenging will be presented. I will begin by considering some of the major limitations of the most commonly used measures of LD_S, the *type-token ratio* (TTR). First, I discuss Heap's law (Heap, 1978) as it applies to linguistics and more specifically to the study of LD_S. According to this law, the more a speaker talks, the less probable it

is that he/she will produce new words. Holding everything else constant, shorter language samples often appear to be more lexically diverse when using measures such as TTR, rendering comparisons across speakers and language samples problematic. I will discuss the assumptions that underlie one of the most widely used approaches to "salvage" TTR, *truncation*, and why results from this technique might be misleading. Subsequently, examples will be provided from the field of communication disorders that illustrate why interpretations based on the TTR might be biased and inconclusive. Finally, I will present four measures from the field of computational linguistics that claim to produce valid and reliable scores for LD_S and (i) control for length effects at least to some degree, (ii) use the whole language sample to estimate a score without discarding any data, and (iii) are accompanied by some evidence for their psychometric properties.

Type token ratio. The most obvious way to measure LD_S would be to count the number of different words (i.e. types) in a language sample. Types are the unique lexical items that are used in a language sample. For example, the sentence "The birds are playing on the branch" contains the types *the, birds, are, playing, on, branch.* If the samples have the same number of total words (i.e. tokens), then their LD_S could be inferred based on their respective number of tokens. However, when the number of tokens is not kept constant, conclusions based strictly on comparisons of the number of types might be misleading and also not meaningful. Is a sample of 50 tokens that contains 40 types less diverse than a sample of 400 tokens that contains 60 types? Quickly it becomes evident that unless the number of tokens is equal, the number of types would reflect both

LD_S as well as the contribution of length. That is, language samples that were longer would be credited with higher LD_S.

To overcome this obstacle, one could consider the ratio of the types divided by the tokens (TTR) to control for length. TTR has been the traditional method for measuring LD_S (Chotlos, 1944; Templin, 1957). However, even though TTR is an improvement compared to counting the number of different words (NDW), it is still inherently flawed because it also varies as a function of sample length (Heap, 1978). As the sample increases, the probability of introducing new words decreases because the vocabulary that characterizes individuals at any given time is considered finite; as a result, the growth of the numerator in the TTR decelerates. However, the denominator (i.e., the number of tokens) always increases steadily with every additional word produced. As the language sample unfolds over time, the TTR decreases monotonically and forms a hyperbolic curve that asymptotically tends to zero. Therefore, comparisons across language samples of different speakers or even across different samples produced by the same speaker will be confounded by sample length.

To solve this problem, researchers have used various algebraic transformations of TTR (e.g., Root TTR, Guiraud, 1960, Corrected TTR, Carroll, 1964, Herdan's Index, Herdan, 1960, see Table 1). However, even though some authors have reported some success using these indices, it has been demonstrated in a number of studies that these tools have also been found to covary with sample length, thus yielding mathematically and conceptually spurious results (Malvern & Richards, 1997; Tweedie & Baayen, 1998; Jarvis, 2002; Vermeer, 2000). With

regards to Root TTR and Corrected TTR, it has been argued that the transformation only serves to rescale the TTR and does not eliminate the problem. According to Malvern et al., Herdan's Index has shortcomings as well because it is based on the invalid assumption that "…the rate of increase of types with increasing token count […] will be proportional to the TTR for any given value of N" (Malvern et al., p. 27). So, these measures fail because they assume there is a constant relationship between types and tokens. Therefore, mathematical corrections such as the ones just presented may reduce but not completely eliminate the problem of TTR.

Some researchers have proposed standardizing the sample size to overcome the problem of measuring relative length. The most common approach to standardizing length has been through truncation. One of the problems with this is that for results to be comparable across studies, researchers have to agree on the number of tokens to estimate TTR. In aphasiology for example, some researchers have used 300 tokens as a "standard" length (Brookshire & Nicholas, 1994; Prins & Bastiaanse, 2004). However, consensus on this issue is low; the main reason being that it is not always feasible to obtain a predetermined number of tokens. For example, the sample length may depend on the discourse genre that is being produced. When individuals are asked to describe a procedure such as planting a flower in the garden, it is not unusual even for neurologically intact adults to produce samples that are less than 200 tokens. In addition, individuals with aphasia often do not produce long samples, especially individuals with non-fluent types of aphasia. So, it is not uncommon for researchers to ignore consensus and

restrict the number of tokens to be equal to the shortest sample in the study. For example, Gordon (2008) followed this approach and restricted sample lengths to the first 200 content word tokens produced.

Some researchers have argued that discarding any amount of text may reduce a language sample's integrity and may lead to spurious results. Youmans (1991) pointed out that during discourse, the introduction of new information (such as new episodes during storytelling) would coincide with a spike of new vocabulary causing peaks in the TTR/(sample length) curve. Alternatively, if a speaker is discussing a topic with only a few new words being produced prior to providing new information, the TTR/(sample length) curve should appear to decrease or plateau. Jointly, the peaks and valleys would make the TTR/(sample length) curve appear less smooth than predicted. This pattern becomes important when considering truncation. TTR results may vary depending on whether a language sample is truncated during a peak or a valley. Therefore, even comparisons across truncated language samples can include more noise than researchers may expect.

Finally, when exploring LD_S in language-impaired populations there are additional considerations. If not all data are used then is possible that the restricted sample might obscure the findings due to clustering of content words. Prins and Bastiaanse (2004) provide an illuminative example of how truncation can distort results because it violates the assumptions of textual homogeneity. This is demonstrated in a sample from one of their participant's with Broca's aphasia answering two questions during an interview:

1) You have problems finding the words? **Yes, yes.** But, as I understand, you also encounter problems when making a sentence? **Yes, it doesn't come . . . at moment when I write er goes that er slow er no.** When you are writing? **Yes, before the time I did know writing down. Er I write down, nothing remembers me.** Yes, but when you really want to, can you speak in correct sentences? **Yes.** Then why don't you do that? **Er, too fast to talk.** What do you mean, too fast? **Er, I too fast to talk, er, I cannot er search for words.** That's why you talk in short sentences? **Yes, `the', `a', `and' I leave out I just leave er.** Do you do that on purpose? **No, oh god, no.** That happens automatically? **Yes, I hear always what I says. Sentences quick I hear. Er `and', `the' I hear er always er what I says wrong.**

2) Okay, something else, it will soon be Sinterklaas and Christmas. **Yes, yes.** Do you have any plans? **Yes, no, plans not not. Sinterklaas shops business. Me purse always empty. Future, no past.** Won't you celebrate Sinterklaas? **No, absolutely.** Don't you do anything? **In the pan tasty things. Snacks . . . tasty.** But no presents? **No, no.** And at Christmas and New Year's Eve, are you going to do something then? **Er eating tasty things, presents Christmas. Drawing numbers, all er getting presents. Ten guilder, ten guilders each.** You are not going out? **I don't know.**

You don't know. **No, we sold house, our house. New about March. Er er we saving pennies.** (pp. 1085)

Notice how the participant produces agrammatic sentences when answering the first question but later on during the interview, his/her responses become telegraphic with higher density of content words. If the focus of the study had been, for example, to contrast verb lexical diversity, then it becomes clear that the results could have been distorted depending on the length of the sample. Such distortions cannot be foreseen and can (and most probably do) introduce noise in the analysis.[2]

Even though problems with TTR have been known and documented for quite a long time, there are still numerous examples of researchers who report and interpret results from analyses using TTR without taking into consideration its limitations. For example, Ertmer et al. (2003) conducted a longitudinal study to investigate the emergence of language skills of a young child with cochlear implants, which were activated at 20 months of age. To track growth, Ertmer et al. used TTR but did not safeguard against its known problems. Further, they cited Templin's (1957) 0.5 norms to assess the rate with which the child's language skills were attained. Based on the results from the TTR at five time intervals and using TTR, the researchers concluded that the child's language skills were

[2] A variation of truncation involves random sampling of *n* tokens from a sample to create a sub-sample, where *n* = the number of tokens in the shortest sample. However, such a method does not solve the problem completely because the integrity of the sample is still not maintained.

different from that of a typically developing child. Specifically, they asserted that the participant had restricted vocabulary compared to a hearing child of the same age.

Gordon (2008) explored the productive vocabulary of individuals with fluent and non-fluent aphasia in the context of the "division of labor" hypothesis (Gordon & Dell, 2003). According to this hypothesis, some words vary on the extent to which they rely on semantic or syntactic contributions for production. This gives rise to the different speech patterns fluent and non-fluent individuals with aphasia demonstrate with regard to the number of function and content words they use. One of the reasons Gordon chose to use TTR was because using and interpreting the results of a more sophisticated tool such as D (presented later) might not have been as easily interpretable. The language samples were truncated to the first 200 content word tokens produced by each participant. However, language samples from two individuals (out of a sample of 16) included less than 200 content words. Both exceptions were speakers with non-fluent aphasia (the first produced 187 content words; the second only 64). Given the known trend of TTR to decrease as a function of language length, one could predict that the nonfluent group that included the shorter language samples would have an "unfair" advantage; that is, holding all else constant TTR's from shorter language samples would have been higher, driving the group's mean TTR higher. Indeed, Gordon reported two sets of results comparing fluent and nonfluent PWA - one including the score of the nonfluent subject who produced 64 content words (and had the highest TTR across groups) and one without it. Not surprisingly, in the

former case she found a significant difference between the two groups; in the latter, she did not. Gordon's study highlights the inconsistency of results that are based on TTR and the problems with truncation.

Lind et al. (2009) also noted the significance of measuring selected aspects of semi-spontaneous discourse in individuals with aphasia and developed a battery of tools to capture clinically relevant aspects of noun and verb production. Even though they acknowledged the limitations of TTR, they opted to use it after truncating language samples instead of using a tool that takes into account the whole language sample. Lind et al.'s justification for choosing TTR was that the measure is simple and it is easily calculated in clinical practice. However, this approach requires one to make the assumption of textual homogeneity, which as discussed earlier (Prins & Bastiaanse, 2004; Youmans, 1991) is not a plausible one.

In another example that highlights the central weakness of TTR, McNeil et al. (2007) found that their participants with aphasia had significantly lower TTR on the Story Retelling Procedure (SRP; Doyle, McNeil, Spencer, Goda, Cottrell, & Lustig, 1998) compared to other discourse samples (e.g., procedural descriptions). McNeil et al. acknowledged that sample length variations may have contributed to this counterintuitive finding. Indeed, given that the number of words elicited using the SRP was significantly greater than any other elicitation procedure, lower TTR were expected compared to procedural descriptions.

Despite its notorious lack of proper psychometric properties and its many limitations, researchers continue to use TTR. Often researchers make no effort to

correct for the TTR's limitations and use it without controlling for length discrepancies (e.g., Heisler, Goffman, & Younger, 2010). On some occasions, researchers attempt to control for length by standardizing the number of utterances (e.g., Corthals, 2010) or time (e.g., Peets, 2009). However, these approaches do not ensure an equal number of tokens across language samples and further, compromise the integrity of the language samples.

Quantifying lexical diversity using sophisticated measures

Recently, a new generation of tools has emerged from the field of computational linguistics that can be used to measure LD_S. With a few exceptions, these measures have been used in limited applications in the field of speech-language pathology. In the following section I will present three estimation techniques that have been recently developed and an older technique, that is based on a complex logarithmic transformation of the TTR.

D. The first measure is D, originally designed by Malvern and Richards (1997) and further developed by McKee, Malvern, and Richards (2000). D combines an algebraic transformation model and curve fitting to estimate LD_S. The "secret" of D is that it embraces TTR's inevitable and thus predictable fall as sample length increases. For example, consider two language samples and how their TTR would decrease as a function of sample length (Figure 1). For a language sample that is contains a more diverse vocabulary, TTR would decrease at a slower pace (squares). However, for a language sample, that consists of the set of lexical items that are being used repeatedly, TTR would decrease faster as a

function of the sample size (circles). D reflects LD_S by capturing how fast TTR decreases. D appears 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. Further, because D is estimated from the whole transcript using repeated random samplings, it is less prone to measurement error due to clustering of novel content words.

Estimating D involves a series of random text samplings to plot an empirical TTR versus number-of-tokens curve for a sample. First, 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, the estimation of D involves solving the following mathematical formula to produce a theoretical curve that maximizes the fit to the empirical TTR curve using the least squares approach (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 result in steeper theoretical curves that fit the empirical curves of samples with poorer LD_S better. 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.

The validity of D scores has been explored in several studies (Durán et al., 2004; Malvern & Richards, 1997, 2000; Richards & Malvern, 1997, 1998). Analyses have been conducted on samples from typical language learners, children with specific language impairment, and second language learners. In these studies, the model produced estimates of D that strongly correlated with well-validated measures of language as well as developmental and demographic variables (e.g., MLU, age, and socioeconomic status).

However, the validity of D score interpretations has also been questioned. In the Owen and Leonard (2002) study, children with and without language impairment were compared in terms of LD. Initially, significant differences between the two groups were found, both when they were matched in terms of age and when they were matched in terms of mean length of utterance. However, when language samples were truncated to the first 250 words, the differences between the groups were attenuated significantly only when the groups were matched for age. The authors interpreted these findings using length to explain the difference in the results patterns. They argued that when children were matched for age, their language samples varied in the average number of words which in turn might had inflated estimates of D. They followed up these results with another experiment for which they estimated D scores based on language samples that were truncated to the first 500 words. Then they compared the mean estimates of D that were based on the two sample sizes and found significant differences. Based on their results Owen and Leonard concluded that "it appears that D does not entirely avoid the problem of sample size influence" (p. 935).

Measure of textual lexical diversity. Another tool that has been proposed recently for estimating LD_S is the measure of textual lexical diversity (MTLD) (McCarthy, 2005). This index is quite different from D as it is designed to address the skepticism about the random sampling process on which the estimation of D is based (Jarvis, 2002). Jarvis argued that because D is based on repeated random samplings, word order is completely ignored and the text is treated as a collection of randomly selected discrete items. McCarthy (2005) further noted that such an approach (i.e., random sampling) towards LD seems to contradict Van Dijk and Kintsch's (1983) theory that asserts that listeners form coherent mental representations based on the inherent structure of texts that hold textual components together. To address the validity threat stemming from this fact, McCarthy (2005) chose to employ a sequential analysis of the language samples for the estimation of MTLD.

MTLD capitalizes on the TTR's predictable fall as a function of a sample's size and according to McCarthy (2005) it is not affected by length variations. Conceptually, for any given sample, MTLD reflects the average number of words in a row for which a certain TTR is maintained. For language samples with high LD_S, multiple words are required to drop TTR below a certain cutoff score because there is a lower propensity of the same words being repeated. Samples with higher LD_S tend to drop TTR below the predetermined value less than samples with low LD_S.

During the estimation process (Figure 2), each word of the language sample is evaluated sequentially for its TTR. 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, the 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. Then, the forward and the reversed MTLD scores are averaged to provide the final MTLD estimate.

McCarthy and Jarvis (2010) investigated several aspects of validity of the MTLD scores. For example, they used texts from 16 registers such as press reportage, popular lore, biographies, official documents, academic prose, science fiction, humor, and textbooks to investigate convergent and divergent validity. Commonly used indices of LD_S were used that included Maas (Maas, 1972), D, and Yule’s K (Yule, 1944). Across all registers, MTLD was found to correlate moderately to strongly with all three indices. Correlations were -.843, .694, and .848, respectively. McCarthy and Jarvis argued that based on these results, convergent validity was supported.

Maas index. Another approach researchers have used in the past to address the inherent flaw of TTR is linearizing. This approach is based on the fact that the TTR curve can be fit relatively well by a logarithmic curve. Figure 3 shows a well-behaved empirical curve (green) that reflects how TTR drops as a function of

sample size. Further, it demonstrates how well a theoretical logarithmic curve can be fit to the data (red).

In theory, if one could transform the relationship between N and TTR to achieve linearity, it would be fairly straightforward to use the machinery of regression analysis to estimate the slope of the line that would be constant regardless of sample size. In that case, estimating the slope of the curve would be sufficient to describe the LD associated with a language sample irrespective of the sample length. Herdan (1960) was the first who applied the logarithmic transformation to the TTR. However, Herdan's index is not stable because it relies on invalid statistical assumptions with respect to the frequency with which new lexical items are produced as the token count increases.

Consequently, more complex logarithmic transformations have been used to change the curved relationship between N and TTR. Indices that are based on more elaborate linearizing formulas include Somers (1966), Maas's (1972), Dugast's (1978), and Tuldava's (1993). Even though the validity evidence of these approaches has been questioned in the past (Hess, Sefton, & Landry, 1986; Tweedie & Baayen, 1998), in a recent study it was shown that one measure from this family of LD indicators, *A* (Maas, 1972) might be a better indicator of LD:

Maas (1972): $$a^2 = \frac{(logTokens - logTypes)}{log^2 Tokens} \quad (2)$$

Specifically, McCarthy and Jarvis (2007) investigated the performance of 14 LD indices including several logarithmically transformed LD indices across a

number of written and spoken genres. In terms of length effects, even though all indices correlated significantly with token count, Maas outperformed all other log indices and demonstrated practically no effect of text length (approx. R^2=.02).

Moving average type token ratio. A new measure of estimating the LD_S was introduced by Covington and McFall (2010) in a paper titled "Cutting the Gordian Knot: the moving-average type-token ratio (MATTR)". MATTR calculates LD_S scores by using a smoothly moving window that estimates type-token ratios for each successive window of fixed length (see Figure 4). Initially, a window length is selected (in this case ten words) and then 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 TTR's are averaged.

Covington and McFall note a number of features that make MATTR an attractive index of LD_S. Unlike other similar approaches that have been proposed in the past, MATTR does not discard any data. For example, MATTR is similar to the mean segment TTR (MSTTR; Johnson, 1944), an approach that estimates the LD of a sample by computing the TTR from successive **non-overlapping** segments of the sample and then averaging them. However, unlike the MSTTR that necessarily discards the data at the end of the sample that are less than the predetermined size of the moving window, MATTR "crawls" successively until it has reached the final token in the sample. Further, MATTR does not rely on any statistical assumptions like the Maas index. In addition, as long as the same window size is used for all the language samples (e.g., 50 tokens) the estimates should be, theoretically, independent of sample length.

Eliciting Language to Measure LD

Effects of language sampling technique. In studies that focus on language sample analysis, researchers have used different elicitation tasks to obtain language samples depending on their research questions. For example, participants have been asked to describe common procedures thus producing activity-focused, step-by-step descriptions of how to achieve a goal (e.g., how to plant a flower; Brady, Armstrong, & Mackenzie, 2005; Caspari & Parkinson, 2000; Longacre, 1996). In other studies, researchers have asked participants to describe pictorial stimuli (Christiansen, 1995; Nicholas & Brookshire 1993; Olness, 2006; Wright & Capilouto, 2009). The type of discourse most often elicited with this method is called eventcast. Eventcasts are narratives that explain a scene of activities (e.g., Cookie Theft Picture). Further, narratives have been elicited through telling of familiar stories (i.e., storytellings) and/or sharing past experiences (i.e., personal narratives or recounts) (Ash, Moore, Antani, McCawley, Work, & Grossman, 2006; Coelho, Grela, Corso, Gamble, & Feinn, 2005; Hough & Barrow, 2003; Ulatowska, North, & Macaluso-Haynes, 1981). Stories are fictionalized, highly structured forms (e.g., Cinderella), whereas recounts are verbal reiterations of an event (e.g., what one did last weekend) (Heath, 1986).

There is general consensus among researchers and clinicians that different types of discourse are associated with different cognitive and linguistic demands (Bliss & McCabe, 2006; Brady et al., 2005; Nicholas & Brookshire, 1993; Ulatowska, Allard, & Chapman, 1990). These differences are reflected in the

verbal output of the speakers. That is, instances of the same discourse type that are elicited using the same technique are often characterized by structural similarities. Therefore, when eliciting more than one type of discourse, performance on some indices that assess microlinguistic (e.g., syntactic complexity) and/or macrolinguistic (e.g., story elements) aspects of verbal production may vary systematically (Li, Volpe, Ritterman, & Williams, 1996; MacLachlan & Chapman, 1988).

Recently, LD_S was also found to vary as a function of the discourse type or the elicitation technique one chooses to use. In a recent study, Fergadiotis, Wright, and Capilouto (2011) examined how LD_S is influenced by the effect of various elicitation techniques in cognitively healthy adults. Four commonly used tasks for research and clinical purposes were included; procedures, eventcasts, storytelling, and recounts. 86 cognitively healthy adults participated in the study and were grouped into two age groups - young (20-29 years old) and old (70-89 years old). Fergadiotis et al. found a LD_S hierarchy that was similar across age groups for the four discourse types. Procedural discourse yielded the lowest LD_S followed by eventcasts and storytelling, with recounts yielding the highest LD_S values.

Why LD_S varies systematically with respect to discourse type is not well understood but several hypotheses have been put forward. Bliss and McCabe (2006) have suggested that storytelling from pictorial stimuli such as wordless picture books are associated with the elicitation of a rich vocabulary because of the contents of the illustrations. For example, tasks that provide pictorial support

may be associated with a higher propensity for eliciting a variety of concrete and high-imageability words (Balota & Chambley, 1984; Grosjean, 1980; Kroll, 1986; Tyler & Wessels, 1983). The additional information provided by the pictorial stimuli may be used to prime the semantic/conceptual content of the target words, and as a result increase their activation leading to easier retrieval.

An alternative explanation is that the scaffolding provided by certain elicitation techniques may serve as a cognitive map or schema. When rich contextual information is provided by the task, speakers may spend fewer resources to retrieve details from memory or plan and organize their discourse. In this case, under the assumption that the autonomous low-level cognitive process of lexical access shares resources with higher order cognitive processes (Rabovskya, Álvarez, Hohlfeld, & Sommer, 2008), lexical access and retrieval may be benefited.

Conversely, when tasks do not provide contextual support, lexical access and retrieval and therefore LD_S may be decreased. This is true especially in the case of people with aphasia, who have limited resources for linguistic processing. For example, Fergadiotis and Wright (2011), who examined the effect of discourse elicitation technique in terms of LD_S in people with aphasia and neurologically intact adults, found a significant interaction between type of discourse and group membership. Specifically, the language samples of the two groups differed significantly more when telling culturally familiar stories in the absence of pictorial stimuli than telling stories based on pictures. Further, with respect to methods that use single versus sequential pictures to elicit discourse,

Wright and Capilouto (2009) demonstrated that narratives elicited from single picture stimuli were associated with lower LD compared to narratives elicited with sequential picture stimuli. In addition, Capilouto, Wright, and Wagovich (2005) argued that sequential stimuli might offer additional temporal and causal information about the depicted story. As a result, when responding to sequential pictures, participants' narratives are often more complex in terms of characters and events. To convey representations that entail a larger number and/or more complicated interactions among the elements of the story would require the introduction of specific vocabulary; which in turn may increase the likelihood of sampling from a wider variety of lexical items and thus producing more diverse vocabulary.

Others have argued that open tasks that do not "restrict" participants in terms of the content might allow speakers to produce language samples with higher LD_S. For example, O'Loughlin (1995) investigated the effects of task format and type on *lexical density* (a ratio of frequency weighted lexical items divided by the sum of lexical and grammatical items which reportedly also measures LD) in neurologically intact adults. O'Loughlin found that personal narratives were associated with higher lexical density compared to picture descriptions and suggested that "…'open' tasks seem to elicit language with a higher degree of lexical density than 'closed' tasks […] because candidates are not constrained by any stimulus material" (p. 234).

Context and discourse production. Currently, there is a lack of a satisfying and coherent account for how specific aspects of elicitation techniques

affect discourse production. In spite of this, there is a general consensus that discourse processing entails the formulation and expression of a communicative intent *within a specific context*, which led Goffman (1981) to argue that discourse is "language in use" and serves a social purpose. More specifically, discourse production involves the translation of conceptual knowledge into discourse structures that are appropriate in a communicative situation (Frederiksen, 1996). This process involves building a conceptual representation of an idea to be communicated within a given context. Van Dijk and Kintsch (1983) termed the product of this process "situation model". Specifically, a situation model is an amalgam of bits and pieces of information from a minimally organized general factual and/or procedural knowledge base (as defined by Tulving, 1972) *and* a context specific communicative intent. Another important component of this process is the construction of a map that conveys the micro- and macrostructure of the intended message that corresponds to the situation model. This map, termed text base, eventually contributes to the selection and formation of linguistic units.

Halliday and Hasan's (1989) work was focused primarily on vocabulary knowledge but their work is closely aligned with van Dijk and Kintsch's, Goffman's and Frederiksen's. Haliday and Hasan argued that the selection of lexical items is heavily influenced by contextual effects and may vary qualitatively and/or quantitatively. They also suggested a decomposition of context into three complex elements: *field*, *tenor*, and *mode*. Field includes the setting and the topic of discourse; tenor refers to the interlocutors, their relationship and objectives; and, mode refers to parameters such as the channel

and type of discourse being produced. For example, because of differences in these three elements, one's use of lexical items might differ when recounting casually a previous experience to friends compared to arguing in favor of a controversial theory at a conference.

Similarly, when asked to tell a story, it is likely that one will produce a narrative with a structured story format that emphasizes the interplay of story elements (i.e., agents, events) along a temporal continuum. If one is asked to describe a procedure, it is probable that they will produce procedural discourse focusing only on the basic steps that need to be conveyed. And, when asked to produce a recount, speakers are more likely to allocate resources to memorial processes to retrieve stored representations. The likelihood of responding using the targeted type of discourse is directly related to a number of factors such as how specific the instructions are and how easy they are to follow, whether the task was modeled or not; and, the quality of the stimuli and the type of feedback participants receive during practice. Under this assumption of elicited discourse type homogeneity, following Halliday and Hasan, elicitation techniques can be considered part of the context within which a speaker produces discourse. Further, systematic manipulation of task elicitation technique would presumably give rise to systematic variation of indices that reflect vocabulary knowledge such as LD.

Validity

Bachman (2003), Kane (1992, 2001, 2006, 2009), Mislevy (2006), and Mislevy and Yin (2009) describe how assessing psychological abilities may be

viewed as a process of evidentiary reasoning, which in turn constitutes a special case of argument (Toulmin, 1958). Specifically, measurement may be viewed as an argument that entails a logical inference from how individuals are perceived to perform under particular conditions (e.g., what they say or do) to the individuals' capacities more broadly. For this study, measuring LD_i (i.e., the LD of an individual) is akin to using an argument to reach conclusions about a speaker's cognitive skills after having observed an LD score. Within this framework, assessing validity becomes equivalent to assessing whether it is reasonable to draw inferences about specific unobserved skills based on observed data.

Toulmin (1958) created a system for judging the rationality of assertions. Figure 5 outlines the structure of a simple argument that consists of seven interrelated components: the claim, the data, the warrant, the qualifier, the backing, the alternative explanation, and the rebuttal evidence. The *claim* is what one wished to support as a valid conclusion. Unless the claim was made wildly, the claim should have a factual foundation upon which it is based; this foundation consists of the *data*. The bolded arrow between the data and the claim represents inference, or in other words the logical bridge from the data to the claim. Toulmin refers to this part of the argument as an "inference-license" or *warrant* (p. 91). For example, in the context of this paper, the claim could be that individual *i* has low LD_i. The claim would be based on data which in this case could be individual *i*'s observed LD score in discourse type *y* that was estimated using approach *z*. Given that the datum and the claim are not the same, a logical step is required to argue that individual *i* has low LD based on his/her observed score.

The legitimacy of the inference depends on the warrant. The warrant corresponds to a statement that can be used to authorize the logical step from the data to the claim. For example, in Figure 5, an inference has to be made to reach the conclusion that an individual has low LD after observing a particular LD score (bold arrow). The inference is based on a warrant according to which "*data reflect the construct of interest*". More often than not, the bearing of the data on the conclusion is not stated explicitly, nevertheless it constitutes an essential part of the argument (if not the most crucial one). Further, the warrant is not the same as the inference because one can make an inference without a warrant or an inference may be based on an erroneous warrant. Theory, prior research, and experience provide the backing for the warrant.

Further, a particular set of data might be open to *alternative explanations* (e.g., a participant has memorized somebody else's lexically rich story). The possibility of alternative explanations necessitates the stipulation of safeguards to better support the step between the data and the claim using the original warrant. Finally, it might be necessary sometimes to go beyond just specifying the data, the warrant and the claim. Depending on the application, it may be required to supplement the argument structure with a *qualifier.* Qualifiers reflect the certainty with which the warrant allows us to draw inferences based on the data.

Toulmin's argument structure is at the core of many current philosophical and psychometric theories of validity that make the case that it is much more useful to talk about the *validity of the inferences drawn based on scores* rather than the validity of tests or tasks. For example, Gorin (2007) defined validity as

"the extent to which test scores provide answers to targeted questions" (p. 456). Zumbo (2007) elaborated by adding "test score validation is an ongoing process wherein one provides evidence to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores about individuals" without which "…any inferences made from a measure are potentially meaningless, inappropriate and of limited usefulness" (p. 48). Bachman and Palmer (1996) further argued, "test developers and test users must provide adequate justification for any interpretation we make of a given test score […] and not simply assert or argue that they are valid" (p. 21). And finally, Borsboom (2005) identified the causal relationship between the manifest variables and the latent construct a key ingredient of validity.

In reality, it has been argued that an assessment argument might be more complex than the one presented earlier and might involve a series of reasoning steps (e.g., Mislevy & Yin, 2009). In terms of LD, for an investigator to make a claim about an individual, he/she has to overcome at least two hurdles identified earlier in the introduction. First, there is uncertainty in the measurement process that stems from the lack of perfect techniques to estimate LD scores. As mentioned earlier, indices of LD have been notorious for systematically misrepresenting the LD_S. So, the extend to which investigators can draw inferences and make claims about the LD_S is not always clear. Consider the following example. Assume that a speech-language pathologist uses TTR to estimate the LD of a language sample of an individual recovering from a stroke. If the score is low, the speech-language pathologist might reach the conclusion that

the individual produced a language sample with poor LD_S. To make this claim, the speech-language pathologist would have to make an inference (knowingly or unknowingly) about the LD_S based on the TTR score he observed. The inference though rests upon the assumption (i.e., warrant) that TTR reflects the sample's LD. However, not only is there no evidence to back up this warrant, but there is ample evidence that TTR is influenced systematically by other factors such as sample length. For the remainder of the paper, the level of inference that involves the logical step of making inferences about the LD_S based on an observed score will be referred to as Level 1.

Then, a second level of inferential reasoning, which will be henceforth referred to as Level 2, will be used to describe the LD_i as it manifests itself across multiple types of discourse. Usually, investigators study language samples because they wish to know something about the individual. However, as noted earlier, it is possible that the type of discourse a speaker chooses to produce and LD might interact. As a result, some types of discourse may be better suited for producing scores that may have greater potential for answering research and clinical questions. For example, a certain type of discourse might reflect LD_i significantly better than others and therefore its potential diagnostic potency might be higher. Of course this level of reasoning, it contingent upon the strength of the inferences that are drawn at the previous level: the strength of the conclusions one reaches about LD_i depends on the capacity of an estimation technique to provide a score that would be a valid indicator of LDs.

Statement of the Problem

LD has been used in a wide range of areas, producing a rich history in speech-language pathology. LD has been used to understand the relationship between phonological processing and the development of reading skills (Dillon and Pisoni, 2001; Smith, 2007); to investigate whether specific instruction techniques are more effective when teaching writing in schools (Cameron et al., 1995); to differentiate typically developing children from children with specific language impairment (Owen & Leonard, 2002; Klee, 1992; Klee, Stokes, Wong, Fletcher & Gavin, 2004; Owen & Leonard, 2002; Thordardottir & Namazi, 2007; Watkins et al., 1995); to screen and identify bilingual children with language deficits (Kapantzoglou et al., 2010); to evaluate the progress of children after cochlear implantation (Ertmer et al., 2002), identify how age of implantation affects subsequent language development Geers et al. (2002), and examine what environments foster faster development of expressive skills after the implantation (Maner-Idrissi, 2009); to differentiate individuals with aphasia from neurologically intact adults (Holmes & Singh, 1996; Lind et al., 2009); to capture the differences between individuals with fluent and nonfluent aphasia (Wright et al., 2003; Fergadiotis & Wright, in press); to validate assessment tools (McNeil et al., 2007); to study the disproportionate impairment of nouns and verbs in aphasia and inform models of language processing (Crepaldi et al., 2011; Gordon, 2008); to assess the efficiency of therapeutic approaches (Rider et al., 2008); and, to assess its potential as an early marker of dementia in individuals with a diagnosis of probable dementia (Bucks et al., 2000).

Confusion often arises in studies of LD because within and across disciplines LD has been used to characterize both a person's knowledge of vocabulary and a quality of a verbal or written product (Yu, 2009) without explicitly distinguishing the two. In this paper, a clear distinction was made between LD_i that refers to the LD as a characteristic of the individual; and LD_S which refers to a language sample. Even though there is a clear relationship between the two, explicitly defining them as distinct concepts alleviates some of the confusion allows us to ask questions that pertain specifically to each one.

Measurement of LD is akin to using an argument to reach conclusions about a speaker or a language sample after having observed an LD score. However, in order to justify a particular score interpretation, one would have to commit to an inference-license, what Toulmin (1958) refers to as a warrant, that would allow this reasoning leap. The warrant could be of the form "the observed score reflects LD_i (or LD_S) and little of anything else" or as Borsboom (2005) put it to stress causality "variations in the attribute causally produce variations in the outcomes of the measurement procedure" (p. 150). To the extent that the warrant is true, one would be allowed to interpret a given score as an indication of the scientific quantity of interest. The accumulation of evidence regarding this aspect of validity, i.e. the degree to which the aforementioned warrant is true, reflects the construct validity of the claims that are made (Bachman & Palmer, 1996; Gorin, 2007; Kane, 1992; Mesick, 1989; Mislevy & Yin, 2009; Zumbo, 2007). Note that in all of the studies cited in the previous paragraph, plus many more, the legitimacy of the warrant had been presupposed.

The validity of LD scores could be threatened at two levels if construct-irrelevant sources of variation 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. First, at Level 1, which focuses on the conclusions drawn about LD_S given an observed score, scores obtained from a specific approach could be found to reflect more than just the LD_S. In that case, it would be difficult to argue in favor of the approach and the construct validity of the conclusions reached when using the specific approach. For example, consider a measure such as the TTR discussed in great detail earlier. It has been shown that there is lack of backing evidence to support the warrant that links the data to the claim; and also, TTR has been shown both mathematically (Heap, 1978) and in practice (e.g., Malvern et al., 2004) to covary with length. That is, it has been shown that the TTR scores will reflect not only the construct of interest, LD, but the effects of length as well. Earlier attempts to solve this problem by applying algebraic formulations to TTR (e.g., Carrol, 1964; Guiraud, 1960; Herdan, 1960) have been also found to yield mathematically and conceptually spurious results (e.g., Malvern & Richards, 1997). In addition, proposed solutions such as truncation (e.g., Prins & Bastiaanse, 2004) are not practical because it is not always feasible to obtain a predetermined number of tokens. Therefore, with respect to TTR scores, the legitimacy of the warrant "observed scores reflect the construct of interest" is seriously challenged to the extent that arguing in favor of the validity of any conclusions reached may be highly problematic.

Because of the widespread interest in conducting research with LD, more sophisticated methodologies have been developed to address the limitations of TTR. Some of these measures have more evidence to justify the validity of their score interpretations (e.g., D*;* Durán et al., 2004; Malvern & Richards, 1997, 2000; Richards & Malvern, 1997a, 1998), some have less (e.g., MTLD; McCarthy, 2005; McCarthy & Jarvis, 2010), and some have none (MATTR; Covington & McFall, 2010) other than face validity. Further, the validity of measures that were, until recently, considered the golden standard in terms of quantifying LD, such as *D,* have been questioned; whereas for others (e.g., Maas index) new supportive evidence has emerged (e.g., McCarthy & Jarvis, 2007).

The methodology that has been utilized to collect validity evidence regarding these tasks has relied primarily on the examination of correlational relationships among variables. Reaching conclusions about construct validity when using solely these techniques has been criticized heavily. For example, Borsboom (2005) refers to such sources of evidence as "circumstantial" (p. 151). Further, he argues, when validity is investigated this way, it ignores the most basic component of validity: the causal relationship between the latent trait and the observed indicators that supposedly underlies the patterns of observed behaviors. In a similar vein, Bollen (1989) argued, "a bivariate relationship is neither a necessary nor a sufficient condition for a causal relationship" (p. 57). Others have taken less of a critical standpoint in terms of correlational approaches to investigating validity (e.g., Angoff, 1988 as cited in Sireci, 2009).

On a different yet related note, conceptualizing and capturing LD in the field of communication disorders and related areas aims at drawing inferences about individuals. To differentiate between the relationship of an observed score and LD_S and the relationship of an observed score and LD_i, I refer to the latter as Level 2. As mentioned earlier, measuring LD_i though cannot occur in isolation from considering the language sample elicitation technique over and above the potential influence of the estimation technique. Depending on the research question, researchers and clinicians have used a variety of approaches to elicit different types of discourse such as descriptions of common procedures and pictorial stimuli (e.g., Brady et al., 2005; Christiansen, 1995) and re-tellings of familiar stories, and/or sharing past experiences (e.g., Ash et al., 2006; Coelho et al., 2005; Hough & Barrow, 2003; Ulatowska et al., 1981). The aforementioned techniques, when implemented carefully, elicit specific types of discourse that fall under the procedural genre (e.g., scripts) or the narrative genre (e.g., eventcasts, storytellings, and recounts), respectively.

However, there is a general agreement that various types of discourse are associated with different cognitive and linguistic demands (Bliss & McCabe, 2006; Brady et al., 2005; Nicholas & Brookshire, 1993; Ulatowska et al., 1990). As a result, performance on microlinguistic and/or macrolinguistic aspects of verbal production may vary systematically (e.g., syntactic complexity, Li et al., 1996; story elements, MacLachlan & Chapman, 1988). Similarly to these aspects of discourse, LD has also been found to vary as a function of discourse type/elicitation technique (e.g., Fergadiotis et al., 2011). Even though the exact

mechanisms that underlie this systematic variation are unclear (Bliss & McCabe, 2006; Capilouto et al., 2005; O'Loughlin, 1995; Wright & Capilouto, 2009), theories of vocabulary knowledge (Halliday & Hasan, 1989) and discourse processing (Frederiksen, 1996; Goffman, 1981; van Dijk & Kintsch, 1983) can account for this finding at a more global level. These theories suggest that the context within which a speaker produces discourse influences the mental processes that may give rise to linguistic phenomena such as LD.

Goals of the study. The experimental control associated with the design of a study is tailored to the needs of the question under investigation to allow for greater inferential power. Often then, the laboratory tasks and techniques that are selected are specific to particular experimental contexts. However, this may introduce a problem. For empirical findings to be informative and generalizable there should be a well-defined understanding of the structure that determines the relationship among the outcomes of different studies**.**

In recent years, several novel techniques have been developed to assess the breadth of one's vocabulary. Though all of the techniques assert to measure lexical diversity, each one is based on its own theoretical assumptions, which are reflected in the computational machinery they employ. Therefore, is not clear 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 lexical diversity related to each other.

In the field of speech-language pathology, researchers and clinicians often use several different methods to elicit various types of discourse. Lexical diversity

can be estimated based on any of these methods. However, it is an empirical question whether the inferences drawn from analyzing one type of discourse relate and generalize to other types of discourse. Here, I examine a corpus of four types of discourse that are commonly used for clinical and research purposes for evidence of a common trait that is captured across all sampling techniques. The goal is to provide a picture of how an individual's productive vocabulary manifests itself across different types of discourse that impose different cognitive and linguistic demands.

The specific aims of the this paper are organized into two levels that correspond to Levels 1 and 2 as they were discussed earlier.

Level 1: Supplement our understanding regarding the validity of the scores generated by different LD estimation techniques. Four techniques will be explored: *D*, Maas, MTLD, and MATTR. Specific questions to be addressed include:

i. Do all the techniques generate scores that are manifestations of the same latent variable (i.e., LD_S)?
ii. Is there a single latent variable determining performance for each estimation technique or are there specific method factors the jointly determine the scores?
iii. What are the magnitude and the nature of the relationships among the observed scores, the LD_S, and the method factors?
iv. Is there evidence that supports the use of a specific estimation technique?

Level 2: Explore the extent to which observed LD scores from different types of discourse are determined by a single construct (i.e., LD_i). Four tasks that are commonly used in clinical and research practice will be used: procedures, picture description, recounting personal experiences, and story re-telling. Specific questions to be addressed include:

i. Is LD_i a unitary construct that underlies the observed scores of a speaker when her/his language is sampled using different techniques?
ii. Are the inferences drawn regarding LD_i equally strong when different elicitation techniques are used? Are some tasks "better" than others?
iii. In terms of clinical and research practice, does measuring a specific type of discourse justify the conclusions reached regarding the individual's LD*i* skills?

Chapter 2

Method

Participants

Language samples from 442 participants were included in the analysis. All participants met the following inclusion criteria for participation in the study: (a) no history of stroke, head injury, or neurogenic disorder, per 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 performance on the Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 2002); and (f) no signs of depression as indicated by a passing score (0-4) on the Geriatric Depression Scale (Brink et al., 1982).

Discourse Elicitation

Stimuli and instructions. Participants' discourse samples used corresponded to four types of discourse: procedures, eventcasts, storytelling, and recounts. The first discourse task, procedures, was designed to elicit procedural discourse that is an activity-focused step-by-step description of how to achieve a goal (Longacre, 1996). The other three tasks were designed to elicit three different types of narrative discourse: eventcasts, storytelling, and recounts. Eventcasts are narratives that explain a scene of activities, stories are fictionalized, highly structured forms, and recounts are verbal reiterations of an event (Heath, 1986). Each experimental discourse task was introduced with a warm-up task. For the procedural discourse task, first, the examiner modeled the task by providing the

steps to make a pot of coffee. Then, participants were asked to provide the steps to: (a) make a peanut butter and jelly sandwich and (b) plant a flower in a garden. For the eventcasts, participants were presented with the Nicholas and Brookshire (1993) single 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*"). A practice task preceded participants' narrative descriptions of the experimental stimuli. The practice task included a brief narrative provided by the examiner describing the Picnic Scene from the Western Aphasia Battery-Revised (WAB-R; Kertesz, 2007) and then, the participant practiced by providing a story for the Cookie Theft picture from the Boston Diagnostic Aphasia Examination-3 (BDAE-3; Goodglass, Kaplan, & Barresi, 2001). During practice, feedback was provided to avoid eliciting a simple description of objects, characters and/or their physical characteristics.

Also, participants viewed and told the story depicted in the wordless picture book *Picnic* (McCully, 1984). This is a story about a family of mice who drive to the forest for a picnic. The baby mouse falls out of the truck on the way to the picnic site; however, the family does not notice and continues on without her. The family eventually realizes the baby mouse has been lost, and the story concludes when the family finds the baby mouse back on the road and decides to have their picnic then and there. Participants were presented with the stimulus book and were allowed as much time as they desired to view it and get familiar with the story. Then, they were asked to "Tell a story that goes with the pictures". Prior to the task, the examiner provided an example of how to tell a story using a

different wordless picture book (The Great Ape; Krahn, 1978). Finally, to elicit the recounts, each participant was asked to recall and share three past experiences: (a) what they did last weekend, (b) what they did during their last holiday and, (c) what they did during their last vacation. Similar to the previous tasks, the examiner initially modeled the task by providing a brief personal narrative about a trip to San Diego, California. For the picture descriptions and the recounts, if the participant stopped after 15 seconds or less, he/she was prompted with ‘‘Can you tell me more?’’

Transcription. Samples were digitally recorded and then orthographically transcribed in the CLAN format (MacWhinney, 2000) by trained research assistants. Samples were then segmented into c-units. A c-unit is a communication unit and includes an independent clause with its modifiers (Loban, 1976); it is commonly used to segment oral discourse samples (Hughes, McGillivray, & Schmidek, 1997). Approximately 10% of the samples were randomly selected and transcribed again for reliability purposes. Intra- and inter-rater word-by-word transcription reliability above 85% was chosen as a criterion for adequate reliability. Nonwords, hesitations, revisions, repetitions, and onomatopoeia were coded via transcription codes in CLAN.

Eventually, separate files were created for each type of discourse in simple Unicode text format that could be used with any of the applications that estimate each of the different LD indices (i.e., D, Maas, MTLD, and MATTR). Procedures included “How to make a peanut butter and jelly sandwich and” and “How to plant a flower in a garden”; eventcasts included the four picture descriptions;

stories included Picnic and Good Dog Carl; and, recounts included participant's narration of three previous experiences. Prior to estimating LD scores, number of types and tokens for each language sample was estimated. Patterns of missing data were also noted.

Estimating Lexical Diversity. *D* was estimated using the voc-*D* program in CLAN. Each participant contributed four input files which corresponded to the four types of discourse elicited. Similarly, MTLD was estimated for each file using a stand-alone application tool, the Gramulator 5.0, developed by McCarthy, Watanabe and Lamkin (in press) with permission from the developers. Maas was estimated using the Gramulator. The last index of LD, MATTR, was estimated using the computer software developed by Covington (2007).

Modeling Approach

LD by its very nature cannot be directly observed nor measured the same way height or weight can. However, LD can be conceptualized as an unobserved latent construct and captured using SEM. Within the SEM framework, latent variables called common factors can be defined using behaviors that represent them and can be directly observed. The observed behaviors are usually scores from tasks; these tasks are commonly referred to as indicators or manifest variables because they reflect the influence of the common factors. By formally defining the relationship of the observed indicators and the underlying factor using a series of equations, one can assume "measurement" of the respective cognitive process.

Confirmatory factor analysis (CFA) is a type of structural equation modeling (SEM) for specifying and exploring the relationship between manifest variables and latent variables to build measurement models. In the CFA model, common factors are the unobservable (i.e., latent) variables that underlie the manifest tasks and are assumed to determine how they vary. The variance of a manifest variable not accounted for by the common factors is referred to as the residual term or uniqueness. There are two sources of variance that combined comprise residual terms: *unique or specific factors*, that represent systematic variance associated with a specific indicator, and *random error* (e) often conceptualized as measurement error. The path diagram in Figure 6 shows the two sources, the factor and the residual term that jointly account for the variance in the indicator.

In most applications of CFA there is more than a single indicator. Let $\mathbf{X_p}$ denote the vector of *p* indicators, $\boldsymbol{\xi_r}$ a vector of factor scores, $\boldsymbol{\Lambda_{pr}}$ the matrix of loadings relating $\mathbf{X_p}$ to $\boldsymbol{\xi_r}$, and ε_p the vector of residual terms. Because specific factors and measurement error are often not distinguished, they are denoted by $\varepsilon_p = e_i + s_{i;}$ (Bollen, 1989; Meredith & Teresi, 2006) giving the following equation:

$$\mathbf{X_p} = \boldsymbol{\Lambda_{pr}}\boldsymbol{\xi_r} + \boldsymbol{\varepsilon_p} \qquad (3)$$

If the model is specified correctly and $\boldsymbol{E(\varepsilon_p) = 0}$ and COV($\boldsymbol{\xi_r}, \boldsymbol{\varepsilon_p}$), then Equation (3) leads to the following equation for factor analysis that models indicator covariances as a function of common and unique factors (Jöreskog, 1969):

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}^{3} \qquad (4)$$

where: $\boldsymbol{\Sigma}$ is a (p x p) population covariance matrix of the manifest variables,

$\boldsymbol{\Lambda}$ is a (p x r) matrix of the predictors' loadings on $\boldsymbol{\xi_r}$,

$\boldsymbol{\Phi}$ is a (r x r) matrix of covariances among the latent variables, and

$\boldsymbol{\Theta}$ is a (p x p) a diagonal matrix of covariances among the residual terms in the model.

If unidimensionality is assumed, $\mathbf{r} = 1$, $\boldsymbol{\xi_1}$ and $\boldsymbol{\Phi}$ have dimensions 1 x 1**,** whereas $\boldsymbol{\Lambda}$ becomes a p x 1 vector.

Conducting analyses to examine validity within the factor analytic framework has several advantages over traditional approaches. When scores from two tasks are correlated to assess validity, there is an implicit assumption that the reason scores correlate is a common underlying trait that influences both tasks in a similar way. For instance, if individual i has high LD_i, she is expected to have high scores on tasks p_1 and p_2 if the tasks are true indicators of LD_i. This scenario is akin to a case of a spurious correlation in which variables correlate because they are both caused by a third variable, LD_i. Using only observed variables to explore validity, one could correlate the scores from p_1 and p_2 and a high correlation between the two tasks could be considered as evidence that there is

[3] Equation (2) can be supplemented by $\boldsymbol{\tau_p}$ that is the vector of measurement intercepts, The inclusion of the intercept is an extension of the basic model, which assumes for most applications, that the intercepts are zero.

indeed a latent variable that could explain the correlation. However, that would be rather weak evidence because the common factor is not included in the model (Bollen, 1989, p. 195).

Conversely, CFA addresses this issue by explicitly modeling the common factor in the model. For example, for the two tasks, a series of equations to denote the relationship among the tasks and the latent factor would be:

$$x_{p1} = \lambda_{p1r1}\xi_{r1} + \varepsilon_{p1} \qquad (5)$$

$$x_{p2} = \lambda_{p2r1}\xi_{r1} + \varepsilon_{p2} \qquad (6)$$

$$E(\varepsilon_p) = E(e_{p2}) = 0 \qquad (7)$$

$$\mathrm{COV}(\xi_{r,}\, \varepsilon_p) \qquad (8)$$

where x_{p1} and x_{p2} are the scores in tasks ***p₁*** and ***p₂,*** respectively,

ξ_{r1} is the common factor that is the same for each individual,

λ_{p1r1} and λ_{p2r1} are the parameters that link the observed scores to the latent factor,

ε_{p1} and ε_{p2} are the residual terms of the tasks.

First, this model explicitly states the presence of a latent common factor, by including the term $\boldsymbol{\xi_{r1}}$ that may be interpreted as the mathematical instantiation of the latent trait. The loadings, λ_{p1r1} and λ_{p2r1}, can be interpreted as regression coefficients expressing the expected change in *x* as for a one unit increase of $\boldsymbol{\xi_{r1}}$. Further, often, the model stipulates that once the effect of the latent variable is

taken into account, there is no more systematic covariance among the residual terms of the observed indicators. That is, the model may be specified to denote that the sole determinant of the observed scores is the common factor.

The path diagram in Figure 7 shows how discourse-specific LD_S relates to the observed scores that are estimated using four estimation techniques (D, Maas, MTLD, and MATTR). Typically, latent variables and observed variables are represented with squares and circles, respectively. Unidirectional lines represent hypothesized direct effects. The arrowhead pointing to the observed variables denotes that the unobserved variable "determines" how scores vary when estimated using different techniques. The residual terms are symbolized with small circles under each indicator and express the amount of variance in the indicators not explained by the substantive factor. Further, there are no two-headed curved arrows linking these residual terms, which would allow them to covary, because this model stipulates there is a single factor that accounts for the intercorrelations among the observed indicators. That is, once the variance of the common factor is accounted for, tasks do not covary any more.

The factor model described is consistent, conceptually, with the idea that latent variables cause the observed scores. Discourse-specific LD_S was defined earlier as "a characteristic of the language sample" and it was argued that the scores obtained when using different estimation techniques are merely reflections of the language sample's LD. Scores often contain a blend of construct-related and potentially construct-irrelevant influences and random error. It was further argued that observed scores are not the same as the sample's LD_S and a reasoning

step is required to make inferences about the sample's LD_S based on the observed scores. The CFA framework offers a language to express the relationship between the latent variable, which, in theory, corresponds to the trait of interest and the observed scores. Also, CFA provides the mathematical machinery to evaluate that expression using fit indices that gauge how well the model accounts for the data. Further, by examining the parameter estimates of the estimated model it is possible to gain an understanding of the magnitude and nature of the relationships among the model elements. Because of these features, the warrant in Toulmin's (1958) argument structure can be recast as a CFA model. In Figure 8, the warrant, "scores reflect the construct of interest", may be supported or refuted using evidence from the CFA model. To find evidence that the model holds and the parameters suggest a strong relationship between the indicators and the factor would provide evidence in support of the plausibility of the interpretive argument.

Following the diagram, the argument could be expressed as follows: for a given language sample, a low score was estimated using technique X; so, an inference can be made that the language sample probably has low LD_S. This inference is based on the warrant that the scores from technique X reflect the construct of interest, in this case discourse-specific LD_S. Upon request to back up that warrant, one could provide evidence based on the CFA model, including the fit of the model and the substantive interpretation of the parameter estimates. Specifically, a supportive model would fit well and the loadings of the manifest variables would be high; that could be interpreted as evidence that the tasks measure a single construct and are strong indicators of the latent variable. This

argument would be valid unless an alternative explanation was offered. In such a case rebuttal evidence would have to be put forward.

Using the single factor (i.e., unidimensional) CFA model to gather evidence regarding the plausibility of the warrant is considered standard practice in psychometric evaluations. Yet, in most applications, the tasks that are used to operationalize the common factor are *not* similar in any way other than the fact that they are designed to measure the same construct. Based on this assumption, the model in path diagram 7 stipulates that there are no residual covariances among the residual terms of the indicators once the substantive factor has been taken into account. However, that might not be the case. Especially in light of the literature review findings according to which some indices reflect more than just the effects of LD (i.e., LD-irrelevant variance such as length effects).

In the next section, a broader framework is presented within which the questions of this study were explored. Specifically, First, I use as a starting point the multitrait-multimethod modeling approach to model potential method effects associated with the estimation techniques and explore how well the reflect the LD_S; and, I discuss the advantages and disadvantages of two popular multitrait-multimethod parameterizations and why one of them was preferred in this paper. Then, I introduce some general aspects of the hierarchical factor analytic framework that was used to answer how LD_i relates to different discourse types.

Multitrait-multimethod approaches. Multimethod measurement refers to the use of more than one method to capture some or all of the constructs or traits of interest. With respect to measurement models, such an approach can have

a key role in the validation process (Eid & Diener, 2006; Campbell & Fiske, 1959). As discussed earlier, validity represents the degree to which theory and empirical evidence justify the outcomes of inferential reasoning and the course of actions that are based on a measurement model (Messick, 1989). Geiser (2008) pointed out that by utilizing a multimethod approach to capture one or more latent traits, several sources (i.e. diverse methodologies) may provide information to formulate the factor. In such cases, the combined information allows researchers to draw stronger conclusions that may be more easily generalized.

Consider for example LD. Assume that a speaker's LD_i was measured using language samples elicited under different conditions and multiple estimation techniques were utilized. Any examiner would feel significantly more confident to argue that a speaker has great LD*i* if the examiner had observed high scores across the board. However, if different conditions were yielding conflicting information, one would normally have less faith in the inferences made based on the results.

A problem arises though, if one considers that scores that are derived using a specific methodology might not be pure estimates of the construct of interest. Scores may be a function of the testing method used. Campbell and Fiske (1959), in an attempt to address this issue, developed the analysis of the multitrait-multimethod (MTMM) matrix. Complementing Campbell and Fiske's approach with appropriate modern methodological approaches allows CFA models (potentially) for a decomposition and examination of variance that is due to traits, variance that is due to methods, and unique or error variance.

Within the context of this study, which focuses on capturing LD at different levels, four indices of LD were used to capture LD_S in four different types of discourse. Therefore, common estimation techniques are used to estimate LD_S in each type of discourse. So, it is possible that some covariation among observed variables might be due to the method of measurement rather than the substantive content of the measure. It is noteworthy, that in this paper, the traits refer to LD_S associated with a specific type of discourse rather than different traits. This is a departure from how MTMM approaches have been used traditionally in the past. Nevertheless, it will be shown in the next section how the machinery of MTMM approaches can be used to accommodate the questions of this paper.

Several kinds of models can be applied to MTMM matrices to reach to conclusions about the potential underlying factorial structure of the data. Two forms of specification though have been prominent and will be presented in the following section along with some of their advantages and disadvantages: the correlated traits-correlated methods (CT-CM) and the correlated traits-correlated uniquenesses (CT-CU).

Correlated traits – correlated methods. Traditionally, within the CFA framework, the MTMM matrices have been analyzed using the correlated traits-correlated methods parameterization (CM). If *i* latent variables are assumed to correspond to types of discourse being measured using *j* estimation methods through *i* x *j* observed indicators (henceforth referred to as type-method units), then following Widaman (1985) and Lance, Noble, and Scullen (2002), equation

(2) can be re-written as:

$$\boldsymbol{TMU}_{ij} = \lambda_{T_{ij}} \boldsymbol{T}_i + \lambda_{M_{ij}} \boldsymbol{M}_j + \varepsilon_{ij} \quad (9)$$

where: $\boldsymbol{TMU}_{ij}$ is the *ij*th type-method unit (i.e., the *ij*th indicator that is designed to measure the *i*th Type using the *j*th Method),

$\lambda_{T_{ij}}$ is the T_i.*ij*th factor loading linking the *ij*th TMU to its respective *i*th Type,

$\lambda_{M_{ij}}$ is the M_j.*ij*th factor loading linking the *ij*th TMU to its respective *i*th method,

$\boldsymbol{T}_i$ is the *i*th Type,

M_j is the *j*th Method, and

ε_{ij} represents systematic variance for the *ij*th indicator and random error.

If the assumption that $\boldsymbol{E(\xi_g, \varepsilon_g) = 0}$ is further specified as $\boldsymbol{E(T_i, \varepsilon_{ij}) = E(M_j, \varepsilon_{ij}) = 0}$, and based on equation (9), equation (4) can be partitioned as:

$$\boldsymbol{\Sigma} = [\boldsymbol{\Lambda_T} | \boldsymbol{\Lambda_M}|] \begin{bmatrix} \boldsymbol{\Phi_{TT}} & \boldsymbol{\Phi_{TM}} \\ \boldsymbol{\Phi_{MT}} & \boldsymbol{\Phi_{MM}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{T}}' \\ \boldsymbol{\Lambda}_{\boldsymbol{M}}' \end{bmatrix} + \boldsymbol{\Theta} \quad (10)$$

Further, if $\boldsymbol{E(T_i, M_j) = 0}$ for identification purposes (Widaman, 1985), then equation (9) can be re-written as:

$$\boldsymbol{\Sigma} = [\boldsymbol{\Lambda}_{\mathbf{T}}|\boldsymbol{\Lambda}_{\mathbf{M}}|] \begin{bmatrix} |\boldsymbol{\Phi}_{\mathbf{TT}}| & \mathbf{0} \\ \mathbf{0} & |\boldsymbol{\Phi}_{\mathbf{MM}}| \end{bmatrix} \begin{bmatrix} \acute{\boldsymbol{\Lambda}}_{\mathbf{T}} \\ \acute{\boldsymbol{\Lambda}}_{\mathbf{M}} \end{bmatrix} + \boldsymbol{\Theta} \quad (10a)$$

where $\boldsymbol{\Sigma}$ is a ((i x j) x (i x j)) covariance matrix of the manifest variables,

$\boldsymbol{\Lambda}_{\mathbf{T}}$ is a ((i x j) x Ti) submatrix that contains the loadings that link the TMU_{ij}'s to their corresponding T_i Types,

$\boldsymbol{\Lambda}_{\mathbf{M}}$ is a ((i x j) x Mj) submatrix that contains the loadings that link the TMU_{ij}'s to their corresponding M_j Methods,

$\boldsymbol{\Phi}_{\mathbf{TT}}$ is a (i x i) symmetric submatrix of covariances among the factors that correspond to Types,

$\boldsymbol{\Phi}_{\mathbf{MM}}$ is a (j x j) symmetric submatrix of covariances among the factors that correspond to Methods, and

$\boldsymbol{\Theta}$ is a ((i x j) x (i x j)) a diagonal matrix of covariances among the residual terms in the model.

If i = 4 and j = 4, the model is consistent with four types of discourse and four methods influencing the observed scores. Also, residual terms are not allowed to covary among them or with any other latent variable in the model. This model corresponds to Widaman's (1985) 3C case within his taxonomy.

Correlated traits – correlated uniquenesses. An alternative CFA parameterization of the multitrait-multimethod approach that has been proposed in response to some commonly observed problems with convergence and admissibility of the CT-CM model, is the correlated traits-correlated uniquenesses

model (CT-CU; Kenny, 1979; Kenny & Kashy, 1992; Marsh, 1989). However, when this parameterization is used, disentangling the sources that contribute to the observed score variation becomes significantly more subjective. With the CT-CM model discussed in the previous section, method effects are defined separately from the residual terms and are captured by obtaining information on how the estimation method performs across multiple types of discourse. This allows researchers to put forward substantive interpretations of the nature of the method factors. With the CT-CU approach, method effects are part of the residual terms; but because the residual terms of the TMU's are an amalgam of random error and method effects, the covariances among them may not be interpreted in a straightforward manner; and nor can they assist substantially with the interpretive nature of the method effects.

Another major limitation of the CT-CU is that it does not allow for testing alternative method factor structures. For example, as opposed to the CT-CM, the CT-CU necessarily assumes orthogonal methods, and therefore one cannot assess whether an orthogonal or oblique method factor structure underlies the data. So, as opposed to the correlated methods approach that allows for a wide range of testable model variants, the CT-CU models appear to be more restricted.

Further, this misspecification may propagate through the model and bias the substantive variance components upwards, in some cases to a significant degree (e.g., Byrne & Goffin, 1993; Kenny & Kashy, 1992). Specifically, if the assumption of orthogonal method factors does not hold, substantive variances and covariances might be overestimated giving a false impression of the relationships

among the substantive factors. Because of these disadvantages, the CT-CM parameterization was preferred to address the goals of the paper.

Modeling Level 1

One of the great advantages of using the CT-CM framework to explore the first goal of the paper was its ability to decompose variance and therefore increase the substantive interpretability and clarity of the model parameters. Widaman (1985) described a series of hierarchically nested models that could be compared to evaluate several hypotheses of substantive interest.

To answer the main questions at Level 1, a series of nested models was specified on an a priori basis. Each model was evaluated in terms of its overall goodness of fit. Further, the nested model relationship of the proposed models was exploited to test statistically if the restricted model in each comparison fit significantly worse than the comparison model.

The rational for specifying the series of models in this paper was based on parsimony. The first model that was fit to the data was a highly restricted unidimensional model which assumed that a single construct determined performance across all variables regardless of type of discourse or estimation technique. The second model relaxed this restriction and assumed that TMU's that were estimated based on the same type of discourse were determined by a discourse-specific LD_S construct. Once the effects of this source were modeled, TMU's were assumed to be uncorrelated. Further, the model in this step assumed that discourse-specific LD_S's were unrelated.

In the next step, a model was specified to test the relationship between the discourse-specific LD_S's. In this step, a model was specified that allowed discourse-specific LD_S's to correlate. The comparison of the misfit associated with the restricted model of the previous step would provide evidence in terms of the relationship between the latent constructs. Specifically, if the model that allowed the discourse-specific LD_S's to correlate fit the data better than the more restricted model, that would suggest that LD_S's derived from specific types of discourse are related. The belief that some unmeasured common cause created the unexplained relationship between them was tested later in the analysis.

Subsequently, the next step included the critical test for the presence of method factors which was based on the comparison of two models: one that stipulated no method factors (i.e., the model fro the previous step and one that stipulated that observed scores of the TMU's were systematically influenced by method factors. The former model was consistent with the hypothesis that the four LD indices measure only discourse-specific LD_S. According to this model, the observed scores of the indicators were functions, solely, of discourse-specific LD_S and the residual terms. In other words, method factors associated with specific estimation techniques were restricted to have zero systematic effects on the scores. The latter model stipulated that the TMU's reflect the additive effects of (i) the underlying discourse-specific LD_S, which they are intended to capture, via the loadings in the upper part of the diagram, (ii) the methods, which are used to operationalize the LD within each type, via the loadings in the lower part of the diagram, and (iii) specific and unreliable variance.

If the restricted model (i.e., no method factors) did not perform significantly worse than the full model (i.e., with method factors), then that would support that the LD estimation techniques produced scores that were free of systematic method effects that would question the validity of interpretation. Conversely, if eliminating the method factors caused significant misfit, the presence of method factors would be suggested. Finally, the last step of addressing the questions of Specific Aim 1 included an exploration of the relationship between the method factors. Specifically, it included a comparison of the best fitting model from the previous step to a model that allowed for the method factors to be correlated.

Hierarchical Factor Analysis

Modeling Level 2

The main questions that this study investigated at Level 2 were (i) whether LD_i is a unitary construct that underlies the observed scores of a speaker when her/his language is sampled using different techniques; (ii) whether the inferences ones draws regarding LD_i are equally strong when using different elicitation techniques; and, (iii) to what extent do the conclusions drawn from studying one type of discourse justify drawing conclusions about performance in other types of discourse.

Given the structure of the data, the modeling approach to explore these questions would have to be able to model LD_i over and above the potential presence of method factors associated with specific techniques; and, over and

above the dependency of the observed indicators that were based on the same type of discourse. One alternative under these circumstances was to employ a bifactor model (also referred to as "hierarchical model"; Yung, Thissen, & McLeod, 1999).

The bifactor factor model was initially developed in the context of research on cognitive abilities and it was an extension of Spearman's conceptualization of intelligence (1904) by Holzinger and Swineford (1937). Bifactor models are potentially applicable in situations when there is a general factor that is hypothesized to account for the covariance among the observed variables; and there are multiple specific factors, each of which is hypothesized to account for the unique influence of the specific factor on a group of observed variables over and above the general factor.

Based on Chen, West, and Sousa (2006), the bifactor model can be specified through the CFA model. The equation linking the observed variables and the factors was shown earlier:

$$\mathbf{X_p} = \mathbf{\Lambda_{pr}}\xi_r + \varepsilon_p \qquad (3)$$

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda'} + \mathbf{\Theta} \qquad (4)$$

where $\mathbf{X_p}$ denotes the vector of *p* indicators,

$\mathbf{\Lambda_{pr}}$ the matrix of loadings relating $\mathbf{X_p}$ to ξ_r,

ξ_r a vector of *r* factor scores,

ε_p the vector of residuals,

$\boldsymbol{\Sigma}$ is a (p x p) population covariance matrix of the manifest variables,

$\boldsymbol{\Lambda}$ is a (p x r) matrix of the predictors' loadings on ξ_r,

$\boldsymbol{\Phi}$ is a (r x r) matrix of covariances among the latent variables, and

$\boldsymbol{\Theta}$ is a (p x p) a diagonal matrix of covariances among the residual terms in the model.

Below, Equation 4 for the bifactor model is expanded. Numbers 1-4 correspond to types of discourse (procedures, eventcasts, storytelling, recounts); letters 1-4 to estimation techniques (D, Maas, MTLD, MATTR); and T stands for the general factor. The first vector includes the observed indicators (i.e., the combinations of types of discourse and estimation techniques). Then, follows the $\boldsymbol{\Lambda}$ matrix of the predictor's loadings on the factors. This matrix is of particular interest because it conveys the hierarchical structure of the model. Every TMU_{ij} is determined by the general factor and then depending on the type of discourse and the estimation technique, it may potentially be determined by two additional factors. First, they may be influenced by factors are that are specific to the type of discourse; and second, they may be influenced by factors that are specific to the estimation technique used to derive the LD score. Further, the next vector includes the factors scores. It is noteworthy that each speaker is characterized by a single factor score for the general factor which reflects the assumption that there is something common driving the LD scores across all TMU's.

$$\begin{bmatrix} TMU_{1a} \\ TMU_{1b} \\ TMU_{1c} \\ TMU_{1d} \\ TMU_{2a} \\ TMU_{2b} \\ TMU_{2c} \\ TMU_{2d} \\ TMU_{3a} \\ TMU_{3b} \\ TMU_{3c} \\ TMU_{4a} \\ TMU_{4b} \\ TMU_{4c} \\ TMU_{4d} \end{bmatrix} = \begin{bmatrix} \lambda_{1a.T} & \lambda_{1a.1} & 0 & 0 & 0 & \lambda_{1a.a} & 0 & 0 & 0 \\ \lambda_{1b.T} & \lambda_{1b.1} & 0 & 0 & 0 & 0 & \lambda_{1b.b} & 0 & 0 \\ \lambda_{1c.T} & \lambda_{1c.1} & 0 & 0 & 0 & 0 & 0 & \lambda_{1c.c} & 0 \\ \lambda_{1d.T} & \lambda_{1d.1} & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{1d.d} \\ \lambda_{2a.T} & 0 & \lambda_{2a.2} & 0 & 0 & \lambda_{2a.a} & 0 & 0 & 0 \\ \lambda_{2b.T} & 0 & \lambda_{2b.2} & 0 & 0 & 0 & \lambda_{2b.b} & 0 & 0 \\ \lambda_{2c.T} & 0 & \lambda_{2c.2} & 0 & 0 & 0 & 0 & \lambda_{2c.c} & 0 \\ \lambda_{2d.T} & 0 & \lambda_{2d.2} & 0 & 0 & 0 & 0 & 0 & \lambda_{2d.d} \\ \lambda_{3a.T} & 0 & 0 & \lambda_{3a.3} & 0 & \lambda_{3a.a} & 0 & 0 & 0 \\ \lambda_{3b.T} & 0 & 0 & \lambda_{3b.3} & 0 & 0 & \lambda_{3b.b} & 0 & 0 \\ \lambda_{3c.T} & 0 & 0 & \lambda_{3c.3} & 0 & 0 & 0 & \lambda_{3c.c} & 0 \\ \lambda_{3d.T} & 0 & 0 & \lambda_{3d.3} & 0 & 0 & 0 & 0 & \lambda_{3d.d} \\ \lambda_{4a.T} & 0 & 0 & 0 & \lambda_{4a.4} & \lambda_{4a.a} & 0 & 0 & 0 \\ \lambda_{4b.T} & 0 & 0 & 0 & \lambda_{4b.4} & 0 & \lambda_{4b.b} & 0 & 0 \\ \lambda_{4c.T} & 0 & 0 & 0 & \lambda_{4c.4} & 0 & 0 & \lambda_{4c.c} & 0 \\ \lambda_{4d.T} & 0 & 0 & 0 & \lambda_{4d.4} & 0 & 0 & 0 & \lambda_{4d.d} \end{bmatrix} \begin{bmatrix} \xi_T \\ \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_a \\ \xi_b \\ \xi_c \\ \xi_d \end{bmatrix} + \begin{bmatrix} \varepsilon_{1\alpha} \\ \varepsilon_{1\beta} \\ \varepsilon_{1c} \\ \varepsilon_{1d} \\ \varepsilon_{2a} \\ \varepsilon_{2b} \\ \varepsilon_{2c} \\ \varepsilon_{2d} \\ \varepsilon_{3a} \\ \varepsilon_{3b} \\ \varepsilon_{3c} \\ \varepsilon_{3d} \\ \varepsilon_{4a} \\ \varepsilon_{4b} \\ \varepsilon_{4c} \\ \varepsilon_{4d} \end{bmatrix}$$

For example TMU_{1a} (which corresponds to LD scores estimated using D based on procedures) is determined by (i) the general factor ξ_T via the loading $\lambda_{1\alpha.T}$, (ii) the discourse specific factor ξ_1 via the loading $\lambda_{1\alpha.1}$, (iii) the method specific factor ξ_a via the loading $\lambda_{1\alpha.a}$, and (iv) a residual variance ε_{1a}.

Even though the complexity of these models is greater because they have two layers of factors, there are several similarities with the CFA models discussed so far. When CFA models are specified, the indicators are the observed scores and the factors are linked to the indicators via their corresponding loadings. In hierarchical CFA, the general factor has direct effects on the indicators. In other words, the general factor is defined by the covariance among all the observed indicators and represents a common trait that underlies performance in all the TMU's. Given that the general factor is uncorrelated with the rest of the factors, the specific factors account for unique variance that is not explained by the general factor. Finally, as in every CFA model there are residual terms that

represent random noise and the unique factors associated with each of the indicators.

Using the hierarchical CFA framework, it is possible to estimate parameters to draw information regarding the relationship between the general factor and the observed variables. The motivation for using the hierarchical CFA was to include in the model a higher order common factor corresponding to discourse-independent LD. Earlier, I discussed how the multi trait – multi method approach could be used to explore the implicit assumption that the LD_S estimation techniques were free of any specific method effects that could introduce construct (i.e., LD_S) irrelevant variance. The Level 2 questions could be explored by testing a hierarchical model versus a model that allows for correlations among the discourse specific factors. The hierarchical model would stipulate that a general factor is responsible for the intercorrelations among the discourse specific factors. The acceptance of this "hierarchical" model would support the hypothesis that LD is a relatively unitary construct that underlies the observed scores of a speaker when his or her language is sampled using different techniques.

Testing whether a single second-order factor can adequately summarize the variance shared among the discourse-specific LD factors can be achieved by comparing the model fit of the two models (Rindskoppf & Rose, 1988). Comparing the two models can be performed using their χ^2 difference to test whether imposing restrictions in the structural part of the model causes a statistically significant change in model fit. In these comparisons, additional fit indices can be consulted to evaluate whether the model misfit is substantial.

Given a hierarchical model that converges, produces admissible parameter estimates, and fits reasonably well, the relationships among the factors and the observed indicators can be explored and can be substantively interpreted.

Chapter 3

Results

Preliminary Analysis

Language samples from 442 participants were included in the study. A summary of their demographic information can be seen in Table 2. Descriptive statistics for the estimated lexical diversity (LD) indices for each type of discourse before the removal of outliers can be seen in Table 3. Also, descriptive statistics for the number of types, tokens and type-token ratios can be found in Table 4.

After importing data in SPSS, they were screened for missing values. The percentage of missing data of the major study variables ranged from .23% to 2.71%. Across variables, the average percentage of missing values was 1.02%. Reasons for missing data in the majority of the cases included recording equipment failures, video format conversion failures, and testers' oversights during task administration and data collection. More information regarding missing data for each variable can be found in Table 5.

Data were screened for univariate outliers; defined as scores that were more than 3.3 *SD*'s beyond the mean. Across all variables, .39% of data points were identified as univariate outliers by inspecting frequency distributions of *z* transformed scores. The language sample of each outlier was inspected to explore the reason for why scores deviated significantly from the mean. In most cases, outliers were generated because participants did not follow task instructions correctly. For example, some participants produced a mixture of expository, narrative and procedural discourse when asked to produce procedures. Other

participants used dialogue to convey all of the interactions of the characters when asked to produce eventcasts. These values were removed and treated as missing data. The number and percentage of missing data for each variable after the removal of outliers can be found in Table 6.

After the removal of univariate outliers, data were screened for multivariate outliers using the Mahalanobis distance statistic, which is distributed as a χ^2 statistic with degrees of freedom equal to the number of variables; multivariate outliers were defined as the cases that were associated with p values less than .001, which would have suggested that the null hypothesis that the specific case comes from the same population as the rest was rejected. Using this approach, no multivariate outliers were identified.

Distributions were visually inspected and assessed in terms of the normality assumption; and, skewness and kurtosis statistics were estimated. Several distributions were noted to be positively skewed and with various degrees of kurtosis but the majority of the distributions were consistent with the normality assumption. Descriptive statistics for the estimated LD variables for each type of discourse after the removal of outliers can be seen in Table 6. Skewness and kurtosis are reported in Table 6.

After outliers were removed, the variances of the major study variables were estimated. When the ratio of the largest to the smallest variance was greater than 10.0, the corresponding observed variables were re-scaled to avoid having to work with an ill-scaled covariance matrix that could lead to convergence problems. Finally, data were exported to Mplus compatible format for data

analysis. The variance-covariance matrix of the transformed LD variables that was modeled can be seen in Table 7.

Main Analysis

The models were estimated in Mplus 6.1 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. A series of hypotheses using nested model comparisons were tested followed by a substantive evaluation of model parameters. To perform the nested model comparisons the scaled difference χ^2 test statistic (Satorra & Bentler, 2001) was used. Although very little (average covariance coverage = 98%, Table 8), missing data were accommodated in all analyses using direct maximum likelihood under the assumption of missingness at random (cf. Enders, 2010).

To assess the fit of the models, several fit indices were taken into account to examine various aspects of model fit (i.e., absolute fit, parsimonious fit, fit relative to the null). Fit indices included the Satorra-Bentler scaled χ^2 statistic (Satorra & Bentler, 1994) to take into account the non-normality of the data. Two additional fit indices that were taken into account included the comparative fit index (CFI; Bentler, 1990), and the root-mean square error of approximation (RMSEA; Cudeck Browne, 1993). Finally, the standard root mean residual (SRMR; Hu & Bentler, 1998) was also included. A good fitting model was expected to have a non-significant χ^2; however, χ^2 is sensitive to sample size and with large enough sizes it is possible to have a significant χ^2 even when the

discrepancies between the observed and reproduced matrices are small (Kline, 2005). Nevertheless, the χ^2 statistic was reported for completeness of results and because χ^2 difference tests were used to evaluate differences between nested models. The CFI and RMSEA are less sensitive to sampling characteristics and take degrees of freedom into account, and therefore were chosen as additional fit indices. Based on published guidelines, good fit was indicated by a CFI value close or higher than .95, a RMSEA value below .08, with the upper bound of the 95% confidence interval below .10, and an SRMR value close to .08 (Brown, 2006; Hu & Bentler, 1999; Kline, 2005). The SRMR reflects the discrepancy between the observed and predicted covariance matrices. In this index, lower values are indicative of a better fit.

Level 1. The first set of analyses addressed the questions regarding the validity of the scores generated by different LD estimation techniques. Four techniques were assessed (*D*; Maas; Measure of Textual Lexical Diversity, MTLD; and Moving Average Type Token Ratio, MATTR) using four types of discourse (procedures, eventcasts, stories, and recounts).

Step 1. The first model (Model 1) that was estimated assumed that the relationships among all the observed variables (i.e., Type Method Units, TMU's) could be accounted for by a single latent construct that could be interpreted as general LD. To test how consistent this conceptualization of LD was with the data, a unidimensional model was specified, according to which one latent factor determined performance across all the observed indicators. To model the assumption of local independence after accounting for the common factor, the

residual variances of the manifest variables were not allowed to covary. Further, to model the lack of any other kind of systematic effect to the data, no other latent factors were specified.

The one-factor model demonstrated a very poor fit to the data $\chi^2(104, N = 442) = 3888.62$, $p < .001$, CFI = .40, RMSEA = .29 (90% confidence bands = .28 and .30), and SRMR = .21. The path diagram for this model can be seen in Figure 9 (the full solution including the standardized and unstandardized parameters estimates can be seen in Table 9).

Step 2. The second model (Model 2) that was tested, was consistent with two hypotheses. First, LD scores derived from the same language sample were determined by a common latent variable. And second, these latent variables (i.e., the LD_S for each type of discourse) were unrelated. To reflect the first hypothesis, Model 2 assumed that the LD observed variables that were derived from the same type of discourse loaded on the same factor. Therefore, four latent factors were specified that corresponded to procedures, eventcasts, storytelling, and recounts (LD_{S1}-LD_{S4}, respectively). Moreover, once the covariance due to the common factors was taken into account, the residual terms of the observed variables were not allowed to covary. To reflect the second hypothesis, the correlations among the latent factors were constrained to be zero.

Model 2 resulted in a large increase in model fit compared to Model 1 across all types of global fit indices ($\chi^2(104, N = 442) = 1112.63$, $p < .001$, CFI = .84, RMSEA = .15 (90% confidence bands = .14 and .16), and SRMR = .26 However, despite the considerable improvement in overall fit, Model 2 did not

demonstrate acceptable fit to the data. The standardized path coefficients of Model 2 are presented in Figure 10. The complete unstandardized and standardized solution of Model 2 can be seen in Tables 10 and 11. In terms of the parameter estimates of the factor loadings, the vast majority of them was of high magnitude (median = .91; range of absolute values = .74 - .98). Due to the poor fit to the data, parameter estimates were not further interpreted.

In the next step, the model was aligned more closely to the a priori expressed hypothesis that discourse-specific LD_S's were related. An inspection of the modification indices in the Mplus output, also suggested that this would be a reasonable next step. Specifically, the highest estimate of how much χ^2 would decrease if a single parameter was freed was associated with the correlation between the two factors that were defined by the eventcasts TMU's and the storytelling TMU's, respectively (approximate $\Delta\chi^2 = 161.87$; expected parameter change = .64).

Step 3. In step 3, a model was tested that was consistent with two hypotheses. First, similar to Model 2, Model 3 assumed that the four estimation techniques that were used to derive the LD scores, measured the same construct. The second hypothesis reflected in Model 3 was that LD_S's that were based on different types of discourse were related.

The specification of Model 3 was similar to Model 2 with one exception. Model 3 included four latent factors, each corresponding to four indices derived from the same type of discourse. The residual terms of the observed variables were not allowed to covary. Therefore, similar to Model 2, Model 3 stipulated that

the four estimation techniques, for each type of discourse, were measuring the same construct and very little of anything else that could cast additional covariation between observed indicators. So, once the covariance due to the common factor was taken into account, the residual terms of the observed variables were constrained to be uncorrelated. The difference between Models 2 and 3 was that, in Model 3, the latent factors were allowed to covary. This latter specification mirrored the assumption that estimated LD_S based on one type of discourse would be related to estimated LD_S based on other types of discourse.

Model 3 resulted in further improvement of model fit compared to Model 2 especially in terms of the standardized root mean square residuals, $\chi^2(98, N = 442) = 757.53$, $p < .001$, CFI = .89, RMSEA = .13 (90% confidence bands = .12 and .14), and SRMR = .08. Using the scaled difference χ^2 test statistic, Model 2 was found to fit the data significantly worse than Model 3, $\Delta\chi^2(6) = 370.29$, $p < .001$. However, despite the improvement in overall fit, Model 3, still, failed to demonstrate acceptable fit to the data. Only SRMR achieved the desired level of fit; specifically, its value was equal to the cut-off score (= .08). The standardized path coefficients of Model 3 are presented in Figure 11. The unstandardized factor loadings and the residual variances can be seen in Table 12; the standardized loadings and the residual variances are presented in Table 13. The intercorrelations of the factors can be seen in Table 14. Due to the poor fit to the data, parameter estimates were not further interpreted because misspecified models often produce biased parameter estimates.

Step 4. The a priori hypothesis that generated the next model was that observed indicators produced by the same estimation technique were related for reasons that were independent of discourse-specific LD_S. In terms of how it was specified, Model 4 included four correlated factors that were defined by the observed variables within each type of discourse (LD_{S1}-LD_{S4}). In addition, four additional factors were specified that were defined by variables that had been estimated using the same estimation technique (henceforth referred to as Method factors; M_a-M_d for D, Maas, MTLD and MATTR respectively). This model was consistent with the hypothesis that the scores in each observed variable were mutually determined by three sources: the LD of the language sample, a method effect, and a residual term. Further, Model 4 stipulated that the method factors were uncorrelated, suggesting that the nature of the method effects was not the same. Finally, the residual terms of all the observed variables were fixed to be equal to zero.

The estimation of Model 4 converged to an inadmissible solution that included two parameter estimates with illogical values (i.e., Heywood cases). Specifically, the residual variance of the observed variable StoMATTR (i.e., estimated MATTR scores based on storytelling) and its loading on its corresponding method factor were estimated to be negative and larger than one, respectively ($\varepsilon^2_{STOMATTR}$ = -1.31, $\lambda_{STOMATTR}$ = 1.16). Based on Chen, Bollen, Paxton, Curran, and Kirby (2001) some causes of Heywood cases that could potentially apply in this case include (i) specification errors that could be associated with extracting more or less factors than necessary, (ii) bad start

values, and/or (iii) extremely low or high population correlations that result in empirical underidentification. The standardized factor loadings and the residual variances of Model 4 are presented in Table 15. The correlations among the factors can be seen in Table 16.

To identify the source of the problem several steps were taken. First, the solution of Model 4 was inspected focusing on the parameters of the offending variable. StoMATTR loaded on two factors. It loaded on a well-defined discourse-specific factor. This LD_S factor was defined by the observed indicators that were based on storytelling. StoMATTR was also allowed to load on a Method factor that was defined by the variables that were estimated using the MATTR technique (i.e., M_d). After looking at its corresponding loadings, it appeared that this specific factor was ill-defined, as evidenced by the very low loadings of the remaining MATTR estimated variables ($\lambda_{ProMATTR.LDs1} < .01$, $\lambda_{EveMATTR.LDs2} = .01$, $\lambda_{RecMATTR.LDs4} < -0.01$). The patterns of the factor loadings of the TMU's that were estimated using MTLD were very similar. They all loaded on a well-defined LD_S discourse-specific factor and an ill-defined method factor, as evidenced by the very low loadings of the three out of four MTLD-estimated TMU's ($\lambda_{ProMTLD.LDs1} = -.15$, $\lambda_{EveMTLD.LDs2} = -.08$, $\lambda_{StoMTLD.LDs3} = 0.10$, $\lambda_{RecMTLD.LDs4} = 0.01$). To further explore whether the inadmissible solution of Model 4 could be caused by over-extraction of factors, the residual variance of StoMATTR was fixed to a very small, positive value (.001) and the model was re-estimated (Model 4a).

With the exception of the significant χ^2, $\chi^2(83, N = 442) = 307.26$, $p < .001$, the rest of the fit indices suggested acceptable fit to the data. CFI was above

the cutoff value (= .96) and RMSEA appeared adequate (= .08, 90% confidence bands = .08 and .09). The SRMR was also consistent with a generally acceptable fit (= .07). The path diagram for this model can be seen in Figure 12 and the unstandardized and standardized factor loadings and residual variances can be seen in Tables, 17 and 18, respectively. The correlations among the factors can be seen in Table 19.

An inspection of the loading patterns revealed that two method factors, that were previously ill-defined in the previous solution, were ill-defined in this solution as well (i.e., had very small loadings). Also, the majority of loadings on these factors were not statistically significant. Further, it was possible that the high factor loadings associated with RecMTLD (i.e., MTLD scores based on recounts) and StoMATTR were artifacts of the parameter estimation process due to a misspecification of the model. So, even though the model demonstrated a good fit to the data, a new model, Model 5, was specified in which the method factors associated with the MATTR and MTLD estimation approaches were eliminated.

Step 5. In the next step, a new model was specified that assumed only D and Maas estimated variables to be determined by method factors. The rest of the specification reflected the same assumptions as Model 4. Model 5 was 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. Further, the correlation of the two method factors was fixed to be zero. The structure of the LD_S factors was similar to Model 4. The model converged to a solution for which,

with the exception of the significant χ^2, $\chi^2(90, N = 442) = 327.69$, $p < .001$, the rest of the fit indices suggested acceptable fit to the data, CFI = .96, RMSEA = .08 (90% confidence bands = .07 and .09), and SRMR = .07. Also, specifying a model with six parameters less did not have a noticeable effect in global fit as evidenced by the fit indices of Model 5 that were almost identical to those of Model 4a. The results from the nested model comparisons also suggested that the method factors associated with MATTR and MTLD could be eliminated without compromising model fit significantly: using the scaled difference χ^2 test statistic, the null hypothesis that Model 5 fit significantly worse than Model 4a was not rejected at the .01 level, $\Delta\chi^2(6) = 15.15$, $p = .02$. Finally, the selection of Model 5 over Model 4a was also based on the fact that the estimation of Model 5 did not require fixing parameters to prevent Heywood cases. The path diagram for this model can be seen in Figure 13 and the unstandardized and standardized factor loadings and residual variances can be seen in Tables, 20 and 21, respectively. The correlations among the factors of Model 5 can be seen in Table 22.

For the subsequent step, the model was modified to reflect the hypothesis that the two method specific sources were related. An inspection of the modification indices in the Mplus output, also suggested that this would be a reasonable next step. The highest estimate of how much χ^2 would decrease if a single parameter was added to the model, was associated with the correlation between the two method factors (approximate $\Delta\chi^2 = 46.29$; expected parameter change = .47).

Step 6. In the next step, a new model was specified, Model 6, that reflected the hypothesis that the method factors MD and MMaas were related. Model 6 was identical to Model 5 with one exception: the correlation among the two method factors was allowed to be freely estimated.

The model converged to a solution for which with the exception of the significant χ^2, $\chi^2(89, N = 442) = 274.86$, $p < .001$, the rest of the fit indices suggested acceptable fit to the data, CFI = .97, RMSEA = .07 (90% confidence bands = .06 and .08), and SRMR = .07. An inspection of the standardized solution revealed that none of the parameter estimates took on out-of-range values and all the estimates were statistically significant. Model 6 and 5 were further compared using the scaled difference χ^2 test statistic to explore whether fixing the correlation between the two method factors to zero resulted in a statistically significant decrease in model fit. Based on the results, Model 6 fit significantly better than model 5, $\Delta\chi^2(1) = 55.49$, $p < .001$. The path diagram for Model 6 is presented in Figure 14 and the unstandardized and standardized factor loadings and residual variances can be seen in Tables, 23 and 24, respectively. The correlations among the factors can be seen in Table 25.

The patterns of factor loadings across all four discourse types yielded very similar results. For the most part, factor loadings from their corresponding type factors were large in magnitude (median = .91; range of absolute values = .464 - .98). Following Bollen (1989), the loadings were used to compare the relative influence of the factors on several TMU's. For example, in procedures, ProD (i.e., D scores based on procedures) and ProMaas both depend on LD_{S1}, but $\lambda_{ProD.LDs1}$ =

.91 and $\lambda_{ProMaas.Lds1}$ = .75, respectively. This suggests that TMU_{1a} is more responsive to T_1 than TMU_{1ab} in standard deviation units. Consistently, the highest loadings were associated with the observed variables that were estimated using the MATTR and MTLD variables ($median_{\lambda MATTR}$ = .96, $median_{\lambda MTLD}$ = .94), followed by D ($median_{\lambda D}$ = .85) and Maas ($median_{\lambda Maas}$ = -.55).

In terms of the method effects, the factor loadings between the D estimated variables and the D method factor were relatively small in magnitude but nevertheless substantial ($median_{\lambda D}$ = .35; range of absolute values = .19 - .38). Further, the factor loadings between the Maas observed variables and the Maas method factor were considerably higher than D ($median_{\lambda Maas}$ = .45; range of absolute values = .27 - .62).

The loadings from the fully standardized solution were also used to gauge the relative influence of two latent factors on the same TMU. For example, TMU RecMaas was determined by two factors: LD_{S4} and M_b. The loadings from these two factors to RecMaas were $\lambda_{RecMaas.LDS4}$ = -.46 and $\lambda_{RecMaas.Mb}$ = .36, respectively. This indicates that RecMaas was more responsive to LD_{S4} than M_b. In general, D estimated variables were less influenced by M_a compared to the LD_S factors (average absolute difference between M_a loadings and T loadings was equal to .46). Conversely, with the exception of procedural discourse, Maas-estimated TMU's were heavily influenced by the M_b and in two cases, the loadings from the M_b factor were similar to or greater than the corresponding LD_S loadings. Overall for Maas-estimated TMU's, the average absolute difference between M_b loadings

and LD_S loadings was equal to .06 which indicates that they were significantly more similar in magnitude compared to D estimated variables.

Further the part of explained variance in TMU's that was uniquely attributable to factors was estimated. The computation of the percentage variance explained by a certain factor involved squaring the appropriate loadings (given the orthogonality of the solution). For example, consider ProD. In this case, by squaring $\lambda_{ProD.LDs1}$, the variance in ProD that was attributed to LD_{S1} was estimated and was found to be equal to 82.62%. Similarly, by squaring $\lambda_{ProD.Ma}$, the variance that was attributed to M_a was estimated (= 3.61%). The explained proportions of variance by each factor can be seen in Table 26.

Level 2

Step 7. Researchers and clinicians often use various sampling approaches to investigate LD_i. One of the main goals of this paper was to explore the extent to which all the indicator variables were manifestations of the same latent construct. Model 6 provided evidence that discourse-specific LD_S factors were correlated. Model 7 reflected the hypothesis that the correlations among discourse-specific LD_S factors were due to a more general factor that determined variation in scores across all the observed indicators, that is, all the combinations of discourse types and LD estimation techniques. Further, Model 7 reflected the hypothesis that once the effects of the general common factor were modeled, the residual covariation among the TMU's of the same discourse type would be determined by specific group factors related to types of discourse, independently.

Model 7 was similar to Model 6. However, it assumed the existence of an unaccounted latent factor within the structure of Model 6. This general factor was specified in addition to the four type-specific factors and the factors that were associated with TMU's that were estimated using the D and Maas techniques (i.e., the method factors). In addition, the type-specific factors were forced to be uncorrelated with each other and with the general factor. Further, the two method factors were correlated with each other but uncorrelated with everything else in the model.

The model converged to a solution for which with the exception of the significant χ^2, $\chi^2(79, N = 442) = 207.262$, $p < .001$, the rest of the fit indices suggested a good fit to the data, CFI = .98, RMSEA = .06 (90% confidence bands = .05 and .07), and SRMR = .06. Model 7 and 6 were further compared using the scaled difference χ^2 test statistic to explore whether fixing parameters to obtain Model 6 had a statistically significant impact on global fit. Based on the results, the null hypothesis that Model 6 did not fit significantly worse than Model 7 was rejected, level, $\Delta\chi^2(1) = 65.79$, $p < .001$. Further, this model specification resulted in a substantial improvement in model fit according to the rest of the fit indices as well. The fully standardized factor loadings and residual variances for Model 7 can be seen in Figure 15.

Table 28 includes the standardized loadings of all of the observed variables in Model 7 (the unstandardized loadings can be seen in Table 27). All variables loaded on a common factor and their corresponding discourse-specific factors. Only the variables associated with D and Maas were allowed to load on a

third factor. In general, the parameter estimates of the "method" part of Model 7 (i.e., the correlation between factors M_a and M_b, and their factor loadings to their corresponding TMU's) were almost identical to the parameter estimates of Model 6. Only minor differences in the second decimal figures were noted. In terms of the factor loadings of the general factor, their magnitude varied significantly as a function of discourse type and estimation technique (average based on absolute estimates = .52; range of absolute values = .19 - .83). The highest loadings were noted for TMU's that were estimated using MATTR, followed by MTLD, D, and Maas. The factor loadings for Maas-estimated TMU's were small (<.29), with the exception of StoMaas (= -.57). In terms of discourse types, the highest loadings were noted for storytelling and eventcasts TMU's. Overall, across discourse types and estimation techniques, StoMATTR had the highest factor loading (= 0.83).

Similar to Step 6, the solution was further explored by estimating the part of explained variables in TMU's that was uniquely attributable to factors. In this case, the highest proportions of variance shared to the general factor were yielded by storytelling TMU's when they were estimated using MATTR and MTLD, 69% and 64%, respectively. The explained proportions of variance by each factor for all variables can be seen in Table 29.

Post-Hoc Analyses

In the next set of analyses, the nature of the method factors was explored. In the analyses so far, complete language samples were analyzed that naturally varied in length. Based on prior literature, it was hypothesized that the method

factors were related to length effects. First, in order to make a causal inference as to whether different lengths of text produce different outcomes, the language samples were manipulated to equate them in terms of length. If the method factors were associated with length that in turn was causing LD-irrelevant variation in the observed variables, then experimentally removing the length variation would eliminate the need to model any method factors.

To illustrate this point, assume that two math items measured the ability to solve math problems but they were written in Spanish. If the items were administered to a US school, performance on both items would depend on two sources: one representing math skill and one representing Spanish proficiency. Therefore, to adequately model the items, two factors would be required: one for students' level on math and one for student's Spanish language proficiency. However, if the items were administered to a sample that was equated in terms of Spanish, modeling the data would not require a Spanish language factor because once math ability was taken into account, the items would not correlate anymore. Now, assume that the math items were measuring math ability, Spanish, and chemistry. Even if the items were administered to a highly proficient sample in Spanish, a single factor model would not be able to account for the data. Further, if solving the items depended on math ability and chemistry knowledge alone, a single factor model again would not be able to account for the data even if the items were administered to a proficient Spanish sample.

So, all language samples were equated in terms of length to test the hypothesis that method factors were associated with length. The language samples

were truncated to 75 words, which was the length of the shortest language sample in the database. The choice to truncate them to 75 words was made to retain the language samples from all 442 participants. Data were treated similarly prior to modeling in terms of screening for missing data, univariate and multivariate outliers, the assumption of normality and the metric of the variance. The same number of missing data was identified. Univariate outliers that were excluded in the first round of analyses were also excluded from this data set. No multivariate outliers were identified.

When data were ready for modeling, two models were fit to the data: Model 3(2) and Model 6(2)[4]. Model 6(2) was identical to Model 6 that was found to have a good fit to the first set of data. It included four correlated factors that had direct effects to observed variables that were associated with the same type of discourse, and; two correlated factors with direct effects to observed variables that were estimated using the D and Maas techniques. All the residual terms of the observed variables were uncorrelated. Model 3(2) had a similar configuration with one major difference: the two correlated factors associated with the D and Maas techniques were removed. Thus, Model 3(2) mirrored Model 3.

The model fit for the four models can be seen in Table 30. Model 6(2) converged to a solution with Heywood cases. The residual covariance and the latent variable covariance matrices were not positive definite. Specifically, the residual variance of ProD and the absolute value of the correlation between the

[4] The number 2 in parentheses denotes that data from the truncated samples were analyzed.

two method factors was more than 1. Further the loadings to the factors associated with the estimation techniques were very weak for 6 out of 8 observed variables determined by the method factors (ranging from .010 to .04). Therefore, despite the good fit of the model to the data, the magnitude of the loadings in conjunction with the out-of-bound parameters that were associated with observed variables loading on the method factors suggested an over-extraction of factors. The standardized factor loadings for Model 6(2) can be seen in Table 31; the intercorrelations of the factors can be seen in Table 32. In contrast, Model 3(2) yielded a very good fit, $\chi^2(79, N = 442) = 300.56$, $p < .001$, CFI = .98, RMSEA = .07 (90% confidence bands = .06 and .08), and SRMR = .03 without specifying additional method factors. The solution can be seen in Figure 16. Tables 33 and 34 contain the unstandardized and standardized loadings of all of the observed variables in Model 3(2). The correlations of the factors can be seen in Table 35. Overall, these results combined with the findings of the main analysis suggest that only with the complete language samples it is necessary to include method factors. When language samples were equated in terms of length, method factors were no longer necessary to model the data.

The second approach attempted to replicate the previous finding statistically by regressing the two method factors on the average language sample length (Length). If the method factors were reflecting length effects, strong correlations would be expected. The model, Model 6a, was similar to Model 6, but the method factors were specified to be uncorrelated. The algorithm converged to an inadmissible solution that included a correlation between the

Maas method factor and Length (= 1.01). In terms of model fit, global fit indices suggested that the model demonstrated a borderline adequate fit, $\chi^2(104, N = 442) = 373.74$, $p < .001$, CFI = .96, RMSEA = .08 (90% confidence bands = .07 and .09), and SRMR = .085.

To further explore the problematic solution, Length was entered in the model but was specified to be uncorrelated with everything else (Model 6b). The model converged to an admissible solution with poor fit: $\chi^2(106, N = 442) = 831.23$, $p < .001$, CFI = .89, RMSEA = .12 (90% confidence bands = .11 and .13), and SRMR = .11. Of particular interest were the modification indices in this model. An inspection of the modification indices, suggested that the highest estimate of how much χ^2 would decrease if a single parameter was freed, was associated with allowing the Maas factor to be regressed on Length (approximate $\Delta\chi^2 = 174.10$).

Next, Length was allowed to predict one method factor at the time. First, the Maas method factor was regressed on Length and the D method factor was constrained to be uncorrelated with Length and the Maas method factor. Model 6c converged to an admissible solution that had less than adequate model fit, $\chi^2(105, N = 442) = 481.45$, $p < .001$, CFI = .94, RMSEA = .09 (90% confidence bands = .8 and .10), and SRMR = .10. The standardized regression coefficient for the Maas method factor on Length was estimated to be equal to .998. Based on the modification indices, the highest estimate of how much χ^2 would decrease if a single parameter was freed, was associated with allowing the D method factor to be regressed on Length as well (approximate $\Delta\chi^2 = 88.75$).

In the next step, a new model, Model 6d, was specified that was similar to the one from the previous step with one exception: the D method factor was regressed on Length holding the correlation between Length and the Maas method factor fixed to zero. This model converged to an admissible solution but its model fit was relatively poor, $\chi^2(105, N = 442) = 741.95$, $p < .001$, CFI = .91, RMSEA = .12 (90% confidence bands = .11 and .13), and SRMR = .10. In this case, the standardized regression coefficient between Length and the D method factor was equal to .58. Not surprisingly, the modification indices suggested that freeing the path between Length and the Maas method factor would have the maximum change in χ^2, approximate $\Delta\chi^2 = 198.76$.

In the last step, the method factor for Maas-estimated TMU's was eliminated from the model and it was replaced by Length. The rationale behind this modification was that Length and the Maas method factor were very highly correlated to the extent that one was redundant (i.e. not contributing information in the model over and above the rest). Therefore, this model was identical to Model 6, but instead of six latent variables, it was specified with five latent variables and an observed variable, Length predicting all the Maas-estimated TMU's. The model (Model 6e) converged to a stable solution that demonstrated borderline adequate fit, $\chi^2(105, N = 442) = 372.89$, $p < .001$, CFI = .96, RMSEA = .08 (90% confidence bands = .07 and .09), and SRMR = .085. The standardized solution for this model can be seen in Table 36. It is noteworthy that in terms of the parameter estimates, the solution of Model 6e was very similar to that of Model 6.

Overall, these results suggest that the Maas method factor could be considered as a latent variable that represents length. The method factor for D, even though it was moderately-strongly related to length, could not be considered a pure length effect.

Chapter 4

Discussion

Level 1

In this part, I discuss the conceptual implications of the model comparisons described in steps 1 to 6. Then, I focus on the substantive interpretation of the parameter estimates of the best fitting model (Model 6) to answer the Level 1 questions.

Step 1. The first model (Model 1) that was estimated assumed that the relationships among all the observed variables (i.e., Type Method Units; TMU's) could be accounted for by a single latent construct that could be interpreted as general lexical diversity (LD). A major assumption that was consistent with this perspective was that once the influence of the general factor was taken into account, the scores on the observed indicators would be independent. If Model 1 were found to fit the data well, and the observed variables were strong indicators of the underlying construct, it would provide evidence that all TMU's were measuring a single trait, LD_i, with no systematic effects due to sampling method or estimation technique. However, the poor fit of the model suggested that there were residual relationships that the unitary factor model did not account for. Further, the poor fit of the unidimensional model suggested that perhaps a more complex factor structure would have to be specified to better represent the underlying structure of the data.

Step 2. The second model that was tested (Model 2) was consistent with two hypotheses. First, LD scores derived from the same language sample were

determined by a common latent variable. In other words, the four estimation techniques for each type of discourse were measuring the same construct. After taking into account the covariance among the tasks due to the discourse-specific LD_S, the observed variables had nothing else in common. The second hypothesis was that LD_S across different types of discourse was unrelated. Conceptually, that would be consistent with the idea that knowing the LD_S of an individual when she or he tells stories would reveal no information regarding that person's LD_S when recounting past experiences.

Model 2 resulted in a large increase in model fit compared to Model 1 but still failed to demonstrate acceptable fit to the data. If this model had been found to account for the data well and the strength of the loadings had been found to be substantial, it would have provided some evidence in favor of the validity of the score interpretations generated by each technique within each type of discourse. However, its failure to fit the data well signaled a failure to obtain evidence that would support the hypothesis that when a language sample from a given discourse is analyzed, every estimation technique yields estimates that are reflections of the LD of the language sample and very little of anything else.

Nevertheless, the contrast of the overall fit of Models 1 and 2 suggested that not specifying distinct factors to account for the intercorrelations among LD variables derived from the same discourse type was associated with a dramatic decrease in model fit.

Step 3. In step 3, a model was tested that was consistent with two hypotheses. First, similar to Model 2, Model 3 assumed that the four estimation

techniques that were used to derive the LD scores were measuring the same construct. An extension of this hypothesis, which would constitute a threat to the validity of the interpretations of the scores generated by different techniques, was that no other systematic sources of covariation would have to be specified. If Model 3 was found to account for the data well, and the factor loadings suggested strong relationships between the underlying trait and the observed LD indicators, this would constitute evidence in favor of the construct validity of the different estimation techniques.

The second hypothesis that was reflected in Model 3 was that LD_S estimates that were based on different types of discourse were related. As discussed in the introduction, there is a general consensus that different types of discourse are associated with different cognitive and linguistic demands that give rise to different patterns of microlinguistic indices such as LD. Model 2 assumed that the magnitude of the differences was such that LD_S across discourse types were not linearly related. In this step, this restriction was relaxed to assess whether, despite the variability across language sampling techniques, the LD scores that were based on different types of discourse were not independent. Instead, Model 3 reflected the belief that individuals with high LD_S in one type of discourse would have a propensity to demonstrate high LD_S in other types of discourse as well.

Model 3 resulted in further improvement of model fit compared to Model 2. This is important because correlations among different types of discourse were a prerequisite condition prior to investigating whether LD_S across different types

of discourse is determined by a single construct in the second part of the analysis. If constraining the correlations had not caused significant additional misfit, then that would have been in stark contrast to any claims that discourse-specific LD_S is determined by a general LD_i construct.

However, despite the improvement in overall fit, Model 3, still, failed to demonstrate acceptable fit to the data. Overall, the goodness of fit of Model 3 did not support the notion that the model was specified correctly; despite the facts that (i) Model 2 was found to have significantly worse fit than Model 3 and, (ii) at least in the case of one fit index, Model 3 demonstrated acceptable fit. Consequently, strong evidence was not found to support that observed scores derived using different estimation approaches and language-sampling techniques were manifestations of a single latent trait for each type of discourse.

Step 4. The a priori hypothesis that generated the next model was that observed indicators that were produced by the same estimation technique were related for reasons that were independent of discourse-specific LD_S. This was one of the main questions of the study. LD_S was measured in procedures, eventcasts, storytelling, and recounts. For each type of discourse, four estimation techniques were used to estimate LD_S. It was possible that some variation of the observed variables might have been due to the estimation technique rather than the substantive portion of the measurement. If a model that allowed for method factors was found to account for the data well, that would imply that scores that were derived using a specific methodology were not pure estimates of the construct of interest (in this case discourse-specific LD_S). Instead, it would

suggest that the scores may be a function, to a greater or lesser extent, of the estimation technique used in addition to discourse-specific LD_S.

In addition to the hypothesis that was consistent with the presence of method factors, Model 4, similar to Model 3, also assumed that the four estimation techniques that were used to derive the LD scores were measuring the same construct in each type of discourse. Also, LD_S across discourse types were believed to be related. After taking into consideration both the content as well as the estimation-specific sources of variation/covariation, observed indicators were expected to be unrelated.

The estimation of Model 4 converged to an inadmissible solution that included parameter estimates with illogical values. To obtain a stable solution, parameters had to be fixed to predetermined values and the model was re-estimated. The new model (Model 4a) demonstrated acceptable fit to the data.

Of greater interpretive importance were the standardized factor loadings in Table 18. Specifically, this solution provided evidence for the presence of only two method factors in the data. First, each observed variable that was estimated using D was determined by three sources: (i) a factor that was defined by variables that corresponded to a specific type of discourse, (ii) a factor that was defined by D estimated measures only, and (iii) residual variance. The case was similar for Maas with the exception that the factor loadings from the Maas method factor to Maas estimated measures were of greater magnitude compared to the factor loadings from the D method factor to the D estimated variables. However, as opposed to the D and Maas methods, the method factors that were

defined by the MTLD and MATTR estimated variables were ill-defined. In other words, the influence of the method factors for the latter two estimation techniques appeared to be negligible for three out of four observed variables based on the magnitude of the corresponding loadings. So, even though the model demonstrated a good fit to the data, a new model, Model 5, was specified in which the method factors associated with the MATTR and MTLD estimation approaches were eliminated.

Step 5. The results from this step suggested that eliminating the two method factors that were associated with the MTLD and the MATTR techniques did not have a significant impact on how well the model could account for the data. The fact that the MTLD and MATTR method factors were not necessary to model the data reinforced the suspicion that Model 4 did not converge due to a possible over-extraction of (unnecessary) factors.

These findings could be interpreted as evidence for arguing that MTLD and MATTR were primarily determined only by discourse-specific LD_S and little of anything else. In contrast, that was not the case with D and Maas estimated variables. Regarding the latter variables, according to the model, there were three sources that had direct effects on them and jointly determined them: (i) factors that were defined by variables that corresponded to a specific type of discourse, (ii) a factor that was defined by D estimated measures only (or Maas estimated variables), and (iii) residual variance.[5]

[5] These findings had very important implications for the validity of the score interpretations generated by different estimation techniques, and will be

Step 6. The model in step 6 was similar to Model 5 with one difference: method factors were allowed to correlate. Thus, Model 6 was consistent with the hypothesis that method factors did not necessarily represent distinct sources of covariation in the data. The rest of the model reflected the same hypotheses as Model 5.

The model converged to a stable solution that overall demonstrated adequate model fit. Importantly, Model 6 demonstrated statistical improvement compared to the more restricted Model 5 as evidenced by the significant $\Delta\chi^2$ test. The preceding steps accompanied by statistical tests of global fit and nested model comparisons seemed to point to the fact that Model 6 was an appropriate representation of the data.

In what follows, the parameters of Model 6 were explored to address the first set of questions of this paper.

MATTR & MTLD. Across all four types of discourse, the scores that were generated by the different techniques were strongly related to their respective content factors, with the exception of the Maas index. Specifically, MATTR and MTLD were consistently very strongly influenced by the discourse-specific LD_S factors, followed by D and then Maas. The first two techniques averaged 91% of the variance shared with the content factor across discourse types. An important implication of this finding was that MTLD and MATTR appeared to be stronger indicators of the LD of a language sample compared to D

---

further discussed in later sections.

and Maas. Therefore, one could argue that, holding everything else constant, these two techniques when applied to samples of the discourse type elicited in this study would provide a more accurate reflection of the LD_S.

More specifically, the highest average proportion of variance shared between measures and content factors across discourse types was associated with MATTR estimated variables. Specifically, the average proportion of variance across discourse types for MATTR-estimated variables was 93%. These results suggest that MATTR, using a window of 50 tokens, was the strongest indicator of the LD_S regardless of type of discourse in this study. This finding is significant because it highlights the potential of this new measure to produce estimates that are valid and reliable indicators of LD_S. Further, this finding is important because currently there is very limited research with this measure.

Regarding MTLD, the findings of this study confirm and expand previous results reported in the literature. For example, in earlier studies, 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). However, the methodology that had been utilized up until this point to collect validity evidence regarding LD indices had relied primarily on the examination of correlational relationships among variables to establish convergent validity; and, as discussed in the introduction, reaching conclusions about construct validity by relying exclusively on correlational approaches comes with limitations. In the current study, TMU's that were estimated using D, Maas, and MTLD loaded strongly on a common factor that represented the LD_S. Based on the structure of

the model and its parameters (i.e., strong factor loadings), it was concluded that a large chunk of the observed correlations among the measures was due to a common source, i.e., discourse-specific LD_S. To the extent that our results are accurate, Model 6 represents a more accurate representation of how TMU's relate which could be used to explain the high correlations that have been observed across other studies in the literature.

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. In particular, the results from Step 5 suggested that TMU's that were generated with these techniques were influenced by a single factor that was, in turn, defined by variables that corresponded to a specific type of discourse. Once the variance explained by the content factors in the observed variables was accounted for by the common factor, residual variances for MATTR- and MTLD-estimated TMU's were relatively small. This finding held across all types of discourse that were examined in this study (average $\varepsilon_{MATTR} = 7\%$, average $\varepsilon_{MTLD} = 11\%$,).

Taken together, the findings that (i) MTLD and MATTR are strong indicators of discourse-specific LD_S, (ii) they do not have systematic effects from construct-irrelevant sources, and (iii) their residual variances are very small, constitute evidence in favor of the validity of their score interpretations at Level 1. When a language sample is elicited, researchers and clinicians can choose from a variety of techniques to estimate the LD_S. When a given technique yields a score, the evaluator, often implicitly, draws an inference about the LD_S of the language

sample. Using Toulmin's framework that was described in the introduction (Validity Section), to justify a particular score interpretation an inference license is required that Toulmin referred to as the warrant. To make the reasoning step from the observed score produced by a specific technique to the inference about the LD_S, a warrant of the form "the observed score reflects the LD of the language sample and little of anything else" would have to be used. Assuming that Model 6 was a good approximation of the true processes that underlie the data, its parameters suggested that for MTLD and MATTR this reasoning step was warranted. Figure 17 demonstrates the argument structure for MATTR informed by the results of the current study.

D. Regarding D, the parameters of Model 6 suggested that the validity of its score interpretations in Level 1 was not as strong as for the two aforementioned measures. First, the average loadings of D-estimated TMU's across discourse types was .83 and 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 the underlying construct, nevertheless they are considerably lower than the respective parameter values of MATTR (i.e., 93%) and MTLD (i.e., 90%).

More importantly, the current study revealed that when language samples were measured using D, there were two sources that determined their scores. First, for each discourse type, scores were determined by the same factor that influenced the scores across all four TMU's. Arguably, this factor represented the LD_S of a specific discourse type. But, unlike MTLD and MATTR, D-estimated

TMU's also reflected the additive effects of a D-specific method factor[6]. Even though on average the proportion of variance that was accounted for by this method factor was not considerably high (= 11%), nonetheless it constitutes a second dimension along which D scores varied systematically.

Further, the findings from this study provide a coherent explanation of 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 the results of a discriminant analysis in which scores generated by these three techniques were used to discriminate different types of discourse. Results suggested that D, Maas, and MTLD contributed unique information to the prediction model based on which 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. That is, MTLD, vocd-D (or HD-D), and Maas all appear to be able to capture unique LD information" (p. 390-391). Based on Model 6, these results are not surprising and could be explained on the basis of the different sources of covariation in the data. These techniques produce scores that are strongly correlated because they are heavily influenced by the same factor. But, D and Maas are also 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.

---

[6] A post hoc analysis suggested that this method factor might have been related to length (see Section "The Nature of Method Factors").

Maas. Based on this study, the interpretation of a Maas score as a clear indication of the LD_S is not warranted. Maas demonstrated the lowest average loading across discourse types (= -.55[7]). The average unique variance attributed to the discourse-specific LD_S factors for TMU's that were measured using Maas was **equal to .46.** This was the lowest average across all four techniques that were utilized in this study. Moreover, Maas-estimated TMU's were influenced heavily by a second method factor. The average proportion of variance that was attributed to the method factor was approximately 22%, more than twice 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. Also, when Maas scores were based on storytelling, the loading from the content factor was only slightly higher than the loading from the Maas method factor ($\lambda_{3b.3}$ = .60, $\lambda_{3b.b}$ = .57). The implication of these findings is that Maas scores were so strongly influenced by a variable other than the construct of interest, they should not be interpreted as valid indicators of the LD of a given language sample.

The post-hoc series of analyses that was conducted explored the nature of the method factors. Two approaches were taken: an experimental and a statistical. First, the language samples were truncated to equate them in terms of length and LD scores were estimated for each type of discourse using the four estimation techniques. Then, two models were fit to the data, one that was consistent with the presence of method factors over and above the content factors, and one that was

[7] As a reminder, the negative value indicates that higher Maas scores are indicative of lower LD.

consistent with content factors only. In the second approach, the average language length across samples was introduced into the model as a predictor of the method factors. If the method factors were associated with length influencing measurement systematically, it was expected that length would be a strong predictor of both method factors.

The Nature of the Method Factors

The post-hoc series of analyses employed two different approaches to explore the nature of the method factors. Overall, the results from both provided converging evidence that both method factors were associated directly or indirectly with length.

Box (1966) argued that "To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)" (p. 629). From that perspective, one could argue that the strongest evidence regarding the nature of the method factors came from experimentally manipulating the data. In the experimental approach, a comparison of Models 3, 3(2), 6 and 6(2) revealed that when the observed scores were based on complete language samples, including 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 not unwarranted.

Another related, yet noteworthy finding was that when language samples were truncated, D and Maas demonstrated significantly higher factor loadings and

in most cases they outperformed the MTLD and MATTR indices. This suggests that by employing truncation D- and Maas-generated scores became strong indicators of LD. Of course, the point here is not to advocate truncation as a means to increase the validity of the scores. Rather, the point is to emphasize that a larger proportion of variance in the Maas scores was determined by the content factors (see more discussion on this in Section "Conclusions") given that they were the sole determinants of the TMU's.

However, even though perturbing the system to discover causal relationships among variables is often preferable to doing it statistically, nevertheless, the statistical approach revealed an interesting finding. After manipulating the data through truncation, and controlling for length effects experimentally, the conclusion that was reached was that the method factors were somehow related to the length. The results of the statistical approach supported this conclusion; but they further suggested that even though the Maas method factor could be considered isomorphic to length, that was not the case with the D method factor, which was moderately-strongly related to length. The implication of this was that the interpretation of the D method factor was not as straightforward.

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 Dugast's U (Dugast, 1978) was monotonically influenced by length. Dugast's U is a notational variant of the Maas a^2 index:

$$a^2 = \frac{logTokens - logTypes}{log^2 Tokens}$$

$$U = \frac{log^2 Tokens}{logTokens - logTypes}$$

and therefore:

$$a^2 = \frac{1}{U}$$

So, Tweedie and Baayen's conclusions about Dugast's U would also apply to scores generated by Maas. Similar conclusions were reached by Cossette (1994) who argued that Dugast's U was not length invariant.

However, our findings are not consistent with the conclusions reached by McCarthy and Jarvis (2007) who argued that length's influence on Maas was minimal. In their study, they used 23 genres from 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 significantly more in spoken genres than written genres. Evidence for this can be found in McCarthy and Jarvis' paper. The authors reported the percentage of variance Maas shared with the mean token size *across all corpora*. However, an examination of Table 7 (p. 480) from the same paper suggests 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 genres separately would be equal to 1% and 10%, respectively. This would suggest that one would expect a much stronger length effect in data such as the data reported in this study. Of course this leaves open the question "why does Maas (or Dugast's U) behave so differently across different modalities" to be addressed in future studies.

However, even though relatively strong evidence was found that the Maas method factor was reflecting length effects, the picture was less clear regarding the D factor. When length was experimentally "factored out", the D method factor was not necessary anymore. This finding suggested a close relationship between length and the D method factor. However, when the same factor was regressed on length, they shared approximately 35% of their variance. If the D method factor

were reflecting pure length effects, higher estimates of the relationship between the D and Maas factors would have been expected.

One possible explanation is that D might not be directly related to length but rather to the number of topics or themes a speaker produces. This hypothesis was first put forward by Richards (2001) when commenting on Owen and Leonard's (2002) study. Owen and Leonard found that for 500-word samples, D scores were *higher* 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. The introduction of new themes in turn, Own and Leonard argued, "would, by necessity, introduce new content words, altering the D scores of the children towards a slightly higher score" (p. 934).

If this was indeed the case, it could also explain why the correlation between the Maas and D method factors in our study correlated only .6. If D and Maas were directly related to length, one might have expected a higher correlation between their 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 exactly length relates to D, the results from Owen and Leonard's study are consistent with the configuration and the

parameter estimates of Model 6. For example, if the loadings from discourse specific LD factors to D-estimated TMU's 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 6. 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.

Clinical and research implications: Level 1

There are several situations in which the technique one uses to estimate LD_S 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 volubility 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 the LD_S. 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_S because of the length effects. Subsequently, that could mask or minimize 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. This is another situation where the decision to use a problematic technique to estimate LD_S could lead to uninterpretable or possibly erroneous conclusions. This could be the case because when holding everything else constant, including any effects of the treatment approach, the selection of the LD index could render the scores susceptible to length effects. It is possible that clinicians could be making judgments about the efficacy of a treatment by misinterpreting length effects as changes in LD_S. These length effects could be manifesting themselves in a systematic fashion as a function of establishing rapport or getting familiar with the task demands.

From the four techniques that examined in this study, to the best of our knowledge, 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.

It is also important to note given that the estimation of LD_S is plagued by the issue of length influence on LD scores, 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_S using one of the methods that were found to be 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 provided evidence suggesting that these techniques generated scores that were strong indicators of the LD of language samples. However, a great advantage of MTLD over MATTR is that, at this point, the former is actively researched. Therefore, more evidence has been accumulated that favors the validity of MTLD. On the other hand, very little is known about MATTR and

future studies should be designed to explore its performance under different conditions. Particularly, given that a great 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. This 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 et al., 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. On the other hand, if the meaning of the scores is easily conveyed, it enables more meaningful communication between clinicians, patients, and their families.

Level 2

In the previous section of the discussion, the emphasis was on the interpretation of Model 6 in terms of its implications regarding measuring the LD of language samples using different estimation techniques. Model 6 included two method factors that determined the scores in TMU's that were estimated using the D and Maas estimation techniques. These factors were found to be associated with each other and quite possibly related, directly or indirectly, to language sample length. Further, Model 6 included four discourse-specific LD_S factors. Each of these factors was the common source of variation in TMU's that were

estimated based on the same type of discourse. Importantly, for the second part of the discussion, these discourse-specific LD_S factors were found to be correlated. However, Model 6 did not make any assumptions regarding the nature of these correlations.

In what follows, the focus is on LD_i and the influence of the discourse type speakers choose to produce in response to specific language elicitation techniques. As discussed in the introduction, there is a general consensus that various types of discourse are associated with different cognitive and linguistic demands (Bliss & McCabe, 2006; Brady et al., 2005; Nicholas & Brookshire, 1993; Ulatowska, Allard, & Chapman, 1990). Several theories suggest that these demands influence the mental processes that may give rise to linguistic variables such as LD. The findings from fitting Model 7 to the data are discussed and interpreted to answer the questions of the second part of this paper. Specifically, the focus is on how individuals' productive vocabulary manifests itself across different types of discourse.

As a reminder, Model 7 was similar to Model 6. However, it assumed an unaccounted latent factor within the structure of Model 6. This general factor was specified in addition to the four discourse-specific factors and the method factors. So, each observed variable was determined by two or three sources: the general factor; the discourse-specific factors; and, for D- and Maas-estimated TMU's, the method factors. More importantly, this model was consistent with the hypothesis that once this common factor was taken into account, the discourse specific-factors would be unrelated.

The addition of the general factor and the additional constraints imposed on the model changed the substantive interpretation of the discourse-specific factors significantly. The general factor reflected the commonality across all 16 observed variables (i.e., TMU's) and represented a general factor of LD_i[8]. The discourse-specific factors were assumed to be uncorrelated with everything else in the model because they captured unique information about the language samples elicited using a specific technique. In other words, the discourse-specific factors captured the "left-over" covariance –among TMU's of the same discourse type– that could not be explained by the general factor.

In terms of the factor loadings of the general factor, their magnitude varied significantly as a function of discourse type and estimation technique. As a reminder, for variables that load on the same factor, higher loadings indicate more responsiveness to latent factor. Also, when an observed variable loads on two latent factors, comparing the loadings can be used to assess the relative influence of the factors. Therefore, loadings were used to compare the relative influence of the factors on the variables. Further, given that all observed variables were determined by orthogonal factors, squaring the loadings provided the part of variance in the measure accounted for by each factor.

---

[8] One could argue that the global factor that is extracted might not reflect LD, but rather a global method factor. That seems unlikely because after specifying the general factor, the model fit improves despite the fact that the correlations among the discourse specific factors are fixed to zero. However, the correlation between the two method factors remains unchanged (taking into account the 95% CI for the two estimated correlations in models 6 and 7). Therefore, it seems more plausible that the general factor represents a global LD factor than a method factor.

Storytelling and eventcasts. LD_i was defined in the introduction as a speaker's capacity to deploy a diverse vocabulary, by accessing and retrieving lexical items from a relatively intact knowledge base (i.e., lexicon) for the production of discourse. A language sample is an instantiation of a discourse type and its LD_S is driven to some extent by the LD_I of the individual; when the language sample intersects with an estimation technique, a numerical estimate is produced that is, possibly among other things, the observed manifestation of the latent LD_i. The types of discourse vary in the extent to which they can reflect the latent trait. Overall, the results from Model 7 suggested that storytelling and eventcasts were the most informative types of discourse of the ones included in this study for measuring the trait that underlies all LD TMU's. In other words, assuming that the general factor represents the LD_i, storytelling and eventcasts were its best indicators.

On average, the storytelling TMU's shared the most variance with the general factor (= 53%). The proportion of variance that was attributed to the general factor in StoMattr and StoMTLD (i.e., the TMU's that did not load on method factors) was equal to 69% and 64%, respectively. In terms of the Maas-estimated TMU for storytelling, it shared the highest proportion of variance with the general factor across all Maas-estimated TMU's (=32%). However, even in this case, the influence of the Maas method factor was significant as 36% of the variance in StoMaas was attributed to the method factor. When LD was estimated using D in storytelling, it still shared 46% of variance with the general factor, despite its influence from its method factor (=15%).

The second strongest indicators of LD_i, according to Model 7, were LD scores that were based on eventcasts. On average, the variables that were based on eventcasts shared a significant proportion of variance with the general common factor (= 38%). Excluding the Maas estimated scores, the average proportion explained by the general factor was 53%. The same figure for storytelling was 59% which suggests that eventcasts-produced scores were not as strong indicators of LD_i as storytelling scores but nevertheless performed relatively well. What was different in eventcasts compared to storytelling was the ordering of the TMU's. EveD shared the largest proportion of variance with the general factor which was approximately 59%. EveMTLD and EveMATTR shared 46% and 53%, respectively. However, even though D-generated scores had the strongest influence from the general factor, they were also influenced by the D method factor. Approximately 13% of variance in the D scores for eventcasts was uniquely attributed to the method factor. The proportion of unique variance that was attributed to the general factor in EveMaas was approximately 8%. For the same TMU, the method factor accounted for approximately 38% of the variance. This finding suggests that even in the case of a language sampling technique that has the potential of yielding scores that could be interpreted as strong indicators of a person's LD_i, the selection of the estimation technique can be crucial.

Why storytelling and eventcasts might be better indicators of the underlying general factor is not very clear and warrants further investigation. In terms of storytelling, it differs from the other types of discourse that were included in this study because it requires participants to communicate about

complex interactions among characters and events. Consider for example the story grammar of Picnic, which follows the schema of a typical western tradition story (Stein, 1978). The story starts with an introduction of the setting and the main characters that provides a temporal, social and physical context (a family of mice is going to have a picnic). Initially, the story elements are in a dynamic equilibrium that soon after is disturbed by an initiating event (the baby mouse falls out of the truck on the way to the picnic site but no one notices her). The initiating event, which is something beyond the control of the main characters, is the trigger that sparks off the episodes in the story. First, it might trigger emotional responses in the protagonists (the baby mouse is sad and lonely, the family worried and concerned) and soon after it leads to a quest to return to the status quo by forming a plan and taking overt action (both parties decide to look for each other). The quest is associated with different levels of tension that eventually result in the climax of the story (the baby mouse hears the truck on the road and runs to find them). The attempt to remediate the situation leads to the consequence that marks the attainment or non-attainment of the characters' goals (they all get together). And finally, there is a conclusion that reveals the protagonists' feelings about their goal attainment or non-attainment and re-sets the story elements in a state of equilibrium (everybody is excited they are together again and decide to have their picnic then and there). Eventcasts make similar demands to storytelling in that they involve description and communication of thematic, temporal sequencing, and cause and effect relationships. They are also goal directed and follow a plot structure that is usually focused on the resolution

of a problem or overcoming an obstacle. One difference is that unlike storytelling, eventcasts have simpler story grammar, usually contain fewer episodes and are not as highly structured as stories.

This complexity –more for storytelling and perhaps less for eventcasts – creates specific demands in terms of the appropriate lexical items that need to be retrieved to produce a good story or eventcast. A speaker must produce vocabulary that is well-tailored to the plot to make the narrative come to life. Importantly, the boundaries of the set of lexical items that would convey the meaning of the story are determined by the elicitation materials and the speaker is "constrained" in a sense to search for specific words. During the task, a speaker may have to retrieve lexical items with the correct shade of meaning to express the emotions, actions and interactions of the protagonists. The narrative may necessitate the selection of the proper words that would signal a change in the mood or tone in the narration as the story or eventcast unfolds and events take place; or, it might call for the use of vocabulary to evoke a visual imagery of the action. It is possible then that LD_i was better reflected in these two types of discourse because (a) the participants were presented with specific target stories that they had to tell, (b) they were all exposed to the same stimuli and the same demands applied to everyone, and (c) the success of storytelling was more contingent upon deploying a diverse vocabulary than the other types of discourse studied in this paper.

Procedures. In this study, the loadings of procedural TMU's averaged approximately .37 and they shared the least variance with the general factor

(14%). If ProMaas, which had the highest method factor influence among the procedural TMU's and the lowest from the general factor, was not included, the average loading was .40 and the average variance shared with the general factor was 16%. Overall, these results would suggest that when language samples were elicited by asking individuals to produce procedural discourse, the LD scores were rather weak indicators of LD_i.

It is possible that the reason procedural-based LD_i scores were poor reflections of LD_i may lie in how lexical items are selected to serve a speaker's communicative goals. By definition, procedural discourse lacks agents and is focused on the steps that have to take place to complete a task (Longacre, 1996). Therefore, its communicative intent is quite different from storytelling. The speaker's purpose is to provide the necessary steps in a clear and concise manner and for this reason its structure is usually significantly less complex. Consider for example the following typical language sample elicited by asking a participant how to make a peanut butter and jelly sandwich:

> Okay, to make a peanut butter and jelly sandwich you want to get out the bread. You want to get the peanut butter and the jelly out. You want to get a plate and a knife. And put all those on the counter. Then, you want to untie the loaf of bread and get out two slices of bread and… I unscrew the peanut butter jar and unscrew the jelly jar. You want to get the peanut butter out first and spread that on the bread. Then, wipe clean the knife on the other slice of

> the bread and then get the jelly out. And put your recommended amount on the other slice of bread. And then you want to put up the jelly. And put up the jar of peanut butter and the bread. Put those back in the cupboard. And then you want to put your both of your slices of bread together. And, then I like to squish them to make it the bread soft so then I… Take your hand and you press it down on the sandwich on the plate and then that's it. Then you have a peanut butter and jelly sandwich.

With respect to productive vocabulary, achieving the communicative goal of procedural discourse may depend less on searching a pool of lexical items to sample words from, and more on providing the steps efficiently. This could be achieved by using words that convey order and mechanics rather than semantics. Usually, these belong in closed-word classes such as prepositions, conjunctions and pronouns. A property of closed sets is that they offer very limited possibility for expansion. Similarly, closed word classes consist of a finite and very limited number of items (when contrasted to open classes such as nouns of adjectives) that are approximately the same across individuals. It is quite plausible that during procedures, speakers relied to a greater extent (compared to other types of discourse) on sampling from a closed set of words rather than open-word classes. Moreover, conveying the steps of a simple procedure such as how to make a peanut butter and jelly sandwich does not require specialized or diverse vocabulary. To the contrary, repeating words may enhance the clarity with which

the process is described. Overall, then, procedures may be uninformative indicators of LD_i because they predispose participants to restrict their LD_i by selecting more closed-class words and using words from a limited set of simple lexical items.

Recounts. LD scores that were based on recounting past experiences did not seem to be strongly responsive to the common general trait that all sixteen TMU's were reflecting. The loadings of D-, Maas-, MTLD-, and MATTR-generated scores included -.19, .44, .44, and .45, respectively. Even though the factor loading of the D-estimated TMU to the general factor was very similar to the MTLD and MATTR's factor loadings, it should be noted that D estimated scores were also systematically influenced by a method factor. Excluding RecMaas, on average, the TMU's that were estimated based on recounts shared approximately 20% with the general common factor.

If LD scores across different combinations of TMU's reflect a general LD_i trait, then the findings suggest that eliciting recounts to get an indication about a speaker's LD_i may be significantly less than ideal. One reason maybe that autonomous low-level cognitive processes such as LD_i may share resources with memorial cognitive processes (Rabovskya, Álvarez, Hohlfeld, & Sommer, 2008). Therefore, it is possible that LD in recounts might be influenced by the ability to activate representations in long-term memory which may infuse construct irrelevant variability in the data. Alteratively, other elicitation techniques such as eventcasts and storytelling may serve as a cognitive map or schema. When rich contextual information is provided by the task (i.e., pictured stimuli as with the

eventcasts and stories), speakers may spend fewer resources retrieving details from memory or planning and organizing their discourse; thus focusing more on accessing and retrieving specific lexical items.

Another possible account for why recounts were found to be weak indicators of LD_i may be due to the elicitation task itself. Describing past experiences such as "What did you do last weekend?" does not tap into LD_i as much as story telling even though both are considered types of narrative discourse genre.

Storytelling (and eventcasts) are more likely to draw heavily on LD resources as the participants access and retrieve the appropriate lexical items to translate their conceptual knowledge about the to-be-communicated story into discourse structures. These structures that would then allow the listener to re-construct the mental representation of the story in her/his mind. Recounts on the other hand, have less externally guided structure. Individuals can choose to convey more impoverished narratives, in terms of story grammar. Recounts are more likely to consist of a simple description of a sequence of events and might subsequently rely less on LD_i resources to convey their meaning (Wetherell, Botting, & Conti-Ramsden, 2007). Indeed, in our data, participants' language samples for the recounts were found to elicit language samples that varied considerably in terms of complexity. Typically, using a specific set of stimuli tends to elicit relatively homogenous language samples in terms of structure. However, eliciting a specific type of discourse is a non-deterministic process that entails both a predictable component and a random element. The structure of the

language sample, then, is expected to vary from one speaker to the next, even if the same procedures are followed. It appeared that recounts were more susceptible to this. This was evident both in the range of the number of words produced using this technique compared to the others (Table 4); and also in the story structure of the elicited samples. Compare for example the language samples of two individuals telling what they did for their last vacation and how they differ not only in terms of length but also in terms of structure. Both are classified as narratives using Longacre's definition (i.e., agents performing actions in chronological order); however, the second speaker produces a story that adheres more closely to Stein's (1978) story grammar discussed earlier:

Speaker A

Last winter I went to meet my significant other in Austin Texas where he was visiting where he was living for awhile. And then we came back to Phoenix for one day before traveling to San Diego for Christmas. We then went to Northern California for a short while. And then just came back and stayed in Phoenix for a few weeks.

Speaker B

On the day after new years we drove south on our way to Florida. And we got as far as a small village in Georgia where there is a resort we stayed at overnight. We found it in some travel book a

very interesting place. You stay in old style cottages. It was fun for me. Bed was too high for my wife. So they had a pair of steps you had to climb. It was an old fashion bed you know. And it was it was a nice cottage a little primitive in some ways. And they we went there for had dinner there. The food was excellent. It's basically used as a resort for hunting and fishing, particularly hunting upland birds pheasants and of course crouse. We left there in the morning and drove down debating which way to go using our G-P-S as usual to get to Amelia Island which is off the north coast of Florida. We were on our way to the Ritz Carlton and the road deteriorated. This is an interstate too. So we backed up and went cross country hit Jacksonville, went around Jacksonville and got to over to Fernandina beach and then to Ritz Carlton. We go there every couple of years, very nice. And we stayed there for about four days. I like it. And she likes it. We usually stay in the club floor because in effect you can eat and drink for free. So it makes it cheaper that way. And we really didn't go out for dinner. We just sort of lazed around. I went to the fitness room every morning. And we just had a very good time. Oh it's not done yet. We left there after four days drove up to Atlanta and stayed at the Hyatt which unfortunately was refurbishing its restaurant. So it was closed. But its in a location in Buckhead if you know Atlanta where there are restaurants all around. Buckhead and from a

> woman's point of view it's sort of like being on fifth avenue shopping wise. Great shopping from a woman's point of view. We went book shopping in the afternoon something everybody in the family loves to do and basically just sort of wandered around and talked and then went home.

As demonstrated in the above examples, the degree to which recounts depend on LD_i may vary considerably across participants depending on how they perceive the task. Given the lack of an externally guided structure, participants may choose to respond by producing more or less elaborate narratives that in turn may rely at various degrees on LD_i.

An alternative explanation for the overall pattern of results from this study could be that the magnitude of the loadings across discourse types was related to the length of the language samples. In other words, stories are stronger indicators of LD because they are longer and potentially provide more information based on which to estimate LD. However, it seems unlikely that this was case. The loadings of procedural discourse were smaller than the rest and indeed they had the smallest mean number of words and range. If the magnitude of the loadings was related to the how long language samples were, then one would expect that recounts would also demonstrate high loadings because they had similar mean number of words as eventcasts which had high loadings. This was not the case as recounts demonstrated significantly lower loadings compared to both storytelling and eventcasts.

Clinical and Research Implications: Level 2

The results of Model 7 may carry significant implications for the selection of appropriate materials when evaluating LD_i for research and clinical purposes. Currently, global indicators of discourse production such as LD_i have been underutilized in normal and impaired populations. Instead, individuals have been studied and evaluated often exclusively using decontextualized tasks that assess language skills in sterile communicative environments. Of course, two reasons for the prevalence of such approaches is the long history of problematic indices of LD_S and the poor understanding of how LD_i manifests itself across different types of discourse. The uncertainty that stemmed from the inconsistent and perplexing patterns of results in previous studies due to the aforementioned reasons cast a lack of confidence in LD scores.

This study has provided initial evidence of how different estimation techniques and types of discourse influence the measurement of LD_i. The magnitude of the loadings of storytelling and eventcasts TMU's on the general factor suggested that large differences in the observed scores would correspond to large differences in the underlying trait. Further, it suggested that observed scores based on these two types of discourse represent less construct-irrelevant variation that could threaten the validity of any conclusions. On the other hand, even large changes in the underlying trait would be expected to cause very small changes in measures based on procedures for example. For example, even if an individual's ability to access and retrieve lexical items during discourse had increased significantly via therapeutic or experimental intervention, using a task such as

procedures would probably mask the change. Therefore, investigators could potentially capitalize on storytelling' and eventcasts' sensitivity to track changes in discourse production and draw more robust conclusions especially if these types of discourse were matched with one of the LD_S estimation techniques that were found not to be influenced systematically by external factors.

Storytelling's greater sensitivity to LD_i differences suggests that it is likely to be more diagnostic. That is, if LD_i differences were expected among groups, then according to Model 7 it would be easier to uncover them using storytelling, relatively to the other types of discourse. This conclusion is consistent with the findings of at least two recent studies. First, Fergadiotis and Wright (2011) examined the effect of discourse elicitation technique in terms of LD_S in individuals with aphasia and neurologically intact adults. Eventcasts and storytelling were used to elicit the language samples and LD scores were estimated using D. Both types of discourse were expected to be associated with differences in the observed LD scores between the groups. Fergadiotis and Wright found a significant interaction between type of discourse and group membership. Specifically, the two groups differed significantly more when telling culturally familiar stories than eventcasts. Similar findings were reported by Kapantzoglou et al. (2010) who examined whether D-estimated scores differed as a function of elicitation type (spontaneous language samples elicited with pictorial support, storytelling task) and group membership (typically developing and language impaired predominately Spanish-speaking children). The authors found that the type of language elicitation procedure influenced D scores as well. Specifically,

even though the group means differed for both story re-tell and spontaneous speech, the difference was statistically significant only for story re-tell (greater between-groups difference and less variability within). Based on this finding, the authors argued that the type of discourse one uses to elicit language samples may have an impact on the diagnostic accuracy of children with specific language impairment and typically developing children.

Conclusions and Future Directions

Measuring a construct of interest, such as LD_i, is an argument that includes drawing inferences and making claims about an individual's capacities after observing his or her performance (e.g., what she/he says or does) under particular conditions (Mislevy & Yin, 2009). Based on the results of this study, eliciting a narrative using storytelling and estimating LD scores using MATTR, yielded scores that had the greatest potential of reflecting a speaker's LD_i. The argument structure for this claim can be seen in Figure 17.

Further, this paper made an explicit distinction among the observed LD score, the LD of a language sample (LD_S), and the LD of an individual (LD_i). First, this was done to alleviate the terminological confusion identified by Yu (2009). The second reason that necessitated this distinction was the uncertainty that surrounded the LD scores that are generated by various sophisticated LD estimation techniques. LD scores generated by such techniques can often yield quite different scores for the same language sample. Therefore, it is not clear which estimation technique could claim the status of the perfect indicator of the

LD_S. If there were such an estimator, then the LD observed score and the LD_S could be treated as isomorphic thus eliminating the need for the additional complexity. However, given that this is not the case, specifying LD_S as a latent variable allowed for exploring the construct validity of these techniques. Other researchers have attempted to explore the validity of LD estimation techniques using experimental and correlational approaches without explicitly distinguishing LD_S and LD_i. Ignoring the difference between the two could lead to the following problematic situation. In the face of evidence that technique X was an excellent indicator of the LD_S, the technique could be implicitly considered as an excellent indicator of LD_i. The current study demonstrated that this oversimplification does not hold. Even if a perfect estimator of LD_S were available, it would be a mistake to consider it a perfect indicator of LD_i. Even though estimation techniques may be accompanied by claims that they yield scores that are valid indicators of an underlying trait and have strong validity coefficients to support the claim, the picture might not be complete unless additional factors are considered that jointly determine the validity of score interpretations.

This dissociation becomes even clearer if one considers that it might be possible to improve the measurement of LD_S while *simultaneously* degrading the measurement of LD_i. This study offered some evidence for this hypothesis. Specifically, a comparison of the structure and discourse-specific factor intercorrelations of Models 6 and 3(2) suggest that truncation might have this effect. As a reminder, Model 6 assumed four discourse-specific factors and two

method factors associated with D and Maas. Further, the discourse-specific factors were allowed to be correlated to reflect the hypothesis that LD_S across different types of discourse were related. Model 3(2) was similar to Model 6 but it specified only discourse-specific factors and it was used with truncated language samples. When Model 3(2) was fit to the data, the loadings of D and Maas TMU's across discourse types increased significantly and the method factors were eliminated. This suggested that truncation had a positive effect for these techniques in terms of reflecting the discourse-specific factors (Table 34). However, comparing the factor intercorrelations in Models 6 and 3(2) suggested a different story in terms of LD_i. The intercorrelations of the discourse-specific factor in Model 3(2) were significantly lower compared to the corresponding intercorrelations in Model 6. This could be interpreted as an indication of a weaker effect from a common cause LD_i. It is quite possible that by truncating the language samples, the estimated scores were more influenced by discourse-specific factors than a general LD_i factor.

Besides arguing in favor of distinguishing between LD_S and LD_i, this finding has clinical and research implications as well Truncation has been advocated as a means to improve the performance of LD indices. In this study it was demonstrated that indeed truncation might yield scores that are more valid indicators of LD_S. However, evidence was also found that by truncating the language samples, the warrant that allows clinicians and researchers to make claims about individuals based on their observed scores, has probably lost some of its power as well. Nevertheless, our findings suggest that the effects of

truncation should be further investigated not only in terms of LD_S but also in terms of making inferences about LD_i. Particularly, given that a minimum of 300 tokens might be an unrealistic target for clinical and research purposes and investigators often choose lower cut-off values[9].

Future investigations should focus on exploring whether the findings generalize beyond the specific language sampling elicitation techniques that where used in this study. Ideally, to generalize conclusions about a universe of items or in this case language sampling elicitation techniques that give rise to specific types of discourse, one would have to randomly sample from that universe. However, this was not the case in this study. The tasks that were used comprised some of the most commonly used types of stimuli for eliciting language samples, and the language sampling procedures were consistent with how researchers and clinicians typically elicit language. Nevertheless, the selection was not random and more evidence would be desirable to make stronger claims about the generalizability of the results of this study.

Also, it might be fruitful to explore how the findings of this study, which was conducted with neurologically intact adults, could be applied to language-impaired individuals. LD_i was defined as the capacity to deploy a diverse vocabulary by accessing and retrieving lexical items for discourse production. For neurologically intact adults it is presumed that they do not differ significantly in

[9] For example, in this study, had I followed the recommendation to restrict language to 300 words for the eventcasts, for example, I would have to discard approximately 42% of the language samples. If the cut-off score was set to 200 tokens I would have to discard 18% of the data points.

terms of the mechanisms that support access and retrieval. However, that is not the case with people with language impairments. For example, individuals with aphasia may exhibit difficulties with accessing the semantic content or retrieving the phonological form of the word. It might be necessary to take into account at which level of processing breakdowns occur to clearly define the construct that is being measured. In a similar vein, it might be important to consider performing a lemma-based analysis to disentangle grammaticality from LD.

Further, a future direction would be to explore the degree of variability of LD_i across time. In this study, LD_i was conceptualized as a trait that is relatively stable across time and situations. However, it is quite possible that LD_i might be characterized by considerable short-term fluctuations caused by situational and/or interactional effects. According to models for the measurement of variability, e. g., models of latent state-trait theory (Steyer, Ferring, & Schmitt, 1992), inter-individual differences on one occasion of measurement (state differences) are caused by three sources of variance: (a) stable inter-individual trait differences, (b) differences in the situations in which people have performed, and (c) the interaction between the people and the situations. Intra-individual differences in task performance between occasions of measurement are explained by the variability of the situations (and/or the interactions of persons and situations) between these occasions. Knowing the extent to which LD_i exhibits large fluctuations across occasions may be critical for exploring its potential for clinical practice.

Gorin (2007) defined validity as "the extent to which test scores provide answers to targeted questions" (p. 456). In this study two aspects of measuring LD were explored: the estimation technique and the type of discourse. Strengths and weaknesses were identified for both sets. The next steps may include building upon the findings of this study to investigate the usefulness of LD scores to answer questions that pertain, for example, to differentiating subgroups among clinical populations or predicting level of impairment. In a similar vein, future investigations could explore the relationship of LD with external criteria of vocabulary knowledge such as tests of vocabulary in children and adults, both in impaired and neurologically intact individuals.

References

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology, 14*(9), 875-892. doi:10.1080/02687030050127685

Angoff, W.H. (1988). *Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), Test validity (pp. 19–32). Hillsdale, NJ: Erlbaum.*

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., & Grossman, M. (2006). Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology, 66*(9), 1405-1413. Doi:10.1212/01.wnl.0000210435.72614.38

Avent, J. R., & Austermann, S. (2003). Reciprocal scaffolding: A context for communication treatment in aphasia. *Aphasiology, 17*(4), 397-404. doi:10.1080/02687030244000743

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (2003). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1–34.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Ballard, K., & Thompson, K. (1999). Treatment and generalization of complex sentence production in agrammatism. *Journal of Speech, Language, and Hearing Research, 42*(3), 690-707.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10(*3), 340-357. doi: 10.1037/0096-1523.10.3.340

Bliss, L. S., & McCabe, A. (2006). Comparison of discourse genres: Clinical implications. *Contemporary Issues in Communication Science and Disorders, 33*(2), 126-137.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47.

Box, E. P. G. (1966). Use and abuse of regression. *Technometrics.*, *8*(4), 625-229.

Bollen, K. A. (1989). *Structural equation modeling with latent variables*. New York: Wiley.

Borsboom, D. (2005). *Measuring the mind: Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Brady, M., Armstrong, L., & Mackenzie, C. (2005). Further evidence on topic use following right hemisphere brain damage: Procedural and descriptive discourse. *Aphasiology, 19*(8), 731-747. doi:10.1080/02687030500141430

Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P., Adey, M. B., & Rose, T.L. (1982). Screening tests for geriatric depression. *Clinical Gerontologist, 1,* 37-44.

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test-retest stability of connected speech measures for adults with aphasia. *Journal of Speech & Hearing Research, 37*(2), 399-407.

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

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analyzing lexical performance. *Aphasiology*, *14*, (1), 71-91.

Cameron, R. M., Wambaugh, J. L., Wright, S. M., & Nessler, C. L. (2006). Effects of a combined semantic/phonologic cueing treatment on word retrieval in discourse. *Aphasiology, 20*(2-4), 269-285. doi:10.1080/02687030500473387

Campbell, D. T, & Fiske, D. W (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Capilouto, G., Wright, H. H., & Wagovich, S. A. (2005). CIU and main event analyses of the structured discourse of older and younger adults. *Journal of Communication Disorders, 38*(6), 431-444. doi:10.1016/j.jcomdis.2005.03.005

Carrell & Monroe, (1993) Learning styles and composition. *Modern Language Journal, 77*, 148-162.

Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, New Jersey: Prentice-Hall.

Caspari, I., & Parkinson, S. R. (2000). Effects of memory impairment on discourse. *Journal of Neurolinguistics*, *13*(1), 15–36. doi:10.1016/S0911-6044(99)00009-3

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10*(2), 157-187.

Chen, F. F., West, S. G., & Sousa, K. H. (2006) A Comparison of Bifactor and Second-Order Models of Quality of Life. Multivariate Behavioral Research, *41(2),* 189–225.

Chomsky, N. (1980). *Rules and representations.* New York: Columbia University Press.

Christiansen, J. A. (1995). Coherence violations and propositional usage in the narratives of fluent aphasics. *Brain and Language, 51*, 291-317.

Coelho, C. A., Grela, B., Corso, M., Gamble, A., & Feinn, R. (2005). Microlinguistic deficits in the narrative discourse of adults with traumatic brain injury. *Brain Injury, 19*(13), 1139-1145. doi:10.1080/02699050500110678

Coelho, C. A. (2002). Story narratives of adults with closed head injury and non-brain-injured adults: Influence of socioeconomic status, elicitation task, and executive functioning. *Journal of Speech, Language, and Hearing Research, 45*(6), 1232-1248. doi:10.1044/1092-4388(2002/099)

Cooper, P. V. (1990). Discourse production and normal aging: Performance on oral picture description tasks. *Journals of Gerontology, 45*(5), 210-214.

Corthals, P. (2010). Nine- to twelve-year olds' metalinguistic awareness of homonymy. *International Journal of Language and Communication Disorders, 45*(1), 121–128.

Covington, M.A. (2007). *CASPR Research Report 2007-05. MATTR User Manual.*

Covington, M.A., & McFall, J.D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics, 17*, 94-100.

Chotlos, J.W. (1944). Studies in language behavior. IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs, 56*, 75-111.

Crepaldi, D., Ingignoli, C., Verga, R. Contardi, A., Semenza, C., and Luzzatti, C. (2011). On nouns, verbs, lexemes, and lemmas: Evidence from the

spontaneous speech of seven aphasic patients. *Aphasiology*, *25*(1), 71–92.

Del Toro, C. M., Altmann, L. J. P., Raymer, A. M., Leon, S., Blonder, L. X., & Rothi, L. J. G. (2008). Changes in aphasic discourse after contrasting treatments for anomia. *Aphasiology, 22*(7-8), 881-892. doi:10.1080/02687030701844204

Dell, G. G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283-321.

Dell, G. S., Oppenheim, G. M., & Kittredge, K. A. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and cognitive processes, 23*(4), 583-608.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and non-aphasic speakers. *Psychological Review, 104 (4)*, 801- 838.

Dillon, C. M. & Pisoni, D. B. (2004). Nonword repetition and reading in deaf children with cochlear implants. *International Congress Series*, *1273*, 304–307.

Doyle, P. J., McNeil, M. R., Spencer, K. A., Goda, A. J., Cottrell, K., & Lustig, A. P. (1998). The effects of concurrent picture presentations on retelling of orally presented stories by adults with aphasia. *Aphasiology*, *12*, 561–574.

Dugast, D. (1978). Sur quoi se fonde la notion d'etendue theoretique du vocabulaire? *Le Francais Moderne, 46*, 25-32.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220-242. doi:10.1093/applin/25.2.220

Eid, M., & Diener, E. (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Ertmer, D.J., Strong, L.M., & Sadagopan, N. (2002). Beginning to communicate after cochlear implantation: oral language development in a young child. *Journal of Speech, Language, and Hearing Research, 46*, 328-40.

Fergadiotis, G., Wright, H. H., & Capilouto, G. (2010, May). *Productive vocabulary across discourse types.* Clinical Aphasiology Conference, Isle of Palms, SC.

Fleming, V. B., & Harris, J. L. (2008). Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology, 22*(7-8), 729-740. doi:10.1080/02687030701803762

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189-198. doi:10.1016/0022-3956(75)90026-6

Frederiksen, N. (1986). Toward a broader conception of human intelligence. *American Psychologist. 41*(4), 445-452. DOI: 10.1037/0003-066X.41.4.445

Garrard P., Maloney L.M., Hodges J.R. & Patterson K.(2005). The effects of very early Alzheimer's disease on the characterics of writing by a renowned author. *Brain, 128*, 250–260.

Geers, A., Spehar, B., & Sedey, A. (2002). Use of speech by children from total communication programs who wear cochlear implants. *American Journal of Speech-Language Pathology, 11*(1), 50-58. DOI: 10.1044/1058-0360(2002/006)

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

Goffman, E. (1981). *Forms of talk*. Oxford: Basil Blackwell.

Gordon, J. K. (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology, 22*(7-8), 839-852. doi:10.1080/02687030701820063

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science: A Multidisciplinary Journal, 27*(1), 1-40. doi:10.1016/S0364-0213(02)001118

Gorin, J. (2007). Reconsidering Issues in Validity Theory: *Educational Researcher*, 36, 456-462. DOI: 10.3102/0013189X07311607

Grela, B. G. (2002). Lexical verb diversity in children with Down syndrome. *Clinical Linguistics and Phonetics. 16*(4), 251-263.

Grimes, N. (2005). *Walt Disney's Cinderella*. New York, NY, USA: Random House.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics. 28*, 267-283.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel.

Halliday, M.A.K., & Hasan, R. (1989) *Language, context, and text: aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Heap, H.S. (1978). *Information retrieval – computational and theoretical aspects*. New York:Academic Press.

Heath, S. B. (1986). Taking a cross-cultural look at narratives. *Topics in Language Disorders, 7*, 84-95.

Heisler, L., Goffman, L., Younger, B. (2010). Lexical and articulatory interactions in children's language production. *Developmental Science*, *13*, (5), 722-730. DOI: 10.1111/j.1467-7687

Herdan, G. (1960). *Type Token Mathematics*. Hague: Mouton.

Hess, C. W., Sefton, K.M. & Landry, R.G., (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research, 29,* 129-34.

Holzinger K.J., & Swineford, F. (1937). The bi-factor method. *Psychometrika.* 2, 41–54.

Honda, R., Mitachi, M., & Watamori, T. S. (1999). Production of discourse in high-functioning individuals with aphasia—with reference to performance on the Japanese CADL. *Aphasiology, 13*(6), 475-493. doi:10.1080/026870399402037

Holmes, D. I., & Singh, S. (1996). A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing, 11,* 133–140.

Hough, M. S., & Barrow, I. (2003). Descriptive discourse abilities of traumatic brain-injured adults. *Aphasiology, 17*(2), 183-191. doi:10.1080/729255221

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hughes, D., McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment*. Eau Claire, WI: Thinking Publications.

Jakobson, R. (1971). Two aspects of language and two types of aphasic disturbances. In R. Jakobson & M. Halle (Eds.), *Fundamentals of language.* The Hague: Mouton.

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

Johnson, W. (1946). *People in quandaries.* New York: Harper.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational measurement. 38*(4), 319-342.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. (pp.39-64). In, R. W. Lissitz (Ed.), *The concept of validity. Revisions, New directions, and Applications.* Charlotte, NC: Information Age Publishing.

Kane, M. (2006). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Westport, CT: Praeger.

Kapantzoglou, M., Fergadiotis, G., & Restrepo, M. A. (August, 2010). *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.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

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

Kempen, G. & Hoenkamp, E. (1987) An incremental procedural grammar for sentence formulation. *Cognitive Science, 11*, 201–58.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*(2), 28-41.

Klee, T., Gavin, W. J., & Stokes, S. F. (2007). Utterance length and lexical diversity in American- and British-English speaking children: What is the evidence for a clinical marker of SLI? In R. Paul (Ed.), *Language disorders from a developmental perspective: Essays in honor of robin S. Chapman.* (pp. 103-140). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Klee, T., Stokes, S. F., Wong, A. M., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language & Hearing Research, 47*(6), 1396-1410.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

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

Kroll, J., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 92-107. doi: 10.1037/0278-7393.12.1.92

Lance, C. E., Noble, C, L, & Scullen, S. E. (2002). A Critique of the Correlated Trait–Correlated Method and Correlated Uniqueness Models for Multitrait-Multimethod Data. *Psychological Methods*, *7*(2), 228–244.

Lapointe, L. L., & Horner, J. (1998). *Reading comprehension battery for aphasia*. Austin, TX: Pro-Ed.

Larfeuil, C., & Le Dorze, G. (1997). An analysis of the word-finding difficulties and of the content of the discourse of recent and chronic aphasic speakers. *Aphasiology, 11*(8), 783-811. doi:10.1080/02687039708250456

Li, E. C., Volpe, A. D., Ritterman, S., & Williams, S. E. (1996). Variation in grammatic complexity across three types of discourse. *Journal of Speech-Language Pathology and Audiology*, *20*, 180–186.

Lind, M., Kristoffersen, K. E., Moen, I., & Simonsen, H. G. (2009). Semi-spontaneous oral text production: Measurements in clinical practice. *Clinical Linguistics & Phonetics, 23*(12), 872-886. doi:10.3109/02699200903040051

Loban, W. D. (1976). *Language development: Kindergarten through grade twelve. NCTE committee on research* (Report No. 18). Urbana, Illinois: National Council of Teachers of English.

Longacre, R. E. (1996). *The grammar of discourse,* (2nd ed.). New York: Plenum Press.

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik, 8*, 73-79.

MacLachlan, B. G., & Chapman, R. S. (1988). Communication breakdowns in normal and language learning-disabled children's conversation and narration. *Journal of Speech & Hearing Disorders*, *53*(1), 2–7.

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

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiolog, 24*(6-8), 856-868. doi:10.1080/02687030903452632

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon, UK: Multilingual Matters.

Maner-Idrissi, G., Rouxel, G., Pajon, C., Dardier, V., Gavornikova-Baligand, Z., Tan-Bescond, G., Godey, B., et al. (2009). Cochlear implant and lexical diversity development in deaf children: intra- and interindividual differences. *Current Psychology Letters/Behaviour, Brain & Cognition*, *25*(2), 75-111.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335– 361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait–multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15,* 47–70.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual lexical diversity* (Doctoral Dissertation). Retrieved from ProQuest Nursing & Allied Health Source database.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459-488. doi:10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392.

McCully, E. A. (1984). *Picnic*. China: Harper Collins.

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

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

McNeil, M. R. (1997). *Clinical management of sensorimotor speech disorders*. New York, NY: Thieme Medical Publishers.

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology, 15*(10-11), 991-1006. doi:10.1080/02687040143000348

McNeil, M. R., & Pratt, S. R. (2001). Defining aphasia: Some theoretical and clinical implications of operating from a formal definition. *Aphasiology, 15*(10), 901-911. doi:10.1080/02687040143000276

McNeil, M. R., Sung, J. E., Yang, D., Pratt, S. R., Fossett, T. R. D., Doyle, P. J., & Pavelko, S. (2007). Comparing connected language elicitation procedures in persons with aphasia: Concurrent validation of the Story Retell Procedure. *Aphasiology, 21*(6-8), 775-790. doi:10.1080/02687030701189980

Menn, L., Reilly,K. F., Hayashi, M., Kamio, A., Fujita, I, & Sasanuma, S. (1998). The Interaction of Preserved Pragmatics and Impaired Syntax in Japanese and English Aphasic Speech. *Brain and Language. 61*, 183–225.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger Publishers.

Mislevy, R. J., & Yin, C. (2009). If Language Is a Complex Adaptive System, What Is Language Assessment? *Language Learning*, *59*(1), 249–267.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.,pp. 13–103). New York: American Council on Education/Macmillan.

Murray, L. (2009, November). *Narrative differences among elderly with clinical depression versus early Alzheimer's.* Poster session presented at the Convention of the American Speech-Language-Hearing Association, New Orleans, LA.

Nation, P. (2007). The Four strands. *Innovation in Language Learning and*

Teaching,1(1), doi: 10.2167/illt039.0

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech & Hearing Research, 36*(2), 338-350.

Olness, G. S. (2006). Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. *Aphasiology, 20*(2-4), 175-187. doi:10.1080/02687030500472710

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing, 12*(2), 217-237. doi:10.1177/026553229501200205

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 & Hearing Research, 45*(5), 927.

Peets, K. F. (2009). Profiles of dysfluency and errors in classroom discourse among children with language impairment. *Journal of communication disorders, 42*, 136–154.

Prins, R. S., Snow, C. E., & Wagenaar, E. (1978). Recovery from aphasia: Spontaneous speech versus language comprehension. *Brain and Language, 6*(2), 192-211. doi:10.1016/0093-934X(78)90058-5

Prins, R., & Bastiaanse, R. (2004). Analyzing the spontaneous speech of aphasic speakers. *Aphasiology, 18*(12), 1075-1091. doi:10.1080/02687030444000534

Rabovskya, M., Carlos J., Álvarez, C. J., Hohlfeld, A., Sommer, W. (2008). Is lexical access autonomous? Evidence from combining overlapping tasks with recording event-related brain potentials. *Brain Research*, *1222*, 156–165.

Ransdell, S., & Åsa Wengelin, Å (2003). Socioeconomic and sociolinguistic predictors of children's L2 and L1 writing quality. *Arob@se*, *1*(2), 22-29. Retrieved from http://lu-research.lub.lu.se

Richards, B. J., and Malvern, D. D. (1997a). *Quantifying Lexical Diversity in the Study of Language Development*. The University of Reading New Bulmershe Papers, Reading.

Richards, B. J., and Malvern, D. D, (1998). *A New Research Tool: Mathematical Modelling in the Measurement of Vocabulary Diversity* (Award reference no. R000221995). Final Report to the Economic and Social Research Council, Swindon, UK.

Rider, J. D., Wright, H. H., Marshall, R. C., & Page, J. L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *American Journal of Speech-Language Pathology, 17*(2), 161-172. doi:10.1044/1058-0360(2008/016)

Rindskoppf, D., & Rose, T. (1988). Some Theory and Applications of Confirmatory Second-Order Factor Analysis. *Multivariate Behavioral Research*, 23, 51-67.

Rubin, C., & Newton, E. (2001). Capture the moment: The Pulitzer Prize photographs. New York: W. W. Norton.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, 54, 228-264.

Lee, S.Y. (2007) *Structural equation modeling: a Bayesian approach*. West Sussex, England: John Wiley & Sons.

Smith, S. L. (2009). Early phonological and lexical markers of reading disabilities. *Reading and Writing.* 22 (1) 25-40.

Somers, H. H. (1966). Statistical methods in literary analysis. In J. Leeds (Ed.), *The computer and literary style* (pp. 128-140). Kent, OH: Kent State University.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201–292.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*, 79-98.

Tabachnick and Fidell (2007) *Using Multivariate Statistics.* New York: Harper Collins Publishers.

Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minneapolis Press.

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

Tomas, J.M., Hontangas, P.M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number

of indicators and correlation among methods. *Multivariate Behavioral Research*. 35, pp. 469–499.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Tuldava, J. (1993). The statistical structure of a text and its readability. In L. Hrebicek and G. Altmann (Eds), *Quantitative text analysis* (pp. 215-27). Trier: Wissenschaftlicher Verlag Trier.

Tulving, W. D. (1972). Episodic and semantic memory. In *Organization of Memory*, (Eds) Tulving, E. & Donaldson, W., pp. 381–403. New York:Academic.

Tyler, L. K., Wesse1s, J. (1983). Quantifying contextual contributions to word recognition processes. *Perception and Psychophysics.* 34, 409-420.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*(5), 323-352.

Ulatowska, H. K., Allard, L., & Bond Chapman, S. (1990). Narrative and procedural discourse in aphasia. In Y. Joanette, & H. H. Brownell (Eds.), *Discourse ability and brain damage: Theoretical and empirical perspectives* (pp. 180-198). New York: Springer-Verlag.

Ulatowska, H. K., North, A. J., & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and Language, 13*(2), 345-371. doi:10.1016/0093-934X(81)90100-0

Van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Van Velzen, M., and Garrard, P. (2008). From hindsight to insight – retrospective analysis of language written by a renowned Alzheimer's patient. *Interdisciplinary Science Reviews*, 33, (4), 278-286.

Verhallen, M., & Scoonen, R. (1998). Lexical knowledge in L1 and L2 of third and fifth graders. *Applied Linguistics. 19*(4), 452-470. DOI: 10.1093/applin/19.4.452

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65-83. doi:10.1191/026553200676636328

Wachal, R. S., & Spreen, O. (1973). Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech, 16*(2), 169-181.

Watkins, R. V., Kelly, *D*. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*, *38*, 1349-1355.

Widaman, K. F. (1985). Hierarchically nested covariance structures models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology, 23*(10), 1295-1308. doi:10.1080/02687030902826844

Wright, H. H., & Capilouto, G. J. (2011). Measuring age-related changes in discourse production. Unpublished raw data.

Wright, H. H., Silverman, S. W., & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology, 17*(5), 443-452. doi:10.1080/02687030344000166

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236-259.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. New York, NY, US: Cambridge University Press.

Yung Y., Thissen D., McLeod L.D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika. 64*, 113–128.

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary management profile. *Language. 67*, 763-789.

Zumbo, B. (2007). Validity: foundational issues and statistical methodology. *Handbook of Statistics*. *26*(6), 45-70.

Tables

Table 1
Transformations of Type Token Ratio

Root Type token ratio[a]	RTTR = sqrtN x TTR
Corrected type token ratio[b]	CTTR = Sqrtn/2 x TTR
Log type token ratio[c]	H = logTTR

[a]Guiraud (1960); [b]Carrol (1964); [c]Herdan (1960)

Table 2
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%
12^{th} 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%
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 3

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

Lexical Diversity Index	*n*	*M*	*SD*	S^2	Range	Skewness	Kurtosis
Procedures							
D	434	37.50	12.63	159.56	14.82 - 96.04	1.07	1.85
Maas	434	121.82	20.09	403.77	74.1 - 178.67	0.16	-0.23
MTLD[a]	434	33.3	9.03	81.53	18.27 - 68.85	0.97	0.97
MATTR[b]	434	0.67	0.05	0	0.53 - 0.81	-0.03	-0.29
Eventcasts							
D	441	62.16	12.28	150.87	31.5 - 123.56	0.81	1.76
Maas	441	106.83	12.65	160.01	68.04 - 160.54	-0.09	0.66
MTLD[a]	441	50.57	11.18	124.9	25.45 - 92.19	0.8	0.9
MATTR[b]	441	0.75	0.03	0	0.66 - 0.83	-0.09	0.08
Storytelling							
D	441	57.78	13.83	191.18	22.78 - 114.71	0.92	1.19
Maas	440	106.82	12.66	160.37	68.04 - 160.54	-0.09	0.66
MTLD[a]	441	43.11	10.33	106.78	21.95 - 83.66	0.92	1.37
MATTR[b]	441	0.73	0.03	0	0.63 - 0.83	-0.1	0.01
Recounts							
D	433	64.4	14.54	211.47	28.44 - 112.15	0.3	0.07
Maas	430	105.25	16.6	275.51	56.06 - 155.77	0.04	-0.09
MTLD[a]	437	53.08	14.19	201.24	0 - 126	0.56	3.63
MATTR[b]	437	0.77	0.03	0	0.66 - 0.86	-0.03	0.04
Valid N (listwise)	423						

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 4

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

Lexical Diversity Index	*n*	*M*	SD	S^2	Range	Skewness	Kurtosis
Procedures							
Types	434	81.21	35.58	1265.93	30 - 404	2.82	17.46
Tokens	434	192.43	109.33	11952.74	53 - 1224	3.25	21.12
TTR[a]	434	0.45	0.08	0.01	0.25 - 0.66	0.3	0.01
Eventcasts							
Types	441	185.96	54.55	2976.15	74 - 507	1.17	3.26
Tokens	441	448.56	192.5	37057.06	118 - 2044	2.2	11.76
TTR[a]	441	0.44	0.06	000	0.25 - 0.63	0.35	0.41
Storytelling							
Types	441	310.18	80.83	6533.06	115 - 639	0.87	1.19
Tokens	441	1048.45	398.97	159173.99	295 - 2997	1.42	3.2
TTR[a]	441	0.31	0.04	0	0.17 - 0.47	0.09	0.27
Recounts							
Types	434	180.82	101.36	10274.03	53 - 875	2.42	9.32
Tokens	434	472.4	432	186625.84	85 - 3775	3.73	18.74
TTR[a]	434	0.44	0.09	0.01	0.2 - 0.71	0.02	0.01
Valid N (listwise)	423						

[a]Type Token ratio

Table 5

Patterns of missing data before and after the removal of outliers for each Type of Discourse

Lexical Diversity Index	Before		After	
	Number of Missing Data	Percentage of Missing Data	Number of Missing Data	Percentage of Missing Data
Procedures				
D	8	1.81%	12	2.71%
Maas	8	1.81%	8	1.81%
MTLD[a]	8	1.81%	10	2.26%
MATTR[b]	8	1.81%	8	1.81%
Eventcasts				
D	1	0.23%	3	0.68%
Maas	1	0.23%	2	0.45%
MTLD[a]	1	0.23%	4	0.90%
MATTR[b]	1	0.23%	1	0.23%
Storytelling				
D	1	0.23%	3	0.68%
Maas	2	0.45%	5	1.13%
MTLD[a]	1	0.23%	5	1.13%
MATTR[b]	1	0.23%	1	0.23%
Recounts				
D	9	2.04%	9	2.04%
Maas	12	2.71%	12	2.71%
MTLD[a]	5	1.13%	12	2.71%
MATTR[b]	5	1.13%	5	1.13%
	Averages			
	4.5	1.02%	1.75	1.41%
	Totals			
	72		95	

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 6

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

Lexical Diversity Index	*n*	*M*	*SD*	S^2	Range	Skewness	Kurtosis
Procedures							
D	430	37.03	11.70	136.95	14.82 - 75.41	0.71	0.29
Maas	434	60.91	10.05	100.94	37.05 - 89.34	0.16	-0.23
MTLD[a]	432	33.15	8.75	76.51	18.27 - 62.94	0.86	0.50
MATTR[b]	434	168.63	13.51	182.41	133.00 - 202.75	-0.03	-0.29
Eventcasts							
D	439	61.91	11.71	137.24	31.50 - 100.34	0.50	0.33
Maas	440	106.70	12.40	153.79	68.04 - 141.49	-0.25	0.15
MTLD[a]	438	50.30	10.72	114.92	25.45 - 86.17	0.64	0.40
MATTR[b]	441	300.02	11.78	138.72	264.00 - 331.20	-0.09	0.08
Storytelling							
D	439	57.53	13.36	178.55	22.78 - 101.63	0.77	0.62
Maas	437	133.72	11.06	122.24	101.30 - 161.12	0.11	-0.36
MTLD[a]	437	42.76	9.72	94.51	21.95 - 75.55	0.68	0.56
MATTR[b]	441	292.85	13.95	194.52	252.40 - 330.80	-0.10	0.01
Recounts							
D	433	64.40	14.54	211.47	28.44 - 112.15	0.30	0.07
Maas	430	105.25	16.60	275.51	56.06 - 155.77	0.04	-0.09
MTLD[a]	430	53.14	12.30	151.20	21.09 - 98.98	0.68	0.61
MATTR[b]	437	306.92	13.00	169.10	265.20 - 344.40	-0.03	0.04
Valid N (listwise)	423						

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 7
Variance-Covariance Matrix of Lexical Diversity Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Procedures																
1. D	141.8															
2. Maas	-77.46	100.71														
3. MTLD[a]	88.05	-58.64	78.19													
4. MATTR[b]	121.96	-84.61	91.34	134.78												
Eventcasts																
5. D	54.01	-20.83	28.36	42.50	139.71											
6. Maas	-8.86	29.83	10.03	-15.45	-45.45	153.41										
7. MTLD[a]	45.57	-28.88	31.82	44.69	100.68	-69.16	119.63									
8. MATTR[b]	60.81	-32.77	38.87	57.39	148.66	-89.25	153.84	251.87								
Storytelling																
9. D	44.56	-5.19	19.84	32.17	110.38	-16.23	67.67	112.67	183.15							
10. Maas	-23.72	31.20	-19.18	-28.05	-44.56	74.22	-46.76	-68.26	-51.85	123.13						
11. MTLD[a]	35.00	-15.45	22.70	31.27	76.50	-28.68	59.67	93.34	117.17	-58.99	100.41					
12. MATTR	64.96	-27.74	39.92	59.31	139.05	-50.02	107.30	174.59	208.68	107.70	171.91	320.67				
Recounts																
13. D	74.44	-39.26	44.98	66.11	68.12	-25.78	53.55	78.48	67.88	-38.38	53.08	95.45	210.45			
14. Maas	-5.05	42.10	14.76	-24.57	-16.77	67.99	-23.35	-37.25	-6.60	60.07	-21.99	-42.75	-43.07	276.41		
15. MTLD[a]	39.07	-36.52	30.92	42.31	41.50	-41.70	47.08	71.06	30.86	-43.72	41.13	77.59	121.14	102.54	158.14	
16. MATTR[b]	54.01	-47.18	44.75	61.52	63.39	-59.60	68.83	104.47	48.89	-61.19	61.07	112.01	182.63	147.05	209.72	338.27

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 8
Covariance Coverage

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Procedures																
1. D	0.97															
2. Maas	0.97	0.98														
3. MTLD[a]	0.97	0.98	0.98													
4. MATTR[b]	0.97	0.98	0.98	0.98												
Eventcasts																
5. D	0.97	0.98	0.97	0.98	0.99											
6. Maas	0.97	0.98	0.97	0.98	0.99	1										
7. MTLD[a]	0.96	0.97	0.97	0.97	0.99	0.99	0.99									
8. MATTR[b]	0.97	0.98	0.98	0.98	0.99	1	0.99	1								
Storytelling																
9. D	0.97	0.98	0.97	0.98	0.99	0.99	0.98	0.99	0.99							
10. Maas	0.97	0.98	0.97	0.98	0.98	0.99	0.98	0.99	0.98	0.99						
11. MTLD[a]	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.99	0.98	0.99					
12. MATTR[b]	0.97	0.98	0.98	0.98	0.99	0.99	0.99	1	0.99	0.99	0.99	1				
Recounts																
13. D	0.96	0.96	0.96	0.96	0.97	0.98	0.97	0.98	0.98	0.97	0.97	0.98	0.98			
14. Maas	0.95	0.96	0.96	0.96	0.97	0.97	0.96	0.97	0.97	0.97	0.96	0.97	0.97	0.97		
15. MTLD[a]	0.95	0.96	0.95	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	
16. MATTR[b]	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.99	0.98	0.97	0.97	0.99

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 9

Solution for Model 1

Type-Method Unit	Unstandardized Parameter Estimates: General Factor λ	Unstandardized Parameter Estimates: Residual Variances	Standardized Parameter Estimates: General Factor λ	Standardized Parameter Estimates: Residual Variances
Procedures				
D	3.83**	122.06	0.33**	0.89
Maas	-1.89**	97.15	-0.19**	0.97
MTLD[a]	2.43**	70.46	0.28**	0.92
MATTR[b]	3.67**	121.22	0.32**	0.90
Eventcasts				
D	8.21**	70.58	0.70**	0.51
Maas	-3.16**	143.44	-0.26**	0.94
MTLD[a]	6.27**	76.14	0.58**	0.66
MATTR[b]	10.26**	146.63	0.65**	0.58
Storytelling				
D	11.78**	42.81	0.87**	0.24
Maas	-6.01**	86.32	-0.54**	0.71
MTLD[a]	9.73**	5.42	0.97**	0.05
MATTR[b]	17.51**	13.89	0.98**	0.04
Recounts				
D	5.84**	176.78	-0.40**	0.84
Maas	-2.62**	267.98	0.37**	0.98
MTLD[a]	4.34**	132.25	0.35**	0.88
MATTR[b]	6.71**	293.65	0.37**	0.87

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 10
Unstandardized Solution for Model 2

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	Residual Variance
Procedures					
D	10.73**				25.43
Maas	-7.40**				45.95
MTLD[a]	8.06**				12.97
MATTR[b]	11.32**				6.54
Eventcasts					
D		9.77**			43.81
Maas		-6.15**			115.86
MTLD[a]		10.26**			13.50
MATTR[b]		14.98**			27.65
Storytelling					
D			11.85**		41.71
Maas			-5.91**		87.54
MTLD[a]			9.82**		4.10
MATTR[b]			17.54**		13.40
Recounts					
D				10.25**	106.15
Maas				-8.17**	208.12
MTLD[a]				11.86**	17.85
MATTR[b]				17.74**	23.75

Note. LD_S = Lexical diversity of the sample; numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Storytelling, and Recounts, respectively.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 11
Standardized Solution for Model 2

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	Residual Variance
Procedures					
D	0.91**				0.18
Maas	-0.74**				0.46
MTLD[a]	0.91**				0.17
MATTR[b]	0.98**				0.05
Eventcasts					
D		0.83**			0.32
Maas		-0.50**			0.75
MTLD[a]		0.94**			0.11
MATTR[b]		0.94**			0.11
Storytelling					
D			0.88**		0.23
Maas			-0.53**		0.72
MTLD[a]			0.98**		0.04
MATTR[b]			0.98**		0.04
Recounts					
D				0.71**	0.50
Maas				-0.49**	0.76
MTLD[a]				0.94**	0.11
MATTR[b]				0.96**	0.07

Note. LD_S = Lexical diversity of the sample; numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Storytelling, and Recounts, respectively.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 12
Unstandardized Factor Loadings and Residual variances for Model 3

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	Residual Variance
Procedures					
D	10.74**				25.23
Maas	-7.40**				45.94
MTLD[a]	8.06**				12.95
MATTR[b]	11.31**				6.68
Eventcasts					
D		9.89**			41.29
Maas		-6.04**			117.23
MTLD[a]		10.16**			15.57
MATTR[b]		15.04**			25.56
Storytelling					
D			11.83**		42.10
Maas			-5.94**		87.15
MTLD[a]			9.78**		4.64
MATTR[b]			17.58**		11.63
Recounts					
D				10.32**	104.69
Maas				-8.15**	208.33
MTLD[a]				11.90**	16.91
MATTR[b]				17.66**	26.49

Note. LD_S = Lexical diversity of the sample; numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Storytelling, and Recounts, respectively.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 13
Standardized Factor Loadings and Residual Variances for Model 3

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	Residual Variance
Procedures					
D	0.91**				0.18
Maas	-0.74**				0.46
MTLD[a]	0.91**				0.17
MATTR[b]	0.98**				0.05
Eventcasts					
D		0.84**			0.30
Maas		-0.49**			0.76
MTLD[a]		0.93**			0.13
MATTR[b]		0.95**			0.10
Storytelling					
D			0.88**		0.23
Maas			-0.54**		0.71
MTLD[a]			0.98**		0.05
MATTR[b]			0.98**		0.04
Recounts					
D				0.71**	0.50
Maas				-0.49**	0.76
MTLD[a]				0.95**	0.11
MATTR[b]				0.96**	0.08

Note. LD_S = Lexical diversity of the sample; numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Storytelling, and Recounts, respectively.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 15

Standardized Factor Loadings and Residual Variances for Model 4

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	M_c λ	M_d λ	Residual Variance
Procedures									
D	0.91**				0.20**				0.14
Maas	-0.74**					0.27**			0.38
MTLD[a]	0.92**						-0.15		0.14
MATTR[b]	0.97**							0.01	0.06
Eventcast									
D		0.83**			0.41**				0.14
Maas		-0.49**				0.65**			0.34
MTLD[a]		0.94**					-0.08		0.11
MATTR[b]		0.94**						0.01	0.11
Storytelling									
D			0.86**		0.36**				0.14
Maas			-0.57**			0.57**			0.35
MTLD[a]			0.98**				-0.10		0.03
MATTR[b]			0.98**					1.16**	-1.31
Recounts									
D				0.73**	0.32**				0.38
Maas				-0.46**		0.36**			0.66
MTLD[a]				0.95**			0.02		0.10
MATTR[b]				0.96**				-0.01	0.08

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, stories, and recounts; M = Method factor; a-d correspond to estimation techniques: D, Maas, MTLD, and MATTR.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 16

Intercorrelations Among Discoure-Specific Factors in Model 4

	LD_{S1}	LD_{S2}	LD_{S3}	LD_{S4}	M_a	M_b	M_c	M_d
LD_{S1}	-							
LD_{S2}	0.36	-						
LD_{S3}	0.28	0.64	-					
LD_{S4}	0.33	0.40	0.36	-				
M_a	0	0	0	0	-			
M_b	0	0	0	0	0	-		
M_c	0	0	0	0	0	0	-	
M_d	0	0	0	0	0	0	0	-

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-d correspond to estimation techniques: D, Maas, MTLD, and MATTR. For all estimated correlations, $p < .001$.

Table 17

Unstandardized Factor Loadings and Residual Variances for Model 4a

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	M_c λ	M_d λ	Residual Variance
Procedures									
D	10.72**				2.35**				19.12
Maas	-7.47**					2.72**			38.06
MTLD[a]	8.09**						-1.32		10.73
MATTR[b]	11.25**							-0.19	7.87
Eventcast									
D		9.62**			4.77**				19.32
Maas		-6.05**				7.99**			51.73
MTLD[a]		10.28**					-0.82		13.42
MATTR[b]		14.92**						-1.05	27.50
Storytelling									
D			10.90**		4.58**				22.55
Maas			-6.53**			6.52**			44.81
MTLD[a]			9.82**				-1.08		3.31
MATTR[b]			17.54**					-3.48**	0.00[c]
Recounts									
D				10.75**	4.67**				82.54
Maas				-7.57**		5.82**			175.34
MTLD[a]				11.90**			0.21		16.59
MATTR[b]				17.68**				0.45	26.64

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, stories, and recounts; a-d correspond to estimation techniques: D, Maas, MTLD, and MATTR. [a]Measure of textual lexical diversity; [b]Moving average type-token ratio. [C]This residual variance was fixed to 0.001 to converge to an admissible solution.

Table 18
Standardized Factor Loadings and Residual Variances for Model 4a

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	M_c λ	M_d λ	Residual Variance
Procedures									
D	0.91**				0.20**				0.14
Maas	-0.74**					0.27**			0.38
MTLD[a]	0.92**						-0.15		0.14
MATTR[b]	0.97**							-0.02	0.06
Eventcast									
D		0.83**			0.41**				0.14
Maas		-0.49**				0.65**			0.34
MTLD[a]		0.94**					-0.08		0.11
MATTR[b]		0.94**						-0.07	0.11
Storytelling									
D			0.86**		0.36**				0.14
Maas			-0.57**			0.57**			0.35
MTLD[a]			0.98**				-0.11		0.03
MATTR[b]			0.98**					-0.19**	0.00[c]
Recounts									
D				0.73**	0.32**				0.38
Maas				-0.46**		0.36**			0.66
MTLD[a]				0.95**			0.02		0.11
MATTR[b]				0.96**				0.02	0.08

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, stories, and recounts; a-d correspond to estimation techniques: D, Maas, MTLD, and MATTR. [a]Measure of textual lexical diversity; [b]Moving average type-token ratio. [C]This residual variance was fixed to 0.001 to converge to an admissible solution.

Table 19

Intercorrelations Among Discourse-Specific Factors in Model 4a

	LD_{S1}	LD_{S2}	LD_{S3}	LD_{S4}	M_a	M_b	M_c	M_d
LD_{S1}	-							
LD_{S2}	0.36	-						
LD_{S3}	0.28	0.64	-					
LD_{S4}	0.33	0.40	0.36	-				
M_a	0	0	0	0	-			
M_b	0	0	0	0	0	-		
M_c	0	0	0	0	0	0	-	
M_d	0	0	0	0	0	0	0	-

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-d correspond to estimation techniques: D, Maas, MTLD, and MATTR. For all estimated correlations, $p < .001$.

Table 20

Unstandardized Factor Loadings and Residual Variances for Model 5

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures							
D	10.74**				2.33**		19.01
Maas	-7.48**					2.73**	37.93
MTLD[a]	8.10**						12.37
MATTR[b]	11.25**						8.11
Eventcast							
D		9.66**			4.80**		18.63
Maas		-6.05**				7.98**	52.07
MTLD[a]		10.22**					14.46
MATTR[b]		14.96**					28.16
Storytelling							
D			10.91**		4.64**		22.07
Maas			-6.55**			6.54**	44.53
MTLD[a]			9.77**				4.81
MATTR[b]			17.60**				11.10
Recounts							
D				10.76**	4.61**		83.03
Maas				-7.58**		5.82**	175.38
MTLD[a]				11.90**			16.58
MATTR[b]				17.64**			27.17

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 21
Standardized Factor Loadings and Residual Variances for Model 5

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures							
D	0.91**				0.20**		0.14
Maas	-0.74**					0.27**	0.37
MTLD[a]	0.92**						0.16
MATTR[b]	0.97**						0.06
Eventcast							
D		0.83**			0.41**		0.14
Maas		-0.49**				0.65**	0.34
MTLD[a]		0.94**					0.12
MATTR[b]		0.94**					0.11
Storytelling							
D			0.86**		0.36**		0.14
Maas			-0.57**			0.57**	0.34
MTLD[a]			0.98**				0.05
MATTR[b]			0.98**				0.04
Recounts							
D				0.73**	0.31**		0.38
Maas				-0.46**		0.36**	0.66
MTLD[a]				0.95**			0.11
MATTR[b]				0.96**			0.08

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 22

Intercorrelations Among Discourse-Specific Factors in Model 5

	LD_{S1}	LD_{S2}	LD_{S3}	LD_{S4}	M_a	M_b
LD_{S1}	-					
LD_{S2}	0.36	-				
LD_{S3}	0.29	0.65	-			
LD_{S4}	0.33	0.40	0.36	-		
M_a	0	0	0	0	-	
M_b	0	0	0	0	0	-

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

Table 23

Unstandardized Factor Loadings and Residual Variances for Model 6

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures							
D	10.76**				2.25**		19.19
Maas	-7.51**					2.74**	37.57
MTLD[a]	8.09**						12.44
MATTR[b]	11.25**						8.08
Eventcast							
D		9.72**			4.45**		21.03
Maas		-6.33**				7.72**	54.51
MTLD[a]		10.23**					14.30
MATTR[b]		14.92**					29.15
Storytelling							
D			10.90**		4.89**		19.34
Maas			-6.90**			6.62**	43.16
MTLD[a]			9.78**				4.69
MATTR[b]			17.58**				11.59
Recounts							
D				10.87**	4.60**		83.03
Maas				-7.56**		5.83**	174.91
MTLD[a]				11.91**			16.47
MATTR[b]				17.64**			27.46

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 24
Standardized Factor Loadings and Residual Variances for Model 6

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures							
D	0.91**				0.19**		0.14
Maas	-0.75**					0.27**	0.37
MTLD[a]	0.92**						0.16
MATTR[b]	0.97**						0.06
Eventcast							
D		0.84**			0.38**		0.16
Maas		-0.51**				0.62**	0.35
MTLD[a]		0.94**					0.12
MATTR[b]		0.94**					0.12
Storytelling							
D			0.86**		0.38**		0.12
Maas			-0.60**			0.57**	0.32
MTLD[a]			0.98**				0.05
MATTR[b]			0.98**				0.04
Recounts							
D				0.73**	0.31**		0.37
Maas				-0.46**		0.36**	0.66
MTLD[a]				0.95**			0.10
MATTR[b]				0.96**			0.08

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 25

Intercorrelations Among Discourse-Specific Factors in Model 6

	LD_{S1}	LD_{S2}	LD_{S3}	LD_{S4}	M_a	M_b
LD_{S1}	-					
LD_{S2}	0.36	-				
LD_{S3}	0.29	0.65	-			
LD_{S4}	0.33	0.40	0.36	-		
M_a	0	0	0	0	-	
M_b	0	0	0	0	.50	-

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

Table 26

Variance Decomposition for Model 6

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures							
D	0.83				0.04		0.14
Maas	0.56					0.07	0.37
MTLD[a]	0.85						0.16
MATTR[b]	0.94						0.06
Eventcast							
D		0.71			0.14		0.16
Maas		0.26				0.38	0.35
MTLD[a]		0.88					0.12
MATTR[b]		0.88					
Storytelling							
D			0.74		0.14		0.12
Maas			0.36			0.32	0.32
MTLD[a]			0.96				0.05
MATTR[b]			0.96				
Recounts							
D				0.53	0.10		0.37
Maas				0.21		0.13	0.66
MTLD[a]				0.90			0.10
MATTR[b]				0.92			

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; a-b correspond to estimation techniques: D and Maas.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 27

Unstandardized Factor Loadings and Residual Variances for Model 7

	LD_i λ	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures								
D	1.00**	9.62**				1.99**		19.91
Maas	-0.53**	-7.01**					2.65**	38.05
MTLD[a]	0.64**	7.41**						12.39
MATTR[b]	0.93**	10.23**						8.03
Eventcast								
D	1.79**		4.44**			4.28**		18.18
Maas	-0.72**		-5.63**				7.68**	50.15
MTLD[a]	1.46**		7.42**					9.33
MATTR[b]	2.28**		9.14**					34.44
Storytelling								
D	1.80**			7.57**		5.24**		11.60
Maas	-1.24**			-1.67*			6.71**	36.41
MTLD[a]	1.59**			5.55**				4.66
MATTR[b]	2.94**			9.33**				11.36
Recounts								
D	1.33**				8.77**	3.82**		85.82
Maas	-0.60**				-6.88**		5.80**	174.66
MTLD[a]	1.10**				10.50**			17.31
MATTR[b]	1.60**				15.76**			24.96

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 28
Standardized Factor Loadings and Residual Variances for Model 7

	LD_i λ	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures								
D	0.43**	0.81**				0.17**		0.14
Maas	-0.27**	-0.70**					0.26**	0.38
MTLD[a]	0.37**	0.84**						0.16
MATTR[b]	0.41**	0.88**						0.06
Eventcast								
D	0.77**		0.38**			0.36**		0.13
Maas	-0.29**		-0.45**				0.62**	0.33
MTLD[a]	0.68**		0.68**					0.08
MATTR[b]	0.73**		0.58**					0.14
Storytelling								
D	0.68**			0.57**		0.39**		0.07
Maas	-0.57**			-0.15**			0.60**	0.29
MTLD[a]	0.80**			0.55**				0.05
MATTR[b]	0.83**			0.52**				0.04
Recounts								
D	0.45**				0.59**	0.26**		0.39
Maas	-0.19**				-0.42**		0.36**	0.66
MTLD[a]	0.44**				0.83**			0.11
MATTR[b]	0.44**				0.86**			0.07

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 29

Variance Decomposition for Model 7

	LD_i λ	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures								
D	0.18	0.66				0.03		0.14
Maas	0.07	0.49					0.07	0.38
MTLD[a]	0.14	0.71						0.16
MATTR[b]	0.17	0.77						0.06
Eventcast								
D	0.59		0.14			0.13		0.13
Maas	0.08		0.20				0.38	0.33
MTLD[a]	0.46		0.46					0.08
MATTR[b]	0.53		0.34					0.14
Storytelling								
D	0.46			0.32		0.15		0.07
Maas	0.32			0.02			0.36	0.29
MTLD[a]	0.64			0.30				0.05
MATTR[b]	0.69			0.27				0.04
Recounts								
D	0.20				0.35	0.07		0.39
Maas	0.04				0.18		0.13	0.66
MTLD[a]	0.19				0.69			0.11
MATTR[b]	0.19				0.74			0.07

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 30

Model Fit for Models 3 and 6 Applied to Complete and Truncated Language Samples

Model	Complete Data	Truncated Data
Model 6	$\chi^2(98) = 757.53$, $p < .001$ CFI = .89 RMSEA = .13 (CI = .12 - .14) SRMR = .08	$\chi^2(98) = 297.30$, $p < .001$ CFI = .98 RMSEA = .07 (CI = .06 .08) SRMR = .03
Model 3	$\chi^2(89) = 274.86$, $p < .001$ CFI = .97 RMSEA = .07 (CI = .06 - .08) SRMR = .07	$\chi^2(89) = 282.69$, $p < .001$ CFI = .98 RMSEA = .07 (CI = .06 - .08) SRMR = .02

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

Table 31

Standardized Factor Loadings and Residual Variances for Model 6(2)

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	M_b λ	Residual Variance
Procedures							
D	0.95**				0.32**		0.00[c]
Maas	-0.92**					0.25	0.09
MTLD[a]	0.86**						0.26
MATTR[b]	0.95**						0.10
Eventcast							
D		0.99**			0.04		0.02
Maas		-0.98**				0.04	0.05
MTLD[a]		0.87**					0.24
MATTR[b]		0.85**					0.28
Storytelling							
D			0.99**		-0.02		0.01
Maas			-0.97**			-0.02	0.07
MTLD[a]			0.86**				0.25
MATTR[b]			0.85**				0.28
Recounts							
D				0.99**	0.03		0.02
Maas				-0.97**		0.01	0.06
MTLD[a]				0.86**			0.27
MATTR[b]				0.85**			0.28

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

[a]Measure of textual lexical diversity; [b]Moving average type-token ratio; [c]Fixed to 0.001 to converge to admissible solution.

Table 32
Intercorrelations Among Factors in Model 6(2)

	LD_{S1}	LD_{S2}	LD_{S3}	LD_{S4}	M_a	M_b
LD_{S1}	-					
LD_{S2}	0.23	-				
LD_{S3}	0.14	0.34	-			
LD_{S4}	0.21	0.17	0.17	-		
M_a	0	0	0	0	-	
M_b	0	0	0	0	-1.12	-

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

Table 33
Unstandardized Factor Loadings for Model 3(2)

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	Residual Variance
Procedures					
D	16.32**				2.74
Maas	-19.01**				22.69
MTLD[a]	16.87**				139.29
MATTR[b]	16.52**				59.09
Eventcasts					
D		15.09**			3.39
Maas		-17.00**			14.20
MTLD[a]		16.58**			87.17
MATTR[b]		12.04**			55.22
Storytelling					
D			14.62**		2.36
Maas			-15.17**		16.98
MTLD[a]			13.17**		58.90
MATTR[b]			12.73**		64.49
Recounts					
D				15.20**	3.92
Maas				-14.83**	14.25
MTLD[a]				17.24**	106.96
MATTR[b]				12.13**	57.25

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 34
Standardized Factor Loadings for Model 3(2)

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	Residual Variance
Procedures					
D	0.99**				0.01
Maas	-0.97**				0.06
MTLD[a]	0.82**				0.33
MATTR[b]	0.91**				0.18
Eventcasts					
D		0.99**			0.01
Maas		-0.98**			0.05
MTLD[a]		0.87**			0.24
MATTR[b]		0.85**			0.28
Storytelling					
D			0.99**		0.01
Maas			-0.97**		0.07
MTLD[a]			0.86**		0.25
MATTR[b]			0.85**		0.28
Recounts					
D				0.99**	0.02
Maas				-0.97**	0.06
MTLD[a]				0.86**	0.27
MATTR[b]				0.85**	0.28

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Table 35
Intercorrelations Among Discourse-Specific Factors in Model 3(2)

	LD_{S1}	LD_{S2}	LD_{S3}	LD_{S4}
LD_{S1}	-			
LD_{S2}	0.22	-		
LD_{S3}	0.12	0.34	-	
LD_{S4}	0.21	0.17	0.17	-

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.

Table 36
Standardized Factor Loadings, regression weights and Residual Variances for Model 6e

	LD_{S1} λ	LD_{S2} λ	LD_{S3} λ	LD_{S4} λ	M_a λ	Length	Residual Variance
Procedures							
D	0.91**				0.22**		0.13
Maas	-0.76**					0.25	0.36
MTLD[a]	0.92**						0.16
MATTR[b]	0.97**						0.06
Eventcast							
D		0.83**			0.36		0.18
Maas		-0.55**				0.55	0.40
MTLD[a]		0.94**					0.12
MATTR[b]		0.94**					0.11
Storytelling							
D			0.86**		0.35		0.18
Maas			-0.61**			0.51	0.40
MTLD[a]			0.98**				0.12
MATTR[b]			0.98**				0.11
Recounts							
D				0.73**	0.36		0.34
Maas				-0.50**		0.49	0.51
MTLD[a]				0.95**			0.10
MATTR[b]				0.96**			0.08

Note. LD_S = Sample Lexical diversity; 1-4 correspond to discourse types: procedures, eventcasts, storytelling, and recounts; M = Method factor; a-b correspond to estimation techniques: D, and Maas. For all estimated correlations, $p < .001$.
[a]Measure of textual lexical diversity; [b]Moving average type-token ratio.

Figures

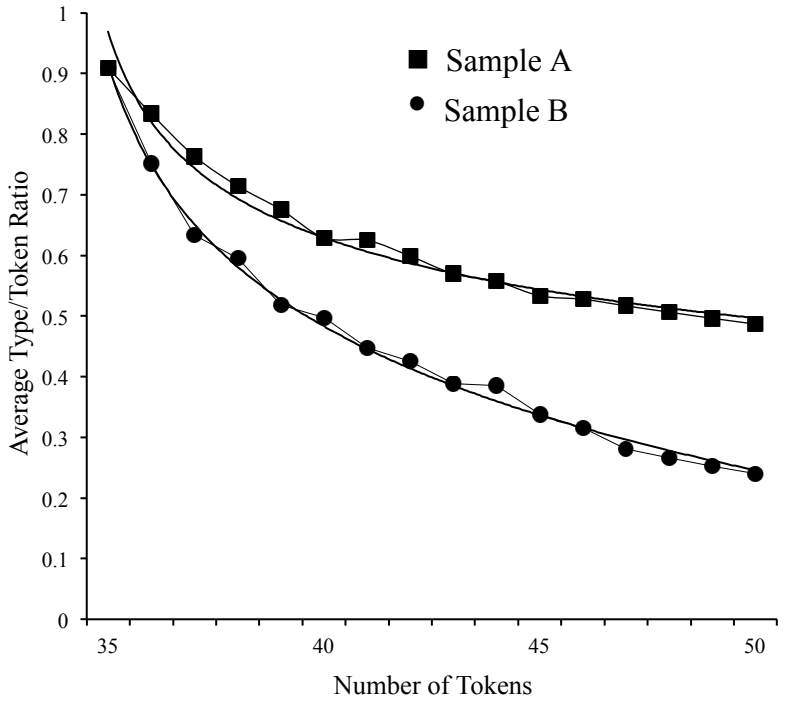

Figure 1. Estimating D. D reflects how fast the average TTR decreases. The slope of the fitted nonlinear curve corresponds to different D values. The steeper the slope of the fitted line, the lower the D value.

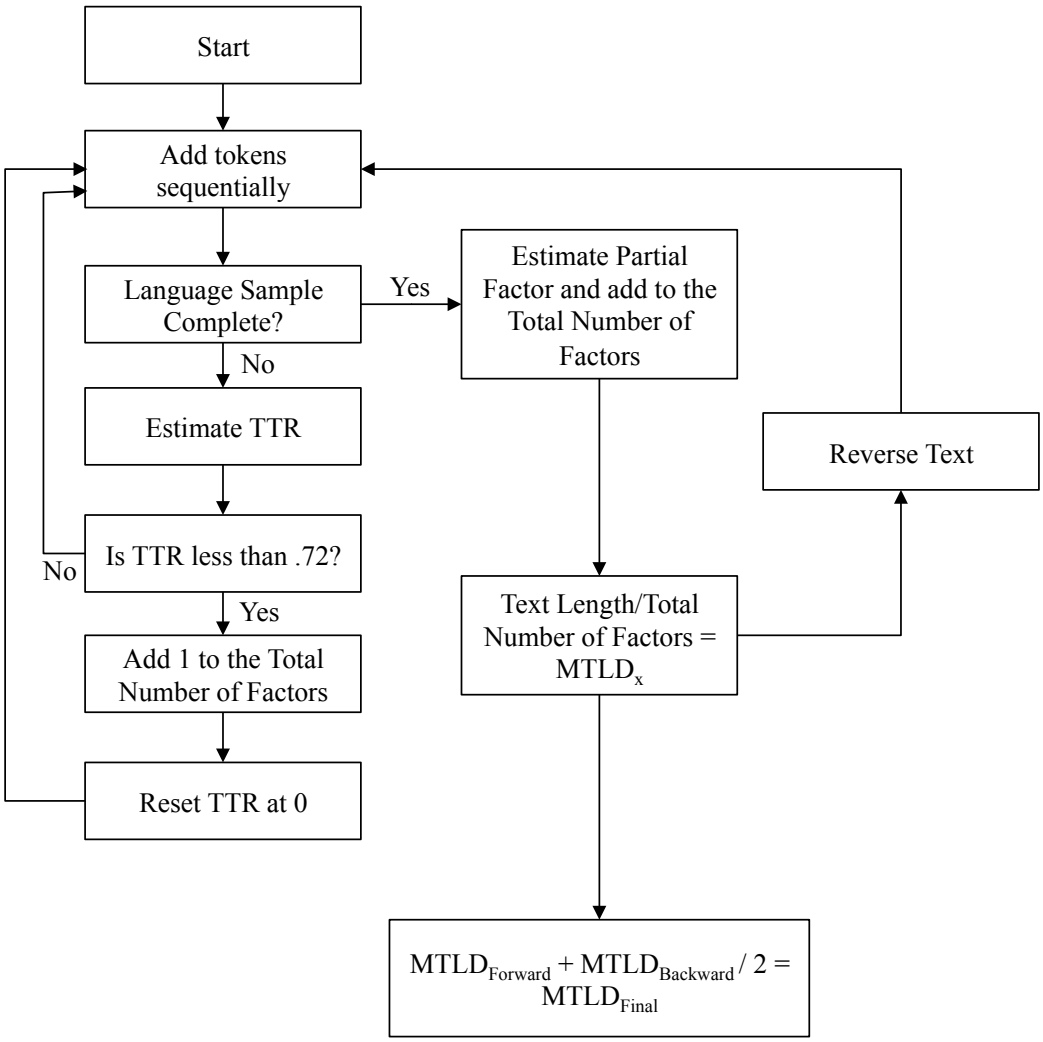

Figure 2. Measure of textual lexical diversity flow chart.

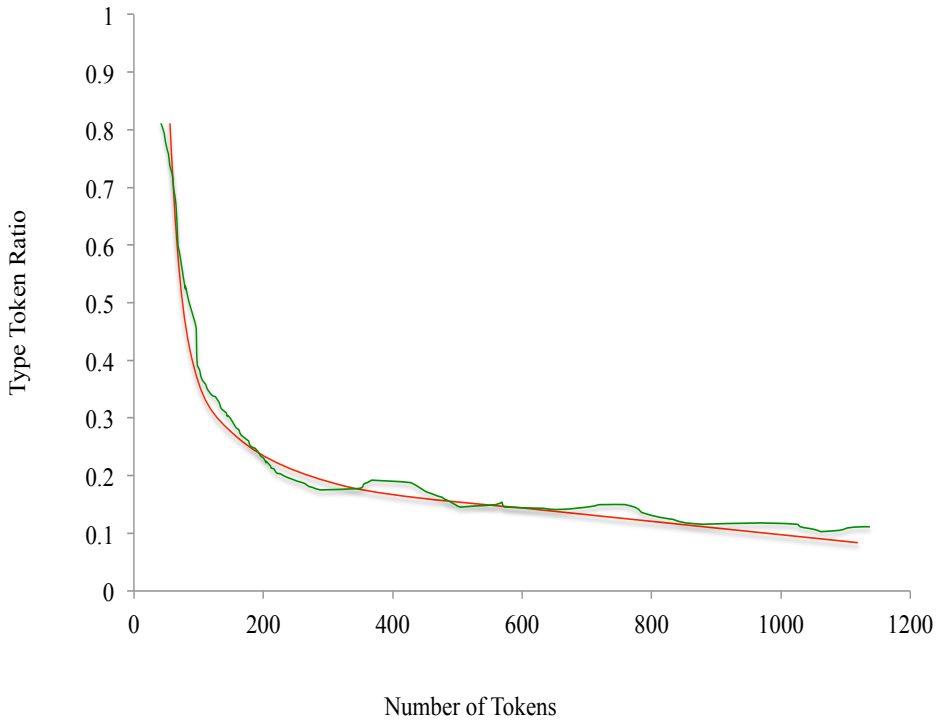

Figure 3. Emprical type-token ratio and fitted logarithmic curve.

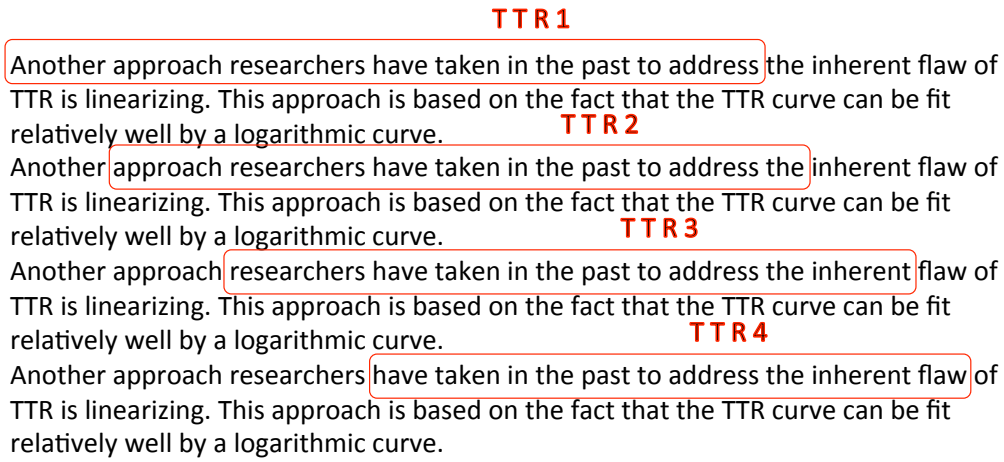

Figure 4. The Moving average type-token ratio (MATTR) in action. MATTR calculates LD scores by using a smoothly moving window that estimates type-token ratios for each successive window of fixed length (here ten words).

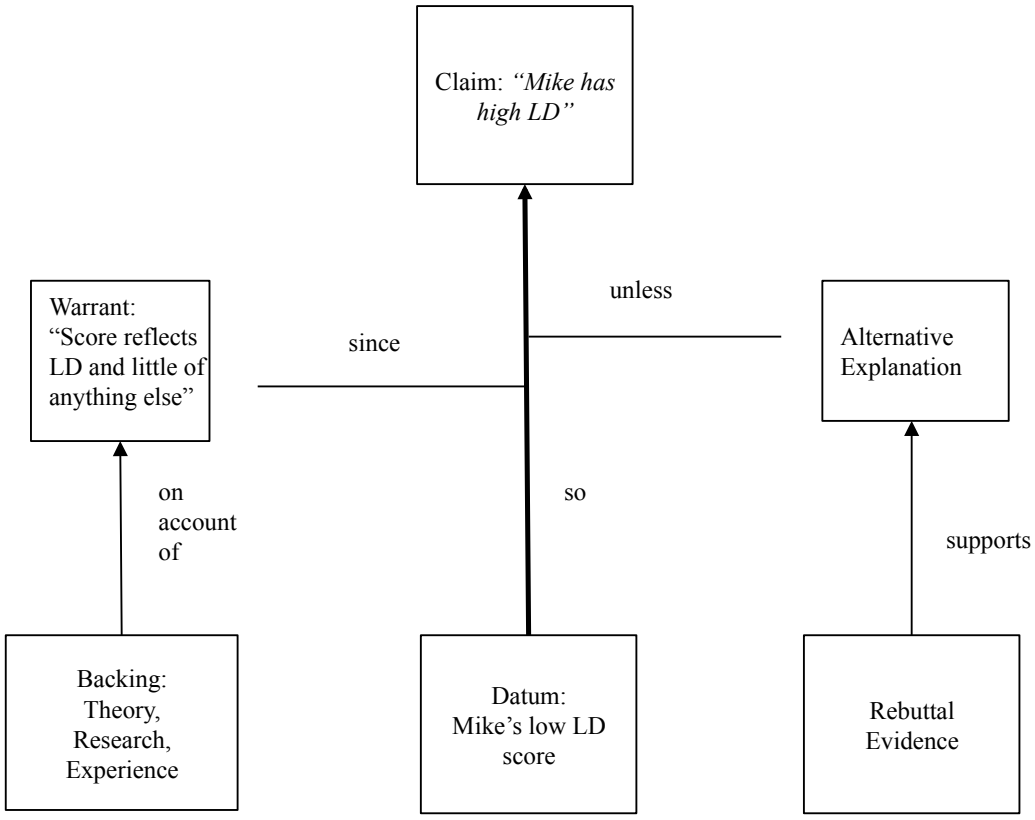

Figure 5. Toulmin's argument structure.

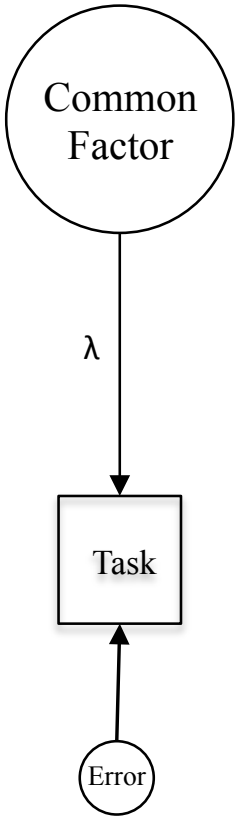

Figure 6. Latent Variable Modeling. Task scores are jointly determined by two sources: the common factor and the residual term. Common factors represent unobserved traits; residual terms represent unique variance.

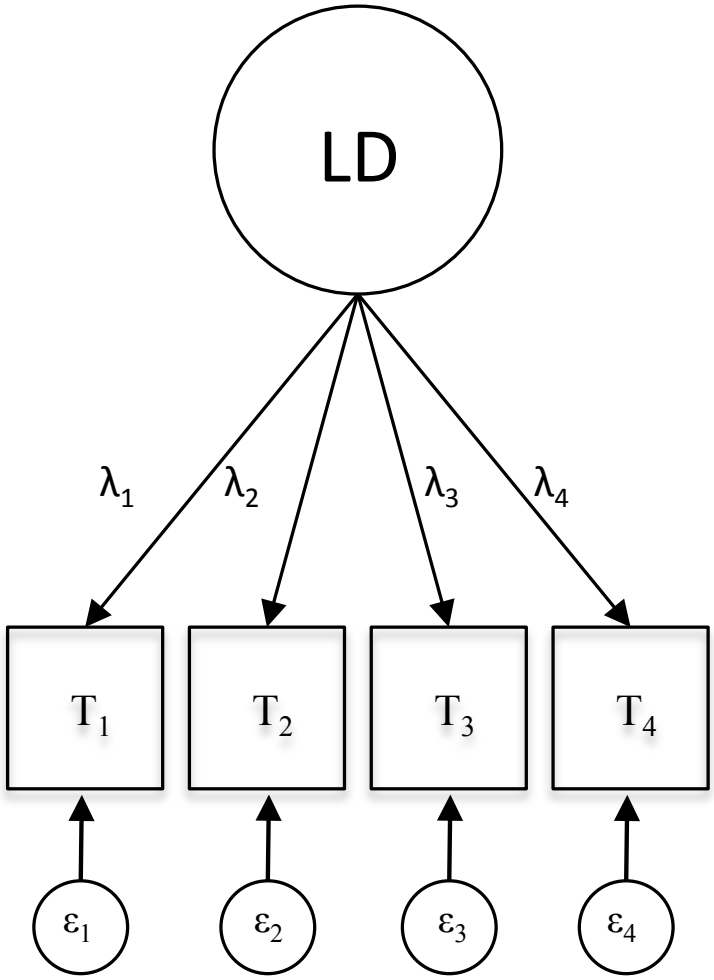

Figure 7. A unidimensional measurement model of lexical diversity (LD). LD determines the observed scores across four types of discourse (T_1 - T_4). Loadings and residual terms are denoted with Greek epsilons and lambdas, respectively.

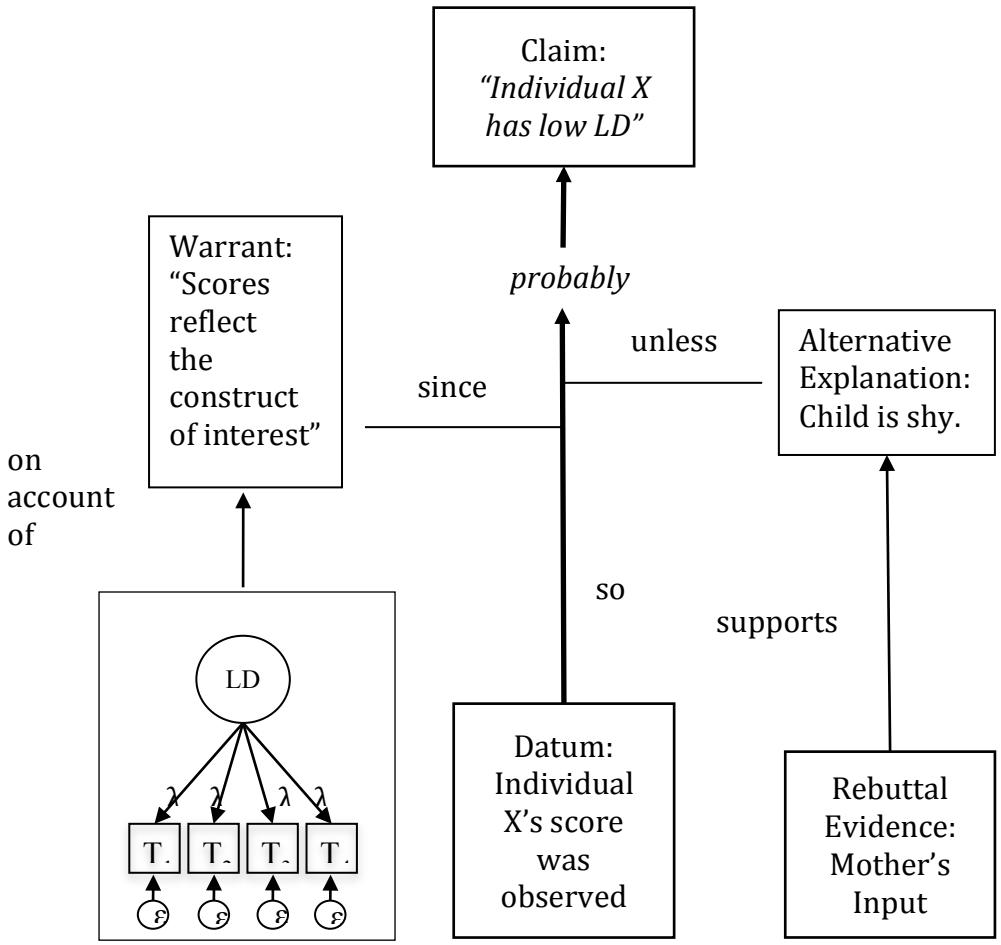

Figure 8. An argument structure for LD. A unidimensional confirmatory factor analytic model could be estimated to assess the extent to which "scores reflect the construct of interest and little of anything else".

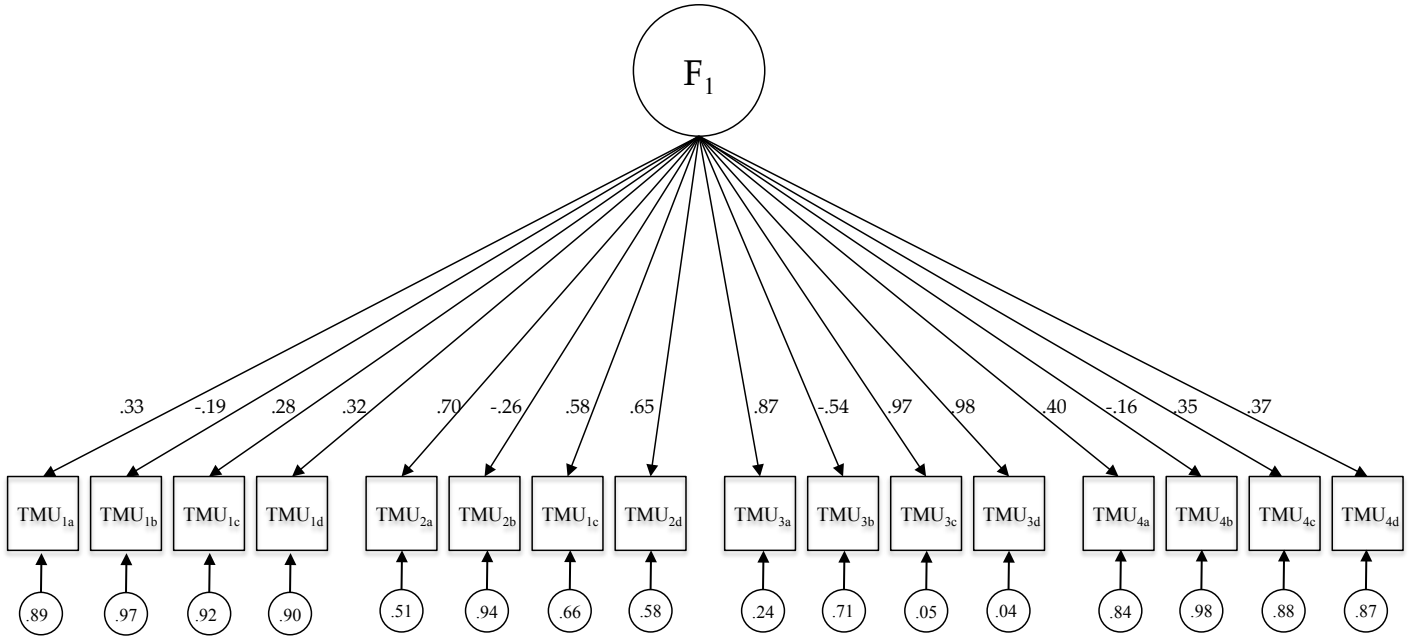

Figure 9. Model 1. TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d correspond to the four estimation techniques: D, Maas, MTLD, and MATTR.

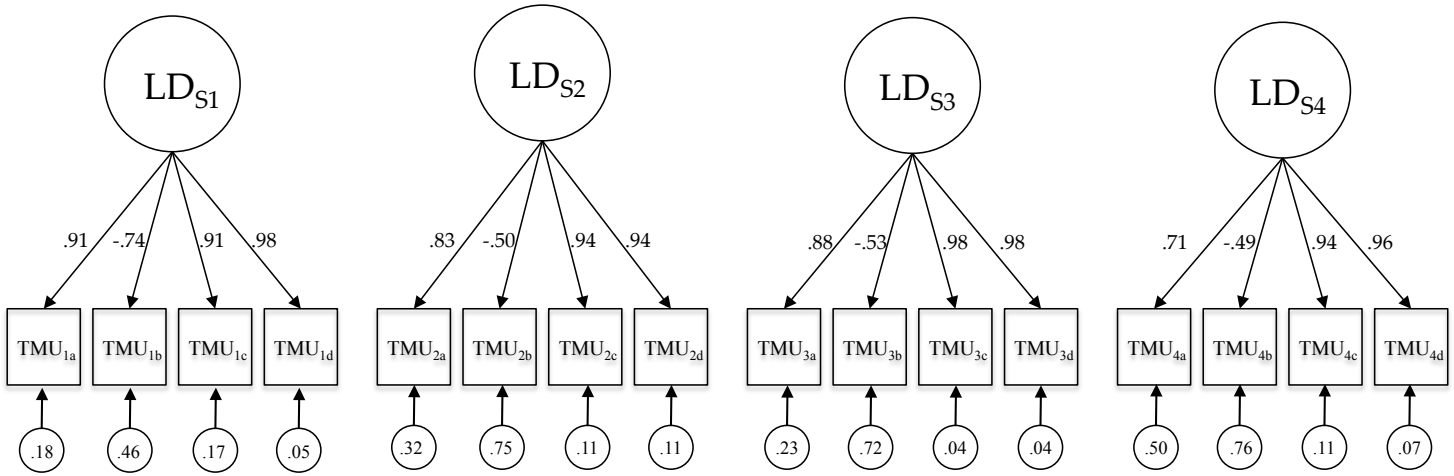

Figure 10. Model 2. TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d correspond to the four estimation techniques: D, Maas, MTLD, and MATTR.

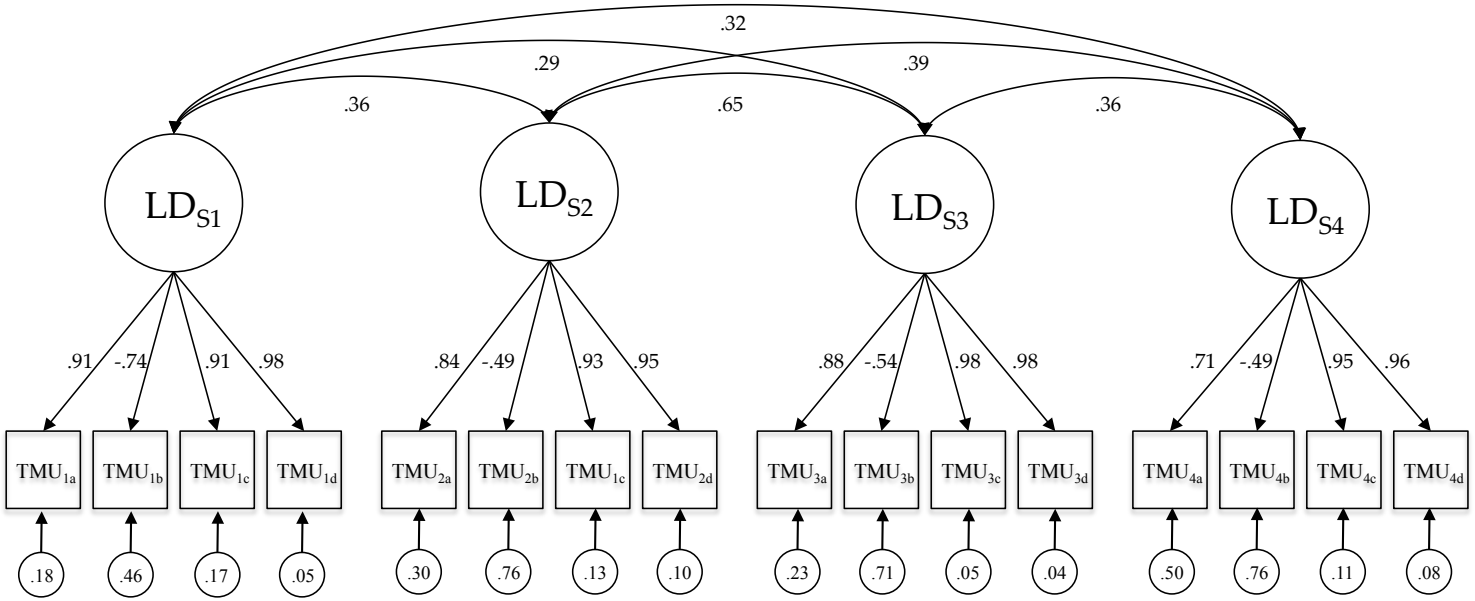

Figure 11. Model 3 TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d correspond to the four estimation techniques: D, Maas, MTLD, and MATTR.

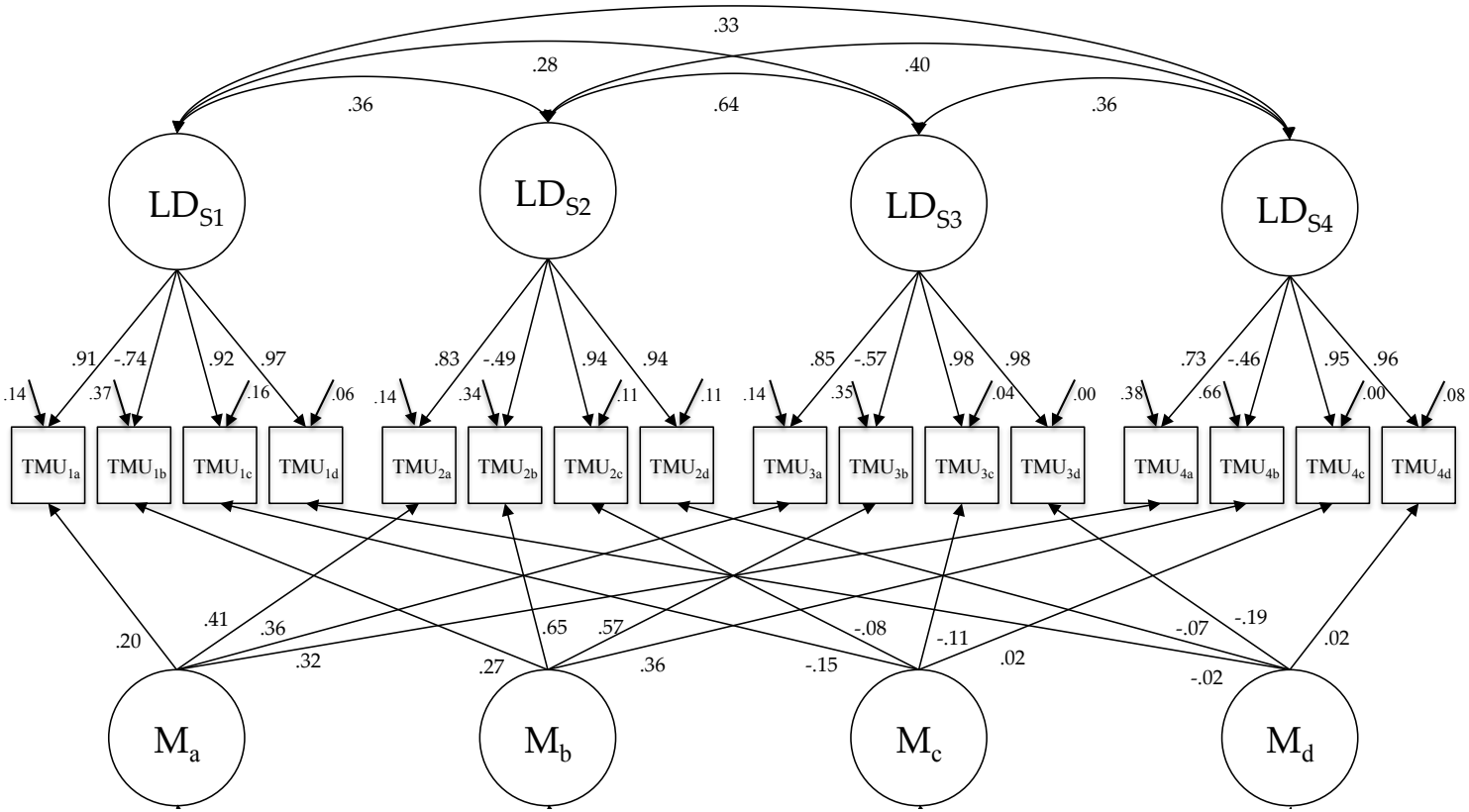

Figure 12. Model 4a. TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d correspond to the four estimation techniques: D, Maas, MTLD, and MATTR.

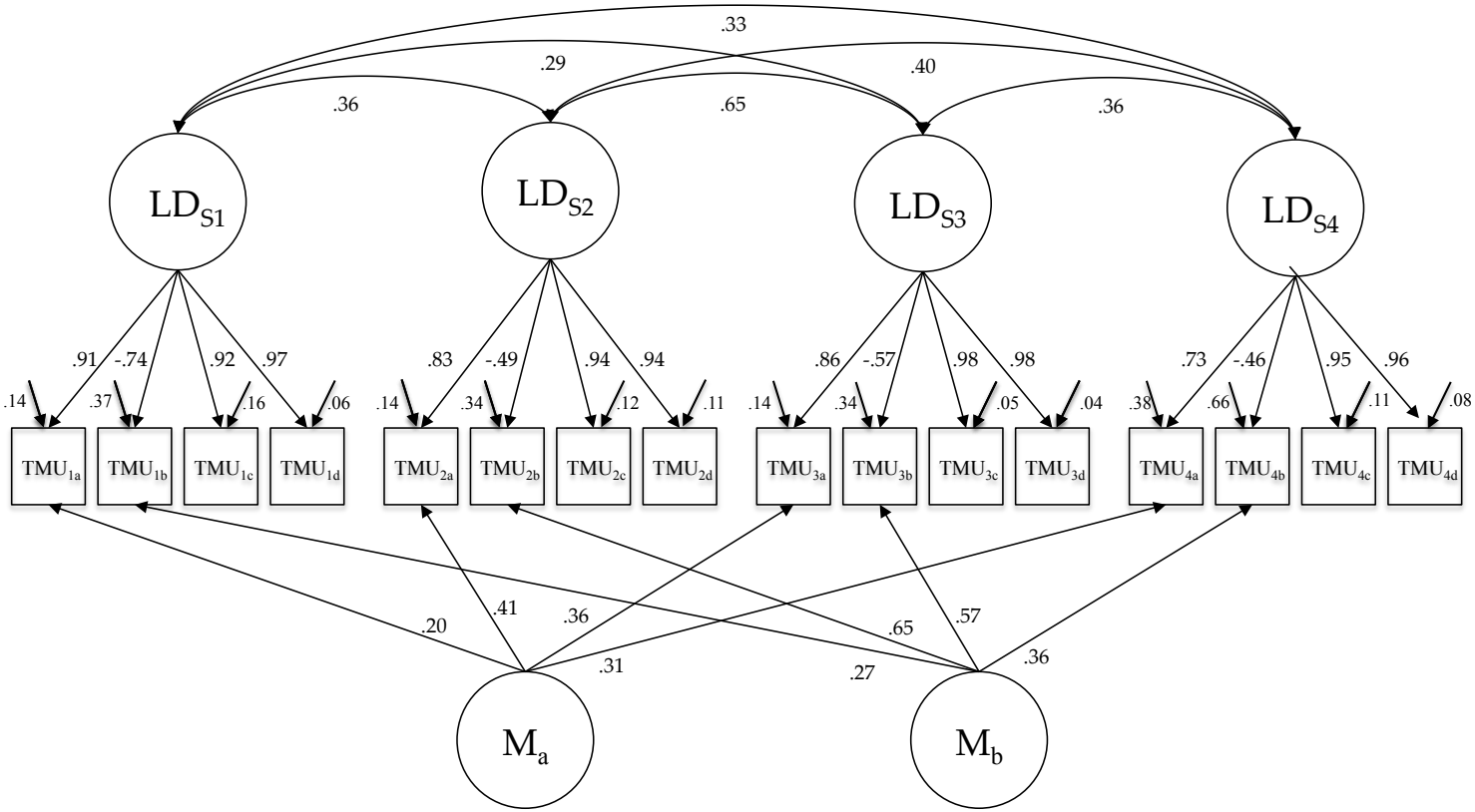

Figure 13. Model 5. TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d correspond to the four estimation techniques: D, Maas, MTLD, and MATTR.

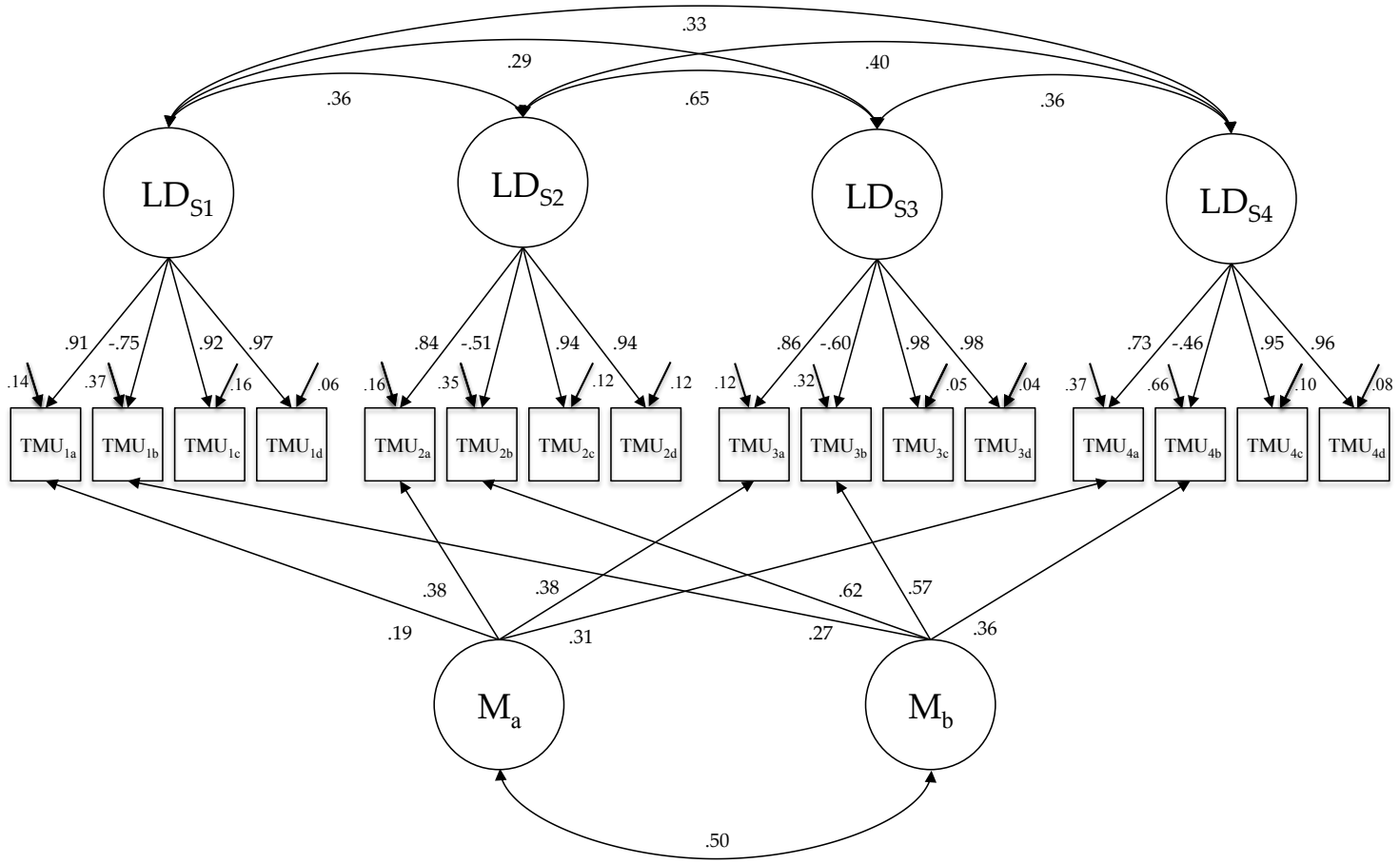

Figure 14. Model 6. TMU = TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d: D, Maas, MTLD, and MATTR, respectively.

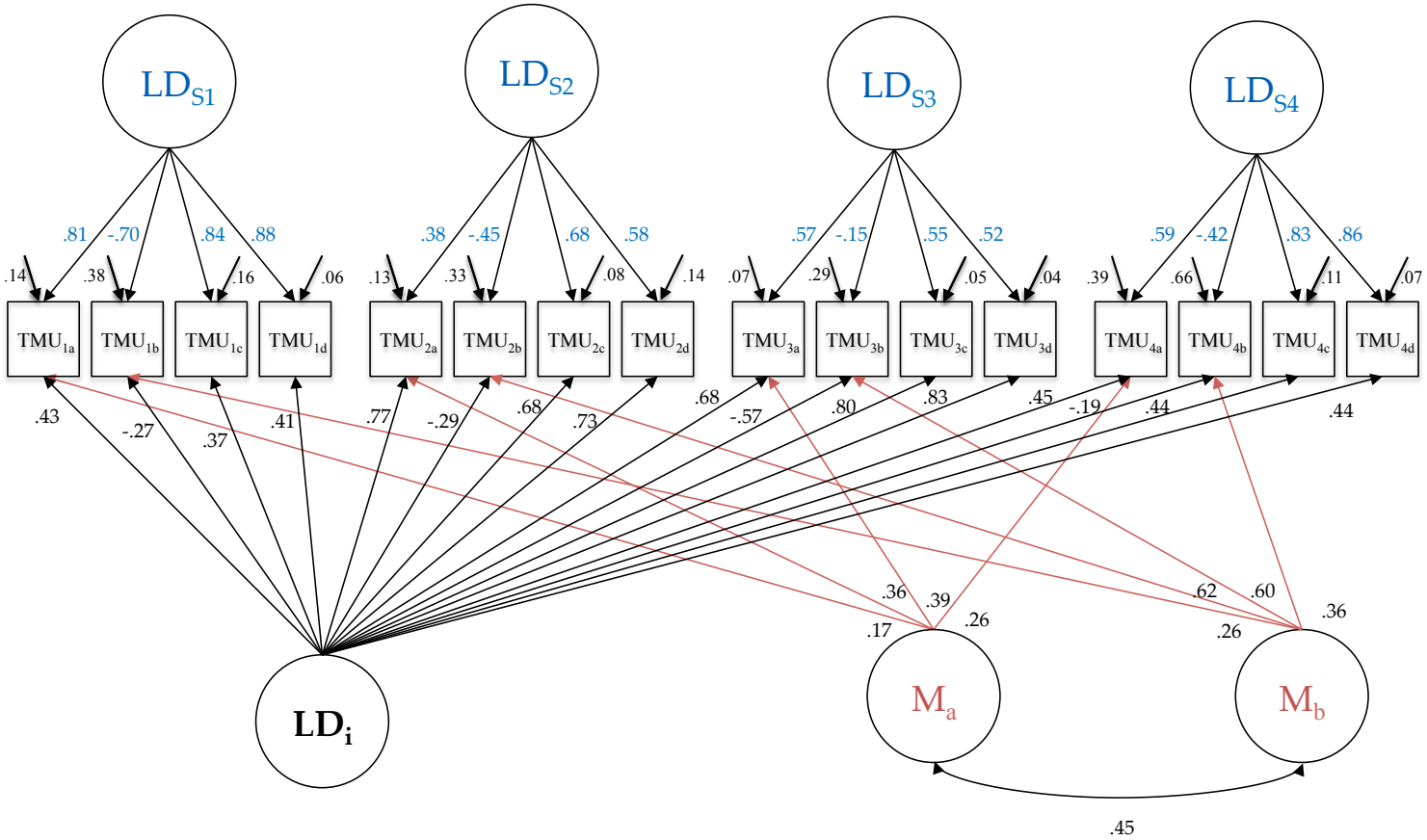

Figure 15. Model 7. TMU = TMU = Type-Method Unit. LD_i = Lexical diversity of the individual. LD_S = Lexical diversity of the sample. Numbers 1-4: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d: D, Maas, MTLD, and MATTR, respectively.

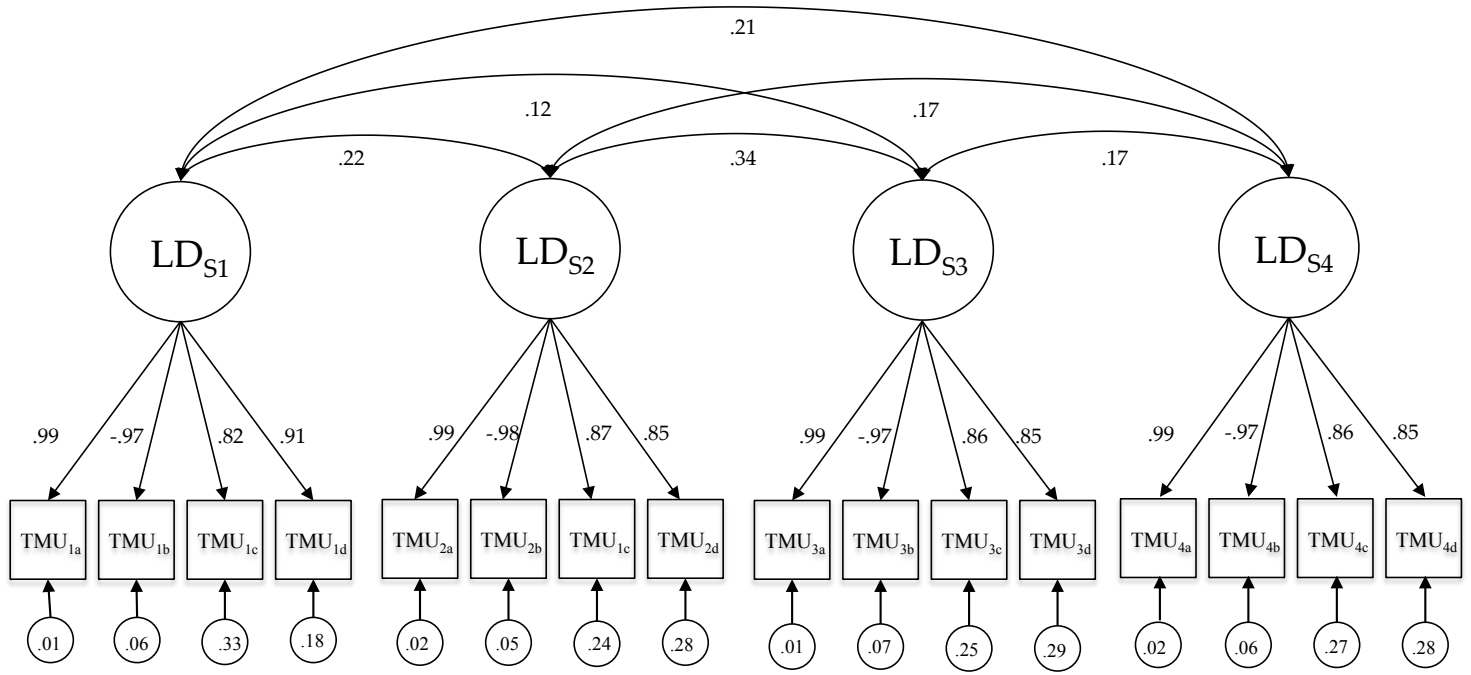

Figure 16. Model 3(2). TMU = Type-Method Unit. LD_S = Lexical diversity of the sample. Numbers 1-4 correspond to four types of discourse: Procedures, Eventcasts, Story-Telling, and Recounts, respectively. Letters a-d correspond to the four estimation techniques: D, Maas, MTLD, and MATTR.

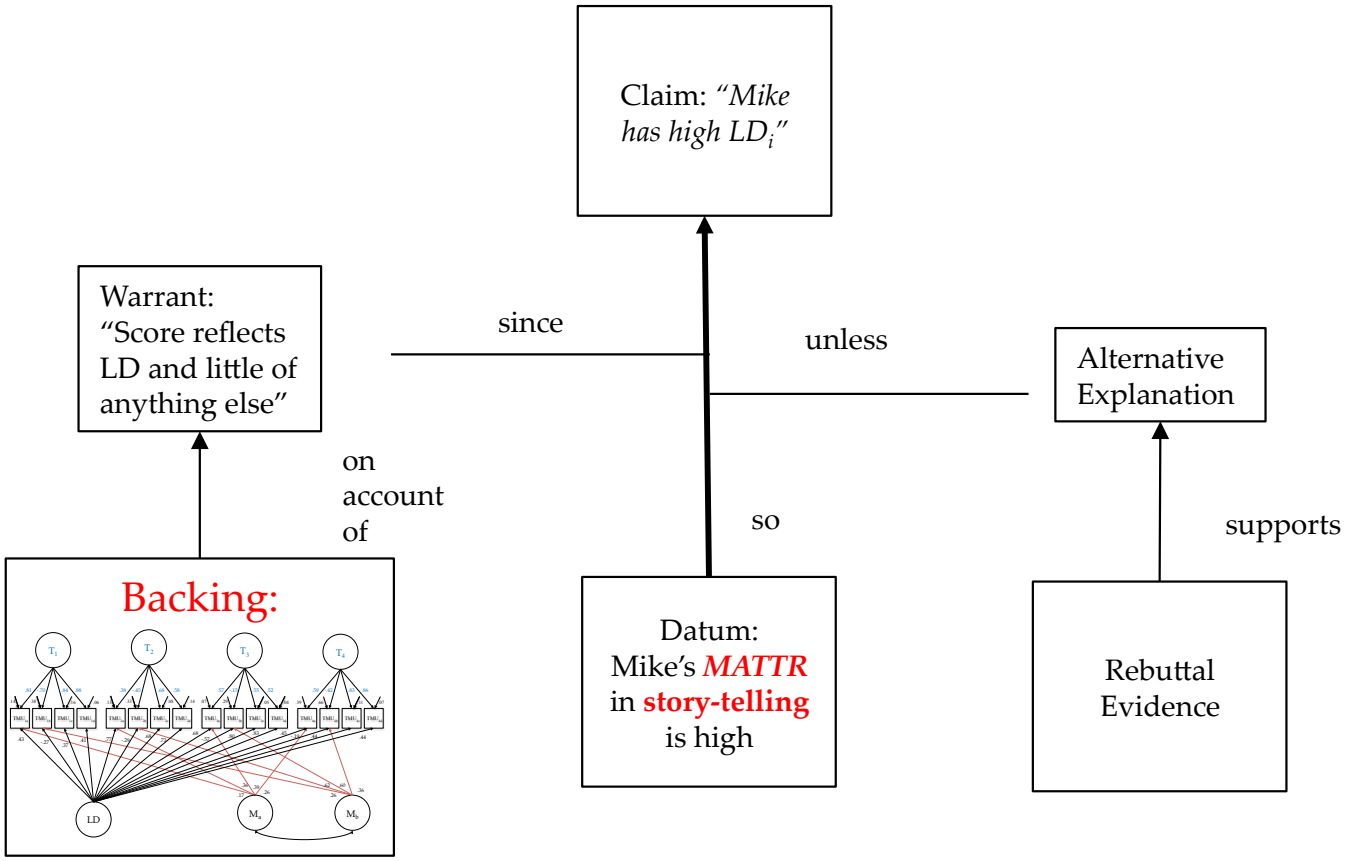

Figure 17. Toulmin's argument structure for the best combination of language sampling and estimation technique